
**Departamento de teoría de la señal**

Trabajo de Fin de Grado

# ANÁLISIS Y PREDICCIÓN PROBABILÍSTICA DE RESULTADOS DEPORTIVOS: TENIS

**Autor: Raúl Gómez-Álvarez Lobón**

**Tutor: Víctor Elvira Arregui**

Leganés, 27 de septiembre de 2015

# Agradecimientos

Se termina el último capítulo de una de las etapas más importantes de mi vida. Llegando al final, tan sólo puedes echar la vista atrás y ver de dónde vienes y lo que te ha costado llegar hasta aquí, algo que no hubiese sido posible sin el apoyo de las personas que voy a mencionar a continuación y que se merecen el agradecimiento y reconocimiento.

A mis padres, Josefa y José Ángel, y a mi hermano Álvaro por su apoyo en todos los sentidos y por estar conmigo desde el principio, también en los momentos de más tensión, épocas de estudio y exámenes. Desde pequeño os daba la lata con que quería ser ingeniero de telecomunicaciones, ya no va a hacer falta que insista más.

A mi novia, María, por ser mi otra mitad. Gracias por estar conmigo cada minuto, por animarme, por quererme, por apoyarme, por creer en mí y por darme los empujones hacia arriba necesarios en los momentos en los que lo necesitaba. Ya hemos llegado hasta aquí, ahora sólo nos queda soñar.

A mi tutor, Víctor, por estar disponible en todo momento para contestar mis dudas, poniendo todos los materiales a mi disposición desde el principio y facilitándome el proceso del trabajo.

A mis compañeros de clase durante estos años, Javi, Elena, Iriberri, Ana, Cris, Lechu, Gonzalo, Fer, Carlos, Andrea, Chan, Manu, Álvaro, Nevado, Rober, Emilio, Estefanía, Saúl, Jaime, Edu, Juancar, no quisiera dejarme a nadie porque sois todos muy grandes. Atrás quedan jornadas maratonianas en la biblioteca, exámenes interminables, montañas de apuntes esperando y anécdotas únicas. Pero ya estamos aquí, lo hemos conseguido.

A toda la demás familia, amigos y compañeros de la universidad que han estado conmigo, gracias por estar ahí durante todos estos años, en las buenas y en las malas.

# Resumen

Hoy en día, deporte, tecnología y apuestas son tres sectores con un gran crecimiento y que juegan un papel importante en la sociedad. En el presente proyecto uniremos los tres con el objetivo de desarrollar un modelo de predicción de resultados de partidos de tenis en los circuitos masculino y femenino que mejore el modelo previo existente.

Hablar de apuestas hoy en día puede parecer sencillo debido a la velocidad y rapidez con la que puedes acceder a ellas. Pero para plantearse un modelo de apuestas a largo plazo hay que contextualizar y analizar numerosos elementos que pueden afectarnos tales como el tipo de apuestas, las distintas casas que existen, las características del deporte al que vamos a apostar o los criterios que vamos a seguir, con el objetivo de obtener el máximo beneficio.

Para poder plantear un modelo con el cual podamos obtener beneficios a la hora de apostar, una vez decididos los factores mencionados anteriormente utilizaremos un enfoque Bayesiano y diferentes herramientas estadísticas a través del teorema de Bayes. Debemos seguir un determinado proceso para poder obtener parámetros que nos digan cómo se van a comportar los jugadores ante distintos partidos y así poder predecir el resultado de un encuentro de manera más sencilla. Este nuevo modelo que vamos a exponer tendrá como base técnicas como el aprendizaje máquina o la validación cruzada aplicadas sobre distintos factores determinantes en el tenis, con el fin de poder utilizar de la mejor manera toda la información a nuestra disposición que obtendremos de las distintas bases de datos. Con todo esto llegaremos al cálculo de la probabilidad que tienen los jugadores de ganar un encuentro, durante una serie de partidos de una o varias temporadas.

Para saber si los resultados de las probabilidades que obtenemos son buenos o no, utilizaremos funciones de coste. Debemos llegar al final del experimento obteniendo buenos resultados en esta parte, para luego poder apostar en una casa de apuestas online y comprobar el funcionamiento de nuestro modelo. El objetivo final es, como el de todo jugador, obtener algún tipo de beneficio.

Mucho se escribe sobre apuestas, estrategias, criterios y distintas formas de apostar. Nosotros estamos convencidos de que con el modelo riguroso que vamos a presentar los resultados, tanto a nivel cualitativo como cuantitativo, serán muy interesantes.

Palabras clave: modelos probabilísticos, apuestas, validación cruzada.

# Abstract

Nowadays, sports, technology and betting are three sectors with high growth and which play an important role in our society. In this project we will join the three together in order to develop a model for predicting outcome of tennis matches in the male and female tours to improve the existing previous model.

Speaking of betting today may seem easy due to the speed and quickness with which you can access them. However, to propose a long-term betting model, we must contextualize and analyze many elements that can affect such as the type of betting, the different sports bookmakers that exist, the characteristics of the sport that we will bet or the criteria that we will follow, in order to obtain the maximum benefit.

To propose a model with which we can obtain benefits when betting, once determined the factors mentioned above we will use a Bayesian approach and different statistical tools through the Bayes theorem. We must follow a certain process to obtain parameters that tell us how they will behave players to different matches and thus to predict the outcome of a match more easily. This new model that we will expose is technically based on the machine or cross-validation learning applied to different determinants in the tennis, in order to be able to use in the best way all the information at our disposal, what we will gain from the different bases of data. With all this, we will come to the calculation of the probability that players win a game, during a series of matches one or more seasons.

To find out if the results of the chances we get are good or not, we will use cost functions. We must get to the end of the experiment performing well in this part, then to to bet in online sports bookmarkers and finally to check the functioning of our model. The ultimate goal is, like every player, get some benefit.

Much is written about betting, strategies, criteria and other forms of betting. We are convinced that with the rigorous model that we will present the results, both qualitatively and quantitatively, will be very interesting.

Keywords: probabilistic models, betting, cross-validation.

# Contenidos

# Lista de figuras

# Lista de tablas

# 1. Introduction

## 1.1. Motivation of the project

Nowadays, sports, technology and betting are directly linked and growing together, playing an important role in our society. First of all, sport, besides being practiced by many people, is currently the most watched activity in the world: football has more than 3500 million of followers worldwide, basketball, around 400 million followers, and tennis, near from 1000 million followers, with an increasing media influence at all levels [1]. Moreover, technology, through internet and telecommunications, allows us to have any result of the remotest matches within hand's reach, wherever, with lots of available statistics and in the format that suits us. Furthermore, betting, which already have an consolidated and prolonged history, grows even more in a context of crisis like the one we live because of people who want to make some quick money through sports betting of all kinds, looking for a way out of its battered economy. it's not by chance that it is getting consolidated in a country like Spain, where the tradition of gaming is being maintained along the last decades, even with non-sports betting such as the Lottery and Euro Millions.

Therefore, the motivation for this project is, joining the passion for sports, technology and betting, to predict tennis matches, using different knowledge of statistics and probability, for betting online. It should also be noted that the sports betting industry is in high growth, having increased online betting 44% in Spain over the last year [2].

In addition to joining the three passions with more development at all levels in our days, we can see that there is a greater incorporation of women into the world of sport, placing different female categories at a high level that makes women's sport grow increasingly. Therefore, there is a motivation for placing women's sport at the place where it should be and to see how level, characteristics and peculiarities of women's tennis have grown.

## 1.2. Objectives

As we initially have an existing that predicts results of sports betting in men's tennis, our goal in this project is to propose new models that allow us to get better results qualitatively and quantitatively than the previous model and also allow us to analyze more in depth the male and female tours.

## 1.3. Structure of the document

- **2. Problem statement.-** This section twigs needed for the project will be made. To do this, a brief history of sports betting will be exposed, and later the sport that we will bet, the means at our disposal and other criteria will all be explained in detail, as well as a situation of the current legal framework in this area will be discussed.

- **3. Design of probabilistic model and machine learning.-** Once bases in paragraph 2 are fixed, we propose a new model that we will justify technically and mathematically, and introduce the process of cross-validation. Further we develop all the factors that make this proposed model.

- **4. Results and evaluation.-** Once the proposed model and fundamental factors are set, we show a wide set of results after application and we compare the results obtained by the previous model, which also will draw new arrangements for the comparative process more conclusive. We seek, in addition to good quantitative results, obtain conclusions about the male and female tours and its players at a qualitative level after the application of the new model.

- **5. Budget and work planning.-** We analyze all costs that the project has led and the definition of its phases, sequencing and timing, including comments on each part.

- **6. Conclusions.-** We carefully expose all the conclusions we have drawn from the development of the project and we will pose potential future research.

- **Bibliograpy**
- **Annex.- Summary**

# 2. Planteamiento del problema

## 2.1. Historia y desarrollo

Puede parecer extraño, pero las apuestas no son algo propio de la actualidad. Tienen su origen en Grecia, aunque después destacaron también en el Imperio Romano, donde ya se hacían apuestas a diferentes deportes y competiciones propias de la época mientras los espectadores animaban. El abanico de apuestas iba desde combates de gladiadores hasta carreras de cuádrigas. [3]

Las apuestas no sólo se mantuvieron en el tiempo, sino que además crecieron en la Edad Media haciéndose muy populares en los torneos de caballeros o el tiro con arco, de manera que fueron las propias apuestas las que conseguían elevar estas competiciones a lo más alto y hacerlas conocidas a todos los niveles. Aquí ya comenzamos a entender que las apuestas en competiciones y el espectáculo mediático son un binomio inseparable.

Sin embargo, no es hasta 1780 cuando comienza de verdad la revolución de las apuestas deportivas. En Inglaterra, y con las carreras de caballos como punta de lanza, comienza a hacerse negocio de manera más directa y se empiezan a abrir lugares destinados específicamente a las apuestas deportivas.

El salto a América se produce en los siglos XIX y XX donde, alrededor de la década de 1930, alcanzan su punto máximo de éxito ayudado por la difusión de los medios de comunicación y periódicos. Es precisamente en el siglo XX cuando llegan definitivamente a Europa, donde su desarrollo fue enorme desde el primer momento [4].

Finalmente, ya en el nuevo siglo y viendo el nicho de negocio que tenía el sector, se decide lanzar las apuestas online, facilitando así el acceso al juego y proporcionando una capacidad de apostar prácticamente inmediata. Esto, unido a la evolución de internet, ha hecho que el negocio de las apuestas no pare de crecer en los últimos años.

En España concretamente, la primera casa de apuestas presencial se abrió en Madrid en 2008 de la mano de Codere [5]. Aunque las apuestas deportivas más tradicionales han sido las carreras de caballos, la pelota vasca o la quiniela, esas tradiciones han ido ampliándose y dando paso a campos deportivos como el fútbol, el baloncesto o el tenis. En nuestro país en 2013/2014 se jugaron 29026.2 millones de €, cantidad que representa el 2.85% del PIB. Si se trata de apuestas online serían 5600.4 millones de €, un 0.6%

del PIB, cantidad que no para de ascender durante los últimos 4 años, ganándole terreno a las apuestas presenciales [6].

Todos estos datos nos demuestran no sólo que el sector de las apuestas deportivas está en continuo crecimiento y que el avance de la tecnología y las comunicaciones lo impulsan todavía más, sino que además las apuestas son algo inherente al ser humano que ha existido prácticamente desde los orígenes con todo tipo de competiciones y modalidades. Mientras sigan existiendo los humanos, seguirá habiendo jugadores entre ellos [7], y por tanto el mercado de las apuestas seguirá existiendo.

## 2.2. Medios para apostar: las casas de apuestas

Las casas de apuestas son el medio a través del cual apostamos nuestro dinero en distintos deportes y competiciones. Todas ellas nos ofrecen diversas modalidades a las que poder jugar otorgando a cada evento una cuota determinada que puede seguir varios formatos: americano, fraccional (utilizado en Reino Unido) y decimal (utilizado en Europa). Para poder establecer comparaciones con el modelo anterior, y dado que nuestras apuestas se van a centrar en casas que operan principalmente en España y en Europa, utilizaremos en todo momento la cuota decimal. Las ganancias con este tipo de cuotas se calculan de la siguiente manera:

$$g = mb \cdot q \tag{2.1}$$

Donde $g$ es ganancia, $mb$ dinero apostado y $q$ la cuota, que está formada por hasta dos decimales. Significa que la cantidad apostada se multiplica por la cuota indicada para obtener la ganancia, o lo que es lo mismo, la cuota decimal indica el número de unidades monetarias a cobrar por cada unidad apostada. [8]

Aplicado a un ejemplo, si la cuota del jugador por el que vamos a apostar es 2 y juego 2€, la cantidad total que obtendría es g = 2 · 2 = 4€. Hay que tener en cuenta que esto no indica las ganancias netas sino las brutas. Así, los beneficios netos los definiríamos:

$$b = mb \cdot (q - 1) \tag{2.2}$$

Por tanto, en el ejemplo anterior, b = mb · (q – 1) = 2 · (2 – 1) = 2€. Este factor es importante y caracteriza a las cuotas decimales, puesto que en ciertas ocasiones podríamos dejarnos llevar por la sensación de que hemos ganado mucho dinero cuando en realidad no es el beneficio neto el que obtenemos, pudiendo crear falsas expectativas. Alguien puede decirnos que ha tenido una ganancia de 101€ en una apuesta y podemos pensar que ha sido exitoso, cuando en realidad lo que le ha podido ocurrir es que ha apostado 100€ a un partido con una cuota muy baja de 1.01€, arriesgando demasiado para obtener un beneficio neto de 1€.

Una vez definido el tipo de cuota que vamos a utilizar vamos a explicar brevemente el funcionamiento de las casas de apuestas. Al inicio, en nuestro país existían dos funcionamientos posibles: por un lado existían aquellas casas que dejaban operar al mercado libremente utilizando un método de intercambio, como era el caso de Betfair, y por otro aquellas que jugaban contra el usuario, calculando probabilidades propias y estableciendo cuotas en consecuencia. Tras la nueva ley del juego online establecida en España, las casas de intercambio no podían operar y pasaron todas a funcionar con un sistema de apuesta contra el usuario. [14]

Apostar contra el usuario significa que la propia casa de apuestas, a través de mecanismos sofisticados, de personal cualificado y de potentes herramientas, establece una probabilidad propia para cada partido. En función de la misma, establece las cuotas para cada uno de los resultados posibles, con las que los usuarios deberán apostar y jugar dinero en función de lo que dicha casa de apuestas ha establecido. La probabilidad que ha calculado la casa de apuestas para un determinado resultado se puede obtener de la forma p = 1 / cuota. Por ejemplo, si en un partido entre Nadal y Federer, la cuota asignada a Nadal es 1.57€, la probabilidad que ha calculado la casa de apuestas de que gane Nadal es de p = 1 / 1.57 = 0.636. Si seguimos este proceso y calculamos las probabilidades de ganar de dos jugadores en un mismo partido, nos damos cuenta de que la suma de ambas probabilidades está por encima de 1. Esto significa que la casa de apuestas se reserva un margen de beneficio que normalmente va entre el 4% y el 15% dependiendo de la competición o el mercado. De esta manera, en las apuestas en las que el método es "la casa contra el usuario", si es el usuario el que consigue acertar un gran número de apuestas con las cuotas que establece la casa, esta siempre se reserva un margen de beneficio que le asegura un negocio próspero. Así, las casas de apuestas, además de vivir del error del jugador, se aseguran no perder en ciertos vasos. [9].

Lo vemos con un ejemplo. El 13/03/2012 se jugó un partido Berdych-Roddick. Las cuotas para apostar a la victoria del primero eran de 1.30€ y del segundo 3.40€ en Bet365. Así, tenemos que $p_1$ = 1 / 1.30 = 0.769 y $p_2$= 1 / 3.40 = 0.294. Si sumamos, $p_{total}$= $p_1$+ $p_2$= 1.063. Esto quiere decir que el margen que se reserva la casa es de alrededor del 6%. Esto nos da un indicador de lo difícil que va a ser ganar dinero apostando contra las casas de apuestas en un sistema de este tipo.

Ahora que ya conocemos la implantación de las casas de apuestas, las cuotas que nos ofrecen y su modo de funcionamiento, podemos ver como en el mercado existen infinidad de ellas ofreciendo servicios de todo tipo con todas las alternativas posibles. Vamos a pasar a analizar algunas de las más conocidas en España para poder posteriormente argumentar cuál es la mejor para desarrollar este proyecto [10]:

- **BET365.** Conocida a nivel europeo e internacional, cuenta con el sistema más desarrollado de apuesta online y apuestas en directo, con cuotas que siempre suelen ser las mejores del mercado. La oferta de apuestas es muy amplia a través de la web, que tiene una velocidad de respuesta y actualización instantánea con una interfaz muy sencilla. No cuenta con demasiadas posibilidades en bonos, regalándote el 100% del dinero que ingreses al abrir tu cuenta hasta 100€.

- **Betfair**. Fue la casa pionera en "trading" o fluctuación de cuotas en el mercado, pero lo tuvieron que retirar por el cambio de leyes en España. Posee una interfaz online sencilla pero poco directa a la hora de apostar. Sus puntos fuertes son la posibilidad de seguir los eventos en directo y las suculentas promociones para nuevos usuarios, llegando a ofrecerte 100€ por 10€ jugados en unas determinadas condiciones.

- **Sportium**. Nacida de la unión de las empresas Cirsa y Ladbrokes, se crea en 2007 pero no es hasta 2013 cuando se inicia en el juego online. Fue la primera casa de apuestas en aparecer en varios formatos, a través de web y aplicaciones para smartphones y tablets. Ofrece una gran variedad de apuestas con más de 60000 mercados, siendo este su punto fuerte. En cuanto a los bonos, Sportium ofrece el 100% de tu primer ingreso hasta 100€.

- **BWIN**. Es una de las casas más mediáticas gracias a su patrocinio de distintos equipos de fútbol. Tiene una interfaz fácil y directa, similar a la de BET365 y muy intuitiva. Pone a nuestra disposición infinidad de estadísticas de jugadores y equipos y eventos en directo, ofreciéndonos variadas promociones. Su punto fuerte son la gran cantidad de métodos de pago que nos permite, muchos más que el resto de casas de apuestas.

- **William Hill**. Una casa con gran tradición en el Reino Unido que ahora da el salto al resto de Europa. Tiene una interfaz fácil, sencilla e intuitiva. Destaca por tener una explicación en la propia web de cada tipo de apuesta, sirviendo como guía para que no nos perdamos en ningún momento. Además de eventos en streaming, ofrece el 100% de tu primer ingreso hasta 100€ al igual que Sportium.

- **Luckia**. Es la casa de apuestas que vamos a analizar de más reciente creación y que opera únicamente en España. A pesar de eso es propiedad de Egasa, empresa con experiencia en el sector del juego. La web presenta una interfaz fácil pero distinta, más interactiva de lo normal. Su punto fuerte está en las promociones de hasta 120€ con tu primer ingreso.

Por tradición (más de 40 años desde que se fundó), experiencia tanto en Europa como en Reino Unido, sus buenas cuotas pero sobre todo, por su servicio online, su interfaz y su seguridad, la casa de apuestas que utilizaremos será BET365. No buscamos en este caso grandes bonos de bienvenida, ni amplios mercados de todos los eventos porque sólo vamos a apostar en tenis. Buscamos algo rápido, seguro, directo y con buenas cuotas, y BET365 es la mejor opción. Además cuenta con protocolos de seguridad, sistemas de pago seguro y cifrado de datos avanzado, muy parecido al de los bancos, que nos protege desde que creamos la cuenta hasta que introducimos o sacamos dinero de la misma, con un servicio técnico de calidad.

