Universidad
Carlos III de Madrid

Ingeniería Informática

PROYECTO FIN DE CARRERA

Sistema de predicción de resultados en eventos deportivos y su aplicación en las apuestas

Autor: Fernando Valera Guardiola

Tutor: David Griol Barres

Leganés, junio de 2013

Título: **Sistema de predicción de resultados en eventos deportivos y su aplicación en las apuestas**

Autor: **Fernando Valera Guardiola**

Director: **David Griol Barres**

EL TRIBUNAL

Presidente: _____

Vocal: _____

Secretario: _____

Realizado el acto de defensa y lectura del Proyecto Fin de Carrera el día __ de _____ de 20__ en Leganés, en la Escuela Politécnica Superior de la Universidad Carlos III de Madrid, acuerda otorgarle la CALIFICACIÓN de

VOCAL

SECRETARIO PRESIDENTE

Resumen

El Proyecto de Final de Carrera que se presenta puede dividirse en dos partes: la primera de ellas tiene como principal objetivo la **creación de un Sistema de Predicción** capaz de predecir el resultado más probable de un partido en un determinado evento deportivo. Después de desarrollar el Sistema de Predicción, se pasará a la segunda parte del proyecto, que será la encargada de la **explotación** de estas predicciones para obtener beneficios en las casas de apuestas. En la fase de explotación, se han utilizado tanto estrategias que son ya utilizadas por los aficionados a las apuestas, como nuevas estrategias que se han ideado para explotar al máximo las características del Sistema de Predicción que se ha desarrollado en la primera fase.

El sistema se centra en la predicción de resultados de dos deportes: **fútbol** (englobando varias competiciones tanto nacionales como internacionales) y **baloncesto** (considerando únicamente la liga americana de baloncesto, también conocida como NBA).

Para el desarrollo del Sistema de Predicción, que estará **basado en clasificadores**, se han recogido datos de un gran número de partidos de los deportes y competiciones nombrados anteriormente. A través de un proceso de **minería de datos** se ha determinado cuáles son los atributos que aportan mayor información para realizar la predicción.

Una vez se tienen todos los datos correctamente etiquetados, se ha utilizado la herramienta **WEKA** para crear diversos modelos de predicción basados en la clasificación. A través de esos modelos generados y tras recoger otra tanda de datos de los eventos deportivos, se ha analizado que modelos eran los que obtenían mejores tasas de acierto.

Para automatizar la tarea de predicción de los resultados, se ha creado un **archivo Excel**, con el que a través de varias **macros** que implementan los clasificadores seleccionados se calcula el resultado más probable del partido y el riesgo estimado que ese resultado lleva asociado.

Posteriormente, se definirán las **estrategias** que se van a seguir para intentar obtener beneficios en las casas de apuestas a través de las predicciones realizadas por el Sistema de Predicción. Entre estas estrategias se ha desarrollado un **Algoritmo Genético**, que a partir de los datos ofrecidos por el Sistema de Predicción es capaz de conseguir una combinación de apuestas para las que se maximiza el beneficio minimizando el riesgo.

Tras definir las estrategias e implementar el algoritmo en el que se basará una de ellas, se realizará un **estudio de beneficios** en el que se podrá ver qué estrategias son capaces de generar beneficios. Además, se podrá comprobar qué competiciones son las que mejores tasas de aciertos tienen y las que más beneficios son capaces de generar

Palabras clave: predicción, resultados, deporte, apuestas, clasificadores

Abstract

This project can be divided into two parts: the first one has as main objective the **creation of a Prediction System** capable of predicting the most likely result of a match in a particular sport event. After developing the Prediction System, we will go to the second part of the project, which will be responsible for the **exploitation** of these predictions for profit in betting. In the operational phase, will be used both strategies that are already used by the betting fans, as new strategies that have been designed to exploit the characteristics of the Prediction System which has been developed in the first phase.

The system focuses on two sports: **football** (considering several national and international competitions) and **basketball** (considering only American Basketball League, also known as NBA).

To develop the Prediction System, which will be **based on classifiers**, we have collected data from a large number of sports matches and competitions listed above. Through a **data mining** process, it has been determined what attributes provide more information for the prediction.

Once we have all the data correctly labeled, we used the **WEKA** tool to create different prediction models based on classification. Through these models generated and after picking up another round of sporting event data, it has been analyzed what models has the best hit rates.

To automate the task of prediction, we have created an **Excel file**, which through several **macros** which implement the selected classifiers, calculate the most likely result of the match and the estimated risk that is associated with that result.

Later, we will define the **strategies** that are going to follow to try to make profit in betting through the predictions made by the Prediction System. For one of these strategies we have developed a **Genetic Algorithm**, which is based on the data provided by the Prediction System, and is able to get a combination of bets for maximizing profit while minimizing risk.

After defining the strategies and implement the algorithm in which one of them will be based, we will make a **study of benefits** in which we can see which strategies are able to generate profits. Additionally, we can see which competitions have the best hit rates and the best profit ratio.

Keywords: prediction, results, sports, bets, classifiers

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

INTRODUCCIÓN

1.1 Introducción

Si hay un negocio que en estos tiempos no se resiente en nuestro país ese el de los juegos de azar, que respaldado por los avances tecnológicos y el boom de Internet de los últimos años, se ha expandido hasta alcanzar cifras de facturación en España que alcanzaron en 2011 los 370 millones de € generados por alrededor de 600.000 jugadores en todo el territorio nacional [1].

La popularidad de estos juegos de azar ha hecho que se creen innumerables páginas web que ofrecen este tipo de juegos a los usuarios para que prueben su suerte e intenten sacar beneficios a partir de una inversión inicial.

Entre todos los tipos de juegos de azar existentes, uno de los más populares son las apuestas deportivas. Las páginas especializadas en este tipo de apuestas ofrecen a los usuarios un gran número de eventos deportivos sobre los que puede realizar todo tipo de predicciones, desde resultados finales de un partido hasta predicciones sobre posiciones en la liga o campeones de torneos.

El gran abanico de posibilidades para apostar que ofrecen las páginas web de apuestas es el principal motivo por el que se pone en marcha este proyecto. Todo ese gran abanico de opciones para apostar puede ser aprovechado por los usuarios para elegir las opciones más seguras y así conseguir beneficios a partir de su inversión inicial.

El gran problema a resolver es cómo se puede saber con seguridad el resultado de un partido. La realidad es que no se puede saber, pero basándonos en la experiencia podríamos estimar si una apuesta que queremos realizar es segura o no. Así es como nace la idea de creación del Sistema de Predicción que se va a desarrollar en este proyecto. Ya que no podemos saber con seguridad el resultado final de un evento deportivo, se va a desarrollar un sistema que sea capaz de decirnos cuál es el resultado más probable de un evento deportivo y cómo de peligroso es apostar por él.

Para desarrollar este sistema se utilizarán en su fase inicial técnicas de minería de datos, que permitirán a partir de una recogida masiva de datos y un posterior procesado de éstos, la extracción de conclusiones que nos aportarán la información necesaria para crear nuestros modelos de predicción basados en clasificadores.

Una vez se tengan desarrollados y probados los modelos de predicción, se pondrán en marcha diferentes estrategias para aprovechar los resultados que arrojen nuestros modelos. Estas estrategias irán desde la utilización de técnicas habituales desarrolladas por apostantes expertos hasta la creación de nuevas estrategias que exploten al máximo las cualidades de los modelos de predicción que han sido generados.

Finalmente, cuando tengamos definidos tanto los modelos de predicción como las estrategias de apuesta, sólo quedará probar la efectividad del sistema y si éste es capaz de generar beneficios a partir de una inversión inicial.

1.2 Objetivos

Al hablar de los objetivos de este proyecto se pueden diferenciar muy claramente dos grandes objetivos, que se pasan a detallar a continuación:

- **Desarrollo de un Sistema de Predicción para Eventos Deportivos:**

El primero de los objetivos a alcanzar en este proyecto es la creación de un **Sistema de Predicción** capaz de predecir el resultado final en un evento deportivo. Este sistema será la base del proyecto, ya que a partir de los resultados predichos por éste, se aplicarán varias estrategias de apuestas con las que se intentará obtener beneficios en las casas de apuestas

Para el desarrollo del Sistema de Predicción se aplicarán técnicas de minería de datos, que nos permitirán a partir de unos datos observados generar una serie de modelos de predicción que estarán basados en clasificadores.

En esta fase del proyecto, la herramienta principal que nos ayudará a analizar los datos y a ver qué clasificadores generan unos mejores modelos predictivos será **WEKA (Waikato Environment for Knowledge Analysis)**[2][3].

- **Generación de estrategias para la explotación del Sistema de Predicción:**

Tras la fase de desarrollo del sistema, después de la cual dispondremos de un sistema capaz de predecir el resultado más probable de un evento deportivo, se pasará a la fase de explotación. En esta fase se aplicarán los resultados predichos por el sistema a distintas **estrategias de apuesta** para intentar conseguir beneficios en las casas de apuestas. En el conjunto de estrategias que se van a utilizar, se han escogido tanto estrategias ya utilizadas por los aficionados a las apuestas, como nuevas estrategias ideadas para explotar al máximo las características del sistema de predicción desarrollado.

Entre las nuevas estrategias que se van a utilizar en esta fase del proyecto, podemos destacar el desarrollo de un **Algoritmo Genético** que se encargará de generar una combinación de apuestas que maximice el beneficio minimizando el riesgo. El algoritmo se nutrirá de los resultados predichos por el sistema de predicción para, utilizando su función de evaluación de apuestas, generar la combinación de apuestas que nos permita obtener el mayor porcentaje de rentabilidad sobre nuestras apuestas posible.

Una vez evaluadas todas las estrategias consideradas, para finalizar y sacar las conclusiones de este proyecto se verificará si estas estrategias son capaces de generar beneficios a partir de los resultados que ha predicho el sistema.

1.3 Planificación

La planificación de este proyecto se divide en seis tareas principales que abarcan desde la fase de planificación y diseño hasta la fase de implementación y test del sistema generado. Según la planificación general del proyecto, éste debería ser desarrollado en un total de ocho meses, teniendo como fecha de inicio el 24/09/2012 y como fecha de fin el 23/05/2013.

Las principales tareas en las que se divide el proyecto son las siguientes:

- **Preliminar:** Fase donde se estudiará la viabilidad del proyecto y donde se recopilará y analizará documentación de sistemas similares al que se quiere desarrollar.
- **Diseño:** Fase en la que se diseña el sistema de predicción y se establecen los datos necesarios para realizar el estudio con clasificadores.
- **Recogida de Datos:** Fase en la que se recopilan los datos de todos los partidos de las competiciones a estudiar para después utilizarlos en el entrenamiento con clasificadores.
- **Desarrollo:** Fase en la que se implementa tanto el Sistema de Predicción como el Algoritmo Genético que será utilizado en las estrategias de apuestas.
- **Estudio de Beneficios:** Fase en la que se verifica que las estrategias utilizadas para aprovechar las predicciones del sistema generan beneficios.
- **Documentación:** Fase en la que se genera la documentación requerida por el proyecto.

A continuación se puede ver en la Figura 1 la planificación del proyecto.

	Nombre de tarea	Duración	Comienzo	Fin	Predecesoras
	Fase Preliminar	**19 días**	**lun 24/09/12**	**jue 18/10/12**	
	Búsqueda y Análisis de Documentación asociada al proyecto	15 días	lun 24/09/12	vie 12/10/12	
	Estudio de Viabilidad	2 días	lun 15/10/12	mar 16/10/12	2
	Recopilación de herramientas necesarias para el proyecto	2 días	mié 17/10/12	jue 18/10/12	3
	Fase Diseño	**3 días**	**vie 19/10/12**	**mar 23/10/12**	**1**
	Diseño del Sistema de Predicción	3 días	vie 19/10/12	mar 23/10/12	
	Diseño de la Fase de Recogida de Datos	3 días	vie 19/10/12	mar 23/10/12	
	Fase de Recogida de Datos	**83 días**	**mié 24/10/12**	**vie 15/02/13**	**5**
	Recogida de datos de los partidos	83 días	mié 24/10/12	vie 15/02/13	
	Adecuación de los datos al fichero de entrenamiento	83 días	mié 24/10/12	vie 15/02/13	
	Fase de Desarrollo	**99 días**	**lun 29/10/12**	**jue 14/03/13**	**5**
	Selección e Implementación de Sistemas de Apuesta	**7 días**	**lun 29/10/12**	**mar 06/11/12**	
	Selección de Estrategias de Apuesta	3 días	lun 29/10/12	mié 31/10/12	
	Adecuación de los sistemas de apuestas al proyecto	4 días	jue 01/11/12	mar 06/11/12	13
	Desarrollo del Algoritmo Genético	**71 días**	**mié 07/11/12**	**mié 13/02/13**	**12**
	Estudio de las características del algoritmo	7 días	mié 07/11/12	jue 15/11/12	
	Implementación	60 días	vie 16/11/12	jue 07/02/13	16
	Ajuste de parametría del algoritmo	2 días	vie 08/02/13	lun 11/02/13	17
	Pruebas	2 días	mar 12/02/13	mié 13/02/13	18
	Desarrollo Hoja de Predicción	**20 días**	**vie 15/02/13**	**jue 14/03/13**	**8**
	Creación de Conjuntos de Entrenamiento	1 día	vie 15/02/13	vie 15/02/13	
	Entrenamiento de conjuntos con WEKA	8 días	lun 18/02/13	mié 27/02/13	21
	Selección de Mejores Clasificadores	1 día	jue 28/02/13	jue 28/02/13	22
	Implementación de Modelos de Predicción	10 días	vie 01/03/13	jue 14/03/13	23
	Estudio de Beneficios	**65 días**	**lun 18/02/13**	**vie 17/05/13**	
	Recogida de datos para el estudio	12 sem.	lun 18/02/13	vie 10/05/13	
	Evaluación de las estrategias de apuesta	3 días	lun 13/05/13	mié 15/05/13	26
	Cálculo de beneficios	2 días	jue 16/05/13	vie 17/05/13	27
	Documentación del Proyecto	4 días	lun 20/05/13	jue 23/05/13	25

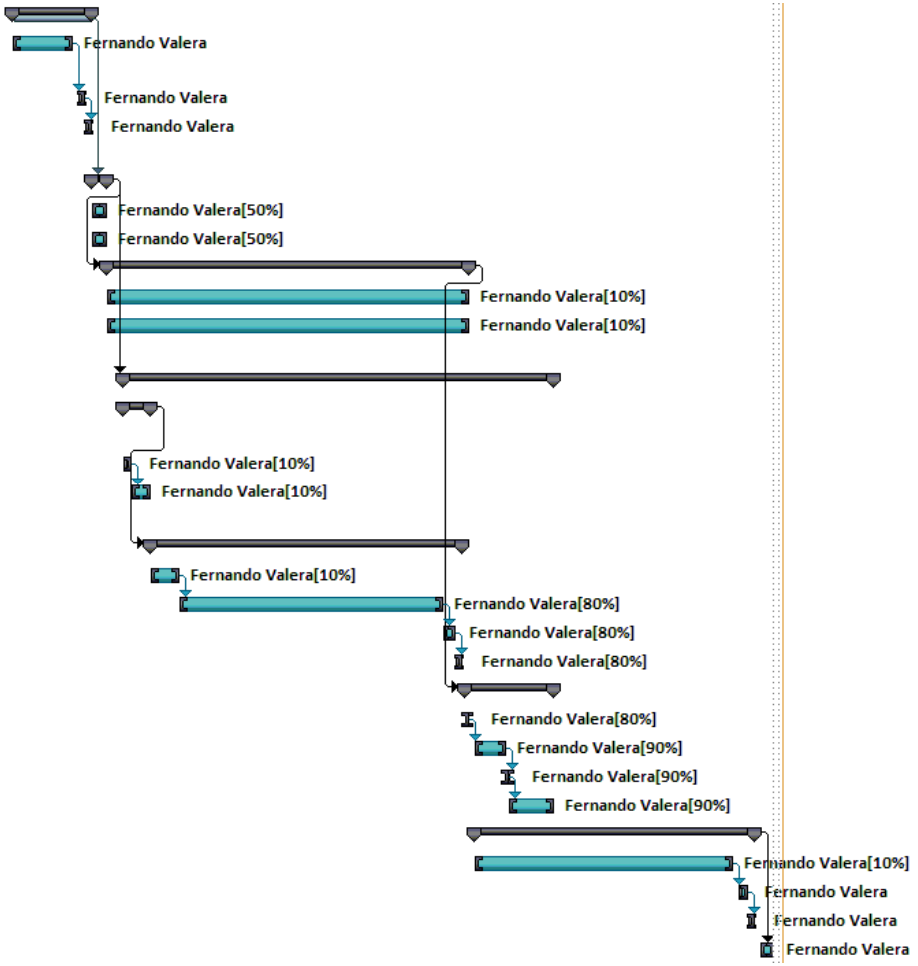

Figura 1. Planificación del Proyecto

1.4 Presupuesto

En este apartado se van a detallar cada una de las partidas presupuestarias que ha tenido el proyecto durante su desarrollo. Los costes se irán agrupando en diversas categorías como se muestra a continuación.

Costes de Personal:

Esta partida presupuestaria hace referencia a los gastos en forma de salarios que hay que pagar al personal asociado al proyecto.

Para establecer los meses de dedicación de cada uno de los intervinientes en el proyecto, se ha seleccionado el periodo total del proyecto para el caso del Ingeniero Junior (el alumno que realiza el proyecto) y un periodo total de un mes para el Ingeniero Senior (el tutor del proyecto). Aunque el Ingeniero Senior realiza tareas durante toda la duración del proyecto se ha decidido en términos de dedicación agrupar todas esas tareas en un mes de trabajo.

Para los costes por mes de cada uno de los intervinientes, se ha tomado un sueldo medio de 20.000€/año para el Ingeniero Junior, que trabaja un total de 40 horas semanales. Para el caso del Ingeniero Senior se ha calculado su coste a partir de un sueldo medio de 12.000€/año, teniendo en cuenta que trabaja 20 horas semanales.

Tras realizar los cálculos asociados al Coste de Personal, que pueden ser consultados en la Tabla 1, podemos concluir que los Costes de Personal del proyecto ascienden a **14.333,28€**

Apellidos y Nombre	N.I.F.	Categoría	Dedicación (hombres mes)	Coste hombre mes	Coste (Euro)
Valera Guardiola, Fernando		Ingeniero Junior	8	1.666,66	13.333,28
Griol Barres, David		Ingeniero Senior (Apoyo, Coaching)	1	1.000,00	1.000,00
				Total	**14.333,28**

Tabla 1. Costes de Personal

Costes asociados al Hardware:

Esta partida presupuestaria hace referencia al coste derivado de la utilización de equipos como ordenadores, impresoras o escáners que han sido usados durante el proyecto

En la Tabla 2 se realiza un desglose del material hardware utilizado para el desarrollo del proyecto. El coste imputado será calculado en función de la amortización del dispositivo. En este caso, al tratarse de hardware se ha decidido que su amortización se realice sobre un periodo de 60 meses.

A continuación se muestra la fórmula utilizada para el coste amortizado. Esta fórmula será utilizada para el cálculo de los costes asociados al hardware y software usado durante el proyecto.

$$\frac{A}{B} \text{xCxD}$$

A = nº de meses desde la fecha de facturación en que el equipo es utilizado

B = periodo de depreciación (X meses)

C = coste del equipo (sin IVA)

D = % del uso que se dedica al proyecto (habitualmente 100%)

Descripción	Coste (Euro)	% Uso dedicado proyecto	Dedicación (meses)	Periodo de depreciación	Coste imputable
Ordenador Portátil	592,09	100	8	60	78,94
Impresora	158,00	100	8	60	21,06
				Total	**100,00**

Tabla 2. Costes asociados al hardware

Costes asociados al Software:

Los costes asociados esta partida, derivan de la utilización de software durante la realización del proyecto. Se tendrán en cuenta todas las herramientas software utilizadas para el desarrollo del proyecto. En este caso, el periodo de depreciación se calculará sobre 36 meses.

Descripción	Coste (Euro)	% Uso dedicado proyecto	Dedicación (meses)	Periodo de depreciación	Coste imputable
WEKA	0	100	8	36	0
MS Office Professional 2010	402,11	100	8	36	89,35
MS Project 2010	0	100	8	36	0
				Total	**89,35**

Tabla 3. Costes asociados al software

Costes de funcionamiento del proyecto:

Esta partida presupuestaria deriva de diversos costes que surgen durante la realización del proyecto. Estos costes hacen referencia a material de oficina, transporte o dietas entre otros. El desglose de esta partida puede verse en la Tabla 4.

Descripción	Empresa	Costes imputable
Material Fungible		40,00
Desplazamientos		90,00
Cartuchos Impresora		35,00
Dietas		30,00
	Total	**195,00**

Tabla 4. Costes de funcionamiento del proyecto

Resumen Final:

Para concluir el apartado presupuestario del proyecto se incluye el resumen final del presupuesto. En él se incluyen todas las partidas que fueron desglosadas en las Tablas 1, 2, 3 y 4 además de otra partida de costes indirectos. Estos costes indirectos hacen referencia a costes derivados de la realización del proyecto, como pueden ser el alquiler de un local de trabajo, recibos de agua y luz, o costes que surgen y que no habían sido previstos al iniciar el proyecto. Para el cálculo de estos costes se ha supuesto que los costes indirectos ascienden a un 15% del total del proyecto.

Como se ve en la Tabla 5, el presupuesto final asciende a **16.925,27€**

Partidas	Total
Personal	14.333,28
Amortización	189,35
Subcontratación de tareas	0,00
Costes de funcionamiento	195,00
Total	**14.717,63**
Costes Indirectos	2.207,64
Total	**16.925,27**

Tabla 5. Resumen del Presupuesto Total

1.5 Material Empleado

Para la realización de este proyecto se han utilizado diversos equipos y materiales, que se van a pasar a detallar a continuación:

- **Ordenador Portátil:**

Este es el equipo principal del proyecto, y en él se realizarán la gran mayoría de las tareas que han sido listadas en la planificación del proyecto (Figura 1).

Las características del equipo son las siguientes:

- Procesador AMD E-300 de doble núcleo 1,3 GHz
- 1Gb de Memoria RAM
- Disco duro 320 GB (5400 rpm S-ATA)
- Tarjeta gráfica AMD Radeon HD 6310M
- Conectividad 802.11 b/g/n

- **Impresora:**

La impresora ha sido un elemento de vital importancia a la hora de la implementación de los algoritmos de clasificación, ya que la impresión en papel tanto de los árboles de decisión como de los coeficientes de la red bayesiana ha agilizado el proceso de implementación.

Las características del equipo son las siguientes:

- Impresión por tinta térmica a petición
- Lenguaje de comandos PLC de HP Nivel 3
- Conexión puerto paralelo
- 8 páginas por minuto (Texto en Negro)

- **Microsoft Office 2010:**

Para el desarrollo de la Hoja de Predicciones será imprescindible la instalación de Microsoft Office 2010, ya que dentro del paquete de aplicaciones trae consigo Microsoft Excel, que será la herramienta sobre la que se desarrollará la hoja de predicciones.

Además, para la elaboración de la documentación del proyecto también será necesario tener instalado Microsoft Word, el cual también está incluido en el paquete Office.

- **Eclipse:**

Para el desarrollo del algoritmo genético, se ha instalado en el ordenador el entorno de desarrollo Eclipse [4]. Mediante este entorno seremos capaces de programar las clases y funciones diseñadas para el algoritmo. Además, al poseer consola de ejecución podremos realizar las pruebas dentro del propio entorno.

Capítulo 2

ESTADO DEL ARTE

2.1 Introducción

El sistema que se ha desarrollado en este proyecto se encarga de realizar predicciones a través de la clasificación de instancias. En el dominio concreto de la predicción de resultados en eventos deportivos, no se han encontrado trabajos que intenten realizar la misma tarea basándose en principios similares a los que se van a utilizar en este proyecto. No obstante, se pueden encontrar estudios similares aplicados a otras disciplinas de conocimiento, desde predicciones meteorológicas a predicciones sobre el desarrollo de enfermedades como puede ser el cáncer.

Este capítulo tiene como principal objetivo resumir los trabajos más destacados en este campo de investigación, en el que se utilizan técnicas de análisis de datos y la creación de clasificadores para realizar labores de diagnóstico o predicción.

Además, se presentarán los proyectos y sistemas más destacados que actualmente se dedican a la predicción de resultados deportivos y a la obtención de beneficios mediante el uso de las apuestas en casas de apuesta online. Este apartado nos servirá para ver el estado actual de sistemas que realizan labores similares al que se quiere desarrollar en este proyecto y para ver las técnicas utilizadas por dichos sistemas para realizar las predicciones.

2.2 Sistemas de Predicción

Según la definición que podemos encontrar en el diccionario de la RAE [5], predecir es *anunciar por revelación, ciencia o conjetura algo que ha de suceder*.

Si tenemos en cuenta la definición que aparece en el diccionario, un sistema de predicción será un sistema que se encarga de realizar un cálculo anticipado de un hecho que va a suceder a partir unas observaciones previas o indicios.

Posiblemente no nos demos cuenta, pero la importancia de estos sistemas es muy grande para el desarrollo de nuestra vida cotidiana, tal y como puede observarse en el gran número de aplicaciones mostradas en las siguientes subsecciones.

2.2.1 Sistemas de Predicción en la vida cotidiana

- **Sistema de predicción de la demanda eléctrica**

Uno de los problemas más importantes de la producción de energía eléctrica es que en todo momento la cantidad de energía producida tiene que estar igualada a la energía demandada. Esto es debido a que la energía eléctrica no se puede almacenar en grandes cantidades, haciendo que en todo momento producción y demanda tengan que estar equilibradas para no provocar problemas en la red [6].

Debido a este problema, es muy importante tener una estimación por día y franja horaria que permita coordinar y gestionar qué cantidad de energía deberá producir cada una de las centrales del parque energético para cubrir las demandas previstas.

En el caso de España, la empresa *Red Eléctrica de España S.A.,* encargada del transporte de la electricidad y operador del sistema del país, debe asegurar la continuidad y seguridad en el suministro eléctrico, manteniendo en constante equilibrio la generación y el consumo eléctrico del país [7].

Las predicciones que realiza la empresa pueden ser vistas en cualquier momento a través de una herramienta gráfica que muestra tanto el consumo previsto como el real, además de la producción programada para cada tramo horario [8] (Figura 2).

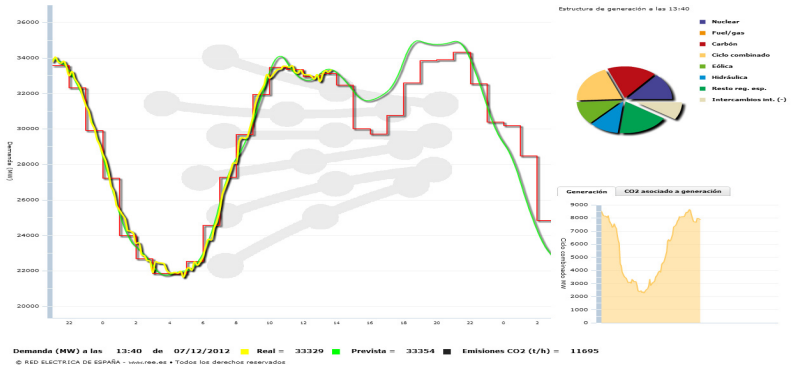

Figura 2. Previsión y evolución de la demanda eléctrica por franja horaria

- **Sistema de Predicción Meteorológico**

Otro de los sistemas de predicción que se usan en la actualidad para informar sobre el tiempo que habrá en una determinada zona en un momento determinado son los sistemas de predicción meteorológica.

Estos sistemas se basan en complejos modelos matemáticos que simulan el comportamiento de la atmósfera dadas unas condiciones previas que han sido observadas por los satélites enviados por las agencias meteorológicas.

En este dominio de aplicación, sistema de predicción puede entenderse como un programa informático que genera una serie de valores atmosféricos que corresponden a un momento y zona determinados. Estos valores se calculan a través de complejas ecuaciones matemáticas que emulan procesos físicos y dinámicos de la atmósfera. Al ser un sistema que utiliza ecuaciones no lineales, los resultados obtenidos sobre el estado de la atmósfera en un lugar y momento determinado son aproximados, lo que hace que la predicción meteorológica no sea fiable al 100%

De todos los modelos de predicción existentes, el más extendido y el que ofrece unos mejores resultados es el GFS (Global Forecast System) [9]. El GFS es un modelo que se actualiza 4 veces al día y que realiza predicciones que alcanzan los 16 días. La fiabilidad de la predicción va decreciendo según la lejanía del día de la predicción, siendo 7 el número de días que se utilizan para catalogar una predicción como fiable.

La fiabilidad de este modelo hace que esté entre los modelos de predicción más usados en todo el mundo. Además, es un modelo con cobertura global que ofrece gratuitamente las salidas de su predicción para que agencias estatales de meteorología los usen para realizar sus predicciones.

En concreto, la Agencia Estatal de Meteorología (AEMET) utiliza estos resultados para elaborar sus predicciones. Este servicio es clave para la seguridad civil, ya que a partir de esta información se generan informes que pueden alertar de situaciones climáticas anómalas que requieran movilizar a los cuerpos y fuerzas de seguridad del estado. Además, se generan informes que son enviados a los organismos de navegación aérea y marítima para que regulen el tráfico en función de las previsiones climatológicas [10].

- **Sistema de Predicción Sísmico**

Otro tipo de sistema de predicción en el que se está trabajando actualmente es el que realiza predicciones sobre terremotos. Actualmente los sistemas que realizan este tipo de predicciones no están muy avanzados, ya que es muy difícil saber con precisión cuándo y dónde se va a producir un terremoto.

Actualmente, para evaluar posibles zonas en las que podría ocurrir un terremoto se utilizan los mapas de evaluación de peligro sísmico, que son mapas que representan la probabilidad de que ocurra un terremoto en una zona determinada. Estos estudios se realizan a partir de registros históricos de terremotos y localización de las fallas, lo que hace que se pueda ver el riesgo que tiene una determinada zona [11].

El problema que tiene este método de predicción es que tan sólo nos indica las zonas de riesgo. En cuanto al periodo de tiempo en el que ocurriría el terremoto, sólo se indica un intervalo de tiempo en el que se estima un terremoto de determinada magnitud. La Figura 3 muestra un ejemplo de este tipo de sistemas. En este caso se puede observar el estudio de probabilidad de terremoto en la Bahía de San Francisco de Estados Unidos.

En la figura se representan las fallas que están presentes en la geografía de la bahía, y éstas están coloreadas con colores más o menos vivos teniendo en cuenta la probabilidad de que esa falla provoque un terremoto de una magnitud de 6,7 o más en los próximos treinta años. Además, en color turquesa aparecen las áreas urbanas cercanas a las fallas, y que en caso de un gran terremoto sufrirían grandes daños.

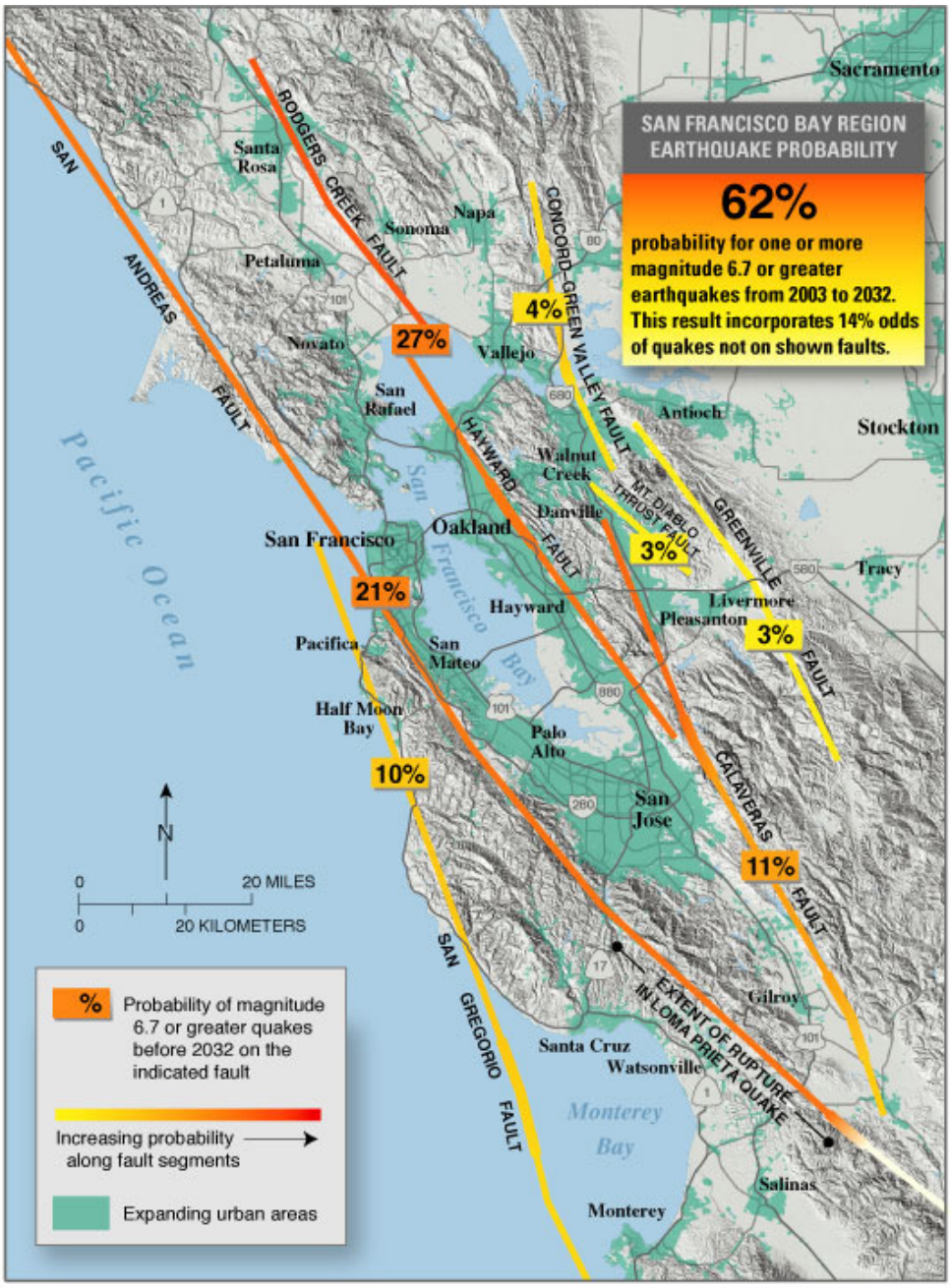

Figura 3. Mapa de evaluación de Riesgo Sísmico

Actualmente hay una gran cantidad de proyectos en curso, todos ellos intentando mejorar la precisión tanto geográfica como temporal del terremoto. Uno de los casos más sonados y que tuvo lugar hace pocos años fue la predicción realizada por el geólogo Giampaolo Giuliani en la ciudad italiana de L'Aquila. El geólogo y su equipo detectaron concentraciones elevadas de gas radón en zonas sísmicas activas cercanas a la ciudad, lo que para ellos, era un claro indicio de que podía haber un terremoto inminente.

Días después de la predicción, un terremoto de magnitud 6.3 sacudió la ciudad de L'Aquila, causando 308 muertos y 1500 heridos. Este caso ha sido uno de los más nombrados en los últimos años, ya que las autoridades italianas denunciaron al geólogo días antes de que el terremoto tuviera lugar por intentar crear una falsa alarma [12].

Después de este terremoto y tras las predicciones realizadas, una de las vías de investigación y desarrollo de sistemas de predicción sísmicos, es la detección de altas concentraciones de radón en zonas sísmicas activas para realizar las predicciones.

2.2.2 Sistemas de predicción basados en clasificadores

Tras describir varios tipos de sistemas de predicción que se usan a diario y que son de vital importancia para el desarrollo de nuestra vida cotidiana, nos vamos a centrar en el estudio de sistemas de predicción basados en clasificadores, que es el tema en el que se engloba este proyecto.

En la actualidad podemos encontrar un gran número de clasificadores de diversos tipos: basados en Reglas, en el Teorema de Bayes, en Redes de Neuronas, y así una larga lista en la que cada uno de los tipos tiene unas características específicas [13].

En este apartado se enumeran varios proyectos relacionados con la obtención de predicciones haciendo uso de clasificadores. Tal y como se describirá, alguno de ellos utilizan las mismas herramientas empleadas en el desarrollo del proyecto que se presenta, por lo que podremos ver las similitudes y diferencias a la hora de usar un tipo de clasificador u otro.

- **Sistema de diagnóstico preventivo del cáncer**

Este sistema presentado en [14] creado por investigadores de la Universidad Técnica Particular de Loja (Ecuador), tiene como objetivo mejorar el diagnóstico en fases iniciales del cáncer. La motivación para el desarrollo de este sistema viene debido al estancamiento que se había producido en el número de supervivientes de cáncer. Este estancamiento se debía en gran medida a la imposibilidad de tener un diagnóstico precoz que hiciera posible empezar un tratamiento contra la enfermedad en las fases iniciales, donde hay más probabilidades de cura.

El sistema proporciona una primera alerta y determina si un individuo, a través de datos como el entorno en el que vive, sus hábitos o antecedentes familiares, es propenso a desarrollar cáncer de pulmón. Si el sistema determinara que una persona es propensa a desarrollar este tipo de cáncer se le comenzarían a realizar chequeos periódicos para que en el caso de que la enfermedad acabe desarrollándose.

Para la generación del sistema de diagnóstico preventivo del cáncer se optó por la utilización de las Redes Bayesianas [15], que es un modelo de clasificación muy usado en el diagnóstico asistido por ordenador. El problema que tiene este modelo es que suele estancarse en mínimos locales, por lo que para la optimización de este sistema se optó por combinar la utilización de las Redes Bayesianas con Algoritmos Evolutivos. Estos algoritmos proporcionarían una mayor precisión en el diagnóstico, ya que son perfectos para las búsquedas globales basadas en grandes poblaciones.

Los datos utilizados para la generación de los modelos probabilísticos fueron atributos concretos de las personas, como puede ser su edad, sexo, factores genéticos o antecedentes familiares. A partir de estos datos el sistema aprende diversas Redes Bayesianas a partir del Algoritmo genético, lo que provocará que según se vayan creando nuevas generaciones se vayan obteniendo redes que sean capaces de realizar un mejor diagnóstico.

Los datos finales de este estudio arrojan que el sistema es capaz de realizar diagnósticos más precisos que los métodos tradicionales que se vienen utilizando hasta ahora, lo que proporcionará un mejor diagnóstico preventivo y por lo tanto un mayor porcentaje de supervivientes de esta enfermedad.

- **Algoritmos para la clasificación de correo electrónico no deseado**

En [16] se presenta un estudio cuyo principal objetivo es comparar diversos clasificadores para comprobar su eficacia a la hora de clasificar correo electrónico no deseado. La motivación para el desarrollo de este proyecto vino dada por la explosión que supuso la aparición del correo electrónico en nuestras vidas. Este sistema de comunicación era rápido y con un coste bajo, lo que hizo que fuera un medio idóneo para la publicidad comercial.

Este tipo de publicidad comercial ha llegado a ser un auténtico quebradero de cabeza para los usuarios habituales del correo electrónico, que veían como todos los días les llegaban decenas de correos con publicidad comercial. Este hecho ha desencadenado que durante todo este tiempo se hayan desarrollado potentes filtros antispam que descarten este tipo de mensajes.

En este caso concreto, el sistema diseñado se basa en el clasificador de Naïve Bayes [17], clasificador que se comparará con el resto de técnicas empleadas en otros sistemas similares. La elección de un clasificador bayesiano es debido a varias características, que hacen a este tipo de clasificadores perfectos para este tipo de problemas. Dichas características son:

- Cada ejemplo observado modifica las probabilidades del suceso. Esto quiere decir que aunque un suceso no concuerde bien con una muestra de sucesos grande, éste no se desechará, ya que lo único que se producirá es una pequeña reducción de probabilidad para la hipótesis dada.

- Este modelo es robusto ante el posible ruido existente en los ejemplos de entrenamiento.

- La clasificación realizada tiene en cuenta la experiencia previa.

Para la creación de este sistema se ha utilizado la herramienta WEKA, que permite la utilización de sus clasificadores para comparar entre ellos cuál realiza mejor las tareas de clasificación de correo electrónico.

El conjunto de entrenamiento utilizado en este proyecto constaba de un total de 4601 instancias, un conjunto de estancias suficiente como para que los modelos generados arrojen resultados interesantes.

Además del mencionado clasificador de Naïve Bayes, se han utilizado varios clasificadores disponibles en WEKA para realizar comparaciones que determinen cuál de los clasificadores trabaja mejor en este problema. Los clasificadores utilizados han sido además del Naïve Bayes, el J48 (árbol de decisión), ADTree (árbol de decisión) y el IBk (clasificador K vecinos más cercanos, KNN).

Tras usar un conjunto de entrenamiento de 34 instancias para evaluar cada uno de los clasificadores, se comprobó que el clasificador Naïve Bayes clasificaba las 34 instancias del conjunto de test correctamente, siendo el mejor de los clasificadores con un 100% de acierto. Esto hace que sea un clasificador perfecto para la tarea de clasificar correo no deseado.

2.3 Herramienta WEKA

Como ya hemos podido ver en el apartado anterior, un gran número de investigadores utilizan la herramienta WEKA para realizar distintas tareas de análisis de datos. Para el proyecto que se quiere llevar a cabo, esta herramienta nos puede ser muy útil, ya que gracias a la gran cantidad de utilidades que posee nos permite realizar todo tipo de tratamiento de datos, elección de atributos para los modelos de predicción y entrenamiento de los clasificadores. Es por ello que WEKA ha sido la herramienta escogida para desarrollar la primera parte del proyecto, en la cual se crearán los modelos de predicción para el sistema.

En este apartado se describen muy brevemente las características principales de esta herramienta para en apartados posteriores explicar más en detalle todas las tareas que han sido realizadas usando algunas de las utilidades de WEKA.

2.3.1 Historia de la herramienta

La herramienta WEKA (Waikato Environment for Knowledge Analysis) [2] es una herramienta desarrollada por la Universidad de Waikato (Hamilton, Nueva Zelanda) y que comenzó a ser desarrollada en el año 1993. El motivo principal por el que se desarrolla esta herramienta es la necesidad de tener un software que fuera capaz de analizar datos procedentes de explotaciones agrícolas con los que realizar estudios sobre los cultivos realizados en ellas. Esta primera versión de WEKA fue realizada en lenguaje C.

Cuatro años más tarde, en 1997, se decide reescribir todo el código original en Java, incluyendo además distintas implementaciones de algoritmos de modelado.

El éxito de la herramienta llega a tal punto que en el año 2005 recibe el galardón *Data Mining and Knowledge Discovery Service* por parte de la ACM (Association for Computer Machinery) [18].

2.3.2 Utilidades

A continuación, vamos a mostrar algunas de las utilidades que presenta la herramienta, y que en alguno de los casos serán utilizadas para el desarrollo de nuestro sistema de predicción.

La Figura 4 muestra la pantalla que aparece cuando abrimos la herramienta y cargamos los datos que serán objeto del estudio. Esta primera pantalla, que se sitúa en la pestaña **Preprocess**, nos permite tanto consultar estadísticas de cada uno de los atributos de los registros cargados como realizar filtros que modifiquen o eliminen a los datos iniciales para dejarlos preparados para el estudio.

Esta primera pantalla servirá para realizar un primer visionado de los datos y así poder detectar anomalías como valores atípicos o simplemente visualizar cómo se distribuyen las instancias en un conjunto de clases determinado.

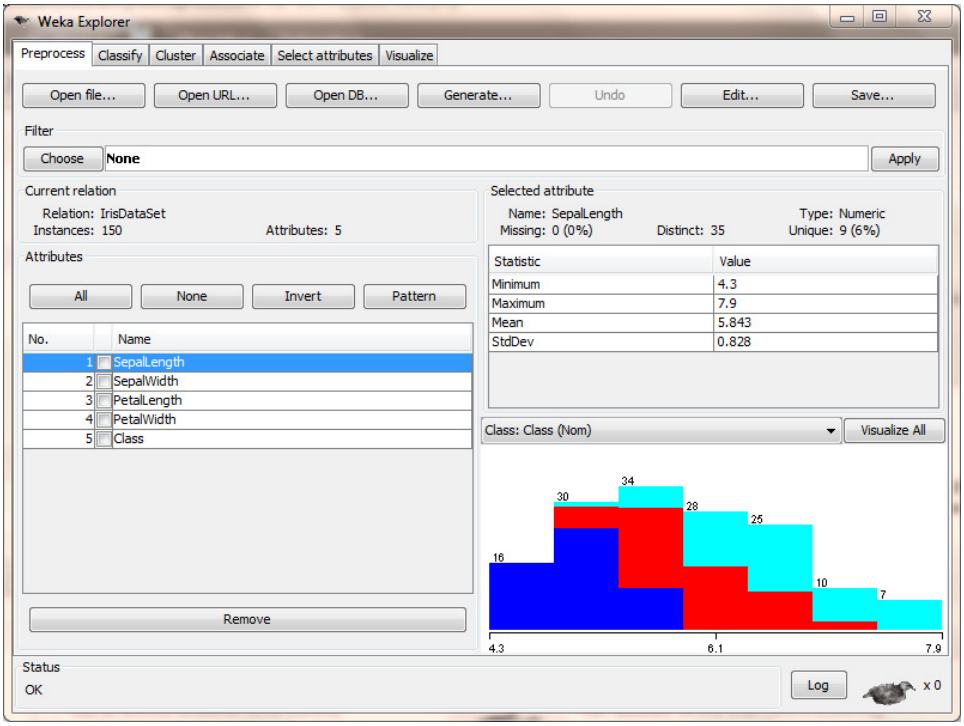

Figura 4. Pestaña Preprocess de WEKA

La pestaña **Classify**, mostrada en la Figura 5, tiene una gran importancia dentro de la herramienta, ya que dentro de ella se realizan todas las labores de clasificación de instancias. Esta clasificación puede ser realizada mediante los clasificadores que la

herramienta tiene implementados. Los clasificadores implementados se dividen en 6 secciones, mostradas en la Figura 6, cada una de las cuales tiene diversos clasificadores que poseen unas características determinadas.

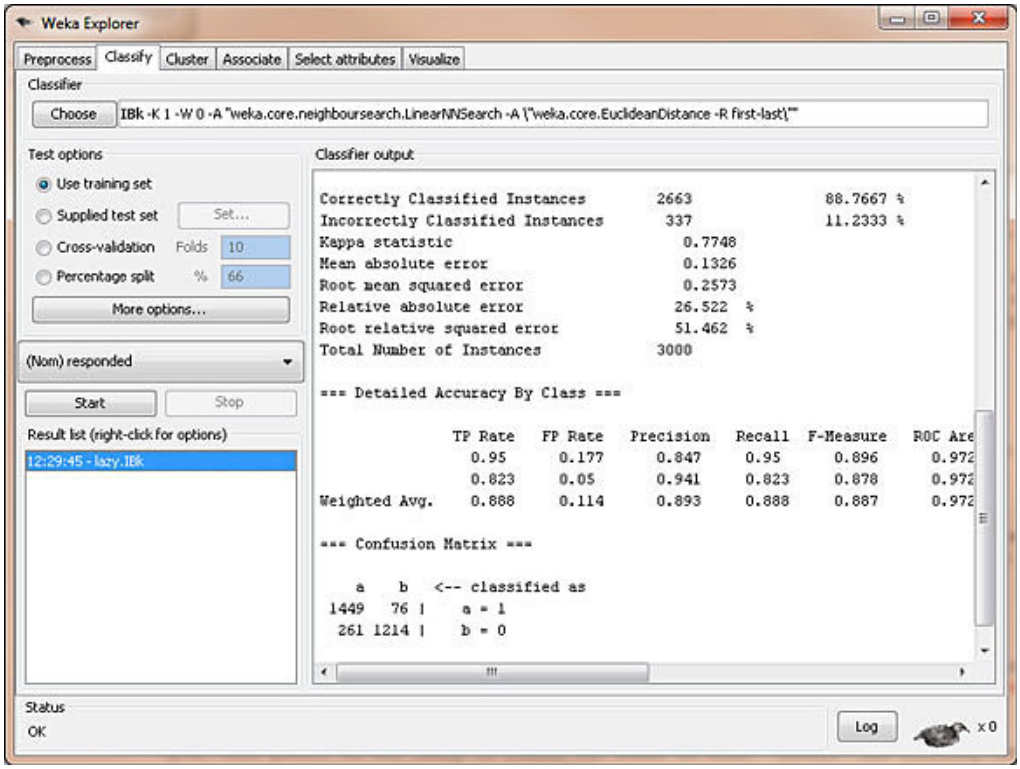

Figura 5. Pestaña Classify de WEKA

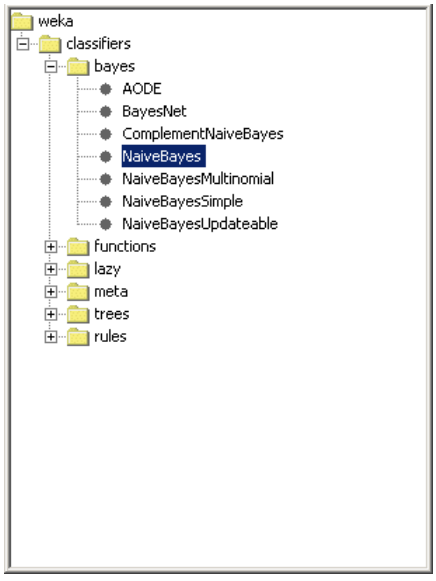

Figura 6. Clasificadores WEKA

Tal y como muestra la Figura 5, podemos ver que la interfaz se puede dividir en tres partes:

- En la zona superior izquierda tenemos varias opciones que nos permiten establecer el tipo de clasificador que queremos utilizar además de poder indicar el tipo de validación que queremos realizar (validación cruzada, validación a partir de fichero de test, partición del conjunto de entrenamiento o el uso del propio conjunto de entrenamiento para evaluar).

- En la zona inferior izquierda tendremos todos los clasificadores creados desde que tenemos la herramienta abierta. Esto nos permitirá comparar los resultados obtenidos con cada uno de los clasificadores y así verificar cuál es el que se ajusta mejor a las características de nuestro problema. En algunos casos, la generación de clasificadores como árboles o redes bayesianas permite la visualización gráfica del clasificador generado por la herramienta. Esta visualización nos permitirá ver de una manera más sencilla el modelo de clasificación creado por WEKA.

- Por último, en la zona derecha tenemos una gran zona en la que se muestran los datos del clasificador seleccionado en la zona inferior izquierda. Esta zona de la interfaz es de gran importancia, ya que es aquí donde aparecen todos los datos estadísticos de los clasificadores creados, lo que nos permitirá realizar los estudios necesarios para ver cuál es el mejor clasificador para utilizar en la solución de nuestro problema. Como estadísticos de interés a la hora de desarrollar el proyecto, nos fijaremos mucho en la tasa de aciertos del clasificador y en la matriz de confusión que se genera. Con esta matriz seremos capaces de ver tanto las instancias que han sido clasificadas correctamente como las que no y así poder ver hacia donde se desvían los errores de la clasificación.

Como se describirá en apartados posteriores, para el desarrollo de este proyecto se han utilizado algoritmos pertenecientes a la mayoría de categorías que se muestran en la Figura 6. A priori no tenemos la seguridad de que un algoritmo en concreto vaya a ser el mejor para resolver un problema, por lo que para encontrar el mejor algoritmo utilizaremos diversos clasificadores para ver cuál se ajusta mejor a las características del problema.

Otra de las utilidades que ofrece la herramienta WEKA es la creación de **clusters** para la clasificación (Figura 7). En Análisis de Clusters es una técnica de Análisis Exploratorio de Datos para resolver problemas de clasificación. Su objeto consiste en ordenar objetos en grupos o clusters de forma que el grado de asociación/similitud entre los miembros del mismo cluster sea más fuerte que el grado de asociación/similitud entre miembros de diferentes clusters. Cada cluster se describe como la clase a la que sus miembros pertenecen. Esta técnica también es muy útil para realizar clasificación de instancias, ya que mediante estas técnicas podemos descubrir agrupaciones que a simple vista no son obvias.

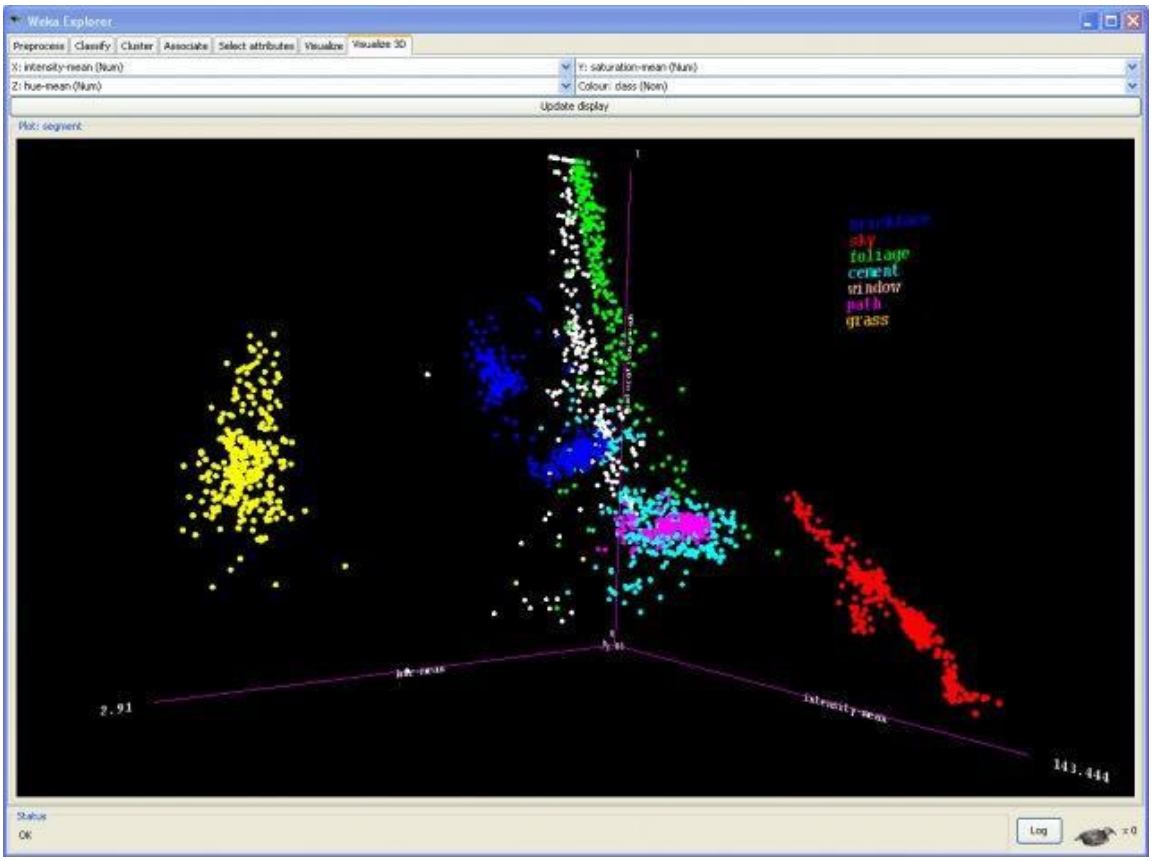

Figura 7. Cluster WEKA

En la Figura 8 se puede ver la pestaña **Associate** de WEKA. Esta pestaña permite realizar asociaciones de atributos con algoritmos predeterminados que posee la herramienta WEKA. Los algoritmos predeterminados que posee la herramienta son totalmente ajustables al problema que se intente resolver, ya que es posible realizar modificaciones en los parámetros del algoritmo para ajustar las propiedades de dicho algoritmo a las características del problema.

Además, para facilitar el análisis de las soluciones, WEKA permite ordenar las reglas resultantes mediante diversos indicadores como la confianza, el apalancamiento o la elevación.

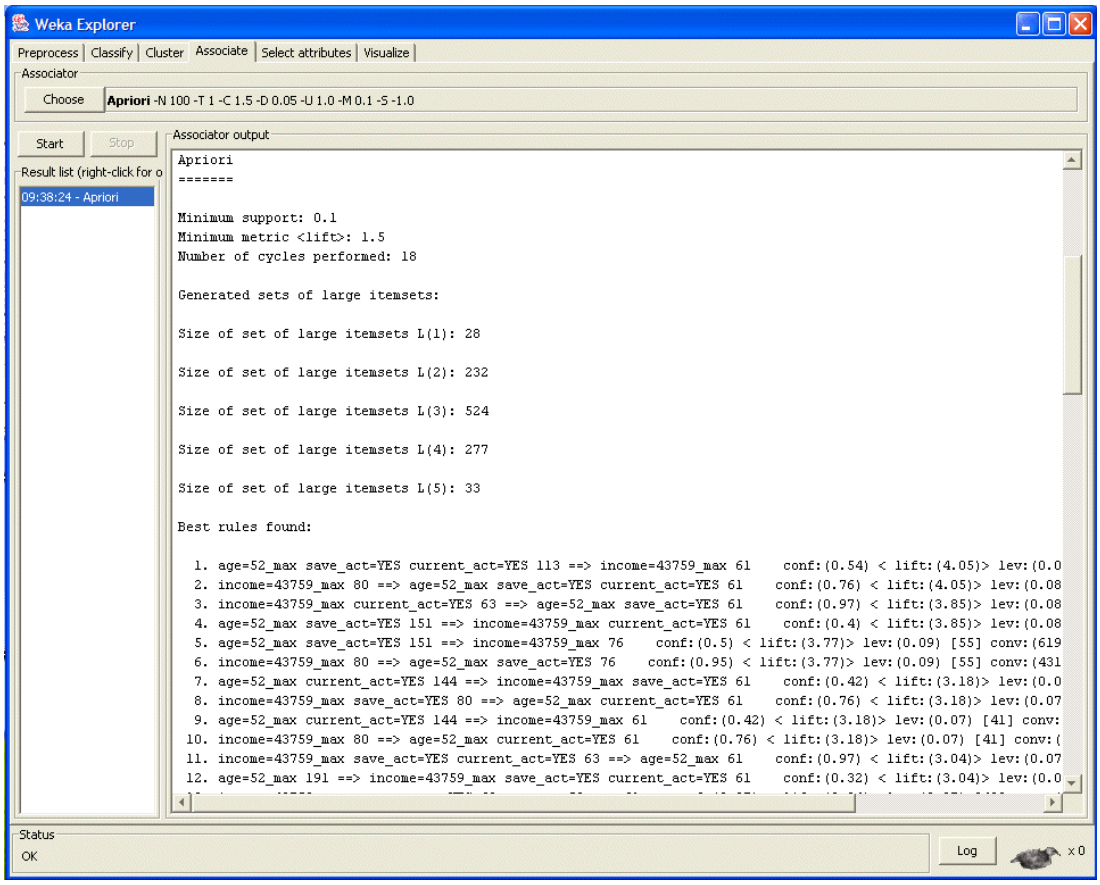

Figura 8. Pestaña Associate de WEKA

La Figura 9 muestra la pestaña **Select Attributes.** Esta pestaña ha sido de vital importancia para el desarrollo del proyecto, ya que esta funcionalidad nos permite clasificar los atributos de un conjunto de datos de acuerdo a diferentes parámetros como la correlación entre atributos y la clase, el valor chi-cuadrado para el atributo o la ganancia.

Esta funcionalidad tendrá al igual que el resto de funcionalidades un conjunto de algoritmos de clasificación de atributos predefinidos, que facilitarán el análisis de relevancia de cada uno de los atributos de los conjuntos de entrenamiento.

Además de los diferentes algoritmos predefinidos para realizar los procesos de análisis de relevancia de atributos, esta pestaña posee también métodos de clasificación de los atributos que permitirán generar listas que faciliten la lectura de los resultados de los estudios al clasificar los datos según los criterios seleccionados.

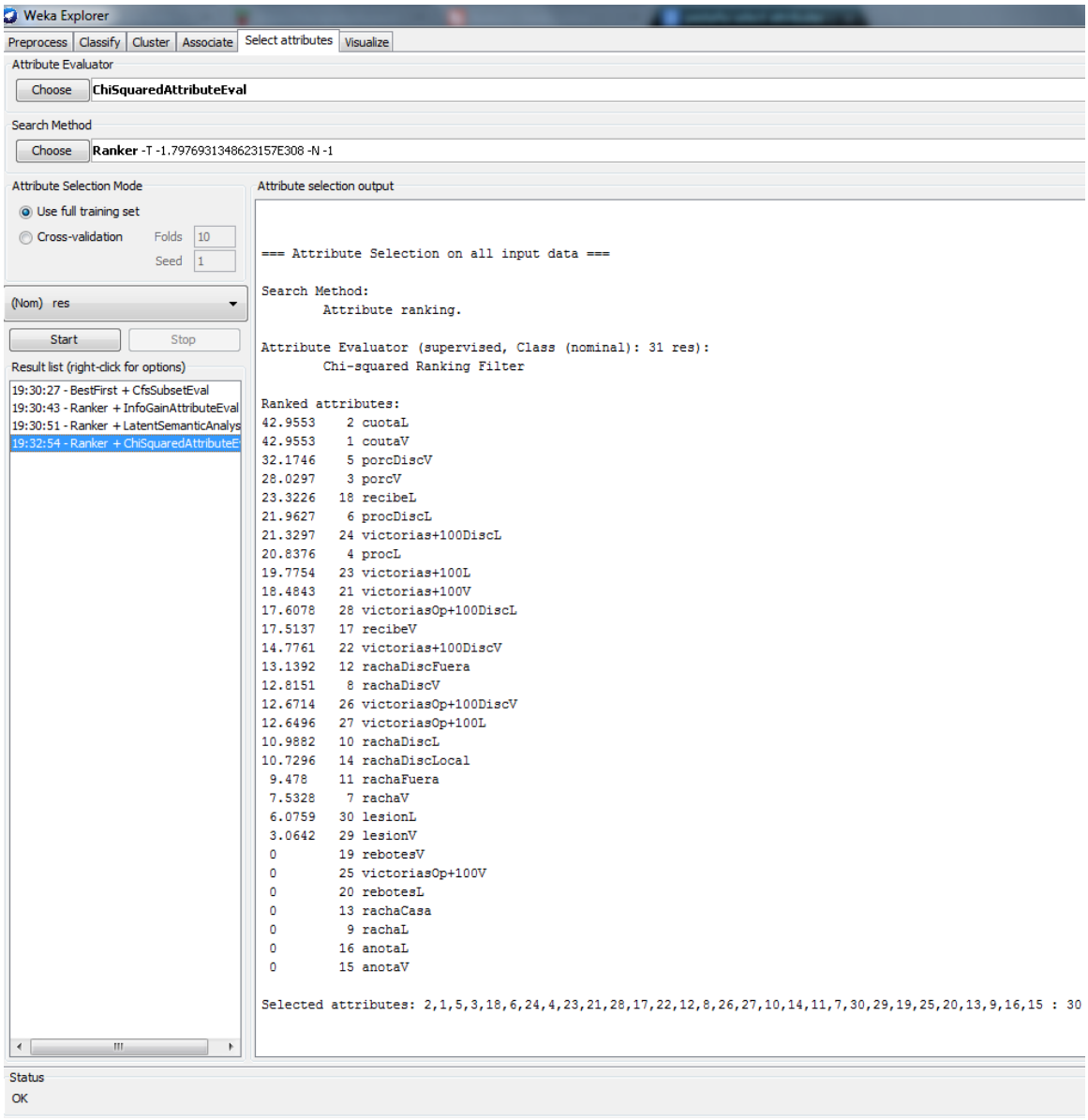

Figura 9. Pestaña Select Attributes de WEKA

Por último y para acabar con la presentación de la herramienta WEKA, que va a ser clave en futuros procesos de análisis de datos del proyecto, sólo quedaría presentar la pestaña **Visualize** (Figura 10). En esta funcionalidad de la herramienta WEKA podremos realizar contrastes entre dos variables del conjunto de entrenamiento para ver si cumplen algún tipo de relación entre ellas y la clase.

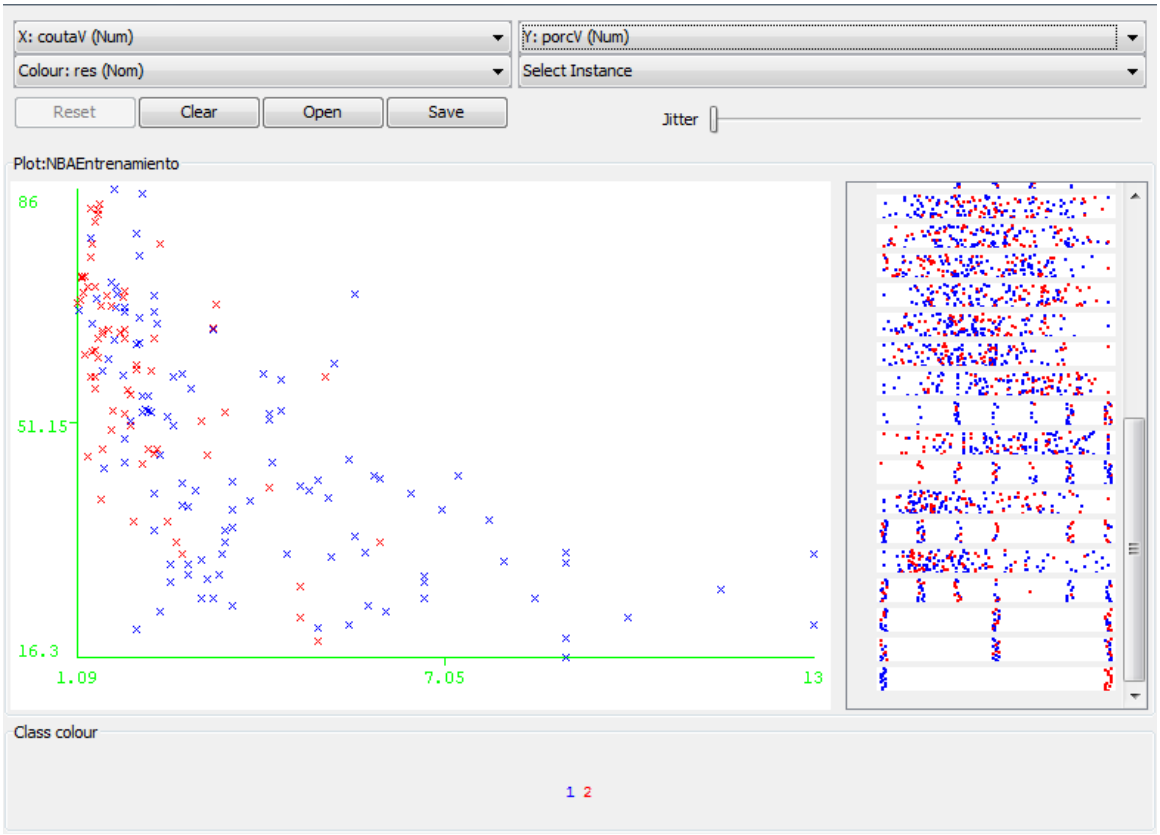

Figura 10. Pestaña Visualize de WEKA

Para el manejo de esta herramienta tan sólo es necesario seleccionar dos variables del conjunto de entrenamiento e indicar la clase por la que se va a clasificar. Al seleccionar estos atributos la herramienta se encarga automáticamente de indicar en las coordenadas *X,Y* los puntos que reflejan los datos de cada una de las instancias del conjunto de entrenamiento.

Esta funcionalidad es muy importante para detectar correlaciones entre pares de atributos del conjunto de entrenamiento, ya que detectar un determinado comportamiento de un par de atributos del conjunto nos puede servir para la posterior toma de decisiones al diseñar los algoritmos de predicción de nuestra herramienta.

Tras explicar brevemente cada una de las funcionalidades principales que nos ofrece la herramienta WEKA, se pasará a explicar de forma muy rápida la historia que ha envuelto el mundo de las apuestas, para así poder situar el contexto en el que se realiza esta actividad y entender mejor los objetivos de este proyecto.

2.3.3 Ficheros de Entrada (ARFF)

Los datos de entrada que utiliza la herramienta para realizar los análisis, deben presentarse en un formato de fichero específico con formato .arff.

Estos ficheros están divididos en tres partes:

- **Nombre de la relación:** Sección que comienza con el marcador *@relation* y que indica el nombre de la relación.

- **Atributos de la relación:** Sección en la que se definen los atributos de la relación. Cada atributo irá precedido del marcador *@attribute.*

- **Datos de la relación:** Finalmente el fichero tendrá una zona en la que se almacenan los datos de los registros que se van a utilizar. Esta zona del fichero va precedida del marcado *@data*.

La creación de estos ficheros es muy sencilla, ya que a partir de un fichero con formato .csv y haciendo unas pequeñas modificaciones, podemos disponer de un fichero .arff listo para ser usado en la herramienta WEKA.

Para el desarrollo de este proyecto se detalla más adelante cómo se han generado los ficheros .arff con los datos de entrenamiento de cada una de las competiciones que se van a estudiar.

2.4 Las Apuestas Deportivas

En esta sección trataremos de ver dónde está el origen de las apuestas deportivas, la evolución que ha sufrido este negocio y nos centraremos en el apartado legal, y más concretamente en las nuevas leyes que han entrado en vigor en España para ver las nuevas condiciones que tienen tanto casas de apuestas como los apostantes a la hora de realizar sus actividades.

2.4.1 Historia de las Apuestas Deportivas

Como ya se había comentado anteriormente, el negocio de las apuestas, tanto por Internet como en locales especializados en este tipo de actividad, ha sufrido un aumento considerable en los últimos años. España no ha sido tradicionalmente un país en el que este tipo de juegos fueran muy populares, ya que hasta hace unos años sólo contaba con su famosa quiniela semanal y las apuestas de caballos, que sólo podían ser realizadas en los hipódromos donde tenían lugar las carreras.

Si quisiéramos hablar sobre el origen de las apuestas deportivas, habría un gran número de personas que opinaría que este tipo de actividad es una nueva tendencia en

nuestra sociedad, pero esto no es cierto. Para encontrar el origen de las apuestas deportivas tendremos que retroceder varios siglos para encontrar los primeros indicios.

La historia comienza en tiempos de la Antigua Grecia. Ellos fueron los que cada cuatro años se congregaban en el estadio para animar a los mejores deportistas de la época en los Juegos Olímpicos. En aquella época ya podíamos encontrar personas dispuestas a apostar por los ganadores de cada una de las disciplinas de los juegos. No obstante, la presencia de las apuestas en la sociedad griega de la época era testimonial.

Si avanzamos unos cuantos siglos y llegamos hasta la época de los romanos también podremos encontrar indicios de apuestas en los espectáculos de la época. Los famosos espectáculos de gladiadores y las carreras de cuadrigas fueron también eventos en los que las apuestas estuvieron presentes. En esta época las apuestas tienen una presencia más importante que en la época griega, pero no llegan a tener un volumen lo suficientemente grande como para considerarlo importante.

Dejamos el Imperio Romano y saltamos hasta la Edad Media, donde también se ha detectado un volumen importante de apuestas. En esta ocasión, estas apuestas se realizaban en los torneos caballerescos, que se volvieron más interesantes a raíz de este tipo de apuestas. Además de los torneos de caballeros, otro tipo de espectáculo en el que se realizaba un gran número de apuestas eran los torneos de tiro con arco, en los que se solía apostar por el ganador.

Si queremos conocer el verdadero origen del fenómeno que son hoy las apuestas deportivas tendremos que irnos a Inglaterra, y concretamente al año 1780, año donde se legalizaron las apuestas en las carreras de caballos.

En años posteriores, y gracias a la evolución de la prensa, los periódicos de Londres comenzaron a crear secciones que iban dedicadas exclusivamente a las apuestas deportivas. Esta tendencia de la realización de apuestas se extendería a las colonias inglesas por todo el mundo, pero sobre todo a las colonias americanas. En la segunda mitad del siglo XIX, las apuestas llegaron a América, haciendo que en ciudades como Filadelfia este tipo de juegos fueran muy populares.

Sin embargo, sería en los años 30 del siglo XX cuando las casas de apuestas empezaron a multiplicarse. El primero de estos locales abrió en la ciudad inglesa de Liverpool, pero una gran cantidad de este tipo de negocios comenzó a extenderse por todo el continente, momento en el que las apuestas llamaron la atención a las masas y se hicieron muy populares.

El momento en el que este negocio se revoluciona y llega hasta lo que conocemos hoy por el negocio de las apuestas, llega con la evolución que sufre Internet en la última década del siglo XX y comienzos de siglo XXI. En este momento se crean empresas especializadas en apuestas deportivas, primero en Canadá y posteriormente en estados Unidos, empresas que idearon una nueva forma de apostar en línea.

Actualmente, la competencia entre las casas de apuestas es muy grande. Como en cualquier empresa, el objetivo que tienen es atraer nuevos clientes y fidelizar a los que ya tienen. Esto hace que en la actualidad los apostantes puedan disfrutar de un gran número de promociones que ofrece cada una de las casas de apuestas.

2.4.2 Legislación vigente en materia de apuestas deportivas

Si hablamos de las apuestas deportivas en España no podemos olvidarnos de la nueva ley de regulación de este tipo de juegos que entró en vigor en mayo de 2011 [19]. Esta ley se creó debido a la desactualización en materia legal que había en España, donde la ley del juego no sufría modificaciones desde 1977. La última legislación tan solo trataba actividades como loterías, casinos y bingos, lo que hacía que hubiera un gran vacío legal en lo que a apuestas se refiere.

Las apuestas online son un mercado que mueve en España una gran cantidad de dinero que está escapando del control del fisco. El principal objetivo de la nueva ley es que las empresas que ofrecen apuestas tengan que registrarse en España y no operar desde paraísos fiscales como Luxemburgo, Gibraltar, Malta o Suiza para así evitar la evasión fiscal.

La formulación de la nueva ley sacará de la ilegalidad a muchas casas de apuestas que operan en España sin ningún tipo de licencia. A partir de ahora las empresas que quieran operar en el mercado de las apuestas tendrán que obtener una licencia.

Una de las principales batallas que quiere librar esta nueva ley de apuestas deportivas en España es la lucha contra el juego de menores de edad y la ludopatía. Por ello se han creado mecanismos muy estrictos para el control de acceso a los niños y la limitación de acceso a aquellas personas que soliciten voluntariamente su autoexclusión de la página.

Esta ley además de regular ciertos aspectos que quedaban en el aire con la antigua ley también tiene muy presentes a los clubes de futbol del país, ya que al igual que ocurre con la Quiniela, los clubes verán cómo una parte de los ingresos de las casas de apuestas revierte directamente en sus arcas. Esto es debido a que las casas de apuestas utilizan a dichos equipos para desarrollar su negocio.

Las empresas que tienen páginas web de apuestas deportivas tendrán un régimen fiscal especial que gravará desde la obtención de la licencia hasta la organización y el reembolso de los premios por parte de los apostantes. Esto garantizará el juego limpio y la transparencia para los usuarios.

Esta ley también contempla la creación de una Comisión Nacional del Juego, que se encargará de autorizar y más tarde, supervisar y controlar, así como sancionar si fuera necesario, este tipo de actividades. Se han establecido multas a las casas de apuestas que no respeten la ley que podrían llegar hasta los 50 millones de euros.

Esta ley ha sido muy controvertida, ya que muchas casas de apuestas dejarán de operar en España ya que no están dispuestas a pagar los impuestos que ha establecido el estado para este tipo de actividades. No podemos olvidar que las empresas que se dedican a este negocio invierten una gran cantidad de dinero en publicidad, sobre todo en prensa escrita.

Las cifras son escandalosas. La Hacienda española recauda anualmente unos 1.700 millones de euros en concepto de juego. El juego privado ha generado unos 100.000 empleos directos. Mientras tanto, las nuevas modalidades de juegos online generaron el

año pasado unos 315 millones de euros de beneficio, pero el juego online, a penas crea puestos de empleo en España.

En España los juegos online ya han enganchado a unas 400.000 personas en todo el país [20].

2.5 Herramientas para apostantes

2.5.1 Software

En este apartado trataremos de mostrar herramientas disponibles para los apostantes que tratan de realizar predicciones de los resultados de todo tipo de deportes. Las herramientas que se mostrarán a continuación son compatibles con el sistema operativo Android, por lo que éstas estarán disponibles tanto en teléfonos móviles como tablets, lo que facilita la consulta de las estadísticas y probabilidades en prácticamente cualquier sitio. Las aplicaciones encontradas son las siguientes:

- **Soccer Prediction:**

Esta aplicación, que puede ser descargada desde la Play Store de Google [21], se encarga de hacer recomendaciones sobre los posibles resultados de un partido de fútbol y la probabilidad que hay de que ese resultado se dé al final del partido. El sistema que utiliza para realizar estas predicciones no se detalla en la información de la aplicación, por lo que no podemos contrastar los métodos utilizados para realizar las predicciones.

La aplicación se centra únicamente en resultados de partidos de fútbol, en los que recomiendan apuestas del equipo ganador y doble oportunidad (estas apuestas consisten en apostar sobre 2 resultados para cubrir un espacio de resultados mayor, reduciendo el beneficio pero aumentando las posibilidades de aceptar).

Esta sencilla aplicación está dividida en tres pestañas. La Figura 11 muestra la pestaña destinada a las predicciones realizadas en semanas anteriores con los resultados finales de los partidos. Las predicciones con fuente en color verde son las que fueron acertadas, mientras que las de color rojo son las que han sido falladas.

Figura 11. Pantalla de resultados (Soccer Prediction)

En la pantalla de predicciones de la aplicación (Figura 12) aparecen las predicciones realizadas para los partidos que tendrán lugar durante la siguiente semana. Como ya se ha comentado anteriormente, en la aplicación se recomiendan dos tipos de apuestas:

- **Apuestas de resultado único:** En este caso la aplicación recomienda una única apuesta al equipo que más probabilidades tiene de ganar. Estas apuestas aparecen con el literal "Home win" o "Guest win".

- **Apuestas doble oportunidad:** Para partidos en los que el resultado final no está tan claro la aplicación recomienda apostar a lo que es conocido en las apuestas como una "doble oportunidad". Este tipo de apuesta consiste en apostar por dos de los tres resultados posibles en un partido de fútbol para así aumentar las probabilidades de acierto. Por el contrario, este aumento de probabilidades de acierto repercute en la cuota que ofrece la casa de apuestas, que en este caso será menor que la cuota que se ofrecería por una apuesta de resultado único. La apuesta de doble oportunidad que ofrece la aplicación puede estar formada por cualquier par de resultados posibles del partido.

Figura 12. Pantalla de predicciones (Soccer Prediction)

Respecto a los porcentajes de acierto que aparecen en las predicciones, tenemos dos tipos de porcentajes. Los porcentajes que aparecen entre paréntesis corresponden a las probabilidades de que la apuesta de doble oportunidad sea acertada. Los porcentajes que no aparecen entre paréntesis reflejan la probabilidad de que el resultado único en el que el equipo con más probabilidades gane acabe siendo correcto.

Según hemos podido observar, la aplicación nunca ofrece una apuesta de resultado único en la que nos recomiende apostar al empate en un partido. Como se verá en apartados posteriores, la predicción de empates es muy complicada, ya que el empate es el resultado menos frecuente de los tres posibles y mientras que una victoria local o visitante puede ser fácil de predecir debido a la superioridad de uno de los dos equipos en sus estadísticas, en los resultados de empate no ocurre por norma general que las estadísticas de los dos equipos estén igualadas.

Finalmente, la tercera de las secciones de la aplicación muestra un registro histórico de las apuestas realizadas y elabora una estadística para ver el porcentaje de aciertos que se ha conseguido (Figura 13). A fecha de 12 de diciembre de 2012 el porcentaje de

aciertos conseguidos por la aplicación ha sido del 66.73%. Teniendo en cuenta que aproximadamente la mitad de las apuestas son de doble oportunidad y la otra mitad son de apuesta única, la probabilidad a priori que se tendría al apostar sería del 50%, lo que demuestra que la aplicación consigue aumentar el porcentaje de acierto muy por encima de la probabilidad a priori.

	Tip	Right / Wrong	Percent
Lay	9791	7141 / 2650	72.93%
Back	3967	2040 / 1927	51.42%
Total	13758	9181 / 4577	66.73%

Figura 13. Pantalla de estadísticas (Soccer Prediction)

- **Bet2Win:**

Bet2Win es un conjunto de aplicaciones que pueden ser descargadas desde la Play Store de Google [22]. Para el análisis de la aplicación se ha utilizado la versión Lite que se puede descargar de forma gratuita y que contiene parte de las funciones que ofrece la versión de pago.

Esta aplicación es muy similar a la se ha mostrado anteriormente, ya que realiza recomendaciones de apuestas, pero en este caso se tienen en cuenta más deportes además de fútbol. Los deportes sobre los que realiza predicciones son fútbol, baloncesto, tenis, hockey hielo y fútbol americano.

Según se informa en las características de la aplicación, las predicciones realizadas son llevadas a cabo por expertos que tienen en cuenta varios factores en cada uno de los partidos como pueden ser las rachas de los equipos, lesiones de los jugadores o traspasos realizados durante la temporada.

La sencillez para el uso de la aplicación es clave, ya que simplemente se limita a mostrar una serie de partidos con las predicciones realizadas (Figura 14).

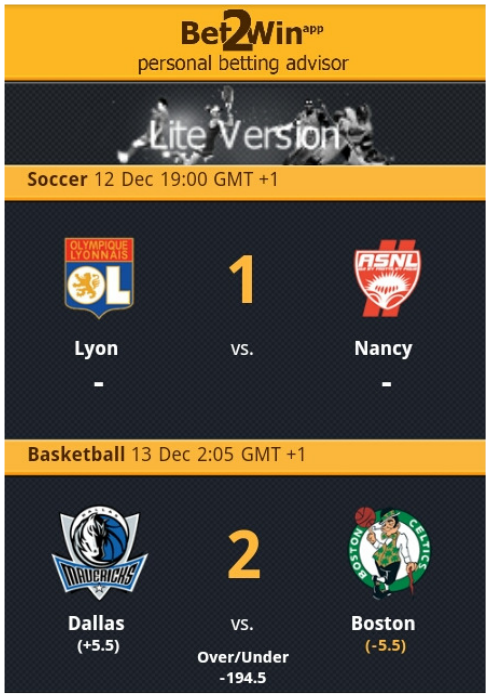

Figura 14. Pantalla de predicciones (Bet2Win)

Para realizar la comparación y verificación de los resultados que habían sido recomendados por la aplicación, se ofrece una sección en la que aparecen los resultados de los encuentros, la recomendación hecha y las ganancias o pérdidas acumuladas si se hubieran hecho caso a las recomendaciones (Figura 15).

	Away	Home	T	H	O/U	
12 Dec	LA Lakers -5.5	Cleveland +5.5	1	h1	+206.5	94:100
12 Dec	Denver -4.5	Detroit +4.5	1	h1	+202.5	101:94
12 Dec	Washington +4.5	New Orleans -4.5	2	h2	-187.5	77:70
11 Dec	Detroit +6.5	Philadelphia -6.5	1	h1	+186.5	97:104
11 Dec	G. State -4.5	Charlotte +4.5	1	h1	+203	104:96
11 Dec	Atlanta +7.5	Miami -7.5	2	h2	-195.5	92:101
10 Dec	Milwaukee +4.5	Brooklyn -4.5	2	h2	+191.5	97:88
10 Dec	Denver +5.5	New York -5.5	2	h2	-202.5	106:112
10 Dec	Utah +6.5	LA Lakers -6.5	2	h1	+205.5	117:110
9 Dec	San Antonio -8.5	Charlotte +8.5	1	h1	-199.5	132:102

Figura 15. Pantalla de resultados (Bet2Win)

Como puede verse en la Figura 15, esta aplicación ofrece predicciones para tres tipos de apuestas:

- **Apuestas de resultado único:** En este caso la aplicación recomienda una única apuesta al equipo que más probabilidades tiene de ganar. La predicción de estas apuestas aparecen con el literal "1" (gana el equipo local) o "2" (gana el equipo visitante) en la columna T.

- **Apuestas con hándicap:** En todos los partidos se suele dar la posibilidad de apostar a apuestas con hándicap, que lo que hacen es añadir puntos a uno de los dos equipos haciendo que el resultado final para esta apuesta sea el resultado final del partido más los puntos que ha añadido la casa de apuestas con el hándicap. Las predicciones aparecen en la columna H con los literales "h1" (gana el equipo local después de aplicar el hándicap) o "h2" (gana el equipo visitante después de aplicar el hándicap).

- **Apuestas de over-under:** Estas apuestas se centran en la suma de los puntos conseguidos por los dos equipos en el partido. La apuesta consistirá en decir si la suma de los puntos será mayor o menos que un número de puntos establecidos por la casa de apuestas. Las predicciones aparecen en la columna "O/U", donde un número precedido de un + supone apostar a la apuesta *Over*, mientras que si el número va precedido de un – la predicción aconseja apostar a la apuesta *Under*.

2.5.2 Blogs y Páginas Web especializadas

Otra de las herramientas muy utilizadas por parte de los apostantes son los blogs y páginas web especializadas. Estas web suelen ofrecer predicciones de todo tipo de eventos deportivos y son muy populares entre los apostantes.

Normalmente son especialistas o simplemente aficionados a las apuestas los que ofrecen la información en estas páginas y simplemente se limitan a informar sobre apuestas que en su opinión son atractivas, bien porque la cuota ofrecida es alta o bien porque las consideran muy seguras.

También es muy normal encontrar este tipo de publicaciones dentro de los diarios deportivos, que aprovechan el tirón que tiene las apuestas deportivas para crear secciones en las que diversos especialistas realizan predicciones sobre todo tipo de deportes. Una de las secciones más seguidas por los apostantes es la que el Diario MARCA ofrece en su página web [23]. En esta sección, cinco especialistas realizan predicciones sobre fútbol, baloncesto, tenis y fútbol americano.

El éxito de este tipo de páginas reside en la sencillez con la que se presenta la información y en los buenos resultados que suelen obtener los expertos, lo que hace que los usuarios vuelvan a visitar la página.

Capítulo 3

SISTEMA DE PREDICCIÓN DE RESULTADOS EN EVENTOS DEPORTIVOS

3.1 Introducción

En este apartado se explica minuciosamente todos los pasos que se han ido siguiendo para la realización de la primera parte del proyecto, cuyo objetivo principal es crear un sistema capaz de realizar predicciones sobre los resultados de determinados eventos deportivos.

En primer lugar, se detalla el proceso seguido, desde la definición de los datos a recoger hasta el estudio de éstos para la generación de los modelos de predicción. Este proceso es clave para el desarrollo del proyecto, ya que serán estos datos a partir de los cuales se desarrollará toda esta primera fase de creación de los modelos de predicción.

Una vez definidos los datos a recoger y realizada dicha recolección, se detalla el proceso de análisis de los datos. Este proceso ha sido realizado con la ayuda de la herramienta WEKA, a partir de la cual comprobaremos qué modelos de clasificación son los más adecuados para la resolución de este problema. Se contrastarán diferentes modelos de clasificación para saber cuál de ellos se amolda mejor a nuestro problema,

comprobando el porcentaje de aciertos en la clasificación y la capacidad de acierto en cada resultado individual (victoria local, empate y victoria visitante).

Una vez se tengan los modelos de clasificación escogidos en función de los parámetros comentados anteriormente, se pasará a implementar dichos modelos en el soporte escogido para el sistema, que será un fichero Excel. Dichos modelos serán implementados en Visual Basic como una macro del fichero Excel. Una vez el usuario introduzca en las celdas los datos sobre la predicción, será la macro la que se encargue de calcular el resultado más probable basándose en los modelos escogidos.

Por último, también se detallarán tanto los datos necesarios para realizar la predicción como el formato de salida de la predicción, ya que necesitamos generar un fichero de predicciones que sirva como entrada del algoritmo genético que se desarrollará en la siguiente fase del proyecto.

3.2 Proceso de Minería de Datos

El primero de los pasos que debemos dar para poner en marcha el proyecto, es definir qué datos nos pueden ser útiles para realizar las predicciones y ver dónde podemos encontrarlos. A continuación se detallarán qué datos se han recogido para cada competición explicando brevemente por qué se cree que dichos datos podrían contribuir a generar un modelo de clasificación con un alto porcentaje de aciertos:

3.2.1 Selección de los datos necesarios para cada competición

- **Copa de Europa de Clubes (Champions League):**

La competición de la Champions League (también conocida como Liga de Campeones) es una competición futbolística europea en la que compiten los mejores equipos del continente. Dependiendo de los Coeficientes UEFA [24] de cada una de las ligas que pertenecen a la UEFA, cada país clasificará a un número de equipos determinados para esta competición. También haciendo referencia al coeficiente los países clasificados entrarán en una ronda preliminar o directamente se clasificarán para la fase de grupos del torneo. Por ello, podemos distinguir tres fases en el torneo:

- Fase Previa: En ella compiten sobre todo equipos de países cuyas ligas tienen un Coeficiente UEFA bajo. También compiten equipos de ligas con coeficientes altos, pero éstos comenzarán a jugar en rondas avanzadas. Esta fase previa se compone de tres rondas eliminatorias a doble partido. Los equipos que consigan ganar la eliminatoria correspondiente a la tercera ronda, pasarán directamente a la siguiente fase del torneo. El resto de equipos pasarán a disputar la otra competición

europea existente, la Europa League, que también será objeto de estudio en este proyecto y que se explicará más adelante.

- ❖ Fase de Grupos: En esta ronda participan los 24 equipos que se clasificaron directamente para la competición y los ocho equipos clasificados desde la Fase Previa. En este punto se forman ocho grupos de cuatro equipos cada uno mediante un sorteo que tiene en cuenta los méritos de los equipos en años anteriores. Esto hace que los equipos de un nivel más alto no se enfrenten entre si hasta la última fase del torneo.

  Esta fase se compone de seis partidos en los que los equipos ganan tres puntos por ganar un partido y un punto por empatarlo. Después de disputarse los seis partidos de cada grupo, pasarán a la siguiente fase los dos mejores equipos de cada grupo. El tercer equipo de cada grupo pasará a jugar la Europa League.

- ❖ Fase Final: Esta fase es disputada por los 16 equipos clasificados desde la Fase de Grupos. Esta fase del torneo se crea a partir de un sorteo en el que se generan ocho partidos que enfrentan al primero de uno de los grupos contra el segundo de otro grupo, evitando siempre que equipos del mismo país se enfrenten entre sí. A partir de este sorteo se genera un cuadro de eliminatorias donde ya no habrá ningún tipo de restricción. Todas las eliminatorias se disputan a doble partido, excepto la final del torneo que se disputa a un único partido en un estadio designado con un año de antelación por la UEFA.

Antes de comenzar la recogida de datos para el estudio hay que intentar reconocer qué factores podrían influir en el resultado de un partido de este torneo. A continuación se explicarán los atributos escogidos y las razones de dicha elección:

- ▪ Tipo de Partido: Este atributo lo que hace es distinguir entre los dos tipos de partidos que se pueden dar en este torneo: **partido dentro de una eliminatoria** o **partido dentro de la fase de grupos**.

  La elección de este atributo para el estudio es debido a varios factores que se detallan a continuación:

  - ❖ En la fase de grupos, cuando uno de los mejores equipos del torneo juega contra otro más débil, hay un alto número de victorias para el mejor de los equipos.

  - ❖ Cuando un equipo que es teóricamente más débil que el resto consigue pasar a la fase final, no está tan claro que al enfrentarse contra un equipo superior éste vaya a ganar sin ningún problema.

  - ❖ Si en una eliminatoria uno de los equipos ha conseguido una ventaja grande para disputar el partido de vuelta, el equipo que consigue dicha

ventaja intenta mantener la renta obtenida en el primer partido, lo que hace que el resultado pueda estar algo condicionado.

- ❖ En ocasiones, cuando un equipo ya ha conseguido clasificarse para la fase final antes de que se disputen todos los partidos de la fase de grupos, este equipo suele disputar los partidos restantes de la fase de grupos con jugadores que no juegan habitualmente para así reservar a sus mejores jugadores para fases posteriores. Esto puede hacer que el resultado del partido no sea el esperado comparándolo con un partido en el que el equipo juega con sus jugadores titulares.

- Cuotas de las Casas de Apuestas: Aquí se engloban tres atributos referentes a las cuotas que pagan las casas de apuestas por cada uno de los tres resultados posibles. Por ello, para nuestro estudio tendremos en cuenta las cuotas que ofrecen las casas de apuestas para la victoria del equipo local, el empate y la victoria del equipo visitante.

  La casa de apuestas que se ha elegido para recoger los datos tanto de esta competición como del resto que se detallarán posteriormente ha sido **Bwin**, una casa de apuestas en la que vamos a poder encontrar además de los datos de las cuotas de partido otros atributos que nos van a ser muy útiles en nuestro estudio y que explicaremos posteriormente.

  La elección de estos atributos para su inclusión en el estudio es debido a que las casas de apuestas realizan análisis previos para determinar esas cuotas. Ese estudio realizado les permite generar una gran cantidad de beneficios, ya que establecen las cuotas a partir de la probabilidad que estiman de que un determinado resultado se lleve a cabo. Este estudio previo para fijar las cuotas, les permite a las casas obtener grandes beneficios cuando el resultado del encuentro es el esperado y minimizar las pérdidas cuando el resultado del partido es totalmente inesperado.

- Previsión Meteorológica: Este atributo tiene como objetivo reflejar el estado meteorológico de la ciudad en el día y hora del partido. Dado que hay muchos factores que se pueden tener en cuenta a la hora de hablar de la previsión meteorológica se ha decidido codificar este campo de la siguiente forma para facilitar el análisis:

  - ❖ **Normal**: Hace referencia a condiciones climáticas en las que no hay precipitaciones y donde la temperatura no sobrepasa los 30ºC y tampoco es inferior a 10ºC.

  - ❖ **Precipitaciones:** En este caso el partido se disputaría con precipitaciones, ya sean en forma de lluvia o nieve y con una temperatura entre 10ºC y 30ºC.

  - ❖ **Frío:** Partido que se disputa con una temperatura entre 0ºC y 10ºC. Puede que haya precipitaciones o no.

- **Muy Frío:** El partido se disputa con una temperatura inferior a 0ºC. Puede que haya precipitaciones o no.

- **Calor:** En este caso el partido se disputaría sin precipitaciones y con una temperatura superior a los 30ºC.

La elección de esta variable se ha realizado en vista de que en muchas ocasiones, equipos de gran prestigio juegan un partido en ciudades del norte de Europa como pueden ser Moscú, San Petersburgo o Trondheim, donde en invierno las temperaturas pueden estar por debajo de 0ºC. Normalmente el partido se disputa con el campo lleno de nieve y con un frío que hace que un equipo acostumbrado a jugar a temperaturas entre los 10ºC y 30ºC tenga muchas dificultades para ganar el partido.

Para intentar anticiparnos a posibles problemas derivados del formato y la codificación de los datos, se ha intentado simplificar al máximo los valores de este atributo. Para ello, se tuvo en cuenta que en los casos en los que las condiciones de juego no son las normales, sería el equipo que juega como local el que tenga ventaja, por tanto se ha decidido modificar los valores de este atributo a sólo dos:

- **Sí:** Partidos en los que las condiciones climatológicas no son normales (precipitaciones, frío, muy frío o calor según la clasificación anterior) y por lo tanto el equipo local tendría una pequeña ventaja derivada de las condiciones climatológicas.

- **No:** Partido disputado bajo condiciones climáticas normales en los que a priori no habría ningún tipo de ventaja para ninguno de los equipos.

- Rachas: Este atributo recogerá las rachas positivas y negativas que lleva un equipo en los seis últimos partidos disputados. Se tendrán en cuenta los partidos de cualquier competición y el valor de este atributo saldrá a partir de esta codificación:

  Se considerará una racha a una sucesión de resultados consecutivos ya sean positivos o negativos. Las rachas las podremos clasificar en dos grupos:

  - **Racha Positiva:** Es una sucesión de partidos en las que el equipo ha conseguido empatar o ganar cada uno de los encuentros disputados. El valor de la racha será calculado sumando **1 punto** por cada partido ganado y **0,33 puntos** por cada partido empatado. Para ilustrar como calcular el valor de una racha pondremos el siguiente ejemplo:

    Consultamos los resultados de un equipo en los seis últimos partidos y obtenemos el siguiente detalle:

Victoria
Empate
Victoria
Empate
Derrota
Victoria

Si consideramos que el primer partido que aparece en la lista es el último que ha disputado el equipo, tendremos que la racha del equipo es:

$$\textit{Racha} = 1 + 0{,}33 + 1 + 0{,}33 = \mathbf{2{,}66}$$

Como podemos observar, se ha dejado de sumar en el partido en el que el equipo perdió, ya que su racha positiva se ve cortada. Por tanto, el valor de la racha obtenida para este equipo es de **2,66 puntos**.

- **Racha Negativa:** Es una sucesión de partidos en las que el equipo ha perdido los encuentros disputados. El valor de la racha será calculado restando **1 punto** para cada partido perdido. Al igual que en el otro tipo de racha, se dejará de sumar puntos cuando el resultado de un partido corresponda a una racha positiva. Un ejemplo en el que se puede ver el cálculo de una racha negativa podría ser el siguiente:

  Consultamos los resultados de un equipo en los seis últimos partidos y obtenemos el siguiente detalle:

Derrota
Derrota
Victoria
Empate
Derrota
Victoria

Si consideramos que el primer partido que aparece en la lista es el último que ha disputado un equipo, tendremos que la racha del equipo es:

$$\textit{Racha} = -1 - 1 = \mathbf{-2}$$

Como podemos observar, se ha dejado de sumar en el partido en el que el equipo ganó, ya que su racha negativa se ve cortada. Por tanto, el valor de la racha obtenida para este equipo es de **-2 puntos**.

Las razones por las que hemos elegido este atributo para nuestro estudio son claras, ya que un equipo que viene con una buena racha, podemos suponer que tendrá más posibilidades de ganar que un equipo que no tiene una buena racha en sus últimos 6 partidos.

La decisión de coger 6 partidos y no otro número diferente se ha hecho a partir de los datos que nos ofrecen las páginas web de las casas de

apuestas, y más concretamente de Bwin. Estas páginas web ofrecen un análisis de cada partido, donde se muestran diversas estadísticas entre las que se encuentran los resultados de los últimos 6 partidos de cada uno de los equipos que disputan un encuentro.

Una vez definida la fórmula para codificar las rachas, se ha detectado un posible futuro problema que se ha decidido abordar en esta fase del proyecto para que quede totalmente solventado. El problema en cuestión es la aparición de valores negativos en los valores de las rachas. Estos valores negativos puede que afecten a nuestro análisis, ya que hay algunos clasificadores (como el Perceptrón Multicapa [25]) que no acepta valores negativos.

Como conocemos el rango de valores dentro del cual puede estar una racha, podremos modificar la fórmula para que no aparezcan valores negativos.

Por la definición de racha realizada anteriormente, una racha puede tener un valor mínimo de -6 y un máximo de 6, que corresponden a perder 6 partidos seguidos o ganar 6 partidos seguidos. Para eliminar totalmente los valores negativos sólo tendremos que sumar 6 a la racha calculada con la fórmula para asegurarnos de que no aparecerá ningún valor negativo para este atributo. El rango de valores pasará de ser [-6, 6] a [0, 12].

- Países de los Equipos: Este atributo reflejará la nacionalidad de los equipos que intervienen en un partido.

  La elección de este atributo se ha hecho en base a que un equipo que pertenece a un país que por tradición tiene equipos de buen nivel, tendrá más posibilidades de ganar un partido que un equipo de un país en el que no hay mucha tradición en este deporte.

- Coeficiente UEFA del país del equipo: Este dato nos permite convertir en un valor numérico el país al que pertenece un equipo. De cara al análisis de los datos, esto nos permite acotar de una forma más eficaz el rango de valores que puede tomar el atributo país de un equipo. Los coeficientes que se utilizarán para cada uno de los países se resumen en la **Tabla 2**. Esta tabla muestra además otros datos como la correspondencia entre país y coeficiente en la temporada 2012-2013.

Después de haber definido qué datos serían adecuados recoger para realizar las predicciones, se ha decidido derivar otros datos al considerar que éstos podrían ayudar a mejorar la precisión de las predicciones. Los nuevos atributos que se han decidido derivar de los ya existentes son los siguientes:

- Racha Discreta de Resultados:

  Al detectar que hay más de 40 valores posibles para las rachas de los equipos, se ha decidido acotar los valores posibles a cuatro, agrupando rachas similares dentro de un grupo. Los valores que se han definido son los que aparecen en la Tabla 6:

	Rango Inferior	Rango Superior
Muy Mala	0	3
Mala	4	6
Buena	6.33	8.99
Muy Buena	9	12

Tabla 6. Correspondencias entre racha de resultados y su discretización

  Lo que queremos conseguir con esta discretización de los datos, es que los algoritmos de clasificación operen con un rango de valores menor, lo que a priori podría facilitar la clasificación de instancias.

- Zona del país en la Clasificación por Coeficientes UEFA:

  Al haber un total de 53 países pertenecientes a la UEFA, cada uno con un coeficiente diferente, se ha considerado agrupar a los países que tengan coeficientes similares, identificando así en qué zona de la clasificación están. La Tabla 7 muestra la conversión utilizada para calcular este atributo.

País	Coeficiente	TOP		País	Coeficiente	TOP
Inglaterra	84,410	TOP4		Bulgaria	14,250	TOP50
España	84,186	TOP4		Hungría	9,750	TOP50
Alemania	75,186	TOP4		Finlandia	9,133	TOP50
Italia	59,981	TOP4		Georgia	8,666	TOP50
Portugal	55,346	TOP10		Bosnia	8,416	TOP50
Francia	54,178	TOP10		Irlanda	7,375	TOP50
Rusia	47,832	TOP10		Eslovenia	7,124	TOP50
Holanda	45,515	TOP10		Lituania	6,875	TOP50
Ucrania	45,133	TOP10		Moldavia	6,749	TOP50
Grecia	37,100	TOP10		Azerbaijan	6,207	TOP50
Turquía	34,050	TOP20		Letonia	5,874	TOP50
Bélgica	32,400	TOP20		Macedonia	5,666	TOP50
Dinamarca	27,525	TOP20		Kazajistán	5,333	TOP50
Suiza	26,800	TOP20		Islandia	5,332	TOP50
Austria	26,325	TOP20		Montenegro	4,375	TOP50
Chipre	25,499	TOP20		Liechtenstein	4,000	TOP50
Israel	22,000	TOP20		Albania	3,916	TOP50
Escocia	21,141	TOP20		Malta	3,083	TOP50
Rep.Checa	20,350	TOP20		Gales	2,749	TOP50
Polonia	19,916	TOP20		Estonia	2,666	TOP50
Croacia	18,874	TOP50		Irlanda del Norte	2,583	TOP50
Rumania	18,824	TOP50		Luxemburgo	2,333	TOP50
Bielorrusia	18,208	TOP50		Armenia	2,208	TOP50
Suecia	15,900	TOP50		Islas Feroe	1,416	RESTO
Eslovaquia	14,874	TOP50		Andorra	1,000	RESTO
Noruega	14,675	TOP50		San Marino	916	RESTO
Serbia	14,250	TOP50				

Tabla 7. Correspondencias entre País, Coeficiente y Zona en el Ranking UEFA

Finalmente, sólo nos queda definir la clase por la que se va a clasificar. Al ser la mayoría de las competiciones pertenecientes al mismo deporte, la clase va a ser la misma y va a estar codificada igual para todas las competiciones de fútbol sobre las que se va a desarrollar este proyecto. La clase, claro está, va a ser el resultado final del partido, que es el dato que buscamos predecir. La codificación elegida para la clase, es la que típicamente se usa en las apuestas deportivas, tal y como muestra la Tabla 8.

Victoria Local ➔ 1 Empate ➔ X Victoria Visitante ➔ 2

Tabla 8. Codificación de la clase por la que vamos a clasificar

- **Liga Europea de la UEFA (Europa League):**

Esta competición, segunda en importancia en el ámbito de la UEFA, agrupa a campeones de las ligas con menor coeficiente en la clasificación de la UEFA y a equipos que han quedado clasificados en posiciones altas de ligas importantes, pero que sin embargo no han llegado a clasificarse para la Champions League.

El formato de la competición define varias fases que se pasan ahora a detallar:

- Fase Previa: Es una fase conformada por cuatro eliminatorias en las que los equipos de ligas menores clasificados para esta competición, juegan partidos eliminatorios con el objetivo de clasificarse para la Fase de Grupos, en la que se podrán enfrentar a equipos de las ligas más potentes del continente. Los partidos de esta fase serán de ida y vuelta, siguiendo las mismas reglas que hay en otros torneos, como por ejemplo la Champions League.

- Fase de Grupos: En esta fase participan 48 equipos, los cuales están divididos en 12 grupos de cuatro equipos cada uno. Cada equipo jugará dos partidos contra el resto de equipos de su mismo grupo, clasificándose para la siguiente fase los dos mejores equipos.

- Fase Final: La fase final contará inicialmente con 32 equipos. 24 de ellos provienen de la Fase de Grupos de esta competición, mientras que los ocho restantes proceden de la Fase de Grupos de la Champions League. Estos ocho equipos son los equipos que hayan quedado en 3ª posición de su respectivo grupo de la Champions League.

  El formato de esta fase será el de eliminatoria, jugándose para cada cruce un partido de ida y vuelta. La final se disputará en un campo designado por la UEFA con al menos un año de antelación y será a partido único.

Una vez se ha definido el formato de la competición, el siguiente paso es definir qué tipo de datos necesitamos para crear los modelos de predicción. Al ser una competición muy similar a la Champions League se ha considerado que los datos a recoger deberían ser los mismos que en la otra competición. Por este motivo no se volverán a explicar todos los datos necesarios, ya que están recogidos en el apartado anterior.

Lo que sí cabe destacar es que este conjunto de datos que vamos a recoger para ambas competiciones nos va a servir en la fase de análisis de los datos para realizar estudios uniendo ambos conjuntos, ya que podría considerarse que ambas competiciones son muy similares y por tanto el modelo de predicción podría ser muy similar.

Más adelante, cuando se explique la fase de análisis de datos, se detallará como se realiza el análisis con los grupos de datos divididos y unidos en un único conjunto de datos.

- **Competición de Liga:**

Al hacer referencia a la competición de liga nos referimos a la competición que se disputa entre diferentes equipos de un país y que juegan en una misma categoría. Para el desarrollo de este proyecto se han tenido en cuenta un conjunto de competiciones de liga que han sido seleccionadas debido a la gran cantidad de información que se puede encontrar de ellas. Las competiciones seleccionadas para este estudio han sido:

- **Liga BBVA** (Primera División Española).
- **Liga Adelante** (Segunda División Española).
- **Ligue 1** (Primera División Francesa).
- **Barclays Premier League** (Primera División Inglesa).
- **Serie A** (Primera División Italiana).
- **Bundesliga** (Primera División Alemana).

Como ya hemos comentado anteriormente, este conjunto de ligas se ha escogido porque, debido a su prestigio (como mejores ligas del mundo), podemos tener a nuestra disposición una gran cantidad de datos de los equipos que forman parte de cada una de ellas.

De cara al análisis, y aunque posteriormente se explicará más en detalle, no se va a recoger la misma cantidad de datos para todas las ligas, ya que se quiere comprobar si se puede generar un modelo único para todas las ligas o si por el contrario cada liga va a necesitar de un modelo específico para realizar las predicciones. Es por ello que se recogerán menos cantidad de datos de la Bundesliga para utilizar esta liga en las pruebas del modelo global de predicción para este tipo de competiciones.

En cuanto a los datos, todos los atributos que se recojan de estas competiciones serán los mismos. Esto es debido a que la similitud que existe entre estas competiciones es tan

grande que no es necesario definir un modelo de recolección de datos diferente para cada una de ellas. Esta similitud reside en el modelo de competición que sigue cada una de ellas.

El formato es sencillo y se podría resumir en que cada equipo tiene que jugar dos partidos contra el resto de equipos de la competición. Por cada partido que gana el equipo recibe tres puntos, mientras que por cada partido que empata el equipo recibe un punto. El ganador de la liga es el que al finalizar la competición tiene más puntos, mientras que los últimos clasificados descienden de división, haciendo que nuevos equipos procedentes de divisiones inferiores jueguen la siguiente temporada de la competición.

Una vez se ha explicado el formato de competición, se van a definir los atributos que a priori podrían facilitar el proceso de clasificación y predicción de resultados. Muchos de los datos que se indicarán ya se han utilizado en las competiciones explicadas anteriormente, por lo que en esos casos no se explicarán de nuevo las razones por las que se ha escogido ese dato. Los datos que se van a recoger para crear el modelo de clasificación de las competiciones de liga son los siguientes:

- Cuotas de las Casas de Apuestas: Una vez más, las tres cuotas de resultados que se ofrecen en los partidos serán recogidas para realizar los análisis posteriores. Al igual que en los anteriores casos, estas cuotas reflejan el estudio previo que han realizado las casas de apuestas sobre el partido y que les permite maximizar las beneficios ajustando las cuotas a las probabilidades que creen que tiene cada uno de los resultados posibles.

- Posición en la Liga: Se ha considerado la posición que ocupa en la liga cada uno de los equipos. Este atributo nos permitirá identificar qué equipos llevan una mejor trayectoria de resultados en la competición, lo que en teoría debería traducirse en un aumento de posibilidades de salir vencedor de un encuentro.

  El dato se recogerá en formato numérico y puede tener distintos valores dependiendo de la competición sobre la que estemos recogiendo los datos. Los posibles valores que puede tomar la posición dependiendo de la competición se resumen en la Tabla 9.

	Primera Posición	Última Posición
Liga BBVA	1	20
Liga Adelante	1	22
Ligue 1	1	20
BP League	1	20
Serie A	1	20
Bundesliga	1	18

Tabla 9. Rango de valores del atributo Posición en la Liga

Como podemos ver en la Tabla 9, los valores que toma el atributo son muy parecidos, pero hay alguna pequeña diferencia en las competiciones de la Liga Adelante y la Bundesliga, donde las ligas no tienen 20 participantes como ocurre en el resto de casos. Esto podrá suponer algún tipo de problema a la hora de analizar los datos, sobre todo para los partidos de la Liga Adelante, ya que habrá muy pocos partidos en el conjunto final de datos que tengan algún equipo en las posiciones 21 y 22. Este problema se ha intentado resolver mediante una discretización de los datos que se expondrá más adelante.

- Tiempo de descanso entre partidos (Partidos entre semana):

Otro de los factores que puede influir en el resultado del partido es el tiempo de descanso que ha tenido cada equipo. Muchos de los participantes de la liga tienen que jugar partidos entre semana de competiciones europeas o de la competición local de copa. Esto hace que los jugadores tengan menos tiempo de descanso, lo que se podría traducir en más cansancio y peor rendimiento para el siguiente partido.

Para la recogida de este dato simplemente identificaremos si cada uno de los equipos ha tenido que jugar algún partido entre semana, sea de la competición que sea. Se recogerá el valor **Sí** cuando el equipo haya jugado entre semana y el valor **No** cuando el equipo no ha jugado ningún partido entre dos partidos de liga.

- Racha de Resultados:

Esta racha de resultados tendrá los mismos formatos y propiedades que las que se han definido en competiciones anteriores. Por tanto, este dato reflejará la racha de resultados de los últimos seis partidos disputados por un equipo tomando un rango de valores que irá desde el 0 (6 partidos seguidos perdidos) y el 12 (6 partidos seguidos ganados).

- Media de Goles anotados en la competición de liga:

El siguiente atributo que se va a definir para nuestro conjunto de datos es la media de goles que mete un equipo en el campeonato de liga. La media será calculada teniendo en cuenta los goles que ha metido un equipo en la presente temporada, y nunca teniendo en cuenta datos de temporadas anteriores.

La inclusión de este dato en nuestro conjunto tiene una razón clara, y es que en este deporte gana el equipo que más goles mete en un partido, por lo que tener una alta media de goles, aumentaría las posibilidades de ganar.

El formato en el que se quiere presentar este dato es un número con hasta dos decimales.

- Media de Goles Recibidos en la competición de liga:

  Este atributo es muy similar al anterior, pero en este caso se calculará la media a partir de los goles que recibe un equipo. Al igual que en el caso anterior las razones por las que se incluye este atributo son claras, y es que un equipo que recibe de media muchos goles tendría menos oportunidades de ganar un partido, ya que estadísticamente necesitaría meter más goles para resultar vendedor de un encuentro.

  El formato en el que se recogerá el dato es igual que el anterior, presentando la información como un número con hasta dos decimales.

- Partido de Alta Rivalidad (Derbi):

  El atributo que se va a explicar a continuación hace referencia a la rivalidad histórica entre equipos. Los partidos catalogados como derbis, son partidos entre dos equipos cuya rivalidad histórica es grande, normalmente debido a que los dos equipos son de la misma ciudad. Por norma general, al ser un partido especial en el que los jugadores suelen rendir a un nivel superior al estar más motivados, los análisis que realizamos para el resto de partidos puede que no sirvan para este tipo de enfrentamientos.

  Lo que se quiere hacer con este atributo es que actúe como flag para identificar partidos que podrían estar fuera de modelo. Por tanto, para la recogida de datos de este atributo se marcará con un **Sí** aquellos partidos entre equipos de la misma ciudad o equipos cuya rivalidad histórica sea grande (por ejemplo, partidos entre Real Madrid y Fútbol Club Barcelona). Se marcará con un **No** el resto de partidos, que en teoría seguirían un modelo de clasificación normal.

  En la fase posterior de análisis de los datos se verá si este atributo nos sirve para mejorar los resultados de las predicciones.

- División:

  Después de haber decidido qué competiciones son las que se van a incluir en nuestro estudio, necesitamos incluir un atributo en nuestro conjunto de datos que identifique la categoría o división de los datos que estamos recogiendo. Aunque sólo una de las competiciones no corresponde a la primera división de un país, es necesario identificar de alguna forma estos casos. Por ello, para aquellos datos de partidos de la Liga Adelante este atributo tendrá el valor 2, mientras que para el resto de ligas tendrá el valor 1.

  Este atributo nos podría ayudar en un futuro a identificar otro tipo de ligas en el caso de que el sistema se amplíe y se añadan ligas de divisiones diferentes a las que hay ahora.

- País de la competición:

  Otro atributo diferenciador que se va a recoger para la competición de liga es el país de la propia competición. Al igual que en el caso anterior, este atributo se va a utilizar para diferenciar a que competición pertenece cada uno de los datos.

  Aunque se quiere intentar generar un modelo global de predicción que sirva para todas las ligas, tanto este atributo como el anterior nos servirán para identificar los datos de cada competición y así poder generar un modelo específico de clasificación para cada una de ellas en el caso de ser necesario.

  Los valores que toma este atributo serán: **España**, **Italia**, **Inglaterra**, **Alemania** y **Francia**.

  Al igual que en las competiciones anteriores se ha considerado la opción de generar atributos derivados de los ya existentes para aumentar la información y simplificarla de cara al proceso de análisis. Los atributos derivados que se van a generar para la competición de liga son:

- Racha Discreta de Resultados:

  Las razones por las que se desea incluir este atributo derivado son las mismas que en competiciones anteriores (gran rango de valores posibles para la racha de resultados). Los valores tras la discretización serán los mismos que se mostraban en la **Tabla 6** de este documento.

- Zona en la clasificación de la liga:

  Debido al amplio rango de valores que puede tener el atributo de clasificación en la liga, se ha decidido distribuir las posiciones dentro de diferentes zonas de la clasificación. Las zonas que se han ideado y sus posiciones correspondientes pueden verse en la Tabla 10.

	Posición Inicial	Posición Final
Champions	1	4
Europa League	5	7
Tranquila	8	10
Media	11	14
Peligro	15	17
Descenso	18	22

Tabla 10. Rango de valores de las zonas en la clasificación de liga

- **Partidos Internacionales (Selecciones Nacionales):**

  En este tipo de encuentros, dos equipos nacionales de fútbol se enfrentarán entre sí en un partido que podrá ser oficial o amistoso.

  Los partidos **amistosos** se refieren a aquellos partidos organizados entre dos federaciones nacionales en los que no hay ningún trofeo o clasificación para un torneo en juego. En estos partidos sólo habrá en juego puntos para la clasificación mundial que elabora la FIFA [26], y que sirve como indicador para ver cómo de bueno es un equipo teniendo en cuenta los resultados de los últimos cuatro años, ponderando con un coeficiente mayor aquellos resultados que se han producido más recientemente.

  En cuanto a los partidos **oficiales**, también son partidos entre dos equipos nacionales, pero en este caso se enmarcan dentro de una competición o ronda preliminar de una competición. Este tipo de competiciones son el Campeonato del Mundo de la FIFA, los torneos continentales (Copa América, Eurocopa, Copa África, etc.), Copa Confederaciones y todas las rondas preliminares de torneos en las que los combinados nacionales participan para obtener la clasificación para alguno de los torneos anteriormente enunciados.

  Una vez situados en el ámbito en el que se van a desarrollar estos partidos los datos que a priori necesitamos para realizar las predicciones van a ser los siguientes:

  - Cuotas de la Casa de Apuestas:

    Al igual que en competiciones anteriores, uno de los elementos claves a la hora de realizar las predicciones, serán las cuotas en las casas de apuestas. Para este caso también se recogerán las cuotas de **victoria local**, **empate** y **victoria visitante**, ya que contienen una información implícita que nos puede servir de gran ayuda a la hora de realizar las predicciones. Sólo tendremos que familiarizarnos un poco con las cuotas que se ofrecen en las casas de apuestas para ver que equipos que son claramente favoritos tienen cuotas de victoria muy bajas, mientras que los equipos más débiles tienen cuotas más altas de lo normal.

  - Ranking en la Clasificación Mundial de la FIFA:

    Como ya hemos explicado anteriormente, la FIFA actualiza mensualmente una clasificación que ella misma elabora y que refleja la situación actual de las selecciones nacionales. Esta clasificación se basa en los resultados de los partidos de los últimos cuatro años, por lo que nos será muy interesante recoger las posiciones en el ranking de los dos equipos que van a participar en un partido. En cada partido internacional que se disputa, cada una de las selecciones nacionales puede conseguir un determinado número de puntos que se calculan mediante la fórmula mostrada en la Figura 16.

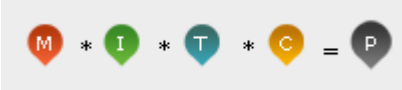

Figura 16. Fórmula para cálculo de puntos en Clasificación Mundial de la FIFA

Los puntos obtenidos al finalizar un partido **(P)** se calculan multiplicando los siguientes parámetros:

M: Se asignan **3 puntos** en caso de victoria y **1 punto** en caso de empate. En caso de derrota se asignan **0 puntos**, por lo que el equipo sumará 0 puntos en la clasificación de la FIFA para ese partido. Si el partido acaba con lanzamientos de penalti, el ganador obtiene **2 puntos** y el perdedor **1 punto**.

I: Este parámetro refleja la importancia del partido, asignando **1 punto** a partidos amistosos, **2.5 puntos** a partidos de clasificación para algún torneo, **3 puntos** para partidos dentro de torneos de la confederación o la copa confederaciones y **4 puntos** para partidos dentro del Mundial de la FIFA.

T: Este parámetro hace referencia a la fuerza de los contendientes, y se calcula restando a 200 la clasificación en el ranking del equipo rival. Excepcionalmente, al líder de la clasificación se le asignan **200 puntos** y a las selecciones clasificadas por debajo del puesto 150 se les asigna un valor mínimo de **50 puntos**. Esto hace que ganar a equipos situados en las primeras posiciones del ranking dé muchos más puntos que ganar a un equipo de la zona baja.

C: Por último este parámetro hace referencia a la calidad de las selecciones que forman parte de una confederación. Mientras que la confederación europea o la sudamericana tienen asignado **1 punto**, otras confederaciones más débiles como la africana o la asiática tienen asignados **0,86 puntos**

- Racha de Resultados:

Otro de los datos que repetimos respecto a otras competiciones es la racha. Como ya habíamos comentado, este atributo a priori podría resultar útil para el modelo de predicción de resultados. El método de cálculo no se variará respecto al explicado en otras competiciones

Una vez tenemos definidos los atributos que consideramos necesarios para la predicción se han definido otros atributos derivados de los primeros que consideramos que pueden facilitar el proceso de clasificación de instancias. Los atributos derivados que hemos definido son:

- Racha Discreta de Resultados:

Al detectar que hay más de cuarenta valores posibles para las rachas de los equipos, se ha decidido acotar los valores posibles a cuatro, agrupando rachas similares dentro de un grupo. La Tabla 11 muestra los valores que se han definido.

	Rango Inferior	Rango Superior
Muy Mala	0	3
Mala	4	6
Buena	6.33	8.99
Muy Buena	9	12

Tabla 11. Correspondencias entre racha numérica y su discretización

Lo que queremos conseguir con esta discretización de los datos es que los algoritmos de clasificación operen con un rango de valores menor, lo que a priori podría facilitar la clasificación de instancias.

- Zona en Ranking Mundial de la FIFA:

  Al haber más de 200 equipos nacionales, el rango de valores que puede tomar el atributo de posición en el Ranking Mundial de la FIFA es muy grande. Para reducir el rango de valores, se ha optado por agrupar a selecciones nacionales que se encuentran en posiciones cercanas en grupos. La Tabla 12 muestra la agrupación que se ha diseñado.

	Posición Superior	Posición Inferior
TOP 10	1	10
TOP 20	11	20
TOP 50	21	50
TOP 100	51	100
TOP 150	101	150
TOP 200	151	200
RESTO	201	207

Tabla 12. Correspondencias entre Posición – Zona del Ranking Mundial de la FIFA

❖ **Baloncesto NBA:**

La siguiente competición que se ha estudiado es la principal liga de baloncesto americana, la National Basketball Association, también conocida como NBA [27].

Esta competición tiene un formato que difiere mucho de las competiciones de baloncesto que se disputan en el resto de países. Para el caso de la NBA, son 30 los equipos divididos en dos conferencias (Este y Oeste) los que forman parte de la competición.

La primera fase (o fase regular) de la competición se disputa bajo un formato de todos contra todos donde cada equipo disputa 82 partidos. Cada equipo jugará cuatro

partidos contra los cuatro equipos más cercanos geográficamente (miembros de la misma división), dos contra equipos que no son de su conferencia y entre tres y cuatro partidos con el resto de equipos. Los ocho mejores equipos de cada conferencia pasarán a la fase final o playoffs.

La fase de playoffs se disputa bajo el formato de eliminatoria al mejor de siete partidos. Los emparejamientos se establecen a partir de las posiciones obtenidas en la fase regular de la competición. La final de la competición tendrá siempre a un equipo de cada conferencia que saldrá a partir de los emparejamientos establecidos al acabar la fase regular y viendo qué equipo a resultado vencedor de todas las eliminatorias anteriores.

Una vez explicado el formato de la competición, seguidamente se analizan algunas de las peculiaridades que posee esta competición. En este caso, la competición que vamos a estudiar es totalmente diferente a las que hemos explicado hasta ahora. En primer lugar cambiamos de deporte y en segundo lugar cambiamos de modelo de resultados, ya que en el baloncesto no es posible que un partido termine en empate. Este hecho a priori es beneficioso, ya que limitar el espacio de resultados implica cometer menos errores en la clasificación. Se tomarán algunos atributos similares a los que se han tomado en las competiciones de fútbol, pero tendremos que buscar otro tipo de atributos que se ajusten a las características del deporte y a las propiedades con las que queremos dotar su modelo.

Los atributos que se han considerado necesarios para realizar las predicciones han sido los siguientes:

- Cuotas de la Casa de Apuestas: Al igual que en el resto de competiciones, las cuotas de la casa de apuestas nos darán una información muy valiosa sobre qué equipo es favorito en un determinado partido. Para este caso sólo habrá que recoger dos cuotas, ya que en este deporte el empate no es posible.

  Al igual que en las competiciones que se han ido explicando, en esta también los equipos que juegan como local suelen tener una mayor probabilidad de ganar un partido. A lo largo de una temporada, más de un 65% de los partidos son ganados por el equipo local, lo que acaba reflejándose en las cuotas. Se podrá ver que un mismo equipo en fechas muy cercanas tendrá una cuota menor si su partido es en su pabellón en comparación a si juega un partido como visitante.

- Porcentaje de Victorias: Este dato refleja el porcentaje de victorias que ha conseguido un equipo sobre el total de partidos que ha disputado en una temporada. La obtención de este dato dotará al modelo de clasificación de una información muy importante sobre lo bueno que es un equipo teniendo en cuenta los partidos que ya ha disputado en una temporada. El dato será recogido en formato porcentaje con una cifra decimal.

- Racha de Resultados: Para este caso en concreto el dato de la racha será calculado de igual forma que en competiciones anteriormente explicadas, con la única salvedad de que en este caso no hay empates.

Debido a la disponibilidad de los datos, para esta competición tenemos para calcular las rachas todos los partidos de la temporada, por lo que no podemos tratar los valores obtenidos para eliminar los números negativos en los modelos. Este atributo, por tanto, no podrá ser utilizado en algunos modelos de clasificación.

- Racha de Resultados como local del equipo que juega en su pabellón:

  Otro tipo de racha que no había podido ser incluida en otras competiciones debido a la disponibilidad de los datos, es la racha que tiene el equipo que juega como local en partidos en su propio pabellón. Este dato nos puede servir para ver la solidez de un equipo cuando juega en su cancha.

  Debido a la importancia de jugar como local, que ya hemos explicado anteriormente con los datos de victorias locales durante la temporada, un equipo con una buena racha de victorias en su campo va a tener más probabilidades de continuar esa racha debido al supuesto buen estado de forma que atraviesan los jugadores.

- Racha de Resultados como visitante del equipo que juega fuera de casa:

  Al igual que en el caso anterior, tenemos fácil acceso a las rachas de cada equipo cuando juegan fuera de su estadio, y este es un dato que nos puede ser muy útil, ya que un equipo que tiene una buena racha de victorias fuera de casa podría neutralizar la ventaja que tiene el equipo local al jugar en su pabellón.

  El método de cálculo tanto de este dato como del anterior sigue la misma fórmula que se ha detallado en el resto de datos de rachas, con la única salvedad de que aquí sí que habrá valores negativos, ya que no existe un mínimo para este parámetro.

- Media de Puntos Anotados por partido: Otro parámetro importante a tener en cuenta en el proceso de recogida de datos es la media de puntos que anota cada equipo de media. Este dato tendrá sólo en cuenta los puntos anotados en los partidos de la temporada en curso. El formato del dato será un número con hasta dos decimales.

- Media de Puntos Recibidos por partido: Es un parámetro similar al comentado anteriormente, sólo que en este caso, el dato que se va a recoger es la media de puntos recibidos por un equipo en los partidos que ha disputado en la temporada en curso. El formato también será igual que el anterior, recogiendo un número con hasta dos decimales.

  La combinación de este dato con el anterior nos ofrecería una información muy valiosa sobre el estilo de juego de un equipo (defensivo u ofensivo), además de poder llegar a intuir el resultado de un partido en el que las medias de los dos equipos son muy dispares.

- Media de Rebotes capturados por partido: Este dato nos indica la media de rebotes que captura un equipo de media por partido. Nos ofrece la capacidad defensiva y de no conceder oportunidades al rival, por lo que en teoría, a mayor número de rebotes capturados por partido, menos oportunidades tendrá el equipo rival para anotar una canasta. El formato que tomará este dato será el de número con hasta dos decimales.

- Porcentaje de Victorias cuando el equipo anota más de 100 puntos:

  Otro dato estadístico que es fácil de encontrar en las estadísticas de la liga, es el porcentaje de victorias que tiene un equipo cuando anota más de 100 puntos. Este dato, combinado junto con la media de puntos que anota un equipo nos puede dar una idea de las probabilidades que tiene un equipo de ganar el partido.

  Si un equipo tiene una media de puntos anotados superior a los 100 puntos y tiene un porcentaje de victorias cuando anota más de 100 puntos alto, lo lógico es que tenga muchas probabilidades de ganar el partido. Además, si el equipo rival recibe de media más de 100 puntos las probabilidades de victoria deberían dispararse.

  El formato en el que se tomará este dato será el de número con hasta un decimal.

- Porcentaje de Victorias cuando el equipo recibe más de 100 puntos:

  Este dato estadístico es similar al anterior, con la diferencia de que el porcentaje calculado hace referencia a las victorias cuando el equipo recibe más de 100 puntos.

  Siguiendo el ejemplo del dato anterior, en este caso si tenemos que un equipo recibe de media más de 100 puntos y además su porcentaje de victorias cuando recibe más de 100 puntos es bajo, esto nos indicará que tendrá pocas probabilidades de ganar el siguiente partido. Si además el equipo tiene una media de puntos anotados inferior a los 100 puntos las probabilidades de ganar por parte de ese equipo serían todavía menores.

  Al igual que el dato anterior, el formato en el que se recogerá el dato será el de número con hasta un decimal.

- Lesiones: Otro de los aspectos que se ha considerado a la hora de realizar las predicciones son las lesiones de jugadores importantes del equipo. En esta competición cada uno de los equipos tiene un quinteto inicial muy definido y que rara vez es modificado. Esos cinco jugadores que comienzan el partido suelen ser los de mayor calidad del equipo, y cuando no disputan un partido el equipo suele bajar su rendimiento.

El formato de competición de la NBA está centrado en conseguir la igualdad de los equipos para favorecer el espectáculo. Por ello, cada año al finalizar la temporada, los equipos peor clasificados escogen en primer lugar los mejores jugadores de institutos y universidades. Esto hace que la competición tienda a igualarse.

Este modelo de competencia en igualdad de condiciones hace que todos los equipos tengan jugadores de gran calidad, por lo que cuando alguno de ellos no juega por alguna lesión sufrida en partidos anteriores, automáticamente el rendimiento del equipo suele bajar.

Por ello también se tendrá en cuenta si de cara a un partido un equipo tiene a alguno de sus jugadores titulares lesionado. Los posibles valores que se han definido para este campo son los siguientes:

- **Sí:** El equipo, con total seguridad no va a contar con alguno de sus jugadores para el siguiente partido.

- **Duda:** Es posible que el equipo no pueda contar con alguno de sus titulares para el siguiente partido.

- **No:** El equipo podrá contar con total seguridad con todos los jugadores para el próximo partido.

La información sobre lesiones está siempre disponible en los dosieres de los equipos que publican en la previa de los partidos en la página oficial de la liga tal y como muestra la Figura 17.

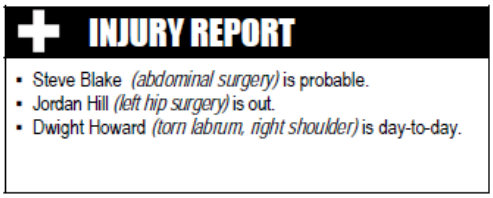

Figura 17. Reporte de lesiones de Los Angeles Lakers (05.02.2013)

Como se puede ver en el reporte de lesiones mostrado, uno de los jugadores titulares del equipo es duda para el siguiente partido (*day-to-day)*, por lo que el dato que se introducirá para su equipo en el apartado lesiones es "Duda".

Al igual que se ha hecho en otras competiciones explicadas anteriormente, para ésta también se han definido atributos derivados de los datos ya definidos. Estos atributos tendrán como objetivo intentar mejorar la clasificación de los algoritmos que utilizaremos en la siguiente fase del proceso de análisis de los datos. Los atributos derivados que se han creado son los siguientes:

- Porcentaje de Victorias Discretizado:

  Como hemos podido ver en otros atributos definidos con anterioridad, uno de los problemas con los que nos podemos encontrar en atributos continuos, es la gran cantidad de valores que pueden llegar a tener. Para discretizar los valores se han diseñado una serie de rangos de valores con los que categorizar cada uno de los porcentajes que obtengamos en la recogida de datos. La tabla de conversión diseñada para traducir los valores se corresponde con la Tabla 13.

	Rango Inferior	Rango Superior
Muy Mal	0%	25%
Mal	25%	35%
Regular	35%	45%
Normal	45%	50%
Bueno	50%	55%
Muy Bueno	55%	65%
Excelente	65%	100%

Tabla 13. Correspondencias entre el porcentaje de victorias y su categorización

- Discretización de Rachas de resultados:

  Ya lo habíamos hecho con otros datos de rachas de otras competiciones, pero ahora que el rango de rachas es mucho más grande que en anteriores casos, tenemos que discretizar el valor de la racha para que el algoritmo de clasificación tenga mejores resultados.

  Esta discretización se va a llevar a cabo para todas las rachas recogidas en esta competición de baloncesto (rachas totales y rachas de partidos como local y visitante). La Tabla 14 muestra los valores y rangos definidos para la discretización.

	Rango Inferior	Rango Superior
Pésima	-∞	-10
Muy Mala	-10	-5
Mala	-5	-1
Regular	-1	1
Buena	1	5
Muy Buena	5	10
Excelente	10	+∞

Tabla 14. Correspondencias entre la racha y su categorización

3.2.2 Recogida de datos

Una vez se han definido las características de los datos que van a ser recogidos para cada una de las competiciones, el siguiente paso es comenzar a recopilar una cantidad suficiente de datos que nos permita comenzar la fase de análisis a partir de la cual generaremos los modelos de predicción de resultados.

El gran problema con el que nos enfrentamos en esta fase es que la recogida de datos no puede realizarse de forma automática. Para cada una de las competiciones la variedad de datos que se recogen impide que se pueda crear un proceso automático sencillo en el que recopilar todos o al menos una gran mayoría de los datos que hemos definido en el punto anterior. Esta dificultad hace que el proceso de recogida de datos tenga que ser efectuado de forma totalmente manual.

Para unificar el proceso de recogida y ya que algunos datos tienen variaciones durante la semana (cuotas de las casas de apuestas o previsión climatológica), se van a recoger los datos el mismo día que se disputa el partido. De esta manera, se asegurará que el dato referente a las casas de apuestas está lo más ajustado posible respecto a la demanda que ha tenido la casa de apuestas.

Las fuentes de datos utilizadas para recoger cada uno de los atributos de las competiciones a estudiar han sido los siguientes:

- **BWIN [28]:** En la página web de la casa de apuestas recogeremos gran cantidad de datos. Espacialmente importantes serán los datos de las cuotas de resultados, que serán recopilados exclusivamente de esta página para que las cuotas siempre guarden relación entre ellas (cada casa de apuestas puede tener un método diferente de cálculo de las cuotas).

  No sólo se recogerán las cuotas de todas las competiciones, sino que debido a la gran cantidad de estadísticas que ofrece la página, de aquí también sacaremos la mayoría de los datos, concretamente los siguientes: tipo de partido, rachas de resultados, posición en la liga, partidos jugados entre semana, media de goles a favor y en contra, división y el país.

- **yr.no [29]**: Para recoger los datos de las previsiones climatológicas se ha utilizado la página web del Instituto Meteorológico de Noruega. La elección de esta fuente es debida a que en dicha página podemos consultar la previsión de cualquier ciudad del planeta por pequeña que ésta sea. Además, la fiabilidad de las previsiones que realiza es tan grande que podemos estar seguros de que el dato que recojamos coincidirá con la climatología en el partido en un gran porcentaje de los casos.

- **UEFA [30]:** Desde la página de la UEFA se recogerán tanto el país al que pertenecen los equipos que participan en competiciones europeas como el Coeficiente UEFA que ha establecido el organismo para cada una de las ligas del continente.

- **FIFA [26]:** Como ya habíamos explicado anteriormente, en el apartado de los partidos internacionales, la posición en el Ranking Mundial de la FIFA será

tomada desde la página oficial del organismo internacional. Esta clasificación se actualiza de manera mensual, recalculando las puntuaciones de cada selección nacional en función de los últimos resultados registrados.

- **NBA [27]:** Todos los datos referentes a la liga americana de baloncesto (excepto las cuotas de la casa de apuestas) serán recogidos desde la sección de estadísticas de la página de la NBA. En dicha página podremos encontrar todas las estadísticas necesarias desglosadas por equipos, lo que nos facilitará la localización de los datos para incorporarlos al conjunto. Además, en los reportes previos que publican los equipos antes de cada partido podremos ver si los equipos cuentan con la baja de alguno de sus jugadores más importantes para el siguiente partido.

Después de identificar los atributos necesarios y la fuente desde la que se van a recoger, se puede comenzar el proceso de recogida de datos. Este proceso durará varios meses, ya que cada semana sólo se pueden recoger como máximo 10 partidos de cada liga de futbol, entre 16 y 24 partidos de competiciones europeas y unos 20 partidos de la NBA. Respecto a los partidos internacionales, no todas las semanas hay partidos de este tipo, por lo que hay que estar especialmente atento a las fechas en las que estos partidos se disputan, ya que hay oportunidades escasas durante el año para recoger información sobre estos encuentros.

3.2.3 Conjuntos de Entrenamiento

Después de concluir con la fase de recogida de datos, el siguiente paso que ha de darse es la definición de los conjuntos de entrenamiento correspondientes a cada competición.

Dependiendo de la frecuencia con la que se disputan los partidos de una determinada competición, se han conseguido elaborar conjuntos de entrenamiento más o menos grandes. Los conjuntos con mayor tamaño son los que pertenecen a la competición de liga de los distintos países estudiados. Algo menos voluminosos son los conjuntos de entrenamiento de las competiciones europeas. Esto es debido a que estas competiciones son disputadas por menos equipos que las ligas y que además no todas las semanas se disputan partidos.

Respecto a los partidos internacionales, se ha tenido que invertir más tiempo que el inicialmente estimado para recoger un conjunto de datos lo suficientemente grande. Las competiciones disputadas durante la época de verano (Mundial y Eurocopa) han permitido recoger un volumen de datos suficientes para realizar un análisis.

Finalmente, el conjunto que falta comentar es el de partidos de la NBA. Respecto a este conjunto cabe destacar que no es tan voluminoso como el resto de conjuntos, pero para cada partido se han recogido una gran cantidad de atributos. Aunque el volumen de partidos no es muy grande (tan sólo 175 partidos) este conjunto nos permitirá sentar las bases de un modelo de predicción, que posteriormente podría ser mejorado alimentando al modelo con más partidos para que sus parámetros sean recalculados.

Con la idea de generar modelos globales de predicción en competiciones similares, se ha decidido agrupar datos de competiciones similares para generar conjuntos de entrenamientos más grandes que ayuden a realizar mejores predicciones de dichas competiciones. Esta agrupación se realizará con los siguientes conjuntos de entrenamiento:

- **Conjunto de Partidos Europeos:** Este conjunto saldrá de la unión de los conjuntos de entrenamiento de los partidos de **Champions League** y de **Europa League**. La razones de esta unión son la similitud entre los conjuntos (ambos conjuntos están formados por los mismos atributos) y que las dos competiciones tienen un formato parecido, lo que hace que los partidos de todas las fases sean similares para las dos competiciones.

- **Conjunto de Partidos de Liga:** En este caso se unirán para formar este conjunto los registros de partidos de la **Liga BBVA**, **Liga Adelante**, **Serie A**, **Premier League**, **Ligue 1** y **Bundesliga**. Para cada uno de los subconjuntos se ha recogido un número diferente de datos, lo que posteriormente nos permitirá comprobar si el modelo global que se quiere generar funciona igual de bien para todas las competiciones.

Como resumen final de este apartado, la Tabla 15 muestra las características principales de cada uno de los conjuntos de entrenamiento.

	Número de Partidos Analizados	Número de Atributos por Partido
Champions League	136	17
Europa League	226	
Partidos Europeos	362	
Liga BBVA	298	23
Liga Adelante	203	
Ligue 1	202	
Serie A	159	
BP League	148	
Bundesliga	80	
Partidos de Liga	1090	
Partidos Internacionales	505	18
NBA	175	26

Tabla 15. Partidos y Atributos analizados por competición

3.2.4 Ficheros ARFF para Conjuntos de Entrenamiento

Una vez tenemos definidos los conjuntos de entrenamiento, tenemos que construir los ficheros .arff para poder iniciar el análisis de datos en la herramienta WEKA.

El formato de entrada de los ficheros WEKA es el .arff. Su estructura es muy sencilla, lo que hace que se puedan generar estos ficheros de forma muy fácil a partir de una hoja Excel. El fichero .arff se divide en las siguientes partes:

- Nombre de la relación: La primera línea del fichero .arff tendrá el nombre con el que se identifica el conjunto de datos que define el fichero. El nombre del conjunto irá precedido del texto **@relation**, como puede apreciarse en la Figura 18.

```
@relation NBAEntrenamiento
```

Figura 18. Definición del nombre para el Conjunto de Entrenamiento NBA

- Atributos: La siguiente parte del fichero está dedicada a la definición de los atributos que van a formar parte del conjunto de entrenamiento. Cada uno de los atributos irá precedido del texto **@attribute**. Para los conjuntos de entrenamiento que vamos a definir se van a definir dos tipos de atributos: **atributos numéricos** y **atributos discretos**. Los atributos de tipo numérico serán definidos al colocar la palabra **real** después del nombre del atributo. Mientras tanto, los atributos discretos tendrán el conjunto de valores que puede tomar el atributo entre llaves justo después del nombre del atributo (ver Figura 19).

```
@attribute coutaV real
@attribute cuotaL real
@attribute porcV real
@attribute procL real
@attribute porcDiscV {MuyMal, Mal, Regular, Normal, Bueno, MuyBueno, Excelente}
@attribute procDiscL {MuyMal, Mal, Regular, Normal, Bueno, MuyBueno, Excelente}
@attribute rachaV real
@attribute rachaDiscV {Pesima, MuyMala, Mala, Regular, Buena, MuyBuena, Excelente}
@attribute rachaL real
@attribute rachaDiscL {Pesima, MuyMala, Mala, Regular, Buena, MuyBuena, Excelente}
@attribute rachaFuera real
@attribute rachaDiscFuera {Pesima, MuyMala, Mala, Regular, Buena, MuyBuena, Excelente}
@attribute rachaCasa real
@attribute rachaDiscLocal {Pesima, MuyMala, Mala, Regular, Buena, MuyBuena, Excelente}
@attribute anotaV real
@attribute anotaL real
@attribute recibeV real
@attribute recibeL real
@attribute rebotesV real
@attribute rebotesL real
@attribute victorias+100V real
@attribute victorias+100DiscV {MuyMal, Mal, Regular, Normal, Bueno, MuyBueno, Excelente}
@attribute victorias+100L real
@attribute victorias+100DiscL {MuyMal, Mal, Regular, Normal, Bueno, MuyBueno, Excelente}
@attribute victoriasOp+100V real
@attribute victoriasOp+100DiscV {MuyMal, Mal, Regular, Normal, Bueno, MuyBueno, Excelente}
@attribute victoriasOp+100L real
@attribute victoriasOp+100DiscL {MuyMal, Mal, Regular, Normal, Bueno, MuyBueno, Excelente}
@attribute lesionV {Si, No, Duda}
@attribute lesionL {Si, No, Duda}
@attribute res {1,2}
```

Figura 19. Definición de atributos para el Conjunto de Entrenamiento NBA

- Datos: La última parte del fichero es en la que se reflejan los datos de todos los partidos del conjunto. Cada partido estará en una línea, y cada atributo irá separado por una coma. Justo antes de empezar la definición de los datos, el fichero tendrá que tener el texto **@data** para que la herramienta WEKA localice dónde comienza la zona de datos (Figura 20).

```
@data

1.5,3.7,5,17,9,PEL,TRANQ,No,No,No,-1,0.99,5,6.99,Mala,Buena,1.33,0.92,1.51,0.94,No,2
2.2,2.9,3,6,19,EL,DESC,No,No,Si,-1,-1,5,5,Mala,Mala,1.33,0.82,1.12,0.97,No,X
2.85,2.7,2.45,16,1,PEL,CHAMP,No,No,Si,3.33,1,9.33,7,MuyBuena,Buena,1.1,1.2,1.46,0.82,No,2
2.1,2.7,3.5,10,14,TRANQ,MEDIA,No,No,No,-1,-3,5,3,Mala,MuyMala,0.89,0.97,1.02,1.28,No,1
2.2,3,2.9,7,20,EL,DESC,No,No,Si,-1,1,5,7,Mala,Buena,1.56,1.15,1.3,1.25,No,2
1.75,3.2,4,12,13,MEDIA,MEDIA,No,No,Si,1,0.33,7,6.33,Buena,Buena,1.23,1.1,1.2,1.41,No,2
1.3,4.2,8,4,8,CHAMP,TRANQ,No,No,No,3.66,1,9.66,7,MuyBuena,Buena,1.38,1.33,0.94,1.2,No,1
1.75,3.2,4,15,11,PEL,MEDIA,No,No,No,-1,-2,5,4,Mala,Mala,1.51,0.84,1.38,1.1,No,1
3,2.9,2.2,21,2,DESC,CHAMP,No,No,Si,-1,4.33,5,10.33,Mala,MuyBuena,0.97,1.51,1.33,0.97,No,1
2.4,2.9,2.7,5,3,EL,CHAMP,No,No,Si,1,2.66,7,8.66,Buena,Buena,1.48,1.46,1.12,0.84,No,X
1.4,3.9,6,18,11,DESC,MEDIA,No,No,No,-2,1.33,4,7.33,Mala,Buena,1.32,1.1,1.55,1.37,No,1
4,3.3,1.7,9,5,TRANQ,EL,No,No,Si,1.99,1.33,7.99,7.33,Buena,Buena,0.97,1.45,0.95,1.1,No,1
1.27,4.3,8.5,4,12,CHAMP,MEDIA,No,No,Si,2.99,1,8.99,7,Buena,Buena,1.42,1.55,0.82,1.4,No,1
2.6,3.1,2.3,13,19,MEDIA,DESC,No,No,Si,-3,1,3,7,MuyMala,Buena,0.87,1,1.15,1.32,No,1
1.17,5.5,11,2,22,CHAMP,DESC,No,No,Si,-1,-3,5,3,Mala,MuyMala,1.5,0.9,1,1.45,No,1
3.8,3.4,1.75,15,3,PEL,CHAMP,No,No,Si,3.33,4.66,9.33,10.66,MuyBuena,MuyBuena,1.02,1.4,1.3,0.92,No,X
2.3,2.8,2.9,20,17,DESC,PEL,No,No,Si,0.33,-1,6.33,5,Buena,Mala,0.82,1.1,0.97,1.5,No,1
1.4,3.9,6,1,10,CHAMP,TRANQ,No,No,No,2,1,8,7,Buena,Buena,1.25,0.95,0.82,1.02,No,1
1.45,3.8,5.4,16,7,PEL,EL,No,No,Si,-4,-2,2,4,MuyMala,Mala,0.97,1.55,1.32,1.32,No,1
1.65,3.2,4.6,21,14,DESC,MEDIA,No,No,Si,2,-1,8,5,Buena,Mala,1.17,1.2,1.25,1.2,No,1
5,3.7,1.5,5,16,EL,PEL,No,No,Si,-1,1,5,7,Mala,Buena,1.41,1.36,1.09,1.51,No,2
```

Figura 20. Datos de partidos para el Conjunto de Entrenamiento Liga Adelante

3.2.5 Atributos relevantes para la clasificación

Después de formar los conjuntos de entrenamiento para cada una de las competiciones, el siguiente paso es ver cuáles de los atributos son realmente relevantes para la clasificación. A partir de esta fase la herramienta WEKA será la protagonista, ya que servirá como apoyo a todo el proceso de análisis de los datos.