Una vez definido el formato de cuota y la casa de apuestas que vamos a utilizar, podemos observar que en todas ellas hay un enorme abanico de posibilidades y modalidades de apuestas, entre las que destacan las siguientes [11]:

- **Apuestas sencillas**. Son las más habituales. Elegimos el deporte, el evento al que queremos apostar y su opción determinada, y la cantidad que vamos a jugar. Ejemplo: apostamos 2€ a que gana Nadal a Murray con una cuota de 1.27€.

- **Apuestas combinadas**. Son aquellas en las que seleccionamos todos los pronósticos que queramos, recibiendo ganancias sólo si acertamos todos ellos. En el momento en el que fallemos uno, no ganaremos nada. El dinero total que podremos ganar saldrá de multiplicar el dinero que vamos a jugar por cada una de las cuotas de los eventos que vamos a añadir a nuestra combinada. Ejemplo: apostamos a que Nadal gana a Federer con una cuota de 1.54€, que Berdych gana a Murray con una cuota de 2.10€ y que Djokovic gana a Del Potro con una cuota de 1.20€. Si jugamos una cantidad de 2€ y acertamos los tres pronósticos, las ganancias serían $g = 2 \cdot 1.54 \cdot 2.10 \cdot 1.20 = 7.76$€. En el momento en el que fallemos uno de esos tres partidos, las ganancias serán 0€.

- **Apuestas de sistema.** Son apuestas en las que, al igual que en las combinadas, podemos seleccionar varios pronósticos a la vez, con la diferencia de que se nos permite fallar algunos de ellos. Existen apuestas de sistema prácticamente infinitas, entre las que destacan el patent o el sistema 2/3. Este último consiste en agrupar las apuestas de dos en dos para selecciones de tres eventos, realizando combinaciones de ganancias por separado. Así, la posibilidad de recuperar algo de dinero es mayor, pero la cantidad apostada es más elevada. Ejemplo: apostamos a que Nadal gana a Federer con una cuota de 1.54€, que

Berdych gana a Murray con una cuota de 2.10€ y que Djokovic gana a Del Potro con una cuota de 1.20€ con un sistema 2/3. Si ganan Nadal y Berdych y pierde Djokovic, los beneficios son de $g_1$= 1.54 · 2.10 = 3.23€. Si ganan Nadal y Djokovic y pierde Berdych, los beneficios son $g_2$= 1.54 · 1.20 = 1.85€. Si ganan Berdych y Djokovic los beneficios son $g_3$= 2.10 · 1.20 = 2.52€. Y por último, si acertamos los 3 partidos, $g_4$= 2 · 1.54 · 2.10 · 1.20 = 7.76€. Sin embargo, la cantidad que estamos jugando, si queremos mantener los 2€ por apuesta, habría que multiplicarla por las tres apuestas combinadas excepcionales, por lo tanto estaríamos apostando 6€. Como hemos dicho antes, nos aseguramos de recibir algo de dinero si no acertamos un pronóstico, pero en el caso de acertar todos, ganaríamos igual que en la combinada (7.76€) habiendo apostado sin embargo el triple de dinero.

- **Apuestas de hándicap europeo**. Son muy típicas en tenis. Las casas de apuestas establecen un hándicap para cada partido que debe restarse al total de juegos que consiga el jugador favorito. Estas apuestas consisten entonces, una vez aplicado el hándicap correspondiente, en acertar qué jugador conseguirá más juegos. Ejemplo: Nadal -5.5 juegos, en un encuentro Nadal-Ferrer. Esto significa que, restándole 5.5 juegos al total de los que gane Nadal en el partido, tendrá que sumar 6 juegos de los que sume Ferrer, o dicho de otra forma, Nadal tendrá que conseguir 6 juegos más que Ferrer en total.

- **Apuestas especiales de BET365**. Bet365 nos ofrece apuestas especiales en tenis como apostar a la existencia o no de tie break en un encuentro o al resultado correcto del primer set.

- **Variedad de apuestas**. Nos encontramos con una amplia variedad de apuestas en todos los deportes. En concreto, en tenis podremos apostar al ganador del primer o el segundo set, a la duración de un partido, al jugador que más saques directos hará o a cuántas faltas de saque se podrán cometer.

Una vez decidida la casa de apuestas en la que vamos a jugar nuestro dinero y explicados los distintos tipos de apuestas, decidimos que utilizaremos las apuestas sencillas. Lo decidimos así porque el nuevo modelo de predicción que vamos a proponer se centrará en el cálculo de los resultados finales de un número elevado de partidos, por lo que es la apuesta más coherente para poder obtener el mejor rendimiento.

## 2.3. Apuestas en tenis

Vamos a pasar a desgranar a continuación el deporte al que vamos a apostar, el tenis, con el objetivo de saber qué características tiene y qué nos pueden aportar. Las características más importantes del tenis que nos pueden influir a la hora de desarrollar un modelo de predicción de resultados de partidos son:

- Es un deporte individual, con la excepción de una pequeña cantidad de partidos que se juegan en dobles y que no vamos a tener en cuenta aquí. Esta característica nos va a permitir calcular la probabilidad de que un jugador gane un partido de una manera más sencilla que en un deporte colectivo, donde hay que tener muchos más factores en cuenta y el proceso sería mucho más complejo, pero también significa que la influencia del estado de un jugador (tanto físico como anímico) a la hora de disputar un encuentro es mucho mayor. Los distintos estados de los jugadores que puedan derivar en buenas o malas rachas se tendrán muy en cuenta en nuestro nuevo modelo de predicción.

- Según el perfil del jugador y su tipo de juego, la superficie sobre la que se dispute el encuentro puede afectar en el resultado final, pudiendo jugar en césped, tierra y superficie dura. Este será otro de los puntos con mayor protagonismo en nuestra propuesta de modelo de predicción.

- Los partidos que se disputan pueden ser más o menos largos, de 3 o 5 sets, llegando a alargarse hasta 6 horas y 11 horas en ocasiones puntuales [12], lo que evidentemente puede influir en el resultado.

- Los tipos de torneo influirán en la preparación y motivación de un jugador. No es lo mismo un torneo importante a nivel mundial que torneos menores, a los que muchos de los mejores jugadores ni siquiera acuden o simplemente se lo toman como test, sin prepararlos al 100%.

- En tenis sólo puede existir un resultado. No hay empate, pudiendo ganar sólo un jugador. Esto nos va a facilitar el cálculo de probabilidades y las apuestas, siendo algo más sencillo que en un deporte donde sí se puedan dar empates y por lo tanto sea más difícil acertar.

# 2.4. Herramientas para apostar: criterio de Kelly

A la hora de apostar, no sólo tenemos que saber en qué casa de apuestas lo vamos a hacer, qué tipo de apuesta nos viene mejor y cuáles son las mejores cuotas, sino que también tenemos que encontrar una estrategia que nos pueda determinar qué cantidad de dinero vamos a jugar a cada partido en función de unas determinadas características.

En el último siglo, con la revolución que supuso la tecnología y el desarrollo de los ordenadores y la computación, las técnicas de trading y los algoritmos han avanzado exponencialmente, hasta el punto de desarrollar varias teorías que nos decían cuánto deberíamos arriesgar en juegos binarios para obtener beneficios. Puesto que el tenis es un deporte con sólo dos resultados posibles como ya hemos dicho, la cantidad de dinero que vamos a apostar a cada encuentro nos la puede definir una de esas teorías: el criterio de Kelly.

Este criterio tiene su origen en 1956, y fue descrito por JL Kelly Jr. [13]. Aunque también se aplica en banca y en mercados financieros, es un criterio muy útil para las apuestas deportivas. El principal objetivo por el que vamos a aplicar este criterio y por el que se comenzó su estudio es el de optimizar cada apuesta y minimizar el riesgo de quiebra. Si apostamos en un juego binario simétrico a largo plazo (es decir, en un periodo de tiempo en el que la fracción de apuestas observadas que logramos acertar es igual a la probabilidad de que cualquier apuesta realizada sea acertada) este criterio nos asegura que no vamos a quebrar y que vamos a maximizar nuestras ganancias, siempre y cuando seamos capaces de obtener una buena probabilidad. Intentando cumplir estas condiciones, Kelly llegó a la conclusión de que resolviendo la siguiente ecuación podríamos obtener la cantidad máxima a arriesgar en cada apuesta [13]:

$$0 = p \cdot \log(1 + f) + (1 - p) \cdot \log(1 - f) \tag{2.3}$$

Ahora bien, aunque las probabilidades que aplicamos para apostar en tenis pueden tener las características de un juego binario a largo plazo, no son simétricas. Para el caso simétrico, la investigación llevada a cabo por Kelly le hizo obtener la siguiente fórmula, que es la que aplicaremos [13]:

$$f = \frac{p \cdot (b + 1) - 1}{b} \tag{2.4}$$

- f=porcentaje de la banca que vamos a jugar. Si tenemos un total de 1€ y f=20%, jugaremos 0.20€ a ese partido en concreto.

- b=cuota de la casa de apuestas, sin ninguna modificación.

- p=probabilidad de que el jugador por el que vamos a apostar gane el partido.

Con un ejemplo, si el jugador tiene una probabilidad del 70% de ganar y una cuota de 1.30€, la cantidad que deberíamos apostar es $f = (0.7 \cdot (1.30 + 1) – 1) / 1.30 = 0.47$, el 47% de la banca que tengamos.

Debido al tipo de deporte al que estamos apostando, el resultado de la fórmula siempre será positivo, 0 si el jugador no tiene ninguna opción de ganar. Además, si la probabilidad de que el jugador al que vamos a apostar es del 50% o inferior, Kelly nos recomienda no apostar, reduciendo así los riesgos.

Evidentemente este criterio no va a funcionar si no somos capaces de calcular bien las probabilidades, pero es una herramienta muy útil que nos va a minimizar los riesgos y maximizar las ganancias.

## 2.5. Marco regulador

A pesar de los siglos de historia que tienen las casas de apuestas, la regulación nunca ha sido su punto fuerte. En países como Reino Unido algunas casas se ganaron el prestigio gracias a su seguridad, sus pagos y su seriedad, pero las leyes no se han ido publicando hasta hace bien poco. En España, hasta finales del año 2010, era un mercado de nueva implantación del que se sabía poco o nada. Precisamente eso provocaba poca fiabilidad y que los usuarios no se introdujesen en el mundo de las apuestas, bien porque tenían fama de ilegales en muchas competiciones, o bien porque no estaba clara su legalidad en un marco general, pese a operar en el libre mercado y dentro de la Unión Europea. Sin embargo, en nuestro país se regula el sector del juego y por ende las casas de apuestas a partir de la ley 13/2011 que tiene los siguientes puntos importantes [14]:

- Regula el sector por parte del Estado por primera vez en estas nuevas condiciones. Hasta ese momento solo se regulaba el juego presencial en casinos y bingos.

- Establece unos mínimos de seguridad, transparencia, fiabilidad e integridad que todas las casas de apuestas tienen que cumplir para obtener una licencia. Esto hace que los jugadores confíen más en las casas de apuestas al estar respaldadas por la ley y que por lo tanto la demanda aumente. Dicha licencia se adquirirá además después de pagar una serie de impuestos y tras una serie de plazos que deben cumplir.

- El Estado utiliza mecanismos de regulación y control para comprobar que todos los requisitos y la legalidad se cumplen. Para ello, controla elementos intermediarios como medios de comunicación para publicitarse o las entidades financieras que dan soporte a los métodos de pago. Esto de nuevo reduce el riesgo de fraude y aumenta la fiabilidad.

- La Comisión Nacional del Juego y el Consejo de Políticas de Juego tendrán más competencias para supervisar el sector y aplicar las sanciones que contempla la ley en caso de irregularidades en normativa de los juegos, licencias y sistemas técnicos de control de operadores.

- Se establecen todas las modalidades de juego y los requisitos que deben cumplir cada una de ellas, y se fijan también determinadas prohibiciones.

- El Estado se garantiza recibir un porcentaje de la recaudación de las apuestas online, sin duda el sector que tiene mayor crecimiento apoyado en la Ley del Deporte Profesional. Además, crea impuestos especiales en este ámbito.

- Los operadores y las casas de apuestas están obligados a la gestión responsable del juego, lo que les obliga a colaborar con la justicia en caso de blanqueo de capitales, reducir cualquier daño potencial a la sociedad o la obligación de proporcionar al usuario toda la información posible a la hora de apostar.

- Se impondrán sanciones a aquellas casas de apuestas que alteren o manipulen sistemas técnicos a la hora de otorgar premios, concedan préstamos, organicen apuestas ilegales o amañen resultados, entre otras actividades delictivas.

- Existe un límite de 3000€ mensuales como cantidad máxima a apostar por una persona.

En general, las casas de apuestas que hemos desgranado en apartados anteriores no han tenido ningún problema para conseguir la licencia, cumpliendo todos los requisitos. Incluso podemos afirmar que se han visto beneficiadas de esta regulación, viendo incrementar sus ingresos por varios motivos. El primero y ya mencionado, porque el jugador se siente más respaldado y apuesta más dinero, teniendo más confianza en el

sistema. En segundo lugar, porque se dejan a un lado las apuestas ilegales que lastraban a muchas casas. Y en tercer y último lugar, porque las casas pueden aprovecharse de medios publicitarios para ofrecer sus servicios públicamente sin inconvenientes.

A nivel de usuario, también los beneficios son importantes. Se puede jugar sin riesgos de caer en la ilegalidad, teniendo además una seguridad jurídica y bancaria respaldada, ya que tanto los fondos como los datos van a estar garantizados.

# 3. Diseño del modelo probabilístico y aprendizaje máquina

## 3.1. Modelos existentes y algoritmos de inferencia

Como ya hemos comentado previamente, vamos a apostar a un deporte individual con dos resultados posibles realizando apuestas simples para cada partido. Para ello tendremos que obtener la probabilidad de ganar que tienen los dos jugadores que juegan un encuentro. La primera opción para esto podría ser no arriesgar y otorgarle las mismas posibilidades de ganar a ambos, un 50% para cada uno, decidiendo exactamente igual que si lanzásemos una moneda al aire [15]. Este ejemplo, aunque lógicamente no es viable para realizar predicciones, nos va ayudar a definir el modelo del cual vamos a partir y que nos va a servir de guía para llegar a los resultados finales.

Para comprobar si una moneda es justa y otorga el 50% de probabilidades a cada uno de sus posibles sucesos (cara o cruz), podemos decir que es suficiente con repetir el experimento un número elevado de veces y anotar sus resultados para comprobarlo. Este sería un enfoque frecuentista [16], en el que nosotros aplicamos un determinado grado de confianza y un margen de error que nos proporcionan el número de veces que tendremos que lanzar para hacer la comprobación de si la moneda es justa o no. Es decir, se repite el experimento un número determinado de veces, se anotan los resultados y se saca una conclusión utilizando únicamente la información obtenida en el ensayo, donde se han fijado unos criterios de decisión a priori que permanecen estáticos durante todo el estudio y que tienen poco peso, dándole una mayor importancia a la verosimilitud adquirida a través de las repeticiones. Sin embargo, desde hace algunas décadas está tomando fuerza una visión alternativa surgida de las aportaciones del matemático Thomas Bayes (aunque tiene sus orígenes en el siglo XVIII) [17]. El enfoque bayesiano, que así se denomina debido a su creador, nos interpreta la probabilidad de una manera más subjetiva. Nos dice que, a la hora de realizar un experimento, el resultado que obtengamos no es algo que podamos afirmar con contundencia, sino que simplemente es una información externa que podemos ir modificando según vayamos teniendo más datos y que nos puede ayudar a confirmar o desmentir ciertas hipótesis previas. En ese caso, se le da mucho más importancia a unos criterios de decisión a priori que podamos establecer, basados en información real pero que pueden modificarse según vamos incorporando nuevos datos, respecto a la verosimilitud del experimento como tal. Así pues, esta información a priori se transformará en probabilidad a posteriori después del proceso, que es la que tenemos en

cuenta a la hora de inferir los datos [18]. En el ejemplo de la moneda, la probabilidad de que se obtenga cara o cruz dependerá del tipo de función prior que nosotros consideremos que tiene que tener (uniforme, normal, etcétera), pudiendo obtener distintos resultados. Será importante por lo tanto discutir qué probabilidad vamos a establecer a priori y con qué criterios, puesto que afectará directamente a los resultados finales.

Para realizar nuestro proyecto, partiremos de una amplia base de datos en formato Excel actualizada a diario que nos aporta la siguiente información [19]:

- Nombre del torneo, tipo y localización geográfica.
- Fecha de cada partido.
- Superficie en la que se juega cada encuentro.
- Ronda del torneo al que pertenece.
- Jugadores que se enfrentan, posición en el ranking y puntos que tienen. Esto sigue siempre un criterio: se sitúan primero los jugadores ganadores y después los perdedores en cada partido.
- Cuota de 7 casas de apuestas distintas, casi todas a nivel europeo, entre las que destacan BET365, Expert, Ladbrokes, Pinnacle Sports o Stan James. El orden es igual que en el punto anterior: primero tenemos la cuota del ganador y a continuación la del perdedor.
- Si el partido se ha jugado o se ha parado por alguna lesión u otros motivos.

El experimento realizado con una moneda nos confirma que debemos aplicar un enfoque bayesiano para obtener los resultados. Como podemos ver, la cantidad de datos que tenemos a nuestro alcance, la posibilidad de modelar a priori algunos parámetros para evitar ciertos problemas que comentaremos después y la capacidad de ir introduciendo información de todo tipo para ajustar el modelo hacen que sea la mejor opción.

Para comenzar el proceso, configuraremos una lista de todos los jugadores que participan en los torneos y otorgaremos a cada uno de ellos un factor de calidad que los defina y tenga en cuenta todos los factores posibles. Con este factor de calidad podremos obtener después la probabilidad de que los jugadores ganen distintos partidos. Para ello, vamos a ir acumulando información partido a partido y, utilizando la regla de Bayes, estableceremos una función a priori que irá aprendiendo con los mismos y nos

dará una información útil en la que basarnos para calcular probabilidades futuras, aplicando finalmente inferencia bayesiana y obteniendo la probabilidad a posteriori.

Por la fórmula de Bayes [20]:

$$p(\Theta|\, Y) = \frac{p(Y\,|\Theta\,) \cdot p(\Theta\,)}{p(Y)} \tag{3.1}$$

Donde $\Theta$ es el vector que define el factor de calidad de los jugadores, mientras $Y$ es el vector que define qué jugador ha ganado el partido, otorgando un valor de 1 al ganador y de 0 al perdedor. Como el proceso va a tratar de maximizar la posterior $p(\Theta \mid Y)$, no necesitamos normalizar y por lo tanto $p(Y)$ no nos afecta a la hora de calcular máximos, por lo que podemos prescindir de ella:

$$p(\Theta|\, Y) = \; p(Y \mid \; \Theta) \cdot p(\Theta\,) \tag{3.2}$$

Para iniciar el proceso, vamos a empezar por el cálculo de $p(Y \mid \Theta)$ (likelihood) aplicando un modelo de probabilidad en el que basarnos. Por las características del tenis, el que mejor se adecúa es el modelo de Bernouilli [21]:

$$f(k;p) = \begin{cases} p & if\ k = 1 \\ 1-p & if\ k = 0 \end{cases} \tag{3.3}$$

Es una distribución discreta con las siguientes características:

- El experimento consiste en n pruebas que se repiten.
- Cada prueba puede tener como resultado k=1 (éxito) con una probabilidad p o k=0 (fracaso) con una probabilidad 1-p, situándose p entre 0 y 1.
- Las pruebas que se repiten son independientes.

Como comprobamos se adecúa perfectamente a lo que buscamos. Calcularemos la probabilidad de que uno de los jugadores gane en cada partido de manera independiente a los partidos anteriores, asumiendo directamente que la probabilidad de que pierda un encuentro es la probabilidad de que lo gane el otro jugador. Esta p va a definirse por la siguiente función que hemos establecido:

$$p_{ij} = \frac{1}{1 + e^{\theta_j - \theta_i}} \tag{3.4}$$

Donde $p_{ij}$ es la probabilidad de que el jugador i gane al jugador j, y $\theta_i$ y $\theta_j$ los factores de calidad de ambos, así como $p_{ji} = 1 - p_{ij}$. Esta es una función que nos da números superiores a 0.5 si el jugador i gana al jugador j, e inferiores si sucede al revés, estableciendo valores en todo momento entre 0 y 1 que nos definen exactamente la probabilidad sin tener que realizar modificaciones.

Con todo ello, definimos la likelihood de la siguiente manera:

$$p(Y \mid \Theta) = \prod_{n=1}^{N} p(Y_n \mid \Theta) = \prod_{n=1}^{N} p_{i(n),j(n)}^{y} \ (1 - p_{i(n),j(n)})^{1-y} \tag{3.5}$$

Donde $Y_n$ nos indica el resultado de forma binaria, siendo $Y_n = 1$ si el jugador *i* gana el partido y $Y_n = 0$ si lo gana el jugador *j*. Ya tenemos la likelihood, el modelo probabilístico y la función de probabilidad que vamos a utilizar. Pero para calcular la probabilidad a posteriori no nos basta con aplicar el teorema de Bayes tajantemente, sino que tenemos un vector $\Theta$ de factores de calidad que no conocemos y que tenemos que inferir. Inferir dichos parámetros consiste en deducirlos mediante evidencias u observaciones y ver qué consecuencias pueden tener a la hora de calcular las probabilidades, para poder calcular futuros parámetros incorporando información nueva al modelo estableciendo así un proceso de aprendizaje [22].

Sin embargo, a la hora de continuar con el proceso de obtención de probabilidades nos encontramos con un problema. Puede darse el caso de que un jugador que no tenga unas buenas cualidades gane el único partido que juegue. Entonces, este jugador podría tener un factor de calidad elevado que no se correspondería con la realidad, puesto que ha ganado el 100% de los partidos que hemos analizado pero eso no le define. Este fenómeno que nos puede ocurrir se denomina sobreajuste, bajo el cual el modelo se

ajusta muy bien a los datos existentes pero tiene un pobre rendimiento a la hora de predecir nuevos resultados, precisamente porque se ha adaptado en exceso a los datos de los que ya se disponía [23].

Para solventar este inconveniente vamos a utilizar una función prior que nos permita modelar estos parámetros, maximizándola posteriormente. Si el objetivo final es tener los mejores valores en la posterior, tendremos que maximizar la likelihood y la prior:

$$
\begin{aligned}
-\log p(\Theta) = & -\log\left(\prod_{n=1}^{M} p(\Theta)\right) = -\sum_{n=1}^{M} \log \frac{\Theta(n)^{k-1} e^{-\frac{\Theta(n)}{s}}}{s^k \gamma(k)} \\
& = (k-1)\sum_{n=1}^{M} ln\big(\Theta(n)\big) \\
& - \sum_{n=1}^{M} \frac{\Theta(n)}{s} - M \cdot k(s) - M \cdot ln\left((k-1)!\right)
\end{aligned}
\tag{3.6}
$$

Donde s=2, k=2, $\gamma$ es una función que calcula el factorial de k y M es el número de jugadores. La función prior que acabamos de definir tiene la siguiente forma:

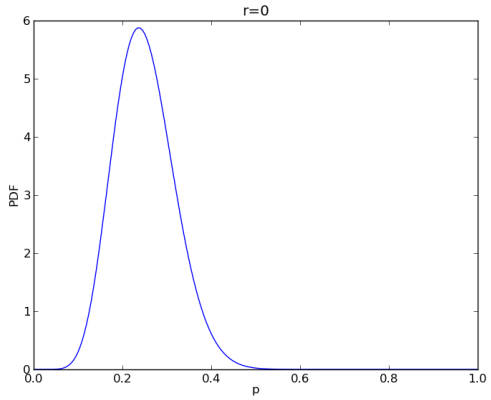


Figura 3.1: Prior.

Hemos elegido esta prior por los siguientes motivos. En primer lugar, sólo nos otorga valores positivos. En segundo lugar necesitamos que, cuando un jugador juega pocos partidos, no le otorguemos el máximo factor de calidad para evitar así el sobreajuste mencionado antes. En esta gráfica podemos observar cómo se evita esto precisamente, otorgando un valor para el inicio del parámetro que nunca es el máximo de la probabilidad. A partir de este máximo, permitimos la evolución del parámetro hacia un

lado u otro de la gráfica. Si un jugador juega y gana más partidos tendrá más probabilidad y por lo tanto mayor factor de calidad, desplazándose hacia la derecha del máximo. Si por el contrario, al acumular partidos los pierde, nos desplazaremos hacia la izquierda del máximo, la probabilidad será menor y el factor de calidad más bajo, pero nunca otorgamos valores extremos en primera instancia, al revés de lo que pasaría con otro tipo de funciones como por ejemplo una exponencial.

Como hemos comentado previamente, el objetivo final es obtener unos parámetros óptimos en la posterior. Esto sólo puede hacerse maximizando dicha función y, por ende, maximizando sus dos componentes, para lo que aplicaremos logaritmos:

$$L = -\log p(\Theta|Y) = -\log p(Y|\Theta)p(\Theta) \tag{3.7}$$

Hasta aquí el desarrollo matemático que nos ha llevado a inferir los parámetros de calidad que necesitábamos para calcular las probabilidades. Este desarrollo se utilizó en el modelo anterior y su proceso se seguirá también en el nuevo modelo que propondremos a continuación. Pero todo ello tiene que tener un soporte para su cálculo y una herramienta con la que poder operar rápidamente con vectores de gran tamaño (cada Excel de la base de datos tiene entre 2000 y 3000 partidos por temporada). Esta herramienta no es otra que Matlab.

Para poder calcular L, programaremos el código que utilizará en última instancia la función *fmincon* [24] que tiene la siguiente forma:

$$x = \text{fmincon}(\text{fun}, x0, A, b, Aeq, beq, lb, ub, \text{nonlcon}, \text{options}) \tag{3.8}$$

Donde los argumentos de entrada que vamos a utilizar son:

- Fun.- introduciremos la función posterior a falta del factor de calidad theta, cuyos valores maximizados nos los devolverá fmincon.
- x0.- Punto inicial de la función
- Lb.- Límites inferiores (lower bound)

Con estos argumentos, fmincon será capaz de recorrer todos los partidos que introduzcamos uno a uno, y nos dará el mínimo correspondiente de la función L, o lo que es lo mismo, el parámetro de calidad de los jugadores maximizado después del proceso de inferencia.

## 3.2. Modelos propuestos

A continuación vamos a presentar el modelo propuesto para desarrollar este proyecto. Dicho modelo va a utilizar el aprendizaje máquina [25], pero además tendrá como base el método de validación cruzada aplicado a una serie de factores que vamos a justificar convenientemente en el presente apartado [26]:


Figura 3.2: Validación cruzada.

La validación cruzada es un procedimiento para realizar distintas pruebas en nuestro modelo, valorar cual da mejores resultados y comprobar que funciona. Para ello, vamos a utilizar unos datos de entrenamiento fijos que estarán alrededor de los 500 partidos (los dos primeros meses de una temporada), para entrenar el modelo (franja azul en la Figura 3.2). Después, con los datos de validación (franja verde en la Figura 3.2), vamos a realizar distintas pruebas, modificando los factores y observando cual puede ser la mejor de las opciones. Una vez que tenemos esto claro y que sabemos qué variante de los factores debemos elegir para obtener los mejores resultados, debemos probar nuestro modelo en unos datos de test (franja roja en la Figura 3.2), que nos dirán la capacidad de dicho modelo para predecir. Tenemos que tener claro que los datos sobre los que hacemos la validación tienen que ser distintos a los datos de test.

La argumentación de los tres factores que vamos a presentar a continuación y que son la base de nuestro modelo tiene la misma línea, y no es otra que saber exactamente, según distintos elementos que pueden afectar al resultado final, qué partidos nos aportan la mejor información y la más útil, cómo seleccionarlos y cómo tratarlos, puesto que no tenemos que olvidar que el objetivo final es inferir unos parámetros de calidad que nos

permitan obtener una probabilidad lo más acertada posible, y para ello debemos contar con una información de máximo rigor.

### 3.2.1. Factor de olvido

Uno de los principales elementos a tener en cuenta en un deporte como el tenis es el estado físico y anímico que puedan tener los jugadores en el momento de jugar un partido. Las distintas rachas que pudieran tener son claves a la hora de disputar un encuentro e influyen directamente en el resultado de un partido, por lo que medirlo correctamente de alguna manera nos puede ayudar mucho si de lo que se trata finalmente es de predecir un resultado. Para poder hacerlo, a la hora de calcular la probabilidad de un partido necesitamos dar una mayor importancia a los encuentros disputados en fechas más cercanas que al resto de encuentros de los que podamos disponer en las bases de datos, puesto que estos partidos disputados hace más tiempo no nos van a aportar demasiada información a la hora de saber qué estado tiene el jugador últimamente. Es posible que un jugador hace un año tuviese una racha buenísima pero, tras pasar por una lesión, haya perdido los tres últimos encuentros porque se encuentre en baja forma. Puede suceder también que un jugador no sea muy bueno pero venga de derrotar a un jugador del Top 3 y se encuentre motivado. O simplemente, puede suceder que un jugador haya ganado los últimos 5 encuentros y por lo tanto podamos deducir que está en buena forma tanto física como anímica, entre otros posibles casos. Por tanto son todos estos partidos que hemos citado los que más información nos van a aportar a la hora de predecir el resultado de otro encuentro que se vaya a jugar próximamente.

Para poder modelar este factor a través de distintos parámetros, lo definimos de la siguiente manera:

$$\alpha = a_0 - (a_0 - b_0)e^{-\beta(t-1)} \tag{3.9}$$

Donde describimos el factor de olvido como una exponencial creciente con un valor mínimo y un valor máximo, que nos permite otorgarle a cada partido una importancia distinta, así como una t que nos indica el tiempo en días que hay entre el día en que se jugaron los partidos y el día actual. Por otro lado β es un factor que nos definirá el crecimiento de la exponencial.

Este factor de olvido, al otorgarle importancias distintas a los encuentros de una o varias temporadas en función del tiempo que hace que se disputaron, nos estará describiendo el

perfil de los jugadores y jugadoras: cuanto más dependa de sus rachas un jugador, más irregular será y mejor efecto surtirá, mientras que si los jugadores son muy regulares nos dará igual coger los últimos partidos u otros anteriores porque los resultados a nivel de calidad serán similares.

Una vez definido, nos interesa analizar este factor en profundidad para ver sus posibles variantes y las distintas modalidades que posteriormente podemos llevar a cabo a la hora de aplicar la validación. De los resultados del modelo anterior se obtuvo la conclusión de que el factor de olvido no conseguía mejorar prácticamente nada. Por lo tanto, debemos aplicar un modelo distinto que haga mejorar los resultados. Podemos modificarlo de la siguiente manera. El valor estándar que se utilizaba y que hasta ahora se había aplicado en modelos anteriores era β=0.03, con valores de $b_0$=0.1 y $a_0$=1. Con esto obtenemos un factor de olvido que a partir de los cuatro meses analizados aproximadamente (t=130, α=0.98) otorga la máxima importancia a todos los partidos, y al resto un peso bastante elevado. Al aplicar el anterior modelo, los resultados con este factor de olvido como hemos comentado no eran muy buenos. Al mantener estos parámetros, la modificación que se hace es mínima: las diferencias entre los resultados con o sin factor de calidad serían muy pequeñas porque estaríamos prácticamente reproduciendo los resultados. Esto nos lleva a no mejorar a pesar de incorporar más datos al análisis. En el presente modelo, lo que haremos será establecer un criterio de pesos distinto, más incisivo, que nos permita probar si hay ciertos partidos que nos aportan más información que otros de una manera más clara. Así, podríamos aplicar varias modificaciones con la siguiente lógica:

- Podríamos modificar el límite superior o el límite inferior. Con esto estaríamos variando la amplitud del factor de olvido y reduciríamos las diferencias entre los distintos partidos.

- Podríamos modificar el parámetro beta para variar la curva de la exponencial. Aquí podemos aumentar o disminuir el valor de β:

    - Si subimos el valor de beta a, por ejemplo, 0.04, conseguiríamos que el crecimiento de la exponencial fuese más rápido, llegaría al máximo antes y abarcaría más partidos con valores altos, dando una alta importancia a los mismos a partir de los tres meses (t=100, α=0.98) por lo que valoraríamos prácticamente todos los partidos como importantes.

    - Si bajamos el valor de beta a, por ejemplo, 0.02, le daríamos más importancia a los partidos de los últimos 3-4 meses del análisis (t=200, α=0.98), aplicando un criterio algo más directo que establece mayores diferencias entre ellos.

En cualquier caso, realizaremos distintas pruebas en la sección 4.2 siguiendo la validación cruzada y justificaremos numéricamente cuál es la mejor elección y qué nos aporta.

### 3.2.2. Factor de superficie

Según el perfil del jugador o jugadora y su tipo de juego, la pista sobre la que se dispute un partido puede afectar en el resultado final. Podemos clasificar las superficies, de más a menos rápidas en césped, dura de interior, dura de exterior y tierra, la más lenta [27] [28]:

- **Césped.-** En este tipo de superficie el bote de la pelota es rápido y deslizante, obligando al jugador a defenderse de los ataques en una posición más incómoda y con mucha menos capacidad de movimiento. Aquí se defienden bien los jugadores a los que les gusta subir a la red, rápidos y con buen saque, dificultando el juego desde el fondo de la pista. El torneo más famoso disputado en esta superficie es Wimbledon. Algunos de los jugadores que mejor se desenvuelven actualmente en ella son Federer o Murray.

- **Superficies duras (de exterior y de interior).-** Son las superficie en la que más partidos se juegan. Las de exterior suelen estar fabricadas en cemento aunque también pueden utilizarse otros materiales resistentes, mientras que las de interior son de materiales sintéticos especiales. El bote de la pelota se caracteriza por ser rápido y no muy alto, complicando la defensa del jugador a la hora de obtener puntos de referencia. Favorecen sobre todo a jugadores más agresivos, con un golpeo potente y buen juego de volea. Entre sí se diferencian en que la pista dura de interior es algo más rápida que la de exterior. El torneo más importante disputado en superficie dura es el Open de Australia. Algunos de los jugadores que mejor se manejan en este tipo de superficies serían Roddick, Federer o Djokovic.

- **Tierra.-** Este tipo de pista formada por arena arcillosa o polvo de ladrillo hace que el juego sea más lento y la velocidad de la pelota se reduzca. El bote de la pelota es mucho más alto y el juego se desarrolla normalmente en el fondo de la pista, favoreciendo por lo tanto a los jugadores con un físico más potente y de corte más defensivo. El torneo más importante jugado en tierra es sin duda el Roland Garros, y el jugador que más destaca en ella es Nadal.

Se nos ocurre por tanto utilizar estas diferencias que tienen los jugadores según la superficie en la que juegan para darle distinto peso a los partidos en función de la

misma, utilizando los criterios de superficies más rápidas y más lentas. Para ello aplicaríamos una matriz simétrica de superficie, S, que nos devolvería un peso distinto en función de la superficie en la que se juegue. Es decir, $S_{ij}$ es el factor que otorgaríamos al partido cuando estamos evaluando la superficie i y el partido se juega en la superficie j, de tal manera que si evaluamos la misma superficie que en la que se juega un encuentro, i=j y el valor que nos retorna S sería el máximo posible. El orden de las superficies en la matriz sería el mismo que hemos establecido según los criterios de más a menos lento. En todo esto podríamos seguir los siguientes criterios o modificaciones:

- Podríamos otorgar pesos entre las distintas superficies de manera regresiva, otorgando los valores de la siguiente manera:

$$S = \begin{pmatrix} 1 & 0.6 & 0.4 & 0.2 \\ 0.6 & 1 & 0.2 & 0.4 \\ 0.4 & 0.2 & 1 & 0.6 \\ 0.2 & 0.4 & 0.6 & 1 \end{pmatrix} \quad (3.10)$$

- Podríamos seleccionar igualmente pesos regresivos, pero tomando valores más extremos para marcar las diferencias:

$$S = \begin{pmatrix} 1 & 0.5 & 0.25 & 0.1 \\ 0.5 & 1 & 0.1 & 0.25 \\ 0.25 & 0.1 & 1 & 0.5 \\ 0.1 & 0.25 & 0.5 & 1 \end{pmatrix} \quad (3.11)$$

- Podemos seguir la lógica contraria a la opción anterior, y otorgar unos valores más altos para que las diferencias sean menores:

$$S = \begin{pmatrix} 1 & 0.75 & 0.5 & 0.25 \\ 0.75 & 1 & 0.25 & 0.5 \\ 0.5 & 0.25 & 1 & 0.75 \\ 0.25 & 0.5 & 0.75 & 1 \end{pmatrix} \quad (3.12)$$

En cualquier caso, a la hora de desarrollar los resultados llevaremos a cabo las distintas alternativas de diseño y veremos cuál de ellas es la que mejor resultados nos puede proporcionar.

### 3.2.3. Factor de enfrentamiento directo head to head

Vamos a desarrollar a continuación el tercer pilar de nuestro modelo, que no es otro que el factor head to head o factor de enfrentamiento directo. Si observamos el histórico de encuentros en los circuitos masculino y femenino, vemos que hay ciertas tendencias en los resultados entre jugadores que podrían llegar a repetirse. Es posible que un jugador, por su estilo de juego, las superficies o porque se prepara algunos partidos específicamente, le tenga tomada la medida a otro y siempre o casi siempre le venza a pesar de tener una peor posición en ranking y peores cualidades. Como ejemplo pueden servir los enfrentamientos Nadal-Djokovic de los últimos tiempos, sobre todo de las temporadas 2011 a 2013. Si nos fijamos en los rankings, partidos ganados, estado físico, etcétera, la mayoría de veces diríamos que Nadal sería el vencedor. Pero no ha sido así, y Djokovic, aunque a priori tiene peores condiciones, ha vencido la mayoría de encuentros contra Nadal convirtiéndose en su bestia negra. Es precisamente este matiz el que nosotros queremos medir con este factor para obtener mejores resultados.

La manera de aplicarlo a nuestro modelo va a seguir la misma línea que los dos anteriores factores: le otorgaremos un peso distinto a los partidos entre jugadores que podamos considerar importantes, para que esta información sea más relevante a la hora de hacer el cálculo del factor de calidad y, posteriormente, de las probabilidades. Las distintas variantes que podemos aplicar vienen dadas por qué peso le vamos a dar exactamente y cuánto van a valer más que en el resto de partidos, pudiendo otorgarle el doble de peso a los enfrentamientos directos respecto al resto de partidos, tres veces más o, incluso, diez veces más. Tendremos que realizar un estudio de nuevo de las distintas variantes para ver cuál nos aporta más y sobre qué jugadores es mejor aplicarlo, y para ello en este caso también tendremos que revisar si los encuentros de los que disponemos son suficientes para llevarlo a cabo y la información es verdaderamente útil.

# 4. Resultados y evaluación.

En este apartado presentaremos una amplia batería de resultados que nos permita comparar, realizar análisis y obtener conclusiones sobre nuestro modelo. Pero para ello, necesitamos dotarnos de ciertas herramientas que, una vez desarrollado el código y obtenidas las probabilidades, nos permita medir su calidad lo mejor posible.

A nivel cuantitativo, tras aplicar el criterio de Kelly con las probabilidades finales que obtengamos, simularemos las apuestas a través de una casa online. Aquí la medición más plausible y directa es ver si somos capaces de obtener o no beneficios después de haber apostado.