Antes de comenzar a probar qué clasificadores son los que mejor se ajustan al tipo de problema que se está planteando, tenemos que analizar si los atributos que hemos recogido son realmente necesarios para la predicción, es decir, si aportan valor al algoritmo de clasificación o no. Para realizar esta tarea utilizaremos la opción de ***Select Attributes*** que ofrece WEKA. Mediante esta opción y utilizando los algoritmos adecuados se podrá hacer una primera criba de atributos que no aportarán nada a los algoritmos de clasificación, por lo que podrían ser eliminados del conjunto de entrenamiento.

Los dos algoritmos que van a ser utilizados para realizar el estudio de la relevancia de cada atributo van a ser:

- **ChiSquaredAttributeEval:** Calcula el valor estadístico Chi-cuadrado de cada atributo con respecto a la clase y así obtiene el nivel de correlación entre

la clase y cada atributo. Esto mostrará que atributos son realmente importantes para realizar las predicciones.

- **GainRatioAttributeEval:** Evalúa cada atributo midiendo su razón de beneficio con respecto a la clase. Será la segunda opción a tener en cuenta a la hora de eliminar atributos del conjunto de entrenamiento.

Para cada uno de los conjuntos de entrenamiento se aplicarán los dos algoritmos presentados arriba combinados con el método de búsqueda **Ranker**, que devuelve una lista ordenada de atributos según la calidad de éstos. El criterio que se seguirá en esta fase para descartar atributos no relevantes, será que un atributo tenga una puntuación de 0 puntos al ser evaluado con los dos algoritmos expuestos anteriormente.

Dado el volumen de datos y tablas que se maneja en esta sección, se ha dispuesto una sección en el anexo final del documento llamada ***ANEXO A: Estudio de Relevancia de Atributos***, en la que pueden consultarse todas las tablas y comentarios del estudio de relevancia de atributos que se ha llevado a cabo.

Para la correcta interpretación de las tablas, hay que especificar que los atributos marcados en rojo son los que se han descartado para la siguiente fase del proyecto, ya que no poseen ninguna correlación con la clase por la que vamos a clasificar.

3.2.6 Entrenamiento de los conjuntos a través de clasificadores

Tras haber realizado un análisis previo en el que se ha intentado descartar los atributos menos significativos, el siguiente paso a dar es el entrenamiento de los conjuntos con diferentes algoritmos de clasificación. Se utilizarán varios algoritmos de clasificación con el objetivo de encontrar el que mejor se ajusta para predecir los resultados de cada conjunto. Como ya se comentó en apartados anteriores, se utilizará la herramienta WEKA para realizar los entrenamientos de los conjuntos a través de los algoritmos que tiene implementados.

Antes de comenzar a exponer los resultados que se han obtenido con cada uno de los algoritmos se va a proceder a enumerar los algoritmos que se van a utilizar y a explicar brevemente las características de cada uno de ellos. Los algoritmos que se van a utilizar en el proyecto son los siguientes:

- **Red Bayesiana (Bayes Net) [31]:**

Una Red Bayesiana es un grafo dirigido acíclico que codifica una distribución de probabilidad conjunta de un grupo de variables aleatorias. Si en la estructura del grafo que representa la Red Bayesiana se encuentra un arco dirigido desde el nodo A hacia el nodo B, se dice que A es padre de B o que B es hijo de A.

Cada arco indica una relación causal, y cada variable independiente definida en la red es independiente de sus no-descendientes dados sus padres.

Además del grafo que representa a la Red Bayesiana cada uno de los nodos de la red posee una tabla de probabilidad condicionada, que relaciona cada variable con sus padres.

- **Regresión Logística (Logistic) [32]:**

La Regresión Logística es un tipo de análisis de regresión que se utiliza para predecir el resultado de una variable en función de un conjunto de variables independientes o predictoras.

La técnica que sigue este clasificador es abordar el problema creado una variable ficticia binaria para representar la pertenencia o no a cualquiera de los dos grupos de observación. Para garantizar que la probabilidad de respuesta está entre 0 y 1 habrá que transformar la variable de respuesta para que siempre se encuentre entre los dos valores deseados.

Para realizar las predicciones, el algoritmo se basará en regresiones sobre las variables independientes para realizar los cálculos de probabilidad.

- **Perceptrón Multicapa (Multilayer Perceptron)[33]:**

Este clasificador surge a partir de las limitaciones que presenta el Perceptrón Simple a la hora de resolver problemas no lineales. Las investigaciones hicieron ver que la combinación de varios perceptrones podría resolver algunos problemas no lineales. La asignación de pesos de las capas ocultas y la retropropagación de errores serían claves para la creación del Perceptrón Multicapa.

La estructura de este clasificador está formada por tres tipos de capas. La **capa de entrada** que se encarga de recibir las señales de entrada y propagarlas hacia la siguiente capa. La **capa de salida** proporciona para cada patrón de entrada una respuesta de la red. La(s) **capa(s) oculta(s)** se encargan de realizar un procesamiento no lineal de los datos recibidos. Cada una de las neuronas de la red suele estar interconectada con todas las neuronas de la siguiente capa.

Para definir la estructura del perceptrón será necesario definir una **función de activación** que determine cuándo se activan las neuronas de cada una de las capas, un **número de neuronas en las capas de entrada y salida** que vendrá dado según el número de variables del problema y el **número de capas y neuronas ocultas** que debe ser determinado por el diseñador de la red, no habiendo un método que nos diga cuál es el número correcto para cada problema.

- **OneR [34]:**

Este clasificador es uno de los más sencillos y rápidos. Para realizar la clasificación de las instancias observadas selecciona el atributo más significativo del conjunto y realiza las predicciones basándose únicamente en los valores de este atributo. Este algoritmo aunque sea muy simple puede llegar a conseguir mejores resultados que otros algoritmos más complejos.

- **J48 [35]:**

El algoritmo J48 que utiliza WEKA es una implementación del algoritmo C4.5, que es uno de los más utilizados en minería de datos. El algoritmo C4.5 genera árboles de clasificación a partir de un conjunto de entrenamiento y a partir del concepto de **entropía de la información** [36]. Este concepto se puede considerar como la cantidad de información que contienen los símbolos usados para representar un dato. Los símbolos con menos probabilidad son los que más información aportan al modelo.

En cada nodo que genera el algoritmo C4.5, se escoge el atributo que divide más eficazmente el conjunto restante en subconjuntos que serán posteriormente clasificados en otro nodo.

Este algoritmo es una mejora del algoritmo ID3, permitiendo en el caso del C4.5 trabajar con datos discretos, lo que es indispensable para la realización de nuestro proyecto.

- **Random Forest[37]:**

El **Random Forest** es una combinación de árboles predictores en los que cada uno de los árboles que lo compone depende de los valores de un vector aleatorio probado independientemente y que posee la misma distribución para cada uno de los árboles.

Esta técnica es muy similar al ***Bagging*** (entrenamiento con varios clasificadores en la que después se promedian los resultados de cada uno de ellos para determinar el resultado final). Para este caso además se escogen atributos aleatorios para el entrenamiento de cada uno de los árboles.

El resultado final de la clasificación se obtendrá a partir de los resultados generados por cada uno de los clasificadores que forman el Random Forest.

Una vez presentados los clasificadores que se van a utilizar para el entrenamiento de los conjuntos elaborados para cada una de las competiciones se va a proceder al entrenamiento para ver cuál de los clasificadores anteriormente citados se ajusta mejor a cada uno de los problemas. La idea inicial es escoger dos algoritmos por cada uno de los conjuntos de competición basándonos en el porcentaje de aciertos que obtenemos con cada uno de ellos y teniendo en cuenta el tiempo que emplean en generar la predicción, ya que en determinados casos como el del Perceptrón Multicapa podemos tener tiempos de predicción muy altos.

Para entrenar los conjuntos a través de los clasificadores que se han escogido se va a utilizar la opción **cross-validation** (validación cruzada) que ofrece WEKA, y más concretamente se utilizarán cinco subconjuntos para el entrenamiento. De este modo la herramienta WEKA dividirá el conjunto inicial en cinco conjuntos para posteriormente usar cuatro conjuntos para el entrenamiento y uno para los tests. Este procedimiento se repetirá cinco veces haciendo que cada subconjunto sea una vez utilizado para el test. El resultado final será un promedio de todas pruebas intermedias. Esta técnica es muy útil para conjunto de datos pequeños ya que permite realizar una prueba más amplia con un conjunto de datos relativamente pequeño.

Cabe recordar también que los conjuntos de entrenamiento no tendrán los atributos que fueron eliminados en la anterior fase, donde se comprobaba la correlación de cada uno de los atributos con la clase. Para escoger el conjunto final de atributos que formarán parte del modelo de predicción se irán realizando pruebas en las que para cada algoritmo se irán utilizando diferentes conjuntos de atributos hasta encontrar el que mejores porcentajes de aciertos genere.

Nuevamente, al igual que ocurrió con el estudio de relevancia de atributos, el volumen de datos y tablas que arroja el estudio del entrenamiento ha hecho que se decida presentar el resultado de cada uno de los entrenamientos de cada competición en el anexo incluido al final del documento llamado ***ANEXO B: Entrenamiento de conjuntos a través de clasificadores***.

Como resumen de esta sección podemos destacar el buen comportamiento de las **Redes Bayesianas** y los **Árboles J48**. En todas las competiciones, estos dos clasificadores han sido los que mejores resultados han obtenido en cuanto a tasa de aciertos. Este buen comportamiento hace que sean los elegidos para encargarse de la predicción de los resultados en todas las competiciones estudiadas.

3.3 Clasificadores escogidos para el Sistema

En esta última sección del Capítulo 3, se recopilan todos los clasificadores que han sido seleccionados para ser implementados en la sección anterior y se muestran las características que poseen cada uno de ellos, mostrando el árbol resultante en el caso de clasificadores basados en el algoritmo J48 y los coeficientes de los nodos en el caso de que el clasificador seleccionado haya sido una red bayesiana. La información se irá presentando dividida en los conjuntos de entrenamiento utilizados.

Cabe recordar también que todos los clasificadores serán o árboles J48 o Redes Bayesianas debido a los buenos resultados que han mostrado en la fase de entrenamiento para todos los conjuntos de datos.

Debido a la gran cantidad de gráficos y tablas que incluye esta sección, todos los datos referentes a los clasificadores escogidos para que se implementen en el Sistema de Preedición serán recogidos al final del documento en el ***ANEXO C: Clasificadores escogidos para el sistema.***

Para resumir el contenido que puede encontrarse en el anexo, se ha diseñado una estructura en la que para cada competición aparecerá la imagen del Árbol J48 que se va a implementar en la hoja de predicción junto con la tabla de coeficientes de la red bayesiana que también va a implementarse en la misma hoja.

Tanto el árbol como los coeficientes de la red han sido extraídos de la herramienta WEKA, por lo que los algoritmos de clasificación que se implementarán serán los mismos que los utilizados para el entrenamiento de los conjuntos en WEKA.

Además, junto con la tabla de coeficientes de la red bayesiana se ha adjuntado los tramos en los que se han dividido ciertos atributos continuos que la red ha discretizado para obtener mejores resultados en la clasificación. Estos tramos también han sido extraídos del análisis realizado con la herramienta WEKA.

3.4 Implementación de la Hoja de Predicción

3.4.1 Introducción

Después de haber elegido el tipo de clasificadores que se van a utilizar para la predicción de cada uno de los conjuntos, el siguiente paso es diseñar el soporte sobre el que se van a realizar las predicciones. En este caso el soporte sobre el que se van a realizar las predicciones va a ser una **Hoja de Excel**. Las razones por las que se ha optado por este soporte son las siguientes:

- La mayoría de usuarios conocerán el funcionamiento de esta herramienta ya que suelen estar habituados a utilizarla tanto en un entorno doméstico como laboral.

- Sencillez a la hora de introducir datos. El sistema de celdas implementado en las hojas Excel permite una fácil navegación entre ellas que se traduce en rapidez a la hora de introducir datos.

- Posibilidad de calcular los atributos derivados a través de las fórmulas que proporciona la propia herramienta.

- El módulo de Visual Basic que incluye la herramienta, permitirá la implementación de los algoritmos de predicción probados en el apartado anterior dentro de la propia hoja de predicciones.

- La posibilidad de tener los datos y los algoritmos dentro de un mismo entorno hace que el tiempo de ejecución de los algoritmos se vea reducido.

- En el caso de que el modelo de predicción sufra algún cambio debido a su evolución o si queremos introducir una nueva competición, el impacto será mínimo, ya que cada competición es independiente respecto al resto. Esto es debido a que para cada competición tenemos definida una pestaña de la hoja y una función en Visual Basic que emula los algoritmos elegidos para realizar la predicción.

Como se puede ver, las ventajas de utilizar una hoja de Excel como soporte de las predicciones son lo bastante relevantes como para tomar la decisión de coger este soporte. La estructura de pestañas y macros que se ha ideado para esta hoja Excel es la siguiente:

- ❖ Una pestaña para cada competición. Permite que cada competición esté completamente separada del resto. Facilita también la ejecución de algoritmos sobre conjuntos de partidos concretos.

- ❖ Una macro para la predicción de resultados de cada competición. Cada una de estas macros contendrán los algoritmos de predicción definidos para cada competición.

- ❖ Una pestaña llamada **Coeficientes UEFA** con la tabla de equivalencias entre el país al que pertenece un equipo y su Coeficiente UEFA y TOP en el ranking. Esta tabla permitirá que el dato del Coeficiente UEFA y el TOP en el ranking se rellene automáticamente al introducir el país de un equipo en un partido de competición europea.

- ❖ Una pestaña llamada **Tabla Bayes** que contendrá todos los coeficientes necesarios para crear la Red Bayesiana que sacará el resultado más probable de los partidos de cada competición.

Para entrar más en detalle, se pasará a explicar en profundidad cada una de las partes enunciadas anteriormente, desde su fase de diseño hasta la fase final en la que tenemos los resultados.

3.4.2 Pestañas de la Hoja de Predicción

- Pestañas de competiciones:

Se han creado diez pestañas en el archivo Excel, nueve de las cuales corresponden a las nueve competiciones que se han estado analizando en fases anteriores. La otra pestaña restante corresponde a la que va a ser utilizada para probar si el modelo global que se ha generado para la competición de liga es válido para partidos que no corresponden a los países analizados.

Cada una de las pestañas tendrá unas columnas específicas para los atributos definidos como necesarios en el análisis por clasificadores realizado en la herramienta WEKA. El usuario introducirá a mano los valores de los atributos que se indican en las cabeceras de columna (excepto para valores autocalculados como los Coeficientes UEFA, la discretización de resultados o zonas en la clasificación en la liga). La zona de datos estará siempre separada por una gruesa línea que la diferenciará de la zona de resultados.

En la zona de resultados tendremos dos columnas por cada clasificador que se utilice para realizar predicciones. Estas dos columnas corresponderán al resultado que ofrece el clasificador y al riesgo estimado de que ese resultado sea el correcto. Para cada competición tendremos cuatro u ocho columnas reservadas para los datos del clasificador, dependiendo si esa competición utiliza un modelo de predicción con dos o cuatro clasificadores.

Finalmente, tendremos dos columnas que reflejarán el veredicto final sobre el resultado y el riesgo de tomar ese resultado como el que se va a dar al final de un partido.

Estas dos columnas serán calculadas mediante fórmulas que tendrán como parámetros de entrada el resto de celdas de predicción.

El sistema que se ha definido para establecer el resultado de estas columnas es el siguiente:

- Predicción:

  Como ya habíamos definido en la sección anterior, cada una de las competiciones tendrá al menos dos clasificadores que devolverán una predicción y un riesgo siguiendo el algoritmo de clasificación correspondiente. Lo que se ha desarrollado para predecir el resultado de un partido es que haya un "**sistema de votaciones**", en el que cada algoritmo emite un "**voto**" sobre el resultado de un partido.

  A la hora de realizar una predicción, lo que buscamos es que la predicción sea lo más segura posible, es decir, que si apostamos por ella tengamos grandes posibilidades de sacar un beneficio. Por ello, para establecer la predicción final miraremos el "voto" que ha emitido cada clasificador y sólo emitiremos un resultado para ese partido si hay unanimidad en la predicción de cada uno de los clasificadores.

  En resumen, si todos los clasificadores llegan a predecir el mismo resultado, éste será tomado como predicción final. En el caso de que haya discrepancias entre clasificadores, ese partido será declarado nulo y por tanto no se ofrecerá un resultado final para él.

  Esta técnica en la que se unen varios clasificadores es muy similar a la técnica de **Bagging (*Bootstrap Aggregating*)** que ya se presentó en secciones anteriores al hablar de los Random Forest. Esta técnica consiste en generar diversos conjuntos de entrenamiento a partir de uno original y entrenar cada uno de ellos a través de un clasificador distinto. La clase final por la que se va a clasificar se decide por votación.

  La diferencia que existe entre esta técnica de Bagging y la que se va a utilizar en este proyecto es que aquí entrenamos el mismo conjunto con distintos algoritmos, y no distintos subconjuntos con distintos algoritmos. En cualquier caso, la fase final de la técnica de Bagging en la que se decide por votación el resultado final de la predicción ha sido incorporada a nuestra técnica de predicción de resultados.

  Para calcular el valor de la predicción se usará una fórmula condicional que verifique que todas las predicciones son iguales. En el caso de serlo, colocará como resultado final la predicción común, mientras que si hay discrepancias colocará el símbolo "-", lo que significaría que no se ofrece predicción para ese partido al haber mucho riesgo en la predicción de un resultado concreto.

❖ Riesgo:

Después de establecer la predicción final de un partido, faltaría por definir el riesgo que supondría tomar esa predicción como resultado final del partido. Para determinar este valor utilizaremos otra fórmula de Excel que nos permitirá ponderar los valores del riesgo que ofrecen cada uno de los clasificadores. En el caso de que el partido se haya declarado como nulo (-) no se calculará valor para el riesgo.

Los pesos por los que ponderaremos los riesgos parciales para obtener el riesgo total serán del **60%** del riesgo para el riesgo parcial de la **Red Bayesiana** y un **40%** para el riesgo parcial del **Árbol J48**.

Las razones por las que se elige un peso tan alto para el valor dado por el clasificador bayesiano es que el valor de riesgo que nos ofrece el árbol J48 no es muy preciso, sobre todo para árboles que están muy ramificados, ya que pueden llegar a tener nodos hoja en los que muy pocos partidos caigan ahí y tengan un valor del riesgo mucho menor que el que de verdad tienen.

Los pesos de 80%-20% se utilizarán para competiciones en la que sólo haya dos clasificadores para la predicción. En el caso de que haya cuatro clasificadores se repartirán los pesos entre los clasificadores del mismo tipo, teniendo un 40% de peso cada clasificador bayesiano y un 10% cada árbol J48.

- Pestaña de Coeficientes UEFA:

Esta pestaña servirá como repositorio de datos y nos facilitará la automatización de algunas tareas de introducción de datos en las correspondientes pestañas. La pestaña contendrá una tabla exactamente igual que la mostrada en la **Tabla 7** de este proyecto y lo que persigue es que los datos de Coeficiente UEFA y el TOP del país en el Ranking UEFA sean datos que se completen automáticamente al escribir el país al que pertenece un equipo.

Para que estos valores se completen automáticamente se utilizará la función de Excel *BuscarV*, que permite realizar búsquedas de datos en tablas que pueden estar en pestañas distintas a la que se está trabajando.

- Pestaña Tabla Bayes:

Esta pestaña servirá como repositorio de datos y en ella se almacenarán los coeficientes de cada uno de los clasificadores bayesianos que se van a definir para realizar las predicciones. Estos coeficientes serán consultados a través de las macros implementadas en Visual Basic para calcular las probabilidades de cada resultado dentro de un mismo partido.

Para que la consulta de estos coeficientes a través de la macro implementada sea lo más rápida posible se ha ideado una estructura de datos que optimiza la consulta de los coeficientes. Dicha estructura tendrá la siguiente forma, que puede verse en la Figura 21:

- En la primera columna sólo podrá escribirse el nombre de una competición. Desde esa fila hasta que aparezca otro nombre de competición diferente, todos los datos que haya entre medias pertenecerán a un clasificador bayesiano de esa competición.

- En la segunda columna sólo podrán ser escritos nombres de atributos. Desde esa fila hasta que aparezca el nombre de otro atributo o competición, todos los datos que haya entre medias pertenecerán al atributo en cuestión.

- En la tercera columna sólo podrán ser escritos los nombres de los grupos en los que se divide cada atributo. Estos grupos salen directamente del análisis realizado en WEKA de los conjuntos entrenados a través de una Red Bayesiana. Para cada uno de estos grupos aparecerán las probabilidades estimadas por WEKA de que ocurra cada uno de los resultados.

Liga Adelante					
	CL		1	X	2
		Baja	0.134	0.01	0.011
		Alta	0.866	0.99	0.989
	CX		1	X	2
		Baja	0.866	0.99	0.989
		Alta	0.134	0.01	0.011
	ZL		1	X	2
		CHAMP	0.223	0.157	0.094
		EL	0.123	0.157	0.198
		TRANQ	0.159	0.12	0.135
		MEDIA	0.123	0.157	0.135
		PEL	0.15	0.139	0.156
		DESC	0.223	0.269	0.281

Figura 21. Coeficientes de la Liga Adelante para la construcción de la Red Bayesiana

3.4.3 Macros

Las macros son un grupo de instrucciones programadas bajo el entorno de Visual Basic for Aplications, cuya tarea principal es la automatización de tareas repetitivas y la resolución de cálculos complejos. Para realizar las tareas de predicción sobre cada una de las competiciones, se ha creado una macro por cada competición. Dicha macro tendrá implementados los algoritmos de clasificación que han sido probados en la herramienta WEKA y que han obtenido los mejores porcentajes de acierto en la clasificación.

Los datos de entrada de cada uno de los algoritmos que implementen las macros estarán presentes en la propia hoja Excel de la competición, y es el usuario el que ha debido insertar los datos necesarios antes de poder ejecutar la macro correspondiente.

Dichos datos serán almacenados en variables creadas en la propia macro y a partir de ahí se realizarán los cálculos y comprobaciones que necesite cada uno de los algoritmos.

Al haber escogido únicamente dos tipos de algoritmos de clasificación la implementación de todos ellos es muy similar para todos los casos. Tan solo variarán las comprobaciones realizadas sobre determinadas variables y la estructura de datos que hay que consultar en la Tabla Bayes para calcular las probabilidades de los distintos resultados posibles. Dada la similitud entre los algoritmos del mismo tipo para cada competición, se va a explicar a continuación la implementación que se ha realizado para un ejemplo de Árbol J48 y de Red Bayesiana.

3.4.4 Implementación de un Árbol J48

El tipo más sencillo de clasificador que vamos a implementar es el Árbol J48. Para implementar este clasificador tan solo hará falta ver la imagen del árbol que nos ofrece WEKA y emularlo en la macro a través de sentencias IF-ELSE. De esta manera conseguiremos implementar de forma sencilla el árbol y que este nos dé un resultado en función del nodo hoja al que hemos llegado.

Además de la predicción necesitamos un valor de riesgo que nos determine cómo de peligroso es apostar a un determinado resultado. La única forma aproximada que tenemos de establecer un valor para el riesgo es tomar los datos del entrenamiento del árbol y ver cuántas predicciones han sido erróneas para cada uno de los nodos hoja.

Si un determinado nodo hoja del árbol ha clasificado diez partidos y cinco han sido fallados, se estimará que el riesgo de tomar el resultado de ese nodo será de un 50%. El problema vendrá en nodos hoja que clasifiquen un número de partidos muy bajo. En muchos casos del entrenamiento, un nodo hoja ha clasificado sólo tres partidos y los tres partidos han sido correctos. Al haber clasificado un número de partidos muy bajo no podemos tomar para ese nodo un riesgo del 0%, ya que posiblemente es irreal. Por ello, para este tipo de nodos se establecerá un riesgo estándar del 25%.

Al ejecutar el código que emula el funcionamiento del árbol J48, el algoritmo irá recorriendo las ramas del árbol hasta llegar a un nodo hoja, que le indicará el resultado y el riesgo que debe aparecer en la hoja Excel correspondiente a la competición que se está estudiando.

3.4.5 Implementación de una Red Bayesiana

La implementación de este algoritmo es algo más complicada que en el caso del árbol J48. Para adaptar el concepto de Red Bayesiana a nuestro problema tendremos que solventar varios problemas y adoptar diversas decisiones de implementación.

Los clasificadores bayesianos se basan en el Teorema de Bayes [38], el cual es capaz de estimar la probabilidad de un suceso aleatorio a través de la fórmula mostrada en la Figura 22.

Sea $\{A_1, A_2, ..., A_i, ..., A_n\}$ un conjunto de sucesos mutuamente excluyentes y exhaustivos, y tales que la probabilidad de cada uno de ellos es distinta de cero (0). Sea B un suceso cualquiera del que se conocen las probabilidades condicionales $P(B|A_i)$. Entonces, la probabilidad $P(A_i|B)$ viene dada por la expresión:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)}$$

donde:

- $P(A_i)$ son las probabilidades a priori.
- $P(B|A_i)$ es la probabilidad de B en la hipótesis A_i.
- $P(A_i|B)$ son las probabilidades a posteriori.

Thomas Bayes (1763)

Figura 22. Teorema de Bayes [38]

El inconveniente que encontramos al intentar aplicar la fórmula del teorema a nuestro problema de predicción, es que las hipótesis que manejamos no son mutuamente excluyentes, es decir, que si por ejemplo en un partido de Competición Europea el equipo local es español, esto no implica que el equipo visitante no lo sea. Este ejemplo se extiende para todos los atributos de los conjuntos que hemos tomado para el entrenamiento.

Para solventar este problema se ha decidido utilizar una variante del clasificador bayesiano que es el clasificador **Naïve Bayes**. Este clasificador es muy similar al clasificador bayesiano, pero en este caso, para simplificar el problema se supone que las variables para la clasificación son independientes en cada una de las clases. Para mejorar los resultados de este algoritmo es preferible que el conjunto de datos de entrada sea discreto, por lo que tomaremos los intervalos que definió la herramienta WEKA al analizar los conjuntos y entrenar la Red Bayesiana.

Para explicar los cálculos que van a ser necesarios para implementar el algoritmo de clasificación bayesiano, se va a poner un ejemplo simplificado en el que se detallarán qué datos necesitaremos y cómo se manipularán para obtener la probabilidad de cada uno de los resultados:

El ejemplo que se va a explicar a continuación trata de predecir mediante el algoritmo de Naïve Bayes si un partido de tenis se disputará o por el contrario se suspenderá debido a las condiciones climatológicas. Los datos de entrada del algoritmo serán diversos indicadores climatológicos, que aunque no lo son han sido tomados como mutuamente excluyentes para su uso en este algoritmo. La Tabla 16 muestra los datos de que disponemos.

P(Cielo | Hay Partido)

Cielo	Si	No
Sol	2/9	3/5
Nubes	4/9	0/5
Lluvia	3/9	2/5

P(Temperatura | Hay Partido)

Temperatura	Si	No
Caliente	2/9	2/5
Templado	4/9	2/5
Frío	3/9	1/5

P(Humedad | Hay Partido)

Humedad	Si	No
Alta	3/9	4/5
Normal	6/9	1/5

P(Viento | Hay Partido)

Viento	Si	No
Si	3/9	3/5
No	6/9	2/5

P(Hay partido)

Si	No
9/14	5/14

Tabla 16. Datos para ejemplo de Redes Bayesianas

Una vez tenemos los datos con los que se entrena la red, imaginemos que queremos saber si un partido se disputará o no. Los datos que tenemos de dicho partido serían los siguientes:

- Cielo = **Sol**
- Temperatura = **Frío**
- Humedad = **Alta**
- Viento = **Si**

La fórmula que se va a aplicar para conocer las probabilidades tiene que ser aplicada tantas veces como valores posibles tenga la clase. Como la clase de este problema tiene dos resultados posibles (Sí y No), tendremos que aplicar la fórmula para esos dos casos y así conocer la probabilidad de que ocurra cada uno de los sucesos. Los cálculos que habría que realizar son los siguientes:

P(Sí | Sol, Frío, Alta, Sí) = 2/9 * 3/9 * 3/9 * 3/9 * 9/14 = **0,0053**

P(No | Sol, Frío, Alta, Sí) = 2/9 * 3/9 * 3/9 * 3/9 * 5/14 = **0,0206**

Como se puede observar, el valor de probabilidad que se obtiene para el caso de que no se juega el partido es mayor que el valor de que sí se juegue en partido. Para mostrar los valores en formato de porcentaje de manera que sean más fáciles de entender normalizaremos los valores para obtener un valor de probabilidad entre 0 y 100, en el que la suma de las dos probabilidades sea del 100%.

P(Sí | Sol, Frío, Alta, Sí) = 0.0053 / (0.0053 + 0.206) = **20,5%**

P(No | Sol, Frío, Alta, Sí) = 0.0206 / (0.0053 + 0.206) = **79,5%**

Después de normalizar los valores podemos ver que hay un 79,5% de probabilidades de que el partido no se dispute, de acuerdo a los datos utilizados en nuestro conjunto de entrenamiento.

Una vez explicado este sencillo ejemplo, sólo hay que emular los cálculos ilustrados en una macro para que el proceso de predicción de resultados sea totalmente automático.

Los valores de las probabilidades de cada uno de los sucesos serán extraídos directamente de la herramienta WEKA (Figura 23), ya que al generar la red bayesiana automáticamente se muestran los coeficientes de cada uno de los atributos.

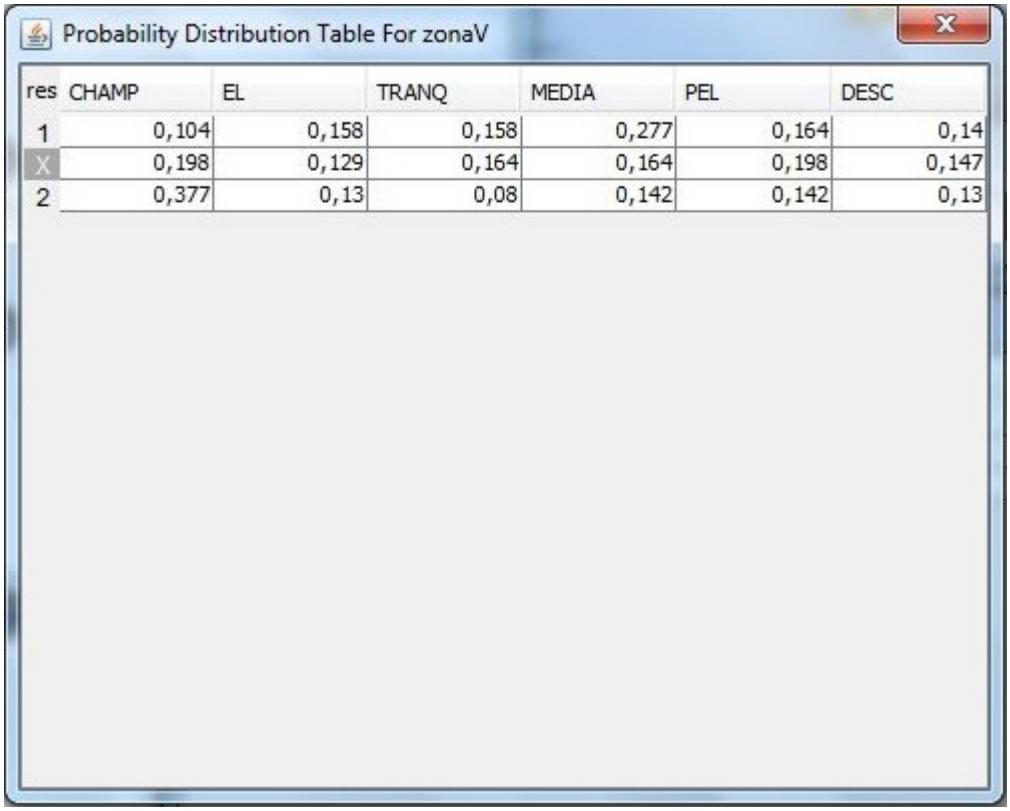

Figura 23. Distribución de probabilidades generada por WEKA para Red Bayesina

Cada una de las distribuciones de probabilidad que genera WEKA en los algoritmos de clasificación bayesianos, serán tomadas y almacenadas según la estructura explicada con anterioridad en la pestaña *Tabla Bayes*. Los coeficientes que contienen dichas tablas serán los que se utilicen para el cálculo de las probabilidades de cada uno de los resultados posibles de un partido.

Los pasos que seguirá la macro para calcular las probabilidades de cada uno de los resultados en las diferentes competiciones a analizar serán los siguientes:

- Lectura desde la pestaña de competición de los datos definidos para la predicción con Redes Bayesianas.

- Discretización de los datos numéricos según los grupos formados por WEKA en las diferentes tablas de distribución de probabilidades.

- Una vez se han discretizado los datos de entrada, se empieza a calcular para cada resultado la probabilidad de que este resultado ocurra al finalizar el partido. Para ello la macro irá buscando los coeficientes de probabilidad que correspondan a cada atributo y los irá multiplicando.

- Tras calcular todos las probabilidades se normaliza el resultado para que los valores estén entre 0% y 100%. Además con esta acción se conseguirá que la suma las probabilidades de todos los resultados posibles sea el 100%.

- La macro tomará como predicción final el resultado que tenga mayor probabilidad.

- El riesgo total de tomar la predicción como resultado final saldrá de restar al 100% la probabilidad obtenida para el resultado de la predicción.

El resultado final tras ejecutar la macro que simula el comportamiento de la red bayesiana será el valor de la predicción de acuerdo al resultado que más probabilidades tiene de ocurrir y el valor del riesgo de tomar esa predicción como la que va a ocurrir. Estos dos valores serán colocados en sus lugares correspondientes de la zona de predicción en la competición que estaba siendo estudiada.

3.5 Conclusiones del capítulo

Como cierre final del capítulo del Sistema de Predicción de Resultados, se van a resaltar las conclusiones más importantes que se han podido sacar hasta este punto del proyecto.

En primer lugar hablaremos sobre el proceso inicial de minería de datos. Este proceso ha sido muy costoso en tiempo, ya que alguna de sus fases (sobre todo la recogida de datos) no ha podido ser automatizada debido a la complejidad del origen de los datos. La dificultad añadida con la que se cuenta a la hora de recoger los datos, es que cada semana sólo pueden recogerse unos pocos registros de partidos de cada competición, lo que eleva el tiempo necesario para tener un conjunto de datos que pueda servir para ser entrenado con clasificadores. La opción de recoger datos de temporadas pasadas se decidió descartar, ya que algunos atributos como las rachas de resultados, si el equipo jugó partidos entre semana o la previsión meteorológica, son difíciles de encontrar si no es en la misma semana de partido.

Una vez se han recogido todos los datos y los conjuntos de entrenamiento han sido formados se ha procedido al análisis de los datos mediante la herramienta WEKA. Mediante esta herramienta se han probado distintos tipos de clasificadores para ver cuál de ellos se ajusta mejor al problema de clasificación que se está planteando. Finalmente tras analizar los resultados obtenidos se ha decidido que por sus buenos resultados y su sencillez a la hora de la implementación se van a utilizar los clasificadores J48 (clasificador tipo árbol) y una Red Bayesiana. Por las características del problema, donde las variables no son mutuamente excluyentes entre sí, se implementará como Red Bayesiana el algoritmo de Naïve Bayes, el cual para obtener las probabilidades de un suceso asume que todas las variables del problema son independientes entre sí.

Finalmente, para implementar los algoritmos seleccionados se ha decidido utilizar como soporte una hoja de cálculo de Microsoft Excel. Los motivos de esta elección se

han basado en la sencillez de la interfaz y los conocimientos previos que tienen la mayoría de usuarios de esta herramienta. Además, el módulo integrado que posibilita la definición e implementación de macros permite poder codificar los algoritmos de clasificación para que los resultados de las predicciones aparezcan en las propias hojas.

Al utilizar varios algoritmos de clasificación como fuente de la predicción de cada uno de los partidos hay que definir una técnica que nos ofrezca la predicción del resultado final de cada partido además de estimar un riesgo de tomar dicha predicción como buena. La técnica que se ha decidido utilizar para estimar estos dos parámetros se basa en la técnica de Bagging (Bootstrap Aggregating), pero introduciendo algunos cambios para ajustar la técnica a las necesidades del problema.

La técnica de Bagging consiste en dividir el conjunto original de datos en varios subconjuntos, y utilizar un clasificador distinto para entrenar cada uno de los subconjuntos. El resultado final de la clasificación se obtendrá al ver cuál es el resultado más probable teniendo en cuenta los resultados parciales que ha obtenido cada uno de los clasificadores.

Para el caso de este problema los cambios que se han introducido a esta técnica son principalmente el uso de un conjunto global (con datos de competiciones de las mismas características que la que se está estudiando) y el de un conjunto específico (subconjunto del conjunto global que contiene sólo datos de la competición que está siendo estudiada). Cada uno de estos conjuntos será entrenado por los dos clasificadores seleccionados en la fase de entrenamiento con la herramienta WEKA. Para el cálculo del resultado final, al evaluar los resultados parciales de cada uno de los clasificadores, como el principal objetivo que se busca con el sistema es que nos dé resultados con seguridad muy altos, sólo se emitirá una predicción cuando la predicción parcial de todos los clasificadores sea la misma. El riesgo de tomar esa predicción como correcta se calculará ponderando los riesgos parciales que ha obtenido cada uno de los clasificadores, con un 80% de peso para los clasificadores bayesianos y un 20% para los árboles J48.

Tras la finalización de la fase de análisis de los datos y diseño e implementación de la hoja de predicción el siguiente paso es probar la efectividad del sistema para realizar predicciones de las competiciones estudiadas. Para ello, se llevará a cabo otro proceso de recogida de datos, en el que se recopilarán los datos definidos para cada competición para después ejecutar las macros y ver si los resultados que arroja el algoritmo son buenos o no.

Capítulo 4

FASE DE PRUEBAS DEL SISTEMA DE PREDICCIÓN

4.1 Introducción

Después de crear la plantilla de predicción de resultado se debe dar otro paso en la ejecución del proyecto para verificar si el sistema que se ha diseñado en la hoja de Microsoft Excel funciona como se esperaba.

Este proceso de test vendrá precedido de una nueva recogida de datos de todas las competiciones a analizar, ya que las pruebas del sistema conviene hacerlas con datos nuevos y no con los que se ha entrenado, ya que los algoritmos de clasificación pueden llegar a "aprender" los datos con los que entrenan, con lo que los resultados obtenidos no serían muy fiables. En esta recogida de datos se reunirán un grupo de partidos de cada una de las competiciones a estudiar. Para este caso no se recopilarán todos los atributos que se recogieron en la fase de análisis de datos. En esta ocasión sólo se tomarán los atributos que han sido elegidos en la fase de entrenamiento de cada uno de los conjuntos y que maximizan la tasa de aciertos de cada uno de los clasificadores.

Para llevar a cabo este proceso de test del sistema, se diseñarán las pruebas a realizar indicando unos objetivos mínimos a alcanzar y unos objetivos esperados, que serán las cifras que se pretendían alcanzar al comenzar este proyecto.

En cuanto a la definición de objetivos, fijaremos varios objetivos para cada una de las competiciones. Debido a las características de los diferentes modelos de predicción generados, los objetivos a alcanzar serán diferentes según el tipo de predicción que vayamos a realizar. En cualquier caso, para intentar asegurar el mejor resultado posible, se van a realizar estudios en los que se van a cubrir determinados resultados a la hora de hacer predicciones, para así poder verificar si cubriendo dichos resultados (que normalmente serán empates) obtenemos unos mejores resultados que si no realizamos dicha acción de cobertura.

De cualquier forma, independientemente al tipo de clasificadores utilizados, tendremos definidos para cada competición un objetivo mínimo y un objetivo esperado. El objetivo mínimo se establece en un porcentaje de aciertos en los que simplemente por azar lograríamos alcanzar una tasa de aciertos determinada. Este objetivo mínimo se situará en el 50% para el conjunto de partidos de la NBA y en un 33,3% para partidos de las diferentes competiciones de fútbol.

Como se ha hablado anteriormente, ante el estudio que se va a realizar en el que se cubrirán los empates de los partidos de las competiciones de fútbol, tendremos que definir otro objetivo mínimo diferentes para este tipo de competiciones. En esta tarea la predicción que se realizará será la unión del resultado de la hoja de predicción junto con el resultado de empate. Es decir, que si la hoja de predicción predice que el equipo visitante ganará el partido, nuestra predicción real será que el equipo visitante ganará el partido o que el partido acabará en empate.

Al cubrir dos de los tres resultados posibles, no se puede dejar el objetivo mínimo de las competiciones de fútbol en un 33,3%, por lo que para este estudio de cobertura de empates el objetivo mínimo será establecido en un 66,6%.

Finalmente, el objetivo deseado de cada una de las competiciones será establecido dependiendo de las expectativas que han generado los diferentes clasificadores en la fase de entrenamiento. Basándonos en las tasas de acierto conseguidas por cada modelo se establecerá dicho objetivo deseado, que representará el porcentaje de aciertos que se desearía obtener en la competición teniendo en cuenta las limitaciones del modelo.

Como ocurrió en el caso anterior, en el caso que estemos realizando el estudio de cobertura de empates, estos objetivos deseados deberán ser modificados, ya que por lógica este objetivo debe ser bastante superior al objetivo mínimo.

A continuación se van a definir tanto las tareas a realizar en el proceso de pruebas como los diferentes datos necesarios para realizar las tareas de análisis de los resultados. Además se definirá el modelo de resultados que se ofrecerá tras analizar los datos, y que incluirá diversas tablas que reflejarán las tasas de aciertos obtenidas, y que irán acompañadas de varios gráficos que mostrarán los datos semanales y acumulados que se han obtenido durante el proceso de pruebas.

4.2 Definición del proceso de pruebas

Antes de comenzar con las pruebas del sistema, se ha de diseñar el proceso, incluyendo tanto las pruebas a realizar como los resultados que se esperan de estas pruebas. Para realizar dichas pruebas se debe poner en marcha un nuevo proceso de recogida de datos que se extenderá durante 12 semanas y en donde se recogerán para cada competición un número determinado de partidos, que contendrán los atributos que fueron especificados para cada competición al entrenar los conjuntos con los diferentes clasificadores. Los atributos que se van a recoger para cada una de las competiciones son los siguientes:

- **Champions League y Europa League:**

  - Cuota de la casa de apuestas para el equipo local.
  - Cuota de la casa de apuestas para el equipo visitante.
  - Coeficiente UEFA de la liga del equipo local.
  - Coeficiente UEFA de la liga del equipo visitante.

  Después de recoger estos datos se derivarán otros como el TOP en el que se encuentran los países de los equipos local y visitante en el Ranking de la UEFA.

- **Partidos Internacionales:**

  - Cuota de la casa de apuestas para el equipo local.
  - Cuota de la casa de apuestas para el equipo visitante.

- **Liga BBVA:**

  - Cuota de la casa de apuestas para el equipo local.
  - Cuota de la casa de apuestas para el equipo visitante.
  - Cuota de la casa de apuestas para el empate.
  - Posición del equipo local en la liga.
  - Posición del equipo visitante en la liga.
  - Goles anotados de media por el equipo visitante.

  Además de los atributos anteriormente mencionados, también se derivarán a partir de los atributos recogidos las zonas de la liga en las que se encuentra el equipo local y el visitante.

- **Liga Adelante, Premier League, Ligue 1, Serie A y Otras Ligas:**

  - Cuota de la casa de apuestas para el equipo local.
  - Cuota de la casa de apuestas para el equipo visitante.
  - Cuota de la casa de apuestas para el empate.
  - Posición del equipo local en la liga.
  - Posición del equipo visitante en la liga.

Al igual que ocurría con los datos pertenecientes a la Liga BBVA, también se derivarán a partir de los atributos recogidos las zonas de la liga en las que se encuentra el equipo local y el visitante.

❖ **NBA:**

- Cuota de la casa de apuestas para el equipo local.
- Cuota de la casa de apuestas para el equipo visitante.
- Porcentaje de victorias en la temporada del equipo local.
- Porcentaje de victorias en la temporada del equipo visitante.

Pare este caso también será necesario derivar el atributo del porcentaje de victorias discretizado. Para esta discretización se seguirán las pautas que se tuvieron en cuenta a la hora del análisis de los datos.

Una vez definidos los datos que son necesarios recoger para realizar el estudio de aciertos en cada una de las competiciones, se va a pasar a definir los objetivos de cada una de estas pruebas. Los objetivos a alcanzar en cada una de las competiciones son los que se muestran en la Tabla 17.

	Objetivo Mínimo	Objetivo Esperado
Champions League	33.3%	60%
Europa League	33.3%	55%
Partidos Internacionales	33.3%	60%
Liga BBVA	33.3%	60%
Liga Adelante	33.3%	50%
Ligue 1	33.3%	55%
Premier League	33.3%	55%
Serie A	33.3%	55%
Otras Ligas	33.3%	50%
NBA	50%	70%

Tabla 17. Definición de Objetivos para las Competiciones estudiadas

En la tabla superior aparecen los objetivos fijados para las comprobaciones de nuestro estudio de la tasa de aciertos de la hoja de predicciones. Como se puede observar, el objetivo mínimo corresponde a la tasa de aciertos que obtendríamos al realizar predicciones al azar. Para el caso de competiciones en las que hay tres resultados posibles para un partido el objetivo mínimo se ha fijado en el 33,3%, mientras que para la competición de baloncesto, al haber sólo dos resultados posibles el objetivo mínimo asciende al 50%.

En cuanto a los objetivos esperados, se han utilizado los resultados obtenidos en el entrenamiento con clasificadores para realizar una estimación de cuál podría ser el objetivo esperado de las pruebas de cada una de las competiciones. El objetivo esperado que se ha fijado es muy similar a la tasa de aciertos que se ha obtenido en el entrenamiento de cada conjunto.

Como se viene avisando durante el apartado, debido a los estudios de cobertura de resultados que se van a realizar para cada una de las competiciones, también es necesario redefinir los dos tipos de objetivos para cada una de las competiciones. La nueva definición para el estudio de cobertura de resultados se refleja en la Tabla 18.

	Objetivo Mínimo	Objetivo Esperado
Champions League	66.6%	85%
Europa League	66.6%	80%
Partidos Internacionales	66.6%	85%
Liga BBVA	66.6%	85%
Liga Adelante	66.6%	75%
Ligue 1	66.6%	80%
Premier League	66.6%	80%
Serie A	66.6%	80%
Otras Ligas	66.6%	75%
NBA	-	-

Tabla 18. Definición de Objetivos para estudio de cobertura de resultados

Como puede observarse en la tabla superior, los objetivos han sido redefinidos para todas las competiciones excepto la NBA. Esto es debido a que como en la competición baloncestística sólo hay dos resultados posibles, no se puede hacer ningún estudio de cobertura de resultados, ya que de cubrir otro resultado diferente al de la predicción se estaría cubriendo todo el espacio de resultados.

Para el resto de competiciones, todas ellas de fútbol, se ha establecido el nuevo objetivo mínimo al 66,6%, ya que al cubrir dos de los tres resultados posibles acertaríamos el pronóstico por azar un 66,6% de las veces.

En cuanto a los objetivos deseados, en esta ocasión no tenemos ninguna referencia de estudios anteriores, por lo que es algo más difícil estimar a qué nivel de tasa de aciertos se puede llegar realizando tareas de cobertura de resultados en las predicciones. Como sí sabemos que en la gran mayoría de competiciones el porcentaje de empates que se produce es de entre el 20% y el 25%, utilizaremos este dato para hacer las predicciones. Finalmente, el objetivo deseado para las competiciones de fútbol en el estudio de cobertura de resultados se ha fijado como el objetivo esperado al no cubrir resultados más un 25%. Ese 25% extra que se añade es el que aporta la cobertura de empates a nuestro estudio.

La definición de estos objetivos nos servirá para medir el éxito de nuestro sistema de predicción, ya que obtener resultados por encima del objetivo esperado sería un gran éxito, mientras que acabar con resultados cercanos al objetivo mínimo sería un fracaso absoluto.

Tras concluir la fase de definición de los objetivos de cada una de las competiciones, se va a pasar a detallar las fases en las que se va a dividir el proceso de prueba del sistema de predicción de resultados.

En primer lugar, tras haber recogido los datos durante el periodo establecido de 12 semanas se realizará un análisis de la calidad de los datos recogidos. Este análisis tiene como objetivo revisar todos los datos recogidos y verificar que no hay valores erróneos dentro del conjunto de test (valores fuera de rango o valores pico que no deberían haberse registrado en una competición en concreto). Tras analizar y verificar que los datos son correctos, el siguiente paso será ejecutar la hoja de predicción para obtener las predicciones de todos los partidos recogidos durante las últimas 12 semanas.

Se debe recordar que la hoja de predicción no siempre efectúa una predicción final sobre el resultado de un partido, ya que en las ocasiones en las que los distintos clasificadores ofrecen distintos pronósticos el sistema no emitirá una predicción al considerar ese partido como un partido de riesgo para emitir una predicción correcta. Por ello, consideraremos el número de partidos en los que no se emite predicción, ya que éstos no deberían entrar en los cálculos de porcentaje de aciertos de cada una de las competiciones.

Tras ejecutar la macro que calcula las predicciones de los partidos habrá que elaborar las tablas que desglosen los datos semana a semana y a partir de las cuales se generarán las gráficas que nos permitirán ver la evolución de los datos a lo largo de estas 12 semanas. El formato de tablas y gráficas que se va a utilizar para recoger los resultados de las pruebas es el que muestra la Tabla 19.

	Partidos		Predicciones Simples		Predicciones Doble Oportunidad	
Semana	Con Predicción	Sin Predicción	Nº Aciertos	%	Nº Aciertos	%
1				0.00%		0.00%
2				0.00%		0.00%
3				0.00%		0.00%
4				0.00%		0.00%
5				0.00%		0.00%
6				0.00%		0.00%
7				0.00%		0.00%
8				0.00%		0.00%
9				0.00%		0.00%
10				0.00%		0.00%
11				0.00%		0.00%
12				0.00%		0.00%
TOTAL	**0**	**0**	**0**	**0.00%**	**0**	**0.00%**

Tabla 19. Ejemplo de tabla de resultados de las pruebas de predicción

En la tabla anterior se puede ver los resultados que vamos a tener en cuenta en las pruebas. En primer lugar tendremos que diferenciar entre los partidos para los que se han hecho predicciones y los que no. Esto permitirá que a la hora de calcular la tasa de aciertos no se tengan en cuenta los partidos para los que no se hace predicción, ya que esos partidos no han sido acertados ni fallados, simplemente están fuera del perímetro de nuestro estudio.

Una vez tenemos localizados los partidos para los que se ha hecho predicción, tendremos que comprobar si la predicción realizada ha sido correcta. Para ello habrá que ir comprobando uno a uno los resultados de los partidos con predicción.

Para registrar los resultados en la tabla se han tenido en cuenta varias consideraciones:

- En la columna **Nº Aciertos** de la sección **Predicciones Simples**, sólo se considerarán aciertos aquellos partidos cuyo resultado coincida exactamente con la predicción que ha realizado el sistema de predicción.

- En la columna **Nº Aciertos** de la sección **Predicciones Doble Oportunidad**, sólo se considerarán aciertos aquellos partidos cuyo resultado coincida exactamente con la predicción que ha realizado el sistema de predicción o aquellos partidos cuyo resultado final ha sido empate. En esta columna registraremos los partidos acertados en la modalidad en la que realizamos cobertura de empates.

- Los **porcentajes de acierto** de cada una de las modalidades serán calculados dividiendo el número de aciertos de cada una de las semanas entre el número de partidos con predicción de esa misma semana.

Para cada competición tendremos una tabla como la anterior que recogerá los resultados obtenidos a lo largo de las 12 semanas. Además de dicha tabla, el estudio contendrá dos gráficas que nos servirán para ver si los resultados obtenidos son satisfactorios o no. En dichas gráficas se comparará la tasa de acierto tanto semanal como acumulado con los objetivos mínimos y recomendados definidos con anterioridad. Esta comparativa nos permitirá ver de un vistazo si se están cumpliendo los objetivos previstos.

Después de definir el procedimiento que se va a llevar a cabo para obtener los datos de las pruebas y generar las estadísticas que nos permitan analizar los resultados, comenzaremos analizando una a una las competiciones de las que se han recogido datos durante estas 12 semanas.

4.3 Fase de Pruebas del Sistema de Predicción

4.3.1 Fase de Pruebas Competición de Champions League

Antes de mostrar los resultados de esta competición hay que mencionar que al ser una competición en la que participan menos equipos se han tenido que recolocar algunos partidos en semanas diferentes a las que tuvieron logar éstos, ya que esta competición no se disputa semanalmente. Como necesitamos tener resultados en 12 semanas consecutivas para después poder aplicar las distintas estrategias de apuesta se ha decidido poner en cada "semana ficticia" un total de seis partidos excepto en la semana 12 donde se han colocado 9. El movimiento de partidos de una semana a otra no altera el porcentaje final de aciertos, ya que esta cifra se basa en los aciertos totales de las 12 semanas.

Los resultados obtenidos para esta competición son los que se muestran en la Tabla 20.

	Partidos		Predicciones Simples		Predicciones Doble Oportunidad	
Semana	Con Predicción	Sin Predicción	Nº Aciertos	%	Nº Aciertos	%
1	2	4	2	100,00%	2	100,00%
2	2	4	1	50,00%	1	50,00%
3	3	3	2	66,67%	3	100,00%
4	1	5	1	100,00%	1	100,00%
5	3	3	2	66,67%	3	100,00%
6	3	3	2	66,67%	3	100,00%
7	1	5	1	100,00%	1	100,00%
8	4	2	3	75,00%	4	100,00%
9	4	2	3	75,00%	4	100,00%
10	1	5	1	100,00%	1	100,00%
11	2	4	1	50,00%	2	100,00%
12	6	3	4	66,67%	5	83,33%
TOTAL	**32**	**43**	**23**	**71,88%**	**30**	**93,75%**

Tabla 20. Resultado pruebas Competición de Champions League

Como se puede observar en la tabla superior se han conseguido unos porcentajes de acierto muy buenos que ponen de manifiesto que los clasificadores escogidos para la predicción de esta competición eran los adecuados.

Cabe destacar un altísimo 93,75% de aciertos en predicciones en las que se cubre el resultado de empate, donde sólo dos de las 32 predicciones han sido erradas.

Al observar el número de predicciones que realiza la hoja de predicción podemos ver que un gran número de partidos se quedan sin predicción debido a que no existe consenso entre los cuatro clasificadores utilizados. Esto muestra la dificultad existente para predecir resultados en esta competición, ya que en más de la mitad de los partidos varios clasificadores no han coincidido en la predicción.

A partir de los datos de la tasa de aciertos para predicciones simples y predicciones con doble oportunidad podemos sacar la cuota mínima que necesitaríamos de la casa de apuestas para obtener beneficios apostando a todos los partidos para los que nuestra hoja de predicción arroja un resultado fruto del consenso entre los cuatro clasificadores. Para este caso las cuotas medias a partir de las cuales se obtendrían beneficios serían las siguientes:

- **Predicción Simple:**

$0{,}7188(x-1) - 0{,}2812 > 0 \rightarrow x > 1/0{,}7188 \rightarrow \mathbf{x > 1{,}39}$

- **Doble Oportunidad:**

$0{,}9375(x-1) - 0{,}0635 > 0 \rightarrow x > 1/0{,}9375 \rightarrow \mathbf{x > 1{,}06}$

Analizando las dos cuotas medias a partir de las cuales se obtendrían beneficios podemos ver que ambas son asequibles, sobre todo la cuota necesaria en apuestas con doble oportunidad, ya que tan solo requiere una media de 1,06.

A continuación se van a mostrar las gráficas que comparan los resultados semanales acumulados con los objetivos mínimos y esperados para los dos tipos de predicciones (Figura 24 y Figura 25).

Como puede observarse en la Figura 24, el porcentaje de aciertos acumulado se mantiene semana a semana muy por encima de los dos objetivos marcados para esta competición, lo que demuestra la calidad del algoritmo diseñado para esta competición.

Además, la gráfica muestra que exceptuando la primera semana donde se obtuvieron resultados muy buenos el acumulado del resto de semanas se mantiene muy estable por encima del 70% lo que hace pensar que en un futuro ese dato se mantendrá estable.

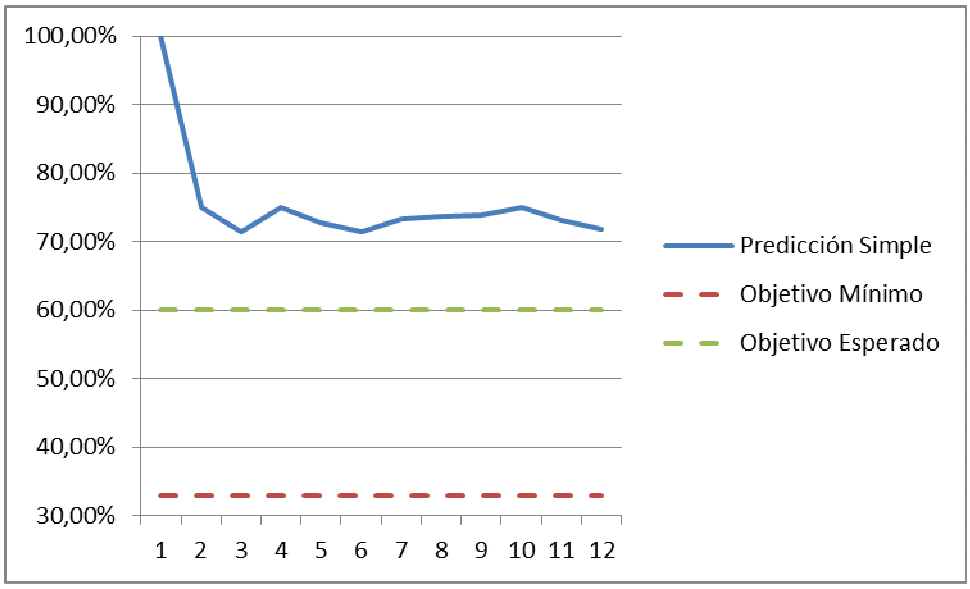

Figura 24. Resultados vs. Objetivos Champions Predicción Simple

En el caso de las apuestas con Doble Oportunidad, la grafica comparativa entre los resultados acumulados cada semana y los objetivos establecidos antes de comenzar las pruebas (Figura 25) muestra resultados muy buenos.

Como ya se había comentado anteriormente el nivel al que llega la tasa de aciertos tras doce semanas es realmente bueno, ya que se sitúa cerca del 95% de acierto. Este dato es muy significativo, ya que permitirá realizar combinaciones de partidos de esta competición para obtener mayores beneficios en la siguiente fase de análisis que corresponde con la aplicación de estrategias de apuestas.

A pesar de estar en la semana 2 por debajo del objetivo esperado, se puede ver que a partir de la semana 8 la tasa de aciertos se estabiliza alrededor del 95% que es el valor final que se ha obtenido en este estudio.

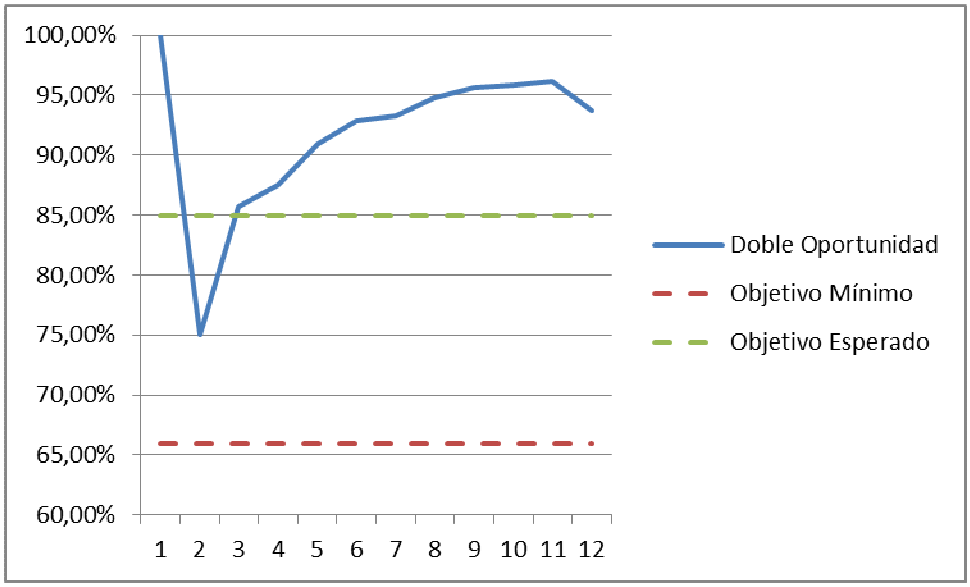

Figura 25. Resultados vs. Objetivos Champions Doble Oportunidad

4.3.2 Fase de Pruebas Competición de Europa League

Al igual que ocurría en la competición de Champions League, esta competición es una competición que no se disputa todas las semanas, como suele pasar con la competición de Liga. Por ello y para su posterior análisis en las estrategias de apuestas que serán explicadas en el próximo capítulo, se ha decidido reubicar los partidos en semanas diferentes a los que en ocasiones han tenido lugar, para así disponer de un total de nueve partidos para cada una de las semanas definidas. Sólo recordar que la nueva ubicación de los partidos no influye en el resultado final de las tasas de acierto, ya que dichas tasas se basan en el número de aciertos totales, por lo que da igual en qué semana tengan lugar dichos aciertos.

Los resultados obtenidos en el estudio de esta competición pueden ser consultados en la Tabla 21.

	Partidos		Predicciones Simples		Predicciones Doble Oportunidad	
Semana	Con Predicción	Sin Predicción	Nº Aciertos	%	Nº Aciertos	%
1	3	6	1	33,33%	3	100,00%
2	5	4	4	80,00%	4	80,00%
3	6	3	3	50,00%	4	66,67%
4	5	4	3	60,00%	5	100,00%
5	4	5	2	50,00%	2	50,00%
6	5	4	2	40,00%	4	80,00%
7	5	4	3	60,00%	5	100,00%
8	5	4	4	80,00%	5	100,00%
9	4	5	4	100,00%	4	100,00%
10	7	2	4	57,14%	6	85,71%
11	6	3	3	50,00%	5	83,33%
12	8	1	6	75,00%	8	100,00%
TOTAL	**63**	**45**	**39**	**61,90%**	**55**	**87,30%**

Tabla 21. Resultado pruebas Competición de Europa League

Como puede observarse en los datos de la Tabla 21, se han conseguido unos porcentajes muy buenos aunque no llegan al nivel de los obtenidos en la competición de la Champions League.

En este caso, la hoja de predicciones no ha desechado tantos partidos como en la anterior competición, lo que significa que los cuatro clasificadores utilizados han llegado al mismo resultado en un porcentaje mayor de ocasiones.

Al igual que en la competición analizada con anterioridad se van a calcular las cuotas medias necesarias para obtener beneficios con cada uno de los dos tipos de predicciones realizadas:

- **Predicción Simple:**

$0{,}619(x-1) - 0{,}381 > 0 \rightarrow x > 1/0{,}619 \rightarrow \mathbf{x > 1{,}61}$

- **Doble Oportunidad:**

$0{,}873(x-1) - 0{,}127 > 0 \rightarrow x > 1/0{,}873 \rightarrow \mathbf{x > 1{,}14}$

Las cuotas medias que se obtienen tras el estudio de las tasas de acierto son algo peores que las obtenidas en la competición de Champions League debido a las menores tasas de acierto en la competición que se está estudiando. No obstante, la tasa obtenida en las predicciones basadas en Doble Oportunidad siguen teniendo una cuota lo suficientemente interesante como para que se puedan obtener beneficios en la próxima fase en la que se pondrán en marcha diferentes estrategias de apuestas.

La Figura 26 muestra los datos acumulados de la tasa de aciertos en la modalidad de predicción simple. Como se puede observar, en las primeras semanas los resultados se mueven por encima y por debajo del objetivo esperado, pero según van pasando las semanas el dato se estabiliza y se sitúa por encima del 60%.

El dato obtenido es bueno ya que se sitúa por encima del objetivo esperado, pero puede que no sea suficiente para la generación de beneficios, hecho que se valorará en el siguiente capítulo de aplicación de las estrategias de apuestas.

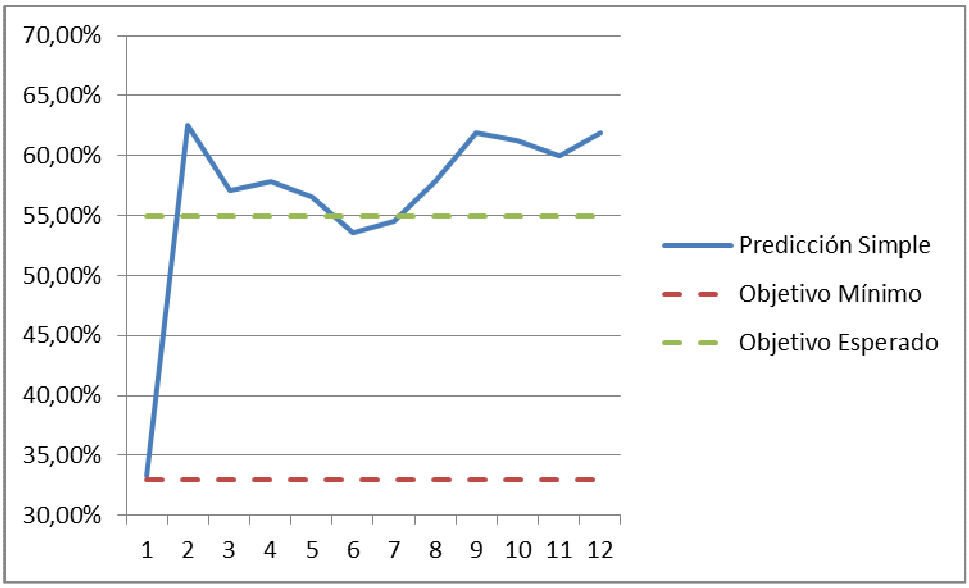

Figura 26. Resultados vs. Objetivos Europa League Predicción Simple

En el caso de apuestas de Doble Oportunidad, la Figura 27 muestra la comparativa de la tasa de aciertos con el método de Doble Oportunidad contra los objetivos mínimo y esperado. Una vez más la tasa de aciertos acumulada se sitúa por encima de ambos objetivos.

Como ha ocurrido en varios de los análisis que ya hemos realizado, en los primeros momentos la tasa oscila hasta que pasadas varias semanas se estabiliza y se sitúa en los niveles alcanzados a finales de la duodécima semana.

Los resultados, aunque no han llegado al nivel de los obtenidos en la competición de la Champions League, son muy buenos. Estos resultados permitirán que puedan obtenerse beneficios en las apuestas a partir de apuestas con cuotas bajas.

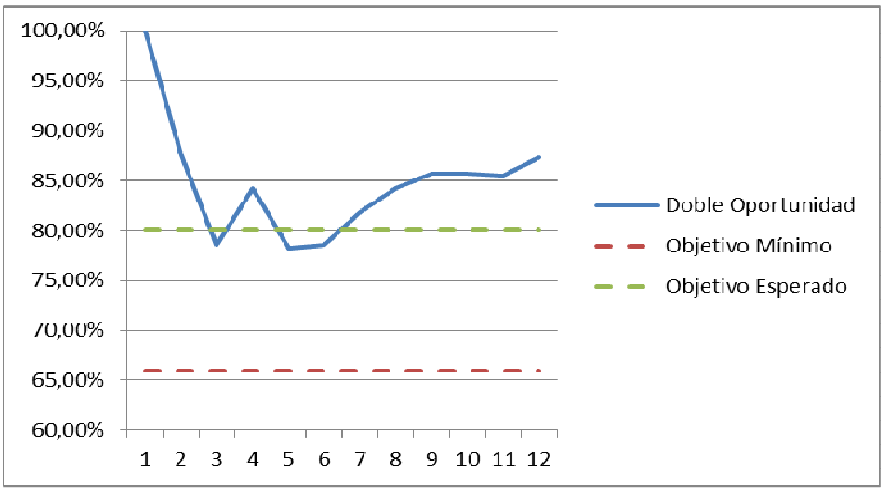

Figura 27. Resultados vs. Objetivos Europa League Doble Oportunidad

4.3.3 Fase de Pruebas Partidos Internacionales

Al igual que ocurría con los partidos de competiciones europeas, los partidos de selecciones internacionales no se disputan semanalmente, por lo que se ha decidido aplicar el mismo método que el utilizado en las competiciones de Champions League y Europa League. Por ello se van a recoger los datos de todos los partidos posibles para después dividirlos en 12 semanas ficticias. Como ocurría en los casos anteriores, esta distribución de los partidos en semanas ficticias no afectará al resultado final de la tasa de aciertos.