Sin embargo, también necesitamos tener como referencia una función de coste que nos devuelva un resultado para ver cómo de bien o de mal estamos obteniendo las probabilidades. Determinamos que esa función de coste será la siguiente [29]:

$$LogLoss = -\frac{1}{n}\sum_{i=1}^{n} y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i) \tag{3.11}$$

En este caso hemos decido utilizar la misma que se utilizó en el modelo anterior para comprobar si los resultados han mejorado. En la función, $y_i$ nos dice el ganador de cada partido (otorgando un valor de 1 al vencedor y 0 al perdedor), $\hat{y}_i$ es la probabilidad que tiene de ganar ese mismo jugador que vence, y *n* es el número de partidos jugados. Las características más importantes de la función Logloss las detallamos a continuación:

- Nuestros resultados serán mejores cuanto más bajo sea el número que nos reporte la función LogLoss.

- En nuestro caso, $y_i$ siempre va a ser 1 dada la disposición de la base de datos en la que el primer jugador siempre gana. Puesto que vamos a extraer toda la información de ahí, el término $(1 - y_i)$ será 0, simplificando la función:

$$LogLoss = -\frac{1}{n}\sum_{i=1}^{n} y_i \log(\hat{y}_i) \tag{3.12}$$

- Si otorgásemos probabilidades altas de ganar a jugadores que después pierden, la penalización es muy grande. Por ejemplo, si pensamos que un jugador tiene una probabilidad de ganar del 80% y gana, y otro jugador tiene el 90% de probabilidad de ganar (y por ende, 10% de perder) y pierde, el resultado sería LogLoss = - (1/2) * (log(0.8) + log(0.1)) = - (1/2) * (-0.111 - 1.151) = 1.26, donde vemos como claramente la penalización del segundo término es elevada. Incluso si le damos una probabilidad de ganar a un jugador del 100% y luego pierde, el término LogLoss sería infinito.

- Si otorgásemos una probabilidad de 0.5 a todos los jugadores en todos los partidos, la función Logloss nos devolvería un valor de 0.69. Esta la casuística con más incertidumbre.

Una vez que contamos con todo lo necesario, pasamos a presentar los resultados.

# 4.1. Previa: resultados del modelo anterior

En este apartado se reproducirán los resultados del modelo anterior con el objetivo de analizarlos para después establecer comparaciones con los obtenidos mediante nuestro modelo propuesto. Pero antes de nada debemos comentar una cuestión importante. El código a ejecutar mediante el que se obtienen los factores de calidad de los jugadores y las distintas probabilidades de los partidos tiene un tiempo de ejecución muy elevado, por lo que necesitamos obtener algún método que lo reduzca para poder hacer pruebas de manera más ágil. Aún así, tenemos que tener en cuenta que estamos manejando entre 2000 y 4000 partidos por ejecución, y que las fórmulas que estamos utilizando los recorren, buscan mínimos, etcétera, por lo que aún así va a ser elevado. El método para reducirlo ha sido utilizar el vector $\Theta$ del factor de calidad del día anterior como punto de partida para volver a calcular fmincon en cada iteración, en vez de volver a calcular el mínimo de todos los datos uno por uno de nuevo. Las mejoras son notables llegando a reducir el tiempo de ejecución a la mitad. Con un ejemplo práctico: tomando como referencia el mismo mes de la temporada 2013, si utilizamos el vector $\Theta$ (factor de calidad) como x0 (punto donde empieza a recorrer la función) el tiempo de ejecución son 12 minutos, mientras que utilizando todos los datos el tiempo de ejecución es de 23 minutos.

Una vez hemos agilizado la ejecución, llevamos el modelo anterior a cabo sobre las temporadas utilizadas en el experimento previo, añadiendo alguna más con el objetivo

de poder sacar mejores conclusiones. Además, aplicamos dicho modelo sobre las mismas temporadas de tenis femenino para ver el efecto que tiene. Los resultados han sido:

| Temporada | Tenis Masculino | Tenis Femenino |
|---|---|---|
| 2013 | 0,6442 | 0,6423 |
| 2012 | 0,6291 | 0,6562 |
| 2011 | 0,6416 | 0,6397 |
| 2011-2012 | 0,6108 | 0,6344 |
| 2012-2013 | 0,6184 | 0,6342 |
| 2011-2013 | 0,6079 | 0,6275 |

Tabla 4.1: LogLoss modelo anterior

Así, las primeras conclusiones que podemos obtener mediante la Tabla 4.1 son:

- Desde el punto de vista de los resultados, la función de coste de una misma temporada nos arroja valores por encima del 0.64, mientras que acumulando dos o más temporadas los resultados están alrededor del 0.62. Esto nos dice que será más interesante centrar nuestros esfuerzos sobre temporadas con este perfil, concretamente analizando en profundidad las temporadas 2012-2013, puesto que son representativas, más actuales y tienen un gran margen de mejora. Hay que dejar claro que las temporadas 2011-2012, 2012-2013 y 2011-2013 no reproducen los resultados de todas las temporadas juntas, sino sólo de la última (2012, 2013 y 2013 respectivamente). Esto lo hacemos para ver el efecto que pudiera tener introducir más partidos en el análisis de una misma temporada y para poder establecer comparaciones.

- En el análisis de varias temporadas juntas, consideramos que es mejor aplicar 2012-2013 que las temporadas 2011-2013 por varios motivos. En primer lugar, porque supone un coste computacional elevado. Y en segundo lugar, si observamos los resultados de la temporada 2012-2013 comparados con la 2013, la mejora en este caso es del 4%, mientras que la mejora del 2011-2013 respecto a 2012-2013 es del 1.5%. Siguiendo esta tendencia podemos pensar que si acumuláramos otra temporada más, los resultados podrían comenzar a empeorar. Esto sucede porque en tres temporadas, después de unos 6000 partidos, los encuentros empiezan a no aportarnos demasiada información útil a la hora de predecir y además sería más difícil modelarlos.

- Observando de igual manera los resultados añadidos del tenis femenino, podemos apreciar que en términos generales nos proporciona peores resultados

que el tenis masculino, situándose por encima de 0.65 de LogLoss en algunos casos.

## 4.2. Resultados cualitativos del modelo propuesto

A continuación desarrollaremos nuestro modelo y profundizaremos en sus aportaciones. A la hora de aplicarlo, tenemos que cumplir al pie de la letra el proceso de validación cruzada descrito en la sección 3. Para la fase de entrenamiento, como ya se comentó, vamos a utilizar alrededor de 500 partidos de la temporada 2012 (los dos primeros meses), tomando el resto de la temporada como datos de validación. En esta validación combinaremos todas las opciones posibles: aplicaremos los factores por separado, de dos en dos o todos juntos, quedándonos con la combinación más favorable y realizando además análisis propios de cada uno de ellos por separado. Finalmente, necesitamos realizar el test sobre un conjunto de datos distinto, que no será otro que la temporada 2013 (con información acumulada desde 2012), es decir, el proceso completo se aplicará sobre 2012-2013. La aplicación de este nuevo modelo nos otorgará nuevos rankings y nuevos valores que tendremos la oportunidad de comparar con el modelo previo. Analizaremos los circuitos femenino y masculino por separado, siguiendo el siguiente esquema:

- A-Factor de olvido
- B-Factor de superficie
- C-Factor de enfrentamiento head to head
- D-Diferencias entre jugadores
- E-Función de coste: LogLoss

Añadido a los tres factores fundamentales de nuestro modelo, nos ha parecido interesante introducir también una medida sobre las diferencias entre jugadores de nuestro ranking. Así, nos interesaría medir la distancia que separa a los mejores jugadores de aquellos con peores cualidades, y compararla con los resultados del modelo anterior con el objetivo de determinar la igualdad o desigualdad entre los jugadores, además del grado en el que se ven afectados por los distintos factores del nuevo modelo en términos del factor de calidad. Para ello, calcularemos la media de dicho factor de calidad de los jugadores del TOP 10 y, posteriormente, la de los jugadores en la franja 90-100 del ranking, restando posteriormente ambas. Esto lo

hacemos porque los jugadores a partir del TOP 100 se diferencian mucho menos entre ellos y su influencia y aportación al análisis no sería determinante a la hora de obtener conclusiones.

Además, antes de profundizar en el análisis es necesario aclarar una cuestión. Es cierto que cuando analizamos el factor de calidad, se podría pensar que el modelo otorga un valor superior a los jugadores que juegan muchos más partidos y disputan más torneos, cuando esto no es exactamente así, al menos no directamente. Evidentemente es cierto que jugar un mayor número de partidos da al jugador más oportunidades de tener factor de calidad elevado, pero se puede demostrar que la relación entre ambos no es tan directa, y que un jugador tiene que ganar la mayoría de los encuentros que juega para tener un factor de calidad elevado, mostrando un modelo coherente y consistente. Vamos a mostrar a continuación varios ejemplos para intentar llegar a una conclusión.

Tomando los datos de la temporada 2012 completa, los siguientes jugadores presentan los valores:

| Jugador | Partidos totales | Partidos ganados | Partidos perdidos | Factor de calidad |
|---|---|---|---|---|
| Djokovic | 79 | 69 | 10 | 6,95 |
| Nadal | 42 | 36 | 6 | 6,87 |
| Dimitrov G. | 40 | 23 | 17 | 3,91 |
| Massu N. | 1 | 0 | 1 | 1,57 |
| Prodon E. | 7 | 0 | 7 | 0,69 |

Tabla 4.2: Relación partidos-factor de calidad tenis masculino 2012.

| Jugador | Partidos totales | Partidos ganados | Partidos perdidos | Factor de calidad |
|---|---|---|---|---|
| Azarenka V. | 67 | 60 | 7 | 7,44 |
| Williams S. | 47 | 44 | 3 | 7,8 |
| Barthel M. | 47 | 21 | 12 | 3,83 |
| Thorpe L. | 1 | 0 | 1 | 1,62 |
| Barrois K. | 7 | 0 | 7 | 0,66 |

Tabla 4.3: Relación partidos-factor de calidad tenis femenino 2012.

En las Tablas 4.2 y 4.3 podemos observar varias cosas:

- Un jugador puede jugar un número muy elevado de partidos, pero sólo consigue un factor de calidad alto si gana muchos de ellos, como son los casos de Djokovic o Azarenka.

- Si un jugador juega un número de partidos importante pero no gana un alto porcentaje de ellos su factor de calidad se queda en un nivel intermedio. Estos casos se ven en los jugadores Dimitrov o Barthel, en contraposición con los casos de Nadal o Serena Williams, que jugando un número de partidos parecidos tienen un factor de calidad mucho más alto porque ganan más.

- Si un jugador juega un partido y pierde, la penalización no es tan grande, como podemos ver en los casos de Massu o Thorpe. Es más grande si no gana al jugar más, como por ejemplo se ve en los casos de Prodon y Barrois. Esto es coherente con lo que presentamos al añadir la función prior.

- No podemos comparar los valores de los factores del circuito femenino y masculino, puesto que los cálculos se realizan entre jugadores distintos, en un número de partidos distintos y en unas condiciones distintas. Podríamos obtener entonces la conclusión de que Serena Williams ganaría a Nadal, y eso no podemos afirmarlo.

### 4.2.1. Tenis masculino

#### A-Factor de olvido

Es en este apartado en el que debemos tomar la decisión de qué modificar y en qué medida para mejorar sustancialmente nuestros resultados. Hemos expuesto en la sección 3 varios parámetros que eran susceptibles de modificarse. Lo que pretendemos es seleccionar los últimos partidos y darle más importancia frente al resto, y para marcar esa diferencia y darle más peso a los partidos más cercanos en el tiempo necesitaremos que la exponencial tenga un crecimiento más lento y no alcance su máximo prácticamente al inicio, como venía haciendo en el modelo anterior. En ese caso, la modificación de los límites superiores e inferiores no nos aportará demasiado, puesto que queremos mantener la máxima importancia de los partidos más cercanos en el tiempo y además que se establezca una diferencia notable con el resto. Para llevar a cabo un cambio real, necesitamos bajar el valor de $\beta$ respecto al otorgado en el anterior modelo de $\beta=0.03$, puesto que subirlo nos otorgaría el efecto contrario y le estaríamos dando a todos los partidos la misma importancia.

Realizamos pruebas entre varios valores: $\beta=0.02$, $\beta=0.01$ y $\beta=0.0025$. Nos damos cuenta de que, cuanto más baja el valor, mejores resultados se obtienen. Finalmente fijamos el valor en $\beta=0.0025$, después de comprobar que seguir bajándolo no nos aporta nada. Este

es un valor más extremo, que nos permite dar una importancia alta y muy similar a los partidos de las últimas 3-4 semanas según avanzamos en el análisis. También nos ayuda a aplicar algo distinto a largo plazo para poder probar qué sucede con esos pesos. Hacemos una prueba con 2012 y los resultados son:

| Modelo | Tenis masculino |
|---|---|
| 2012 Modelo anterior | 0,6291 |
| 2012 Modelo propuesto (con factor de olvido) | 0,6034 |

Tabla 4.4: Evolución LogLoss con factor de olvido. Tenis masculino 2012.

Como se puede apreciar en la Tabla 4.4, la mejora es aceptable respecto al modelo anterior al introducir este factor, alrededor de un 4%. Esto significa que las rachas son un factor determinante en el circuito masculino y que los jugadores dependen de su estado en el momento de jugar, habiéndose modificado el factor de calidad convenientemente para obtener unos mejores resultados.

## B-Factor de superficie

Tras definir en la sección 3 todos los tipos de superficies, ventajas e inconvenientes según el perfil de los jugadores, nos toca definir cuál de todas las alternativas propuestas nos ofrece mejores resultados. Cabría pensar a priori que la superficie es un factor muy importante e influyente y que cuantas más diferencias se marquen entre la superficie en la que se esté jugando un encuentro y el resto de encuentros jugados en superficies distintas, mejor. Pero al realizar las pruebas, nos encontramos con que los mejores resultados no los obtenemos de esta manera, ni aplicando un efecto mínimo de diferencia, sino finalmente con la matriz de valores intermedios definida en (3.10), cuyos resultados de la función LogLoss presentamos a continuación:

| Modelo | Tenis masculino |
|---|---|
| 2012 Modelo anterior | 0,6291 |
| 2012 Modelo propuesto (con factor de superficie) | 0,6223 |

Tabla 4.5: Evolución LogLoss con factor de superficie. Tenis masculino 2012.

En la Tabla 4.5 vemos que las mejoras son de alrededor del 1.2%. La superficie va a ser algo influyente en este circuito pero no todo lo que esperábamos. Esto puede deberse a que nuestro criterio de clasificación de superficies y nuestra matriz no ajuste de manera óptima los valores que definen a los jugadores.

Antes de cerrar este apartado, vamos a realizar un análisis específico sobre cada uno de los tipos de superficie que tratamos en este experimento. Para ello, vamos a aplicar este factor en los partidos de test y vamos a analizar cómo se comportan los jugadores en cada una de ellas por separado con el objetivo de dilucidar qué jugadores dominan más unas superficies que otras, qué comportamiento tienen y qué factores son influyentes. Analizaremos los 5 mejores jugadores en cada uno de los casos para poder concretar más, concentrando todos los partidos de una misma superficie como si se jugasen seguidos.

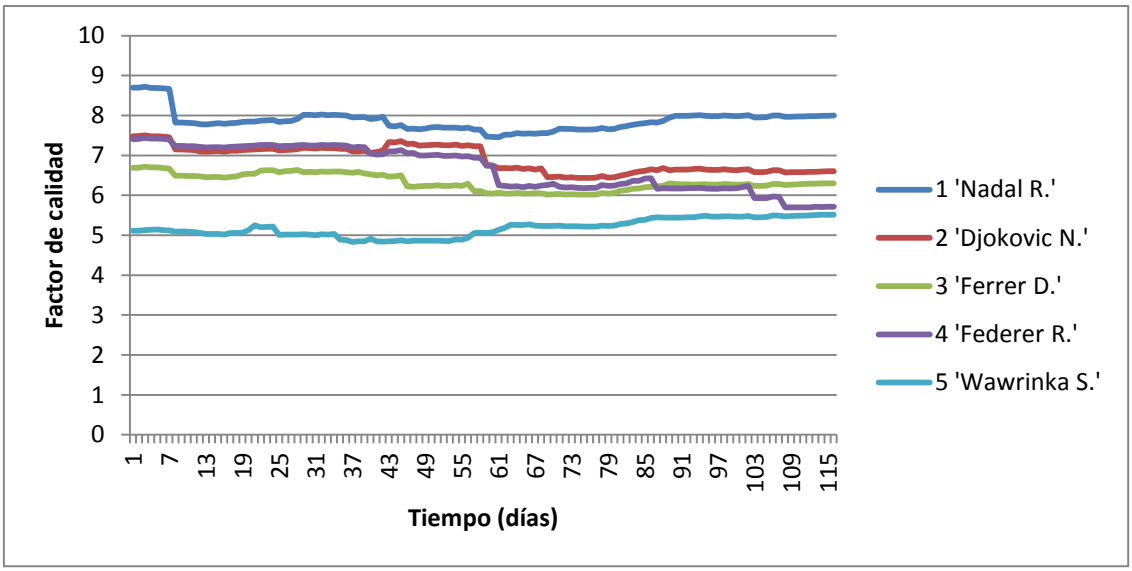


Figura 4.1: Factor de calidad en tierra. Tenis masculino 2012-2013.

El 35% los encuentros de una temporada (115 días) aproximadamente se disputan en tierra. Esto la convierte en una de las superficies más importantes haciendo que, si un jugador la domina bien, tenga muchas posibilidades de situarse en lo más alto del ranking. Como podemos comprobar en la Figura 4.1, el absoluto dominador en todo momento de los partidos en esta superficie es Nadal, de principio a fin. Ya decíamos que, por sus características propias de juego, más profundo y menos directo, y por su capacidad defensiva, es una superficie que le viene muy bien. Por otro lado, jugadores de talla mundial como Federer sufren más en este tipo de pistas al tener un juego más directo y rápido. Además, jugadores que en el ranking total suelen ocupar posiciones entre la 8 y la 10 como Wawrinka, destaca en este tipo de pistas situándose en el Top 5 de nuestro ranking y consiguiendo mantener una regularidad destacable.

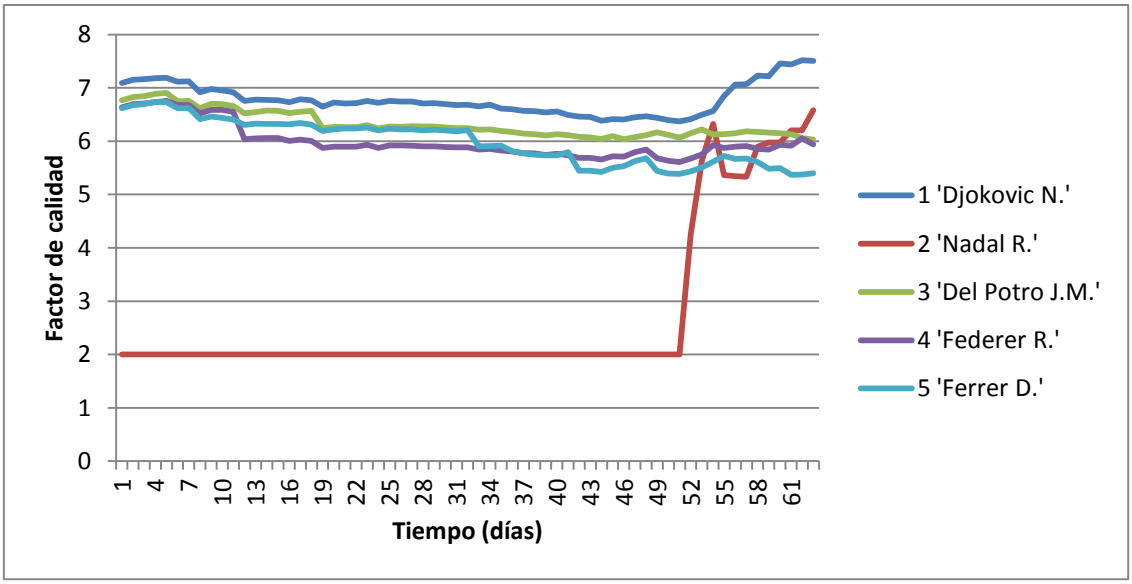


Figura 4.2: Factor de calidad en pista dura de interior. Tenis masculino 2012-2013.