Los resultados obtenidos tras el análisis de los partidos internacionales han sido recogidos en la Tabla 22.

	Partidos		Predicciones Simples		Predicciones Doble Oportunidad	
Semana	Con Predicción	Sin Predicción	Nº Aciertos	%	Nº Aciertos	%
1	7	3	3	42,86%	6	85,71%
2	7	3	4	57,14%	5	71,43%
3	6	4	3	50,00%	4	66,67%
4	9	1	5	55,56%	5	55,56%
5	8	2	8	100,00%	8	100,00%
6	8	2	4	50,00%	6	75,00%
7	8	2	4	50,00%	8	100,00%
8	9	1	5	55,56%	8	88,89%
9	8	2	6	75,00%	6	75,00%
10	7	3	4	57,14%	7	100,00%
11	8	2	7	87,50%	8	100,00%
12	9	1	7	77,78%	9	100,00%
TOTAL	**94**	**26**	**60**	**63,83%**	**80**	**85,11%**

Tabla 22. Resultado pruebas Partidos Internacionales

Los resultados obtenidos en este estudio son muy satisfactorios, ya que se obtienen muy buenas tasas de acierto tanto para el método de Predicciones Simples como para el de Doble Oportunidad.

Estos resultados son aún mejores si nos fijamos en que el 78% de los partidos analizados tuvieron una predicción por parte del sistema, un dato mucho más elevado que en el resto de competiciones. Este hecho se produce por la menor presión que tiene el sistema para esta competición, ya que tan sólo dos clasificadores intervienen en la predicción de los resultados de este conjunto.

Respecto a las cuotas necesarias para generar beneficios, para esta competición tenemos los siguientes resultados:

- **Predicción Simple:**

$0{,}6383(x-1) - 0{,}3617 > 0 \rightarrow x > 1/0{,}6383 \rightarrow \mathbf{x > 1{,}56}$

- **Doble Oportunidad:**

$0{,}8511(x-1) - 0{,}1489 > 0 \rightarrow x > 1/0{,}8511 \rightarrow \mathbf{x > 1{,}17}$

Con estos resultados obtenidos conseguimos reducir un poco la cuota media necesaria para la generación de beneficios en el sistema de Predicción Simple, mientras que para Doble Oportunidad la cuota necesaria aumenta unas centésimas.

Respecto a la consecución de los objetivos marcados para esta competición, a continuación se van a mostrar las gráficas a partir de las cuales se ha analizado la comparativa (Figura 105 y Figura 106).

La comparación de resultados contra los objetivos establecidos para esta competición vuelve a arrojar resultados por encima de los esperados (Figura 28). Como ha ocurrido en el análisis de otras competiciones, no es hasta pasadas unas semanas cuando la tasa de aciertos se estabiliza y se sitúa entre el 60%-65%. No obstante, los resultados se sitúan muy cerca del objetivo esperado y no muy por encima como ha ocurrido en las competiciones analizadas anteriormente.

Estos resultados vienen propiciados por el gran número de partidos para los que el sistema de predicción ha arrojado un resultado para el partido. En anteriores competiciones, al tener más clasificadores participando de la predicción, tenemos unos resultados más conservadores sobre los que apostar.

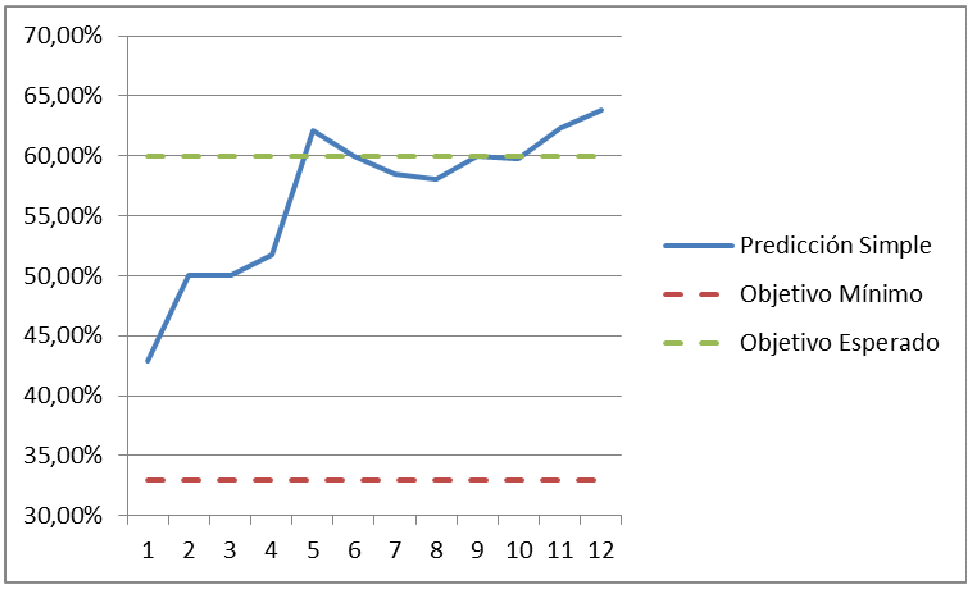

Figura 28. Resultados vs. Objetivos Partidos Internacionales Predicción Simple

Los resultados de la tasa de aciertos con el método de Doble Oportunidad (Figura 29) muestran que los resultados obtenidos por este conjunto son buenos, ya que la tasa supera el objetivo esperado (aunque sólo sea en la última semana).

El problema que se percibe en este estudio es que la tasa de aciertos no se ha llegado a estabilizar a lo largo de las semanas. Como se puede observar en el gráfico la tasa registra importantes subidas desde la cuarta semana. Esto hace que no sepamos con certeza si el nivel del 85% de aciertos que se ha alcanzado es real o por el contrario la tasa seguiría subiendo a lo largo del tiempo o se estabilizara por debajo de la barrera del objetivo deseado.

En cualquier caso, la valoración final sobre los datos obtenidos sobre este conjunto es buena, ya que como se ha indicado anteriormente, los resultados superan los objetivos esperados, aunque está por ver si con este porcentaje de acierto se consigue generar algún beneficio mediante la apuesta a partidos de selecciones internacionales.

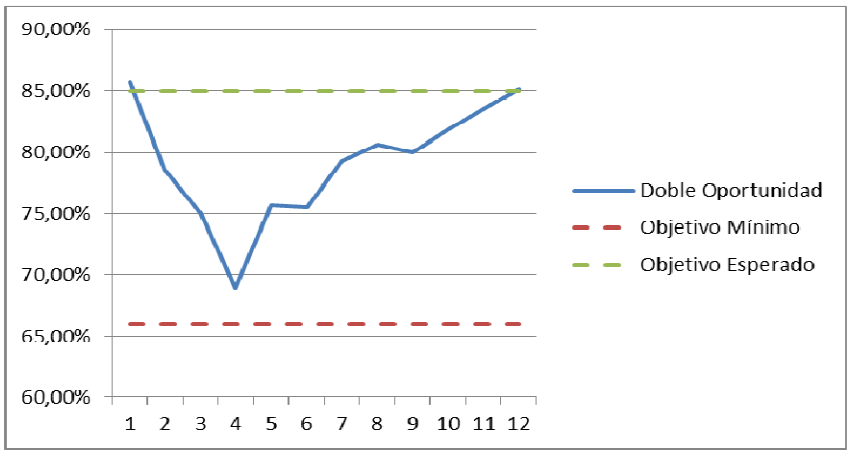

Figura 29. Resultados vs. Objetivos Partidos Internacionales Doble Oportunidad

4.3.4 Fase de Pruebas Liga BBVA

Con el análisis de los partidos de la Liga BBVA se comienza el análisis de las competiciones de liga, competiciones que si se disputan semanalmente (salvo semanas en las que hay compromisos de partidos de selecciones internacionales). Para este caso, las 12 semanas analizadas contendrán partidos disputados en dichas semanas, sin ser necesario realizar ninguna reorganización como ocurría en competiciones anteriores.

Debido a que los calendarios de las competiciones de liga son algo irregulares, puede darse el caso de que un mismo equipo juegue en una misma semana dos partidos, ya que en ocasiones los partidos de liga no sólo se disputan los sábados y domingos.

Los resultados obtenidos tras analizar las predicciones realizadas sobre este conjunto de partidos han sido los registrados en la Tabla 23.

	Partidos		Predicciones Simples		Predicciones Doble Oportunidad	
Semana	Con Predicción	Sin Predicción	Nº Aciertos	%	Nº Aciertos	%
1	8	2	1	12,50%	4	50,00%
2	7	3	7	100,00%	7	100,00%
3	10	0	6	60,00%	10	100,00%
4	5	5	4	80,00%	5	100,00%
5	7	3	4	57,14%	6	85,71%
6	10	0	9	90,00%	10	100,00%
7	8	2	2	25,00%	6	75,00%
8	8	2	4	50,00%	6	75,00%
9	8	2	7	87,50%	8	100,00%
10	9	1	8	88,89%	9	100,00%
11	9	1	3	33,33%	4	44,44%
12	9	1	8	88,89%	9	100,00%
TOTAL	**98**	**22**	**63**	**64,29%**	**84**	**85,71%**

Tabla 23. Resultado pruebas Liga BBVA

Los resultados obtenidos en este estudio son muy satisfactorios, ya que se obtienen muy buenas tasas de acierto tanto para el método de Predicciones Simples como para el de Doble Oportunidad.

Nuevamente se obtiene un porcentaje muy alto de partidos en los que el sistema de predicción ha arrojado un resultado para el partido. Las razones por las que se da este hecho no tienen nada que ver con la menor presión en la predicción, ya que esta competición tiene cuatro clasificadores para realizar la predicción. Para este caso, las razones que podrían explicar este gran porcentaje de consenso entre los cuatro clasificadores puede ser que los conjuntos con los que se entrenaron los clasificadores eran los más grandes de todos, lo que podría implicar que los resultados obtenidos están muy correlacionados con el estilo de juego y resultados previsibles de la competición.

Respecto a las cuotas necesarias para generar beneficios, para esta competición tenemos los siguientes resultados:

- **Predicción Simple:**

$0{,}6429(x-1) - 0{,}3571 > 0 \rightarrow x > 1/0{,}6429 \rightarrow \mathbf{x > 1{,}55}$

- **Doble Oportunidad:**

$0{,}8571(x-1) - 0{,}1429 > 0 \rightarrow x > 1/0{,}8571 \rightarrow \mathbf{x > 1{,}16}$

Con estos resultados se consiguen cuotas bastante aceptables a partir de las cuales se pueden obtener beneficios en las casas de apuestas.

Respecto a la consecución de los objetivos marcados para esta competición, a continuación se van a mostrar las gráficas a partir de las cuales se ha analizado la comparativa (Figura 30 y Figura 31).

Tras analizar los resultados obtenidos con las predicciones realizadas mediante el método de Predicción Simple, se pueden ver unos buenos resultados situados por encima del objetivo esperado para esta competición.

Como muestra la Figura 30, la cuota ha conseguido estabilizarse a la cuarta semana, estableciéndose en niveles cercanos al objetivo esperado. Una vez más los resultados son buenos tras comprobar que la tasa de aciertos se sitúa por encima del objetivo esperado para esta competición.

Los resultados obtenidos podrían garantizar unos buenos beneficios en la fase de explotación del sistema en las casas de apuestas. Estos beneficios residirían en las buenas cuotas que se ofrecen para partidos en los que los equipos no son los primeros clasificados en la clasificación de la competición (normalmente puestos reservados para el Real Madrid y el F.C. Barcelona). Para estos dos equipos las cuotas suelen ser bajas ya que suelen ganar la mayoría de partidos de la competición de liga.

A falta de los datos de generación de beneficios del modelo desarrollado para esta competición, podemos concluir con una valoración positiva de la tasa de aciertos que por el momento ha conseguido el modelo para la competición de la Liga BBVA.

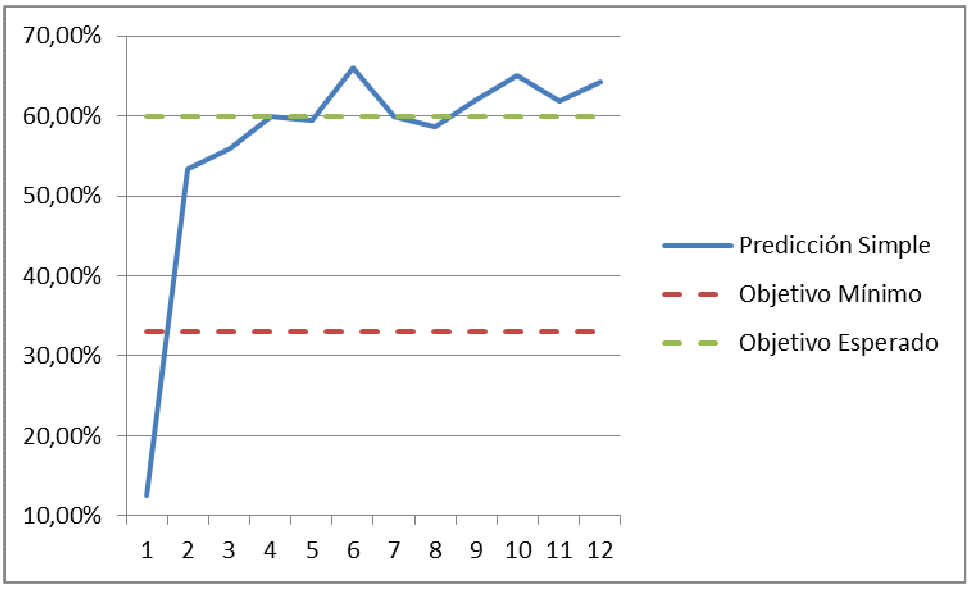

Figura 30. Resultados vs. Objetivos Liga BBVA Predicción Simple

Los datos obtenidos para el estudio de la estrategia de predicción basada en la Doble Oportunidad arrojan parámetros muy similares a los conseguidos con el método de predicción simple (Figura 31).

Al igual que ocurre en la predicción simple, la tasa parece estabilizarse alrededor de la cuarta semana, situándose en niveles entre el 85% - 90%.

Una vez más, los datos se sitúan por encima del objetivo esperado tras la duodécima semana, lo que muestra el éxito momentáneo del método utilizado para esta competición a falta del estudio de beneficios que se realizará en el capítulo posterior.

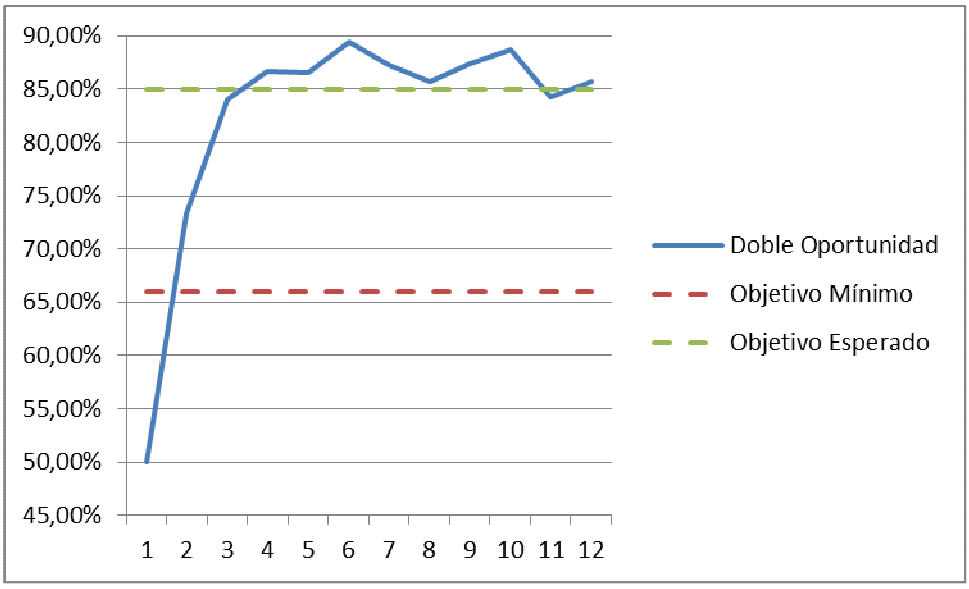

Figura 31. Resultados vs. Objetivos Liga BBVA Doble Oportunidad

4.3.5 Fase de Pruebas Liga Adelante

A continuación se mostrarán los datos obtenidos con las predicciones del sistema sobre la segunda división de fútbol española, también conocida como Liga Adelante.

Si recordamos las consideraciones realizadas sobre esta competición en secciones anteriores se podrá recordar que no se hacían predicciones muy buenas debido al caos que se producía en los resultados de esta competición. La baza que se juega en este estudio es que el sistema de predicción no arroje predicción para muchos de los partidos. Esto podría ocurrir al no haber consenso entre los cuatro clasificadores utilizados por el sistema para realizar las predicciones. Un alto porcentaje de partidos para los que no hay predicción nos aseguraría que los partidos para los que se realiza una predicción tienen una alta seguridad en comparación con el resto de partidos de la competición.

Los resultados obtenidos tras analizar las predicciones de esta competición son los ilustrados en la Tabla 24.

	Partidos		Predicciones Simples		Predicciones Doble Oportunidad	
Semana	Con Predicción	Sin Predicción	Nº Aciertos	%	Nº Aciertos	%
1	4	6	2	50,00%	3	75,00%
2	4	6	2	50,00%	4	100,00%
3	5	5	4	80,00%	4	80,00%
4	5	5	0	0,00%	2	40,00%
5	3	7	1	33,33%	2	66,67%
6	7	3	4	57,14%	5	71,43%
7	6	4	4	66,67%	6	100,00%
8	3	7	3	100,00%	3	100,00%
9	6	4	4	66,67%	5	83,33%
10	4	6	2	50,00%	3	75,00%
11	6	4	5	83,33%	5	83,33%
12	5	5	4	80,00%	4	80,00%
TOTAL	**58**	**62**	**35**	**60,34%**	**46**	**79,31%**

Tabla 24. Resultado pruebas Liga Adelante

Tal y como se había anotado antes, un buen resultado en la tasa de aciertos dependía del porcentaje de partidos para los que el sistema arrojaba una predicción. En este caso, la tasa de partidos con predicción es inferior al 50% lo que ha hecho que la tasa de acierto ascienda por encima del 60% al usar la predicción simple y se sitúe cerca del 80% para la doble oportunidad. Como veremos posteriormente, estas tasas se encuentran por encima de los objetivos esperados para esta competición.

Respecto a las cuotas necesarias para generar beneficios, para esta competición tenemos los siguientes resultados:

- **Predicción Simple:**

$0{,}6034(x-1) - 0{,}3966 > 0 \rightarrow x > 1/0{,}6034 \rightarrow \mathbf{x > 1{,}65}$

- **Doble Oportunidad:**

$0{,}7931(x-1) - 0{,}2069 > 0 \rightarrow x > 1/0{,}7931 \rightarrow \mathbf{x > 1{,}26}$

Al reducir las tasas de aciertos para la competición de la Liga Adelante, se puede ver cómo aumentan las cuotas medias para generar beneficios con esta competición. A pesar de esta subida de cuotas hay que mostrarse muy optimistas, ya que las cuotas que se ofrecen en las casas de apuestas para esta competición son bastante más elevadas que en el resto de competiciones. Estas tasas más altas derivan de la dificultad de acertar los resultados para esta competición debido a la distribución caótica de los resultados que se producen, que en muchas ocasiones carecen de lógica alguna.

En cuanto a la comparativa entre la tasa de aciertos obtenida en esta competición y los objetivos establecidos antes de comenzar los análisis, podemos ver los resultados en las siguientes gráficas (Figura 32 y Figura 33).

Los resultados obtenidos en el estudio de la tasa acumulada de aciertos a lo largo de las 12 semanas en las que se ha centrado, arrojan muy buenos datos, muy por encima del objetivo esperado para esta competición (Figura 32).

El único dato negativo que se encuentra en el estudio es que no se puede determinar con certeza si la tasa de aciertos puede tomarse como la real, ya que a diferencia de otras competiciones, la tasa de aciertos no ha logrado estabilizarse y muestra una clara tendencia al alza. Por tanto no se está en condiciones de saber si la tasa a lo largo del tiempo se quedará en el 60% registrado o si por el contrario seguirá subiendo o tenderá a estabilizarse en una franja más baja de la que se encuentra tras 12 semanas de estudio.

Hará falta esperar a los datos sobre posibles beneficios que arroje el estudio mediante estrategias de apuestas para medir el éxito del método utilizado por el clasificador.

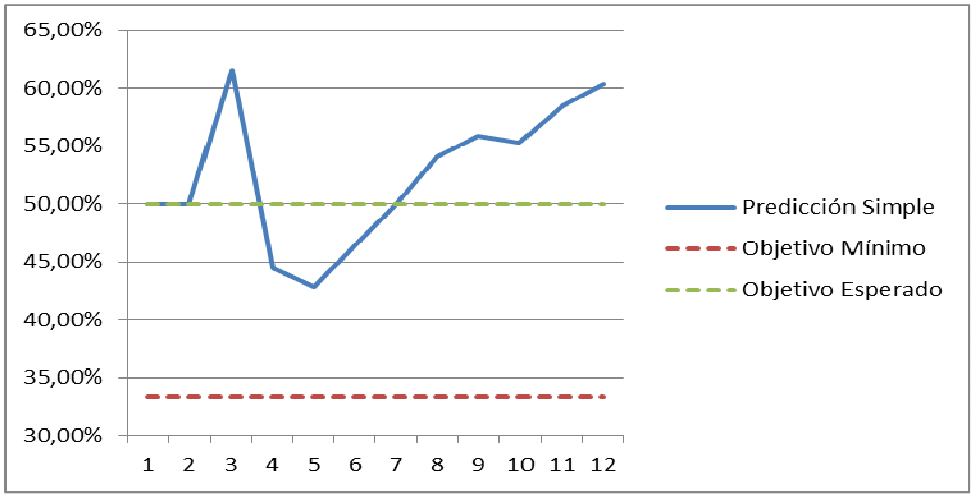

Figura 32. Resultados vs. Objetivos Liga Adelante Predicción Simple

En cuanto a los resultados que pueden verse en la Figura 33 que reflejan la tasa de aciertos del método de Doble Oportunidad, se puede ver una tendencia más clara que en el caso de la Predicción Simple. Para este caso se puede ver como la tasa se estabiliza a partir de la semana 7 en torno a la franja del 75% - 80%. Este hecho hace pensar que el valor obtenido en este estudio se mantendrá estable en esa franja a lo largo del tiempo.

El buen resultado obtenido para la tasa de aciertos en el método de Doble Oportunidad junto con las buenas cuotas que se ofrecen en esta competición en las casas de apuestas, hace pensar que se podrían obtener buenos beneficios a partir de esta competición.

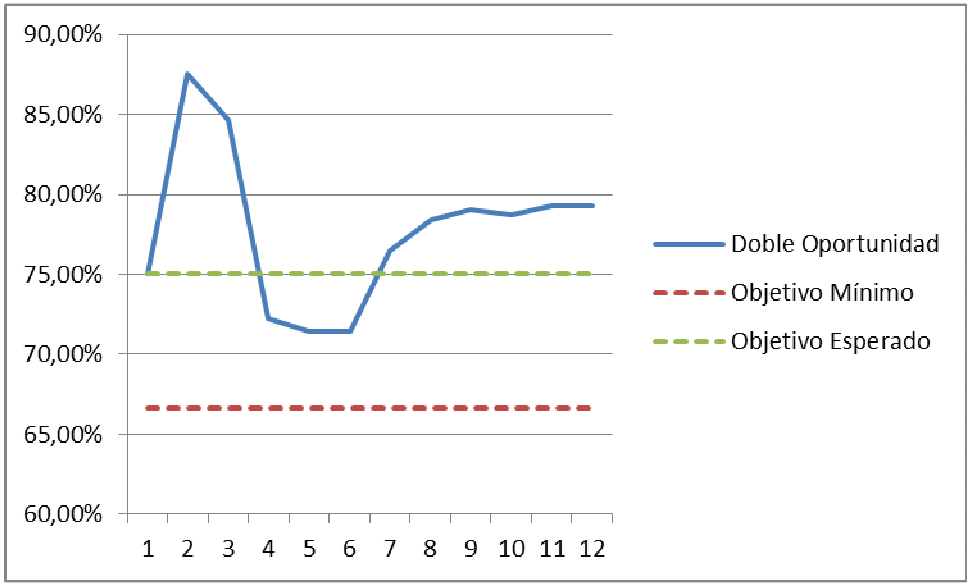

Figura 33. Resultados vs. Objetivos Liga Adelante Doble Oportunidad

4.3.6 Fase de Pruebas Ligue 1

Las características del estudio de la tasa de aciertos para esta competición son exactamente las mismas que las del resto de competiciones de liga incluidas en el sistema de predicción de resultados.

Los datos de la tasa de aciertos obtenidos para la primera división francesa de fútbol también conocida como Ligue 1 aparecen en la Tabla 25.

	Partidos		Predicciones Simples		Predicciones Doble Oportunidad	
Semana	Con Predicción	Sin Predicción	Nº Aciertos	%	Nº Aciertos	%
1	9	1	4	44,44%	7	77,78%
2	8	2	3	37,50%	4	50,00%
3	10	0	6	60,00%	8	80,00%
4	8	2	4	50,00%	6	75,00%
5	8	2	2	25,00%	6	75,00%
6	9	1	5	55,56%	8	88,89%
7	10	0	7	70,00%	9	90,00%
8	7	3	4	57,14%	5	71,43%
9	7	3	4	57,14%	5	71,43%
10	8	2	1	12,50%	6	75,00%
11	8	2	6	75,00%	7	87,50%
12	9	1	4	44,44%	8	88,89%
TOTAL	**101**	**19**	**50**	**49,50%**	**79**	**78,22%**

Tabla 25. Resultado pruebas Ligue 1

Como se puede ver en los resultados de esta competición, las tasas de acierto alcanzadas están por debajo de los niveles esperados. La principal explicación para este hecho es el gran porcentaje de partidos para los que el sistema ha arrojado una predicción. En concreto un 84% de los partidos analizados han tenido una predicción calculada por el sistema, fruto del consenso de los cuatro clasificadores que se encargan de generar las predicciones para esta competición. Este alto porcentaje de consenso puede residir en la utilización de atributos muy similares del conjunto para realizar las predicciones.

Respecto a las cuotas necesarias para generar beneficios, para esta competición tenemos los siguientes resultados:

- **Predicción Simple:**

$0{,}495(x-1) - 0{,}505 > 0 \rightarrow x > 1/0{,}495 \rightarrow \mathbf{x > 2{,}02}$

- **Doble Oportunidad:**

$0{,}7822(x-1) - 0{,}2178 > 0 \rightarrow x > 1/0{,}7822 \rightarrow \mathbf{x > 1{,}27}$

Como puede observarse en los cálculos superiores la cuota media necesaria para sacar beneficios utilizando el sistema de predicción simple se ha disparado hasta una cuota de 2,02, una cuota muy elevada que hace que sea prácticamente imposible generar beneficios a través de este sistema.

La cuota media necesaria para generar beneficios que se ha calculado a partir de la tasa de aciertos del sistema de doble oportunidad arroja resultados más esperanzadores, ya que la cuota media se sitúa muy próxima a las cuotas calculadas en otras competiciones.

A continuación se mostrará la comparativa entre las tasas de acierto conseguidas y los objetivos establecidos antes del estudio (Figura 34 y Figura 35).

Como puede observarse en la Figura 34, los resultados no son todo lo buenos que se esperaban. Aunque la tasa de aciertos se sitúa muy por encima del objetivo mínimo, está no ha sido capaz de superar el objetivo esperado en ninguna de las semanas del estudio, quedando un 5% por debajo del objetivo fijado.

Como ya habíamos adelantado anteriormente, este hecho se produce principalmente por el alto grado de consenso que ha registrado el sistema para esta competición y que no ha impedido que partidos con un riesgo alto sean rechazados para apostar por el sistema.

Además, como puede observarse en la gráfica, la tasa no llega a estabilizarse del todo a lo largo de las 12 semanas, por lo que no se puede saber con certeza si el nivel del 50% alcanzado es el definitivo o si por el contrario la tasa va a sufrir alguna corrección hacia arriba o hacia abajo.

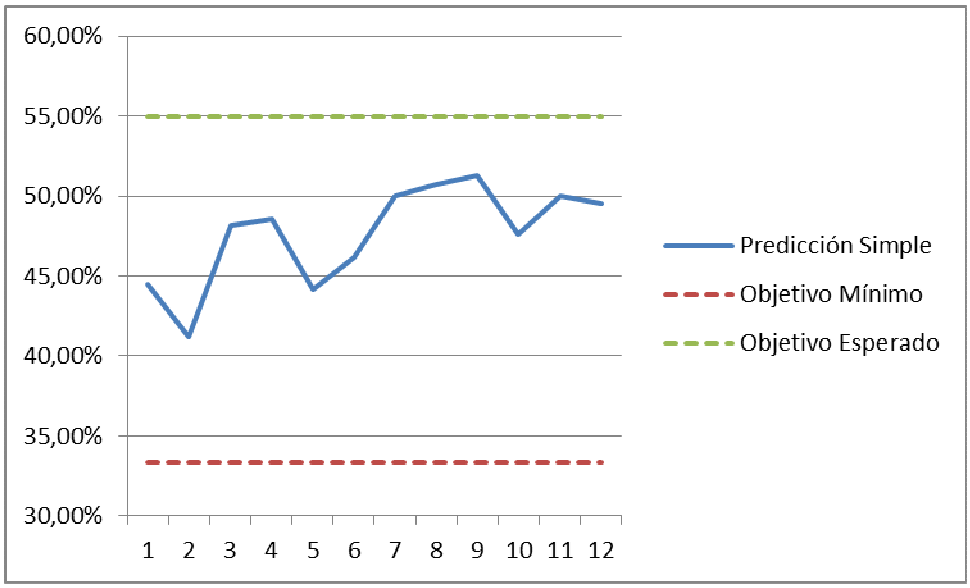

Figura 34. Resultados vs. Objetivos Ligue 1 Predicción Simple

Si nos fijamos ahora en la Figura 35 creada a partir del análisis de la tasa de aciertos para el sistema de Doble Oportunidad, se puede ver que aunque los datos son algo mejores que en el caso anterior, éstos no llegan a alcanzar el objetivo esperado para esta competición.

La tasa parece haberse estabilizado sobre la séptima semana en la franja entre el 75% - 80%, por lo que no se espera que haya grandes variaciones para esta competición a lo largo del tiempo.

De momento, y con los resultados obtenidos para esta competición, estos son los peores resultados registrados en todas las competiciones analizadas. Estos datos muestran que habrá cierta dificultad a la hora de generar beneficios con esta competición, pero no quita que ciertos partidos de bajo riesgo sean tomados para ser combinados con partidos de otras competiciones para la generación de beneficios en la siguiente fase de aplicación de estrategias de apuestas.

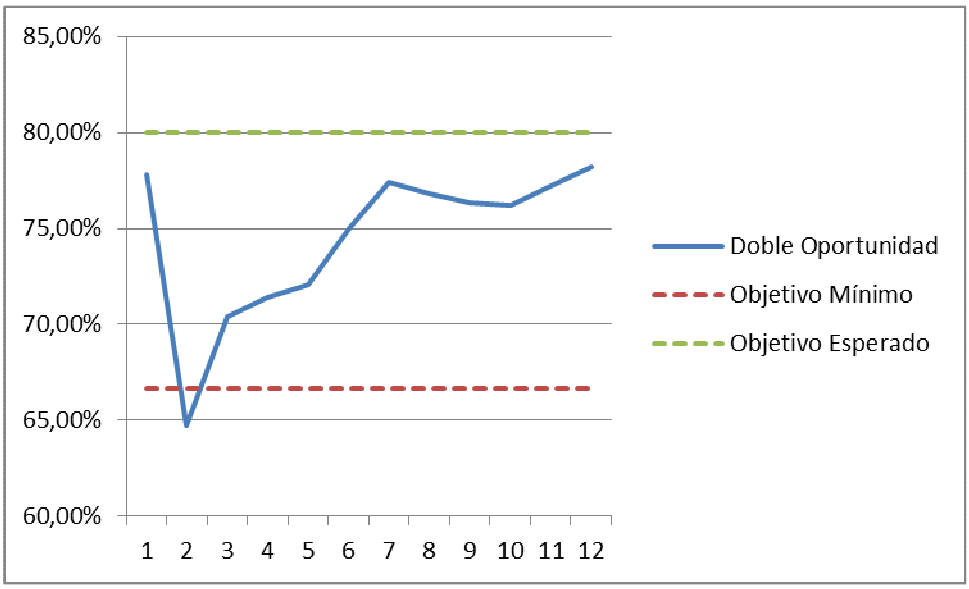

Figura 35. Resultados vs. Objetivos Ligue 1 Doble Oportunidad

4.3.7 Fase de Pruebas Premier League

La siguiente competición en ser analizada será la primera división inglesa de fútbol, también conocida como Premier League. Las características de este estudio son iguales que las del resto de competiciones de liga que se han analizado ya.

Los datos obtenidos tras ejecutar la hoja de predicciones y comprobar los resultados de los partidos aparecen resumidos en la Tabla 26.

	Partidos		Predicciones Simples		Predicciones Doble Oportunidad	
Semana	Con Predicción	Sin Predicción	Nº Aciertos	%	Nº Aciertos	%
1	4	6	1	25,00%	2	50,00%
2	3	7	2	66,67%	3	100,00%
3	5	5	4	80,00%	4	80,00%
4	6	4	5	83,33%	5	83,33%
5	6	4	6	100,00%	6	100,00%
6	6	4	5	83,33%	6	100,00%
7	7	3	6	85,71%	7	100,00%
8	6	4	1	16,67%	4	66,67%
9	6	4	4	66,67%	5	83,33%
10	7	3	4	57,14%	6	85,71%
11	7	3	4	57,14%	5	71,43%
12	5	5	3	60,00%	5	100,00%
TOTAL	**68**	**52**	**45**	**66,18%**	**58**	**85,29%**

Tabla 26. Resultado pruebas Premier League

Como puede observarse en la Tabla 26, los buenos datos de la tasa de aciertos van acompañados nuevamente de un discreto porcentaje de partidos en los que el sistema arroja predicción. El bajo porcentaje de partidos con predicción asegura de nuevo que los partidos con mayor riesgo en la predicción han sido descartados debido a que los cuatro clasificadores no han llegado a un consenso ante la posibilidad de que el resultado final del partido no sea el más probable.

Respecto a las cuotas necesarias para generar beneficios, para esta competición tenemos los siguientes resultados:

- **Predicción Simple:**

$0{,}6618(x-1) - 0{,}3382 > 0 \rightarrow x > 1/0{,}6618 \rightarrow \mathbf{x > 1{,}51}$

- **Doble Oportunidad:**

$0{,}8529(x-1) - 0{,}2178 > 0 \rightarrow x > 1/0{,}8529 \rightarrow \mathbf{x > 1{,}17}$

Los resultados obtenidos en el estudio de las cuotas medias necesarias para la generación de beneficios muestran unos grandes resultados, ya que se han obtenido las mejores cuotas en el método de Predicción Simple para todo el conjunto de ligas analizadas hasta el momento. Además la cuota obtenida con el método de Doble Oportunidad se sitúa muy cercana a la registrada en la Liga BBVA.

Estas cuotas sitúan a esta competición entre las mejores para la generación de beneficios, ya que las cuotas medias que se deben asumir para obtener beneficios se aproximan bastante a las ofrecidas por las casas de apuestas.

En cuanto a la comparativa de las tasas de acierto con los objetivos establecidos para esta competición, se va a mostrar a continuación las gráficas que muestran los datos de cada uno de los dos métodos utilizados (Figura 36 y Figura 37).

La Figura 113 que refleja la tasa de aciertos acumulada durante las 12 semanas de estudio, ratifica los datos que se habían mostrado anteriormente en la Tabla 26. Tras una primera semana en la que se obtuvieron unos muy malos resultados, la tasa ha ido ascendiendo hasta alcanzar un pico cercano al 80% en la semana 7, para después relajarse algo por debajo del 70%.

La tasa no parece estabilizada del todo, pero la tendencia que muestra no parece hacer peligrar el objetivo esperado del 55%. Se espera que la estabilización de la tasa pueda situarse en torno al 65%.

Estos datos no hacen más que ratificar que a pesar de tener clasificadores que no obtienen unos resultados espectaculares, al combinarlos pueden generar resultados realmente buenos.

En este caso no se optó por establecer el objetivo esperado en el 60% (como ocurría en la Liga BBVA), ya que en esta competición no se llegaban a obtener resultados tan buenos como en la competición española. Finalmente, los resultados obtenidos han demostrado que los datos finales no dependen tanto de la tasa de aciertos del clasificador, y se centran más en qué partidos son elegidos por el sistema como los más seguros para emitir una predicción.

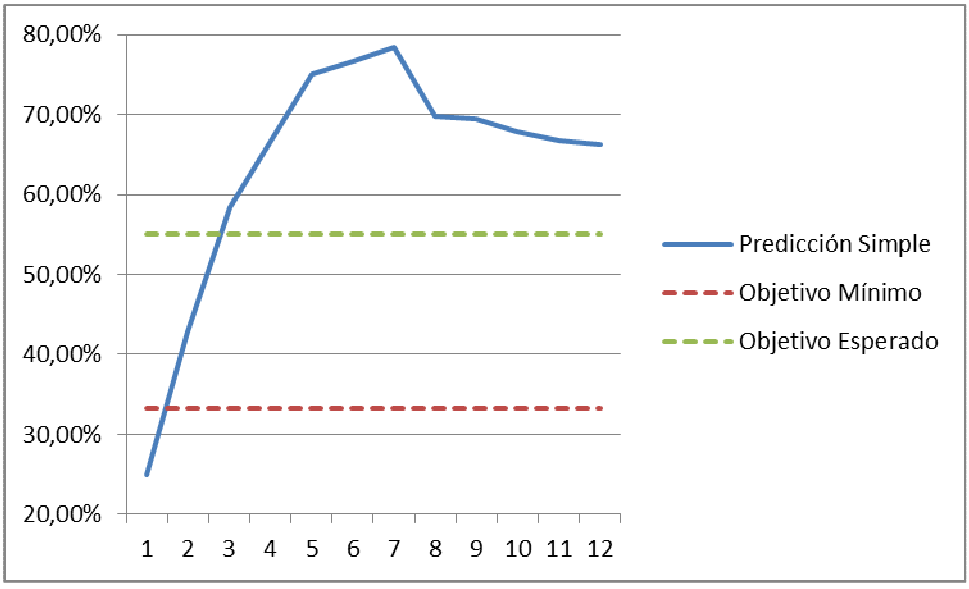

Figura 36. Resultados vs. Objetivos Premier League Predicción Simple

Los resultados obtenidos con el método de Doble Oportunidad son tan buenos como los que se han mostrado del método de Predicción Simple. Al igual que en el caso anterior, la tasa se recupera de una primera semana con sólo un 50% de aciertos y llega a situarse cerca del 90% en la séptima semana para después estabilizarse en torno al 85%, que es donde se ha situado la tasa de aciertos al final de la duodécima semana (Figura 37).

Estos datos muestran la fortaleza del modelo de predicción diseñado para esta competición, además se señalar que esta competición será una de las más utilizadas por las estrategias de apuestas a la hora de combinar partidos para generar beneficios.

Por último, cabe comentar que la estabilización de la tasa es prácticamente total en la franja del 85%, por lo que a largo plazo no se esperan modificaciones en este parámetro.

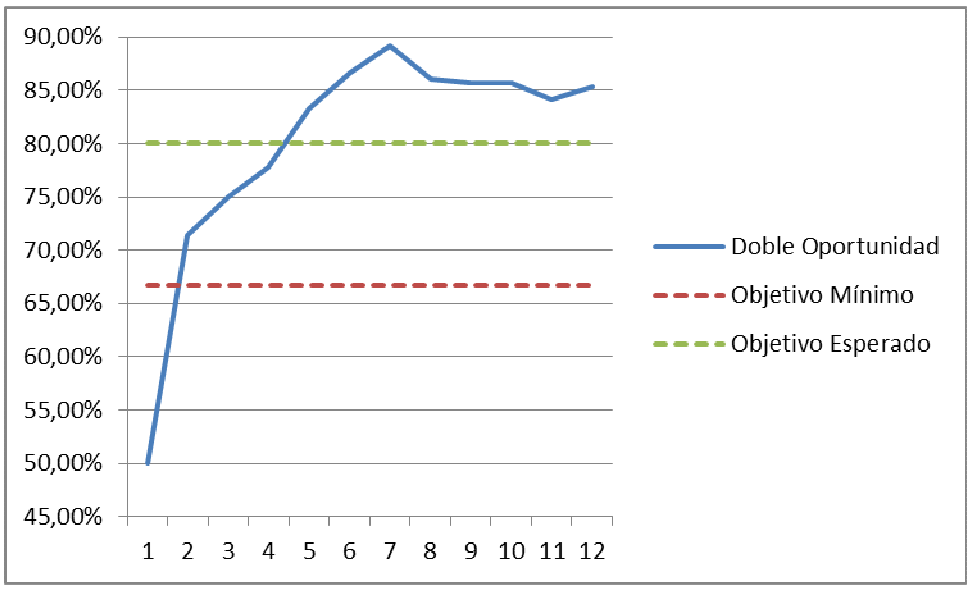

Figura 37. Resultados vs. Objetivos Premier League Doble Oportunidad

4.3.8 Fase de Pruebas Serie A

La siguiente de las competiciones en ser analizada será la primera división de fútbol italiana, también conocida como la Serie A. Al igual que ha ocurrido con el resto de competiciones de liga analizadas, para este caso tendremos diez partidos semanales, en los que en ocasiones, debido a la distribución del calendario de la competición puede coincidir que un mismo equipo juegue más de una vez en cada semana.

Los resultados que se han obtenido tras el análisis de las predicciones del sistema han sido los que aparecen en la Tabla 27.

	Partidos		Predicciones Simples		Predicciones Doble Oportunidad	
Semana	Con Predicción	Sin Predicción	Nº Aciertos	%	Nº Aciertos	%
1	7	3	5	71,43%	6	85,71%
2	6	4	4	66,67%	5	83,33%
3	6	4	3	50,00%	6	100,00%
4	4	6	2	50,00%	4	100,00%
5	6	4	3	50,00%	3	50,00%
6	6	4	4	66,67%	5	83,33%
7	5	5	4	80,00%	5	100,00%
8	5	5	4	80,00%	4	80,00%
9	7	3	4	57,14%	7	100,00%
10	7	3	5	71,43%	6	85,71%
11	7	3	6	85,71%	7	100,00%
12	7	3	1	14,29%	3	42,86%
TOTAL	**73**	**47**	**45**	**61,64%**	**61**	**83,56%**

Tabla 27. Resultado pruebas Serie A

Los resultados de la Tabla 27 muestran los datos obtenidos tras la ejecución de nuestro sistema de predicción para la competición de la Serie A. Para empezar a analizar los resultados, hay que fijarse en la tasa de partidos para los que el sistema ha ofrecido una predicción fruto del consenso de los cuatro clasificadores. El sistema ha ofrecido una predicción para el 60% de los partidos, una tasa no muy alta aunque si superior al de las competiciones que han obtenido mejores tasas de acierto.

Si nos fijamos en las tasas de acierto obtenidas, éstas se sitúan muy por encima de los objetivos marcados para esta competición. Estos buenos resultados residen principalmente en la capacidad del sistema de clasificación para descartar aquellos partidos que suponen cierto riesgo a nuestras apuestas.

Respecto a las cuotas medias necesarias para generar beneficios, para esta competición tenemos los siguientes resultados:

- **Predicción Simple:**

$0{,}6164(x-1) - 0{,}3836 > 0 \rightarrow x > 1/0{,}6164 \rightarrow \mathbf{x > 1{,}62}$

- **Doble Oportunidad:**

$0{,}8356(x-1) - 0{,}1644 > 0 \rightarrow x > 1/0{,}8356 \rightarrow \mathbf{x > 1{,}19}$

Tras analizar las cuotas medias necesarias para la obtención de beneficios en esta competición, podemos observar que tenemos buenas cuotas aunque podrían no ser suficientes para la generación de beneficios en siguientes fases del proyecto, ya que las cuotas que se ofrecen en esta competición no suelen ser excesivamente elevadas. En cualquier caso se esperará a la evaluación de las estrategias de apuestas para valorar correctamente si el modelo desarrollado es capaz de generar beneficios.

A continuación se van a mostrar las gráficas que recogen la comparativa entre las tasas de acierto obtenidas y los objetivos establecidos para esta competición (Figura 38 y Figura 39):

En la Figura 38 se puede ver la comparativa entre la tasa de aciertos para el método de predicción simple y los objetivos fijados para esta competición. Como se puede observar el comportamiento de la tasa de aciertos es excelente, situándose en todo momento por encima del objetivo esperado y estabilizándose alrededor del 65% a partir de la séptima semana de análisis.

Estos datos muestran la estabilidad del modelo desarrollado, ya que no se esperan movimientos futuros en la tasa de aciertos.

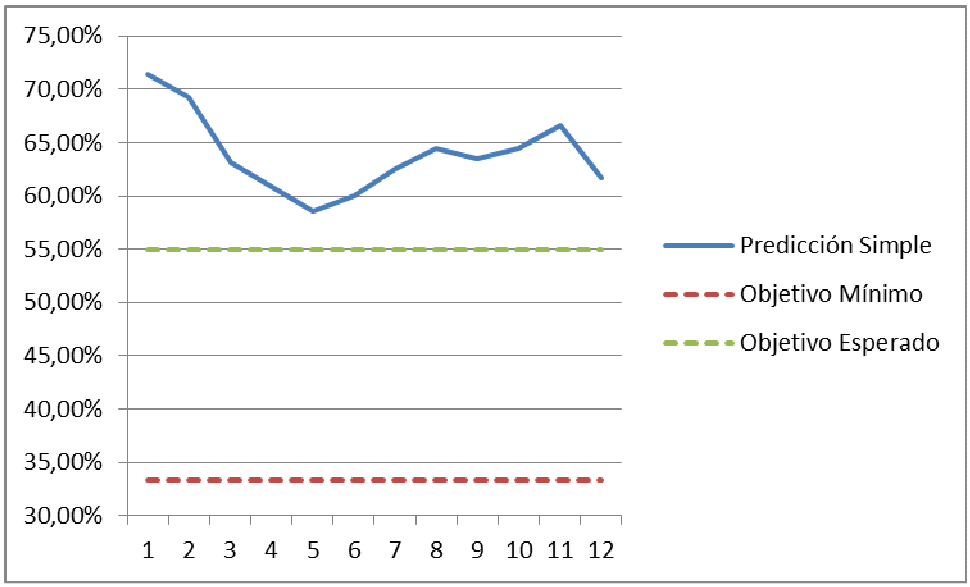

Figura 38. Resultados vs. Objetivos Serie A Predicción Simple

En cuanto a la Figura 38, que muestra la comparativa entre la tasa de aciertos obtenida con el método de Doble Oportunidad y los objetivos establecidos para esta competición se pueden ver datos similares a los obtenidos en el método de Predicción Simple, en el que tenemos una tasa por encima de los objetivos a lo largo de las doce semanas. La estabilización de la tasa es algo más débil que en el caso anterior, pero logra establecerse en la franja cercana al 85% de aciertos, una tasa que podría ser muy buena para optar a obtener beneficios en las fases posteriores del proyecto.

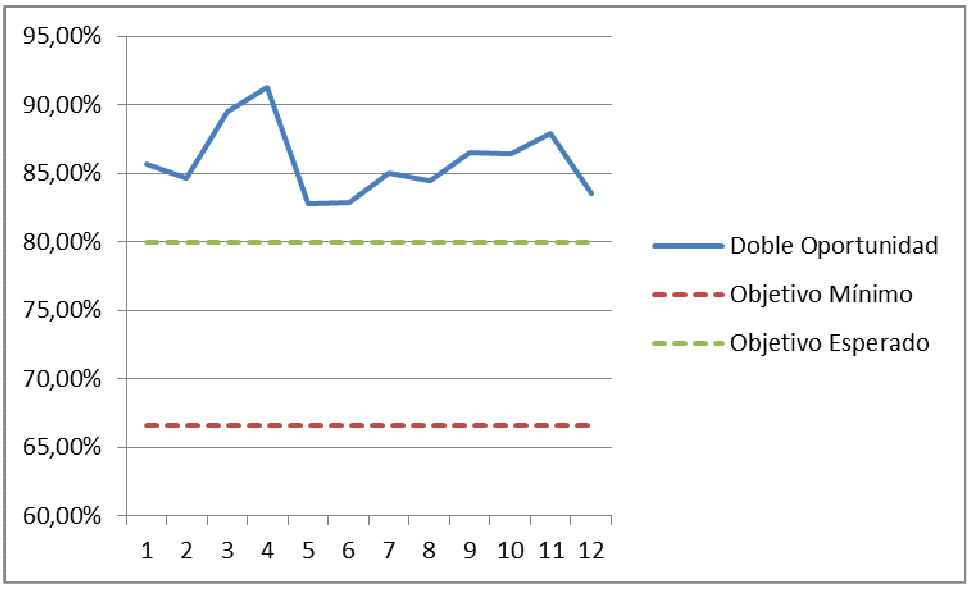

Figura 39. Resultados vs. Objetivos Serie A Doble Oportunidad

4.3.9 Fase de Pruebas Otras Ligas

Antes de entrar en materia de análisis hay que recordar el propósito por el que se generó el conjunto denominado "Otras Ligas". Si recordamos, se desarrollaron dos clasificadores que estaban basados en los datos de los partidos de la competición de liga de varios países europeos. Lo que se quería probar con esos clasificadores era ver si se podían aplicar a cualquier competición de liga del mundo para obtener unos buenos porcentajes de acierto en la predicción de resultados.

Para probar la efectividad de este modelo se han elegido ciento veinte partidos en su mayoría de la liga alemana de fútbol, aunque también se cuenta con algunos partidos de la primera división portuguesa. Se ejecutará la macro que realiza las predicciones para este conjunto y se comprobará si los resultados obtenidos por el sistema de predicción son iguales a los que tuvieron lugar al finalizar cada uno de los partidos analizados. Los datos obtenidos tras el análisis realizado se muestran en la Tabla 28.

	Partidos		Predicciones Simples		Predicciones Doble Oportunidad	
Semana	Con Predicción	Sin Predicción	Nº Aciertos	%	Nº Aciertos	%
1	7	3	3	42,86%	6	85,71%
2	8	2	6	75,00%	6	75,00%
3	10	0	4	40,00%	8	80,00%
4	10	0	5	50,00%	8	80,00%
5	10	0	6	60,00%	8	80,00%
6	10	0	4	40,00%	6	60,00%
7	9	1	4	44,44%	7	77,78%
8	9	1	3	33,33%	5	55,56%
9	9	1	7	77,78%	9	100,00%
10	7	3	3	42,86%	5	71,43%
11	9	1	5	55,56%	7	77,78%
12	10	0	5	50,00%	9	90,00%
TOTAL	**108**	**12**	**55**	**50,93%**	**84**	**77,78%**

Tabla 28. Resultado pruebas Otras Ligas

Lo primero en lo que hay que fijarse a la hora de analizar el modelo es el alto porcentaje de partidos para los que el sistema ha arrojado predicción. Esto tiene una explicación lógica, y es que para estos partidos tan solo intervienen dos clasificadores en la predicción, por lo que es más sencillo que se pongan de acuerdo entre ellos a la hora de emitir un resultado.

Esta facilidad a la hora de emitir resultados hace que las tasas de acierto bajen, ya que los partidos con más riesgo no son rechazados por el sistema al coincidir los dos clasificadores en el veredicto emitido.

Respecto a las cuotas medias necesarias para generar beneficios, para esta competición tenemos los siguientes resultados:

- **Predicción Simple:**

$0{,}5093(x-1) - 0{,}3836 > 0 \rightarrow x > 1/0{,}5093 \rightarrow \mathbf{x > 1{,}96}$

- **Doble Oportunidad:**

$0{,}7778(x-1) - 0{,}1644 > 0 \rightarrow x > 1/0{,}7778 \rightarrow \mathbf{x > 1{,}28}$

Como se puede ver en los cálculos realizados en la parte superior, las cuotas medias necesarias para generar beneficio son bastante altas, y a primera vista no parece que puedan compensar las buenas cuotas que se ofrecen en competiciones de liga que no son de primera línea mundial. No obstante habrá que ver en el posterior capítulo la capacidad de este modelo para generar beneficios.

En cuanto a la comparativa entre las tasas obtenidas y los objetivos fijados para este conjunto de partidos, a continuación se mostrarán unas gráficas que mostrarán los resultados obtenidos en este estudio (Figura 40 y Figura 41).

Como puede observarse en la Figura 40, la tasa de aciertos se sitúa prácticamente en el límite establecido por el objetivo esperado de las predicciones del sistema. Aunque esta tasa se sitúa por encima del objetivo hay que tener en cuenta que el objetivo para este conjunto de partidos había sido establecido en el 50% debido a los pobres resultados registrados en la fase de entrenamiento.

La estabilización del valor de la tasa nos hace pensar que este parámetro va a situarse en el 50% a lo largo del tiempo.

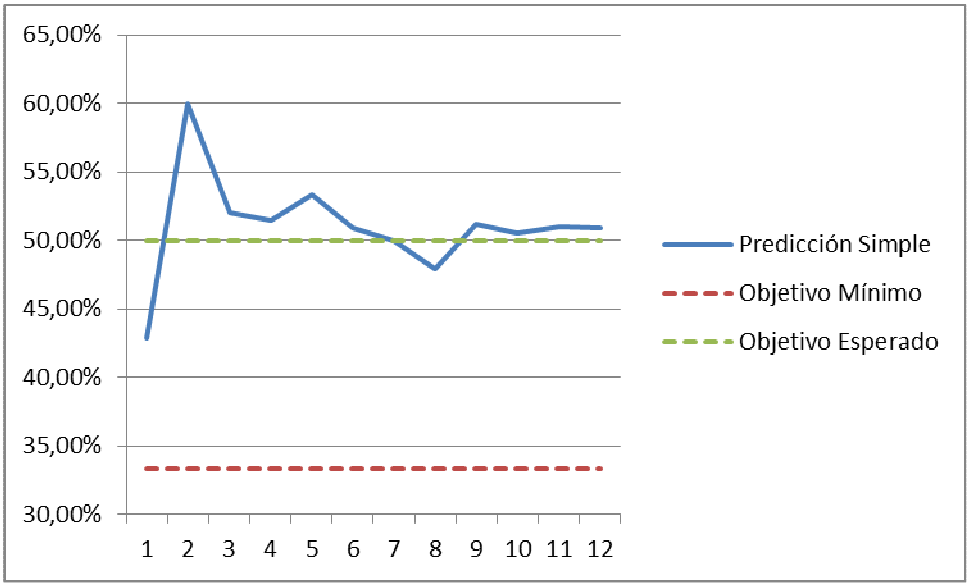

Figura 40. Resultados vs. Objetivos Otras Ligas Predicción Simple

A la hora de llevar a cabo este estudio se ha detectado una peculiaridad que no se había dado en el resto de conjuntos, y es que el sistema ha arrojado varios empates como predicción a algunos partidos del conjunto. Esto ha hecho que a la hora de aplicar el método de doble oportunidad se optara por sólo predecir sobre ese resultado para los pocos partidos que han registrado esta casuística.

Si nos fijamos ahora en la Figura 41 que muestra la comparativa entre la tasa de aciertos y los objetivos fijados podemos ver unos resultados algo mejores que para el método de predicción simple. En este caso la tasa ha logrado estabilizarse con un diferencial algo superior al que se ha registrado en el estudio con el método anterior. Esto hace que los resultados obtenidos sean algo más prometedores que en el caso anterior, aunque siguen estando lejos de los que se han obtenido en otras competiciones analizadas anteriormente.

Una vez más, habrá que esperar al análisis de beneficios del modelo para catalogar como éxito o no el sistema desarrollado para este conjunto de partidos.

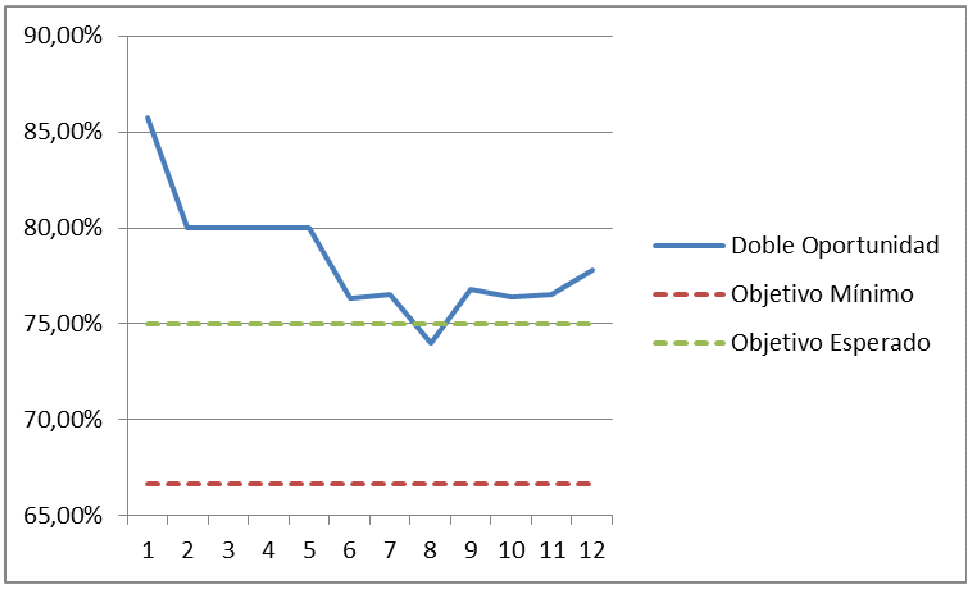

Figura 41. Resultados vs. Objetivos Otras Ligas Doble Oportunidad

4.3.10 Fase de Pruebas NBA

Por último y para concluir con este apartado, se van a analizar los resultados obtenidos al analizar las predicciones realizadas para el sistema para el conjunto de partidos de la competición de la NBA.

Esta competición presenta algunas peculiaridades, ya que en este caso al haber únicamente dos resultados posibles para cada uno de los partidos, no se podrá apostar utilizando el método de la Doble Oportunidad. Además, los objetivos establecidos para este conjunto son algo más altos debidos a la reducción del espacio de resultados posibles para cada uno de los partidos. Los resultados obtenidos se muestran en la Tabla 29.

	Partidos		Predicciones Simples	
Semana	Con Predicción	Sin Predicción	Nº Aciertos	%
1	20	0	10	50,00%
2	20	0	16	80,00%
3	20	0	15	75,00%
4	20	0	13	65,00%
5	20	0	14	70,00%
6	20	0	12	60,00%
7	20	0	14	70,00%
8	20	0	15	75,00%
9	20	0	16	80,00%
10	20	0	17	85,00%
11	20	0	12	60,00%
12	20	0	14	70,00%
TOTAL	**240**	**0**	**168**	**70,00%**

Tabla 29. Resultado pruebas NBA

Como puede observarse en la tabla posterior, el consenso que ha habido entre los dos clasificadores ha sido total, ya que los doscientos cuarenta partidos analizados han tenido una predicción por parte del sistema.

Esto hace que nuestras estimaciones sobre la tasa de aciertos se van muy perjudicadas, ya que no habrá ningún partido con alto riesgo que sea rechazado por el sistema para apostar.

En cualquier caso, aunque no se llegue al objetivo esperado la tasa del 70% de aciertos obtenida está veinte puntos por encima del objetivo mínimo que representa la tasa de aciertos que se podría alcanzar por azar.

La efectividad mostrada por el sistema es más que aceptable, ya que excepto en la primera semana de pruebas, las tasas de acierto semanales se han situado en valores por encima del 60%.

Respecto a las cuotas medias necesarias para generar beneficios, para esta competición tenemos los siguientes resultados:

- **Predicción Simple:**

$0{,}7(x-1) - 0{,}3 > 0$ ➔ $x > 1/0{,}7$ ➔ **$x > 1{,}42$**

En cuanto a la cuota media necesaria para obtener beneficios en esta competición tenemos un valor que podría ser algo alto en comparación con las cuotas que ofrecen las casas de apuestas.

Por último se comparará la tasa semanal acumulada generada por este modelo con los objetivos que habían sido establecido para esta competición. La Figura 42 muestra la comparativa entre estos parámetros.

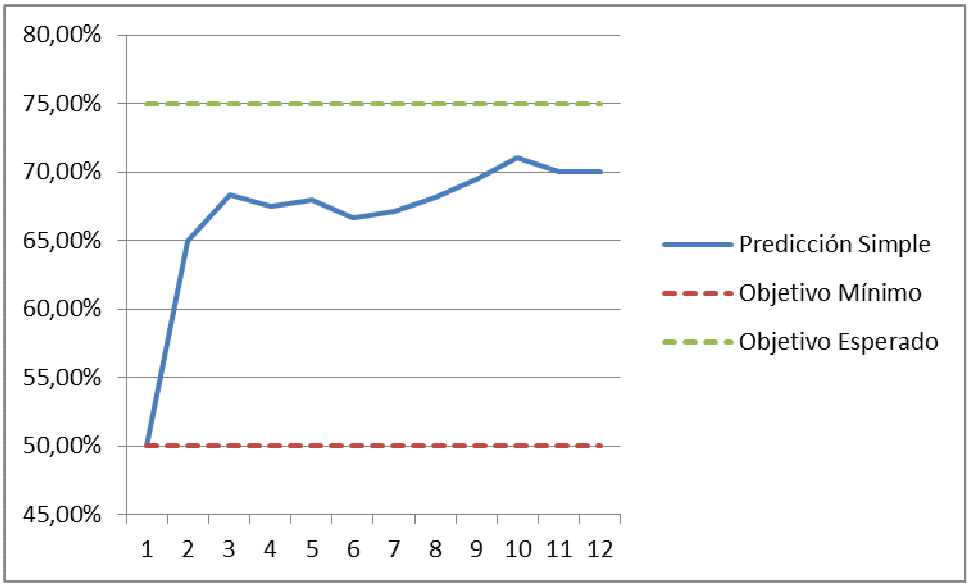

Figura 42. Resultados vs. Objetivos NBA Predicción Simple

Como se puede observar en la Figura 42, la tasa de aciertos obtenida con el estudio de este conjunto se ha logrado estabilizar en torno al 70% de aciertos tras los malos datos de las primeras semanas.

La tasa obtenida está por debajo del objetivo esperado para esta competición, pero muy por encima del objetivo mínimo establecido. Este hecho, que no estaba previsto, se ha debido principalmente a la nula eliminación de partidos por parte del sistema de predicción dado el 100% de consenso al que han llegado los dos clasificadores utilizados en el sistema.

Como en ocasiones anteriores, habrá que esperar para ver el conjunto de partidos de esta competición es capaz de generar beneficios con alguna de las estrategias que se presentarán posteriormente o al menos que ciertos partidos en combinación con otros de diferentes competiciones logren al combinarse buenos beneficios para el estudio.

4.4 Conclusiones

Tras concluir las pruebas del sistema de predicción implementado sobre la hoja Excel, se han podido extraer una gran cantidad de conclusiones que son muy valiosas para el estudio que se esta realizando.

En primer lugar hay que hablar de los resultados obtenidos para la tasa de aciertos de cada uno de los modelos. Estos resultados han sido francamente buenos en la mayoría para la mayoría de competiciones estudiadas. Tan solo las competiciones de la primera división francesa de fútbol (Ligue 1) y la competición de baloncesto de la NBA no han llegado a alcanzar los objetivos esperados que fueron fijados tras el entrenamiento de los conjuntos mediante los clasificadores seleccionados para implementar el modelo.

Se ha podido observar que hay una gran correlación entre buenas tasas de acierto y el consenso de los clasificadores que participan en la predicción. Como se pudo ver en los datos del estudio, las dos competiciones que no alcanzaron el objetivo esperado tenían un elevado grado de consenso entre los clasificadores que formaban su modelo de predicción. Este elevado consenso ha hecho que los partidos de más riesgo de estas competiciones no sean desechados, como ha ocurrido en el resto de conjuntos. Entre los puntos a mejorar del sistema, habrá que plantear cómo conseguir que no haya un consenso elevado, que en el caso de la NBA ha llegado a ser del 100% de los partidos, lo que hace muy difícil que se obtengan buenos resultados en las predicciones.

Otra de las conclusiones clave que pueden sacarse del estudio realizado en este apartado es el buen dato de aciertos que desprende nuestro estudio de cobertura de resultados, estudio realizado mediante la aplicación de la estrategia de predicción de Doble Oportunidad. Como se ha podido observar en los datos del estudio, la aplicación de esta estrategia reduce el riesgo de fallo, reduciendo también los posibles beneficios que se pueden obtener mediante la apuesta de ese resultado. Esta estrategia será tenida muy en cuenta en el siguiente capítulo donde parte de las estrategias de apuesta seguidas estarán basadas en el sistema de predicción de Doble Oportunidad.

Por último y ya mirando hacia el siguiente capítulo, en el que se hablará sobre la explotación del sistema de predicción mediante diferentes estrategias de apuesta, hablaremos sobre la cuota media para generar beneficios que ha sido calculada para cada una de las competiciones estudiadas. Esta cuota varía según la tasa de aciertos obtenida en cada competición. Si observamos las cuotas obtenidas parece que se necesitará la combinación de varios partidos para poder obtener beneficios, ya que las cuotas medias necesarias parecen situarse por debajo de la media ofrecida por las casas de apuestas. En cualquier caso, se definirán diferentes estrategias de apuesta que contemplarán casos en los que se realicen apuestas simples y otros en los que se realicen apuestas combinadas, de manera que cubramos ampliamente todas las opciones que tenemos a nuestra disposición.

Capítulo 5

EXPLOTACIÓN DEL SISTEMA A TRAVÉS DE CASAS DE APUESTAS

5.1 Introducción

En este nuevo capítulo se tratará de comprobar si el sistema de predicción es capaz de generar beneficios aplicando los resultados del sistema y diversos métodos de apuestas en el entorno de las casas de apuestas.

Para realizar esta comprobación se usará el conjunto de partidos que se utilizó en la fase de pruebas del sistema de predicción para ver el porcentaje de aciertos que se obtenía en cada una de las competiciones. Al estar los partidos del conjunto divididos por semanas, se podrán establecer conjuntos de partidos sobre los que hacer apuestas en cada una de las semanas. Dependiendo del resultado de la ejecución del sistema de predicción y de la estrategia que sea usada para extraer beneficios, tendremos diferentes conjuntos de partidos sobre los que se harán apuestas de manera individual o combinada.

Las distintas estrategias que se van a probar en esta sección son en su mayoría estrategias ampliamente conocidas por los apostantes. Alguna de ellas incluso es recomendada y ofrecida de manera automática por algunas casas de apuestas debido a su popularidad.

Además de estas estrategias ya conocidas por la mayoría de los apostantes, se ha desarrollado una nueva estrategia basada en un **algoritmo genético**. Este algoritmo genético tendrá como entrada las predicciones de nuestro sistema de predicción y el riesgo estimado para cada uno de los partidos y creará lo que se ha denominado como **Cartera de Apuestas**. Esta cartera de apuestas representa al conjunto de apuestas que maximiza la relación entre beneficio y riesgo al ser combinadas en una casa de apuestas. Para realizar este proceso de maximización se utiliza una función de evaluación que combina los parámetros mencionados de beneficio esperado y riesgo de la apuesta que será explicada más adelante junto con todo el proceso de creación del algoritmo.

Al finalizar el análisis que se va a exponer en este capítulo seremos capaces de determinar si el sistema diseñado es capaz de generar algún beneficio con alguna de las estrategias propuestas.

5.2 Estrategias de Apuestas

En esta sección se explicarán detalladamente qué estrategias se van a seguir para intentar obtener beneficios con el sistema de predicción. Para cada una de las estrategias que va a ser utilizadas se incluirá una explicación del sistema que siguen y algunas ventajas e inconvenientes del uso de éstas. Además, se detallará un código de cuatro caracteres que servirán para identificar cada una de las estrategias en las tablas de datos que se muestren posteriormente.

5.2.1 Partidos Individuales con Apuesta Simple (PIAS)

Esta estrategia es la más simple de todas las que se van a utilizar. El sistema que se va a utilizar es coger todos los resultados que predice el sistema de predicción y apostar una cantidad fija estipulada con antelación. Esta cantidad fija será la misma para todas las apuestas y competiciones.

La ventaja principal que ofrece este modelo de apuestas propuesto es la sencillez con la que se seleccionan los partidos y el importe a apostar, ya que el criterio de selección de partidos se basa en la elección de todos aquellos partidos para los que el sistema ha arrojado una predicción, mientras que el dinero a apostar siempre va a ser una cantidad fija que por simplicidad asumiremos que es 1€.

Si nos fijamos en los inconvenientes que presenta este modelo de predicción es que las cuotas que se van a manejar en cada uno de los partidos pueden no ser lo suficientemente altas como para poder cubrir los fallos que tenga nuestro sistema de predicción. Con esta afirmación lo que se quiere hacer ver es que si se quiere tener beneficios con un sistema que tiene un 50% de aciertos, la tasa media de las apuestas acertadas debe ser superior a dos para que los beneficios obtenidos contrarresten las pérdidas. Si nos fijamos en las cuotas que manejamos en cada uno de los conjuntos podremos observar que éstas pueden no ser lo suficientemente altas como para que esta estrategia genere beneficios.

5.2.2 Partidos Individuales con Doble Oportunidad (PIDO)

Esta estrategia de apuestas es prácticamente igual que la anterior, pero en este caso en lugar de apostar sólo al resultado que arroja el sistema de predicción, también se apostará al empate (en las competiciones que sea posible) para así reducir el riesgo de la apuesta, aunque esto conlleve una reducción del beneficio potencial que se puede obtener con ésta.

Por tanto la estrategia a seguir será apostar a todos los partidos para los que el sistema arroja una predicción. Para cada partido habrá dos apuestas, una para el resultado que predice el sistema de predicción además del empate. En cuanto la cantidad a apostar, en este caso tendremos que calcular el dinero que se coloca en cada apuesta para asegurar que con cualquiera de las dos apuestas que realicemos vamos a ganar el mismo dinero. La fórmula que se va a seguir para calcular las cantidades a apostar es muy sencilla y se rige por el siguiente razonamiento:

Imaginemos que nuestro sistema de predicción dice que para un determinado partido el resultado más probable es que gane el equipo local y que la cuota que ofrece la casa de apuestas para ese resultado es de 1,5, mientras que la cuota de empate es 3,3.

Si seguimos la estrategia de apuestas explicada en la parte superior tendríamos que apostar a la victoria local y al empate, el problema que queda por resolver es cuanto dinero distribuimos para cada una de las apuestas. Para asegurarnos de que tanto si gana el equipo local como si hay empate ganamos el mismo dinero realizaremos las siguientes operaciones:

Suma de Cuotas = 1,5 + 3,3 = **4,8**

Coeficiente Victoria Local = 1 – (1,5 / 4,8) = **0,6875**

Coeficiente Empate = 1 – (3,3 / 4,8) = **0,3125**

Suma de Dinero Disponible = **1€**

Dinero Apuesta Victoria Local = 0,6875 * 1€ = **0,6875€**

Dinero Apuesta Empate = 0,3125 * 1€ = **0,3125€**

Para comprobar que los cálculos son correctos comprobaremos si al ganar las dos apuestas el beneficio es el mismo:

Victoria Local = 0,6875€ * 1,5 = **1,03125€**

Empate = 0,3125€ * 3,3 = **1,03125€**

Una vez explicada la estrategia que se va a seguir con este modelo sólo falta detectar las ventajas y desventajas de su uso. Por la parte de las ventajas, seguimos teniendo como en el caso anterior un modelo sencillo, tan sólo alterado por la necesidad de calcular cuánto dinero se dedica a cada apuesta. Al ser un razonamiento sencillo el que se usa para calcular la distribución de dinero entre las dos apuestas seguiremos considerando que la complejidad de esta estrategia es baja. Otra de las ventajas que presenta este modelo es que al apostar dos resultados posibles aumentan las posibilidades de acierto.

En el caso de los inconvenientes que presenta esta estrategia tenemos el mismo inconveniente que en el caso anterior, ya que al dividir la inversión en dos resultados estamos produciendo una reducción de la cuota de la apuesta. La ventaja que obteníamos con el aumento de la tasa de aciertos lo contrarresta una bajada del beneficio potencial que podemos llegar a obtener. Aunque aumentemos la tasa de aciertos habrá que ver si ese aumento es contrarrestado por la bajada de cuotas que implica la utilización de esta estrategia.

5.2.3 Combinada de Competición (CC)

En esta nueva estrategia de apuestas introducimos el concepto de apuesta combinada [39]. Como bien apunta la referencia, una apuesta combinada es una apuesta en la que para obtener beneficios hay que acertar todos y cada uno de los partidos que forman parte de la apuesta. Tiene como ventaja que los posibles beneficios son mayores que los de una apuesta simple, debido a que la cuota de la apuesta es la multiplicación de cada una de las cuotas independientes de las apuestas que forman la combinada. La gran desventaja es que en cuanto se falla uno de los resultados la apuesta se pierde, por lo que el riesgo es mayor que en una apuesta simple.

El objetivo de esta estrategia es seleccionar los mejores partidos de una determinada competición y combinarlos en una sola apuesta para obtener mayores beneficios que con una apuesta normal. Cuando se habla de seleccionar los mejores partidos, hablamos de seleccionar los partidos que tengan un menor riesgo, para así minimizar tanto como sea posible la probabilidad de fallo de la apuesta combinada. Por tanto, para llevar a cabo esta estrategia se seleccionará los tres partidos con menor riesgo de cada competición y se combinarán entre ellos (siempre separando entre competiciones).

Las ventajas de este modelo son una vez más la sencillez que presenta, ya que sólo tendremos que fijarnos en qué partidos son los que tienen menor tasa de riesgo para seleccionar los partidos que finalmente conformarán la apuesta combinada. Además al combinar las apuestas obtendremos mejores tasas que si apostamos individualmente a cada uno de los partidos de la combinada.

Por el lado de los inconvenientes tenemos el gran inconveniente de que aumentamos el riesgo de fallar a medida que añadimos nuevas apuestas a la combinada. Para hacernos una idea, si tenemos dos apuestas que individualmente tienen un 50% de posibilidad de fallo, al combinarlas tendremos una probabilidad de fallo del 75%.

5.2.4 Combinada Variada (CV)

Esta nueva estrategia es exactamente igual que la anterior con la única salvedad de que en este caso se permitirá combinar partidos de diferentes competiciones con el objetivo de minimizar al máximo posible el riesgo que tengamos de fallar la apuesta combinada. Por ello, se elegirán tres combinadas de tres partidos cada una, donde estarán los partidos con menor riesgo de todo el conjunto semanal de partidos. Los partidos se combinarán en orden aleatorio, ya que las diferencias de riesgo entre los nueve partidos no serán muy grandes.

Las ventajas son exactamente las mismas que en la estrategia anterior, ya que la complejidad de este modelo es muy baja, permitiendo elegir de forma muy sencilla los partidos que formarán las apuestas combinadas. Además, al permitir combinar partidos de diferentes competiciones, lograremos reducir el riesgo de cada una de las combinadas.

En cuanto a los inconvenientes, tenemos el gran problema de operar con apuestas combinadas, que aunque aumentan los posibles beneficios lo hacen a costa de aumentar el riesgo de las apuestas.

5.2.5 Apuestas de Sistema en Función del Riesgo (ASFR)

Para desarrollar esta nueva estrategia tenemos que introducir el concepto de Apuesta de Sistema [40]. Una Apuesta de Sistema es una apuesta combinada en la que el fallo de algún partido puede no significar la pérdida de la apuesta. El tipo de apuesta de sistema más conocido y el cual se va a utilizar en esta estrategia es el Sistema 2/3 [41].

Esta apuesta combinada lo que hace es agrupar de dos en dos, tres partidos seleccionados por el usuario. Esto hace que se generen tres apuestas combinadas de dos partidos cada una. Si se falla uno de los partidos habrá dos apuestas que resultarán perdedoras, pero en cambio una de ellas devolverá dinero al usuario. En ocasiones, si la cuota es baja, el dinero retornado al usuario puede ser menor al que invirtió en un primer momento, pero al menos consigue no perder todo lo apostado como ocurriría con una apuesta combinada normal.

Para llevar a cabo esta estrategia se seleccionará para cada competición los tres partidos con menor riesgo para combinarlos como una apuesta de sistema 2/3. La cantidad que se apostará en cada una de las competiciones será fija, y por simplicidad se tomará 1€ como cantidad a apostar.

La principal ventaja de esta estrategia es el aumento de la seguridad de no perder todo lo invertido, al introducir la posibilidad de fallar uno de los resultados.

En cuanto a los inconvenientes ocurre lo mismo que viene pasando con el resto de estrategias para apostar, y es que el descenso del riesgo implica también un descenso en la posible rentabilidad de la apuesta.

5.2.6 Selección Genética con Lucky 15 (SG15)

La nueva estrategia que se va a explicar a continuación es el sistema Lucky 15, uno de los sistemas de apuestas más conocidos por los usuarios de las casas de apuestas. Este sistema realiza quince apuestas a partir de cuatro encuentros elegidos por el usuario. La distribución de esas quince apuestas es la siguiente:

- 4 apuestas simples.
- 6 apuestas dobles.
- 4 apuestas triples.
- 1 apuesta cuádruple.

El procedimiento que se va a utilizar para seleccionar los cuatro partidos que conformarán las apuestas semanales, tendrá asociado la ejecución de un algoritmo genético que se ha implementado específicamente para obtener los cuatro partidos que al ser combinados obtienen la mejor relación beneficio/riesgo.

Este algoritmo genético tiene como principal función elegir los mejores partidos en base a una función de evaluación o fitness. Esa función de fitness ha sido diseñada para primar un riesgo bajo en la apuesta sobre un elevado beneficio, ya que el objetivo de la selección de estos partidos es elaborar una apuesta combinada con los partidos seleccionados. Esta apuesta combinada estará limitada a cuatro partidos, pero se permitirá al algoritmo que genere combinaciones de menos número de partidos para así no limitar que se consigan mejores relaciones entre beneficios y riesgo de las apuestas.

Antes de comenzar a explicar los fundamentos y el funcionamiento del algoritmo, se ha de explicar un concepto con el que se va a trabajar a lo largo de la aplicación de esta estrategia. El concepto que se va a introducir es el de **Cartera de Apuestas**. Para explicar de manera sencilla qué es la cartera de apuestas, vamos a establecer un símil con el mundo de la bolsa de valores.

Imaginemos a un inversor que posee acciones de un determinado índice bursátil. Ese conjunto de acciones conforma su Cartera de Acciones. Esa cartera posee ciertas acciones del índice, acciones que el inversor ha elegido por diversas razones como pueden ser la rentabilidad de los dividendos de la acción, o la predicción de una futura subida del precio de esas acciones.

Si trasladamos el ejemplo de la cartera de acciones al mundo de las apuestas tendremos que los partidos son nuestras acciones y el conjunto de competiciones es nuestro índice bursátil. Por tanto nuestra cartera de apuestas es un conjunto de apuestas que han sido seleccionadas a través de la función fitness del algoritmo genético y que son el mejor conjunto seleccionable según el criterio de la función de evaluación.

Antes de pasar a detallar el proceso de creación y las características del algoritmo genético, se va a hacer un alto para explicar cuál es la función de evaluación o fitness que se ha seleccionado para que evalúe las distintas carteras de apuestas que generará el algoritmo genético. La función escogida ha sido la que se muestra en la Figura 43.

$$Fitness = \sqrt{Cuota}/Riesgo$$

Figura 43. Función Fitness del Algoritmo Genético

El objetivo del algoritmo genético será maximizar el valor de la función fitness que ha sido definida. Actualmente sabemos tanto la cuota como el riesgo de una apuesta individual, pero hay que establecer cómo se puede calcular la cuota y el riesgo de las apuestas de la cartera cuando son combinadas entre ellas. Esta operación es muy sencilla, ya que la cuota de la cartera será la multiplicación de las cuotas de cada una de las apuestas individuales. Para el caso del riesgo, tendremos que el riesgo total de la cartera se calcula con la fórmula que aparece en la Figura 44.

$$Riesgo = (1 - \prod_{k=1}^{k=n} (1 - ProbAcierto_k))$$

Figura 44. Riesgo de la Cartera de Apuestas

En cuanto a la fórmula que refleja el cálculo del riesgo de la cartera de apuestas, la probabilidad de acierto de cada uno de los partidos ($ProbAcierto_k$), puede ser fácilmente calculada a partir del riesgo de cada partido individual, ya que la probabilidad de acierto del partido es (1 – Riesgo Partido).

En cuanto a las ventajas que podemos encontrar con sistema Lucky 15 y la selección de partidos a través del algoritmo genético, tenemos que al introducir en la apuesta partidos con cuotas interesantes, podamos permitirnos el lujo de fallar alguno éstos para obtener beneficios. Además, al ser un sistema tan popular para los apostantes, las casas de apuestas suelen ofrecer esta apuesta entre sus opciones de combinación, por lo que no es necesario que el usuario vaya combinando una a una todas ellas.

En el apartado de inconvenientes, ocurre lo mismo que en las apuestas de sistema, y es que al reducir el riesgo, reducimos la posible rentabilidad de ala apuesta. Además, si los partidos que incluimos en la apuesta tienen cuotas bajas, al fallar alguno de los encuentros perderemos gran parte de lo invertido. El otro inconveniente de este sistema es que al tener un total de quince apuestas se necesita una inversión mayor inicial, ya que en muchas ocasiones las casas de apuestas establecen un límite inferior al dinero que se puede apostar en una apuesta, que normalmente suele ser de 1€ o 2€.

5.3 Diseño e implementación del Algoritmo Genético

Llegados a este punto, vamos a hacer un alto en el camino, dejando de lado un momento las estrategias definidas para explicar en detalle el diseño e implementación del algoritmo genético, el cual nos permitirá generar carteras de apuestas cuya relación entre beneficio y riego es máxima.

Un Algoritmo Genético es un método de búsqueda que imita la Teoría de la Evolución Biológica de Darwin para la resolución de problemas [42]. El algoritmo que ha sido desarrollado en este proyecto ha sido implementado en Java a través de un conjunto de clases que recrean los principios fundamentales que todo algoritmo genético debe tener: **selección**, **cruce** y **mutación** de los individuos.

Antes de entrar en detalle con el diseño del algoritmo, se va a explicar en qué consisten los algoritmos genéticos y cuál es su estructura y funcionamiento.

5.3.1 Estructura y operadores de los Algoritmos Genéticos

Para la resolución de un determinado problema se ha de seguir un proceso con unas fases ya predefinidas (Figura 45).

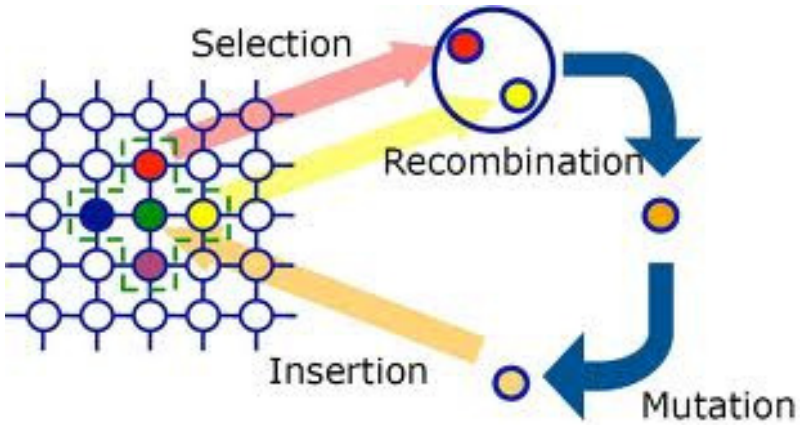

Figura 45. Fases Algoritmo Genético

En primer ligar partiremos de un **conjunto inicial** de individuos que normalmente son generados de manera aleatoria de acuerdo a la estructura definida para el individuo.

Tras generar el conjunto inicial se han de **evaluar** a dichos individuos a través de una función de evaluación, también conocida como función ***fitness*** Dicha función de evaluación es la que se querrá maximizar para lograr la resolución del problema.

Después de evaluar a todos los individuos se ha de **seleccionar** a aquellos individuos que se reproducirán para generar la siguiente generación. El método de selección tiene que tener en cuenta que aquellos individuos cuya evaluación es más alta (y por lo tanto se adaptan mejor a la solución del problema) tienen que tener más probabilidades de ser seleccionados.

El algoritmo debe implementar un **operador de cruce**, cuya función es la de generar nuevos individuos a partir de los que han sido seleccionados para la reproducción. Este operador de cruce partirá los individuos sobre un punto común determinado y recombinará cada una de las partes para generar nuevos individuos (véase la Figura 46).

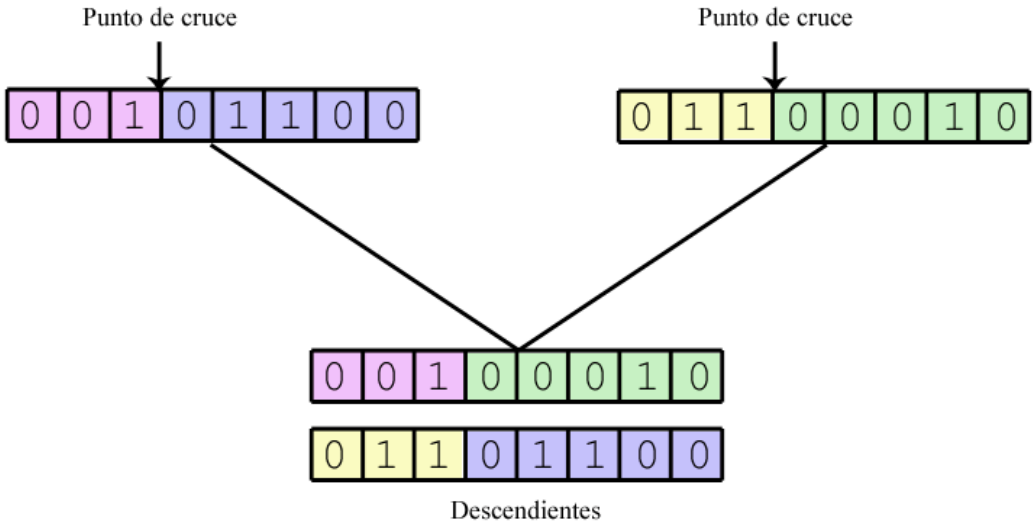

Figura 46. Operador de Cruce del Algoritmo Genético

Una vez se han generado los nuevos descendientes se ha de aplicar el **operador de mutación**. La función de este operador es la de cambiar alguno de los genes de los individuos para así agilizar el proceso de optimización de la función de evaluación. Es muy común que en las primeras generaciones que crea el algoritmo la probabilidad de mutación de los genes sea alta, ya que así se evita que la función de evaluación caiga en algún máximo local que le impida seguir evolucionando. En el caso de individuos con genes binarios, cuando un gen muta pasa de tener el valor 0 al 1 o viceversa.

Después de que a todos los descendientes se les haya aplicado el operador de mutación, el algoritmo vuelve a repetir los pasos anteriores para generar una nueva generación. En ocasiones, para mantener en la población el mejor de los individuos generados hasta el momento y evitar que sus características se pierdan a lo largo de las generaciones, se suele implementar una pequeña función que se encarga de que el mejor de los individuos siempre esté presente en la siguiente generación. A esta maniobra se la conoce como **elitismo**.

5.3.2 Algoritmo Genético de selección de apuestas en carteras

Una vez se ha explicado los objetivos generales y los operadores que conforman un algoritmo genético, se va a explicar de forma detallada el algoritmo desarrollado para este proyecto.

El algoritmo desarrollado en Java consta de cinco clases:

- **Apuesta.java:** Esta es la clase que representa al conjunto de apuestas que forma una cartera. Tiene un único atributo que es un array de bytes, el cual representa qué apuestas individuales del conjunto son las que forman la cartera. Cuando una posición de ese array está con el valor 1, significa que la apuesta asociada a esa posición está dentro de la cartera. En cambio, si esa posición tiene el valor 0, significa que la apuesta asociada a esa posición no está dentro de la cartera.

- **CarteraApuestas.java:** Esta clase es la encargada de gestionar todo el proceso de ejecución del algoritmo. Aquí se define el tamaño de la población, el número de generaciones a crear, el número de individuos que lucharán en el sistema de torneos, los puntos de cruce y la probabilidad de la mutación.

  Además, será esta clase la que se encargará de ir llamando a cada uno de los operadores para que vayan transformando la población inicial en las nuevas generaciones.

  Por último, también se encargará de evaluar la última generación para mostrar cuál es el conjunto de apuestas que mejor relación entre beneficios y riesgo ha conseguido generar. Este conjunto será el que se utilice en la fase de explotación del sistema de predicción para estudiar si genera beneficios o no.

- **EvaluarApuesta.java:** Esta clase contiene las funciones necesarias para el cálculo del beneficio y riesgo de cada una de las carteras de apuestas. Concretamente, la función llamada *evaluación* reproduce los cálculos mostrados en la Figura 44 para la evaluación de cada una de las carteras.

- **TorneoApuestas.java:** Esta clase es la encargada de implementar el torneo del algoritmo, que servirá para seleccionar a los individuos más válidos para la fase de reproducción o cruce. Las carteras de apuestas que se seleccionarán para que participen en el torneo serán elegidas de manera aleatoria, y sólo la mejor cartera (en términos de la función de evaluación) de cada uno de los torneos pasará a la fase de reproducción del algoritmo.

- **ReproducirApuestas.java:** Esta clase es la encargada de coger a los padres seleccionados a través de los torneos y cruzarlos para generar los nuevos descendientes. El punto de corte de cada par de individuos será elegido al azar al realizar cada uno de los cruces. Cuando se han generado todos los descendientes se les aplica el operador de mutación para que estén listos para ser el nuevo conjunto inicial de individuos de la siguiente generación.

Una vez introducidas las clases Java que conforman el algoritmo se va a profundizar en la explicación de cada uno de los operadores y parámetros de éste. Para que la explicación sea más sencilla, ésta va a comenzar con los parámetros fundamentales del algoritmo para después explicar cada una de las fases que podían verse en la Figura 45.

❖ Parámetros del Algoritmo:

A continuación se van a explicar los parámetros de funcionamiento del algoritmo. Es muy importante detallar el valor y la evolución de estos parámetros según pasan las generaciones creadas, ya que el objetivo es ir aumentando la presión selectiva según pasan las generaciones. Este aumento en la presión selectiva se ha querido implementar para que en las primeras generaciones del algoritmo éste realice una búsqueda a ciegas y no se quede estancado en máximos locales. Según vayan pasando las generaciones y la presión vaya aumentando, el algoritmo irá escogiendo las mejores carteras para cruzarlas y así obtener nuevas carteras con mejores cuotas y menos riesgo.

Los principales parámetros de funcionamiento que se controlan desde la clase *CarteraApuestas* y que son los que maneja el algoritmo genético en cada uno de los operadores son los siguientes:

- **num_padres:** Atributo que almacena el número de padres de cada una de las generaciones que se crearán. El valor de este atributo es **1000** y no variará durante la ejecución del algoritmo.

- **num_generaciones:** Atributo que almacena el número de generaciones que creará el algoritmo. El valor que toma este atributo es **120** y no variará durante la ejecución del algoritmo.

- **num_pelean:** Atributo que almacena el número de individuos que luchan en los torneos de selección de individuos que se generan antes de los cruces. El valor inicial de este atributo es **4**, pero el sistema incrementa este valor cada veinte generaciones añadiendo un luchador más. Este mecanismo hará que la presión selectiva sea menor en las primeras generaciones (los peores individuos tienen más posibilidades de reproducirse, por lo que sus características no se perderán tan rápidamente en las primeras generaciones).

- **mutacion:** Atributo que almacena la probabilidad de que un gen del individuo mute (uno de los bytes del array de la cartera cambia de valor). En la primera generación la probabilidad de mutar es del **15%**, pero el sistema actualiza esta probabilidad cada diez generaciones restando un 1% cada vez. Este sistema hace que en las primeras generaciones los cambios de genes sean más frecuentes y por lo tanto aparezcan y desaparezcan más espontáneamente apuestas en el conjunto de la cartera, evitando así que el algoritmo quede estancando en máximos locales.

- **num_partes:** Atributo que almacena el número de cortes que se realizan en los individuos al aplicarles el operador de cruce. El valor del atributo es 1 y no se modifica durante la ejecución del algoritmo.

Tras explicar los diferentes parámetros de funcionamiento del algoritmo, se va a pasar a detallar cada una de las fases de éste.

- **Generación de la Población Inicial:**

La población inicial que se ha decidido tener para la resolución de este problema es de mil individuos. Este número nos permitirá tener un conjunto de carteras lo suficientemente diversas como para que el algoritmo funcione correctamente y obtenga buenos resultados.

Antes de explicar cómo se genera la población inicial necesitamos explicar de dónde se toman las apuestas que van a ser analizadas por el algoritmo.

Cuando ejecutamos el algoritmo genético, éste lo primero que hace es buscar un fichero llamado Apuestas.csv que se tiene que encontrar en la misma carpeta que las clases Java. Este fichero contiene en su interior toda la información necesaria sobre las

apuestas que van a ser analizadas y es el que carga el algoritmo en una lista para poder operar con sus valores.

El fichero consta de cuatro columnas:

- Columna A: Columna en la que se almacenan los equipos que van a disputar el partido.

- Columna B: Columna en la que se almacena la cuota del resultado predicho por el sistema de predicción

- Columna C: Riesgo estimado por el sistema de predicción de apostar al resultado que ha sido predicho.

- Columna D: Resultado predicho por el sistema de predicción.

Al ejecutar el algoritmo, éste accede al fichero y almacena cada una de sus líneas en una lista (proceso realizado a través del método *leerArchivo* de la clase *CarteraApuestas*). Posteriormente, para separar cada una de los parámetros de esas líneas, se utiliza la función *listToArray* de la clase *CarteraApuestas*, que separa cada una de las columnas del fichero original y las guarda en un array bidimensional. Ese array bidimensional es el que manejará el algoritmo para realizar los cálculos de cuotas y riesgos de cada uno de los partidos. Es muy importante tener en cuenta que la posición de los partidos en este array es la misma que en el fichero original y que en el array de bytes que conforma una cartera de apuestas, por lo que si en el array de bytes de la cartera de apuestas la primera posición está a 1 significará que la primera apuesta del fichero está dentro de la cartera de apuestas.

Una vez explicado de dónde toma los datos de apuestas el algoritmo pasaremos a explicar la creación de la población inicial de individuos. La creación de este conjunto inicial se hace en la clase *CarteraApuestas* a través de la llamada al método *crearApuestas* de la clase *Apuestas*. Este método se encarga de rellenar el array de bytes de una cartera de apuestas con los valores 0 y 1, teniendo en cuenta que la probabilidad de que aparezca un 1 es del 10%. Esta operación se repite mil veces hasta generar los mil individuos de la población inicial.

En este momento de la ejecución disponemos de un conjunto de mil carteras de apuestas. Cada cartera de apuestas tiene un atributo que es un array de bytes y que representa qué apuestas individuales son tenidas en cuenta dentro de esa cartera. La longitud de ese array de bytes es la misma que el número de filas que tiene el fichero Apuestas.csv, ya que cada posición del array representa todas las apuestas que pueden ser tenidas en cuenta.

- **Selección de individuos para el cruce (Torneos):**

Una vez tenemos el conjunto inicial con los mil individuos, el siguiente paso a dar es la selección de los mejores individuos para que estos sean los padres de la siguiente generación. Para ello se ha decidido implementar un torneo en el que el número de individuos que luchan sea creciente para así ir aumentando la presión selectiva con el paso de las generaciones. Como ya se explicó en la sección de parámetros del algoritmo,

el número de individuos que pelean va aumentando en una unidad cada veinte generaciones.

La implementación del torneo se ha desarrollado de manera muy sencilla. En primer lugar se coge la lista de padres de la generación actual y se seleccionan al azar tantos padres como el valor que indique *num_pelean* en ese momento. Una vez se han seleccionado los individuos se evalúa como de buenos son utilizando la función fitness definida para este problema y que ha sido implementada en el método *evaluación* de la clase *EvaluarApuesta*. El individuo que tenga mejor puntuación, será el que se añadirá a la lista de individuos que se reproducirán en la siguiente fase.

Para completar esta fase el algoritmo genera mil torneos de los que saldrán los mil padres de los individuos de la siguiente generación.

- **Operador de Cruce:**

Llegados a este punto tenemos un conjunto de mil padres que han sido seleccionados a través de los torneos generados por el algoritmo. El siguiente paso es ir cogiendo a dichos padres de dos en dos para cruzarlos y generar dos nuevos descendientes.

Para escoger a un par de padres, el algoritmo genera dos números aleatorios y toma las dos carteras de apuestas que hay en esas posiciones de la lista. Acto seguido vuelve a generar un número aleatorio para situar el punto de corte de los individuos y proceder al cruce de las partes para generar dos nuevos descendientes. Este proceso se ha implementado en el método *cruzar* de la clase *ReproducirApuestas*.

Los descendientes que se van generando se van incorporando a una nueva lista que será sobre la que se aplicará el operador de mutación que dará fin al proceso de creación de una nueva generación.

- **Mutación:**

Para concluir con la creación de una generación se aplicará sobre los descendientes generados el operador de mutación que ha sido implementado en el método *mutar* de la clase *ReproducirApuestas*. Este método coge cada uno de los descendientes y sobre su array de bytes que tienen como atributo y que representa las apuestas que hay en esa cartera van modificando o no el byte de cada posición. Para realizar la mutación se genera un número aleatorio, y si éste está por debajo de la mutación almacenada en la variable *mutacion*, el gen que se está analizando en ese momento cambiará su valor de 0 a 1 o viceversa.

El algoritmo recorre todos los arrays de bytes de los mil descendientes generados para ir aplicando la mutación de sus genes cuando el sistema así lo determina. Cuando termina con esta operación tendremos disponible una nueva generación.

- **Elitismo:**

Para que no se pierda el mejor resultado obtenido a lo largo de todas las generaciones, se ha implementado en el código un mecanismo por el cual después de la mutación se elimina la peor de las carteras de apuestas que hay en el conjunto actual y se

inserta la mejor cartera hasta el momento conseguida. De esta manera nos aseguramos que cuando lleguemos a la última generación vamos a tener siempre el mejor resultado de los obtenidos durante toda la ejecución del algoritmo.

Este mecanismo ha sido implementado en la clase *CarteraApuestas* dentro del método main y podrá ser encontrado fácilmente en el código a través de los comentarios incluidos en él.

5.3.3 Salida del Algoritmo Genético

Tras la ejecución del algoritmo éste nos ofrecerá una salida por pantalla en la que se mostrarán los partidos seleccionados con el resultado de la predicción y la cuota individual de cada uno de ellos ofrecida por las casas de apuestas. Adicionalmente se mostrará la cuota resultante de combinar todos los partidos en una apuesta combinada.

Este resultado que se muestra por pantalla y del que se muestra un ejemplo en la Figura 47, será el que se utilizará en los estudios de beneficios con las estrategias mencionadas en apartados anteriores.

```
Apuestas seleccionadas:

Partido: Celta - R.Madrid || Resultado: 2 || Cuota: 1.4
Partido: Sevilla - Barcelona || Resultado: 2 || Cuota: 1.6
Partido: Valladolid - Atletico || Resultado: 2 || Cuota: 1.85
Partido: Atletico - R.Sociedad || Resultado: 1 || Cuota: 1.5
Partido: Zaragoza - Osasuna || Resultado: 1 || Cuota: 2.15

Cuota combinada: 13.36
```

Figura 47. Salida del Algoritmo Genético

5.4 Resultados obtenidos tras la aplicación de estrategias de apuesta

Tras las secciones en las que se han explicado todas las estrategias que se van a aplicar para aprovechar las predicciones de nuestro sistema de predicción de resultados, se van a mostrar a continuación los datos de beneficios que se han obtenido combinando las predicciones del sistema con las estrategias anteriormente explicadas.

A continuación se muestran varios cuadros con el resumen de beneficios y pérdidas de cada una de las estrategias durante las doce semanas del estudio. Si se quiere entrar más en detalle y ver los resultados de beneficios para cada una de las competiciones o partidos individuales, se recomienda consultar los ficheros adjuntos que se encuentran en la ruta de carpetas *Archivos PFC / Resultados* donde se encontrará la información más detallada.

5.4.1 Estudio de Beneficios

Tras la explicación de cada una de las seis estrategias que se van a aplicar a los resultados obtenidos, se ha realizado un estudio de beneficios en el que se comprobó si cada una de las estrategias era capaz de generar beneficios al apostar en cada una de las competiciones.

Debido al gran volumen de datos y tablas que se ha generado para calcular los beneficios de cada estrategia en cada competición, se va a mostrar en esta sección a modo de resumen los datos finales obtenidos. A continuación se mostrarán dos tablas. La primera de ellas mostrará la relación de beneficios y rentabilidad para cada estrategia en cada competición (Tabla 30). Como dos de las estrategias combinan apuestas de diversas competiciones, se mostrará en una tabla a parte (Tabla 31) los resultados obtenidos semanalmente para estas dos estrategias (CV y SG15).

En la Tabla 30 se muestran los beneficios y pérdidas obtenidos al aplicar las estrategias PIAS, PIDO, CC y ASFR a los resultados predichos por el sistema de predicción. Lo más destacado del estudio de rentabilidad realizado es que en tres de las cuatros estrategias se han conseguido obtener beneficios, y en dos de ellas la rentabilidad se sitúa por encima del 15% sobre el total apostado.

De entre todas las competiciones analizadas sólo la Premier League inglesa ha conseguido sacar beneficios con las cuatro estrategias utilizadas. Además, estos beneficios han sido realmente buenos para esta competición, debido a las buenas cuotas que se ofrecían para estos partidos en las casas de apuestas.

Por otro lado tenemos que la competición de la Europa League ha sido la única competición que no ha generado ningún beneficio con las estrategias de apuestas aplicadas. Esto es debido a la baja tasa de aciertos que se ha tenido en esta competición, que unido a bajas cuotas para partidos con claros favoritos han hecho que fuera muy difícil generar algún tipo de beneficio.

De las cuatro estrategias analizadas podemos destacar dos: CC y ASFR. Si recordamos lo explicado en las secciones anteriores, la estrategia CC escoge los tres partidos con menor riesgo de una competición y los combina en una única apuesta de 1€. Esta operación se repitió cada semana en cada una de las competiciones y ofreció una rentabilidad al finalizar el estudio del 19,17%. Estos datos tan buenos fueron obtenidos principalmente por la alta tasa de aciertos obtenida en la NBA y las buenas cuotas ofrecidas para la Ligue 1 francesa, que a pesar de no tener buenas tasas de acierto ha conseguido generar buena parte de los beneficios obtenidos con esta estrategia.

En cuanto a la estrategia ASFR, en la que se realizaba una apuesta de sistema de 1€ por cada semana y competición analizada, se ha obtenido un 16,66% de rentabilidad. Si tenemos en cuenta que las cuotas de las apuestas de sistema son menores que las apuestas combinadas, tenemos que el dato del 16,66% es realmente bueno, ya que se ha conseguido una rentabilidad similar a la de la estrategia CC con menos riesgo que el que se ha tomado en la otra estrategia.

	PIAS	PIDO	CC	ASFR
Champions League	-1,56 €	0,12 €	-1,87 €	1,11 €
Europa League	-0,27 €	-0,80 €	-3,52 €	-0,38 €
Internacional	-4,65 €	-2,80 €	5,41 €	2,65 €
Liga BBVA	8,91 €	0,84 €	2,79 €	-0,03 €
Liga Adelante	6,08 €	-2,76 €	0,75 €	2,92 €
Ligue 1	-10,24 €	-7,49 €	9,43 €	3,54 €
Premier League	18,97 €	4,06 €	4,05 €	3,35 €
Serie A	3,70 €	-1,84 €	-4,37 €	-0,55 €
Otras Ligas	0,29 €	-1,63 €	2,07 €	2,53 €
NBA	-3,29 €	-	8,26 €	4,85 €
TOTAL	17,94 €	-12,30 €	23,00 €	19,99 €
Total Apostado	935 €	695 €	120 €	120 €
Rentabilidad	1,92%	-1,77%	19,17%	16,66%

Tabla 30. Estudio de Rentabilidad de las Estrategias (I)

Por último se va a analizar la Tabla 31 en la que aparecen los datos del estudio de beneficios realizado para las estrategias CV y SG15.

En primer lugar se van a analizar los datos de la estrategia CV, que si recordamos su objetivo, trata de formar tres apuestas combinadas de tres partidos de cualquiera de las competiciones. El importe apostado cada semana es de 3€ (1€ por cada una de las apuestas combinadas seleccionadas). Las nueve apuestas seleccionadas son las que menos riesgo tienen según el sistema de predicción. La rentabilidad obtenida con esta estrategia alcanza el 36,39%. Este dato tan alto de rentabilidad se puede explicar gracias a la alta tasa de aciertos que se ha conseguido con este sistema (gracias sobre todo al acierto en partidos de la NBA). Además, los fallos en la predicción que han producido pérdidas han sido bien contrarrestados con los aciertos y beneficios obtenidos con las combinadas acertadas.

Si los datos de rentabilidad eran buenos para la estrategia CV, los obtenidos para la estrategia SG15 son aún mejores. Si recordamos, esta estrategia selecciona cuatro partidos a través de la ejecución del algoritmo genético desarrollado en este proyecto y los combina a través de una apuesta Lucky 15 (una combinada de cuatro partidos, cuatro combinadas de tres partidos, seis combinadas de dos partidos y cuatro apuestas simples). La rentabilidad obtenida con esta estrategia ha sido del 56,72%.

Las claves de la obtención de esta rentabilidad con la estrategia SG15 residen en dos puntos. En primer lugar tenemos que hablar de la apuesta Lucky 15. Esta combinación de apuestas hace que aunque se produzca algún fallo en uno de los cuatro pronósticos sea posible conseguir beneficio, aunque este sea pequeño. Para que esto ocurra se deben tener partidos con cuotas aceptables, y es ahí donde entra en juego el segundo punto clave de esta buena rentabilidad: la selección del algoritmo genético.

Si prestamos atención a las apuestas seleccionadas por el algoritmo genético en las doce semanas estudiadas, las cuotas que posee cada apuesta son realmente buenas para el

riesgo tan bajo que suelen tener. Estas cuotas tan buenas hacen que al fallar alguno de los cuatro pronósticos se sigan obteniendo pequeños beneficios.

En el caso de acertar los cuatro partidos, la rentabilidad de esa semana se dispara más allá del 100%, lo que genera una buena cantidad de beneficios capaz de tapar posibles pérdidas.

	CV	SG15
Semana 1	-1,30 €	-2,03 €
Semana 2	-1,32 €	3,07 €
Semana 3	-3,00 €	-0,72 €
Semana 4	2,37 €	6,15 €
Semana 5	0,23 €	-3,00 €
Semana 6	0,08 €	0,92 €
Semana 7	0,09 €	-0,03 €
Semana 8	3,38 €	4,54 €
Semana 9	0,36 €	2,48 €
Semana 10	6,46 €	5,36 €
Semana 11	2,08 €	0,44 €
Semana 12	3,67 €	3,24 €
TOTAL	13,10 €	20,42 €
Total Apostado	36 €	36 €
Rentabilidad	36,39%	56,72%

Tabla 31. Estudio de Rentabilidad de las Estrategias (II)

5.5 Conclusiones

Tras realizar el análisis de rentabilidad de cada una de las estrategias utilizadas, podemos resaltar varios puntos clave, que son los pilares en los que se sientan las bases de los buenos resultados obtenidos en el proyecto.

Si repasamos por orden cronológico, el primero de los pilares clave en este proyecto ha sido el sistema de predicción de resultados implementado en la hoja Excel. A partir de las predicciones del sistema y de su estimación de riesgo en cada una de las apuestas hemos sido capaces de elegir partidos con un riesgo bajo y con una cuota interesante que nos ha permitido obtener beneficios.

En segundo lugar hay que destacar las estrategias puestas en marcha para la obtención de beneficios en las casas de apuestas. Tanto las estrategias que suelen ser utilizadas por los apostantes, como la creada ad-hoc para este proyecto (Lucky 15 con selección por algoritmo genético) han generado resultados realmente buenos. En el caso de la estrategia que combinaba la apuesta Lucky 15 con la selección por algoritmo genético, los resultados destacan por su alta rentabilidad con un riesgo bajo, ya que el

algoritmo genético se ha encargado de seleccionar las apuestas con una tasa de riesgo baja en comparación con la cuota que ofrecía la casa de apuestas.

Capítulo 6

CONCLUSIONES Y TRABAJO FUTURO

6.1 Introducción

Tras exponer los principales aspectos del proyecto y presentar los resultados obtenidos, es el momento de recopilar y analizar las conclusiones a las que se ha llegado.

Además, de cara a evolucionar el sistema mejorando las características actuales e implementando otras nuevas, se va a exponer en este mismo apartado los trabajos futuros que se quieren llevar a cabo.

Este trabajo futuro engloba varias áreas de mejora dentro del proyecto. La primera de ellas se centrará en la mejora del sistema de predicción, ya que como se ha podido observar en el desarrollo del proyecto, el sistema tiene varios puntos débiles como la dificultad para predecir empates. Se intentará que esos puntos débiles dejen de serlo y proporcionen mejores resultados al sistema.

Además, se tratará de evolucionar la estructura del proyecto añadiendo nuevos deportes y estrategias a la oferta ya existente. Esta ampliación de la oferta de deportes y estrategias proporcionará una reducción en el riesgo que se toma al realizar las apuestas, ya que al tener un espacio más grande de apuestas donde elegir, tendremos más opciones de escoger apuestas seguras para nuestras combinaciones.

Por último, se explicarán los proyectos de difusión que se han ideado para la herramienta. Estos proyectos tendrán como función principal poner a disposición de los usuarios las herramientas necesarias para que puedan utilizar el sistema desde ordenadores o dispositivos móviles.

6.2 Conclusiones Finales del Proyecto

Para sacar las conclusiones finale, tenemos que echar la vista atrás y ver qué objetivos habían sido fijados al iniciar el proyecto. Dos grandes objetivos eran la base de este proyecto: **el desarrollo de un sistema de predicción de resultados en eventos deportivos** y **la generación de estrategias para la explotación del sistema de predicción**.

En cuanto al primero de los dos objetivos, podemos observar que el sistema de predicción desarrollado es un **sistema sencillo** capaz de realizar predicciones sobre los partidos con una **tasa de acierto muy buena**. Como pudo verse en la fase de pruebas de los algoritmos de clasificación, la combinación de los algoritmos ha hecho que el sistema logre **tasas de acierto muy por encima** de la tasa que se conseguiría por puro **azar**. El caso más destacado de entre todas las copeticiones analizadas ha sido la **Champions League**, donde el sistema ha logrado un **71,88%** de aciertos en **predicción simple** y un **93,75%** de aciertos en la modalidad de **Doble Oportunidad**.

Aunque el elemento más destacado del sistema de predicción es la predicción que éste calcula, no podemos olvidar el otro elemento clave de este sistema, que le da un valor añadido a la hora de seleccionar qué predicciones son las más seguras. Este elemento es el **riesgo calculado** por el sistema. Como también se ha podido observar en la fase de estudio de beneficios, aquellas **estrategias que tenían en cuenta el riesgo** de las apuestas **han conseguido muy buenos resultados** derivados de la buena estimación del riesgo que realiza el sistema sistema.

La elección de **Microsoft Excel** como soporte para el sistema de predicción ha sido un acierto, ya que además de aportar **sencillez** a la introducción de datos y a la implementación de los clasificadores, ha permitido **reducir tiempos** en el análisis de los resultados, ya que a partir de la hoja de predicción se han podido analizar tanto las tasas de aciertos como los beneficios generados por cada una de las estrategias utilizadas.

Por otro lado, y como **punto a mejorar**, habrá que buscar una solución a la dificultad que tiene el sistema de predicción para **predecir los empates**. Este resultado, al ser el que ocurre con menos frecuencia tiene poca presencia en los conjuntos de entrenamiento, que en ocasiones son muy pequeños. Por ello, para intentar solucionar este hecho, habría que volver a realizar una recogida de datos que amplíe los conjuntos de

entrenamiento para después volver a generar los clasificadores. Al tener un conjunto de entrenamiento más amplio hay más posibilidades de que este resultado sea más predecible por cada uno de los clasificadores.

En cuanto al segundo de los grandes objetivos marcados para este proyecto, la generación de estrategias para explotar los resultados que nos ofrece el sistema de predicción, hemos podido ver que la variedad de las estrategias utilidas ha permitido sacar el máximo partido a las características del sistema.

En concreto, las estrategias **CV** (Combinada Variada) y **SG15** (selección por algoritmo genético y apuesta Lucky 15) han obtenido **grandes rentabilidades** al haber sido aplicadas a los resultados calculados por el sistema de predicción. En concreto la estrategia **SG15** ha conseguido una rentabilidad en 3 meses del **56,72%**, generando **20,42€** a partir de **36€** apostados.

Clave en los buenos resultados de la estrategia **SG15** ha sido el **algoritmo genético** desarrollado para el proyecto. Su **flexibilidad** ha sido muy importante en la fase de pruebas del proyecto, ya que a través de la parametrización del algoritmo se ha podido ajustar la selección de apuestas realizada por éste a las necesidades de las estrategias a aplicar. Esta flexibilidad permite que si se necesita realizar cualquier cambio en las características del algoritmo debido a la utilización de nuevas estrategias, éstos sean fáciles y rápidos de implementar.

Las selecciones de apuestas que realiza el algoritmo genético son muy buenas, ya que **escoge apuestas con buenas cuotas y con un riesgo bajo** en relación con la cuota que se nos está ofreciendo. De cualquier forma, la **función fitness** implementada en el algoritmo hace que siempre se ponga por delante un riesgo bajo a una buena cuota, ya que nuestra máxima es **buscar la seguridad antes que los beneficios**.

6.3 Evolución del Sistema de Predicción

El primero de los puntos a mejorar es el sistema de predicción. Si recordamos los problemas que han surgido a lo largo de la realización del proyecto, uno de los más importantes era la imposibilidad del sistema para predecir empates. Esta deficiencia en el sistema nos ha hecho tener que realizar ajustes manuales en las apuestas realizadas. Estos ajustes han conducido a la utilización de sistemas de doble oportunidad en las estrategias utilizadas para intentar cubrir los resultados de empate y evitar posibles pérdidas derivadas de esta deficiencia del sistema.

Durante la explicación del desarrollo del sistema se detectó el fallo que hacía que el resultado de empate no fuera fácilmente detectable. En las competiciones futbolísticas, alrededor del 20% de los partidos acaban en empate. Al ser el resultado que menos se repite en cada una de las competiciones, es muy difícil que en conjuntos de entrenamiento pequeños seamos capaces de sacar reglas que sean capaces de predecir cuándo un partido va a acabar en empate.

La solución a este problema es muy clara y pasa por seguir recogiendo datos para incorporarlos al conjunto de entrenamiento y actualizar tanto los árboles de decisión como las redes bayesianas creadas para que las predicciones que realice el sistema obtengan unas tasas de acierto mayores que las que se han obtenido en esta primera versión del sistema.

6.4 Incorporación de nuevas competiciones

Otra de las mejoras que se pretende añadir al sistema, es el aumento de las competiciones tanto de deportes que ya han sido estudiados en la primera versión del sistema como en nuevos deportes que puedan aportar apuestas de valor para nuestros sistemas.

Entre los deportes que se desean añadir a la nueva versión del sistema destaca el tenis. Este deporte es junto al fútbol y el baloncesto el deporte estrella de las casas de apuestas. El gran número de partidos y torneos disponibles para apostar en las casas de apuestas hacen de este deporte el candidato principal a ser añadido al sistema de predicción. Además del gran número de apuestas que podemos encontrar en las casas de apuestas, también disponemos de mucha información que puede ser utilizada para la creación de los modelos de predicción de este deporte. Sistemas como *OnCourt* [43], ofrecen gran cantidad de estadísticas de los jugadores del circuito profesional que pueden ser tenidas en cuenta para tomar las predicciones.

En cuanto a competiciones de deportes que ya han sido analizados se baraja la posibilidad de añadir más competiciones baloncestísticas, ya que como se ha podido ver con la NBA, se pueden obtener muy buenas cuotas con un riesgo relativamente bajo. Competiciones como la Liga ACB o la Euroliga serán tenidas en cuenta en la siguiente versión del sistema de predicción.

6.5 Difusión del sistema

Una vez desarrollado el sistema y tras haber realizado el estudio de beneficios que demuestra que el sistema es capaz de generar dinero, el objetivo es que los aficionados a las apuestas puedan aprovecharse de las predicciones y estrategias que tan buen resultado han dado en la fase de pruebas.

Se tiene pensado que la utilización del sistema pueda ser llevada a cabo bajo dos vertientes:

- Aplicación para dispositivos móviles.
- Blog de apuestas.

Al realizar el estudio del estado del arte, pudimos darnos cuenta de que las dos modalidades que más utilizaban los aficionados a las apuestas a la hora de buscar consejo

para efectuar sus apuestas eran las aplicaciones para dispositivos móviles y los blogs especializados en apuestas deportivas. Es por ello que, viendo el éxito de estas dos modalidades se haya pensado en ellas para que los usuarios se beneficien de las predicciones del sistema.

En cuanto a la aplicación para dispositivos móviles, el objetivo sería diseñar una aplicación que reproduzca las reglas implementadas en el sistema de predicción creado sobre la hoja Excel para que un usuario pueda conocer el resultado más probable de un partido. Esto permitiría al usuario utilizar la aplicación en los locales de casas de apuestas, donde introduciendo las cuotas que ofrece la casa y algunos datos como la clasificación del equipo, sería capaz de conocer el pronóstico del sistema y ver si le conviene o no apostar a ese encuentro.

En cuanto a la opción del Blog de Apuestas, es una opción más sencilla y rápida de implementar. Además, dado el grado de avance de los nuevos dispositivos móviles, este blog también podría consultarse desde este tipo de dispositivos, por lo que el usuario podría consultar las apuestas recomendadas en cualquier momento. La desventaja que tiene el blog respecto a la aplicación para dispositivos móviles es que puede que haya apuestas que no estén reflejadas en el blog y para las que el usuario quiera conocer una predicción y una estimación del riesgo. Es por ello que se considera la opción de la aplicación como la más atractiva aunque sea más costosa en términos de tiempo de desarrollo.

ANEXO

7.1 ANEXO A: Estudio de Relevancia de Atributos

7.1.1 Análisis de relevancia de atributos para el conjunto de Champions League:

Como podemos observar en la Tabla 32, se han podido descartar cuatro atributos debido a la correlación nula que tienen con la clase por la que vamos a clasificar. Estos cuatro atributos son la cuota de empate de la casa de apuestas, las rachas de resultados numéricas del equipo local y visitante y el Coeficiente UEFA del equipo visitante.

	ChiSquared	GainRatio
Cuota Local	30,9485	0,1913
Cuota Visitante	28,6854	0,1890
TOP Local	22,8505	0,0679
Coeficiente UEFA Local	16,5876	0,1133
TOP Visitante	15,4346	0,0493
Racha Discreta Local	4,6050	0,0165
Racha Discreta Visitante	4,0627	0,0155
Tipo de Partido	1,9971	0,0115
Tiempo	0,0443	0,0002
Cuota Empate	0	0
Racha Visitante	0	0
Coeficiente UEFA Visitante	0	0
Racha Local	0	0

Tabla 32. Análisis de Relevancia de Atributos (Conjunto Champions League)

El análisis nos desvela también que otros cuatro atributos presentan una correlación muy baja. De momento se mantendrán dichos atributos en el modelo, pero no se descarta eliminarlos en fases posteriores del análisis si se detecta que no aportan nada al modelo

de predicción. Estos cuatro atributos son las rachas discretizadas del equipo local y visitante, las condiciones meteorológicas en las que se disputará un partido y el tipo de partido que se disputa (liguilla o eliminatoria).

7.1.2 Análisis de relevancia de atributos para el conjunto de Europa League:

Los datos que arroja el estudio de relevancia de atributos sobre el conjunto de partidos de la Europa League son bastante similares a los hechos sobre el conjunto de Champions League (Tabla 33). En este caso, los atributos descartados son exactamente los mismos, a excepción de la Cuota de Empate recogida de la casa de apuestas, que para el caso de este conjunto ha mostrado una correlación con la clase interesante como para tenerla en cuenta en las próximas fases del análisis.

	ChiSquared	GainRatio
Cuota Visitante	46,823	0,164
Cuota Local	39,449	0,158
Coeficiente UEFA Local	32,218	0,107
TOP Local	29,779	0,048
Cuota Empate	22,919	0,080
Racha Discreta Visitante	14,690	0,033
TOP Visitante	11,659	0,018
Racha Discreta Local	7,155	0,014
Tiempo	2,332	0,007
Tipo de Partido	1,812	0,006
Racha Local	0	0
Coeficiente UEFA Visitante	0	0
Racha Visitante	0	0

Tabla 33. Análisis de relevancia de atributos (Conjunto Europa League)

7.1.3 Análisis de relevancia de atributos para Competiciones Europeas:

Tras realizar el análisis de relevancia del conjunto de datos que une los partidos de Champions League y Europa League, se puede observar en la Tabla 34 que los resultados obtenidos son muy similares a los de los conjuntos por separado. Para este caso, se descartarán los datos de Racha de resultados numérica y el Coeficiente UEFA del equipo visitante debido a que no tienen correlación alguna con la clase por la que vamos a clasificar. Estos resultados son exactamente los mismos que se obtuvieron al analizar el conjunto de entrenamiento de la Europa League, lo que hace ver que al unir los conjuntos

la correlación de los atributos ha sufrido variaciones mínimas. Esto hará que dispongamos de un conjunto de entrenamiento mayor, lo que implica que los clasificadores que probemos estarán entrenados por un conjunto de instancias mayor.

	ChiSquared	GainRatio
Cuota Visitante	46,823	0,164
Cuota Local	39,449	0,158
Coeficiente UEFA Local	32,218	0,107
TOP Local	29,779	0,048
Cuota Empate	22,919	0,080
Racha Discreta Visitante	14,690	0,033
TOP Visitante	11,659	0,018
Racha Discreta Local	7,155	0,014
Tiempo	2,332	0,007
Tipo de Partido	1,812	0,006
Racha Visitante	0	0
Coeficiente UEFA Visitante	0	0
Racha Local	0	0

Tabla 34. Análisis de relevancia de atributos (Conjunto Competiciones Europeas)

7.1.4 Análisis de relevancia de atributos para el conjunto de Liga BBVA:

Después de haber analizado la relevancia de los atributos de los conjuntos correspondientes a partidos de competiciones europeas, se va a pasar a analizar cada uno de los conjuntos referentes a competiciones de liga.

En este primer caso se muestra el análisis de relevancia de atributos para el conjunto de atributos de la Liga BBVA. Como se puede ver en la Tabla 35, cuatro atributos van a ser eliminados en la siguiente fase de análisis debido a la correlación nula que guardan con la clase. Estos cuatro atributos son la Cuota de Empate que ofrece la casa de apuestas, los goles que recibe de media tanto el equipo local como el visitante y la Racha de resultados numérica que tiene el equipo local.

	ChiSquared	GainRatio
Cuota Visitante	78,331	0,159
Cuota Local	69,335	0,205
Goles Anotados Visitante	47,164	0,227
Zona Visitante	31,195	0,020
Clasificación Visitante	26,803	0,086
Goles Anotados Local	22,350	0,131
Racha Visitante	20,823	0,097
Clasificación Local	20,419	0,124
Zona Local	17,837	0,017
Racha Discreta Visitante	13,883	0,019
Tiempo	7,335	0,018
Racha Discreta Local	7,143	0,010
Derbi	4,565	0,027
Jugó entre semana Visitante	1,310	0,004
Jugó entre semana Local	0,811	0,002
Cuota Empate	0	0
Goles recibidos Local	0	0
Goles recibidos Visitante	0	0
Racha Local	0	0

Tabla 35. Análisis de relevancia de atributos (Conjunto Liga BBVA)

Otro de los datos que arroja el análisis es que tenemos varios atributos que tienen una mínima correlación con la clase. Estos atributos, entre los que se encuentra el dato de si un equipo ha disputado partidos entre semana, de momento van a ser considerados para las próximas fases del análisis, pero no se descarta que se prescinda de ellos si no aportan valor a las predicciones del sistema.

7.1.5 Análisis de relevancia de atributos para el conjunto de Liga Adelante:

La Tabla 36 muestra el análisis de relevancia de atributos para el conjunto de partidos de la Liga Adelante. Como puede observarse, los resultados no son todo lo buenos que se esperaban, ya que una gran cantidad de atributos no guardan correlación con la clase, mientras que el resto, en su mayoría no guardan una correlación muy alta.

	ChiSquared	GainRatio
Zona Visitante	20,202	0,028
Cuota Empate	13,491	0,185
Cuota Local	13,491	0,185
Racha Discreta Local	9,728	0,021
Zona Local	6,300	0,009
Jugó entre semana Local	3,875	0,041
Racha discreta Visitante	2,687	0,005
Jugó entre semana Visitante	2,202	0,012
Tiempo	1,885	0,016
Derbi	0,908	0,013
Cuota Visitante	0	0
Clasificación Local	0	0
Clasificación Visitante	0	0
Goles anotados Local	0	0
Goles anotados Visitante	0	0
Goles recibidos Visitante	0	0
Goles recibidos Local	0	0
Racha Visitante	0	0
Racha Local	0	0

Tabla 36. Análisis de relevancia de atributos (Conjunto Liga Adelante)

Las conclusiones que podemos sacar tras realizar este análisis es que los clasificadores que se van a probar a continuación con este conjunto, no van a realizar buenas predicciones. Se esperará al entrenamiento con los clasificadores para tomar alguna decisión en el caso de no obtener buenos resultados.

Por otro lado, los atributos que serán descartados para la siguiente fase de entrenamiento de los conjuntos serán la cuota del equipo visitante que ofrece la casa de apuestas, las clasificaciones en la liga del equipo local y visitante, la media de goles anotados y recibidos por ambos equipos y las rachas de resultados en formato numérico de ambos equipos.

7.1.6 Análisis de relevancia de atributos para el conjunto de Ligue 1:

Tras realizar el análisis de relevancia para el conjunto de partidos de la liga francesa de fútbol (Ligue 1), se puede observar en la Tabla 37 que al igual que ha ocurrido en el caso de la Liga Adelante, una gran cantidad de atributos no guardan relación con la clase por la que vamos a clasificar. Estos atributos, que serán descartados para la siguiente fase de entrenamiento de los clasificadores son las clasificaciones en el campeonato de liga de

los equipos local y visitante, la media de goles anotados y recibidos por ambos equipos y la racha de resultados en formato numérico también para ambos equipos.

	ChiSquared	GainRatio
Cuota Visitante	53,268	0,213
Cuota Local	47,525	0,196
Zona Visitante	18,867	0,026
Zona Local	16,162	0,023
Cuota Empate	16,013	0,233
Racha Discreta Visitante	9,847	0,029
Tiempo	6,119	0,026
Jugó entre semana Local	6,104	0,305
Racha Discreta Local	2,217	0,004
Jugó entre semana Visitante	1,207	0,004
Derbi	0,893	0,010
Clasificación Visitante	0	0
Clasificación Local	0	0
Goles Anotados Visitante	0	0
Goles anotados Local	0	0
Goles recibidos Visitante	0	0
Goles recibidos Local	0	0
Racha Visitante	0	0
Racha Local	0	0

Tabla 37. Análisis de relevancia de atributos (Conjunto Ligue 1)

A diferencia de los resultados que se han podido observar en el conjunto de la Liga Adelante, en este análisis encontramos atributos que tienen una correlación con la clase bastante buena. Estos atributos que destacan por su correlación son las cuotas del equipo local y visitante que ofrece la casa de apuestas.

En la siguiente fase, donde se pondrá en marcha el entrenamiento de los conjuntos a través de los diferentes clasificadores que se seleccionarán para los experimentos, se puede asegurar que las dos cuotas mencionadas anteriormente formarán parte de los datos necesarios para que los clasificadores elaboren sus predicciones. Estas dos cuotas se combinarán con el resto de atributos relevantes para intentar obtener el mayor porcentaje de aciertos posible.

7.1.7 Análisis de relevancia de atributos para el conjunto de Premier League:

El siguiente conjunto en ser analizado es el de los partidos de la Premier League Inglesa. Como viene pasando con el resto de conjuntos de entrenamiento de las competiciones de liga, la Tabla 38 muestra que tenemos un gran número de atributos que van a ser descartados para la siguiente fase debido a la baja correlación con la clase. Los atributos que muestran una correlación nula para los dos test son la cuota de empate que ofrece la casa de apuestas, la clasificación en liga del equipo local, la media de goles anotados por ambos equipos, la media de goles recibidos por el equipo local y las rachas en formato numérico de los dos equipos.

	ChiSquared	GainRatio
Cuota Local	39,364	0,208
Cuota Visitante	38,297	0,205
Zona Visitante	33,002	0,065
Goles recibidos visitante	22,936	0,129
Clasificación Visitante	22,197	0,150
Zona Local	15,292	0,028
Jugó entre semana Local	11,031	0,058
Racha Discreta Local	10,298	0,039
Racha Discreta Visitante	9,148	0,036
Derbi	0,795	0,005
Tiempo	0,196	0,001
Jugó entre semana Visitante	0,005	0
Cuota Empate	0	0
Clasificación Local	0	0
Goles anotados local	0	0
Goles recibidos Local	0	0
Goles anotados Visitante	0	0
Racha Visitante	0	0
Racha Local	0	0

Tabla 38. Análisis de relevancia de atributos (Conjunto Premier League)

En este conjunto de datos también tenemos un caso especial en el que el atributo que refleja si el equipo visitante jugó un partido entre semana muestra una correlación nula para el test GainRatio, mientras que en el test ChiSquared muestra una mínima correlación. Para este caso, al ser el valor del test ChiSquared muy próximo a cero, se eliminará también este atributo para la siguiente fase.

En cuanto al resto de atributos, se puede observar que el conjunto posee una serie de atributos que muestran una correlación bastante interesante con la clase, hecho que puede propiciar buenos resultados al entrenar este conjunto con los algoritmos que se utilizarán

en las fases posteriores. Muy a tener en cuenta debido a su alta correlación son los valores de las cuotas de victoria del equipo local y visitante que ofrece la casa de apuestas.

7.1.8 Análisis de relevancia de atributos para el conjunto de la Serie A:

Tras analizar el conjunto de datos correspondiente a la Serie A italiana, se puede observar en la Tabla 39 que con diferencia éste es el conjunto con el que obtenemos peores resultados en la relevancia de atributos. No sólo hay una gran cantidad de atributos que no son relevantes para nuestro problema, sino que hay varios atributos que en el resto de los grupos estudiados eran muy relevantes y que en este caso no lo son. Ese es el caso de las cuotas de las casas de apuestas, que para este caso no muestran ninguna correlación con la clase.

	ChiSquared	GainRatio
Zona Visitante	23,913	0,042
Zona Local	12,453	0,029
Racha Discreta Visitante	4,421	0,013
Racha Discreta Local	4,214	0,010
Tiempo	3,280	0,016
Derbi	1,758	0,029
Jugó entre semana Local	0,972	0,004
Jugó entre semana Visitante	0,301	0,001
Cuota Visitante	0	0
Clasificación Local	0	0
Cuota Local	0	0
Cuota Empate	0	0
Goles anotados Visitante	0	0
Goles anotados Local	0	0
Goles recibidos Visitante	0	0
Goles recibidos Local	0	0
Clasificación Visitante	0	0
Racha Visitante	0	0
Racha Local	0	0

Tabla 39. Análisis de relevancia de atributos (Conjunto Serie A)

Estos problemas de relevancia de atributos pueden ser debidos a dos factores. El primero de ellos es que se necesite un conjunto de entrenamiento mayor para que algunos atributos muestren su verdadera relevancia para resolver el problema. El otro factor que podría influir en el estudio es que los datos de partidos recogidos no siguen un patrón concreto en lo que al resultado se refiere. En el caso de otras competiciones, el equipo que tiene una cuota menor es el que suele conseguir la victoria, y si además este equipo

es el que juega como local, las oportunidades aumentan todavía más. En los datos recogidos en la competición italiana se han visto muchos resultados sorpresa, por lo que a los algoritmos de clasificación les costará tener buenos porcentajes de acierto.

Para este conjunto, los atributos que han mostrado mejor correlación con la clase han sido las zonas en la liga del equipo local y visitante.

7.1.9 Análisis de relevancia de atributos para el conjunto de Competición de Liga:

El último de los conjuntos referente a competiciones de fútbol, es el conjunto de entrenamiento que agrupa los partidos de todas las ligas de las que se han recogido datos. Como ya se comentó anteriormente, este conjunto va a ser usado para intentar crear un modelo de predicción global, que pueda ser aplicado a cualquier competición de liga.

Los resultados obtenidos tras verificar la relevancia de los atributos, han sido muy satisfactorios, ya que tenemos un conjunto de atributos grandes que mantiene una gran correlación con la clase por la que se está clasificando. Como se puede ver en la Tabla 40, los atributos de las cuotas que ofrece la casa de apuestas para el equipo local y visitante son los que tienen una gran correlación con la clase, por lo que éstos presumiblemente formarán parte de los conjuntos que se utilizarán para entrenar los datos mediante clasificadores.

	ChiSquared	GainRatio
Cuota Local	223,815	0,094
Cuota Visitante	213,117	0,082
Zona Visitante	80,917	0,020
Clasificación Visitante	64,814	0,056
Goles anotados Visitante	59,203	0,064
Cuota Empate	45,553	0,054
Zona Local	37,377	0,009
Clasificación Local	30,228	0,034
Goles anotados Local	21,543	0,082
Racha discreta Visitante	16,896	0,006
Racha Local	16,826	0,041
Racha discreta Local	14,256	0,005
Jugó entre semana Local	10,091	0,001
Tiempo	8,862	0,006
Jugó entre semana Visitante	2,473	0,001
Derbi	1,267	0,002
Goles recibidos Visitante	0	0
Racha Visitante	0	0
Goles recibidos Local	0	0

Tabla 40. Análisis de relevancia de atributos (Conjunto Competición de Liga)

Por otro lado, tenemos tres atributos que al no tener ninguna correlación con la clase serán descartados para la próxima fase del proyecto, en la que se entrenarán los conjuntos con diferentes clasificadores para ver con cuál obtenemos mejores resultados. Estos tres atributos que van a ser descartados son la media de goles que reciben los equipos local y visitante y la racha de resultados expresada de forma numérica.

7.1.10 Análisis de relevancia de atributos para el conjunto de Partidos Internacionales:

El último de los conjuntos de competiciones futbolísticas que queda por analizar es el de los partidos de selecciones internacionales. El estudio sobre relevancia de atributos, mostrado en la Tabla 41, revela que algunos de los atributos que forman parte de éste tienen una gran relevancia, lo que será muy beneficioso a la hora de crear clasificadores que tengan buenas tasas de acierto en sus predicciones. Estos atributos son las cuotas del equipo local y visitante que ofrecen las casas de apuestas.

	ChiSquared	GainRatio
Cuota Local	163.070	0.159
Cuota Visitante	145.671	0.151
Clasificación FIFA Visitante	60.079	0.075
TOP Local	57.730	0.036
TOP Visitante	53.787	0.035
Clasificación FIFA Local	39.146	0.095
Cuota Empate	21.118	0.032
Racha Discreta Visitante	19.488	0.015
Racha Discreta Local	18.847	0.014
Racha Local	15.832	0.167
Confederación Local	11.421	0.011
Confederación Visitante	11.034	0.009
Tipo Partido	0.770	0.001
Racha Visitante	0	0
Coef. Confederación Visitante	0	0
Coef. Confederación Local	0	0

Tabla 41. Análisis de relevancia de atributos (Conjunto Partidos Internacionales)

Por otro lado, tenemos un conjunto de tres atributos que debido a la correlación nula con la clase serán descartados para las siguientes fases del proceso. Estos tres atributos son la racha del equipo visitante y los coeficientes que se han asignado a cada una de las confederaciones a las que puede pertenecer una selección nacional de fútbol.

En cuanto al resto de atributos, han mostrado una correlación más que aceptable con la clase, por lo que se tendrán en cuenta para pasos futuros del desarrollo del sistema de predicción de resultados. La combinación de estos atributos con aquellos que muestran

una mayor correlación con la clase, puede ser muy beneficiosa para los clasificadores, ya que puede aumentar la tasa de aciertos de éstos.

7.1.11 Análisis de relevancia de atributos para el conjunto de NBA:

Después de haber analizado todos los conjuntos formados por partidos de fútbol, sólo queda analizar el conjunto formado con los datos recogidos de los partidos de la liga americana de baloncesto (NBA). Como podemos ver en la Tabla 42 y tal y como se había explicado en apartados anteriores, este conjunto de datos cuenta con un número de atributos mayor que el resto.

	ChiSquared	GainRatio
Cuota Local	42,955	0,224
Cuota Visitante	42,955	0,224
Discretización Victorias Visitante	32,174	0,052
% Victorias Visitante	28,029	0,127
Puntos Recibidos Local	23,322	0,102
Discretización Victorias Local	21,962	0,035
Discretización Victorias +100 Local	21,329	0,046
% Victorias Local	20,837	0,089
% Victorias +100 Local	19,775	0,125
% Victorias +100 Visitante	18,484	0,078
Discretización Victorias Rival +100 Local	17,607	0,038
Puntos Recibidos Visitante	17,513	0,076
Discretización Victorias + 100 Visitante	14,776	0,047
Racha Discreta Fuera de Casa	13,139	0,029
Racha Discreta Visitante	12,815	0,035
Discretización Victorias Rival +100 Visitante	12,671	0,024
% Victorias Rival + 100 Local	12,649	0,104
Racha Discreta Local	10,988	0,023
Racha Discreta en Casa	10,729	0,029
Racha Fuera de Casa	9,478	0,103
Racha Visitante	7,532	0,138
Lesiones en equipo Local	6,075	0,020
Lesiones en equipo Visitante	3,064	0,010
Rebotes Visitante	0	0
% Victorias Rival +100 Visitante	0	0
Rebotes Local	0	0
Racha en Casa	0	0
Racha Local	0	0
Puntos Anotados Local	0	0
Puntos Anotados Visitante	0	0

Tabla 42. Análisis de relevancia de atributos (Conjunto NBA)

El resultado obtenido tras realizar el análisis de relevancia de los atributos arroja buenos resultados, ya que la mayoría de los atributos guardan una correlación aceptable

con la clase con la que se clasifica. Tan solo siete atributos muestran una correlación nula, por lo que serán descartados para fases posteriores del análisis de los datos. Estos atributos son la media de rebotes del equipo local y visitante, la media de puntos anotados por el equipo local y visitante, el porcentaje de victorias del equipo visitante cuando encaja más de 100 puntos, la racha en casa del equipo local y la racha total de resultados del equipo local.

Una vez realizado el análisis de atributos de todos los conjuntos de entrenamiento se ha logrado descartar una serie de atributos que como se ha podido ver, no guardan ninguna correlación con la clase por la que se va a clasificar en fases posteriores. Esto no hace más que facilitar el trabajo, ya que se pueden apartar los atributos sin correlación de cara a las fases posteriores del análisis. Al tener ya definidos los conjuntos de entrenamiento para cada una de las competiciones, podemos pasar a la siguiente fase del proyecto, que será el entrenamiento de los conjuntos de cada una de las competiciones a través de alguno de los clasificadores que nos ofrece la herramienta WEKA.

7.2 ANEXO B: Entrenamiento de conjuntos a través de Clasificadores

7.2.1 Conjunto de Partidos de Competiciones Europeas:

A continuación se van a presentar los resultados obtenidos con los seis clasificadores para el conjunto que recoge los partidos de las competiciones europeas (Champions League y Europa League).

Red Bayesiana (Bayes Net):

Una vez entrenada una red bayesiana se han obtenido los resultados que se muestran en la Figura 48. El **porcentaje de aciertos** que se ha conseguido es bueno, ya que se ha conseguido un 56,19% de pronósticos acertados, que hacen referencia a 127 aciertos sobre 226 predicciones realizadas.

```
Correctly Classified Instances         127               56.1947 %
Incorrectly Classified Instances        99               43.8053 %
Kappa statistic                          0.3098
Mean absolute error                      0.3344
Root mean squared error                  0.4673
Relative absolute error                 79.3875 %
Root relative squared error            101.8524 %
Total Number of Instances              226

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.664     0.284      0.689     0.664     0.676      0.697    1
                 0.281     0.148      0.39      0.281     0.327      0.505    X
                 0.644     0.246      0.481     0.644     0.551      0.753    2
Weighted Avg.    0.562     0.24       0.559     0.562     0.555      0.664

=== Confusion Matrix ===

  a  b  c   <-- classified as
 73 14 23 |  a = 1
 23 16 18 |  b = X
 10 11 38 |  c = 2
```

Figura 48. Resultados Bayes Net - Conjunto partidos europeos WEKA

El **estadístico kappa** para este conjunto muestra un valor de 0,309, que simplemente nos dice que la correlación de los elementos del conjunto no es perfecta (se necesita un valor de kappa próximo a 1). De todas formas, el valor que ofrece el estudio nos dice que existe algo de correlación entre los partidos del grupo.

El siguiente parámetro que se va a analizar son las tasas de **True Positives** (es decir, partidos que terminaron con el mismo resultado proporcionado po el clasificador). Como se puede observar en la figura 20, las tasas para los resultados de victoria local y visitante están cercanas a los 2/3, lo que es un muy buen dato. En cambio para el resultado de empate tan solo se tiene un 28.1% para esta medida. Para minimizar los riesgos de acertar este tipo de resultados en fases posteriores se tratará de forma específica la predicción de empates en función de los resultados que se obtengan del entrenamiento con los clasificadores.

Por último, se analizará la **precisión** de este modelo para cada uno de los resultados (es decir, de los partidos que han sido clasificados con un determinado resultado, cuántos han terminado con dicho resultado en su marcador). En este caso podemos ver que para los resultados que implican la victoria del equipo local se ha conseguido una precisión del 68,9%, por lo que es un gran resultado teniendo en cuenta que el resultado que se da con más frecuencia es la victoria del equipo local.

Regresión Logística (Logistic):

Una vez realizado el entrenamiento con el clasificador de regresión logística pasamos a analizar los datos que nos da la herramienta WEKA (Figura 49).

```
Correctly Classified Instances         126               55.7522 %
Incorrectly Classified Instances       100               44.2478 %
Kappa statistic                          0.2401
Mean absolute error                      0.3692
Root mean squared error                  0.4357
Relative absolute error                 87.6248 %
Root relative squared error             94.9605 %
Total Number of Instances              226

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.845     0.552      0.592     0.845     0.697      0.718    1
                 0         0.012      0         0         0          0.542    X
                 0.559     0.204      0.493     0.559     0.524      0.746    2
Weighted Avg.    0.558     0.325      0.417     0.558     0.476      0.681

=== Confusion Matrix ===

  a  b  c   <-- classified as
 93  2 15 |  a = 1
 38  0 19 |  b = X
 26  0 33 |  c = 2
```

Figura 49. Resultados Logistic - Conjunto partidos europeos WEKA

El **porcentaje de aciertos** conseguido es muy similar al del conseguido con la Red Bayesiana. En este caso el clasificador ha fallado en una predicción más que en el caso anterior, situando la tasa de acierto en un 55,75%, que no deja de ser nada mala.