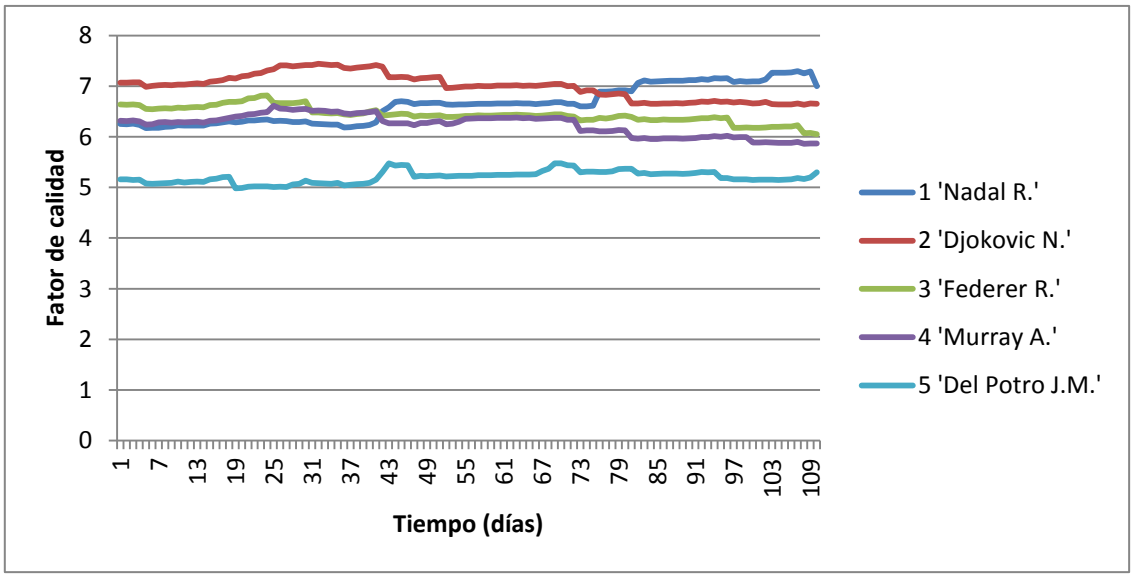


Figura 4.3: Factor de calidad en pista dura de exterior. Tenis masculino 2012-2013.

El porcentaje de partidos jugados en pista dura de interior es del 20%, mientras que en duras de exterior es del 35%, una cifra igual a la de los partidos jugados en tierra. Aquí Djokovic consigue obtener buenos números, sobre todo en pistas de interior donde consigue mejorar en todo momento como vemos en la Figura 4.2. Sin embargo, el mejor en pista dura de exterior es Nadal de nuevo, como se observa en la Figura 4.3 arrebatándole la primera posición a final de temporada al ganarle a Djokovic la semifinal del Master de Montreal. Si observamos además la Figura 4.2 de nuevo,, Nadal está en segunda posición porque no disputa torneos en esta superficie hasta final de temporada, cuando obtiene un buen valor. En general, Nadal consigue imponer su físico

y su excelente juego desde el fondo de la pista para imponerse en este tipo de superficies algo más rápidas que la tierra.

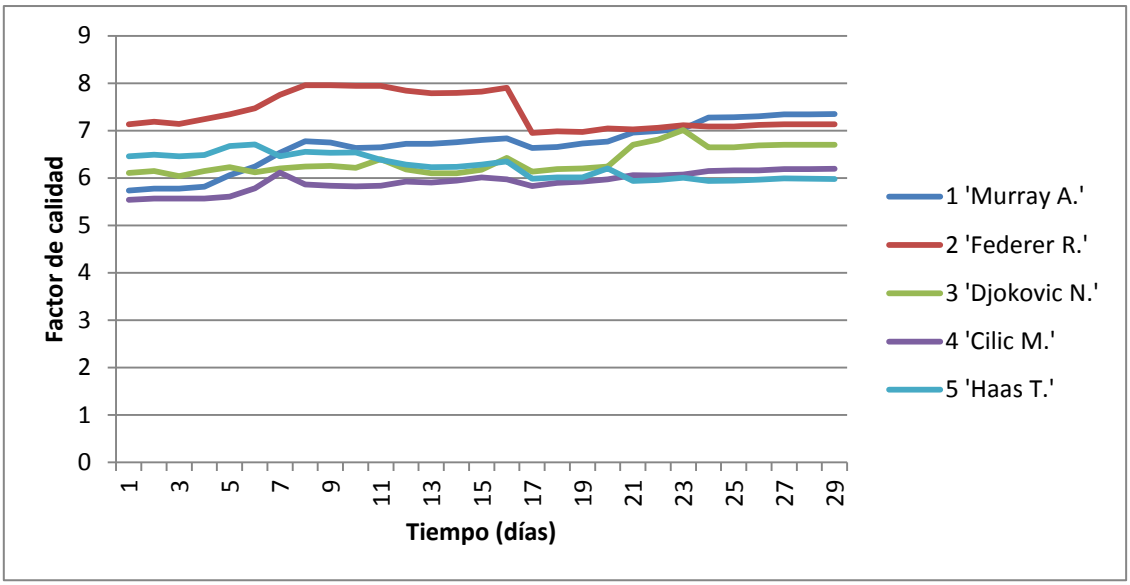


Figura 4.4: Factor de calidad en césped. Tenis masculino 2012-2013.

Es la superficie sobre la que menos encuentros se juegan, un 10% aproximadamente. Podemos ver en la Figura 4.4 que los dos mejores jugadores sobre esta superficie según nuestro ranking son Murray y Federer, tal y como presagiábamos en la sección 3. Su tipo de juego menos físico, más rápido y directo hace que se desenvuelvan mejor. Se observa además como jugadores con menor calidad como Cilic o Haas consiguen meterse en el Top 5, aprovechando muy bien sus cualidades. Pero un detalle llama la atención: Nadal, claro dominador en todas las facetas hasta ahora, no aparece en este Top 5. Esto es porque es una superficie que tradicionalmente no se le ha dado muy bien, a pesar de ganar algún torneo importante como Wimbledon. Pero como el número de partidos jugados en césped comparado con el resto de superficies, donde Nadal tiene dominio absoluto, es muy pequeño, le permite mantenerse en el nivel más alto de la clasificación a nivel global. También hay que destacar el excelente nivel de Djokovic, presente en el Top 3 en todas las superficies.

## C-Factor de enfrentamiento head to head

Lo primero que hacemos antes de aplicar este factor es comprobar que efectivamente tenemos material suficiente para poder llevarlo a cabo y decidir sobre qué jugadores lo vamos a hacer. Lo más interesante desde el punto de vista de las apuestas sería aplicarlo a los mejores jugadores. Por un lado, porque entre jugadores por debajo del TOP 10 no

es un fenómeno que se repita, y por otro, porque dentro de ese mismo TOP 10, además de sí producirse estos efectos en bastantes ocasiones, los tipos de partidos hacen que las cuotas a la hora de apostar sean más elevadas y sea mucho más interesante centrarnos en ellos, tratándose de encuentros más difíciles de predecir. Comprobamos además, que desde 2012, los enfrentamientos entre miembros del TOP 10 (tomado al finalizar 2013) se suceden, llegando a repetirse hasta 8 y 9 veces en algún caso. Después de analizar detenidamente el volumen de partidos y tras realizar la validación correspondiente, decidimos aplicar un factor diez veces superior, de tal manera que el peso para los partidos entre los mismos jugadores sería de 1 y para el resto de partidos de 0.1:

| Partido | Probabilidad Modelo anterior | Probabilidad modelo propuesto (con factor de enfrentamiento) |
|---|---|---|
| Berdych - Murray | 0,3436 | 0,5243 |
| Murray - Federer | 0,1981 | 0,292 |
| Ferrer - Tipsarevic | 0,786 | 0,8058 |
| Del Potro - Federer | 0,2662 | 0,3655 |

Tabla 4.6: Evolución probabilidad con factor de enfrentamiento. Tenis masculino 2012.

Posteriormente, aplicando exclusivamente este factor en los datos de test, obtenemos los siguientes resultados:

| Partido | Probabilidad modelo anterior | Probabilidad modelo propuesto (con factor de enfrentamiento) |
|---|---|---|
| Djokovic - Nadal | 0,4025 | 0,5012 |
| Djokovic - Wawrinka | 0,8497 | 0,8685 |
| Nadal - Federer | 0,6736 | 0,7746 |
| Murray - Federer | 0,2685 | 0,3593 |

Tabla 4.7: Evolución probabilidad con factor de enfrentamiento.

Tenis masculino 2012-2013.

Aquí el resultado de este factor es muy bueno, llegando a acertar partidos muy difíciles que antes no acertábamos en ningún caso como Djokovic-Nadal y mejorando otros resultados como los encuentros Djokovic-Wawrinka o Nadal-Federer, reflejados en la Tabla 4.7. Hemos conseguido establecer un modelo que nos predice muy bien enfrentamientos directos entre los mejores jugadores.

## D-Diferencias entre jugadores

| Modelo | Tenis Masculino |
|---|---|
| 2012-2013 Modelo anterior | 2,57 |
| 2012-2013 Modelo propuesto (con todos los factores) | 2,53 |

Tabla 4.8: Evolución diferencias entre jugadores. Tenis masculino.

En cuanto a las diferencias entre jugadores del Top 10 y del rango 90-100, podemos observar en la tabla 4.8 como la variación es leve tras aplicar el nuevo modelo con todos los factores, obteniendo resultados similares entre los mejores jugadores y entre los que se sitúan en el rango 90-100 en el nuevo ranking. Todo ello quiere decir que el modelo que hemos aplicado mejora a nivel general pero no establece diferencias grandes entre los jugadores, lo que nos lleva a la conclusión de que dicho modelo define bien un circuito masculino igualado y competitivo.

## E-Función de coste: LogLoss

Antes de mostrar los resultados, queremos dejar definido qué combinación de los tres factores del modelo propuesto es mejor aplicar. Haciendo una prueba sobre 2012 con las mejoras por separado, se sitúan alrededor del 0.61 de LogLoss, mientras que la prueba con todas las mejoras nos da resultados de 0.60, por lo que lo haremos de esta manera. El factor que más aporta es el factor de olvido, mientras que la aportación del resto es más reducida.

| Temporada | Tenis Masculino |
|---|---|
| 2013 (Modelo anterior) | 0,6442 |
| 2012-2013 (Modelo anterior) | 0,6185 |
| 2012-2013 (Modelo propuesto, con todos los factores) | 0,6013 |

Tabla 4.9: Evolución LogLoss. Tenis masculino.

Se puede ver en la Tabla 4.9 que las mejoras han sido notables. En 2012-2013 hemos obtenido un descenso del 3% respecto a la temporada homóloga con el modelo anterior y del 7% respecto al 2013 también con el modelo anterior. Esto nos dice que los factores utilizados en el modelo nuevo tienen un mejor efecto sobre el comportamiento

de los jugadores, proporcionándonos mejores resultados en factor de calidad, probabilidad y, por tanto, en LogLoss.

Finalmente, el Top 10 de nuestro modelo queda de la siguiente manera:

| Posición | Jugador | Factor de calidad |
|---|---|---|
| 1 | 'Nadal R.' | 6,39 |
| 2 | 'Djokovic N.' | 6,21 |
| 3 | 'Federer R.' | 5,55 |
| 4 | 'Ferrer D.' | 5,27 |
| 5 | 'Murray A.' | 5,25 |
| 6 | 'Del Potro J.M.' | 4,96 |
| 7 | 'Berdych T.' | 4,77 |
| 8 | 'Wawrinka S.' | 4,35 |
| 9 | 'Tsonga J.W.' | 4,34 |
| 10 | 'Gasquet R.' | 4,31 |

Tabla 4.10: Top 10 final. Tenis masculino.

Con dos claros dominadores como son Nadal y Djokovic, estando el primero por encima, y un ranking bastante igualado como acabamos de definir que llevará a que los partidos sean más emocionantes.

## 4.2.2. Tenis femenino

### A-Factor de olvido

Aquí debemos decidir de nuevo qué variantes de nuestro modelo son las mejores para aplicar en tenis femenino. Vamos a mantener la línea del tenis masculino de sólo modificar el parámetro beta, pero realizamos de nuevo las pruebas para comprobar que lo mejor es bajarlo y dejarlo en un valor de $\beta$=0.0025. Efectivamente esto sucede así, siendo los resultados los siguientes:

| Modelo | Tenis femenino |
|---|---|
| 2012 Modelo anterior | 0,6562 |
| 2012 Modelo propuesto (con factor de olvido) | 0,6200 |

Tabla 4.11: Evolución LogLoss con factor de olvido. Tenis femenino 2012.

Las mejoras son notables, reduciendo un 6% la función de coste respecto al modelo anterior reflejado en la Tabla 4.11. Esto nos habla de un circuito femenino en el que las mujeres son muy irregulares y dependen mucho de sus rachas y de su estado a la hora de disputar un encuentro, presentando un factor de calidad muy variable respecto al anterior modelo pero más ajustado a su perfil irregular.

## B-Factor de superficie

De igual manera, debemos analizar qué matriz S nos puede aportar más a la hora de obtener resultados en el tenis femenino. Tras distintas pruebas, aplicamos la misma matriz de superficies que en el tenis masculino, puesto que es la que mejores resultados nos otorga.

| Modelo | Tenis femenino |
|---|---|
| 2012 Modelo anterior | 0,6562 |
| 2012 Modelo propuesto (con factor de superficie) | 0,6525 |

Tabla 4.12: Evolución LogLoss con factor de superficie. Tenis femenino 2012.

Las mejoras obtenidas en este caso son del 0.6%, como puede verse en la Tabla 4.12. Hemos comprobado que la matriz definida en (3.10) es la que mejor puede definirnos a las jugadoras a la hora de obtener resultados, lo que nos dice que, por muchos cambios de superficie que haya o le demos distintos pesos más o menos grandes, las jugadoras no se ven afectadas y no mejoran los resultados, adaptándose bastante bien a las cualidades de cada tipo de superficie y notando poco los cambios.

Pasamos a continuación a mostrar los resultados en los partidos de test que nos han proporcionado las distintas superficies.

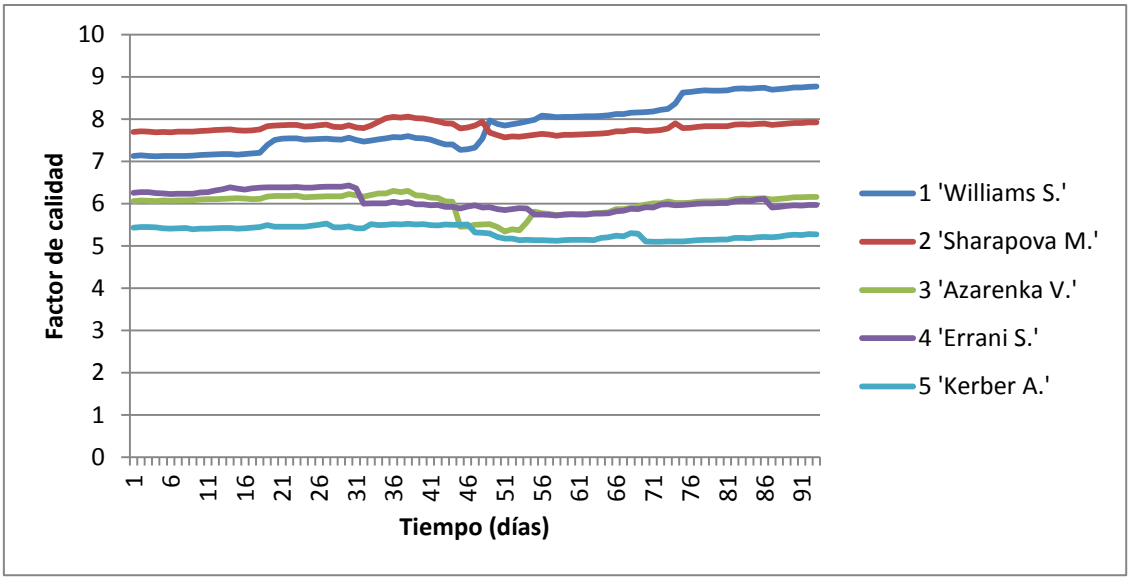


Figura 4.5. Factor de calidad en tierra. Tenis femenino 2012-2013.

En el circuito femenino se juegan el 30% aproximadamente de los encuentros en tierra (95 días). Como podemos ver en la Figura 4.5, las dos claras dominadoras en ese terreno son Serena Williams y Sharapova, estando la primera bastante por encima a final de temporada. El intercambio de posiciones entre ambas sucede a mitad de temporada, cuando Williams gana en cuartos de final del Open de Madrid a Sharapova, arrebatándole la primera posición.

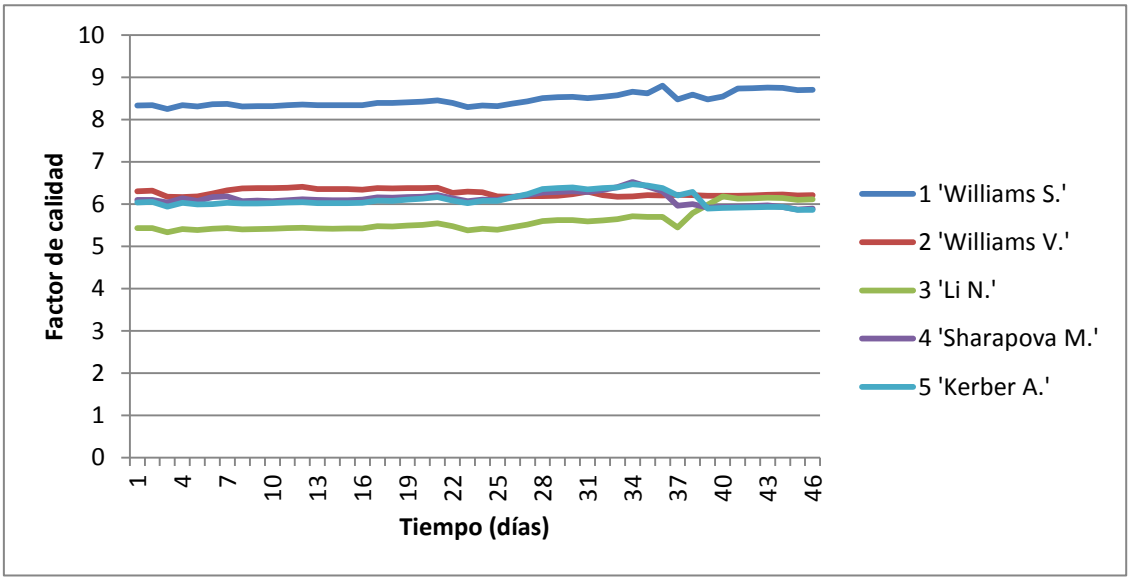


Figura 4.6: Factor de calidad en pista dura de interior. Tenis femenino 2012-2013.