A continuación podemos ver que el **estadístico kappa** sí que es algo menor que en el caso anterior. Para este clasificador tenemos un valor de kappa de 0,240, lo que muestra que los atributos escogidos para este clasificador no están altamente correlacionados.

En cuanto a las tasas de **true positives** que recoge la herramienta, podemos destacar el alto porcentaje obtenido para los partidos en los que gana el equipo local (84,5%) y el equipo visitante (55,9%). Para el caso de los empates no se ha conseguido acertar ninguno de ellos. Este último dato, aunque no es nada bueno, no debe preocupar, ya que como ya anunciamos anteriormente se gestionará la predicción de empates a través de estrategias fuera de los modelos de clasificación.

Finalmente, los datos de **precisión** también son buenos, ya que tenemos un 59,2% para partidos en los que consigue la victoria el equipo local y un 49,3% para partidos en los que el equipo visitante sale vencedor.

Multilayer Perceptron (Perceptrón Multicapa):

La Figura 50 muestra los resultados obtenidos con el conjunto de partidos de las competiciones europeas y la utilización del perceptrón multicapa. Los resultados obtenidos son algo peores que en los dos casos anteriores, obteniendo una **tasa de aciertos** del 50%.

```
Correctly Classified Instances         113               50      %
Incorrectly Classified Instances       113               50      %
Kappa statistic                          0.1529
Mean absolute error                      0.3599
Root mean squared error                  0.4817
Relative absolute error                 85.4217 %
Root relative squared error            104.9905 %
Total Number of Instances              226

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.791     0.543      0.58      0.791     0.669      0.664    1
                 0.123     0.118      0.259     0.123     0.167      0.542    X
                 0.322     0.18       0.388     0.322     0.352      0.677    2
Weighted Avg.    0.5       0.341      0.449     0.5       0.46       0.636

=== Confusion Matrix ===

  a  b  c   <-- classified as
 87 10 13 |  a = 1
 33  7 17 |  b = X
 30 10 19 |  c = 2
```

Figura 50. Resultados Multilayer Perceptron - Conjunto partidos europeos WEKA

En cuanto al **estadístico kappa**, tenemos que para este caso el valor del estadístico es menor que en el de los dos casos anteriores, obteniendo en este caso un valor de 0,152 que muestra que los atributos escogidos en este conjunto para realizar el entrenamiento no tienen casi correlación entre ellos.

En cuanto a las tasas de **true positives,** los datos siguen la tendencia de los parámetros ya analizados, ya que aunque los resultados no son del todo malos, son peores que los de los clasificadores utilizados anteriormente, teniendo un 79,1% para las victorias del equipo local, un 12,3% para empates y finalmente un 32,2% para victorias del equipo visitante.

Por último, queda analizar el parámetro de **precisión** para cada uno de los resultados posibles. Para el caso de victorias del equipo local este parámetro arroja una precisión del 58%. Respecto a las victorias del equipo visitante se ha obtenido una precisión del 38,8% mientras que para los empates tan solo se llega al 25,9%, que aunque es bajo mejora el valor obtenido mediante la regresión logística.

OneR:

El clasificador OneR, que es el más simple de los que vamos a utilizar, ha conseguido igualar los resultados del que hasta ahora era el mejor de los clasificadores, la red bayesiana, tal y como muestra la Figura 51. El **porcentaje de aciertos** que ha conseguido ha sido del 56,19%.

```
Correctly Classified Instances         127               56.1947 %
Incorrectly Classified Instances        99               43.8053 %
Kappa statistic                          0.2686
Mean absolute error                      0.292
Root mean squared error                  0.5404
Relative absolute error                 69.3173 %
Root relative squared error            117.7906 %
Total Number of Instances              226

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.8       0.457      0.624     0.8       0.701      0.672    1
                 0.123     0.077      0.35      0.123     0.182      0.523    X
                 0.542     0.198      0.492     0.542     0.516      0.672    2
Weighted Avg.    0.562     0.293      0.521     0.562     0.522      0.634

=== Confusion Matrix ===

  a  b  c   <-- classified as
 88  6 16 |  a = 1
 33  7 17 |  b = X
 20  7 32 |  c = 2
```

Figura 51. Resultados One R - Conjunto partidos europeos WEKA

En términos del **estadístico kappa** que nos ofrece el estudio, el valor está algo por debajo del de la Red Bayesiana, situándose en 0,268, lo que se traduce en una débil correlación entre los atributos de la clase.

Si prestamos atención a las tasas de **true positives** que nos ofrece este estudio encontramos buenas tasas sobre todo para los partidos que han terminado con victoria del equipo local. En ese caso se ha conseguido una tasa de true positives del 80%, mientras que también es muy destacable la tasa del 56,2% que se ha obtenido para los partidos que han terminado con victoria visitante. La tasa de true positives para los empates sigue la misma tendencia que con el resto de clasificadores, habiendo obtenido tan solo un 12,3% para los partidos que han acabado con este resultado.

Finalmente, si atendemos al atributo de **precisión** los resultados también son bastante aceptables. En concreto se ha obtenido un 62,4% para el caso de los partidos con victoria local, un 52,1% para los partidos con victoria visitante y un 35% para partidos que han concluido con empate.

J48:

El siguiente de los clasificadores que se ha utilizado es el árbol J48. Este algoritmo de clasificación ha conseguido situarse al nivel de los algoritmos con los mejores resultados para este conjunto de entrenamiento, tal y como muestra la Figura 52. En concreto, ha conseguido un **porcentaje de aciertos** del 56,17%. Este porcentaje logra igualar a los mejores resultados obtenidos por la red bayesiana y el algoritmo OneR

```
Correctly Classified Instances         127               56.1947 %
Incorrectly Classified Instances        99               43.8053 %
Kappa statistic                          0.2518
Mean absolute error                      0.3632
Root mean squared error                  0.4457
Relative absolute error                 86.2151 %
Root relative squared error             97.1384 %
Total Number of Instances              226

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.836     0.534      0.597     0.836     0.697      0.713    1
                 0.053     0.03       0.375     0.053     0.092      0.51     X
                 0.542     0.192      0.5       0.542     0.52       0.731    2
Weighted Avg.    0.562     0.318      0.516     0.562     0.498      0.666

=== Confusion Matrix ===

  a  b  c   <-- classified as
 92  4 14 |  a = 1
 36  3 18 |  b = X
 26  1 32 |  c = 2
```

Figura 52. Resultados J48 - Conjunto partidos europeos WEKA

En cuanto al **estadístico kappa**, para este clasificador tenemos un resultado de 0,251, un valor muy similar al que se ha obtenido con el algoritmo OneR.

Respecto a las tasas de **true positives** que nos arroja el estudio podemos ver que para el resultado de victoria del equipo local se ha logrado obtener un altísimo 83,6%, mientras que para el resultado de victoria del equipo visitante tenemos un 54,2%. Como en clasificadores anteriores, la tasa para el empate registra un bajo 5,3%.

Finalmente, si miramos el atributo de **precisión** podremos ver que no se han obtenido valores muy altos excepto para los empates, donde se ha conseguido alcanzar una cifra relativamente alta si la comparamos con clasificadores anteriores, ya que en este caso se ha logrado un 37,5% de precisión. Mientras tanto, la precisión en partidos con victoria local es del 59,7% y en partidos con victoria visitante asciende hasta un 50%.

Random Forest:

Por último, y para concluir el entrenamiento del conjunto de entrenamiento que recopila los partidos de competiciones europeas se utilizará el Random Forest para intentar generar un modelo que permita realizar las predicciones de manera fiable. Lamentablemente, como se puede observar en las estadísticas que nos ofrece WEKA en la Figura 53, el Random Forest no consigue buenos resultados para este conjunto. Concretamente ha conseguido un **porcentaje de aciertos** del 49,11%, que aunque es superior al 33%, que son las probabilidades que tenemos a priori de acertar, se queda lejos de los resultados de otros clasificadores utilizados.

```
Correctly Classified Instances         111               49.115  %
Incorrectly Classified Instances       115               50.885  %
Kappa statistic                          0.1885
Mean absolute error                      0.3548
Root mean squared error                  0.5159
Relative absolute error                 84.2261 %
Root relative squared error            112.4538 %
Total Number of Instances              226

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.627     0.388      0.605     0.627     0.616      0.659    1
                 0.175     0.237      0.2       0.175     0.187      0.441    X
                 0.542     0.18       0.516     0.542     0.529      0.733    2
Weighted Avg.    0.491     0.295      0.48      0.491     0.485      0.623

=== Confusion Matrix ===

  a  b  c   <-- classified as
 69 28 13 |  a = 1
 30 10 17 |  b = X
 15 12 32 |  c = 2
```

Figura 53. Resultados Random Forest - Conjunto partidos europeos WEKA

Si se observa el **estadístico kappa**, también se puede ver que tiene un valor muy bajo (0,188), lo que hace ver que los atributos no mantienen una gran correlación entre ellos.

En cuanto a las tasas de **true positives**, tenemos niveles razonables para las victorias de equipo local o visitante (62,7% y 54,2% respectivamente) mientras que la tasa para los empates se sitúa en el 17,5%.

Por último, respecto a la **precisión** de cada uno de los resultados, tendremos que para victorias del equipo local tenemos una precisión del 60,5%, para victorias del equipo visitante será del 51,6% mientras que para los empates la precisión quedará en el 20%.

Resumen de Clasificadores para el conjunto de Competiciones Europeas:

Una vez se ha entrenado el conjunto de entrenamiento con cada uno de los seis clasificadores, estamos en disposición de comparar los resultados para ver qué clasificadores se adaptan mejor a nuestro problema. La Tabla 43 resume los resultados obtenidos con cada uno de los clasificadores.

	% Acierto	Kappa	Precisión Media
Red Bayesiana	56.19%	0.309	55.9%
Regresión Logística	55.75%	0.240	41.7%
Perceptrón Multicapa	50%	0.152	44.9%
OneR	56.19%	0.268	52.1%
J48	56.19%	0.251	51.6%
Random Forest	49.11%	0.188	48%

Tabla 43. Análisis del entrenamiento (Conjunto Competiciones Europeas)

Tras analizar conjuntamente los resultados de cada uno de los entrenamientos con clasificadores podemos sacar como conclusión que los clasificadores que mejor se adaptan al problema son la Red Bayesiana, el algoritmo OneR y el J48.

Entre estos tres clasificadores hemos escogido dos para la fase de implementación. La Red Bayesiana será escogida en primer lugar, dados sus buenos resultados tanto en el porcentaje de aciertos como en el resto de atributos. El caso del segundo clasificador es más complicado, ya que tanto el algoritmo OneR como el J48 tienen resultados similares.

Si pasamos a analizar estos dos últimos clasificadores en detalle, se va a optar por escoger el árbol J48, ya que las reglas que ha generado el algoritmo OneR parece que producen aciertos de manera aleatoria. La Figura 54 muestra las reglas generadas por dicho algoritmo.

```
=== Classifier model (full training set) ===

cuotaV:
        < 2.075 -> 2
        < 2.225 -> X
        < 2.425 -> 2
        < 3.125 -> 1
        < 3.75  -> 2
        < 4.825 -> 1
        < 5.3   -> X
        >= 5.3  -> 1
```

Figura 54. Reglas OneR - Conjunto partidos europeos WEKA

Como se puede observar, hay reglas que no tienen mucha lógica, como por ejemplo que para valores entre 2,075 y 2,225 el resultado sea un empate, mientras que para valores inmediatamente superiores e inferiores de la cuota en las casas de apuestas del equipo visitante, el resultado sea la victoria visitante.

7.2.2 Conjunto de Partidos de Champions League:

A continuación se van a presentar los resultados obtenidos con los seis clasificadores para el conjunto que recoge los partidos de la máxima competición continental europea, la Champions League:

Red Bayesiana (Bayes Net):

Después de haber entrenado el conjunto de entrenamiento mediante una red bayesiana se han obtenido los resultados que se muestran en la Figura 55. El **porcentaje de aciertos** que se ha conseguido es bueno, ya que se ha conseguido un 56,61% de pronósticos acertados.

```
Correctly Classified Instances          77               56.6176 %
Incorrectly Classified Instances        59               43.3824 %
Kappa statistic                          0.3241
Mean absolute error                      0.3365
Root mean squared error                  0.4514
Relative absolute error                 79.1785 %
Root relative squared error             97.9482 %
Total Number of Instances              136

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.641     0.236      0.707     0.641     0.672      0.719    1
                 0.313     0.154      0.385     0.313     0.345      0.586    X
                 0.65      0.271      0.5       0.65      0.565      0.749    2
Weighted Avg.    0.566     0.227      0.57      0.566     0.564      0.696

=== Confusion Matrix ===

  a  b  c   <-- classified as
 41  9 14 |  a = 1
 10 10 12 |  b = X
  7  7 26 |  c = 2
```

Figura 55. Resultados Bayes Net - Conjunto Champions League WEKA

El **estadístico kappa** para este conjunto muestra un valor de 0,324, lo que muestra una correlación aceptable entre los atributos utilizados.

En cuanto a la tasa de **True Positives**, el estudio refleja muy buenos resultados, obteniendo una tasa del 64,1% para victorias locales, el 65% para victorias visitantes y un 31,3% para los empates.

Por último, se analizará la **precisión** de este modelo para cada uno de los resultados En este caso podemos ver que para los resultados que implican la victoria del equipo local se ha conseguido una precisión del 70,7%, un gran resultado teniendo en cuenta que en esta competición el resultado más repetido es la victoria del equipo local. En cuanto a

las victorias del equipo visitante tenemos un 50% de precisión y en los empates la precisión alcanza el 38,5%.

Regresión Logística (Logistic):

El siguiente de los clasificadores a estudiar es el de Regresión Logística, cuyos resultados se muestran en la Figura 56.

```
Correctly Classified Instances          70               51.4706 %
Incorrectly Classified Instances        66               48.5294 %
Kappa statistic                          0.231
Mean absolute error                      0.3625
Root mean squared error                  0.4463
Relative absolute error                 85.279  %
Root relative squared error             96.8446 %
Total Number of Instances              136

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.625     0.375      0.597     0.625     0.611      0.691    1
                 0.25      0.192      0.286     0.25      0.267      0.604    X
                 0.55      0.198      0.537     0.55      0.543      0.729    2
Weighted Avg.    0.515     0.28       0.506     0.515     0.51       0.682

=== Confusion Matrix ===

  a  b  c   <-- classified as
 40 13 11 |  a = 1
 16  8  8 |  b = X
 11  7 22 |  c = 2
```

Figura 56. Resultados Logistic - Conjunto Champions League WEKA

El **porcentaje de aciertos** ha bajado con respecto a la Red Bayesiana, situando la tasa de aciertos de este clasificador en el 51,47%

En cuanto al **estadístico kappa**, se puede observar que es inferior al del estudio anterior con la Red Bayesiana. Para este clasificador tenemos un valor de kappa de 0.231, lo que muestra que los atributos escogidos para este clasificador no están altamente correlacionados.

En cuanto a las tasas de **true positives** que recoge la herramienta, se puedo observar que las tasas son discretas comparadas con el ejemplo anterior. Concretamente tenemos una tasa del 62,5% para victorias locales, un 55% para victorias visitantes y un 25% para los empates.

Finalmente, los datos de **precisión** también son aceptables, ya que tenemos un 59,7% para partidos en los que consigue la victoria el equipo local y un 53,7% para partidos en los que el equipo visitante sale vencedor. En partidos en los que el resultado es un empate tendremos una precisión estimada del 0.286.

Multilayer Perceptron (Perceptrón Multicapa):

A continuación se ha decidido entrenar el conjunto de partidos de Champions League con un perceptrón multicapa. Los resultados obtenidos (Figura 57) son buenos en términos de **tasa de aciertos**, ya que se ha obtenido un 53,67% de aciertos. En cambio, en términos de predicción los resultados no son tan buenos, ya que este modelo es incapaz de predecir empates como puede observarse en la matriz de confusión.

```
Correctly Classified Instances          73               53.6765 %
Incorrectly Classified Instances        63               46.3235 %
Kappa statistic                          0.2547
Mean absolute error                      0.3752
Root mean squared error                  0.4359
Relative absolute error                 88.2721 %
Root relative squared error             94.6028 %
Total Number of Instances              136

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.672     0.306      0.662     0.672     0.667      0.699    1
                 0         0          0         0         0          0.517    X
                 0.75      0.427      0.423     0.75      0.541      0.697    2
Weighted Avg.    0.537     0.269      0.436     0.537     0.473      0.655

=== Confusion Matrix ===

  a  b  c   <-- classified as
 43  0 21 |  a = 1
 12  0 20 |  b = X
 10  0 30 |  c = 2
```

Figura 57. Resultados Multilayer Perceptron - Conjunto Champions League WEKA

En cuanto al **estadístico kappa**, el valor para este clasificador se aproxima mucho al conseguido por el clasificador de regresión logística. Para el perceptrón multicapa se ha obtenido un valor de 0,254.

En cuanto a las tasas de **true positives** tenemos una tasa muy buena para partidos en los que el equipo visitante consigue la victoria (75%). Mientras tanto, la tasa para partidos que acaban con victoria del equipo local es del 67,2%. Como este modelo no realiza predicciones de empates, no se tiene tasa de true positives para este resultado.

Por último queda analizar el parámetro de **precisión**. En este caso la precisión para empates es del 0%, ya que ninguno de los empates ha sido clasificado correctamente ante la imposibilidad del modelo para predecir este tipo de resultados. En el caso de las victorias del equipo local, la precisión asciende a un 66,2%, mientras que para victorias del equipo visitante la precisión es del 42,3%.

OneR:

Los datos de la Figura 58 muestran los resultados del algoritmo OneR aplicado al conjunto de entrenamiento de los partidos de Champions League. El **porcentaje de aciertos** que ha conseguido ha sido del 50%.

```
Correctly Classified Instances          68               50      %
Incorrectly Classified Instances        68               50      %
Kappa statistic                          0.1916
Mean absolute error                      0.3333
Root mean squared error                  0.5774
Relative absolute error                 78.422  %
Root relative squared error            125.2874 %
Total Number of Instances              136

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.641     0.403      0.586     0.641     0.612      0.619    1
                 0.063     0.058      0.25      0.063     0.1        0.502    X
                 0.625     0.344      0.431     0.625     0.51       0.641    2
Weighted Avg.    0.5       0.304      0.461     0.5       0.462      0.598

=== Confusion Matrix ===

  a  b  c   <-- classified as
 41  4 19 |  a = 1
 16  2 14 |  b = X
 13  2 25 |  c = 2
```

Figura 58. Resultados One R - Conjunto Champions League WEKA

En términos del **estadístico kappa** que nos ofrece el estudio, el valor está muy por debajo del de la red bayesiana, situándose en 0,191, lo que se traduce en una débil correlación entre los atributos de la clase.

Si prestamos atención a las tasas de **true positives** que nos ofrece este estudio encontramos tasas aceptables para los dos tipos de victorias. Se ha conseguido una tasa de true positives del 64,1% para los partidos que acaban con victoria local, mientras que tenemos una tasa del 56,2% para los partidos que han terminado con victoria visitante. La tasa de true positives para los empates sigue la misma tendencia que con el resto de clasificadores, habiendo obtenido tan solo un 6,3% para los partidos que han acabado con este resultado.

Finalmente, si atendemos al atributo de **precisión** los resultados también son bastante aceptables. En concreto se ha obtenido un 58,6% para el caso de los partidos con victoria local y un 43,1% para los partidos con victoria visitante. Mientras tanto, la precisión en los partidos que han acabado con empate es del 25%.

J48:

El siguiente de los clasificadores que se ha utilizado es el árbol J48. Este algoritmo de clasificación no ha logrado alcanzar las tasas de acierto que ha conseguido la Red Bayesiana y el Perceptrón Multicapa (Figura 59), pero las deficiencias en el modelo que presenta el perceptrón hacen que este árbol sea un algoritmo a tener en cuenta. La **tasa de aciertos** que alcanza este algoritmo es del 52,94%.

```
Correctly Classified Instances          72               52.9412 %
Incorrectly Classified Instances        64               47.0588 %
Kappa statistic                          0.2609
Mean absolute error                      0.3746
Root mean squared error                  0.4444
Relative absolute error                 88.1372 %
Root relative squared error             96.4441 %
Total Number of Instances              136

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.578     0.236      0.685     0.578     0.627      0.659    1
                 0.031     0.01       0.5       0.031     0.059      0.503    X
                 0.85      0.479      0.425     0.85      0.567      0.686    2
Weighted Avg.    0.529     0.254      0.565     0.529     0.476      0.63

=== Confusion Matrix ===

  a  b  c   <-- classified as
 37  1 26 |  a = 1
 11  1 20 |  b = X
  6  0 34 |  c = 2
```

Figura 59. Resultados J48 - Conjunto Champions League WEKA

En cuanto al **estadístico kappa**, para este clasificador tenemos un resultado de 0,260, un valor muy similar al que se ha obtenido con otros algoritmos.

Si miramos las tasas de **true positives** que nos arroja el estudio podemos ver que para el resultado de victoria del equipo visitante se ha logrado obtener un altísimo 85%, mientras que para el resultado de victoria del equipo visitante tenemos un 57,8%. Como en clasificadores anteriores, la tasa para el empate registra un bajo 3,1%.

Finalmente, si miramos el atributo de **precisión** podremos ver que para resultados en los que el equipo local ha resultado vencedor tenemos una precisión del 68,5%, mientras que para victorias del equipo visitante se ha llegado a alcanzar un 42,5% de precisión. Aunque para el dato de precisión en los empates tenemos un 50%, cabe resaltar que el algoritmo sólo ha clasificado dos resultados como empate, siendo uno de ellos clasificado de manera correcta. En este algoritmo podemos ver que ocurre lo mismo que en el Perceptrón Multicapa, y el algoritmo no es capaz de predecir empates con facilidad.

Random Forest:

Por último, y para concluir el entrenamiento del conjunto de entrenamiento que recopila los partidos de la Champions League se utilizará el Random Forest para entrenar el conjunto. En este caso los resultados obtenidos son mejores que con el conjunto de Competiciones Europeas (ver Figura 53). En este caso la **tasa de acierto** asciende al 51,32%.

```
Correctly Classified Instances         116               51.3274 %
Incorrectly Classified Instances       110               48.6726 %
Kappa statistic                          0.1867
Mean absolute error                      0.3709
Root mean squared error                  0.4678
Relative absolute error                 88.0519 %
Root relative squared error            101.9655 %
Total Number of Instances              226

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.764     0.5        0.592     0.764     0.667      0.659    1
                 0.14      0.154      0.235     0.14      0.176      0.472    X
                 0.407     0.156      0.48      0.407     0.44       0.716    2
Weighted Avg.    0.513     0.323      0.473     0.513     0.484      0.627

=== Confusion Matrix ===

  a  b  c   <-- classified as
 84 15 11 |  a = 1
 34  8 15 |  b = X
 24 11 24 |  c = 2
```

Figura 60. Resultados Random Forest - Conjunto Champions League WEKA

Si se observa el **estadístico kappa**, se podrá ver que la correlación entre los atributos del conjunto es muy baja, lo que demuestra el valor 0,186. Esto puede mostrar que los resultados obtenidos en los test han sido obtenidos correctamente por mera casualidad. Esto hace que este algoritmo quede prácticamente descartado como algoritmo a implementar para realizar las predicciones de esta competición.

En cuanto a las tasas de **true positives**, tenemos una buena tasa para las victorias locales (76,4%) y unas tasas algo peores para las victorias visitantes (40,7%) y los empates (14%).

Por último, si miramos la **precisión** de cada uno de los resultados, tendremos porcentajes de precisión muy discretos. Para los partidos ganados por el equipo local tendremos un pobre 59,2% de precisión. En los partidos en los que gana el equipo visitante la precisión caerá hasta el 48%, mientras que para los partidos que acaban con empate la precisión será del 23,5%.

Resumen de Clasificadores para el conjunto de Champions League:

Una vez se ha entrenado el conjunto de entrenamiento con cada uno de los seis clasificadores, estamos en disposición de comparar los resultados para ver qué clasificadores se adaptan mejor a nuestro problema. La Tabla 44 resume los resultados obtenidos con cada uno de los clasificadores.

	% Acierto	Kappa	Precisión Media
Red Bayesiana	56.61%	0.324	57.0%
Regresión Logística	51.47%	0.231	50.6%
Perceptrón Multicapa	53.67%	0.254	43.6%
OneR	50.00%	0.191	46.1%
J48	52.94%	0.260	56.5%
Random Forest	51.32%	0.186	47.3%

Tabla 44. Análisis del entrenamiento (Conjunto Champions League)

Tras realizar el análisis individual de cada uno de los clasificadores, ahora es el momento de realizar un análisis conjunto teniendo en cuenta los datos que ofrece la tabla superior.

El clasificador que ofrece unos mejores resultados vuelve a ser la Red Bayesiana, con un 56,61% de acierto y las mejores tasas de true positive y precisión. Además, el estadístico kappa para este algoritmo es del 0,324, lo que muestra que los atributos seleccionados para este estudio están más correlacionados que los que han sido utilizados en el resto de algoritmos de clasificación.

El otro clasificador, que se va a coger para ser implementado en fases posteriores va a ser el algoritmo del árbol J48. Aunque el Perceptrón Multicapa tiene mejores resultados en cuanto a porcentaje de acierto, se ha decidido coger el algoritmo J48 debido a la buena precisión de sus predicciones. Además, su buena tasa de *true positives* para las victorias de equipos visitantes (85%) hacen que sea un algoritmo apto para las pretensiones que se están buscando. Además, el buen porcentaje de aciertos que da este algoritmo (52,94%) hace que sea el otro algoritmo elegido para la predicción de los resultados de la competición de la Champions League.

Al igual que ha ocurrido con los algoritmos elegidos en el conjunto de las competiciones europeas, para este caso también se van a escoger a la Red Bayesiana y al árbol J48 para realizar las predicciones de los partidos de la competición de la Champions League.

7.2.3 Conjunto de Partidos de Europa League:

La Figura 33 muestra los resultados obtenidos con los seis clasificadores para el conjunto que recoge los partidos de la otra competición europea que se quiere analizar en este proyecto, la Europa League.

Red Bayesiana (Bayes Net):

Una vez más comenzamos el proceso de análisis de los datos de cada uno de los clasificadores, en este caso para el conjunto de partidos de la Europa League. El primero de los clasificadores que van a ser utilizados es la Red Bayesiana (Figura 61), que en este caso nos proporciona una **tasa de aciertos** muy buena del 56,19%.

```
Correctly Classified Instances         127               56.1947 %
Incorrectly Classified Instances        99               43.8053 %
Kappa statistic                          0.3098
Mean absolute error                      0.3346
Root mean squared error                  0.4673
Relative absolute error                 79.4221 %
Root relative squared error            101.8579 %
Total Number of Instances              226

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.664     0.284      0.689     0.664     0.676      0.697    1
                 0.281     0.148      0.39      0.281     0.327      0.506    X
                 0.644     0.246      0.481     0.644     0.551      0.753    2
Weighted Avg.    0.562     0.24       0.559     0.562     0.555      0.664

=== Confusion Matrix ===

  a  b  c   <-- classified as
 73 14 23 |  a = 1
 23 16 18 |  b = X
 10 11 38 |  c = 2
```

Figura 61. Resultados Bayes Net - Conjunto Europa League WEKA

El **estadístico kappa** para este conjunto muestra un valor de 0,309, lo que muestra una correlación aceptable entre los atributos utilizados.

En cuanto a la tasa de **True Positives**, el estudio refleja muy buenos resultados, obteniendo una tasa del 66,4% para victorias locales, el 64,4% para victorias visitantes y un 28,1% para los empates.

Por último, se analizará la **precisión** de este modelo para cada uno de los resultados En este caso podemos ver que para los resultados que implican la victoria del equipo local se ha conseguido una precisión del 68,9%, que de nuevo es un gran resultado teniendo en cuenta que en esta competición el resultado más repetido es la victoria del

equipo local. En cuanto a las victorias del equipo visitante tenemos un 48,1% de precisión y en los empates la precisión alcanza el 39%.

Regresión Logística (Logistic):

El siguiente de los clasificadores a estudiar es el de Regresión Logística (Figura 62). Comenzando por el **porcentaje de aciertos**, este clasificador ha llegado a alcanzar una tasa de aciertos del 55,30%.

```
Correctly Classified Instances         125               55.3097 %
Incorrectly Classified Instances       101               44.6903 %
Kappa statistic                          0.2392
Mean absolute error                      0.3695
Root mean squared error                  0.4345
Relative absolute error                 87.6505 %
Root relative squared error             94.7008 %
Total Number of Instances              226

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.818     0.534      0.592     0.818     0.687      0.731    1
                 0.035     0.03       0.286     0.035     0.063      0.491    X
                 0.559     0.204      0.493     0.559     0.524      0.761    2
Weighted Avg.    0.553     0.321      0.489     0.553     0.487      0.678

=== Confusion Matrix ===

  a  b  c   <-- classified as
 90  3 17 |  a = 1
 38  2 17 |  b = X
 24  2 33 |  c = 2
```

Figura 62. Resultados Logistic - Conjunto Europa League WEKA

En cuanto al **estadístico kappa**, ha reducido su valor respecto al del estudio anterior con la Red Bayesiana. Para este clasificador tenemos un valor de kappa de 0,239, lo que muestra que los atributos escogidos para este clasificador tienen una ligera correlación entre ellos.

En cuanto a las tasas de **true positives** que recoge la herramienta, se puedo observar que las tasas son muy buenas, sobre todo la tasa para partidos en los que el equipo local ha conseguido salir vencedor. Para este tipo de partidos la tasa asciende al 81,8%. Mientras tanto, la tasa para partidos en los que el equipo visitante sale vencedor la tasa es del 55,9%. Finalmente la tasa de true positives para partidos que acaban en empate es del 3,5% y mantiene la tendencia de conjuntos y algoritmos utilizados con anterioridad.

Por último, los datos de **precisión** también son aceptables, ya que están dentro de la media que se ha podido observar en experimentos con conjuntos y algoritmos anteriores. La precisión para los partidos en los que el equipo local sale vencedor es del 59,2%,

mientras que para partidos en los que el equipo visitante es el vencedor la precisión se queda en el 49,3%. La precisión para partidos que acaban en empate sería del 28,6%.

Multilayer Perceptron (Perceptrón Multicapa):

A continuación se ha entrenado el conjunto de partidos de la competición de Europa League con un perceptrón multicapa (Figura 63). El resultado obtenido en términos de **tasa de aciertos es muy satisfactorio**, ya que se ha obtenido un 55,30% de aciertos, lo que significa igualar la tasa de aciertos del algoritmo de regresión logística. Por otro lado, una vez más el perceptrón multicapa muestra deficiencias a la hora de predecir los empates, lo que hace que sea un modelo muy limitado para la predicción de resultados en esta competición.

```
Correctly Classified Instances         125               55.3097 %
Incorrectly Classified Instances       101               44.6903 %
Kappa statistic                          0.2165
Mean absolute error                      0.3712
Root mean squared error                  0.4532
Relative absolute error                 88.1105 %
Root relative squared error             98.7797 %
Total Number of Instances              226

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.873     0.629      0.568     0.873     0.688      0.668    1
                 0.018     0.03       0.167     0.018     0.032      0.495    X
                 0.475     0.138      0.549     0.475     0.509      0.708    2
Weighted Avg.    0.553     0.35       0.462     0.553     0.476      0.635

=== Confusion Matrix ===

  a  b  c   <-- classified as
 96  4 10 |  a = 1
 43  1 13 |  b = X
 30  1 28 |  c = 2
```

Figura 63. Resultados Multilayer Perceptron - Conjunto Europa League WEKA

En cuanto al **estadístico kappa**, el valor para este clasificador se aproxima mucho a los conseguidos por otros clasificadores como el de regresión logística. Para el perceptrón multicapa se ha obtenido un valor de 0,216

En cuanto a las tasas de **true positives** tenemos una tasa muy buena para partidos en los que el equipo local consigue la victoria (87,3%). Mientras tanto, la tasa para partidos que acaban con victoria del equipo visitante es del 47,5%. Otro problema que puede observarse en la matriz de confusión de este modelo es que la mayoría de los resultados son clasificados como victoria del equipo local, lo que hace que el porcentaje de aciertos sea grande, ya que este resultado es el que más se repite. No obstante no parece que los resultados que ofrece puedan servir para predecir resultados en un conjunto de test mayor.

Por último, queda analizar el parámetro de **precisión**. Para los partidos clasificados como victoria del local tenemos un 56,8% de precisión, mientras que para partidos con victoria visitante la precisión se queda en un 54,9%. Para partidos cuyo resultado ha sido de empate la precisión será del 16,7%.

OneR:

Los datos de la Figura 64 muestran los resultados del algoritmo OneR aplicado al conjunto de entrenamiento de los partidos de Europa League. El **porcentaje de aciertos** que ha conseguido ha sido del 53,98%.

```
Correctly Classified Instances         122               53.9823 %
Incorrectly Classified Instances       104               46.0177 %
Kappa statistic                          0.2211
Mean absolute error                      0.3068
Root mean squared error                  0.5539
Relative absolute error                 72.8065 %
Root relative squared error            120.7304 %
Total Number of Instances              226

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.818     0.509      0.604     0.818     0.695      0.655    1
                 0.035     0.083      0.125     0.035     0.055      0.476    X
                 0.508     0.186      0.492     0.508     0.5        0.661    2
Weighted Avg.    0.54      0.317      0.454     0.54      0.483      0.611

=== Confusion Matrix ===

  a  b  c   <-- classified as
 90  7 13 |  a = 1
 37  2 18 |  b = X
 22  7 30 |  c = 2
```

Figura 64. Resultados One R - Conjunto Europa League WEKA

En términos del **estadístico kappa** que nos ofrece el estudio, el valor está algo por debajo la media del resto de clasificadores, situándose en 0,221, lo que se traduce en una débil correlación entre los atributos de la clase.

Si prestamos atención a las tasas de **true positives** que nos ofrece este estudio encontramos tasas aceptables para los dos tipos de victorias. Se ha conseguido una tasa de true positives del 81,8% para partidos que han sido clasificados como victoria del equipo local, mientras que tenemos una tasa del 50,8% para los partidos que han sido clasificados como victoria visitante. La tasa de true positives para los empates sigue la misma tendencia que con el resto de clasificadores, habiendo obtenido tan solo un 3,5% para los partidos que han acabado con este resultado.

Finalmente, si atendemos al atributo de **precisión** los resultados también son buenos, ya que tenemos una precisión para victorias locales del 60,4%. No tan buena aunque

aceptable, es la precisión de los partidos que acaban con victoria visitante que es del 49,2%. Por último, la precisión para partidos que acaban con el resultado de empate es del 12,5%.

J48:

El siguiente de los clasificadores en ser utilizado es el árbol J48. Este algoritmo de clasificación ha conseguido igualar en términos de tasa de aciertos a los resultados alcanzados por la Red Bayesiana (Figura 65). La **tasa de aciertos** que alcanza este algoritmo es del 56,19%.

```
Correctly Classified Instances         127               56.1947 %
Incorrectly Classified Instances        99               43.8053 %
Kappa statistic                          0.2689
Mean absolute error                      0.3658
Root mean squared error                  0.4576
Relative absolute error                 86.7086 %
Root relative squared error             99.5934 %
Total Number of Instances              226

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.791     0.466      0.617     0.791     0.693      0.614    1
                 0.123     0.112      0.269     0.123     0.169      0.383    X
                 0.559     0.156      0.559     0.559     0.559      0.714    2
Weighted Avg.    0.562     0.296      0.514     0.562     0.526      0.582

=== Confusion Matrix ===

  a  b  c   <-- classified as
 87 10 13 |  a = 1
 37  7 13 |  b = X
 17  9 33 |  c = 2
```

Figura 65. Resultados J48 - Conjunto Europa League WEKA

En cuanto al **estadístico kappa**, para este clasificador tenemos un resultado de 0.268, un valor algo por debajo del conseguido por la Red Bayesiana, aunque sigue mostrando una pequeña correlación entre los atributos del conjunto.

Si miramos las tasas de **true positives** que nos arroja el estudio podemos ver que para el resultado de victoria del equipo local se ha logrado obtener un 79,1%, mientras que para el resultado de victoria del equipo visitante tenemos un 55,9%. Al igual que en clasificadores anteriores, la tasa para el empate registra un bajo 12,3%.

Finalmente, si miramos el atributo de **precisión** podremos ver que para resultados en los que el equipo local ha resultado vencedor tenemos una precisión del 61,7%, mientras que para victorias del equipo visitante se ha llegado a alcanzar un 55,9% de precisión. Por último, si miramos la precisión de los encuentros que han acabado en empate podremos

ver que la precisión obtenida para este conjunto es del 26,9%, que podría considerarse un valor alto se compara con los datos de precisión obtenidos con otros algoritmos.

Random Forest:

Por último, y para concluir el entrenamiento del conjunto de entrenamiento que recopila los partidos de la Europa League se utilizará un Random Forest para entrenar el conjunto. Para este algoritmo se obtienen resultados algo peores que en el resto de algoritmos utilizados para este conjunto (Figura 66). En este caso la **tasa de acierto** asciende al 50,88%.

```
Correctly Classified Instances         115               50.885  %
Incorrectly Classified Instances       111               49.115  %
Kappa statistic                          0.2081
Mean absolute error                      0.3523
Root mean squared error                  0.4923
Relative absolute error                 83.6374 %
Root relative squared error            107.3011 %
Total Number of Instances              226

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.7       0.379      0.636     0.7       0.667      0.682    1
                 0.228     0.237      0.245     0.228     0.236      0.444    X
                 0.424     0.162      0.481     0.424     0.45       0.745    2
Weighted Avg.    0.509     0.287      0.497     0.509     0.502      0.639

=== Confusion Matrix ===

  a  b  c   <-- classified as
 77 21 12 |  a = 1
 29 13 15 |  b = X
 15 19 25 |  c = 2
```

Figura 66. Resultados Random Forest - Conjunto Europa League WEKA

Si se observa el **estadístico kappa**, se podrá ver que la correlación entre los atributos del conjunto es muy baja, lo que demuestra el valor 0,208. Como en ocasiones anteriores en las que se ha utilizado el Random Forest, un dato tan bajo del estadístico kappa podría indicar que los resultados obtenidos en los tests han sido acertados por mera casualidad, lo que resta puntos al algoritmo para ser elegido para ser implementado como modelo de predicción.

En cuanto a las tasas de **true positives**, tenemos una buena tasa para las victorias locales (70%) y unas tasas algo peores para las victorias visitantes (42,4%) y los empates (22,8%).

Por último, si miramos la **precisión** de cada uno de los resultados, tendremos porcentajes de precisión muy discretos. Para los partidos ganados por el equipo local tendremos un pobre 63,6% de precisión. En los partidos en los que gana el equipo visitante la precisión caerá hasta el 48,1%, mientras que para los partidos que acaban con

empate la precisión será del 24,5%. Como se puede ver, el Random Forest es posiblemente el mejor algoritmo para predecir los empates, como refleja su buena precisión en la clasificación de empates.

Resumen de Clasificadores para el conjunto de Europa League:

Una vez se ha entrenado el conjunto de entrenamiento con cada uno de los seis clasificadores, estamos en disposición de comparar los resultados para ver qué clasificadores se adaptan mejor a nuestro problema. En la Tabla 45 se recogen los resultados obtenidos con cada uno de los clasificadores.

	% Acierto	Kappa	Precisión Media
Red Bayesiana	56.19%	0.309	55.9%
Regresión Logística	55.30%	0.239	48.9%
Perceptrón Multicapa	55.30%	0.216	46.2%
OneR	53.98%	0.221	45.4%
J48	56.19%	0.268	51.4%
Random Forest	50.88%	0.208	49.7%

Tabla 45. Análisis del entrenamiento (Conjunto Europa League)

Tras entrenar al conjunto de partidos de la competición de la Europa League, se va a proceder al análisis de los datos que aparecen en la tabla superior para así poder ver qué clasificadores pueden ser escogidos para ser implementados en la siguiente fase del proyecto.

Una vez más, la Red Bayesiana se ha mostrado como el clasificador más fiable, en primer lugar porque ha obtenido el mejor porcentaje de aciertos y precisión media de los resultados clasificados. Además su buen valor en el estadístico kappa nos muestra que la correlación entre los atributos utilizados en el estudio es más que aceptable para tomar este algoritmo como modelo a implementar para realizar las predicciones.

En cuanto al segundo de los clasificadores que se va a elegir parar ser implementado en fases posteriores, en esta ocasión no hay tantas dudas como en ocasiones anteriores. En este caso se va a elegir al igual que en ocasiones anteriores al algoritmo J48, ya que tiene una tasa de aciertos igual que la de la Red Bayesiana. Además, este algoritmo logra una precisión media superior al 50% y tiene un valor para el estadístico kappa de 0,268, un valor muy por encima al que han conseguido el resto de algoritmos que han sido probados.

Una vez analizados ya tres conjuntos de partidos, se puede ver que la fiabilidad de los algoritmos de clasificación J48 y el que se basa en Redes Bayesianas son los que mejor se adaptan a las características del problema planteado. Por este motivo no nos extrañará el ver que estos dos algoritmos pueden ser también seleccionados en los entrenamientos del resto de conjuntos de partidos que faltan por analizar.

7.2.4 Conjunto de Partidos de Ligas:

A continuación se van a presentar los resultados obtenidos con los 6 clasificadores para el conjunto que recoge los partidos las competiciones de liga que han sido utilizadas para el estudio del problema.

Red Bayesiana (Bayes Net):

Para este conjunto que agrupa los partidos de diferentes ligas europeas se van a volver a evaluar los resultados del entrenamiento de los conjuntos con los seis clasificadores elegidos. El primero de los clasificadores que van a ser utilizados es la Red Bayesiana (Figura 67), que en este caso nos proporciona una **tasa de aciertos** muy buena del 55,59%.

```
Correctly Classified Instances         606               55.5963 %
Incorrectly Classified Instances       484               44.4037 %
Kappa statistic                          0.2345
Mean absolute error                      0.3511
Root mean squared error                  0.4461
Relative absolute error                 83.7626 %
Root relative squared error             97.4613 %
Total Number of Instances             1090

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.836     0.552      0.595     0.836     0.696      0.698    1
                 0.092     0.045      0.403     0.092     0.15       0.586    X
                 0.468     0.176      0.482     0.468     0.475      0.701    2
Weighted Avg.    0.556     0.328      0.518     0.556     0.503      0.671

=== Confusion Matrix ===

   a   b   c   <-- classified as
 449  15  73 |   a = 1
 177  25  69 |   b = X
 128  22 132 |   c = 2
```

Figura 67. Resultados Bayes Net - Conjunto Ligas WEKA

El **estadístico kappa** para este conjunto muestra un valor de 0,234, un valor algo bajo si lo comparamos con los resultados obtenidos con otros entrenamientos realizados con Redes Bayesianas. En cualquier caso hay que quedar a la espera del resto de resultados para poder evaluar lo bueno o malo que es este dato.

En cuanto a la tasa de **True Positives**, el estudio refleja muy buenos resultados, obteniendo un altísimo 83,6% para victorias locales, un 46,8% para victorias visitantes y un 9,2% para los empates.

Por último, se analizará la **precisión** de este modelo para cada uno de los resultados. Cabe destacar el buen resultado que se obtiene en la predicción de empates, que tiene una predicción del 40,3%, un valor muy alto si lo comparamos con el resto de estudios realizados en otras competiciones. La precisión en partidos clasificados como victoria

local es del 59,5%, mientras que en partidos clasificados como victoria visitante se tiene un 48,2% de precisión en la predicción.

Regresión Logística (Logistic):

El siguiente de los clasificadores que se va a utilizar en el entrenamiento de este conjunto será el clasificador basado en regresión logística (Figura 68). Si empezamos analizando el **porcentaje de aciertos**, este clasificador ha llegado a alcanzar una tasa de aciertos del 54,31%.

```
Correctly Classified Instances         592               54.3119 %
Incorrectly Classified Instances       498               45.6881 %
Kappa statistic                          0.1969
Mean absolute error                      0.3733
Root mean squared error                  0.4348
Relative absolute error                 89.0646 %
Root relative squared error             95.0016 %
Total Number of Instances             1090

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.862     0.624      0.573     0.862     0.688      0.7      1
                 0.129     0.09       0.321     0.129     0.184      0.597    X
                 0.333     0.098      0.543     0.333     0.413      0.71     2
Weighted Avg.    0.543     0.355      0.503     0.543     0.492      0.677

=== Confusion Matrix ===

   a   b   c   <-- classified as
 463  38  36 |   a = 1
 193  35  43 |   b = X
 152  36  94 |   c = 2
```

Figura 68. Resultados Logistic - Conjunto Ligas WEKA

En cuanto al **estadístico kappa**, ha reducido su valor respecto al del estudio anterior con la Red Bayesiana. Para este clasificador tenemos un valor de kappa de 0,196, lo que muestra que los atributos escogidos para este clasificador tienen poca correlación entre ellos.

En cuanto a las tasas de **true positives** que recoge la herramienta, se puedo observar que las tasas son muy buenas, sobre todo la tasa para partidos en los que el equipo local ha conseguido salir vencedor. Para este tipo de partidos la tasa asciende al 86,2%. Mientras tanto, la tasa para partidos en los que el equipo visitante sale vencedor la tasa es del 33,3%. Finalmente la tasa de true positives para partidos que acaban en empate es del 12,9. Cabe destacar la baja tasa de los partidos que acaban con victoria del equipo visitante.

Por último, los datos de **precisión** son buenos, y como ya ocurría con la Red Bayesiana, se han obtenido unos datos bastante buenos para la predicción de empates.

Para este tipo de resultados se ha obtenido una precisión del 32,1%. Mientras tanto la precisión en victorias locales es del 57,3% y la de victorias visitantes es del 54,3%.

Multilayer Perceptron (Perceptrón Multicapa):

El siguiente de los clasificadores en ser utilizado ha sido el Perceptrón Multicapa (Figura 69). En este caso los resultados obtenidos no han sido tan buenos como los obtenidos con los otros dos clasificadores. Para este caso, la **tasa de aciertos** se ha situado en un 51,28%.

```
Correctly Classified Instances         559               51.2844 %
Incorrectly Classified Instances       531               48.7156 %
Kappa statistic                          0.1922
Mean absolute error                      0.3831
Root mean squared error                  0.4421
Relative absolute error                 91.4156 %
Root relative squared error             96.5896 %
Total Number of Instances             1090

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.719     0.483      0.591     0.719     0.649      0.684    1
                 0.24      0.178      0.308     0.24      0.27       0.575    X
                 0.383     0.146      0.478     0.383     0.425      0.692    2
Weighted Avg.    0.513     0.32       0.491     0.513     0.497      0.659

=== Confusion Matrix ===

   a   b   c   <-- classified as
 386  86  65 |   a = 1
 153  65  53 |   b = X
 114  60 108 |   c = 2
```

Figura 69. Resultados Multilayer Perceptron - Conjunto Ligas WEKA

En cuanto al **estadístico kappa**, el valor que se ha obtenido en este estudio no es del todo bueno, ya que el estadístico toma el valor 0,192, lo que indica una débil correlación entre los atributos del conjunto entrenado con el perceptrón multicapa.

En cuanto a las tasas de **true positives** tenemos tasas inferiores a las conseguidas por los modelos ya utilizados para entrenar este conjunto. En concreto tendremos una tasa del 71,9% en el caso de partidos que han acabado con victoria local, un 38,3% para partidos que finalizaron con victoria visitante y un 24% para partidos que concluyeron en empate.

Por último queda analizar el parámetro de **precisión**. Para los partidos clasificados como victoria del equipo local tenemos un 59,1% de precisión, mientras que para partidos con victoria visitante la precisión se queda en un 47,8%. Para partidos cuyo resultado ha sido de empate la precisión será del 30,8%.

Se puede observar que de nuevo la precisión sobre los empates ha subido en relación a otros conjuntos de entrenamiento en detrimento de la precisión obtenida por

predicciones de victoria del equipo visitante. Esto puede deberse a que al tener un conjunto de entrenamiento mayor, los algoritmos son capaces de sacar mejores conclusiones para posteriormente aplicarlas a la predicción de los partidos.

OneR:

Los datos mostrados en la Figura 70, corresponden a los resultados del entrenamiento con el clasificador OneR aplicado al conjunto de partido de ligas. El **porcentaje de aciertos** que ha conseguido ha sido del 52,75%.

```
Correctly Classified Instances         575               52.7523 %
Incorrectly Classified Instances       515               47.2477 %
Kappa statistic                          0.1805
Mean absolute error                      0.315
Root mean squared error                  0.5612
Relative absolute error                 75.1548 %
Root relative squared error            122.6131 %
Total Number of Instances             1090

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.819     0.6        0.57      0.819     0.672      0.61     1
                 0.111     0.088      0.294     0.111     0.161      0.511    X
                 0.372     0.137      0.486     0.372     0.422      0.617    2
Weighted Avg.    0.528     0.353      0.48      0.528     0.48       0.587

=== Confusion Matrix ===

   a   b   c   <-- classified as
 440  41  56 |   a = 1
 186  30  55 |   b = X
 146  31 105 |   c = 2
```

Figura 70. Resultados One R - Conjunto Ligas WEKA

En términos del **estadístico kappa**, de nuevo nos vuelve a ofrecer un valor muy bajo. En este caso el valor del estadístico es del 0,180, lo que muestra que la correlación entre los atributos utilizados en el conjunto de entrenamiento es muy débil.

Si prestamos atención a las tasas de **true positives** que nos ofrece este estudio encontramos tasas aceptables tan sólo para victorias del equipo local. Se ha conseguido una tasa de true positives del 81,9% para partidos que han sido clasificados como victoria del equipo local. Mientras tanto, la tasa en partidos que finalizan con victoria visitante baja hasta el 37,2% y la tasa en partidos empatado sólo llega al 11,1%. Estos dos últimos valores son muy bajos en comparación de los obtenidos por otros clasificadores.

Finalmente, si atendemos al atributo de **precisión** los resultados son relativamente buenos, ya que tenemos una precisión para victorias locales del 57%. No tan buena aunque aceptable, es la precisión de los partidos clasificados como victoria visitante que es del 49,6% y de los partidos clasificados como empate, que es del 29,4%

J48:

El siguiente de los clasificadores en ser utilizado es el árbol J48 (Figura 71). Este algoritmo de clasificación ha conseguido unos resultados muy buenos, como ya había demostrado al entrenar otros conjuntos. La **tasa de aciertos** que alcanza este algoritmo es del 56,05%.

```
Correctly Classified Instances         611               56.055  %
Incorrectly Classified Instances       479               43.945  %
Kappa statistic                          0.2246
Mean absolute error                      0.3869
Root mean squared error                  0.4445
Relative absolute error                 92.3061 %
Root relative squared error             97.1076 %
Total Number of Instances             1090

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.885     0.617      0.582     0.885     0.702      0.62     1
                 0.092     0.046      0.397     0.092     0.15       0.501    X
                 0.394     0.124      0.526     0.394     0.45       0.647    2
Weighted Avg.    0.561     0.347      0.522     0.561     0.5        0.597

=== Confusion Matrix ===

   a   b   c   <-- classified as
 475  14  48 |   a = 1
 194  25  52 |   b = X
 147  24 111 |   c = 2
```

Figura 71. Resultados J48 - Conjunto Ligas WEKA

En cuanto al **estadístico kappa**, para este clasificador tenemos un resultado de 0,224, un valor algo por debajo del conseguido por la Red Bayesiana, aunque sigue mostrando una pequeña correlación entre los atributos del conjunto.

Si miramos las tasas de **true positives** que nos arroja el estudio podemos ver que para el resultado de victoria del equipo local se ha logrado obtener un 88,5%, mientras que para el resultado de victoria del equipo visitante tenemos un 39,4%. Al igual que pudo verse en clasificadores y conjuntos anteriores, la tasa para el empate registra un bajo 9,2%.

Finalmente, si miramos el atributo de **precisión** podremos ver que para resultados en los que el equipo local ha resultado vencedor tenemos una precisión del 58,2%, mientras que para victorias del equipo visitante se ha llegado a alcanzar un 52,6% de precisión. Por último, si miramos la precisión de los encuentros que han acabado en empate podremos ver que la precisión obtenida para este conjunto es del 39,7%, un dato que podría considerarse bastante bueno si tenemos en cuenta que el empate es el resultado que menos frecuencia tiene en este tipo de competición.

Random Forest:

Por último, y para concluir el entrenamiento del conjunto de entrenamiento que recopila los partidos de la competición de Liga, se utilizará un Random Forest para entrenar el conjunto (Figura 72). Para este algoritmo se obtienen bastante peores si los comparamos con los resultados obtenidos con los demás clasificadores. En este caso la **tasa de acierto** asciende al 45,68%.

```
Correctly Classified Instances         498               45.6881 %
Incorrectly Classified Instances       592               54.3119 %
Kappa statistic                          0.1346
Mean absolute error                      0.3698
Root mean squared error                  0.5256
Relative absolute error                 88.2439 %
Root relative squared error            114.8365 %
Total Number of Instances             1090

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.588     0.409      0.583     0.588     0.586      0.628    1
                 0.306     0.245      0.292     0.306     0.299      0.55     X
                 0.351     0.204      0.375     0.351     0.363      0.61     2
Weighted Avg.    0.457     0.315      0.457     0.457     0.457      0.604

=== Confusion Matrix ===

   a   b   c   <-- classified as
 316 124  97 |   a = 1
 120  83  68 |   b = X
 106  77  99 |   c = 2
```

Figura 72. Resultados Random Forest - Conjunto Ligas WEKA

Si se observa el **estadístico kappa**, se podrá ver que la correlación entre los atributos del conjunto es muy baja, lo que demuestra el valor 0,134. Como en ocasiones anteriores en las que se ha utilizado el Random Forest, un dato tan bajo del estadístico kappa podría indicar que los resultados obtenidos en los tests han sido acertados por mera casualidad, lo que resta puntos al algoritmo para ser elegido para ser implementado como modelo de predicción.

En cuanto a las tasas de **true positives**, tenemos tasas, muy bajas para los tres tipo de resultados. Para las victorias locales la tasa es del 58,8% y tendremos unas tasas algo peores para las victorias visitantes (35,1%) y los empates (30,6%).

Por último, si miramos la **precisión** de cada uno de los resultados, tendremos porcentajes de precisión muy discretos. Para los partidos ganados por el equipo local tendremos un pobre 58,3% de precisión. En los partidos en los que gana el equipo visitante la precisión caerá hasta el 37,5%, mientras que para los partidos que acaban con empate la precisión será del 29,2%.

Resumen de Clasificadores para el conjunto de Partidos de Liga:

Una vez se ha entrenado el conjunto de entrenamiento con cada uno de los seis clasificadores, estamos en disposición de comparar los resultados para ver qué clasificadores se adaptan mejor a nuestro problema. En la Tabla 46 se recogen los resultados obtenidos con cada uno de los clasificadores:

	% Acierto	Kappa	Precisión Media
Red Bayesiana	55.59%	0.234	51.8%
Regresión Logística	54.31%	0.196	50.3%
Perceptrón Multicapa	51.28%	0.192	49.1%
OneR	52.75%	0.180	48.0%
J48	56.05%	0.224	52.2%
Random Forest	46.68%	0.134	45.7%

Tabla 46. Análisis del entrenamiento (Conjunto Ligas)

Una vez más, y ya van cuatro, los dos algoritmos que mejores resultados han ofrecido al entrenar el conjunto de entrenamiento con los partidos de liga han sido la Red Bayesiana y el árbol J48.

En este caso ha sido el árbol J48 el que nos ofrece unos mejores resultados en términos de tasa de aciertos y precisión media, donde ha llegado a obtener un 56,05% y 52,2% respectivamente. Además la correlación del conjunto de atributos es aceptable como para aceptar que los resultados acertados no han sido fruto de la casualidad.

Para el caso de la Red Bayesiana los resultados obtenidos también han sido muy buenos. En este caso se han obtenido un 55,59% de tasa de aciertos, un valor muy cercano al del árbol J48. En cuanto a la precisión, este algoritmo ha obtenido una precisión media del 51,8%, valor muy cercano también al del algoritmo J48. Mientras tanto, es la correlación del conjunto de atributos donde la Red Bayesiana bate al resto de algoritmos. Aunque el resultado obtenido por el estadístico kappa no es extremadamente alto, puede ser aceptado para poder realizar las predicciones de los partidos.

Por último, cabe destacar la buena capacidad de estos dos algoritmos para predecir empates, lo que muestra los valores cercanos al 33% en la precisión de la predicción de empates, un valor muy por encima del conseguido en otro tipo de competiciones.

7.2.5 Conjunto de Partidos de la Liga BBVA:

A continuación se van a presentar los resultados obtenidos con los seis clasificadores para el conjunto que recoge los partidos la primera división española, también conocida como la Liga BBVA:

Red Bayesiana (Bayes Net):

Para el conjunto de partidos de la Liga BBVA el primero de los clasificadores que van a ser utilizados es la Red Bayesiana (Figura 73), que en este caso nos proporciona una **tasa de aciertos** muy buena del 60,40%.

```
Correctly Classified Instances         180               60.4027 %
Incorrectly Classified Instances       118               39.5973 %
Kappa statistic                          0.213
Mean absolute error                      0.3241
Root mean squared error                  0.4265
Relative absolute error                 82.068  %
Root relative squared error             95.981  %
Total Number of Instances              298

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.903     0.684      0.621     0.903     0.736      0.646    1
                 0         0          0         0         0          0.549    X
                 0.397     0.123      0.534     0.397     0.456      0.621    2
Weighted Avg.    0.604     0.411      0.484     0.604     0.527      0.621

=== Confusion Matrix ===

   a   b   c   <-- classified as
 149   0  16 |   a = 1
  44   0  11 |   b = X
  47   0  31 |   c = 2
```

Figura 73. Resultados Bayes Net - Conjunto Liga BBVA WEKA

El **estadístico kappa** para este conjunto muestra un valor de 0,213, un valor algo bajo si lo comparamos con los resultados obtenidos con otros entrenamientos realizados con Redes Bayesianas.

En cuanto a la tasa de **True Positives**, el estudio refleja muy buenos resultados, obteniendo un altísimo 90,3% para victorias locales y un 39,7% para victorias visitantes. Ante la imposibilidad de la red de predecir empates, los datos reflejan un 0% para partidos que han concluido en empate.

Por último, se analizará la **precisión** de este modelo para cada uno de los resultados. Destacar el buen dato de precisión para las predicciones de victorias locales, que alcanzan una precisión del 62,1%, mientras que para las victorias visitantes la tasa se queda en un 53,4%. Al no haber realizado ninguna predicción de empate no hay dato de precisión para los partidos clasificados como empate.

Regresión Logística (Logistic):

El siguiente de los clasificadores que se va a utilizar en el entrenamiento de este conjunto será el clasificador basado en regresión logística (Figura 74). Si empezamos analizando el **porcentaje de aciertos**, este clasificador ha llegado a alcanzar una tasa de

aciertos del 55,36%, un porcentaje cinco puntos inferior al conseguido por la Red Bayesiana.

```
Correctly Classified Instances         165               55.3691 %
Incorrectly Classified Instances       133               44.6309 %
Kappa statistic                          0.1445
Mean absolute error                      0.3646
Root mean squared error                  0.4435
Relative absolute error                 92.4073 %
Root relative squared error             99.9259 %
Total Number of Instances              298

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.83      0.662      0.609     0.83      0.703      0.638    1
                 0.036     0.053      0.133     0.036     0.057      0.517    X
                 0.333     0.145      0.448     0.333     0.382      0.6      2
Weighted Avg.    0.554     0.414      0.479     0.554     0.5        0.606

=== Confusion Matrix ===

   a   b   c   <-- classified as
 137   9  19 |   a = 1
  40   2  13 |   b = X
  48   4  26 |   c = 2
```

Figura 74. Resultados Logistic - Conjunto Liga BBVA WEKA

En cuanto al **estadístico kappa**, ha reducido su valor respecto al del estudio anterior con la Red Bayesiana. Para este clasificador tenemos un valor de kappa de 0,144, lo que muestra que los atributos escogidos para este clasificador tienen muy poca correlación entre ellos.

En cuanto a las tasas de **true positives** que recoge la herramienta, se puedo observar que las tasas son muy buenas, sobre todo la tasa para partidos en los que el equipo local ha conseguido salir vencedor. Para este tipo de partidos la tasa asciende al 83%. Mientras tanto, la tasa para partidos en los que el equipo visitante sale vencedor la tasa es del 33,3%. Finalmente la tasa de true positives para partidos que acaban en empate es del 3,6%.

Por último, los datos de **precisión** son buenos, obteniéndose una precisión en las victorias locales del 60,9%. La precisión para las predicciones de victoria del equipo visitante es algo más baja y se sitúa en el 44,8%, mientras que para las predicciones de empate es del 13,3%.

Este modelo es capaz de predecir algunos empates, pero su baja correlación entre atributos del conjunto hace sospechar que los aciertos han podido ser simple azar.

Multilayer Perceptron (Perceptrón Multicapa):

El siguiente de los clasificadores en ser utilizado ha sido el Perceptrón Multicapa (Figura 75). Los resultados obtenidos son bastante similares a los que ofrecía el algoritmo de regresión logística. Para este caso, la **tasa de aciertos** se ha situado en un 55,70%, tasa ligeramente superior al algoritmo de regresión logística.

```
Correctly Classified Instances         166               55.7047 %
Incorrectly Classified Instances       132               44.2953 %
Kappa statistic                          0.1805
Mean absolute error                      0.3544
Root mean squared error                  0.4803
Relative absolute error                 89.7383 %
Root relative squared error            108.2184 %
Total Number of Instances              298

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.8       0.579      0.632     0.8       0.706      0.65     1
                 0.145     0.078      0.296     0.145     0.195      0.548    X
                 0.333     0.164      0.419     0.333     0.371      0.59     2
Weighted Avg.    0.557     0.378      0.514     0.557     0.524      0.615

=== Confusion Matrix ===

   a   b   c   <-- classified as
 132  10  23 |   a = 1
  34   8  13 |   b = X
  43   9  26 |   c = 2
```

Figura 75. Resultados Multilayer Perceptron - Conjunto Liga BBVA WEKA

En cuanto al **estadístico kappa**, el valor que se ha obtenido en este estudio no es del todo bueno, ya que el estadístico toma el valor 0,180, lo que indica una débil correlación entre los atributos del conjunto entrenado con el perceptrón multicapa.

En cuanto a las tasas de **true positives** tenemos tasas similares a las conseguidas por los modelos ya utilizados para entrenar este conjunto. En concreto tendremos una tasa del 80% en el caso de partidos que han acabado con victoria local, un 33,3% para partidos que finalizaron con victoria visitante y un bajo 14,5% para partidos que concluyeron en empate.

Por último queda analizar el parámetro de **precisión**. Para los partidos clasificados como victoria del equipo local tenemos un buen 63,2% de precisión, mientras que para partidos con victoria visitante la precisión se queda en un 41,9%. Para partidos cuyo resultado ha sido de empate la precisión será del 29,6%.

La baja correlación entre los atributos del conjunto y los pocos empates que ha podido predecir el modelo hace que quede prácticamente descartado para ser implementado en las siguientes fases del proceso y forme parte del sistema de predicción de resultados para la competición de la Liga BBVA.

OneR:

Los datos mostrados en la Figura 76 corresponden a los resultados del entrenamiento con el clasificador OneR aplicado al conjunto de partido de la Liga BBVA. El **porcentaje de aciertos** que ha conseguido ha sido del 58,38%.

```
Correctly Classified Instances         174               58.3893 %
Incorrectly Classified Instances       124               41.6107 %
Kappa statistic                          0.1632
Mean absolute error                      0.2774
Root mean squared error                  0.5267
Relative absolute error                 70.3201 %
Root relative squared error            118.6646 %
Total Number of Instances              298

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.891     0.744      0.598     0.891     0.715      0.573    1
                 0         0          0         0         0          0.5      X
                 0.346     0.114      0.519     0.346     0.415      0.616    2
Weighted Avg.    0.584     0.442      0.467     0.584     0.505      0.571

=== Confusion Matrix ===

   a   b   c   <-- classified as
 147   0  18 |   a = 1
  48   0   7 |   b = X
  51   0  27 |   c = 2
```

Figura 76. Resultados One R - Conjunto Liga BBVA WEKA

En términos del **estadístico kappa**, de nuevo nos vuelve a ofrecer un valor bajísimo. En este caso el valor del estadístico es del 0,163, lo que muestra que la correlación entre los atributos utilizados en el conjunto de entrenamiento es muy débil.

Si prestamos atención a las tasas de **true positives** que nos ofrece este estudio encontramos tasas aceptables tan sólo para victorias del equipo local. Se ha conseguido una tasa de true positives del 89,1% para partidos que han sido clasificados como victoria del equipo local. Mientras tanto, la tasa en partidos que finalizan con victoria visitante baja hasta el 34,6%. Al no ser capaz de predecir ningún empate, como puede verse en la matriz de confusión, la tasa de true positives será del 0%.

Finalmente, si atendemos al atributo de **precisión** los resultados son relativamente buenos, ya que tenemos una precisión para victorias locales del 59,8%. Para el caso de predicciones de victoria visitante tendremos una precisión del 51,9%. Al no ser el algoritmo capaz de predecir empates no se tiene dato de precisión para este tipo de resultado.

J48:

El siguiente de los clasificadores en ser utilizado es el árbol J48 (Figura 77). Este algoritmo de clasificación ha conseguido de nuevo unos resultados muy buenos, como ya había ocurrido con conjuntos anteriores. La **tasa de aciertos** que alcanza este algoritmo es del 59,06%.

```
Correctly Classified Instances         176               59.0604 %
Incorrectly Classified Instances       122               40.9396 %
Kappa statistic                          0.1765
Mean absolute error                      0.3552
Root mean squared error                  0.4449
Relative absolute error                 90.0331 %
Root relative squared error            100.2321 %
Total Number of Instances              298

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.903     0.752      0.598     0.903      0.72       0.58     1
                 0.036     0.037      0.182     0.036      0.061      0.54     X
                 0.321     0.059      0.658     0.321      0.431      0.601    2
Weighted Avg.    0.591     0.439      0.537     0.591      0.523      0.578

=== Confusion Matrix ===

   a   b   c   <-- classified as
 149   5  11 |   a = 1
  51   2   2 |   b = X
  49   4  25 |   c = 2
```

Figura 77. Resultados J48 - Conjunto Liga BBVA WEKA

En cuanto al **estadístico kappa**, para este clasificador tenemos un resultado de 0,176, un valor algo por debajo del conseguido por la Red Bayesiana, aunque sigue mostrando una pequeña correlación entre los atributos del conjunto.

Si miramos las tasas de **true positives** que nos arroja el estudio podemos ver que para el resultado de victoria del equipo local se ha logrado obtener un 90,3%, mientras que para el resultado de victoria del equipo visitante tenemos un 32,1%. Al igual que pudo verse en clasificadores y conjuntos anteriores, la tasa para el empate registra un muy bajo 3,6%.

Finalmente, si miramos el atributo de **precisión** podremos ver que el algoritmo ha sido capaz de obtener muy buenos resultados. Concretamente ha conseguido obtener un 59,8% de precisión en predicciones de victoria del equipo local y un sorprendente 65,8% de precisión en predicciones de victoria del equipo visitante. Mientras tanto, la precisión de los empates tan solo puede alcanzar un 18,2%.

Los buenos datos de precisión para predicciones de victoria visitante colocan a este algoritmo como un claro candidato a ser implementado para el sistema de predicción de resultados de la Liga BBVA.

Random Forest:

Por último, y para concluir el entrenamiento del conjunto de entrenamiento que recopila los partidos de la competición de Liga BBVA, se utilizará un Random Forest para entrenar el conjunto. Para este algoritmo se obtienen bastante peores si los comparamos con los resultados obtenidos con los demás clasificadores. En este caso la **tasa de acierto** asciende al 50,33% (Figura 78).

```
Correctly Classified Instances         150               50.3356 %
Incorrectly Classified Instances       148               49.6644 %
Kappa statistic                          0.1494
Mean absolute error                      0.3585
Root mean squared error                  0.4956
Relative absolute error                 90.8796 %
Root relative squared error            111.6532 %
Total Number of Instances              298

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.642    0.481      0.624     0.642     0.633      0.606    1
                 0.2      0.148      0.234     0.2       0.216      0.535    X
                 0.423    0.218      0.407     0.423     0.415      0.588    2
Weighted Avg.    0.503    0.351      0.495     0.503     0.499      0.588

=== Confusion Matrix ===

   a   b   c   <-- classified as
 106  24  35 |   a = 1
  31  11  13 |   b = X
  33  12  33 |   c = 2
```

Figura 78. Resultados Random Forest - Conjunto Liga BBVA WEKA

Si se observa el **estadístico kappa**, se podrá ver que la correlación entre los atributos del conjunto es muy baja, lo que demuestra el valor 0,149. Como en ocasiones anteriores en las que se ha utilizado el Random Forest, un dato tan bajo del estadístico kappa podría indicar que los resultados obtenidos en los tests han sido acertados por mera casualidad, pero al tener el resto de algoritmos un valor del estadístico también bajo, no se dará tanta importancia como en ocasiones anteriores.

En cuanto a las tasas de **true positives**, tenemos tasas, muy bajas para los tres tipo de resultados. Para las victorias locales la tasa es del 64,2% y tendremos unas tasas algo peores para las victorias visitantes (42,3%) y los empates (20%).

Por último, si miramos la **precisión** de cada uno de los resultados, tendremos porcentajes de precisión relativamente buenos. Para los partidos ganados por el equipo local tendremos un 62,4% de precisión. En los partidos en los que se predice victoria del

equipo visitante la precisión caerá hasta el 40,7%, mientras que para los partidos que acaban con empate la precisión será del 23,4%.

Resumen de Clasificadores para el conjunto de Liga BBVA:

Una vez se ha entrenado el conjunto de entrenamiento con cada uno de los 6 clasificadores, estamos en disposición de comparar los resultados para ver qué clasificadores se adaptan mejor a nuestro problema. En la Tabla 47 se recogen los resultados obtenidos con cada uno de los clasificadores:

	% Acierto	Kappa	Precisión Media
Red Bayesiana	60.40%	0.213	48.4%
Regresión Logística	55.36%	0.144	47.9%
Perceptrón Multicapa	55.70%	0.180	51.4%
OneR	58.38%	0.163	46.7%
J48	59.06%	0.176	53.7%
Random Forest	50.33%	0.149	49.5%

Tabla 47. Análisis del entrenamiento (Conjunto Liga BBVA)

Para el conjunto de partidos de la Liga BBVA se escogerán otros dos algoritmos que sirvan para crear el sistema de predicción de los partidos de esta competición. Fijándonos en los datos de la tabla resumen de la parte superior los algoritmos seleccionados serán los siguientes:

En primer lugar se escogerá a la Red Bayesiana, ya que a pesar de no poder predecir empates ha logrado un 60,40% de acierto, el valor más alto conseguido hasta ahora en todos los conjuntos entrenados. Las carencias que presenta este clasificador a la hora de predecir empates se intentarán subsanar en posteriores fases del proceso, a través de estrategias de predicción que serán aplicadas en la fase de explotación del sistema.

El segundo de los clasificadores seleccionado será de nuevo el árbol J48, ya que además de tener el segundo porcentaje de acierto más alto tiene la mejor media de precisión (53,7%). Además, esta media de precisión destaca por su alta precisión en predicciones de partidos ganados por el equipo visitante, donde la precisión del modelo llega a alcanzar el 65,8%. Al igual que ocurría con la Red Bayesiana, las carencias que tiene este modelo en la predicción de empates, intentarán ser subsanadas en la fase de explotación del sistema, donde se intentará cubrir de alguna manera el resultado de empate.

Por tanto, de nuevo serán la Red Bayesiana y el algoritmo J48 los algoritmos que serán implementados para desarrollar el sistema de predicción de partidos de la Liga BBVA.

7.2.6 Conjunto de Partidos de la Liga Adelante:

A continuación se van a presentar los resultados obtenidos con los seis clasificadores para el conjunto que recoge los partidos la segunda división española, también conocida como la Liga Adelante:

Red Bayesiana (Bayes Net):

Para el conjunto de partidos de la Liga Adelante el primero de los clasificadores que van a ser utilizados es la Red Bayesiana, que en este caso nos proporciona una **tasa de aciertos** muy discreta del 50,73% (Figura 79). Este valor tan bajo puede ser debido a que esta competición tiene una distribución de resultados algo caótico, ya que los equipos mejor clasificados pierden en algunas ocasiones contra equipos de la zona baja de la tabla. Esto puede hacer que no se encuentre correlación entre los datos de los partidos y el resultado final de estos.

```
Correctly Classified Instances         103               50.7389 %
Incorrectly Classified Instances       100               49.2611 %
Kappa statistic                          0.0445
Mean absolute error                      0.397
Root mean squared error                  0.4465
Relative absolute error                 97.4429 %
Root relative squared error             99.0121 %
Total Number of Instances              203

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.869     0.833      0.538     0.869     0.664      0.589    1
                 0.137     0.066      0.412     0.137     0.206      0.645    X
                 0.067     0.063      0.231     0.067     0.103      0.579    2
Weighted Avg.    0.507     0.47       0.438     0.507     0.425      0.601

=== Confusion Matrix ===

  a  b  c   <-- classified as
 93  7  7 |  a = 1
 41  7  3 |  b = X
 39  3  3 |  c = 2
```

Figura 79. Resultados Bayes Net - Conjunto Liga Adelante WEKA

El **estadístico kappa** para este conjunto muestra un bajísimo valor de 0,044, que puede ser debido a las razones expuestas en el punto anterior.

En cuanto a la tasa de **True Positives**, el estudio refleja buenos resultados sólo para resultados en los que el equipo local sale vencedor. Concretamente se tendrá un 86,9% para victorias locales, un pobre 6,7% para victorias visitantes y un 13,7% para empates.

Por último, se analizará la **precisión** de este modelo para cada uno de los resultados, donde se tiene una precisión del 53,8% para predicciones de victoria del equipo local, un 23,1% para victorias del equipo visitante y un buen 41,2% para empates.

Regresión Logística (Logistic):

El siguiente de los clasificadores que se va a utilizar en el entrenamiento de este conjunto será el clasificador basado en regresión logística. Si empezamos analizando el **porcentaje de aciertos**, este clasificador ha llegado a alcanzar una tasa de aciertos del 48,76% (Figura 80), lo que nos hace ver que los resultados para este conjunto de Liga Adelante van a ser bastante peores debido a la distribución caótica que sufren los partidos de esta competición.

```
Correctly Classified Instances          99               48.7685 %
Incorrectly Classified Instances       104               51.2315 %
Kappa statistic                          0.0558
Mean absolute error                      0.392
Root mean squared error                  0.4575
Relative absolute error                 96.2487 %
Root relative squared error            101.4546 %
Total Number of Instances              203

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.776     0.729      0.542     0.776     0.638      0.586    1
                 0.294     0.151      0.395     0.294     0.337      0.629    X
                 0.022     0.07       0.083     0.022     0.035      0.475    2
Weighted Avg.    0.488     0.438      0.404     0.488     0.429      0.572

=== Confusion Matrix ===

  a  b  c   <-- classified as
 83 16  8 |  a = 1
 33 15  3 |  b = X
 37  7  1 |  c = 2
```

Figura 80. Resultados Logistic - Conjunto Liga Adelante WEKA

En cuanto al **estadístico kappa**, ha reducido su valor respecto al del estudio anterior con la Red Bayesiana. Para este clasificador tenemos un valor de kappa de 0,055, lo que muestra que los atributos escogidos para este clasificador tienen una correlación casi nula.

En cuanto a las tasas de **true positives** que recoge la herramienta, se puedo observar que las tasas son discretas, sobre todo la tasa para partidos en los que el equipo visitante ha conseguido salir vencedor. Para este tipo de partidos la tasa asciende al 2,2%. Mientras tanto, la tasa para partidos en los que el equipo local sale vencedor la tasa es del 77,6%. Finalmente la tasa de true positives para partidos que acaban en empate es del 29,4%.

Por último, los datos de **precisión** son buenos para los casos de victoria local y empate, obteniéndose una precisión en las victorias locales del 54,2%. La precisión para las predicciones de victoria del equipo visitante es algo más baja y se sitúa en el 8,3%, mientras que para las predicciones de empate es del 39,5%.

Multilayer Perceptron (Perceptrón Multicapa):

El siguiente de los clasificadores en ser utilizado ha sido el Perceptrón Multicapa. Los resultados obtenidos siguen la tónica que los clasificadores utilizados anteriormente. Para este caso, la **tasa de aciertos** se ha situado en un 44,82% (Figura 81), tasa muy baja que confirma las teorías sobre la distribución de los resultados de esta competición.

```
Correctly Classified Instances          91               44.8276 %
Incorrectly Classified Instances       112               55.1724 %
Kappa statistic                          0.0712
Mean absolute error                      0.4002
Root mean squared error                  0.4687
Relative absolute error                 98.2447 %
Root relative squared error            103.9386 %
Total Number of Instances              203

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.645     0.5        0.59      0.645     0.616      0.56     1
                 0.353     0.257      0.316     0.353     0.333      0.605    X
                 0.089     0.158      0.138     0.089     0.108      0.425    2
Weighted Avg.    0.448     0.363      0.421     0.448     0.432      0.542

=== Confusion Matrix ===

  a  b  c   <-- classified as
 69 24 14 |  a = 1
 22 18 11 |  b = X
 26 15  4 |  c = 2
```

Figura 81. Resultados Multilayer Perceptron - Conjunto Liga Adelante WEKA

En cuanto al **estadístico kappa**, el valor que se ha obtenido en este estudio no nada bueno, ya que el estadístico toma el valor 0,071, lo que indica una correlación casi nula entre los atributos del conjunto entrenado con el perceptrón multicapa.

En cuanto a las tasas de **true positives** tenemos tasas similares a las conseguidas por los modelos ya utilizados para entrenar este conjunto. En concreto tendremos una tasa del 64,5% en el caso de partidos que han acabado con victoria local, un bajo 8,9% para partidos que finalizaron con victoria visitante y un 35,3% para partidos que concluyeron en empate.

Por último queda analizar el parámetro de **precisión**. Para los partidos clasificados como victoria del equipo local tenemos un 59% de precisión, mientras que para partidos con victoria visitante la precisión se queda en un pobre 13,8%. Para partidos cuyo resultado ha sido de empate la precisión será del 31,6%.

OneR:

Los datos mostrados en la parte superior corresponden a los resultados del entrenamiento con el clasificador OneR aplicado al conjunto de partidos de la Liga Adelante. El **porcentaje de aciertos** que ha conseguido ha sido del 49,26% (Figura 82).

```
Correctly Classified Instances         100               49.2611 %
Incorrectly Classified Instances       103               50.7389 %
Kappa statistic                          0.0712
Mean absolute error                      0.3383
Root mean squared error                  0.5816
Relative absolute error                 83.0433 %
Root relative squared error            128.9414 %
Total Number of Instances              203

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.832     0.656      0.586     0.832     0.687      0.588    1
                 0.078     0.125      0.174     0.078     0.108      0.477    X
                 0.156     0.133      0.25      0.156     0.192      0.511    2
Weighted Avg.    0.493     0.407      0.408     0.493     0.432      0.543

=== Confusion Matrix ===

  a  b  c   <-- classified as
 89 10  8 |  a = 1
 34  4 13 |  b = X
 29  9  7 |  c = 2
```

Figura 82. Resultados One R - Conjunto Liga Adelante WEKA

En términos del **estadístico kappa**, de nuevo nos vuelve a ofrecer un valor bajísimo. En este caso el valor del estadístico es del 0,071, lo que muestra que la correlación entre los atributos utilizados en el conjunto de entrenamiento es prácticamente nula.

Si prestamos atención a las tasas de **true positives** que nos ofrece este estudio encontramos tasas aceptables tan sólo para victorias del equipo local. Se ha conseguido una tasa de true positives del 83,2% para partidos que han sido clasificados como victoria del equipo local. Mientras tanto, la tasa en partidos que finalizan con victoria visitante baja hasta el 15,6%. Finalmente la tasa de true positives para empates se sitúa en el 7,8%

Finalmente, si atendemos al atributo de **precisión** los resultados son relativamente buenos, ya que tenemos una precisión para victorias locales del 58,6%. Para el caso de predicciones de victoria visitante tendremos una precisión del 25%. Por último, la precisión de este modelo para predecir empates es del 17,4%.

J48:

El siguiente de los clasificadores en ser utilizado es el árbol J48. Este algoritmo de clasificación ha conseguido de nuevo unos resultados muy buenos dadas las características de este conjunto. La **tasa de aciertos** que alcanza este algoritmo es del 50,73% (Figura 83).

```
Correctly Classified Instances         103               50.7389 %
Incorrectly Classified Instances       100               49.2611 %
Kappa statistic                          0.139
Mean absolute error                      0.386
Root mean squared error                  0.5024
Relative absolute error                 94.7385 %
Root relative squared error            111.3647 %
Total Number of Instances              203

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.729     0.563      0.591     0.729     0.653      0.552    1
                 0.412     0.204      0.404     0.412     0.408      0.59     X
                 0.089     0.095      0.211     0.089     0.125      0.383    2
Weighted Avg.    0.507     0.369      0.46      0.507     0.474      0.524

=== Confusion Matrix ===

  a  b  c   <-- classified as
 78 20  9 |  a = 1
 24 21  6 |  b = X
 30 11  4 |  c = 2
```

Figura 83. Resultados J48 - Conjunto Liga Adelante WEKA

En cuanto al **estadístico kappa**, para este clasificador tenemos un resultado de 0,139, el valor más alto obtenido por cualquiera de los clasificadores que ha entrenado este conjunto.

Si miramos las tasas de **true positives** que nos arroja el estudio podemos ver que para el resultado de victoria del equipo local se ha logrado obtener un 72,9%, mientras que para el resultado de victoria del equipo visitante tenemos un pobre 8,9%. Mientras tanto, la tasa de true positives para empates ha quedado registrada en un 41,2%.

Finalmente, si miramos el atributo de **precisión** podremos ver que el algoritmo ha sido capaz de obtener muy buenos resultados. Concretamente ha conseguido obtener un 59,1% de precisión en predicciones de victoria del equipo local y un bajo 21,1% de precisión en predicciones de victoria del equipo visitante. Mientras tanto, la precisión de los empates alcanza un 40,4%.

Random Forest:

Por último, y para concluir el entrenamiento del conjunto de entrenamiento que recopila los partidos de la competición de Liga Adelante, se utilizará un Random Forest para entrenar el conjunto. Para este algoritmo se obtienen bastante peores resultados si los comparamos con los obtenidos con los demás clasificadores. En este caso la **tasa de acierto** asciende al 40,88% (Figura 84).

```
Correctly Classified Instances          83               40.8867 %
Incorrectly Classified Instances       120               59.1133 %
Kappa statistic                          0.011
Mean absolute error                      0.4102
Root mean squared error                  0.528
Relative absolute error                100.6961 %
Root relative squared error            117.0921 %
Total Number of Instances              203

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.57      0.563      0.53      0.57      0.55       0.483    1
                 0.275     0.25       0.269     0.275     0.272      0.52     X
                 0.178     0.177      0.222     0.178     0.198      0.508    2
Weighted Avg.    0.409     0.399      0.396     0.409     0.402      0.498

=== Confusion Matrix ===

  a  b  c   <-- classified as
 61 29 17 |  a = 1
 26 14 11 |  b = X
 28  9  8 |  c = 2
```

Figura 84. Resultados Random Forest - Conjunto Liga Adelante WEKA

Si se observa el **estadístico kappa**, se podrá ver que la correlación entre los atributos del conjunto es muy baja, lo que demuestra el valor 0,011. Como viene siendo habitual en este conjunto de partidos, este valor del estadístico kappa muestra una relación prácticamente nula entre los atributos del conjunto.

En cuanto a las tasas de **true positives**, tenemos tasas, muy bajas para los tres tipo de resultados. Para las victorias locales la tasa es del 57% y tendremos unas tasas algo peores para las victorias visitantes (17,8%) y los empates (27,5%).

Por último, si miramos la **precisión** de cada uno de los resultados, tendremos porcentajes de precisión malos, sobre todo para las predicciones de victoria visitante. Para los partidos ganados por el equipo local tendremos un 53% de precisión. En los partidos en los que se predice victoria del equipo visitante la precisión caerá hasta el 22,2%, mientras que para los partidos que acaban con empate la precisión será del 26,9%.

Resumen de Clasificadores para el conjunto de Liga Adelante:

Una vez se ha entrenado el conjunto de entrenamiento con cada uno de los 6 clasificadores, estamos en disposición de comparar los resultados para ver qué clasificadores se adaptan mejor a nuestro problema. En la Tabla 48 se recogen los resultados obtenidos con cada uno de los clasificadores:

	% Acierto	Kappa	Precisión Media
Red Bayesiana	50.73%	0.044	43.8%
Regresión Logística	48.76%	0.055	40.4%
Perceptrón Multicapa	44.82%	0.071	42.1%
OneR	49.26%	0.071	40.8%
J48	50.73%	0.139	46.0%
Random Forest	40.88%	0.011	39.6%

Tabla 48. Análisis del entrenamiento (Conjunto Liga Adelante)

A continuación se pasará a seleccionar los dos algoritmos que se utilizarán para implementar el sistema de predicción de resultados de la Liga Adelante. Viendo los datos de la tabla hay dos clasificadores que destacan sobre el resto por distintos motivos. Estos dos clasificadores de nuevo son la Red Bayesiana y el algoritmo J48.

Los datos obtenidos en este estudio no son muy buenos, ya que apenas se ha conseguido un 50% de aciertos sobre el conjunto de entrenamiento, lo que puede crear problemas en la fase de explotación del sistema, ya que puede que se consigan muy pocos aciertos de esta competición.

En cualquier caso se va a seleccionar para implementar el modelo en primer lugar al árbol J48, ya que es el que tiene una correlación entre atributos más alta de entre el conjunto de algoritmos y además ha obtenido el máximo porcentaje de aciertos en los tests.

Por otro lado se seleccionará también la Red Bayesiana, ya que es el segundo mejor algoritmo si nos basamos en términos de porcentaje de acierto y precisión media de los resultados.

Una vez seleccionados los algoritmos que se van a implementar en fases posteriores de este proyecto, hay que tener muy en cuenta los malos resultados obtenidos para este conjunto en relación con el resto de conjuntos ya entrenados. Como ya se comentó antes, estos datos pueden ser debidos a que en esta competición los equipos favoritos consiguen resultados menos esperados que en el resto de competiciones, lo que hace que predecir un resultado sea altamente complicado.

7.2.7 Conjunto de Partidos de la Ligue 1:

A continuación se van a presentar los resultados obtenidos con los 6 clasificadores para el conjunto que recoge los partidos la primera división francesa, también conocida como la Ligue 1:

Red Bayesiana (Bayes Net):

Para el conjunto de partidos de la Ligue 1, el primero de los clasificadores que van a ser utilizados es la Red Bayesiana, que en este caso nos proporciona una **tasa de aciertos** del 55,94% (Figura 85). Este valor se podría considerar bueno, pero habrá que quedar a la espera de los resultados con el resto de clasificadores para ver si este clasificador es elegido para ser implementado.

```
Correctly Classified Instances         113               55.9406 %
Incorrectly Classified Instances        89               44.0594 %
Kappa statistic                          0.2984
Mean absolute error                      0.3783
Root mean squared error                  0.4531
Relative absolute error                 86.5797 %
Root relative squared error             96.9416 %
Total Number of Instances              202

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.859     0.47       0.57      0.859     0.685      0.653    1
                 0.034     0.042      0.25      0.034     0.06       0.446    X
                 0.655     0.194      0.576     0.655     0.613      0.771    2
Weighted Avg.    0.559     0.266      0.478     0.559     0.482      0.627

=== Confusion Matrix ===

  a  b  c   <-- classified as
 73  2 10 |  a = 1
 39  2 18 |  b = X
 16  4 38 |  c = 2
```

Figura 85. Resultados Bayes Net - Conjunto Ligue 1 WEKA

El **estadístico kappa** para este conjunto muestra un valor de 0,298, un valor más que aceptable como y que está por encima de la media registrada en el entrenamiento de otros conjuntos.

En cuanto a la tasa de **True Positives**, el estudio refleja buenos resultados sólo para resultados en los que el equipo local sale vencedor. Concretamente se tendrá un 85,9% para victorias locales, un buen 65,5% para victorias visitantes y un 3,4% para empates.

Por último, se analizará la **precisión** de este modelo para cada uno de los resultados, donde se tiene una precisión del 57% para predicciones de victoria del equipo local, un muy buen 57,6% para victorias del equipo visitante y un 25% para empates.

Regresión Logística (Logistic):

El siguiente de los clasificadores que se va a utilizar en el entrenamiento de este conjunto será el clasificador basado en regresión logística. Si empezamos analizando el **porcentaje de aciertos**, este clasificador ha llegado a alcanzar una tasa de aciertos del 46,53% (Figura 86), lo que coloca a este clasificador a mucha distancia si comparamos los resultados obtenidos con los de la Red Bayesiana.

```
Correctly Classified Instances          94               46.5347 %
Incorrectly Classified Instances       108               53.4653 %
Kappa statistic                          0.1753
Mean absolute error                      0.3939
Root mean squared error                  0.4553
Relative absolute error                 90.1422 %
Root relative squared error             97.4113 %
Total Number of Instances              202

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.612     0.376      0.542     0.612     0.575      0.68     1
                 0.186     0.217      0.262     0.186     0.218      0.513    X
                 0.534     0.229      0.484     0.534     0.508      0.714    2
Weighted Avg.    0.465     0.287      0.444     0.465     0.451      0.641

=== Confusion Matrix ===

  a  b  c   <-- classified as
 52 19 14 |  a = 1
 29 11 19 |  b = X
 15 12 31 |  c = 2
```

Figura 86. Resultados Logistic - Conjunto Ligue 1 WEKA

En cuanto al **estadístico kappa**, ha reducido su valor respecto al del estudio anterior con la Red Bayesiana. Para este clasificador tenemos un valor de kappa de 0,175, lo que muestra que los atributos escogidos para este clasificador tienen una correlación baja entre ellos.

En cuanto a las tasas de **true positives** que recoge la herramienta, se puedo observar que las tasas son buenas, sobre todo la tasa para partidos en los que el equipo local ha conseguido salir vencedor. Para este tipo de partidos la tasa asciende al 61,2%. Mientras tanto, la tasa para partidos en los que el equipo visitante sale vencedor la tasa es del 53,4%. Finalmente la tasa de true positives para partidos que acaban en empate es del 18,6%.

Por último, los datos de **precisión** son buenos para los casos de victoria local y visitante, obteniéndose una precisión en las victorias locales del 54,2%. La precisión para

las predicciones de victoria del equipo visitante es algo más baja y se sitúa en el 48,9%, mientras que para las predicciones de empate es del 26,2%.

Multilayer Perceptron (Perceptrón Multicapa):

El siguiente de los clasificadores en ser utilizado ha sido el Perceptrón Multicapa. Los resultados obtenidos siguen sin ser brillante como ya habíamos podido ver con el clasificador de regresión logística. Para este caso, la **tasa de aciertos** se ha situado en un 49% (Figura 87), tasa relativamente baja si la comparamos con otros conjuntos de datos entrenados con anterioridad.

```
Correctly Classified Instances          99               49.0099 %
Incorrectly Classified Instances       103               50.9901 %
Kappa statistic                          0.2135
Mean absolute error                      0.3904
Root mean squared error                  0.4625
Relative absolute error                 89.3496 %
Root relative squared error             98.9425 %
Total Number of Instances              202

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.612     0.376      0.542     0.612      0.575      0.658    1
                 0.237     0.203      0.326     0.237      0.275      0.538    X
                 0.569     0.208      0.524     0.569      0.545      0.712    2
Weighted Avg.    0.49      0.277      0.473     0.49       0.479      0.638

=== Confusion Matrix ===

  a  b  c   <-- classified as
 52 19 14 |  a = 1
 29 14 16 |  b = X
 15 10 33 |  c = 2
```

Figura 87. Resultados Multilayer Perceptron - Conjunto Ligue 1 WEKA

En cuanto al **estadístico kappa**, el valor que se ha obtenido en este relativamente bueno, ya que el estadístico toma el valor 0,213, lo que indica una correlación aceptable entre los atributos de la clase.

En cuanto a las tasas de **true positives** tenemos tasas similares a las conseguidas por los modelos ya utilizados para entrenar este conjunto. En concreto tendremos una tasa del 61,2% en el caso de partidos que han acabado con victoria local, un buen 56,9% para partidos que finalizaron con victoria visitante y un 23,7% para partidos que concluyeron en empate.

Por último queda analizar el parámetro de **precisión**. Para los partidos clasificados como victoria del equipo local tenemos un 54,2% de precisión, mientras que para partidos con victoria visitante la precisión se queda en un 52,4%. Para partidos cuyo resultado ha sido de empate la precisión será del 32,6%.

OneR:

Los datos mostrados en la parte superior corresponden a los resultados del entrenamiento con el clasificador OneR aplicado al conjunto de partidos de la Ligue 1. El **porcentaje de aciertos** que ha conseguido ha sido del 50,49% (Figura 88).

```
Correctly Classified Instances         102               50.495  %
Incorrectly Classified Instances       100               49.505  %
Kappa statistic                          0.2177
Mean absolute error                      0.33
Root mean squared error                  0.5745
Relative absolute error                 75.5241 %
Root relative squared error            122.9069 %
Total Number of Instances              202

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.776     0.462      0.55      0.776     0.644      0.657    1
                 0.119     0.133      0.269     0.119     0.165      0.493    X
                 0.5       0.188      0.518     0.5       0.509      0.656    2
Weighted Avg.    0.505     0.287      0.459     0.505     0.465      0.609

=== Confusion Matrix ===

  a  b  c   <-- classified as
 66 10  9 |  a = 1
 34  7 18 |  b = X
 20  9 29 |  c = 2
```

Figura 88. Resultados One R - Conjunto Ligue 1 WEKA

En términos del **estadístico kappa**, de nuevo nos vuelve a ofrecer un valor aceptable. En este caso el valor del estadístico es del 0,217, lo que muestra que la correlación entre los atributos utilizados en el conjunto de entrenamiento es aceptable.

Si prestamos atención a las tasas de **true positives** que nos ofrece este estudio encontramos buenas tasas para las victorias del equipo local y visitante. Se ha conseguido una tasa de true positives del 77,6% para partidos que han sido clasificados como victoria del equipo local. Mientras tanto, la tasa en partidos que finalizan con victoria visitante baja hasta el 50%. Finalmente la tasa de true positives para empates se sitúa en el 11,9%

Finalmente, si atendemos al atributo de **precisión** los resultados son relativamente buenos, ya que tenemos una precisión para victorias locales del 55%. Para el caso de predicciones de victoria visitante tendremos una precisión del 51,8%. Por último, la precisión de este modelo para predecir empates es del 26,9%.

J48:

El siguiente de los clasificadores en ser utilizado es el árbol J48. Este algoritmo de clasificación ha conseguido de nuevo uno de los mejores resultados de entrenamiento de este conjunto de la Ligue 1. La **tasa de aciertos** que alcanza este algoritmo es del 57,42% (Figura 89).

```
Correctly Classified Instances         116               57.4257 %
Incorrectly Classified Instances        86               42.5743 %
Kappa statistic                          0.3172
Mean absolute error                      0.3787
Root mean squared error                  0.4359
Relative absolute error                 86.6548 %
Root relative squared error             93.2532 %
Total Number of Instances              202

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.894     0.504      0.563     0.894      0.691      0.654    1
                 0         0          0         0          0          0.477    X
                 0.69      0.188      0.597     0.69       0.64       0.694    2
Weighted Avg.    0.574     0.266      0.408     0.574      0.474      0.614

=== Confusion Matrix ===

  a  b  c   <-- classified as
 76  0  9 |  a = 1
 41  0 18 |  b = X
 18  0 40 |  c = 2
```

Figura 89. Resultados J48 - Conjunto Ligue 1 WEKA

En cuanto al **estadístico kappa**, para este clasificador tenemos un resultado de 0,317, el valor más alto obtenido por cualquiera de los clasificadores que ha entrenado este conjunto.

Si miramos las tasas de **true positives** que nos arroja el estudio podemos ver que para el resultado de victoria del equipo local se ha logrado obtener un 89,4%, mientras que para el resultado de victoria del equipo visitante tenemos un gran 69%. Mientras tanto, la tasa de true positives para empates ha quedado registrada en un 0%, ya que el algoritmo no es capaz de predecir empates.

Finalmente, si miramos el atributo de **precisión** podremos ver que el algoritmo ha sido capaz de obtener muy buenos resultados. Concretamente ha conseguido obtener un 56,3% de precisión en predicciones de victoria del equipo local y un muy buen 59,7% de precisión en predicciones de victoria del equipo visitante. Mientras tanto, la precisión de los empates no puede ser determinada debido a la imposibilidad del algoritmo de predecir empates.

Random Forest:

Por último, y para concluir el entrenamiento del conjunto de entrenamiento que recopila los partidos de la competición de Ligue 1 francesa, se utilizará un Random Forest para entrenar el conjunto. Para este algoritmo se obtienen malos resultados, unos resultados similares a los peores recogidos en otros clasificadores utilizados para este conjunto. En este caso la **tasa de acierto** asciende al 43,56% (Figura 90).

```
Correctly Classified Instances          88               43.5644 %
Incorrectly Classified Instances       114               56.4356 %
Kappa statistic                          0.1342
Mean absolute error                      0.3803
Root mean squared error                  0.5468
Relative absolute error                 87.027  %
Root relative squared error            116.9971 %
Total Number of Instances              202

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.494     0.41       0.467     0.494     0.48       0.593    1
                 0.356     0.273      0.35      0.356     0.353      0.555    X
                 0.431     0.188      0.481     0.431     0.455      0.655    2
Weighted Avg.    0.436     0.306      0.437     0.436     0.436      0.6

=== Confusion Matrix ===

  a  b  c   <-- classified as
 42 26 17 |  a = 1
 28 21 10 |  b = X
 20 13 25 |  c = 2
```

Figura 90. Resultados Random Forest - Conjunto Ligue 1 WEKA

Si se observa el **estadístico kappa**, se podrá ver que la correlación entre los atributos del conjunto es muy baja, lo que demuestra el valor 0,134. Este valor del estadístico kappa muestra una relación prácticamente nula entre los atributos del conjunto.

En cuanto a las tasas de **true positives**, tenemos tasas, muy bajas para los tres tipo de resultados, si bien para los empates es un buen resultado si se compara con los datos obtenidos con otros clasificadores. Para las victorias locales la tasa es del 49,4% y tendremos unas tasas algo peores para las victorias visitantes (43,1%) y los empates (35,6%).

Por último, si miramos la **precisión** de cada uno de los resultados, tendremos porcentajes de precisión malos, sobre todo para las predicciones de victoria local. Para los partidos ganados por el equipo local tendremos un 46,7% de precisión. En los partidos en los que se predice victoria del equipo visitante la precisión sube hasta el 48,1%, mientras que para los partidos que acaban con empate la precisión será del 35%.

Resumen de Clasificadores para el conjunto de Ligue 1:

Una vez se ha entrenado el conjunto de entrenamiento con cada uno de los seis clasificadores, estamos en disposición de comparar los resultados para ver qué clasificadores se adaptan mejor a nuestro problema. En la Tabla 49 se recogen los resultados obtenidos con cada uno de los clasificadores:

	% Acierto	Kappa	Precisión Media
Red Bayesiana	55.94%	0.298	47.8%
Regresión Logística	46.53%	0.175	44.4%
Perceptrón Multicapa	49.00%	0.213	47.3%
OneR	50.49%	0.217	45.9%
J48	57.42%	0.317	40.8%
Random Forest	43.56%	0.134	43.7%

Tabla 49. Análisis del entrenamiento (Conjunto Ligue 1)

A continuación se seleccionarán los dos algoritmos que se utilizarán para implementar el sistema de predicción de resultados de la competición de la Ligue 1. Viendo los datos de la tabla hay dos clasificadores que destacan sobre el resto, sobre todo en el apartado de la tasa de aciertos. Estos dos clasificadores de nuevo son la Red Bayesiana y el algoritmo J48.

Aunque el algoritmo J48 no es capaz de predecir empates, se ha situado como el mejor algoritmo en cuanto a porcentaje de aciertos, situándose en el 57,42% de aciertos. Como ya ha ocurrido en ocasiones anteriores, las carencias que presenta el algoritmo a la hora de predecir empates se intentarán subsanar en pasos posteriores del proceso, intentando cubrir los empates en el proceso de predicción de resultados para así evitar fallos en la predicción de partidos que finalmente acaben en empate. Es importante reseñar que a pesar de no acertar ninguno de los empates incluidos en los test, el clasificador a conseguido una tasa de aciertos cercana al 60%, lo que nos hace pensar que si se cubren correctamente los posibles empates a la hora de predecir los resultados, podremos conseguir tasas de acierto mucho más altas que las obtenidas en el entrenamiento de este conjunto.

Por otro lado también se elegirá la Red Bayesiana para ser implementada dentro del sistema de predicción de resultados de la competición de Ligue 1. Las razones que nos llevan a coger este clasificador en lugar de cualquiera de los otros 4 son principalmente la buena tasa de aciertos que ha registrado este clasificador y que tiene la precisión más alta entre todo el conjunto de clasificadores utilizados para entrenar este conjunto.

7.2.8 Conjunto de Partidos de la Premier League:

A continuación se van a presentar los resultados obtenidos con los seis clasificadores para el conjunto que recoge los partidos la primera división inglesa, también conocida como la Premier League:

Red Bayesiana (Bayes Net):

Para el conjunto de partidos de la Premier League, el primero de los clasificadores que van a ser utilizados es la Red Bayesiana, que en este caso nos proporciona una muy buena **tasa de aciertos** del 58,10%. Este valor cercano al 60% es bastante bueno, pero se tendrá que esperar a tener todos los resultados para poder hacer una valoración final de los resultados (Figura 91).

```
Correctly Classified Instances          86               58.1081 %
Incorrectly Classified Instances        62               41.8919 %
Kappa statistic                          0.336
Mean absolute error                      0.3498
Root mean squared error                  0.4708
Relative absolute error                 81.6584 %
Root relative squared error            101.7383 %
Total Number of Instances              148

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.773     0.354      0.638     0.773     0.699      0.666    1
                 0.408     0.152      0.571     0.408     0.476      0.555    X
                 0.455     0.157      0.455     0.455     0.455      0.627    2
Weighted Avg.    0.581     0.243      0.575     0.581     0.571      0.621

=== Confusion Matrix ===

  a  b  c   <-- classified as
 51  6  9 |  a = 1
 20 20  9 |  b = X
  9  9 15 |  c = 2
```

Figura 91. Resultados Bayes Net - Conjunto Premier League WEKA

El **estadístico kappa** para este conjunto muestra un valor de 0,336, un valor de los más altos obtenidos en todos los entrenamientos realizados hasta ahora con todos los conjuntos de partidos. Este valor nos indica una correlación más que aceptable entre los atributos del conjunto.

En cuanto a la tasa de **True Positives**, el estudio refleja buenos resultados en dos de los tres resultados posibles. Concretamente se tendrá un 77,3% para victorias locales y un buen 40,8% para empates. Mientras tanto, la tasa de true positives para partidos que acaben con victoria del equipo visitante será del 45,5.

Por último, en la **precisión** encontramos excelentes resultados, donde se tiene una precisión del 63,8% para predicciones de victoria del equipo local, un aceptable 45,5% para victorias del equipo visitante y un gran 57,1% para empates.

Regresión Logística (Logistic):

El siguiente de los clasificadores que se va a utilizar en el entrenamiento de este conjunto será el clasificador basado en regresión logística. Si empezamos analizando el **porcentaje de aciertos**, este clasificador ha llegado a alcanzar una tasa de aciertos del 52,02%, un buen resultado pero que queda lejos del resultado obtenido por la Red Bayesiana (Figura 92).

```
Correctly Classified Instances          77               52.027  %
Incorrectly Classified Instances        71               47.973  %
Kappa statistic                          0.239
Mean absolute error                      0.3531
Root mean squared error                  0.4399
Relative absolute error                 82.4258 %
Root relative squared error             95.07   %
Total Number of Instances              148

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.727     0.341      0.632     0.727      0.676      0.775    1
                 0.367     0.263      0.409     0.367      0.387      0.595    X
                 0.333     0.148      0.393     0.333      0.361      0.723    2
Weighted Avg.    0.52      0.272      0.505     0.52       0.51       0.704

=== Confusion Matrix ===

  a  b  c   <-- classified as
 48 13  5 |  a = 1
 19 18 12 |  b = X
  9 13 11 |  c = 2
```

Figura 92. Resultados Logistic - Conjunto Premier League WEKA

En cuanto al **estadístico kappa**, ha reducido su valor respecto al del estudio anterior con la Red Bayesiana. Para este clasificador tenemos un valor de kappa de 0,239, lo que muestra que los atributos escogidos para este clasificador tienen una correlación aceptable entre ellos, pero menos que en el caso anterior.

En cuanto a las tasas de **true positives** que recoge la herramienta, se puedo observar que las tasas son buenas, sobre todo la tasa para partidos en los que el equipo local ha conseguido salir vencedor. Para este tipo de partidos la tasa asciende al 72,7%. Mientras tanto, la tasa para partidos en los que el equipo visitante sale vencedor la tasa es del 33,3%. Finalmente la tasa de true positives para partidos que acaban en empate es del 36,7%.

Por último, los datos de **precisión** son buenos para los casos de victoria local y empate, obteniéndose una precisión en las victorias locales del 63,2%. La precisión para las predicciones de victoria del equipo visitante es algo más baja y se sitúa en el 39,3%, mientras que para las predicciones de empate es del 40,9%.

Multilayer Perceptron (Perceptrón Multicapa):

El siguiente de los clasificadores en ser utilizado ha sido el Perceptrón Multicapa. Los resultados obtenidos son bastante buenos si tenemos en cuenta la irregularidad que ha mostrado este clasificador en sus diferentes tests. Para este caso, la **tasa de aciertos** se ha situado en un 53,37%, tasa relativamente buena si la comparamos con otros conjuntos de datos entrenados con anterioridad (Figura 93)

```
Correctly Classified Instances          79               53.3784 %
Incorrectly Classified Instances        69               46.6216 %
Kappa statistic                          0.2587
Mean absolute error                      0.3348
Root mean squared error                  0.4724
Relative absolute error                 78.1662 %
Root relative squared error            102.1135 %
Total Number of Instances              148

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.697     0.366      0.605     0.697     0.648      0.721    1
                 0.469     0.232      0.5       0.469     0.484      0.663    X
                 0.303     0.139      0.385     0.303     0.339      0.635    2
Weighted Avg.    0.534     0.271      0.521     0.534     0.525      0.683

=== Confusion Matrix ===

  a  b  c   <-- classified as
 46 13  7 |  a = 1
 17 23  9 |  b = X
 13 10 10 |  c = 2
```

Figura 93. Resultados Multilayer Perceptron - Conjunto Premier League WEKA

En cuanto al **estadístico kappa**, el valor que se ha obtenido es relativamente bueno, ya que el estadístico toma el valor 0,258, lo que indica una correlación aceptable entre los atributos de la clase.

En cuanto a las tasas de **true positives** tenemos tasas similares a las conseguidas por los modelos ya utilizados para entrenar este conjunto. En concreto tendremos una tasa del 69,7% en el caso de partidos que han acabado con victoria local, un discreto 30,3% para partidos que finalizaron con victoria visitante y un buen 46,9% para partidos que concluyeron en empate.

Por último queda analizar el parámetro de **precisión**. Para los partidos clasificados como victoria del equipo local tenemos un 60,5% de precisión, mientras que para partidos con victoria visitante la precisión se queda en un 38,5%. Por otro lado se pueden presentar unos datos de precisión para empates extraordinarios, donde la precisión para estos resultados alcanza el 50%.

OneR:

Los datos mostrados en la parte superior corresponden a los resultados del entrenamiento con el clasificador OneR aplicado al conjunto de partidos de la Premier League. El **porcentaje de aciertos** que ha conseguido ha sido del 54,72% (Figura 94).

```
Correctly Classified Instances          81               54.7297 %
Incorrectly Classified Instances        67               45.2703 %
Kappa statistic                          0.2765
Mean absolute error                      0.3018
Root mean squared error                  0.5494
Relative absolute error                 70.4534 %
Root relative squared error            118.7402 %
Total Number of Instances              148

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.833     0.341      0.663     0.833     0.738      0.746    1
                 0.327     0.202      0.444     0.327     0.376      0.562    X
                 0.303     0.165      0.345     0.303     0.323      0.569    2
Weighted Avg.    0.547     0.256      0.52      0.547     0.526      0.646

=== Confusion Matrix ===

  a  b  c   <-- classified as
 55  6  5 |  a = 1
 19 16 14 |  b = X
  9 14 10 |  c = 2
```

Figura 94. Resultados One R - Conjunto Premier League WEKA

En términos del **estadístico kappa**, de nuevo nos vuelve a ofrecer un valor aceptable. En este caso el valor del estadístico es del 0,276, lo que muestra que la correlación entre los atributos utilizados en el conjunto de entrenamiento es aceptable.

Si prestamos atención a las tasas de **true positives** que nos ofrece este estudio encontramos buenas tasas para las victorias del equipo local. Se ha conseguido una tasa de true positives del 83,3% para partidos que han sido clasificados como victoria del equipo local. Mientras tanto, la tasa en partidos que finalizan con victoria visitante baja hasta el 30,3%. Finalmente la tasa de true positives para empates se sitúa en el 32,7%

Finalmente, si atendemos al atributo de **precisión** los resultados son relativamente buenos, ya que tenemos una precisión para victorias locales del 66,3%. Para el caso de predicciones de victoria visitante tendremos una precisión del 34,5%. Por último, la precisión de este modelo para predecir empates es del 44,4%, un dato que sigue demostrando que para este conjunto se está encontrando facilidad para predecir empates.

J48:

El siguiente de los clasificadores en ser utilizado es el árbol J48. Este algoritmo de clasificación ha conseguido de nuevo uno de los mejores resultados de entrenamiento de este conjunto de la Premier League. La **tasa de aciertos** que alcanza este algoritmo es del 56,75% (Figura 95).

```
Correctly Classified Instances          84               56.7568 %
Incorrectly Classified Instances        64               43.2432 %
Kappa statistic                          0.2918
Mean absolute error                      0.3551
Root mean squared error                  0.4665
Relative absolute error                 82.8413 %
Root relative squared error            100.7662 %
Total Number of Instances              148

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.818     0.488      0.574     0.818     0.675      0.617    1
                 0.429     0.131      0.618     0.429     0.506      0.613    X
                 0.273     0.096      0.45      0.273     0.34       0.534    2
Weighted Avg.    0.568     0.282      0.561     0.568     0.544      0.597

=== Confusion Matrix ===

  a  b  c   <-- classified as
 54  6  6 |  a = 1
 23 21  5 |  b = X
 17  7  9 |  c = 2
```

Figura 95. Resultados J48 - Conjunto Premier League WEKA

En cuanto al **estadístico kappa**, para este clasificador tenemos un resultado de 0,291, que no es el valor más alto obtenido para este conjunto pero se sitúa entre los más altos.

Si miramos las tasas de **true positives** que nos arroja el estudio podemos ver que para el resultado de victoria del equipo local se ha logrado obtener un 81,8%, mientras que para el resultado de victoria del equipo visitante tenemos un pobre 27,3%. Mientras tanto, la tasa de true positives para empates ha quedado registrada en un 42,9%.

Finalmente, si miramos el atributo de **precisión** podremos ver que el algoritmo ha sido capaz de obtener resultados excepcionales. Concretamente ha conseguido obtener un 57,4% de precisión en predicciones de victoria del equipo local y un aceptable 45% de precisión en predicciones de victoria del equipo visitante. Mientras tanto, la precisión de los empates ha registrado un nuevo récord, al haber alcanzado una precisión del 61,8%, tras ser acertados 21 empates de los 34 que han tenido lugar.

Random Forest:

Por último, y para concluir el entrenamiento del conjunto de entrenamiento que recopila los partidos de la competición de la Premier League inglesa, se utilizará un Random Forest para entrenar el conjunto. Para este algoritmo se obtienen malos resultados, unos resultados similares a los peores recogidos en otros clasificadores utilizados para este conjunto. En este caso la **tasa de acierto** asciende al 48,64% (Figura 96).

```
Correctly Classified Instances          72               48.6486 %
Incorrectly Classified Instances        76               51.3514 %
Kappa statistic                          0.1959
Mean absolute error                      0.3713
Root mean squared error                  0.5214
Relative absolute error                 86.6655 %
Root relative squared error            112.6838 %
Total Number of Instances              148

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.606     0.378      0.563     0.606     0.584      0.635    1
                 0.367     0.253      0.419     0.367     0.391      0.551    X
                 0.424     0.174      0.412     0.424     0.418      0.571    2
Weighted Avg.    0.486     0.291      0.482     0.486     0.483      0.593

=== Confusion Matrix ===

  a  b  c   <-- classified as
 40 18  8 |  a = 1
 19 18 12 |  b = X
 12  7 14 |  c = 2
```

Figura 96. Resultados Random Forest - Conjunto Premier League WEKA

Si se observa el **estadístico kappa**, se podrá ver que la correlación entre los atributos del conjunto es muy baja, lo que demuestra el valor 0,195. Este valor del estadístico kappa muestra una correlación baja entre los atributos del conjunto.

En cuanto a las tasas de **true positives**, tenemos tasas, bajas para los tres tipos de resultados, si bien para los empates es un resultado aceptable si se compara con los datos obtenidos con otros clasificadores. Para las victorias locales la tasa es del 60,6% y tendremos unas tasas algo peores para las victorias visitantes (42,4%) y los empates (36,7%).

Por último, si miramos la **precisión** de cada uno de los resultados, tendremos porcentajes de precisión buenos, sobre todo para las predicciones de victoria local y empate. Para los partidos ganados por el equipo local tendremos un 56,3% de precisión. En los partidos en los que se predice victoria del equipo visitante la precisión se queda en el 41,2%, mientras que para los partidos que acaban con empate la precisión será del 41,9%.

Resumen de Clasificadores para el conjunto de Premier League:

Una vez se ha entrenado el conjunto de entrenamiento con cada uno de los seis clasificadores, estamos en disposición de comparar los resultados para ver qué clasificadores se adaptan mejor a nuestro problema. En la Tabla 50 se recogen los resultados obtenidos con cada uno de los clasificadores:

	% Acierto	Kappa	Precisión Media
Red Bayesiana	58.10%	0.336	57.5%
Regresión Logística	52.02%	0.239	50.5%
Perceptrón Multicapa	53.37%	0.258	52.1%
OneR	54.72%	0.276	52.0%
J48	56.75%	0.291	56.1%
Random Forest	48.64%	0.195	48.2%

Tabla 50. Análisis del entrenamiento (Conjunto Premier League)

A continuación se seleccionarán los dos algoritmos que se utilizarán para implementar el sistema de predicción de resultados de la competición de la Premier League. Viendo los datos de la tabla hay dos clasificadores que destacan sobre el resto, tanto en el apartado de la tasa de aciertos como en el de precisión. Estos dos clasificadores de nuevo son la Red Bayesiana y el algoritmo J48.

En esta ocasión el algoritmo J48 ha conseguido evitar los problemas de predicción de empates que ha tenido a la hora de entrenar otros conjuntos de datos pertenecientes a otras competiciones. No sólo se ha conseguido evitar el problema de la predicción de empates, sino que además se ha conseguido una precisión en la predicción de empates de más del 60%, lo que sitúa a este algoritmo como el más fiable para predecir este tipo de resultados. Esta característica, además de la buena tasa de aciertos que se sitúa en el 56,75% hace que se haya seleccionado este algoritmo para ser implementado en el sistema de predicción de resultados de la competición de la Premier League.

Por otro lado también se elegirá la Red Bayesiana para ser implementada dentro del sistema de predicción de resultados de la competición de la Premier League. Las razones que nos llevan a coger este clasificador son la buena tasa de aciertos que se ha conseguido en el entrenamiento con este clasificador (58,10%), el buen valor del estadístico kappa (0,336) que muestra una correlación más que aceptable para este conjunto y la gran precisión que se alcanza con este algoritmo (un 57,5% de media con un valor superior al 57% para la predicción de empates).

7.2.9 Conjunto de Partidos de la Serie A:

A continuación se van a presentar los resultados obtenidos con los 6 clasificadores para el conjunto que recoge los partidos la primera división italiana, también conocida como la Serie A:

Para el conjunto de partidos de la Serie A, el primero de los clasificadores que van a ser utilizados es la Red Bayesiana, que en este caso nos proporciona una buena **tasa de aciertos** del 55,34% (Figura 97). Como se puede observar también en la matriz de confusión, este algoritmo no es capaz de predecir empates con facilidad, lo que nos hará estar atentos a las siguientes pruebas para ver si hay que realizar algún tipo de corrección en las estrategias para cubrir este tipo de marcador.

Red Bayesiana (Bayes Net):

```
Correctly Classified Instances          88               55.3459 %
Incorrectly Classified Instances        71               44.6541 %
Kappa statistic                          0.1868
Mean absolute error                      0.3811
Root mean squared error                  0.4647
Relative absolute error                 93.0708 %
Root relative squared error            102.8108 %
Total Number of Instances              159

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.843     0.632      0.593     0.843      0.697      0.648    1
                 0.083     0.049      0.333     0.083      0.133      0.508    X
                 0.375     0.143      0.469     0.375      0.417      0.614    2
Weighted Avg.    0.553     0.377      0.503     0.553      0.499      0.607

=== Confusion Matrix ===

  a  b  c   <-- classified as
 70  4  9 |  a = 1
 25  3  8 |  b = X
 23  2 15 |  c = 2
```

Figura 97. Resultados Bayes Net - Conjunto Serie A WEKA

El **estadístico kappa** para este conjunto muestra un valor de 0,186, un valor discreto, con el que tendremos que ver si el resto de clasificadores son capaces de superar.

En cuanto a la tasa de **True Positives**, el estudio refleja buenos resultados para victorias locales. Concretamente se tendrá un buen 84,3% para victorias locales. Mientras tanto, se tendrá un 37,5% en victorias visitantes y un 8,3% para empates.

Por último, en la **precisión** encontramos buenos resultados, donde se tiene una precisión del 59,3% para predicciones de victoria del equipo local, un aceptable 46,9% para victorias del equipo visitante y un 33,3% para empates.

Regresión Logística (Logistic):

El siguiente de los clasificadores que se va a utilizar en el entrenamiento de este conjunto será el clasificador basado en regresión logística. Si empezamos analizando el **porcentaje de aciertos**, este clasificador ha llegado a alcanzar una pobre tasa de aciertos

del 48,42% (Figura 98), un buen resultado pero que queda lejos del resultado obtenido por la Red Bayesiana.

```
Correctly Classified Instances          77               48.4277 %
Incorrectly Classified Instances        82               51.5723 %
Kappa statistic                          0.1305
Mean absolute error                      0.3822
Root mean squared error                  0.4861
Relative absolute error                 93.3381 %
Root relative squared error            107.5287 %
Total Number of Instances              159

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.699    0.474      0.617     0.699     0.655      0.593    1
                 0.194    0.179      0.241     0.194     0.215      0.484    X
                 0.3      0.202      0.333     0.3       0.316      0.618    2
Weighted Avg.    0.484    0.339      0.461     0.484     0.47       0.574

=== Confusion Matrix ===

  a  b  c   <-- classified as
 58 12 13 |  a = 1
 18  7 11 |  b = X
 18 10 12 |  c = 2
```

Figura 98. Resultados Logistic - Conjunto Serie A WEKA

En cuanto al **estadístico kappa**, ha reducido su valor respecto al del estudio anterior con la Red Bayesiana. Para este clasificador tenemos un valor de kappa de 0,130, lo que muestra que los atributos escogidos para este clasificador tienen una pobre correlación entre ellos, pero menos que en el caso anterior.

En cuanto a las tasas de **true positives** que recoge la herramienta, se puedo observar que las tasas son discretas. Para partidos en los que el equipo local consigue la victoria la tasa asciende al 69,9%. Mientras tanto, la tasa para partidos en los que el equipo visitante sale vencedor la tasa es del 30%. Finalmente la tasa de true positives para partidos que acaban en empate es del 19,4%.

Por último, los datos de **precisión** son buenos para los casos de victoria local, obteniéndose una precisión en las victorias locales del 61,7%. La precisión para las predicciones de victoria del equipo visitante es algo más baja y se sitúa en el 33,3%, mientras que para las predicciones de empate es del 24,1%.

Multilayer Perceptron (Perceptrón Multicapa):

El siguiente de los clasificadores en ser utilizado ha sido el Perceptrón Multicapa. Los resultados obtenidos son bastante malos, ya que como podemos ver no se obtienen tasas de aciertos muy buenas. Para este caso, la **tasa de aciertos** se ha situado en un

39,62% (Figura 99). Esta tasa es la peor registrada entre todos los entrenamientos con clasificadores que han sido realizados.

```
Correctly Classified Instances          63               39.6226 %
Incorrectly Classified Instances        96               60.3774 %
Kappa statistic                          0.0449
Mean absolute error                      0.4061
Root mean squared error                  0.5798
Relative absolute error                 99.0513 %
Root relative squared error            128.2556 %
Total Number of Instances              159

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.47      0.447      0.534     0.47       0.5       0.532    1
                 0.361     0.301      0.26      0.361      0.302     0.528    X
                 0.275     0.21       0.306     0.275      0.289     0.539    2
Weighted Avg.    0.396     0.354      0.415     0.396      0.402     0.533

=== Confusion Matrix ===

  a  b  c   <-- classified as
 39 27 17 |  a = 1
 15 13  8 |  b = X
 19 10 11 |  c = 2
```

Figura 99. Resultados Multilayer Perceptron - Conjunto Serie A WEKA

En cuanto al **estadístico kappa**, el valor que se ha obtenido es muy bajo, ya que el estadístico toma el valor 0,044, lo que indica una correlación prácticamente nula entre los atributos del conjunto.

En cuanto a las tasas de **true positives,** los resultados son malos si los comparamos con otros entrenamientos realizados con otros clasificadores. Para partidos que acaban con victoria del equipo local la tasa se situaba en el 47%. Para el caso de partidos en los que es el equipo visitante el que sale vencedor se tendrá una tasa del 27,5%, mientras que la tasa para partidos que concluyen en empate se queda en un 36,1%.

Por último queda analizar el parámetro de **precisión**. Para los partidos clasificados como victoria del equipo local tenemos un 53,4% de precisión, mientras que para partidos con victoria visitante la precisión se queda en un 30,6%. Finalmente, si miramos las tasas de precisión para los partidos que han sido clasificados como empate tendremos un 30,6%.

OneR:

Los datos mostrados en la parte superior corresponden a los resultados del entrenamiento con el clasificador OneR aplicado al conjunto de partidos de la Serie A. El **porcentaje de aciertos** que ha conseguido ha sido del 48,42% (Figura 100).

```
Correctly Classified Instances          77               48.4277 %
Incorrectly Classified Instances        82               51.5723 %
Kappa statistic                          0.0967
Mean absolute error                      0.3438
Root mean squared error                  0.5864
Relative absolute error                 83.9635 %
Root relative squared error            129.6922 %
Total Number of Instances              159

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.735     0.579      0.581     0.735     0.649      0.578    1
                 0         0.057      0         0         0          0.472    X
                 0.4       0.261      0.34      0.4       0.368      0.57     2
Weighted Avg.    0.484     0.381      0.389     0.484     0.431      0.552

=== Confusion Matrix ===

  a  b  c   <-- classified as
 61  6 16 |  a = 1
 21  0 15 |  b = X
 23  1 16 |  c = 2
```

Figura 100. Resultados One R - Conjunto Serie A WEKA

En términos del **estadístico kappa**, de nuevo nos vuelve a ofrecer un valor muy bajo. En este caso el valor del estadístico es del 0,096, lo que muestra que la correlación entre los atributos utilizados en el conjunto de entrenamiento es prácticamente nula.

Si prestamos atención a las tasas de **true positives** que nos ofrece este estudio encontramos buenas tasas para las victorias del equipo local. Se ha conseguido una tasa de true positives del 73,5% para partidos que han sido clasificados como victoria del equipo local. Mientras tanto, la tasa en partidos que finalizan con victoria visitante baja hasta el 40%. Finalmente la tasa de true positives para empates se sitúa en el 0%, después de que ninguno de los partidos acabados en empate fueran clasificados como tales.

Finalmente, si atendemos al atributo de **precisión** los resultados son nuevamente discretos, ya que tenemos una precisión para victorias locales del 58,1%. Para el caso de predicciones de victoria visitante tendremos una precisión del 34%. Por último, la precisión de este modelo para predecir empates vuelve a ser del 0%, ya que todos los partidos que han sido clasificados como empates no han logrado acabar con este resultado.

J48:

El siguiente de los clasificadores en ser utilizado es el árbol J48. Este algoritmo de clasificación ha conseguido de nuevo uno de los mejores resultados de entrenamiento de este conjunto de la Serie A. La **tasa de aciertos** que alcanza este algoritmo es del 55,34% (Figura 101).

```
Correctly Classified Instances          88               55.3459 %
Incorrectly Classified Instances        71               44.6541 %
Kappa statistic                          0.2019
Mean absolute error                      0.3809
Root mean squared error                  0.4622
Relative absolute error                 93.0213 %
Root relative squared error            102.2312 %
Total Number of Instances              159

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.807     0.592      0.598     0.807      0.687      0.589    1
                 0.083     0.057      0.3       0.083      0.13       0.475    X
                 0.45      0.16       0.486     0.45       0.468      0.595    2
Weighted Avg.    0.553     0.362      0.503     0.553      0.506      0.565

=== Confusion Matrix ===

  a  b  c   <-- classified as
 67  4 12 |  a = 1
 26  3  7 |  b = X
 19  3 18 |  c = 2
```

Figura 101. Resultados J48 - Conjunto Serie A WEKA

En cuanto al **estadístico kappa**, para este clasificador tenemos un resultado de 0,201, que se sitúa entre los valores más altos para este estadístico para este conjunto de partidos.

Si miramos las tasas de **true positives** que nos arroja el estudio podemos ver que para el resultado de victoria del equipo local se ha logrado obtener un 80,7%, mientras que para el resultado de victoria del equipo visitante tenemos un 45%. Mientras tanto, la tasa de true positives para empates ha quedado registrada en un 8,3%.

Finalmente, si miramos el atributo de **precisión** podremos ver que el algoritmo ha sido capaz de obtener buenos resultados. Concretamente ha conseguido obtener un 59,8% de precisión en predicciones de victoria del equipo local y un aceptable 48,6% de precisión en predicciones de victoria del equipo visitante. Mientras tanto, la precisión de los empates se sitúa en un discreto 30% que está en la media de otros entrenamientos realizados con otros conjuntos y clasificadores.

Random Forest:

Por último, y para concluir el entrenamiento del conjunto de entrenamiento que recopila los partidos de la competición de la Serie A italiana, se utilizará un Random Forest para entrenar el conjunto. Para este algoritmo se obtienen malos resultados, unos resultados que indican la debilidad de este clasificador para ser utilizado para la predicción de resultados de este conjunto. En este caso la **tasa de acierto** asciende al 47,79% (Figura 102).

```
Correctly Classified Instances          76               47.7987 %
Incorrectly Classified Instances        83               52.2013 %
Kappa statistic                          0.1388
Mean absolute error                      0.3591
Root mean squared error                  0.5288
Relative absolute error                 87.6838 %
Root relative squared error            116.9575 %
Total Number of Instances              159

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.639    0.447      0.609     0.639     0.624      0.619    1
                 0.25     0.22       0.25      0.25      0.25       0.473    X
                 0.35     0.185      0.389     0.35      0.368      0.634    2
Weighted Avg.    0.478    0.33       0.472     0.478     0.475      0.589

=== Confusion Matrix ===

  a  b  c   <-- classified as
 53 16 14 |  a = 1
 19  9  8 |  b = X
 15 11 14 |  c = 2
```

Figura 102. Resultados Random Forest - Conjunto Serie A WEKA

Si se observa el **estadístico kappa**, se podrá ver que la correlación entre los atributos del conjunto es muy baja, lo que demuestra el valor 0,138. Este valor del estadístico kappa muestra una correlación baja entre los atributos del conjunto.

En cuanto a las tasas de **true positives**, tenemos tasas bajas para los tres tipos de resultados. Para las victorias locales la tasa es del 63,9% y tendremos unas tasas algo peores para las victorias visitantes (35%) y los empates (25%).

Por último, si miramos la **precisión** de cada uno de los resultados, tendremos porcentajes de precisión buenos, sobre todo para las predicciones de victoria local. Para los partidos ganados por el equipo local tendremos un 60,9% de precisión. En los partidos en los que se predice victoria del equipo visitante la precisión se queda en el 38,9%, mientras que para los partidos que acaban con empate la precisión será del 25%.

Resumen de Clasificadores para el conjunto de Serie A:

Una vez se ha entrenado el conjunto de entrenamiento con cada uno de los seis clasificadores, estamos en disposición de comparar los resultados para ver qué clasificadores se adaptan mejor a nuestro problema. En la Tabla 51 se recogen los resultados obtenidos con cada uno de los clasificadores:

	% Acierto	Kappa	Precisión Media
Red Bayesiana	55.34%	0.186	50.3%
Regresión Logística	48.42%	0.130	46.1%
Perceptrón Multicapa	39.62%	0.044	41.5%
OneR	48.42%	0.096	38.9%
J48	55.34%	0.201	50.3%
Random Forest	47.79%	0.138	47.2%

Tabla 51. Análisis del entrenamiento (Conjunto Serie A)

A continuación se seleccionarán los dos algoritmos que se utilizarán para implementar el sistema de predicción de resultados de la competición de la Serie A italiana. Viendo los datos de la tabla hay dos clasificadores que destacan sobre el resto, tanto en el apartado de la tasa de aciertos como en el de precisión. Estos dos clasificadores de nuevo son la Red Bayesiana y el algoritmo J48.

En esta ocasión el algoritmo J48 ha conseguido sobresalir sobre el resto (con permiso de la Red Bayesiana) manteniendo unas muy buenas tasas tanto de acierto como de precisión de las predicciones. La tasa del 55,34% sitúa a este algoritmo muy por encima de los competidores que no han sido elegidos para la fase de implementación. Además, es uno de los clasificadores que mantiene una correlación aceptable entre los atributos de su conjunto.

Por otro lado también se elegirá la Red Bayesiana para ser implementada dentro del sistema de predicción de resultados de la competición de la Serie A italiana. Las razones que nos llevan a coger este clasificador son que posee la mejor tasa de aciertos (igual tasa que el árbol J48) situándose en el 58,10%, un valor aceptable del estadístico kappa (0,186) que muestra una correlación aceptable para este conjunto y la buena precisión que se alcanza con este algoritmo (un 50,3% de media), valor que está muy por encima de los valores ofrecidos por el resto de algoritmos utilizados, que lamentablemente han presentado unos resultados muy pobres, lo que hace que no se vayan a tener en cuenta en la siguiente fase de implementación de los modelos de predicción para esta competición.

7.2.10 Conjunto de Partidos Internacionales:

A continuación se van a presentar los resultados obtenidos con los seis clasificadores para el conjunto que recoge los partidos de selecciones internacionales. Partidos tanto amistosos como oficiales de las principales competiciones mundiales:

Red Bayesiana (Bayes Net):

Para el conjunto de partidos de selecciones nacionales, el primero de los clasificadores que van a ser utilizados es la Red Bayesiana, que en este caso nos proporciona una buena **tasa de aciertos** del 56,43% (Figura 103). Como se puede observar también en la matriz de confusión, este algoritmo no es capaz de predecir empates con facilidad, lo que nos hará estar atentos a las siguientes pruebas para ver si

hay que realizar algún tipo de corrección en las estrategias para cubrir este tipo de marcador.

```
Correctly Classified Instances         285               56.4356 %
Incorrectly Classified Instances       220               43.5644 %
Kappa statistic                          0.298
Mean absolute error                      0.3324
Root mean squared error                  0.4411
Relative absolute error                 77.1207 %
Root relative squared error             95.031  %
Total Number of Instances              505

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.656     0.276      0.638     0.656     0.647      0.754    1
                 0.044     0.018      0.417     0.044     0.079      0.568    X
                 0.79      0.404      0.511     0.79      0.621      0.759    2
Weighted Avg.    0.564     0.262      0.544     0.564     0.51       0.714

=== Confusion Matrix ===

   a   b   c   <-- classified as
 141   4  70 |   a = 1
  46   5  63 |   b = X
  34   3 139 |   c = 2
```

Figura 103. Resultados Bayes Net - Conjunto Partidos Internacionales WEKA

El **estadístico kappa** para este conjunto muestra un valor de 0,298, un buen valor que salvo sorpresa será uno de los mejores para este conjunto de datos si nos basamos en experiencias con conjuntos y clasificadores anteriores.

En cuanto a la tasa de **True Positives**, el estudio refleja buenos resultados para victorias visitantes. Concretamente se tendrá un buen 79% para victorias visitantes. Mientras tanto, se tendrá un 65,6% en victorias locales y un 4,4% para empates.

Por último, en la **precisión** encontramos buenos resultados, donde se tiene una precisión del 63,8% para predicciones de victoria del equipo local, un aceptable 51,1% para victorias del equipo visitante y un buen 41,7% para empates.

Regresión Logística (Logistic):

El siguiente de los clasificadores que se va a utilizar en el entrenamiento de este conjunto será el clasificador basado en regresión logística. Si empezamos analizando el **porcentaje de aciertos**, este clasificador ha llegado a alcanzar una buena tasa de aciertos del 51,48%, un buen resultado pero que queda lejos del resultado obtenido por la Red Bayesiana (Figura 104).

```
Correctly Classified Instances         260               51.4851 %
Incorrectly Classified Instances       245               48.5149 %
Kappa statistic                          0.2263
Mean absolute error                      0.3589
Root mean squared error                  0.452
Relative absolute error                 83.2653 %
Root relative squared error             97.3866 %
Total Number of Instances              505

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.66      0.362      0.575      0.66      0.615      0.719    1
                 0.079     0.118      0.164      0.079     0.107      0.5      X
                 0.619     0.286      0.537      0.619     0.575      0.747    2
Weighted Avg.    0.515     0.28       0.469      0.515     0.486      0.68

=== Confusion Matrix ===

   a   b   c   <-- classified as
 142  30  43 |   a = 1
  54   9  51 |   b = X
  51  16 109 |   c = 2
```

Figura 104. Resultados Logistic - Conjunto Partidos Internacionales WEKA

En cuanto al **estadístico kappa**, ha reducido su valor respecto al del estudio anterior con la Red Bayesiana. Para este clasificador tenemos un valor de kappa de 0,226, lo que muestra que los atributos escogidos para este clasificador tienen una correlación aceptable entre ellos.

En cuanto a las tasas de **true positives** que recoge la herramienta, se puedo observar que las tasas son buenas en dos de los resultados posibles. Para partidos en los que el equipo local consigue la victoria la tasa asciende al 66%. Mientras tanto, la tasa para partidos en los que el equipo visitante sale vencedor la tasa es del 61,9%. Finalmente la tasa de true positives para partidos que acaban en empate es del 7,9%.

Por último, los datos de **precisión** son buenos para los casos de victoria local y visitante, obteniéndose una precisión en las victorias locales del 57,5%. La precisión para las predicciones de victoria del equipo visitante es algo más baja y se sitúa en el 53,7%, mientras que para las predicciones de empate es del 16,4%.

Multilayer Perceptron (Perceptrón Multicapa):

El siguiente de los clasificadores en ser utilizado ha sido el Perceptrón Multicapa. Los resultados obtenidos son bastante buenos, sobre todo si los comparamos con conjuntos anteriormente entrenados con este clasificador. Para este caso, la **tasa de aciertos** se ha situado en un 53,86% (Figura 105). Esta tasa es hasta el momento la segunda mejor de los clasificadores usados hasta el momento.

```
Correctly Classified Instances         272               53.8614 %
Incorrectly Classified Instances       233               46.1386 %
Kappa statistic                          0.2627
Mean absolute error                      0.3468
Root mean squared error                  0.4385
Relative absolute error                 80.4544 %
Root relative squared error             94.4751 %
Total Number of Instances              505

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.712    0.386      0.577     0.712     0.638      0.747    1
                 0.132    0.118      0.246     0.132     0.171      0.602    X
                 0.591    0.228      0.581     0.591     0.586      0.754    2
Weighted Avg.    0.539    0.27       0.504     0.539     0.514      0.717

=== Confusion Matrix ===

   a   b   c   <-- classified as
 153  27  35 |   a = 1
  59  15  40 |   b = X
  53  19 104 |   c = 2
```

Figura 105. Resultados Multilayer Perceptron – Conjunto Part. Internacionales WEKA

En cuanto al **estadístico kappa**, el valor que se ha obtenido es bueno, ya que el estadístico toma el valor 0,262, lo que indica una correlación aceptable entre los atributos del conjunto.

En cuanto a las tasas de **true positives** los resultados son buenos, sobre todo para partidos que acaban con victoria local. Para este tipo de partidos la tasa se situaba en el 71,2%. Para el caso de partidos en los que es el equipo visitante el que sale vencedor se tendrá una tasa del 59,1%, mientras que la tasa para partidos que concluyen en empate se queda en un 13,2%.

Por último queda analizar el parámetro de **precisión**. Para los partidos clasificados como victoria del equipo local tenemos un 57,7% de precisión, mientras que para partidos con victoria visitante la precisión asciende hasta un 58,1%. Finalmente, si miramos las tasas de precisión para los partidos que han sido clasificados como empate tendremos un 24,6%.

OneR:

Los datos mostrados en la parte superior corresponden a los resultados del entrenamiento con el clasificador OneR aplicado al conjunto de partidos de selecciones nacionales. El **porcentaje de aciertos** que ha conseguido ha sido del 52,27% (Figura 106).

```
Correctly Classified Instances         264               52.2772 %
Incorrectly Classified Instances       241               47.7228 %
Kappa statistic                          0.2455
Mean absolute error                      0.3182
Root mean squared error                  0.564
Relative absolute error                 73.81   %
Root relative squared error            121.5209 %
Total Number of Instances              505

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.656     0.331      0.595     0.656     0.624      0.662    1
                 0.167     0.133      0.268     0.167     0.205      0.517    X
                 0.591     0.283      0.528     0.591     0.558      0.654    2
Weighted Avg.    0.523     0.269      0.498     0.523     0.506      0.627

=== Confusion Matrix ===

   a   b   c   <-- classified as
 141  26  48 |   a = 1
  50  19  45 |   b = X
  46  26 104 |   c = 2
```

Figura 106. Resultados One R - Conjunto Partidos Internacionales WEKA

En términos del **estadístico kappa**, de nuevo nos vuelve a ofrecer un valor más que aceptable. En este caso el valor del estadístico es del 0,245, lo que muestra que la correlación entre los atributos utilizados en el conjunto de entrenamiento aceptable para tenerla en cuenta para el estudio.

Si prestamos atención a las tasas de **true positives** que nos ofrece este estudio encontramos unas tasas aceptables para dos de los tres resultados. Se ha conseguido una tasa de true positives del 65,6% para partidos que han sido clasificados como victoria del equipo local. Mientras tanto, la tasa en partidos que finalizan con victoria visitante baja hasta el 59,1%. Finalmente la tasa de true positives para empates se sitúa en el 16,7%.

Finalmente, si atendemos al atributo de **precisión** los resultados son nuevamente aceptables, ya que tenemos una precisión para victorias locales del 59,5%. Para el caso de predicciones de victoria visitante tendremos una precisión del 52,8%. Por último, la precisión de este modelo para predecir empates es del 26,8%.

J48:

El siguiente de los clasificadores en ser utilizado es el árbol J48. Este algoritmo de clasificación ha conseguido de nuevo uno de los mejores resultados de entrenamiento de este conjunto de partidos internacionales. La **tasa de aciertos** que alcanza este algoritmo es un excepcional 59,40% (Figura 107).

```
Correctly Classified Instances         300               59.4059 %
Incorrectly Classified Instances       205               40.5941 %
Kappa statistic                          0.3331
Mean absolute error                      0.3751
Root mean squared error                  0.4341
Relative absolute error                 87.0243 %
Root relative squared error             93.5311 %
Total Number of Instances              505

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.791     0.379      0.607     0.791     0.687      0.686    1
                 0         0          0         0         0          0.46     X
                 0.739     0.289      0.578     0.739     0.648      0.711    2
Weighted Avg.    0.594     0.262      0.46      0.594     0.518      0.644

=== Confusion Matrix ===

   a   b   c   <-- classified as
 170   0  45 |   a = 1
  64   0  50 |   b = X
  46   0 130 |   c = 2
```

Figura 107. Resultados J48 - Conjunto Partidos Internacionales WEKA

En cuanto al **estadístico kappa**, para este clasificador tenemos un resultado de 0,333, que se sitúa entre los valores más altos para este estadístico para este conjunto de partidos.

Si miramos las tasas de **true positives** que nos arroja el estudio podemos ver que para el resultado de victoria del equipo local se ha logrado obtener un 79,1%, mientras que para el resultado de victoria del equipo visitante tenemos un gran 73,9%. Mientras tanto, la tasa de true positives para empates ha quedado registrada en un 0%, debido a la imposibilidad del modelo para predecir empates. Al igual que se ha mencionado en ocasiones anteriores, se intentará cubrir este resultado para minimizar la tasa de fallos que se produce debido a este resultado.

Finalmente, si miramos el atributo de **precisión** podremos ver que el algoritmo ha sido capaz de obtener buenos resultados. Concretamente ha conseguido obtener un 60,7% de precisión en predicciones de victoria del equipo local y un buen 57,8% de precisión en predicciones de victoria del equipo visitante. Mientras tanto, la precisión de los empates no se puede determinar debido a la imposibilidad del modelo para predecir este tipo de resultados.

Random Forest:

Por último, y para concluir el entrenamiento del conjunto de entrenamiento que recopila los partidos de selecciones nacionales en diferentes competiciones, se utilizará un Random Forest para entrenar el conjunto. Para este algoritmo se obtienen buenos resultados, pero que no llegan al nivel que han alcanzado otros algoritmos utilizados con anterioridad. En este caso la **tasa de acierto** asciende al 50,29% (Figura 108).

```
Correctly Classified Instances         254               50.297  %
Incorrectly Classified Instances       251               49.703  %
Kappa statistic                          0.2227
Mean absolute error                      0.3549
Root mean squared error                  0.5063
Relative absolute error                 82.3489 %
Root relative squared error            109.0685 %
Total Number of Instances              505

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.684     0.307      0.623     0.684     0.652      0.719    1
                 0.175     0.202      0.202     0.175     0.188      0.497    X
                 0.494     0.252      0.512     0.494     0.503      0.665    2
Weighted Avg.    0.503     0.264      0.489     0.503     0.495      0.65

=== Confusion Matrix ===

   a   b   c   <-- classified as
 147  32  36 |   a = 1
  47  20  47 |   b = X
  42  47  87 |   c = 2
```

Figura 108. Resultados Random Forest - Conjunto Partidos Internacionales WEKA

Si se observa el **estadístico kappa**, se podrá ver que la correlación entre los atributos del conjunto es aceptable, lo que demuestra el valor 0,222.