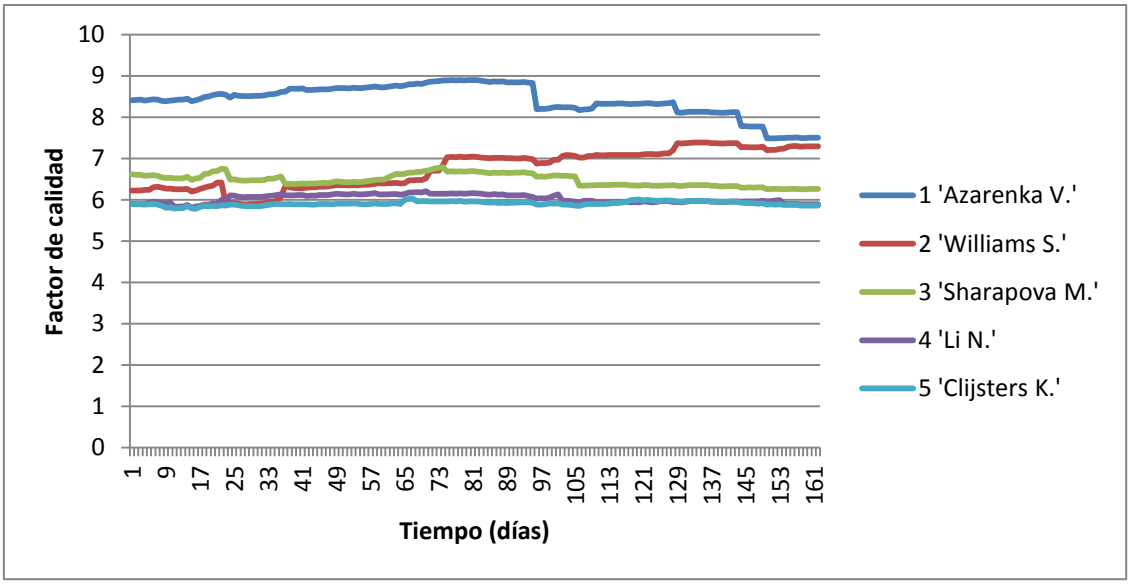


Figura 4.7: Factor de calidad en pista dura de exterior. Tenis femenino 2012-2013.

Aquí vuelve a repetirse la misma línea, reflejada en las Figura 4.6 y 4.7. Por un lado, en las pistas duras de interior, que representan el 15% de los juegos, la dominadora indiscutible es Serena Williams a más de 2 puntos de calidad del resto, entrando en el TOP 5 jugadoras de corte más físico y rápido como Venus Williams. Y por otro, en las pistas duras de exterior, a pesar de que la mejor jugadora es Azarenka, Serena Williams se queda muy cerca de ella, estando ambas muy por encima del resto.

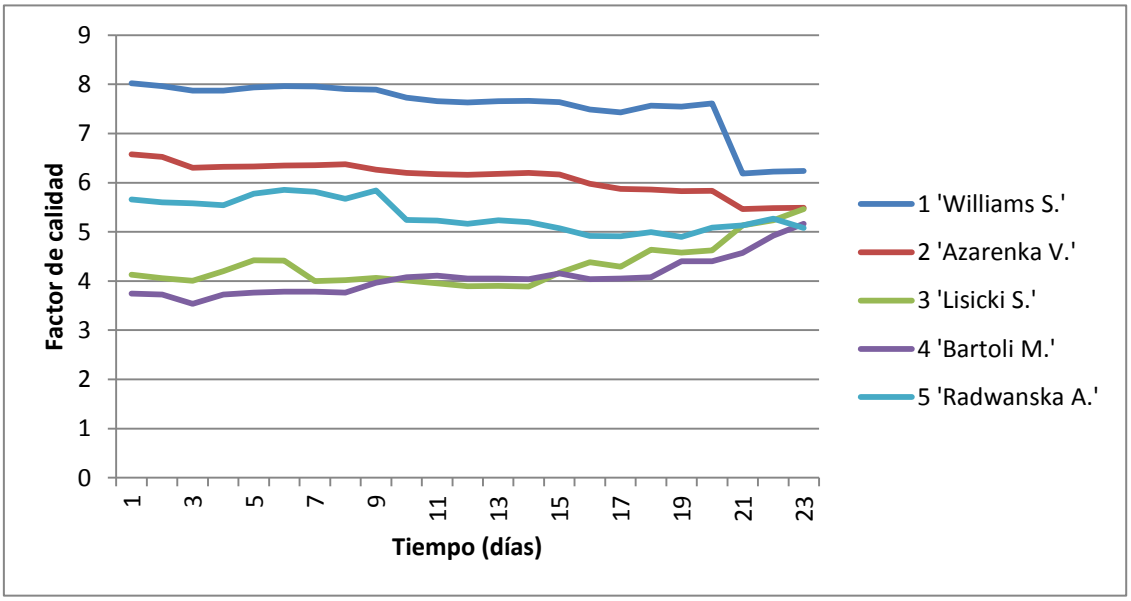


Figura 4.8: Factor de calidad en césped. Tenis femenino 2012-2013.

Una vez más, vemos en la Figura 4.8 como la clara dominadora es Serena Williams, en una superficie donde se juegan apegas el 8% de los partidos. El descenso de su valor al

final coincide con que en la temporada 2013 pierde Wimbledon, a pesar de lo cual se sigue manteniendo líder indiscutible. También dan la talla jugadoras importantes como Azarenka. Podemos observar que este tipo de superficie es en la que más cambios en el ranking hay dado sus condiciones específicas, entrando en el Top 5 jugadoras como Bartoli o Lisicki que habitualmente ni siquiera están en el Top 10.

A nivel general, Serena Williams domina de manera muy clara en superficies que suponen el 53% de los partidos, mientras que es segunda en la superficie en la que se disputa el 47% restante. Esto nos confirma que Serena Williams domina todo tipo de juego y va a ser mucho más difícil que alguna jugadora pueda derrotarla.

## C-Factor de enfrentamiento head to head

Siguiendo la misma línea que en apartado masculino, analizamos que tenemos encuentros suficientes como para poder aplicar este factor. Como el número de partidos es menor, se podría pensar en reducir el peso de los enfrentamientos directos a un valor tres veces superior, pero tras distintas pruebas, comprobamos que no mejora en absoluto y mantenemos un factor con diez veces más peso con buenos resultados:

| Partido | Probabilidad Modelo anterior | Probabilidad modelo propuesto (con factor de enfrentamiento) |
|---|---|---|
| Sharapova - Azarenka | 0,3497 | 0,3626 |
| Williams - Azarenka | 0,4822 | 0,509 |
| Li - Kerber | 0,4566 | 0,4736 |
| Errani - Stosur | 0,6106 | 0,6267 |

Tabla 4.13: Evolución probabilidad con factor de enfrentamiento. Tenis femenino-2012.

Al realizar el test con este factor obtenemos:

| Partido | Probabilidad Modelo anterior | Probabilidad modelo propuesto (con factor de enfrentamiento) |
|---|---|---|
| Sharapova - Azarenka | 0,3876 | 0,5193 |
| Kerber - Radwanska | 0,3983 | 0,5071 |
| Li - Azarenka | 0,2203 | 0,3178 |
| Williams- Radwanska | 0,902 | 0,9133 |

Tabla 4.14: Evolución probabilidad con factor de enfrentamiento.
Tenis femenino 2012-2013.

Acertando ahora encuentros difíciles como Sharapova-Azarenka que antes nos quedábamos muy lejos de acertar. Nos fijamos además que, en general, este factor tiene buen resultado en otros partidos del rango de posiciones del 4 al 10, obteniendo excelentes resultados concretos en partidos como Radwanska-Kerber.

## D-Diferencias entre jugadoras

| Modelo | Tenis femenino |
|---|---|
| 2012-2013 Modelo anterior | 2,76 |
| 2012-2013 Modelo propuesto (con todos los factores) | 2,47 |

Tabla 4.15: Evolución diferencias entre jugadoras. Tenis femenino.

En cuanto a las diferencias entre las jugadoras del Top10 y el Top100, observamos en la tabla 4.15 que la reducción ha sido drástica respecto al modelo anterior, del 11%. Una vez que hemos aplicado el modelo nuevo con todos los factores, la igualdad entre las jugadoras ha aumentado notablemente. Concretamente, en el nuevo ranking, las mejores jugadoras situadas en el Top 10 tienen descensos del factor de calidad, mientras que las jugadoras situadas en el rango 90-100 modifican sus valores muy poco e incluso los mejoran en algunos casos, lo que hace que la diferencia se reduzca por este lado. Esto significa que nuestro modelo tiene más efecto sobre unas jugadoras que sobre otras, manifestando una irregularidad notoria en función de los factores que tengamos en cuenta.

## E-Función de coste: LogLoss

Antes de mostrar los resultados, realizamos de nuevo las distintas combinaciones entre los factores de nuestros modelos para decidir cuál aplicamos. Haciendo una prueba sobre 2012 con las mejoras por separado, se sitúan en este caso alrededor de 0.62, mientras que la prueba con todas las mejoras nos da resultados de 0.613, por lo que lo haremos de la última forma. El factor que más aporta de nuevo es el factor de olvido, mientras que la aportación del resto es más reducida.

| Temporada | Tenis femenino |
|---|---|
| 2013 (Modelo anterior) | 0,6423 |
| 2012-2013 (Modelo anterior) | 0,6342 |
| 2012-2013 (Modelo propuesto, con todos los factores) | 0,6132 |

Tabla 4.16: Evolución LogLoss. Tenis femenino

Como podemos observar en la Tabla 4.16, tenemos unas mejoras del 5% respecto a 2013 y del 3% respecto al homólogo 2012-2013. El resultado es aceptable pero los valores finales se sitúan por encima del 0.61, haciéndonos ver que el modelo que hemos aplicado define bien a las jugadoras pero no lo suficiente como para obtener unos resultados mucho mejores. Esto nos lleva a pensar que el comportamiento de las jugadoras podría depender además de otros factores que no hemos tenido en cuenta en nuestro modelo, complicando mucho más las predicciones y los procesos. Finalmente, el TOP10 tras la aplicación de nuestro modelo con todos los factores queda:

| Posición | Jugadora | Factor de calidad |
|---|---|---|
| 1 | 'Williams S.' | 6,90 |
| 2 | 'Azarenka V.' | 6,50 |
| 3 | 'Sharapova M.' | 6,07 |
| 4 | 'Li N.' | 5,09 |
| 5 | 'Radwanska A.' | 5,01 |
| 6 | 'Errani S.' | 4,46 |
| 7 | 'Kerber A.' | 4,40 |
| 8 | 'Williams V.' | 4,39 |
| 9 | 'Kvitova P.' | 4,27 |
| 10 | 'Wozniacki C.' | 4,07 |

Tabla 4.17: Top 10 final. Tenis femenino

Donde destaca por encima de todas las jugadoras Serena Williams y, además, un TOP3 bastante superior al resto.

## 4.2.3. Comparación tenis masculino y femenino

### A-Factor de olvido

| Modelo | Tenis masculino | Tenis femenino |
|---|---|---|
| 2012 Modelo anterior | 0,6291 | 0,6562 |
| 2012 Modelo propuesto (con factor de olvido) | 0,6034 | 0,6200 |

Tabla 4.18: Comparativa evolución LogLoss con factor de olvido.

Al utilizar finalmente un factor de olvido que le da mayor importancia a los partidos de las últimas semanas, podemos ver claramente en la tabla 4.18 como, aunque los resultados finales son mejores en el circuito masculino, el porcentaje de mejora es mayor en el tenis femenino (un 6% frente a un 4%). La diferencia entre el tenis masculino y el tenis femenino se han reducido de 0.027 a 0.013, más de la mitad. Esto significa que las rachas de las jugadoras nos aportan una información muy útil a la hora de calcular probabilidades, y que los partidos en las últimas semanas son más determinantes para predecir que en el tenis masculino, donde la aportación no deja de ser buena. Lo que nos indica esto es que el circuito femenino es más irregular y que las jugadoras dependen más de su estado físico y anímico a la hora de jugar un partido.

### B. Factor de superficie

| Modelo | Tenis masculino | Tenis femenino |
|---|---|---|
| 2012 Modelo anterior | 0,6291 | 0,6562 |
| 2012 Modelo propuesto (con factor de superficie) | 0,6223 | 0,6525 |

Tabla 4.19: Comparativa evolución LogLoss factor de superficie

Como podemos observar en la Tabla 4.19, dentro de la leve mejoría, este factor afecta más al tenis masculino, con una mejora del 1% frente al 0.6% del femenino. Esto quiere decir que los resultados en este circuito van a depender más del tipo de pista en el que se juegue, y que por el contrario la jugadoras en tenis femenino se van a adaptar mejor a

los distintos tipos de pistas, notando menos los cambios y adaptándose igualmente bien a distintos tipos de juego y sus condiciones, más lento, rápido o profundo.

Aún así, es evidente que los resultados después de aplicar el factor de superficie no son los que nos gustarían. Hemos planteado tres alternativas que parecen no haber sido suficientes para optimizar la inferencia de los parámetros, por lo que quizá, para futuros experimentos, haya que aumentar los grados de libertad y haya que hacer un mayor número de pruebas, incluso 40 o 50, o cambiar el tipo de matriz por una S asimétrica, para saber de verdad qué importancia debemos darle a cada superficie para mejorar los resultados.

Además, del análisis separado por superficies que hemos realizado podemos concluir que Nadal y Serena Williams son los jugadores que mejores resultados obtienen en ambos circuitos al aplicar el factor de superficie. Aunque bien es cierto que Nadal destaca bastante menos y tiene rivales más directos como Djokovic, al contrario que Selena, que no tiene ninguna jugadora que le haga sombra de manera directa.

## C. Factor de enfrentamiento directo head to head

A continuación mostramos algunos ejemplos de los buenos resultados que nos ha arrojado la aplicación de este factor en ambos circuitos:

| Partido | Probabilidad Modelo anterior | Probabilidad modelo propuesto (con factor de enfrentamiento) |
|---|---|---|
| Djokovic - Nadal | 0,4025 | 0,5012 |
| Murray - Federer | 0,2685 | 0,3593 |
| Kerber - Radwanska | 0,3983 | 0,5017 |
| Li - Azarenka | 0,2203 | 0,3178 |

Tabla 4.20: Comparativa evolución probabilidad con factor de enfrentamiento

En general, obtenemos resultados similares tanto en tenis femenino como en masculino, reflejados en la Tabla 4.20, donde conseguimos acertar partidos como Nadal-Djokovic o Kerber–Radwanska, que antes los errábamos, mejorando también probabilidades que antes teníamos muy bajas y que nos podían hacer perder mucho dinero, como los encuentros Murray-Federer o Li-Azarenka. Hemos conseguido solucionar bastante bien el matiz de los enfrentamientos directos.

## D. Diferencias entre jugadores

| Modelo | Tenis masculino | Tenis femenino |
|---|---|---|
| 2012-2013 Modelo anterior | 2,57 | 2,76 |
| 2012-2013 Modelo propuesto (con todos los factores) | 2,53 | 2,47 |

Tabla 4.21: Comparativa evolución diferencias entre jugadores.

Finalmente, podemos observar en la Tabla 4.21 como las diferencias tras aplicar el nuevo modelo son ahora menores en el tenis femenino. Esto significa que el nuevo modelo ha tenido efecto y nos ha llevado hacia un tenis femenino mucho más igualado de lo que se nos presentaba al inicio del experimento. Pero como hemos explicado, esa igualdad sólo ha sido posible por la gran adaptación por la parte de las jugadoras del TOP 10 al mismo que no ha tenido sin embargo el rango 90-100. Por ello cabe pensar que tanto el factor de superficie como el factor de olvido y el factor head to head han supuesto un efecto desigual en las distintas jugadoras que nos hace concluir que el tenis masculino ha mantenido la igualdad tras la aplicación del nuevo modelo mientras que el circuito femenino se presenta mucho más irregular, variable e impredecible ante determinados factores. Los mejores jugadores en ambos circuitos tras aplicar el nuevo modelo son Nadal, seguido de cerca por Djokovic en el circuito masculino, y Serena Williams, clara dominadora en el circuito femenino.

## E. Análisis de función de coste: LogLoss

Mostramos a continuación un gráfico comparativo, tras aplicar los tres factores en conjunto en el nuevo modelo:

| Modelo | Tenis masculino | Tenis femenino |
|---|---|---|
| 2013 Modelo anterior | 0,6442 | 0,6423 |
| 2012-2013 Modelo anterior | 0,6184 | 0,6342 |
| 2012-2013 Modelo propuesto (con todos los factores) | 0,6013 | 0,6132 |

Tabla 4.22: Comparativa evolución LogLoss

Partíamos de unos resultados del modelo previo parejos en 2013 e inferiores en el tenis masculino en 2012-2013. Ahora con el nuevo modelo se confirma a través de la Tabla 4.22 esta tendencia en las diferencias y conseguimos dejar la función de coste en 0.6 en el tenis masculino, mientras que en el tenis femenino se queda cerca del 0.61, una mejora grande pero no suficiente, que ha hecho que aunque las mejoras respecto al modelo anterior sean iguales (del 3% aproximadamente) no se lleguen a los resultados esperados. Esto significa que, mientras que en el tenis masculino este modelo nuevo proporciona factores de calidad que definen bien a los jugadores en cada situación, en el tenis femenino nos aporta bastante pero se queda corto, probablemente porque necesitaríamos tener en cuenta otros factores que no hemos incluido ahora que necesitarían de una profunda investigación.

## 4.3. Resultados cuantitativos finales

En esta sección, con la que cerraremos el apartado de presentación de resultados, utilizaremos las probabilidades para realizar una simulación sobre las casas de apuestas y ver si conseguiríamos o no obtener beneficios. El análisis se realizará por separado en tenis masculino y femenino, como se ha venido haciendo hasta ahora.

Hay que destacar antes de comenzar que es muy complicado ganar a largo plazo a las casas de apuestas, no solamente porque tengan una plantilla muy amplia trabajando constantemente sobre ello para no perder dinero, sino porque en todos los partidos la probabilidad total que marcan es superior a 1 para asegurarse las ganancias. Vamos a mostrar a continuación los resultados que obtendría la casa de apuestas BET365 (que es la que seleccionamos en la sección 2) con nuestra función LogLoss:

| Temporada | Masculino | Femenino |
|---|---|---|
| 2013 | 0,4936 | 0,4913 |
| 2012-2013 | 0,4894 | 0,4974 |

Tabla 4.23: LogLoss Bet365

En la Tabla 4.23, como ya decíamos, podemos observar unos resultados muy buenos en ambos circuitos muy difíciles de superar. Debemos también puntualizar que los resultados de la función de coste del resto de casas de apuestas son muy similares, confirmando que utilizar la casa BET365 para apostar es una decisión correcta por todo lo expuesto en la sección 2.

Antes de comenzar, vamos a intentar analizar el modelo que hemos establecido tras aplicar la validación en la temporada 2012. Esto lo hacemos con el objetivo de estudiar dónde funciona mejor finalmente, qué partidos acertamos con más facilidad o dónde no debemos arriesgar tanto para, en su posterior aplicación en 2013, comprobar si podemos extenderlo y nos proporciona buenos resultados.