En cuanto a las tasas de **true positives**, tenemos tasas aceptables para los tres tipos de resultados. Para las victorias locales la tasa es del 68,4% y tendremos unas tasas algo peores para las victorias visitantes (49,4%) y los empates (17,5%).

Por último, si miramos la **precisión** de cada uno de los resultados, tendremos porcentajes de precisión buenos, sobre todo para las predicciones de victoria local. Para los partidos ganados por el equipo local tendremos un 62,3% de precisión. En los partidos en los que se predice victoria del equipo visitante la precisión se queda en el 51,2%, mientras que para los partidos que acaban con empate la precisión será del 20,2%.

Resumen de Clasificadores para el conjunto de Partidos Internacionales:

Una vez se ha entrenado el conjunto de entrenamiento con cada uno de los seis clasificadores, estamos en disposición de comparar los resultados para ver qué clasificadores se adaptan mejor a nuestro problema. En la tabla 52 se recogen los resultados obtenidos con cada uno de los clasificadores:

	% Acierto	Kappa	Precisión Media
Red Bayesiana	56.43%	0.298	54.4%
Regresión Logística	51.48%	0.226	46.9%
Perceptrón Multicapa	53.86%	0.262	50.4%
OneR	52.27%	0.245	49.8%
J48	59.40%	0.333	46.0%
Random Forest	50.29%	0.222	48.9%

Tabla 52. Análisis del entrenamiento (Conjunto Partidos Internacionales)

A continuación se seleccionarán los dos algoritmos que se utilizarán para implementar el sistema de predicción de resultados del conjunto de partidos de selecciones nacionales disputados en conjuntos de partidos amistosos y oficiales. Viendo los datos de la tabla de nuevo son los dos clasificadores habituales los que destacan sobre el resto, aunque uno de ellos presenta pequeños problemas en términos de precisión.

En esta ocasión el algoritmo J48 ha conseguido destacar sobre el resto de clasificadores manteniendo unas muy buenas tasas de acierto. La tasa del 59,40% sitúa a este algoritmo muy por encima de los competidores que no han sido elegidos para la fase de implementación. Además, es uno de los clasificadores que mantiene una correlación más que aceptable entre los atributos de su conjunto. El problema que tenemos en esta ocasión no es nuevo, y deriva de la imposibilidad del árbol generado de predecir empates. Por esta deficiencia la precisión del modelo cae al 46%. Como se intentará en fases posteriores intentar solventar este percance cubriendo los resultados de empate a la hora de predecir los resultados, el dato del 46% no es significativo, ya que tras cubrir los empates la tasa debería subir por encima del 70%.

Por otro lado también se elegirá la Red Bayesiana para ser implementada dentro del sistema de predicción de resultados de la competición de los partidos internacionales. Las razones que nos llevan a coger este clasificador son que posee la segunda mejor tasa de aciertos situándose en el 56,43%, un valor aceptable del estadístico kappa (0.298) que muestra una correlación aceptable para este conjunto y la buena precisión que se alcanza con este algoritmo (un 54,4% de media), valor que está muy por encima de los valores ofrecidos por el resto de algoritmos utilizados, que de nuevo han vuelto a presentar unos resultados muy pobres, lo que hace que no se tengan en cuenta en la siguiente fase de implementación de los modelos de predicción para esta competición.

7.2.11 Conjunto de la NBA:

A continuación se van a presentar los resultados obtenidos con los 6 clasificadores para el conjunto que recoge los partidos de la máxima competición del baloncesto norteamericano, la NBA:

Red Bayesiana (Bayes Net):

En primer lugar y antes de analizar los datos, sólo recordar que los datos que se han obtenido en los entrenamientos de este conjunto no pueden ser comparados con los de las competiciones futbolísticas, ya que en este caso sólo tenemos dos resultados posibles. Al ser el espacio de resultados más reducido aumentarán las tasas de acierto y precisión de cada uno de los clasificadores usados. El primero de los clasificadores que va a ser usado es la Red Bayesiana, que en este caso nos proporciona una buena **tasa de aciertos** del 76,43% (Figura 109).

```
Correctly Classified Instances         133               76.4368 %
Incorrectly Classified Instances        41               23.5632 %
Kappa statistic                          0.4958
Mean absolute error                      0.2774
Root mean squared error                  0.4636
Relative absolute error                 58.2073 %
Root relative squared error             94.9908 %
Total Number of Instances              174

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.84      0.353      0.788     0.84      0.813      0.76     1
                 0.647     0.16       0.721     0.647     0.682      0.76     2
Weighted Avg.    0.764     0.278      0.762     0.764     0.762      0.76

=== Confusion Matrix ===

  a  b   <-- classified as
 89 17 |  a = 1
 24 44 |  b = 2
```

Figura 109. Resultados Bayes Net - Conjunto NBA WEKA

El **estadístico kappa** para este conjunto muestra un valor de 0,495, una buena correlación que certifica la relación entre los atributos utilizados en este entrenamiento.

En cuanto a la tasa de **True Positives**, el estudio refleja buenos resultados para victorias locales. Concretamente se tendrá un buen 84% para victorias locales. Mientras tanto, se tendrá un 64,7% en victorias visitantes.

Por último, en la **precisión** encontramos buenos resultados, donde se tiene una precisión del 78,8% para predicciones de victoria del equipo local, y un 72,1% para victorias del equipo visitante.

Regresión Logística (Logistic):

El siguiente de los clasificadores que se va a utilizar en el entrenamiento de este conjunto será el clasificador basado en regresión logística. Si empezamos analizando el **porcentaje de aciertos**, este clasificador ha llegado a alcanzar una buena tasa de aciertos del 71,83% (Figura 110), un buen resultado pero que queda lejos del resultado obtenido por la Red Bayesiana.

```
Correctly Classified Instances         125               71.8391 %
Incorrectly Classified Instances        49               28.1609 %
Kappa statistic                          0.407
Mean absolute error                      0.3611
Root mean squared error                  0.4491
Relative absolute error                 75.7825 %
Root relative squared error             92.012  %
Total Number of Instances              174

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.774     0.368      0.766     0.774     0.77       0.741    1
                 0.632     0.226      0.642     0.632     0.637      0.741    2
Weighted Avg.    0.718     0.312      0.718     0.718     0.718      0.741

=== Confusion Matrix ===

  a  b   <-- classified as
 82 24 |  a = 1
 25 43 |  b = 2
```

Figura 110. Resultados Logistic - Conjunto NBA WEKA

En cuanto al **estadístico kappa**, ha reducido su valor respecto al del estudio anterior con la Red Bayesiana. Para este clasificador tenemos un valor de kappa de 0,407, lo que muestra que los atributos escogidos para este clasificador tienen una buena correlación entre ellos.

En cuanto a las tasas de **true positives** que recoge la herramienta, se puedo observar que las tasas son buenas en los dos de los resultados posibles. Para partidos en los que el equipo local consigue la victoria la tasa asciende al 77,4%. Mientras tanto, la tasa para partidos en los que el equipo visitante sale vencedor la tasa es del 63,2%.

Por último, los datos de **precisión** son buenos, sobre todo para las predicciones de victoria local, obteniéndose una precisión en las victorias locales del 76,6%. La precisión para las predicciones de victoria del equipo visitante es algo más baja y se sitúa en el 64,2%.

Multilayer Perceptron (Perceptrón Multicapa):

El siguiente de los clasificadores en ser utilizado ha sido el Perceptrón Multicapa. Los resultados obtenidos no son tan buenos como en los dos casos anteriores. Para este caso, la **tasa de aciertos** se ha situado en un 67,81% (Figura 111). Esta tasa se queda lejos de los dos clasificadores ya utilizados para entrenar el conjunto.

```
Correctly Classified Instances         118               67.8161 %
Incorrectly Classified Instances        56               32.1839 %
Kappa statistic                          0.338
Mean absolute error                      0.361
Root mean squared error                  0.5357
Relative absolute error                 75.7206 %
Root relative squared error            109.7892 %
Total Number of Instances              174

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.698     0.353      0.755     0.698     0.725      0.685     1
                 0.647     0.302      0.579     0.647     0.611      0.685     2
Weighted Avg.    0.678     0.333      0.686     0.678     0.681      0.685

=== Confusion Matrix ===

  a  b   <-- classified as
 74 32 |  a = 1
 24 44 |  b = 2
```

Figura 111. Resultados Multilayer Perceptron – Conjunto NBA WEKA

En cuanto al **estadístico kappa**, el valor que se ha obtenido es bueno, ya que el estadístico toma el valor 0,338, lo que indica una correlación aceptable entre los atributos del conjunto. No obstante vuelve a quedar lejos del valor obtenido en los dos clasificadores anteriores.

En cuanto a las tasas de **true positives** los resultados son buenos, pero de nuevo se quedan lejos de los dos clasificadores anteriores. Para partidos que acaban con victoria local la tasa se situaba en el 69,8%. Para el caso de partidos en los que es el equipo visitante el que sale vencedor se tendrá una tasa del 64,7%.

Por último queda analizar el parámetro de **precisión**. Para los partidos clasificados como victoria del equipo local tenemos un buen 75,5% de precisión, mientras que para partidos con victoria visitante la precisión se queda en un 57,9%.

OneR:

Los datos mostrados en la parte superior corresponden a los resultados del entrenamiento con el clasificador OneR aplicado al conjunto de partidos la NBA. El **porcentaje de aciertos** que ha conseguido ha sido del 69,54% (Figura 112).

```
Correctly Classified Instances         121               69.5402 %
Incorrectly Classified Instances        53               30.4598 %
Kappa statistic                          0.3228
Mean absolute error                      0.3046
Root mean squared error                  0.5519
Relative absolute error                 63.9163 %
Root relative squared error            113.092  %
Total Number of Instances              174

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.849     0.544      0.709     0.849     0.773      0.652    1
                 0.456     0.151      0.66      0.456     0.539      0.652    2
Weighted Avg.    0.695     0.39       0.689     0.695     0.681      0.652

=== Confusion Matrix ===

  a  b   <-- classified as
 90 16 |  a = 1
 37 31 |  b = 2
```

Figura 112. Resultados One R - Conjunto NBA WEKA

En términos del **estadístico kappa**, de nuevo nos vuelve a ofrecer un valor más que aceptable, aunque algo bajo si lo comparamos con los clasificadores usados con anterioridad. En este caso el valor del estadístico es del 0,322, lo que muestra que la correlación entre los atributos utilizados en el conjunto de entrenamiento aceptable para tenerla en cuenta para el estudio.

Si prestamos atención a las tasas de **true positives** que nos ofrece este estudio encontramos unas tasas aceptables para las victorias de equipos locales. Se ha conseguido una tasa de true positives del 84,9% para partidos que han sido clasificados como victoria del equipo local. Mientras tanto, la tasa en partidos que finalizan con victoria visitante baja hasta el 45,6%.

Finalmente, si atendemos al atributo de **precisión** los resultados son nuevamente aceptables, ya que tenemos una precisión para victorias locales del 70,9%. Para el caso de predicciones de victoria visitante tendremos una precisión del 66%.

J48:

El siguiente de los clasificadores en ser utilizado es el árbol J48. Este algoritmo de clasificación ha conseguido de nuevo unos resultados excelentes que le sitúan de nuevo como uno de los clasificadores elegidos para ser implementados. La **tasa de aciertos** que alcanza este algoritmo es un gran 72,98% (Figura 113).

```
Correctly Classified Instances         127               72.9885 %
Incorrectly Classified Instances        47               27.0115 %
Kappa statistic                          0.3787
Mean absolute error                      0.3653
Root mean squared error                  0.4532
Relative absolute error                 76.6339 %
Root relative squared error             92.8785 %
Total Number of Instances              174

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.934     0.588      0.712     0.934      0.808      0.677    1
                 0.412     0.066      0.8       0.412      0.544      0.677    2
Weighted Avg.    0.73      0.384      0.747     0.73       0.705      0.677

=== Confusion Matrix ===

  a  b   <-- classified as
 99  7 |  a = 1
 40 28 |  b = 2
```

Figura 113. Resultados J48 - Conjunto NBA WEKA

En cuanto al **estadístico kappa**, para este clasificador tenemos un resultado de 0,378, que muestra una correlación aceptable entre los atributos utilizados para el entrenamiento.

Si miramos las tasas de **true positives** que nos arroja el estudio podemos ver que para el resultado de victoria del equipo local se ha logrado obtener un grandísimo 93,4%, mientras que para el resultado de victoria del equipo visitante tenemos un pobre 41,2%.

Finalmente, si miramos el atributo de **precisión** podremos ver que el algoritmo ha sido capaz de obtener buenos resultados. Concretamente ha conseguido obtener un 71,2% de precisión en predicciones de victoria del equipo local y un buenísimo 80% de precisión en predicciones de victoria del equipo visitante. Estos dos últimos datos muestran la gran fiabilidad que da el algoritmo para la predicción de resultados, consiguiendo una precisión media del 74,7%

Random Forest:

Por último, y para concluir el entrenamiento del conjunto que recopila los partidos de la NBA, se utilizará un Random Fores. Para este algoritmo se obtienen buenos resultados, en todos los parámetros que se están midiendo. En este caso la **tasa de acierto** asciende al 71,26% (Figura 114).

```
Correctly Classified Instances         124               71.2644 %
Incorrectly Classified Instances        50               28.7356 %
Kappa statistic                          0.3734
Mean absolute error                      0.3704
Root mean squared error                  0.4523
Relative absolute error                 77.7237 %
Root relative squared error             92.682  %
Total Number of Instances              174

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.83      0.471      0.733     0.83      0.779      0.74     1
                 0.529     0.17       0.667     0.529     0.59       0.739    2
Weighted Avg.    0.713     0.353      0.707     0.713     0.705      0.739

=== Confusion Matrix ===

  a  b   <-- classified as
 88 18 |  a = 1
 32 36 |  b = 2
```

Figura 114. Resultados Random Forest - Conjunto NBA WEKA

Si se observa el **estadístico kappa**, se podrá ver que la correlación entre los atributos del conjunto es buena, lo que demuestra el valor 0,373.

En cuanto a las tasas de **true positives**, tenemos tasas buenas para los partidos que acaban con victoria local. Para las victorias locales la tasa es del 83% y tendremos unas tasas algo peores para las victorias visitantes (52,9%).

Por último, si miramos la **precisión** de cada uno de los resultados, tendremos porcentajes de precisión buenos, sobre todo para las predicciones de victoria local. Para los partidos ganados por el equipo local tendremos un 73,3% de precisión. En los partidos en los que se predice victoria del equipo visitante la precisión se queda en el 66,7%.

Resumen de Clasificadores para el conjunto de la NBA:

Una vez se ha entrenado el conjunto de entrenamiento con cada uno de los seis clasificadores, estamos en disposición de comparar los resultados para ver qué clasificadores se adaptan mejor a nuestro problema. En la Tabla 53 se recogen los resultados obtenidos con cada uno de los clasificadores:

	% Acierto	Kappa	Precisión Media
Red Bayesiana	76,43%	0.495	76.2%
Regresión Logística	71,83%	0.407	71.8%
Perceptrón Multicapa	67.81%	0.338	68.6%
OneR	69.54%	0.322	68.9%
J48	72.98%	0.378	74.7%
Random Forest	71.26%	0.373	70.7%

Tabla 53. Análisis del entrenamiento (Conjunto NBA)

Una vez entrenado el conjunto de datos con los partidos correspondientes a la liga norteamericana de baloncesto (NBA) es el momento de seleccionar qué dos clasificadores son los que mejor se han adaptado al problema planteado. Estos dos clasificadores escogidos serán implementados para que formen parte del sistema de predicción de resultados para los partidos de la NBA.

El mejor clasificador si nos fijamos en los datos obtenidos ha sido la Red Bayesiana, ya que tiene la mejor tasa de aciertos y precisión además de mostrar sus atributos la mayor correlación de entre todos los clasificadores y conjuntos entrenados. Los buenos datos de porcentaje de aciertos y precisión media podrían asegurar en el sistema de predicción que tres de cada cuatro partidos en los que se realice predicción sean acertados por el sistema.

El otro clasificador que fue elegido para ser implementado dentro del sistema de predicción de resultados es el árbol J48. Este algoritmo ha conseguido la segunda mejor tasa de aciertos de todo el conjunto de clasificadores, además de tener unos grandes datos de precisión, donde la precisión de partidos que se han clasificado como empate ha llegado al 80%. Además, aunque no tiene el segundo mejor valor para el estadístico kappa, su valor para este parámetro nos hace ver que los atributos que han sido elegidos para entrenar el modela tienen una correlación más que aceptable para formar parte del estudio.

Como se puede apreciar, la disminución del espacio de resultados respecto a las competiciones de fútbol (se ha pasado de 3 resultados posibles a 2) ha hecho que aumenten todos los parámetros relevantes para el estudio de manera considerable. Este aumento también repercutirá presumiblemente en unas mejores tasas de aciertos en la fase de test para este conjunto de partidos.

7.3 ANEXO C: Clasificadores escogidos para el sistema

7.3.1 Clasificadores Competiciones Europeas

Los clasificadores elegidos para el conjunto que recopila los partidos de competiciones europeas (Champions League y Europa League) han sido los siguientes:

J48:

Como se puede apreciar en la Figura 115, el árbol sólo necesita los datos de las cuotas del equipo local y visitante para hacer las predicciones. La deficiencia que presenta este modelo es que no puede predecir empates, por lo que este punto tendrá que ser tratado en la fase de explotación del modelo, donde se intentará cubrir este resultado para evitar que la tasa de fallos sea demasiado alta.

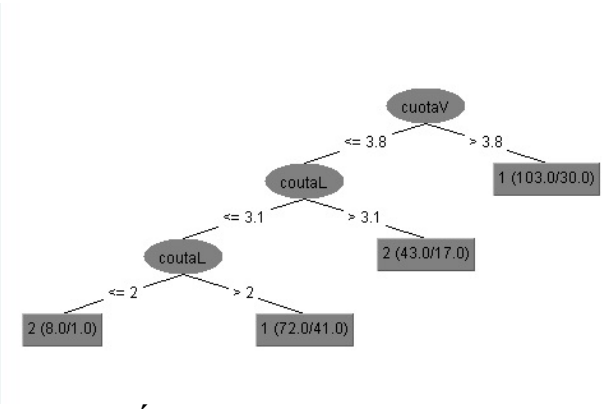

Figura 115. Árbol J48 – Competiciones Europeas

Red Bayesiana:

Para este clasificador bayesiano se han seleccionado cinco atributos, los cuales maximizaban la tasa de aciertos del clasificador (Tabla 54). Estos atributos son las tasas de victoria local y visitante de las casas de apuestas, el coeficiente UEFA del país al que pertenece el equipo local y el TOP al que pertenece el país del equipo local y visitante.

Cuota Local		1	X	2
	Baja	0,518	0,267	0,058
	Alta	0,482	0,733	0,942
Cuota Visitante		1	X	2
	Baja	0,338	0,595	0,875
	Alta	0,662	0,405	0,125
Coef. Local		1	X	2
	Baja	0,221	0,284	0,642
	Alta	0,779	0,716	0,358
TOP Local		1	X	2
	TOP4	0,316	0,277	0,102
	TOP10	0,342	0,378	0,160
	TOP20	0,236	0,176	0,350
	TOP30	0,004	0,008	0,008
	TOP50	0,102	0,160	0,350
TOP Visitante		1	X	2
	TOP4	0,200	0,328	0,268
	TOP10	0,289	0,277	0,415
	TOP20	0,253	0,261	0,187
	TOP30	0,004	0,008	0,008
	TOP50	0,253	0,126	0,122

Tabla 54. Coeficientes Red Bayesiana – Competiciones Europeas

Los rangos que dividen en Alta-Baja los valores de las cuotas y del coeficiente UEFA del equipo local son los siguientes:

Cuota Local: 1,635 (División entre Baja-Alta)

Cuota Visitante: 3,85 (División entre Baja-Alta)

Coeficiente UEFA Local: 29962,5 (División entre Baja-Alta)

7.3.2 Clasificadores Champions League

Los dos clasificadores que han sido elegidos para implementar el sistema de predicción para la competición de la Champions League han sido los siguientes:

J48:

Como podemos apreciar en el árbol mostrado en la Figura 116, el clasificador J48 ha generado un árbol muy sencillo que sólo necesita un nodo para decidir sobre el resultado del partido. El nodo utiliza el atributo de la cuota del equipo local para realizar las predicciones. Como puede observarse, la tasa de aciertos para partidos en los que se predice una victoria local es cercana al 80%, mientras que para victorias visitantes se aproxima al 50%.

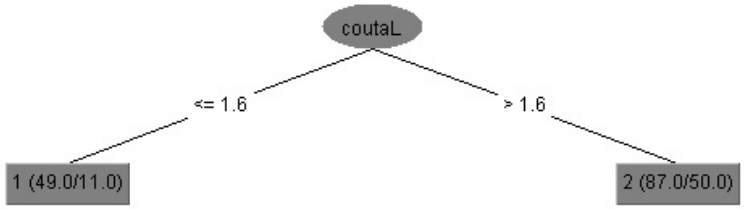

Figura 116. Árbol J48 – Champions League

Una vez más, para evitar que el algoritmo acumule un gran número de errores por fallo en la predicción de empates, se cubrirá este tipo de resultado en la fase de explotación del sistema. De esta manera se intentará conseguir una tasa de aciertos de más del 70% de los partidos analizados.

Red Bayesiana:

Para este clasificador bayesiana se han seleccionado cuatro atributos, los cuales maximizaban la tasa de aciertos del clasificador (Tabla 55). Estos atributos son las tasas de victoria local y visitante de las casas de apuestas y el TOP al que pertenece el país del equipo local y visitante.

Cuota Local		1	X	2
	Baja	0,592	0,258	0,085
	Alta	0,408	0,742	0,915
Cuota Visitante		1	X	2
	Baja	0,454	0,742	0,939
	Alta	0,546	0,258	0,061
TOP Local		1	X	2
	TOP4	0,519	0,507	0,176
	TOP10	0,278	0,246	0,365
	TOP20	0,117	0,195	0,159
	TOP50	0,087	0,050	0,300
TOP Visitante		1	X	2
	TOP4	0,248	0,536	0,529
	TOP10	0,429	0,217	0,294
	TOP20	0,207	0,166	0,065
	TOP50	0,117	0,079	0,112

Tabla 55. Coeficientes Red Bayesiana – Champions League

Los rangos que dividen en Alta-Baja los valores de las cuotas son los siguientes:

Cuota Local: 1,625 (División entre Baja-Alta)

Cuota Visitante: 5,375 (División entre Baja-Alta)

7.3.3 Clasificadores Europa League

Los dos clasificadores seleccionados para esta competición de Europa League han sido los siguientes:

J48:

En esta ocasión el árbol que ha generado la herramienta WEKA es algo más complejo que el de los conjuntos anteriormente mostrados, tal y como se observa en la Figura 117. En cualquier caso, aunque el árbol es más profundo que el de conjuntos anteriores, tan sólo se han utilizado tres atributos para construirlo (las tasas de las casas de apuestas para las victorias del equipo local y visitante además de el coeficiente UEFA del país al que pertenece el equipo local).

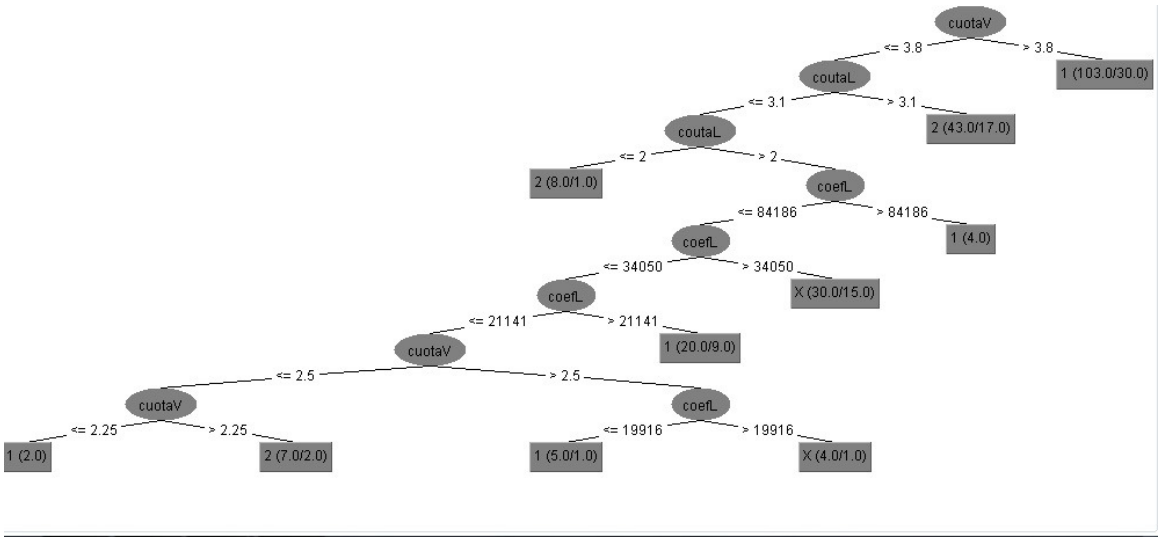

Figura 117. Árbol J48 – Europa League

La profundidad y la cantidad de nodos que posee el árbol generado es que el algoritmo ha estimado oportuno dividir cada uno de los atributos en varias franjas de valores, por lo que cada vez que se decide evaluar una de esas franjas debe haber una nueva división del nodo padre.

A diferencia de casos anteriores, este árbol es capaz de predecir empates, no obstante habrá que estar atentos en la fase de pruebas si estas predicciones son o no correctas, ya que se podría necesitar cubrir este resultado para evitar que la tasa de fallos sea muy elevada.

Red Bayesiana:

En cuanto a la Red Bayesiana que ha sido diseñada a partir del conjunto de datos de la competición de la Europa League, podemos ver en la Tabla 56 la tabla de coeficientes de la red, coeficientes que serán utilizados para el cálculo del resultado más probable de los tres posibles.

Para este clasificador bayesiana se han seleccionado cinco atributos, los cuales maximizaban la tasa de aciertos del clasificador. Estos atributos son las cuotas de victoria local y visitante de las casas de apuestas, el TOP al que pertenece el país del equipo local y visitante y el coeficiente UEFA del país al que pertenece el equipo local

Cuota Local		1	X	2
	Baja	0,518	0,267	0,058
	Alta	0,482	0,733	0,942
Cuota Visitante		1	X	2
	Baja	0,338	0,595	0,875
	Alta	0,662	0,405	0,125

TOP Local		1	X	2
	TOP4	0,317	0,28	0,107
	TOP10	0,344	0,381	0,189
	TOP20	0,237	0,178	0,352
	TOP50	0,103	0,161	0,352
TOP Visitante		1	X	2
	TOP4	0,201	0,331	0,270
	TOP10	0,290	0,280	0,418
	TOP20	0,254	0,263	0,189
	TOP50	0,254	0,127	0,123
Coef. Local		1	X	2
	Baja	0.221	0.284	0.642
	Alta	0.779	0.716	0.358

Tabla 56. Coeficientes Red Bayesiana – Europa League

Los rangos que dividen en Alta-Baja los valores de las cuotas y del coeficiente UEFA del equipo local son los siguientes:

Cuota Local: 1,635 (División entre Baja-Alta)

Cuota Visitante: 3,85 (División entre Baja-Alta)

Coeficiente UEFA Local: 29962,5 (División entre Baja-Alta)

7.3.4 Clasificadores Competición de Liga

El conjunto de competición de liga recoge los datos de todos los partidos de las principales ligas del mundo. Para este conjunto de partidos también se han escogido dos clasificadores que serán los siguientes:

J48:

Como se puede ver en el árbol mostrado en la Figura 118, no es tan profundo como el del conjunto anterior, pero uno de sus niveles está muy ramificado debido a la inclusión del atributo de la Zona del equipo visitante.

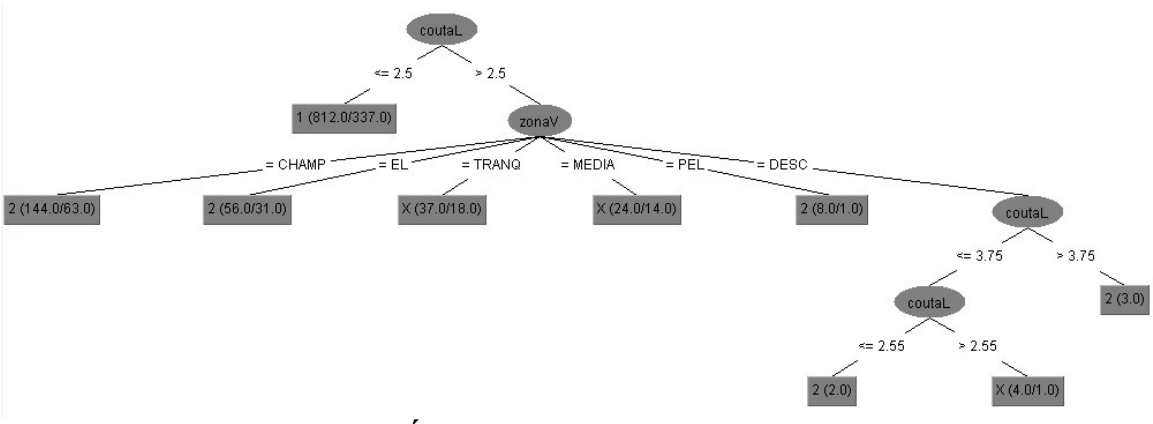

Figura 118. Árbol J48 – Competición de Liga

Para este clasificador se han utilizado tan solo dos atributos, que son la cuota que ofrecen las casas de apuestas por el equipo local y la zona de la clasificación en la que se encuentra el equipo visitante.

Finalmente, cabe destacar la capacidad del árbol para predecir empates, aunque la precisión que se ha registrado a la hora de acertar las predicciones no ha sido muy elevada. Por ello, se esperará a la fase de pruebas para determinar si es necesario o no realizar una cobertura de los empates en cada uno de los partidos.

Red Bayesiana:

En cuanto a la Red Bayesiana que ha sido diseñada a partir del conjunto de datos de la competición Liga, podemos ver en la Tabla 57 los coeficientes de la red, coeficientes que serán utilizados para el cálculo del resultado más probable de los tres posibles.

Cuota Local		1	X	2
	Baja	0,227	0,039	0,030
	Media	0,656	0,670	0,485
	Alta	0,116	0,292	0,485
Cuota Visitante		1	X	2
	Baja	0,031	0,071	0,248
	Media	0,205	0,372	0,382
	Alta	0,565	0,526	0,343
	MuyAlta	0,199	0,031	0,026
Zona Visitante		1	X	2
	CHAMP	0,121	0,195	0,356
	EL	0,158	0,115	0,156
	TRANQ	0,153	0,214	0,104
	MEDIA	0,244	0,184	0,160
	PEL	0,160	0,122	0,118
	DESC	0,164	0,170	0,107

Tabla 57. Coeficientes Red Bayesiana – Competición de Liga

Para este clasificador bayesiano se han seleccionado los tres atributos, que maximizaban la tasa de aciertos del clasificador. Estos atributos son las cuotas de victoria local y visitante de las casas de apuestas y la zona de la clasificación en la que se encuentra el equipo visitante.

Los rangos que dividen en los distintos subconjuntos los valores de los atributos son los siguientes:

Cuota Local: 1,475 (División entre Baja-Media), 2,525 (División entre Media-Alta)

Cuota Visitante: 2,025 (División entre Baja-Media), 3,175 (División entre Media-Baja), 7,075 (División entre Alta-Muy Alta).

7.3.5 Clasificadores Liga BBVA

El conjunto de la Liga BBVA recoge los datos de un conjunto de partidos de la primera división española de fútbol. Para este conjunto de partidos también se han escogido dos clasificadores que serán los siguientes:

J48:

Como se puede apreciar en la Figura 119, el árbol generado por la herramienta WEKA y su algoritmo J48 es muy sencillo. La predicción del resultado sólo tendrá en cuenta si la cuota del equipo visitante es mayor que dos. Si la cuota del equipo visitante es mayor que dos, el árbol clasificará el encuentro como victoria del equipo visitante.

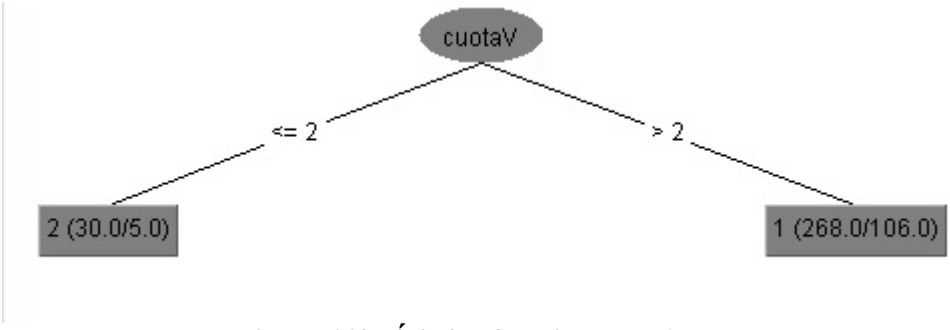

Figura 119. Árbol J48 – Liga BBVA

Si analizamos cuidadosamente lo que implica un árbol tan sencillo nos daremos cuenta de que este algoritmo ha tomado una actitud muy conservadora, y sólo predice una victoria del equipo visitante cuando las cuotas son bajas. Al ser la victoria del equipo local el resultado más probable, para casos en los que la cuota visitante es mayor que dos el algoritmo ha decidido clasificar a los partidos como victoria visitante, ya que así será como se obtenga una mayor tasa de aciertos.

Por el contrario encontramos una gran desventaja en este árbol, y es su imposibilidad de predecir empates, contrariedad que viene afectando en distintos modelos de predicción ya presentados con anterioridad. Al igual que se ha enunciado en modelos anteriores que

tenían el mismo problema, se intentará cubrir los resultados de empate para intentar reducir la tasa de fallos y así poder alcanzar un nivel de aciertos superior al 70%.

Tanto este clasificador como la red bayesiana que se expondrá a continuación, se utilizarán en conjunto con los modelos generados para las competiciones de liga para conformar el sistema de predicción de resultados para la competición de la Liga BBVA.

Red Bayesiana:

La Tabla 58 muestra los coeficientes obtenidos para la Red Bayesiana generada por la herramienta WEKA. Estos coeficientes serán utilizados posteriormente para la predicción del resultado que tiene más probabilidades de darse según la red bayesiana entrenada.

Cuota Local		1	X	2
	Baja	0,141	0,009	0,006
	Media	0,838	0.947	0,686
	Alta	0,021	0.044	0,308
Cuota Visitante		1	X	2
	Baja	0,021	0,044	0,321
	Media	0,688	0,912	0,610
	Alta	0,291	0,044	0,069
Zona Visitante		1	X	2
	CHAMP	0,104	0,198	0,377
	EL	0,158	0,129	0,164
	TRANQ	0,158	0,164	0,080
	MEDIA	0,277	0,164	0,142
	PEL	0,164	0,198	0,142
	DESC	0,140	0,147	0,130
Goles Anotados Visitante		1	X	2
	Baja	0,979	0,955	0,715
	Alta	0,021	0,045	0,285

Tabla 58. Coeficientes Red Bayesiana – Liga BBVA

Para esta red se han escogido 4 atributos, que son las cuotas de victoria del equipo local y visitante que ofrecen las casas de apuestas, la zona de la clasificación en la que se encuentra el equipo visitante y la media de goles anotados por el equipo visitante.

Los rangos que dividen en los distintos subconjuntos los valores de los atributos son los siguientes:

Cuota Local: 1,475 (División entre Baja-Media), 2,525 (División entre Media-Alta)

Cuota Visitante: 2,025 (División entre Baja-Media), 3,175 (División entre Media-Baja), 7,075 (División entre Alta-Muy Alta).

Goles Anotados Visitante: 2,105 (División entre Baja-Alta).

7.3.6 Clasificadores Liga Adelante

La siguiente competición para la que se van a mostrar sus clasificadores es la segunda división española de fútbol, también conocida como Liga Adelante. Para esta competición se han generado estos dos clasificadores:

J48:

Como se puede apreciar en la Figura 120, el clasificador J48 ha generado un árbol muy ramificado que tiene gran cantidad de nodos hojas. Ya se había hablado en apartados anteriores sobre la dificultad de predecir resultados en la Liga Adelante, ya que en esta competición los resultados no suelen guardar una cierta lógica como puede ocurrir en otro tipo de competiciones, donde el equipo favorito es el que suele salir vencedor del partido en una gran mayoría de los casos. Al ser una competición "caótica" el clasificador ha necesitado de muchas más reglas para maximizar, de ahí que el árbol tenga una gran cantidad de nodos hoja.

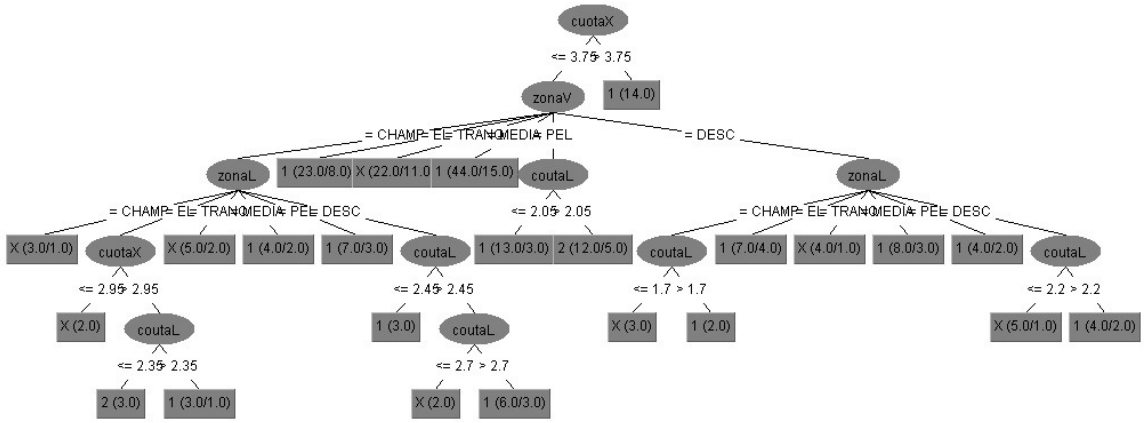

Figura 120. Árbol J48 – Liga Adelante

Además, la utilización de los atributos que hacen referencia a la zona en la clasificación de los dos equipos que disputan el partido ha hecho que el árbol esté muy ramificado, ya que para cada atributo hay 6 valores posibles.

Los atributos que necesita el árbol para realizar las predicciones son las cuotas de victoria local y empate que ofrece la casa de apuestas y las zonas en la clasificación en las que se encuentra tanto el equipo local como el visitante.

En este caso el árbol generado es capaz de predecir empates, ya que varios de sus nodos hojas se encargan de clasificar este tipo de resultado.

Red Bayesiana:

La Tabla 59 muestra los coeficientes obtenidos para la Red Bayesiana generada por la herramienta WEKA. Estos coeficientes serán utilizados posteriormente para la predicción del resultado que tiene más probabilidades de darse según la red bayesiana.

Cuota Local		1	X	2
	Baja	0.134	0.01	0.011
	Alta	0.866	0.99	0.989
Cuota Empate		1	X	2
	Baja	0.866	0.99	0.989
	Alta	0.134	0.01	0.011
Zona Local		1	X	2
	CHAMP	0.223	0.157	0.094
	EL	0.123	0.157	0.198
	TRANQ	0.159	0.120	0.135
	MEDIA	0.123	0.157	0.135
	PEL	0.150	0.139	0.156
	DESC	0.223	0.269	0.281

Tabla 59. Coeficientes Red Bayesiana – Liga Adelante

Para esta red se han escogido tres atributos, que son las cuotas de victoria del equipo local y la cuota de empate que ofrecen las casas de apuestas. Además, también se utilizará la zona de la clasificación en la que se encuentra el equipo local para realizar los cálculos que estime cuál de los tres resultados es el más probable que se dé.

Los rangos que dividen en los distintos subconjuntos los valores de los atributos son los siguientes:

Cuota Local: 1,475 (División entre Baja-Alta)

Cuota Empate: 0,775 (División entre Baja-Alta)

Con respecto a la utilización de estos dos clasificadores habrá que estar muy atentos, ya que la distribución de los resultados que suelen darse en esta competición puede hacer que sea necesario cubrir ciertos resultados para evitar que la tasa de aciertos sea muy baja. Esta decisión será tomada en la fase de pruebas tras analizar las tasas de aciertos que produce el sistema de predicción de resultados para esta competición.

7.3.7 Clasificadores Ligue 1

La siguiente competición para la que se van a mostrar sus clasificadores es la primera división francesa de fútbol, también conocida como Ligue 1. Para esta competición se han generado estos dos clasificadores:

J48:

La Figura 121 muestra el árbol generado por el algoritmo J48 para la competición de la Ligue 1. Este árbol es muy sencillo, como ya ha ocurrido en ocasiones anteriores, y se limita a simplemente comprobar sobre el nodo raíz el valor del atributo que recoge la cuota del equipo visitante. Si el valor de esta cuota es superior a 3, se clasificará el partido como victoria local, y en caso contrario se clasificará el partido como victoria visitante.

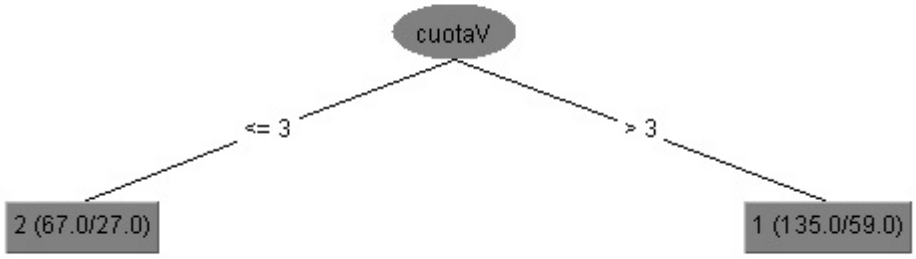

Figura 121. Árbol J48 – Ligue 1

La razón por la que aparece un árbol con una sola comprobación localizada en el nodo raíz es debido a que uno de los atributos está altamente correlacionado con la clase con la que se está clasificando. Por este motivo, la máxima tasa de aciertos se dará cuando se utilice este atributo sin ningún otro que distorsione los resultados.

De nuevo el inconveniente que aparece con este clasificador es que no vamos a ser capaces de predecir empates al utilizarlo. Una vez más, la solución será diseñada a partir de los resultados obtenidos al realizar las pruebas de clasificación con nuevos partidos de la competición. Si se comprueba que la tasa de aciertos no es buena, se puede optar por cubrir el resultado de empate para que la tasa de aciertos no se vea tan perjudicada por la deficiencia del modelo diseñado.

Red Bayesiana:

La Tabla 60 muestra los coeficientes obtenidos para la Red Bayesiana generada por la herramienta WEKA. Estos coeficientes serán utilizados posteriormente para la predicción del resultado que tiene más probabilidades de darse según la red bayesiana entrenada.

Cuota Visitante		1	X	2
	Baja	0.110	0.308	0.686
	Alta	0.890	0.692	0.314
Zona Visitante		1	X	2
	CHAMP	0.131	0.218	0.303
	EL	0.131	0.089	0.238
	TRANQ	0.153	0.218	0.107
	MEDIA	0.256	0.169	0.139
	PEL	0.165	0.153	0.074
	DESC	0.165	0.153	0.139

Tabla 60. Coeficientes Red Bayesiana – Ligue 1

Para esta red se han escogido tan solo dos atributos, que son las cuotas de victoria del equipo visitante que ofrecen las casas de apuestas y la zona de la clasificación en la que se encuentra el equipo visitante. A partir de estos dos atributos y más concretamente de los coeficientes asignados por la red para cada uno de los valores de los atributos, seremos capaces de calcular cuál es el resultado más probable del partido.

Los rangos que dividen en los distintos subconjuntos los valores de los atributos son los siguientes:

Cuota Visitante: 3,025 (División entre Baja-Alta)

Como se puede observar en esta competición en particular y en algunos otros clasificadores mostrados en general, muchos son los clasificadores que utilizan atributos correspondientes a datos del equipo visitante. Esto es debido a que a priori, el resultado más probable es la victoria del equipo local, pero la lógica nos dice que cuando un buen equipo juegue como visitante las probabilidades de que se dé una victoria local disminuyen aunque siguen siendo altas.

Es por ello que muchos algoritmos se basan sobre todo en las cuotas ofrecidas por las casas de apuestas para la victoria visitante o la zona en la que se encuentra el equipo visitante dentro de la clasificación de la competición. Incluso hemos llegado a ver que en la competición de la Liga BBVA se utilizaba la media de goles anotados por el equipo visitante para realizar los cálculos de probabilidad de los resultados posibles en la Red Bayesiana.

7.3.8 Clasificadores Premier League

La siguiente competición para la que se van a mostrar sus clasificadores es la primera división inglesa de fútbol, también conocida como Premier League. Para esta competición se han generado estos dos clasificadores:

J48:

Como se aprecia en la Figura 122, el clasificador generado a partir del algoritmo J48 es un árbol muy simple que necesita tan solo un atributo para clasificar los partidos con el resultado más probable. El único atributo utilizado para la clasificación de los encuentros de esta competición es la cuota de victoria del equipo local que ofrece la casa de apuestas.

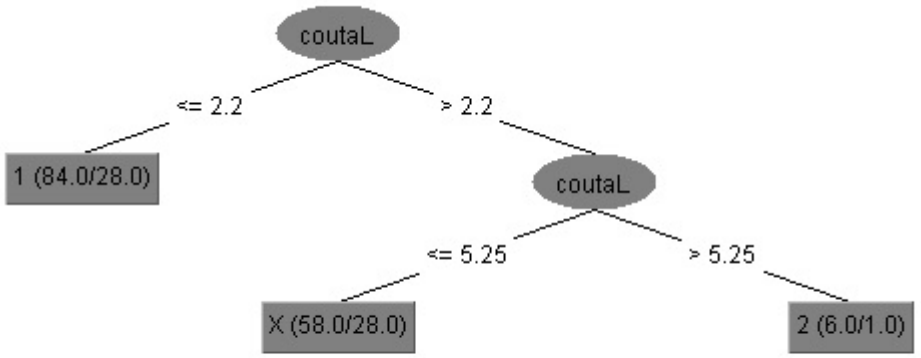

Figura 122. Árbol J48 – Premier League

Las características principales que presenta este árbol generado es que es capaz de predecir empates, ya que uno de sus nodos se encarga de clasificar este tipo de resultados. Además esta clasificación de empates se realiza en unos rangos de valores del atributo estudiado que permiten que estos empates no sean clasificados de manera residual (en ocasiones otros clasificadores clasifican partidos como empate en casos muy contados).

Por el contrario, las desventajas que presenta este clasificador son el escaso número de victorias visitantes que a priori se van a clasificar. Para que un partido sea clasificado como victoria visitante su cuota de victoria de equipo local tiene que ser superior a 5,25, un valor muy alto para este tipo de atributo. Como se puede ver en los nodos hoja del árbol, tan solo ha habido seis partidos de los 148 estudiados que han sido clasificados como victoria visitante. Esto implicará que pueda ser necesario cubrir ciertas victorias visitantes para intentar reducir la tasa de fallos si esta es muy elevada.

Red Bayesiana:

La Tabla 61 muestra los coeficientes obtenidos para la Red Bayesiana generada por la herramienta WEKA. Estos coeficientes serán utilizados posteriormente para la predicción del resultado que tiene más probabilidades de darse según la red bayesiana entrenada.

Cuota Visitante		1	X	2
	Baja	0,142	0,590	0,691
	Alta	0,858	0,410	0,309

Zona Local		1	X	2
	CHAMP	0,254	0,183	0,125
	EL	0,152	0,125	0,208
	TRANQ	0,123	0,087	0,097
	MEDIA	0,254	0,240	0,153
	PEL	0,138	0,106	0,264
	DESC	0,080	0,260	0,153

Tabla 61. Coeficientes Red Bayesiana – Premier League

Para esta red se han escogido tan solo dos atributos, que son las cuotas de victoria del equipo visitante que ofrecen las casas de apuestas y la zona de la clasificación en la que se encuentra el equipo local. A partir de estos dos atributos y más concretamente de los coeficientes asignados por la red para cada uno de los valores de los atributos, seremos capaces de calcular cuál es el resultado más probable del partido.

Los rangos que dividen en los distintos subconjuntos los valores de los atributos son los siguientes:

Cuota Visitante: 3,075 (División entre Baja-Alta)

En este caso la red generada es muy sencilla, en la que sólo son necesarios dos atributos para calcular qué resultado es el más probable para un partido con unos parámetros concretos. Esta simplicidad beneficia al proyecto de cara a la fase de implementación, ya que una red bayesiana sencilla ahorrará tiempos en la fase de implementación al tener que realizar un menor número de comprobaciones de atributos.

Analizando los coeficientes de los dos atributos de la red también se puede ver que no será muy normal que la red clasifique un partido como empate, ya que el único caso en el que el coeficiente de empate es superior al de los otros dos resultados es cuando el equipo local se encuentra en posiciones de descenso. Todos estos problemas serán tenidos en cuanta en las siguientes fases del proyecto para maximizar las tasas de acierto en cada una de las competiciones.

7.3.9 Clasificadores Serie A

La siguiente competición para la que se van a mostrar sus clasificadores es la primera división italiana de fútbol, también conocida como Serie A. Para esta competición se han generado estos dos clasificadores:

J48:

En árbol que se presenta en la Figura 123 es el que se ha generado a partir del algoritmo J48 para el conjunto de partidos de la Serie A italiana.

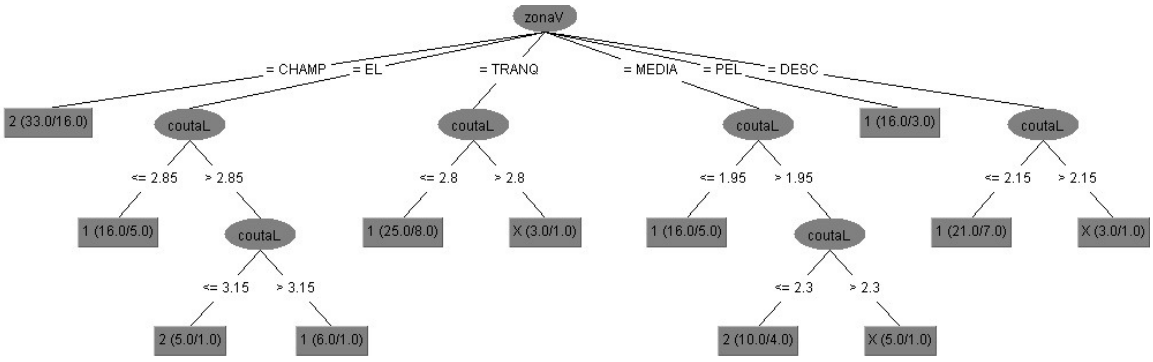

Figura 123. Árbol J48 – Serie A

Como se puede observar, es un árbol muy ramificado, hecho derivado de la utilización del atributo de la zona de la clasificación en la que se encuentra el equipo visitante. Este atributo tiene seis valores posibles, lo que hace que el nodo raíz se divida en seis ramas distintas.

El árbol tan solo necesita dos atributos para realizar las predicciones de resultados. Estos dos atributos son la ya mencionada zona en la clasificación del equipo visitante y la cuota facilitada por las casas de apuestas para la victoria del equipo local.

Las características del árbol le permiten predecir los tres tipos de resultados, aunque para el caso de los empates no se han registrado un gran número de predicciones de este tipo. Esto hará que haya que estar atentos en las próximas fases de prueba de los distintos clasificadores para así poder minimizar los errores que pueda provocar la escasez de predicciones con este resultado. Recordar que este resultado se suele dar en un 20% de los partidos, por lo que su no predicción implica tener ya ese mismo porcentaje de errores por defecto en el modelo.

Red Bayesiana:

La Tabla 62 muestra los coeficientes obtenidos para la Red Bayesiana generada por la herramienta WEKA. Estos coeficientes serán utilizados posteriormente para la predicción del resultado que tiene más probabilidades de darse según la red bayesiana entrenada.

Zona Local		1	X	2
	CHAMP	0,290	0,179	0,108
	EL	0,121	0,047	0,167
	TRANQ	0,121	0,160	0,167
	MEDIA	0,192	0,198	0,206
	PEL	0,156	0,274	0,167
	DESC	0,121	0,142	0,186

Zona Visitante		1	X	2
	CHAMP	0,094	0,179	0,363
	EL	0,192	0,198	0,186
	TRANQ	0,192	0,142	0,088
	MEDIA	0,192	0,217	0,206
	PEL	0,156	0,047	0,069
	DESC	0,174	0,217	0,088

Tabla 62. Coeficientes Red Bayesiana – Serie A

Para esta red se han escogido tan solo dos atributos, que son las zonas de la clasificación en la que se encuentra tanto el equipo local como el visitante. Este es el primer caso en el que ninguna de las cuotas ofrecidas por la casa de apuestas para los tres resultados posibles no forman parte del modelo de predicción.

Si bien es cierto, una comparación entre las zonas en la clasificación de los dos equipos es una buena comprobación para determinar el resultado de un partido, ya que cuanto mejor esté clasificado un equipo más probabilidades a priori debería tener para ganar un partido.

Si se analizan los coeficientes de la red, se podrá observar que están bien equilibrados para que se puedan predecir los tres tipos de resultados. No obstante, como hay que aplicar en el cálculo un coeficiente que refleja la probabilidad con la que se ha dado un resultado históricamente, al tener el empate el menor de estos coeficientes, puede provocar que no se clasifiquen muchos partidos con este resultado. Este hecho provocará que de nuevo se tenga que poner especial atención en las pruebas del modelo para ver si esta problemática afecta en exceso a los resultados obtenidos.

7.3.10 Clasificadores Partidos Internacionales

El siguiente conjunto de datos que va a ser analizado es el que corresponde a los partidos de las selecciones nacionales y que son tomados tanto de partidos amistosos como de competiciones oficiales. Para esta competición se han generado estos dos clasificadores:

J48:

La figura 124 muestra que de nuevo un árbol muy sencillo ha sido el que ha maximizado la tasa de aciertos para el conjunto que estaba siendo estudiado. Para este caso, el algoritmo sólo ha necesitado de la cuota del equipo local para realizar la clasificación de todos los encuentros del conjunto.

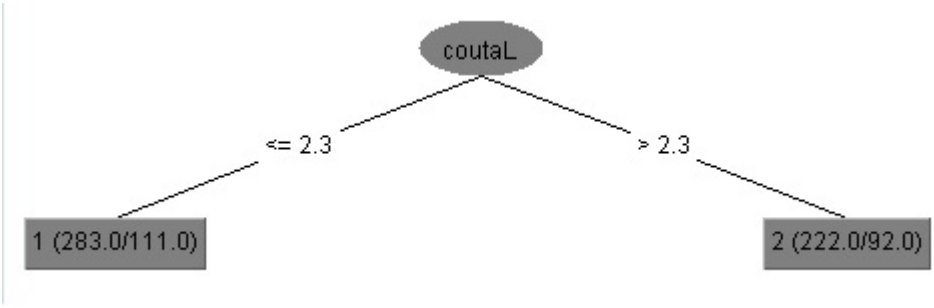

Figura 124. Árbol J48 – Partidos Internacionales

El gran problema de tener árboles tan sencillos en los que sólo el nodo raíz es el que discrimina entre un resultado u otro es que no seremos capaces de predecir empates con este algoritmo. Sin embargo, si nos fijamos en los porcentajes de acierto que consigue cada nodo hoja, tendremos que para el caso de partidos en los que se predice la victoria del equipo local tendremos más de un 60% de acierto, mientras que la tasa para partidos en los que se predice una victoria del equipo visitante la tasa de aciertos también se queda cerca del 60%. Si a eso sumamos que gran parte de los errores de clasificación están siendo debidos a que no se pueden clasificar empates con este modelo, tendremos que las tasas de aciertos en donde cubrimos el resultado de empate van a ser muy elevadas. En cualquier caso, será en la fase de pruebas de los modelos donde se tomarán las decisiones correspondientes para intentar minimizar los fallos producidos por las características que presenta este modelo.

Por último analizar el valor de la cuota de victoria del equipo local que se ha establecido como corte para clasificar los partidos. En este caso la cuota ha sido establecida a 2.3, una cuota relativamente alta para ser del equipo local, lo que hará que el algoritmo tienda a predecir más victorias locales que visitantes.

Red Bayesiana:

La Tabla 53 muestra los coeficientes obtenidos para la Red Bayesiana generada por la herramienta WEKA. Estos coeficientes serán utilizados posteriormente para la predicción del resultado que tiene más probabilidades de darse según la red bayesiana entrenada.

Cuota Local		1	X	2
	Baja	0,520	0,186	0,054
	Media	0,436	0,610	0,527
	Alta	0,044	0,203	0,420
Cuota Visitante		1	X	2
	Baja	0,044	0,169	0,403
	Media	0,409	0,593	0,527
	Alta	0,547	0,238	0,070

Tabla 63. Coeficientes Red Bayesiana – Partidos Internacionales

Para esta red se han escogido tan solo dos atributos, que son las cuotas que ofrecen las casas de apuestas para las victorias del equipo local y visitante. Los rangos que dividen en los distintos subconjuntos los valores de los atributos son los siguientes:

Cuota Local: 1,59 (División entre Baja-Media), 3,95 (División entre Media-Alta)

Cuota Visitante: 1,815 (División entre Baja-Media), 5,125 (División entre Media-Alta)

Si se analizan los coeficientes de la red, se podrá observar que están muy bien equilibrados para que se puedan predecir los tres tipos de resultados. Como se puede ver para cada una de las divisiones que se ha hecho en el atributo predomina un tipo de resultado en cuanto a valor del coeficiente (cuando la cuota está en la franja Media el coeficiente de empate es el más alto, mientras que si la cuota está en una de las otras dos franjas predomina la victoria del equipo local o visitante, dependiendo del atributo que se esté mirando).

El hecho de que estén tan bien equilibrados los coeficientes de la red bayesiana nos hace pensar que se van a obtener unos resultados muy prometedores en la fase de pruebas, ya que no tendremos muchos problemas derivados de la mala predicción de ciertos resultados como se ha podido ver en ejemplo de conjuntos entrenados anteriormente.

7.3.11 Clasificadores NBA

El siguiente conjunto de datos que va a ser analizado es el que corresponde a los partidos de la competición más importante del baloncesto norteamericano, que es la NBA. Para esta competición se han generado estos dos clasificadores:

J48:

Una vez más un árbol muy sencillo ha sido el que ha conseguido maximizar la tasa de aciertos para el conjunto que estaba siendo estudiado (Figura 125). Para este conjunto de la NBA, el algoritmo sólo ha necesitado la cuota ofrecida por la casa de apuesta para la victoria del equipo local para discriminar y realizar las predicciones.

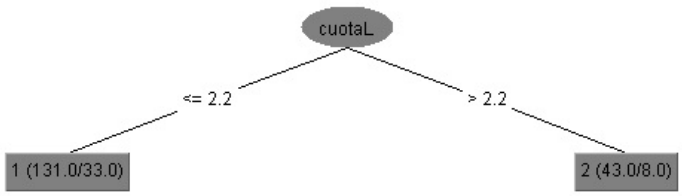

Figura 125. Árbol J48 – NBA

Al contrario que en conjuntos anteriores, el tener un árbol tan simple trae multitud de ventajas. En primer lugar, recordar que se está analizando un conjunto de partidos que pertenecen a una competición de baloncesto, por lo que sólo hay dos resultados posibles, y no tres como venía ocurriendo en los anteriores conjuntos de fútbol estudiados. Esta

variación en el espacio de resultados hace que con una sola comprobación el árbol sea capaz de clasificar entre los dos conjuntos de resultados posibles, además logrando unas tasas de acierto muy buenas.

Como se puede observar en los datos ofrecidos por los nodos hoja, la tasa de acierto conseguida al realizar el entrenamiento del clasificador ha sido del 74% para partidos clasificados como victoria local y un 81% para partidos clasificados como victoria visitante, unos porcentajes muy por encima del 50% de posibilidades que se tienen a priori en un partido de baloncesto.

Por último, cabe analizar el valor de la cuota de victoria del equipo local que se ha establecido como corte para clasificar los partidos. En este caso la cuota ha sido establecida a 2,2, una cuota bastante habitual para esta competición, y que suele ser encontrada para casos en los que el equipo local no ocupa un buen lugar en las clasificaciones de su conferencia.

Red Bayesiana:

La Tabla 64 muestra los coeficientes obtenidos para la Red Bayesiana generada por la herramienta WEKA. Estos coeficientes serán utilizados posteriormente para la predicción del resultado que tiene más probabilidades de darse según la red bayesiana entrenada.

Cuota Local		1	2
	Baja	0,921	0,486
	Alta	0,079	0,514
Cuota Visitante		1	2
	Baja	0,079	0,514
	Alta	0,921	0,486
% Victorias Local		1	2
	Baja	0,397	0,746
	Alta	0,603	0,254
%Victorias Visitante		1	2
	Baja	0,565	0,167
	Alta	0,435	0,833
%Victorias Discreto Local		1	2
	MuyMal	0,059	0,203
	Mal	0,123	0,189
	Regular	0,178	0,259
	Normal	0,078	0,105
	Bueno	0,215	0,049
	MuyBueno	0,269	0,105
	Excelente	0,269	0,091

%Victorias Discreto Visitante	1	2
MuyMal	0,142	0,035
Mal	0,196	0,063
Regular	0,196	0,063
Normal	0,050	0,119
Bueno	0,114	0,077
MuyBueno	0,123	0,287
Excelente	0,178	0,357

Tabla 64. Coeficientes Red Bayesiana – NBA

Esta Red Bayesiana es la más compleja de todas las que se han generado, ya que posee hasta 6 atributos diferentes para realizar la clasificación de los partidos. Los rangos que dividen en los distintos subconjuntos los valores de los atributos son los siguientes:

Cuota Local: 2,225 (División entre Baja-Alta)

Cuota Visitante: 1,685 (División entre Baja-Alta)

%Victorias Local: 48,3 (División entre Baja-Alta)

%Victorias Visitante: 45,95 (División entre Baja-Alta)

Si se analizan los coeficientes de la red, se podrá observar que están muy bien equilibrados para que se puedan predecir los dos tipos de resultados en igualdad de condiciones. La gran cantidad de atributos que tiene esta red hace que el número de posibles combinaciones de coeficientes sea lo suficientemente grande como para asegurarnos de que vamos a tener el espacio de resultados correctamente diversificado.

GLOSARIO

Apuesta Combinada: *Apuesta múltiple que incluye varios partidos o eventos*

Apuesta de Sistema: *Apuesta en la que a pesar de combinar varios eventos, el fallo de uno de ellos no implica la pérdida de la apuesta*

Árbol de Decisión: *Modelo de predicción utilizado en Inteligencia Artificial. Las decisiones del modelo se toman en función del nodo en el que nos encontramos y las bifurcaciones disponibles*

ASFR: *Apuesta de Sistema en Función del Riesgo*

CC: *Combinada de Competición*

Clasificador: *Algoritmo utilizado para asignar un elemento entrante no etiquetado en una categoría concreta conocida*

Cuota: *Indicador de las ganancias potenciales asociadas a una apuesta*

CV *Combinada Variada*

Doble Oportunidad: *Tipo de apuesta en la que se apuesta por dos de los tres resultados posibles. Normalmente uno de esos dos resultados es el empate.*

FIFA: *Fédération Internationale de Football Association*

J48: *Implementación en la herramienta WEKA del algoritmo de clasificación C4.5*

NBA: *National Basket Association*

Perceptrón Multicapa: *Red neuronal artificial formada por varias capas que permite resolver problemas que no son linealmente separables*

PIAS: *Partido Individual Apuesta Simple*

PIDO: *Partido Individual Doble Oportunidad*

Red Bayesiana: *Modelo probabilístico que relaciona un conjunto de variables aleatorias mediante un grafo dirigido que indica explícitamente influencia causal*

SG15: *Selección Genética con Lucky 15*

UEFA: Union of European Football Associations

WEKA: *Waikato Environment for Knowledge Analysis*

REFERENCIAS

[1] *Diario El Pais.* **Hacienda apuesta y gana – Ramón Muñoz. Disponible [Internet]:** <http://economia.elpais.com/economia/2012/05/25/actualidad/1337971909_363586.html> **[27 de mayo de 2012]**

[2] *WEKA*. **Machine Learning Group at University of Waikato. Disponible [Internet]:** <http://www.cs.waikato.ac.nz/ml/index.html>

[3] *Ian Witten, Eibe Frank, Mark Hall*. **Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann Publishers, 2011.**

[4] *Eclipse.* **The Eclipse Foundation open source community website. Disponible [Web]: <**http://www.eclipse.org/>

[5] *Real Academia Española.* **Diccionario de la Legua Española – Vigésima segunda edición. Disponible [Internet]: <**http://lema.rae.es/drae/>

[6] *Red Eléctrica de España S.A.* **Operación del Sistema. Disponible [Internet]:** <http://www.ree.es/operacion/operacion_sistema.asp>

[7] *Red Eléctrica de España S.A.* **Perfil de la Empresa. Disponible [Internet]:** <http://www.ree.es/quien_es/presentacion.asp>

[8] *Red Eléctrica de España S.A.* **Demanda de Energía Eléctrica en Tiempo Real. Disponible [Internet]:** <https://demanda.ree.es/demanda.html>

[9] *M. Kanamitsu, J.C. Alpert, K.A. Campana, P.M. Caplan, D.G. Deaven, M. Iredell, B. Katz, H.-L. Pan, J. Sela and G.H. White.* **Recent changes implemented into the Global Forecast System at NMC, 1991.**

[10] *Agencia Estatal de Meteorología*. **Actividades. Disponible [Internet]:** <http://www.aemet.es/es/quienes_somos/funciones>

[11] *Earth Magacine.* **Earthquake prediction: Gone and back again, 7 de abril de 2009. Disponible [Internet]: <**http://www.earthmagazine.org/article/earthquake-prediction-gone-and-back-again>

[12] *Telegraph*. **Italian earthquake: anger mounts over ignored warnings. Disponible [Internet]:** <http://www.telegraph.co.uk/travel/destinations/europe/italy/centralitaly/5115073/Italian-earthquake-anger-mounts-over-ignored-warnings.html> **[7 de abril 2009]**

[13] *Basilio Sierra Araujo.* **Aprendizaje Automático: Conceptos Básicos y Avanzados, Pearson Prentice Hall, 2006.**

[14] *Universidad Técnica Particular de Loja*. **Minería de Datos con WEKA para el diagnóstico preventivo del cáncer. Disponible [Internet]:** <http://www.utpl.edu.ec/eccblog/?p=3913> **[13 de febrero de 2011]**

[15] *Nir Friedman, Dan Geiger y Moises Goldszmidt.* **Bayesian Network Classifier**s. **Disponible [Internet]:** <http://link.springer.com/article/10.1023/A%3A1007465528199>

[16] *Slideshare*. **Comparación del Algoritmo Naïve Bayes con otras metodologías para la clasificación de correo electrónico no deseado. Disponible [Internet]:** <http://www.slideshare.net/paalvarador/estado-del-arte-4380689>

[17] *I.Rish.* **An empirical study of the Naïve Bayes Classifier. Disponible [Internet]:** <http://www.cc.gatech.edu/home/isbell/classes/reading/papers/Rish.pdf>

[18] *Association for Computer Machinery.* **Web principal de la asociación. Disponible [Internet]: <**http://www.acm.org/>

[19] *Boletín Oficial del Estado.* **Ley 13/2011, de 27 de mayo, de regulación del juego. Disponible [Internet]: <**http://www.boe.es/boe/dias/2011/05/28/pdfs/BOE-A-2011-9280.pdf>

[20] *La Voz de Málaga.* **Los juegos online seducen a 40.000 españoles. Disponible [Internet]: <**http://www.laopiniondemalaga.es/sociedad/2011/01/31/juegos-online-seducen-400000-espanoles/398490.html>

[21] *Play Store*. **Soccer Prediction. Disponible [Internet]:** <https://play.google.com/store/apps/details?id=de.sportwettenblogger.de.vorhersage&feature=search_result>

[22] *Play Store*. **Bet2Win. Disponible [Internet]:** <https://play.google.com/store/search?q=bet2win>

[23] *MARCA*. **Las Cracks. Disponible [Internet]:** <http://www.marca.com/blogs/los-cracks/>

[24] *Wikipedia*. **Coeficientes UEFA. Disponible [Internet]:** <http://es.wikipedia.org/wiki/Coeficientes_UEFA>

[25] *B.B.Chaudhuri, U.Bhattacharya.* **Efficient training and improved performance of Multilayer Perceptron in pattern classification. Disponible [Internet]:** <http://www.sciencedirect.com/science/article/pii/S0925231200003052>

[26] *Fédération Internationale de Football Association* (*FIFA)*. **Clasificación Mundial de la FIFA. Disponible [Internet]:** <http://es.fifa.com/worldranking/rankingtable/index.html>

[27] *NBA.* **Página oficial de la Liga Norteamericana de Baloncesto. Disponible [Internet]:** <http://www.nba.com/>

[28] *BWIN.* **Casa de Apuestas Online. Disponible [Internet]:** <https://www.bwin.es/>

[29] yr.no. **Información meteorológica proporcionada por el instituto meteorológico y la televisión pública noruega NRK. Disponible [Internet]:** <http://www.yr.no/>

[30] *UEFA.* **Coeficientes Nacionales 2012/2013. Disponible [Internet]:** <http://es.uefa.com/memberassociations/uefarankings/country/index.html>

[31] RR Bouckaert. **Bayesian Network Classifiers in WEKA. Disponible [Internet]:** <http://www.cs.waikato.ac.nz/ml/publications/2004/uow-cs-wp-2004-14.pdf> **[1 de Septiembre de 2004]**

[32] *BCS WHFREEMAN.* **Logistic Regression. Disponible [Internet]:** <http://bcs.whfreeman.com/ips5e/content/cat_080/pdf/moore16.pdf>

[33] *John A. Bullinaria.* **Learning in Multilayer Perceptrons. Back-Propagation. Disponible [Internet]:** <http://www.cs.bham.ac.uk/~jxb/INC/l7.pdf> **[2012]**

[34] *Gaya Buddhinath, Damien Derry*. **A simple enhancement to One Rule Classification. Disponible [Internet]:** <http://www.buddhinath.net/OtherLinks/Documents/Improved%20OneR%20Algorithm.pdf>

[35] *Sam Drazin, Matt Montag.* **Decision tree analysis using WEKA. Disponible [Internet]:** <http://www.samdrazin.com/classes/een548/project2report.pdf>

[36] *C. E. Shannon.* **A mathematical theory of communication, Bell System Technical Journal, vol. 27, pp. 379-423 and 623-656, July and October, 1948.**

[37] *Ned Horning.* **Introduction to decision trees and random forests. Disponible [Internet]:** <http://www.whrc.org/education/indonesia/pdf/DecisionTrees_RandomForest_v2.pdf>

[38] *Wikipedia.* **Teorema de Bayes. Disponible [Internet]:** <http://es.wikipedia.org/wiki/Teorema_de_Bayes>

[39] *Webapuestas.* **Apuesta Combinada. Disponible [Internet]:** <http://www.webapuestas.com/apuesta-combinada.html>

[40] *Webapuestas.* **Apuestas de Sistema. Disponible [Internet]:** <http://www.webapuestas.com/guia-tutorial-apuestas/tipos-apuestas/apuestas-de-sistema/index.html>

[41] *Webapuestas.* **Sistema 2/3. Disponible [Internet]:** <http://www.webapuestas.com/guia-tutorial-apuestas/tipos-apuestas/apuestas-de-sistema/sistema-2-3.html>

[42] *Universidad del País Vasco.* **Algoritmos Genéticos. Disponible [Internet]:** <http://www.sc.ehu.es/ccwbayes/docencia/mmcc/docs/temageneticos.pdf>

[43] OnCourt. **Software con información histórica de partidos de tenis. Disponible [Internet]:** <http://www.oncourt.info/>