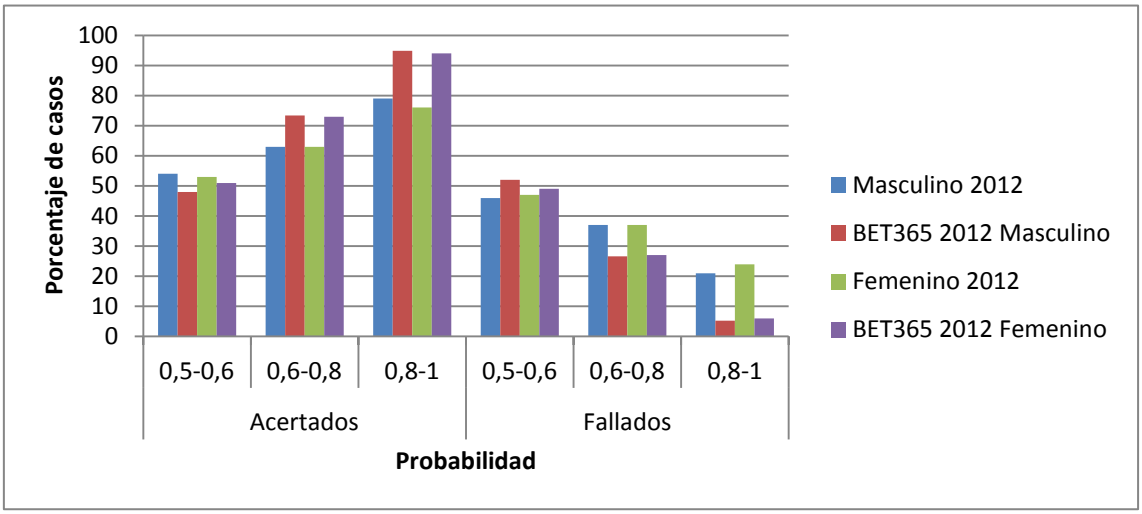


Figura 4.9: Histograma

En la Figura 4.9 definimos el porcentaje de partidos acertados y fallados en diversas franjas de probabilidad, tanto en nuestro modelo como por parte de la casa de apuestas. Aquí hay varias cosas a tener en cuenta. Por un lado, al terminar nuestra fase de validación, vemos que en porcentaje de probabilidad tenemos más éxito en el tenis masculino que en el femenino, con un porcentaje mínimo pero suficiente para marcar diferencias. Y por otro lado, las casas de apuestas en ambos circuitos superan con creces nuestros resultados. Pero hay algo curioso, y es que cuando damos una probabilidad entre 0.5 y 0.6, somos capaces de obtener mejores resultados que cuando es la casa de apuestas quien tiene que dar una probabilidad de ese tipo. Este porcentaje superior puede suceder porque, al aplicar nuestro modelo, somos capaces de acertar muchos resultados que con el modelo previo estaban al borde de ser acertados (con probabilidad de 0.4-0.5), mientras que el resto conseguimos mantenerlos por lo general, unido a que a la casa de apuestas le cuesta mucho acertar partidos con esta probabilidad debido a su complejidad.

Esto nos abre una pequeña grieta a nuestro favor. Al aplicar los resultados cuantitativos, vamos intentar por lo tanto tener en cuenta estos baremos para realizar apuestas en las franjas que nosotros creemos más interesantes. Estableceremos también comparativas de los resultados en los 3 últimos meses de la temporada puesto que, si el aprendizaje máquina funciona, el modelo debería haber aprendido y obtendría en ese rango mejores

resultados. Por su puesto, todo esto tiene que probarse en datos distintos (2013 en nuestro caso), con la esperanza de que las tendencias se cumplan como en 2012.

### 4.3.1. Tenis masculino

Vamos a hacer pruebas con el modelo de 2012-2013 de validación cruzada comentado, estableciendo distintas comparativas. A continuación mostramos los resultados que obtendría BET365 aplicando sus propias probabilidades, jugando 1€/partido como cantidad estándar:

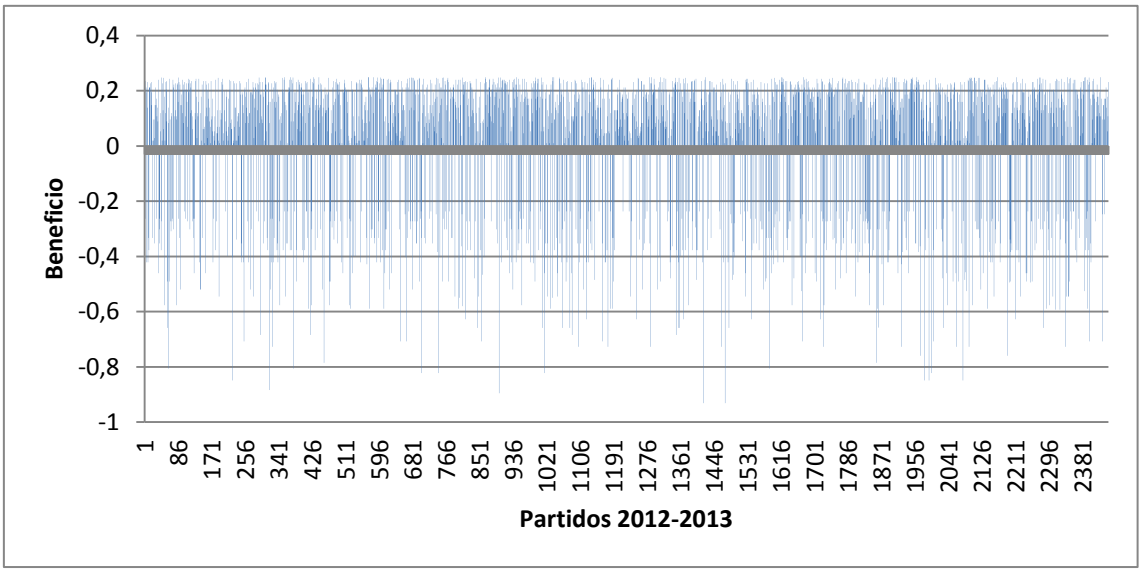


Figura 4.10: Dinero ganado/perdido Bet365. Tenis masculino 2012-2013.

Las gráficas que presentemos para analizar las pérdidas y ganancias tendrán el formato de la Figura 4.10, donde presentamos un eje de abscisas donde recogemos los partidos jugados en la temporada correspondiente por orden cronológico, mientras en el eje de ordenadas reflejamos las ganancias o pérdidas que obtenemos de cada uno de esos encuentros.

Claramente podemos ver en la figura 4.10 como los beneficios son holgados, con una media de 1,3 céntimos ganados por partido, aunque estas ganancias se reducen a la mitad si se trata de los 3 últimos meses. Es curioso observar como la ganancia máxima por partido apenas supera los 0.2€, pero que estas ganancias se repiten en muchísimos partidos, lo que hace que al final el dinero se vaya acumulando y el balance sea positivo.

Por otra parte, los resultados cuantitativos que nos ha dado nuestro modelo son:

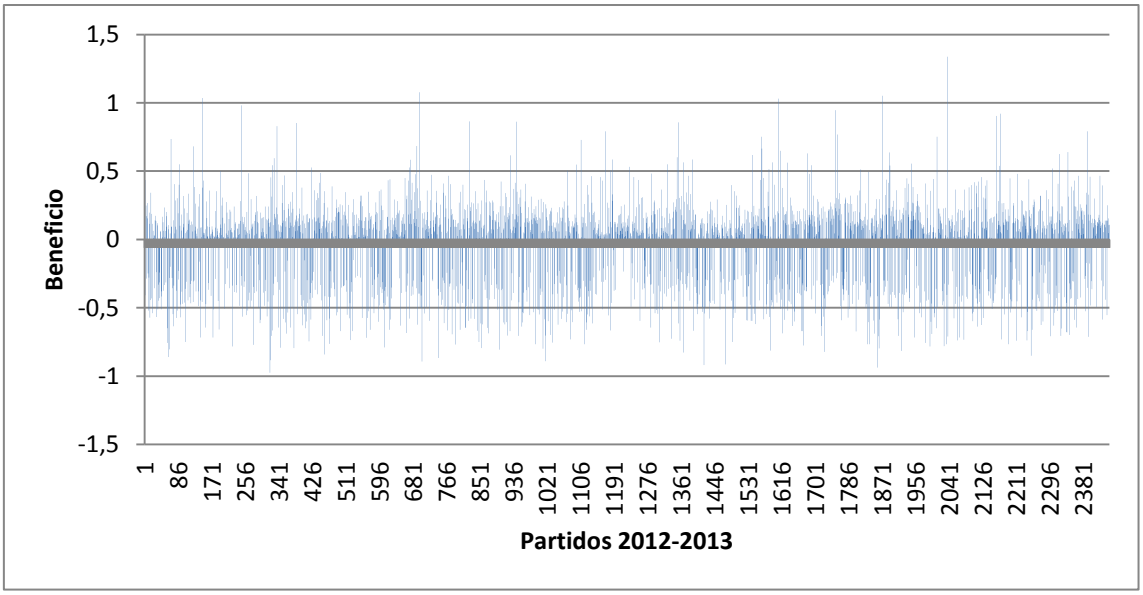


Figura 4.11: Dinero ganado/perdido 2012-2013. Tenis masculino

Aquí, en la figura 4.11, se muestran unos resultados más igualados entre pérdidas y ganancias. Podemos observar varios aspectos positivos, como que varias veces se obtienen beneficios por encima de 1€ ganado en un mismo partido y que existen zonas donde ganamos más dinero, como puede ser alrededor del partido 1210 y al final de temporada.

En balance, perdemos algo menos de 4 céntimos por partido, mientras que en los 3 últimos meses, cuando el modelo ha optimizado el aprendizaje máquina, las pérdidas se reducen a 2 céntimos y medio de € por partido. Evidentemente, no son malos resultados después de 2443 partidos apostados, frente a las condiciones de las casas de apuestas. Aún así, no son resultados óptimos y tenemos que buscar alguna manera de exprimir mejor nuestro modelo en función de todas las conclusiones obtenidas mediante la validación.

A nivel de ganancias conseguimos situarnos más cerca de la casa de apuestas, pero el problema lo tenemos en las pérdidas. Para paliar esto y llegar a obtener beneficios, podemos basarnos en el modelo que hemos desarrollado con la validación en 2012. En este, la franja de probabilidad en la que más porcentajes de partidos acertamos respecto a la casa de apuestas es la 0.5-0.6. Podemos aprovechar esta diferencia a nuestro favor para probar en los datos de test y apostar solamente en partidos que estén en esta franja, y comprobar así si los datos inferidos nos proporcionan buenos resultados consiguiendo superar a la casa de apuestas:

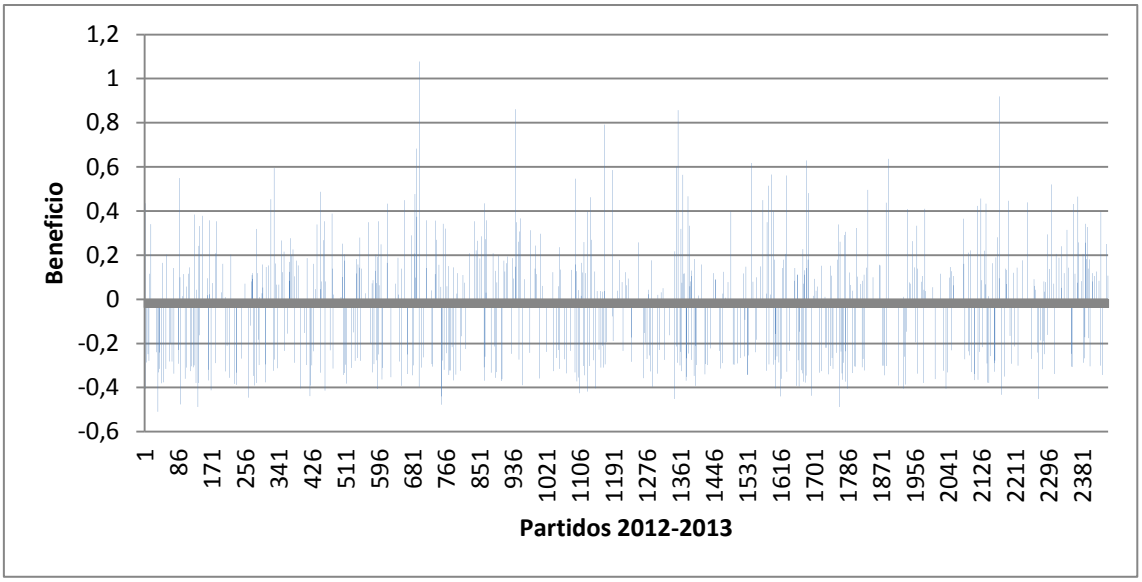


Figura 4.12: Dinero ganado/perdido 2012-2013 con umbral. Tenis masculino.

Como podemos observar en la Figura 4.12, arriesgamos menos dinero, lo que nos ayuda a quitarnos prácticamente la lacra de las pérdidas, y además mejoramos muchísimo los números hasta dejar la balanza en prácticamente 0€: las pérdidas son de 0.1 céntimos de € por partido, prácticamente inapreciables. Eso sí, cuando el análisis lo hacemos en los últimos 3 meses, por fin conseguimos obtener beneficios de 1.2 céntimos de € de media por partido, algo que se asemeja a las ganancias de la casa de apuestas. Estos resultados, teniendo en cuenta las condiciones a las que nos enfrentábamos, pueden calificarse de excelentes. Los mostramos a continuación:

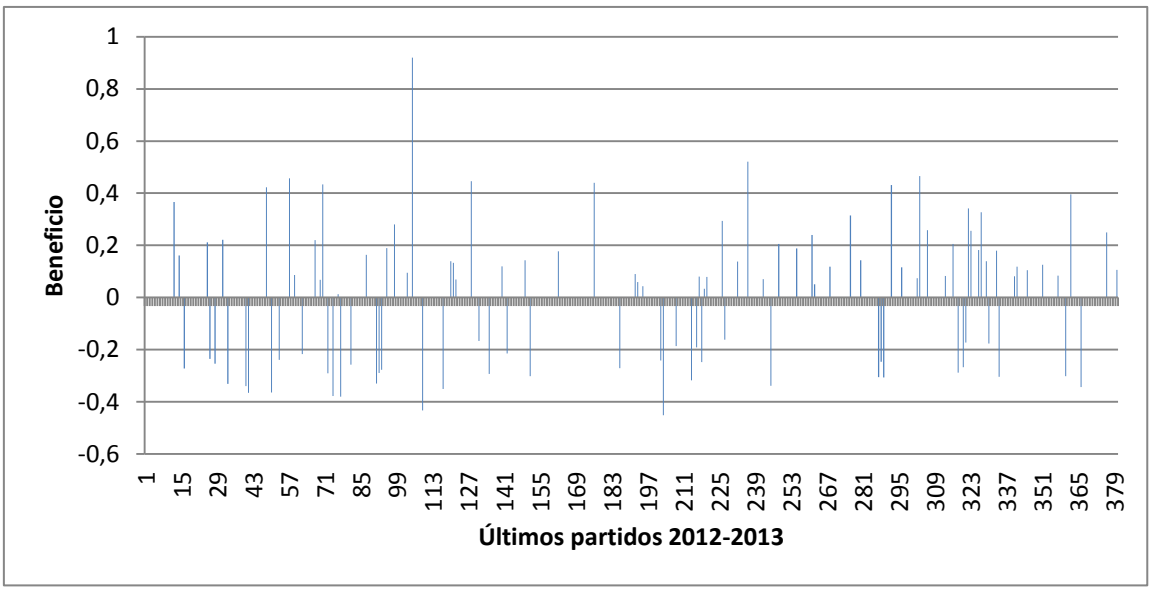


Figura 4.13: Dinero ganado/perdido con umbral.
Tenis masculino 2012-2013 últimos 3 meses

En la Figura 4.13 podemos observar que, además de obtener beneficios a final de temporada, hay una franja de unos 40 partidos entre el 240 y el 281 que conseguimos acertar todos los encuentros a los que apostamos, repitiéndose este fenómeno en más ocasiones.

## 4.3.2. Tenis femenino

Vamos a mostrar primero los resultados que obtendrían las casas de apuestas jugando de nuevo una cantidad estándar de 1€/partido:

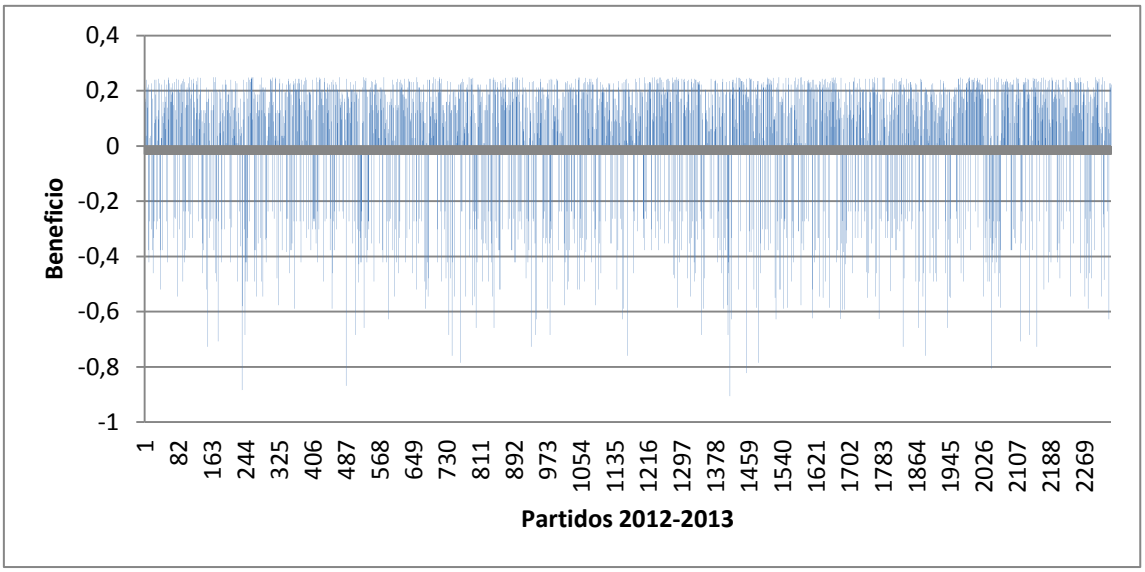


Figura 4.14: Dinero ganado/perdido Bet365. Tenis femenino 2012-2013.

Marca picos máximos en poco más de 0.2, observando que las pérdidas que pueden llegar a tener son bajas y que los partidos en los que ganaría las apuestas serían elevados en toda la temporada. La ganancia total es de 0.02€ por partido, elevándose a los últimos 3 meses a 0.045€ por partido, unas cantidades bastante elevadas.

En cuanto a nuestros resultados aplicando el modelo desarrollado son:

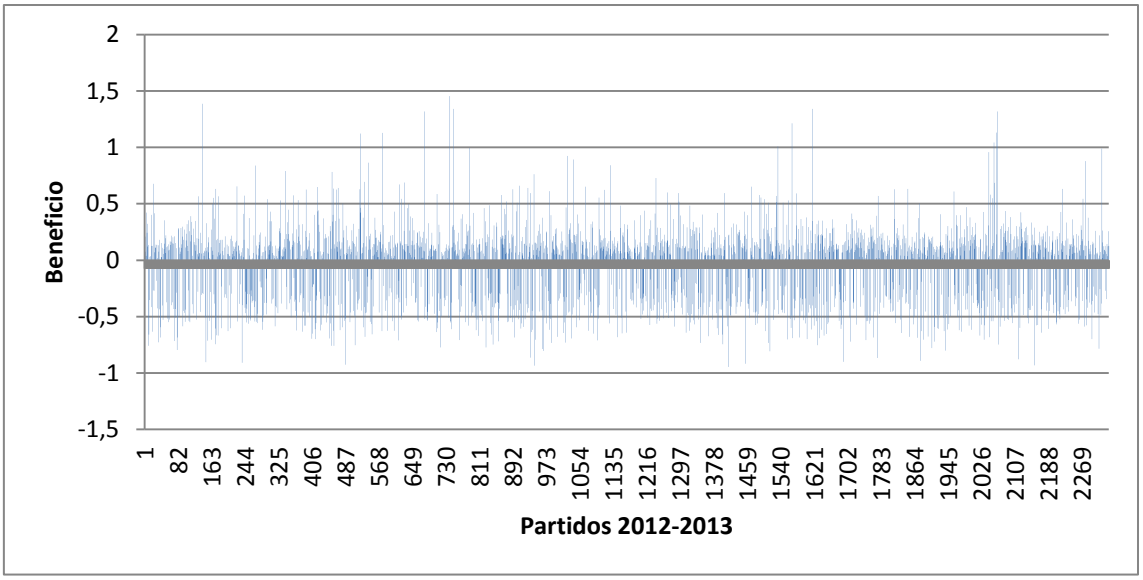


Figura 4.15: Dinero ganado/perdido 2012-2013. Tenis femenino.

Marcamos unos resultados bastante equilibrados, obteniendo en algunos encuentros más de 1€ de beneficio neto. Eso sí, en la Figura 4.15 se observa que muchas de las pérdidas son mayores de 0.5€ en 1 partido, una cantidad que a la larga nos lastra y deja los resultados en pérdidas de alrededor de 3 céntimos por partido.

Aunque la mejora de nuestro modelo no ha sido muy grande, en la franja 0.5-0.6 nuestros resultados son mejores que la casa de apuestas, por lo que vamos a aplicar las mismas mejoras que en el tenis masculino.

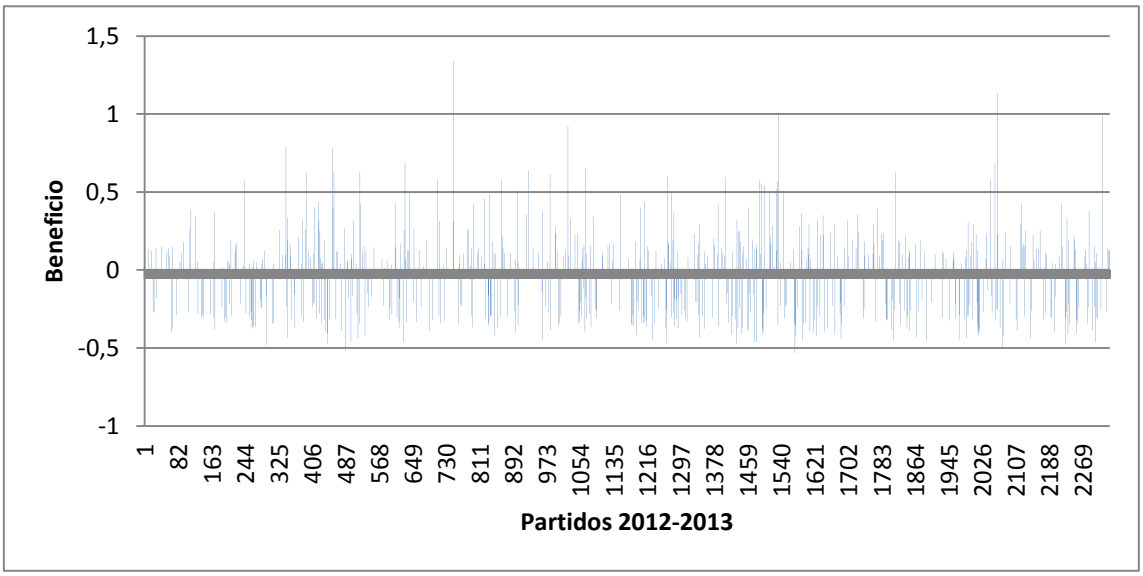


Figura 4.16: Dinero ganado/perdido con umbral. Tenis femenino 2012-2013.

Conseguimos rebajar las pérdidas a medio céntimo por partido, como se puede ver en la Figura 4.16, pero en los últimos 3 meses triplicamos esta cifra hasta perder 1.5 céntimos por partido. No conseguimos dejar el balance a 0 ni ganar nada, pero conseguimos minimizar las pérdidas todo lo posible.

### 4.3.3. Comparación tenis masculino y femenino

Aunque los resultados eran igualados aplicando el modelo anterior, los resultados del nuevo modelo hacían presagiar mejores números a nivel cuantitativo en el tenis masculino que en el femenino. Finalmente, confirmándose lo que preveíamos, podemos observar que los resultados a nivel cuantitativo son mejores en tenis masculino, donde los números de los tres últimos meses mejoran hasta dar beneficios de 1.2 céntimos por partido mientras en el tenis femenino seguíamos perdiendo hasta 1.5 céntimos por partido.

Así, tras comprobar la efectividad del modelo simulando en una casa de apuestas real online, obtenemos tres conclusiones claras:

- Los números que presentan las casas de apuestas son muy difíciles de superar a nivel general. Pero si exprimimos nuestro modelo al máximo conseguimos obtener beneficios de 1.2 céntimos de media por partido en los últimos tres meses en tenis masculino, todo un éxito teniendo en cuenta a qué nos enfrentamos.

- Los resultados son notablemente mejores en el tenis masculino que en el tenis femenino.

- Aunque las pérdidas en tenis masculino son casi inapreciables, 0.1 céntimos/partido, si variamos la cantidad apostada por partido y la aumentamos al doble o el triple, al hacerlo sobre un número muy grande de encuentros, estas mínimas diferencias se magnifican y no mejoramos los beneficios en ningún momento, sino que aumentamos las pérdidas.

- Concluimos que no es rentable apostar en el tenis femenino. Además de que con nuestro modelo no hemos sido capaces de desarrollar buenos resultados, las cuotas no son las más favorables y los buenos resultados de las casas de apuestas en este ámbito nos lo ponen aún más complicado.

# 5. Presupuesto y planificación del trabajo

En el presente capítulo detallaremos fase por fase la planificación que hemos seguido para desarrollar el proyecto, apoyada en un diagrama de Gantt. Además, detallaremos en un presupuesto los datos y los gastos correspondientes, justificando las decisiones que estimemos oportunas.

## 5.1. Presupuesto

En este apartado vamos a presentar un presupuesto del proyecto dejando claro que, al ser un proyecto de desarrollo a nivel de aplicación, vamos a reportar únicamente los costes de mano de obra del ingeniero que lo desarrolla, sin introducir materiales ni equipos (ordenador, licencias de software, etcétera) que entendemos que están ya implícitos y asumidos por el propio departamento o empresa que ofrece el producto. El presupuesto es el siguiente, contando con que contratamos a un ingeniero con un salario al mes de 2144.25€ (según la media que nos dice el COIT, Colegio Oficial de Ingenieros de Telecomunicación, que cobra un ingeniero cada mes con menos de 5 años de experiencia):

Presupuesto

| | |
|---|---|
| "Dirección" | Calle Río Duero |
| "Ciudad" | Leganés |
| "Provincia" | Madrid |
| "Telefóno" | 625959100 |
| "Codigo Post | 28913 |
| "DNI" | 53457478G |

| | | | |
|---|---|---|---|
| Fecha Solicitud | 15-12-14 | Método de pago | Transferencia bancaria |
| Solicitado por: | Cliente | | |

| Producto | Cantidad | Descripción | Precio / mes | Precio |
|---|---|---|---|---|
| 1 | | PROYECTO APUESTAS DEPORTIVAS | 2.144,00 € | 19.296,00 € |
| 2 | | | | |
| 3 | | | | |
| 4 | | | | |
| 5 | | | | |
| 6 | | | | |
| 7 | | | | |
| 8 | | | | |
| 9 | | | | |
| 10 | | | | |

| | | | |
|---|---|---|---|
| | | Subtotal | 19.296,00 € |
| *Si tiene alguna duda sobre este presupuesto* | 21,00% | IVA | 4.052,16 € |
| *no dude en comunicarse con nosotros* | | Costes de Envio | |
| | | Seguro | |
| | | Total | 23.348,16 € |

Figura 5.1: Presupuesto

# 5.2. Planificación

Hemos realizado nuestro proyecto en cinco fases que han evolucionado de la siguiente manera:

- **Fase 1: Documentación y estudio.-** Para comenzar, dedicamos un tiempo al estudio de toda la documentación en varios formatos acerca de apuestas deportivas, modelos probabilísticos, aprendizaje máquina, propiedades del tenis y una serie de conceptos que necesitamos conocer a fondo antes de implementar un modelo propio y que nos van a proporcionar una mayor capacidad para después plantear alternativas o añadir mejoras.

- **Fase 2: Reproducción del modelo existente.-** Revisamos a fondo el modelo anterior existente para analizar su estructura, sus resultados, lo que podemos mejorar, las partes que tienen un mayor margen de mejora y las posibles modificaciones para un modelo nuevo

- **Fase 3: Propuesta de modelo nuevo.-** Desarrollamos un modelo nuevo con el objetivo de mejorar el anterior y lo aplicamos para tenis femenino y masculino.

- **Fase 4: Resultados finales y simulación.-** En esta parte recogemos los análisis finales y resultados de la aplicación de nuestro modelo, llevamos a cabo la simulación sobre números reales y obtenemos los resultados a nivel cuantitativo.

- **Fase 5: Memoria del proyecto.-** Recogemos todos los resultados y los analizamos al detalle, introduciendo comparaciones, gráficas y alternativas a las distintas soluciones para plasmarlo en un documento escrito.

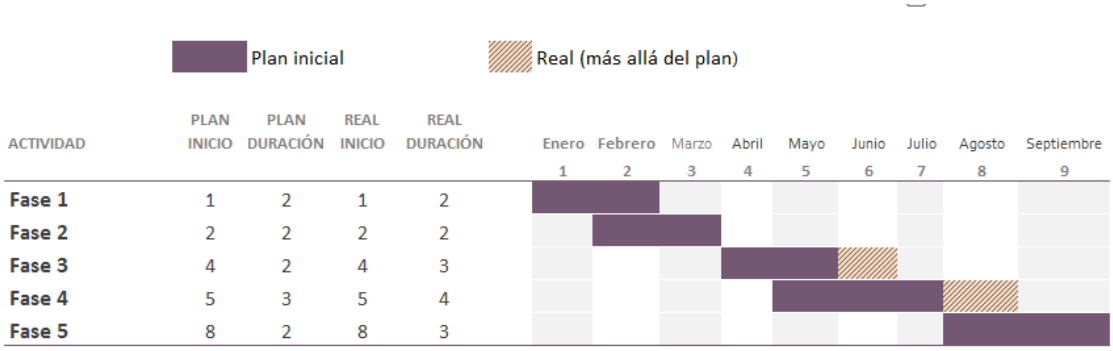


Figura 5.2: Diagrama de Gantt

Nuestro proyecto estaba programado para tener una duración de 9 meses, tal y como refleja la Figura 5.2. En los 3 primeros, recabaríamos toda la información posible y reproduciríamos el modelo anterior. En los 2 siguientes, desarrollaríamos una propuesta nueva de modelo, mientras en paralelo se obtendrían resultados según fuesemos avanzando en la obtención del mismo. Y para acabar, mientras terminásemos de obtener los resultados, iríamos escribiendo la memoria del trabajo, realizando un análisis muy detallado, introduciendo variables nuevas de medida y plasmándolo todo por escrito. Finalmente, el proyecto ha durado los 9 meses que esperábamos. Sin embargo, en la fase 3, de desarrollo del nuevo modelo, utilizamos un mes de más porque creíamos que merecía la pena detenerse un poco más debido a las numerosas alternativas que podíamos ofrecer, para exprimir al máximo el modelo que teníamos delante y poder presentar unos resultados de calidad. Veíamos por supuesto que este mes adicional no suponía ningún estrago a la hora de entregar en fecha el proyecto y que los tiempos estaban perfectamente controlados, como así ha sido.

# 6. Conclusions and future work.

## 6.1. Conclusions.

Since the beginning, our intention was to reproduce the last model, analyze it and propose a new model to improve the first one as much as possible, with the ultimate goal of obtaining clear conclusions about the male and female tours, getting good numbers to bet on sports bookmakers. We analyze what we have conclude in each part:

- A wider range of results has made us conclude that, to apply our model and to develop by machine-learning, it's better to use two consecutive seasons than just one, because the information we provide is continually useful and we are improving the predictions the study of those same seasons goes forward

- The analysis of the results of male and female tours shows clear differences between both. We find that men's tennis is more equaled, more competitive and with better results at LogLoss after the application of our model, though still with room for improvement. However, women's tennis players shows a more irregular LogLoss results which are not expected following the implementation of the new model, though there's still room for improvement for both tours. This is because in women's tennis could affect more factors that have not been considered in this model and that would lead us to a more complex model that would be more difficult to calculate probabilities.

- Important factors of our model give us specific interesting conclusions. At first, the forgetting factor affects the women's tour more than male's, although both tours report great strides. This tells us that women players are more dependent on their spells and physical and mental state in which they are at matches. Furthermore, although the surface factor does not improve everything we would have liked, qualitatively it tells us that, in spite of women players notice the different types of surface in his game, is in men's tennis where players suffer more these changes. Players like Nadal notice an imbalance in the quality factor when playing on grass, and players like Djokovic or Federer, on clay. Finally, factor clash head to head does not bring great benefits globally, but has had great results in the matches played between players TOP10. We are now capable of hitting very difficult games to predict due to their equality as Nadal-Djokovic, Sharapova-Azarenka, Wawrinka Ferrer-Radwanska and Kerber previously gone astray.

- Although it is true that the differences are noticeable between women's and men's tennis, we can conclude that Serena Williams on the one hand, and Nadal and Djokovic on the other hand, are indisputable leaders of both tours in all circumstances of the analyzed seasons.

- Fully squeezing our model and applying cross-validation, we managed to minimize losses when betting to almost 0 and we have even made a profit in the last three months of the 2012-2013 season, something very difficult to achieve. We could only get this profit it in men's tennis since, in addition to our model has provided better numbers at this tour, betting bookmarks put it more complicated in women's tennis, being a not very recommendable tour to bet at these conditions. It is true that benefits in male tennis are not very big, but we have improved the results of the above experiment and we managed to find a crack in the sportsbook, albeit minimal.

- We can finally conclude after all that it is very complicated to obtain big benefits in a bet system against the sports bookmark. Therefore, we must be aware of this everytime, and recomend to every people of this field betting with the maximum rigor and solidity, to establish a determined amount of money for betting that should never be exceeded, and never betting just emotionally, without a solid argument, to avoid to be ruined.

- Finally, we can say that the objectives have been met. We have made a very thorough analysis of the male and female tours obtaining clear conclusions and contrasted from a new model with three pillars (forgetting factor, area factor and factor of direct confrontation), we have added new elements of comparison as the average difference and measuring the differences between players and also quantitatively, we have obtained good results, fulfilling the above in Section 1.

In conclusion we would like to expose the difficulties we encountered when developing the project. On the one hand, although it has reduced the runtime code, we were driving over 4000 games in most executions, which made for each test we had to wait 20 to 25 hours to get the results, depending implementation. This represents almost exclusive dedication of resources for most of the day and also at night. On the other hand, when the online betting sector is a relatively new development in our country and laws are just four years, it is more difficult to find specific quality information on it and its characteristics in Spain, even though it is certainly there are references to internationally valid for our case.

# 6.2. Future work

Next, we expose some proposals to expand this model in the future:

- Regarding the surface factor, a proposal is to apply more degrees of freedom in the matrix and make a deeper study, for parameters that are better aligned with what really affects them to the players. Also the classification criteria of the tracks in slow and fast could be modified by introducing any other.
- Update and further testing with forgetting factor by varying the upper and lower limits or the application window.
- Modify the model using other prior or varying the $\gamma$ function used.
- Entering any parameter of study, as temperature, or deepen existing studies that link ace, double fault or breakdown service with the winner of a match.
- Expanding the set of results to new seasons, even making a historial of seasons from 2000 to relate concepts and compare the classic tennis today.
- Extending the model to double tennis.
- Including combined or special beting into our field of betting.
- Improve the model especially for women's tennis. It deserves its own in-depth study that will take all the performance that we could not take with the model proposed in this project.

# Bibliografía


[1] Robert Gordon University Aberdeen ranking (2015), Gran Bretaña

[2] Dirección General de la Ordenación del Juego, Ministerio de Hacienda y Administraciones Públicas (2014). Memoria Anual 2014.

[3] Sports Betting: Past, Present and Future - Part 1 by Jeremy Martin.

[4] Gómez-Roso Jareño, J.M. (2014). *Origen de las apuestas deportivas*. Trabajo de fin de Grado. Castilla la Mancha: Facultad de Derecho y Ciencias Sociales de la Universidad de Castilla la Mancha

[5] El confidencial (17 de abril del 2008). Madrid abre su primera casa de apuestas al estilo inglés. Recuperado de: http://www.elconfidencial.com/

[6]Gómez Yáñez, J.A. (2014). *Anuario del juego en España 2013-2014*. Fundación Codere.

[7] Fontbona, M. (2008). *Historia del juego en España: de la Hispania Romana hasta nuestros días*. Flor del viento.

[8] iApuestas – portal de apuestas (2005). El apasionante mundo de las apuestas deportivas. Publicaciones iApuestas.

[9] Buchdahl, J. (2013). *How to Find a Black Cat in a Coal Cellar.* HighStakes

[10] Casas de apuestas. Web-site: http://www.lamejorcasadeapuestas.es/

[11] Tipos de apuestas deportivas. Web-site: http://www.webapuestas.com/guia-tutorial-apuestas/tipos-apuestas/index.html

[12] Rigueria, A. (2015). 6h 43’: El segundo partido individual más largo de la historia. Recuperado de: http://www.mundodeportivo.com/

[13] Kelly´s Criterion. Kelly, J. L. (1956). "A New Interpretation of Information Rate". Bell System Technical Journal 35 (4): 917926.doi:10.1002/j.1538-7305.1956.tb03809.x

[14] Ley 13/2011, de 27 de mayo, de regulación del juego.

[15] Checking whether a coin is fair. Web-site: https://en.wikipedia.org/wiki/Checking_whether_a_coin_is_fair#cite_ref-1

[16] Silva Ayçaguer L. C., Muñoz Villegas A. (2000). "*Debate sobre métodos frecuentistas vs bayesianos*" en XVII Reunión Científica SEE, Santiago de Compostela

[17] Mcgrayne, S.B. (2012). *La teoría que nunca murió*. Drakontos.

[18] David J.C. MacKay. Information Theory, Interference, and Learning Algorithms. University of Cambridge. Version 7.2 2005.

[19] Database. Web-site: http://www.tennis-data.co.uk/alldata.php

[20] Walpole R.E., Myers R.H., Myers S.L. (1999). *Probabilidad y estadística para ingenieros.* Pearson Educación.

[21] Canavos G.C. (1998). *Probabilidad y estadística. Aplicaciones y métodos*. Mcgraw Hill.

[22] Gómez-Villegas, M.A. (2005). *Inferencia estadística*. Ediciones Díaz de Santos.

[23] Cárdenas-Montes, M. (2006). "Sobreajuste – overfitting" en *Ciemat (Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas).*

[24] Fmincon. Web-site: http://es.mathworks.com/help/optim/ug/fmincon.html

[25] Rasmussen, C.E., Ghahramani, Z. Lecture 1 and 2. Probabilistic Regression. Machine Learning. Universidad de Cambridge.

[26] Hastie T., Tibshirani R., Friedman J. (2009). *The Elements of Statistical Learning: Data mining, Inference and prediction.* Springer, second edition.

[27] Schmidt López, E.D. (2007). "La superficie de las canchas de tenis, clave en el estilo de juego a desarrollar" en La Jornada Michocoan, México.

[28] Báez, R.A. (2010). *Superficies en las canchas de tenis*. Tennis TOP 10

[29] Log Loss. Web-site: https://www.kaggle.com/wiki/LogarithmicLoss

# Annex

# A. Summary

## 1. Introduction

The motivation of this project is to unite the passion for the sport, telecommunications and betting and develop a model to predict tennis match results with different knowledge of statistics and probability, and incorporate women's tennis to the analysis.

The main objective of this project is to develop a different model of predicting outcome in tennis matches to improve the previously existing, in addition to allow further analysis of the male and female tours. For this, we divide this document, including the problem statement, a technical and mathematical justification of the solution by providing alternatives, a wide presentation of results, a detailed budget and a schedule of work and finally a section containing all the conclusions.

## 2. Problem statement

Betting has a long tradition in our history, but it’s not until XIX and XX centuries when finally arriving in Europe, presenting the successful form of online betting and in the new century.

In Spain, the sports betting are in a sector that is growing exponentially. In our country in 2013/2014 they were played 29.026 millions of €, an amount that represents 2.85% of GDP. Among them, online betting is the fastest growing variant, accounting for 0.6% of GDP, about 5.600 millions of €. All this shows that, in addition to research to develop a favorable context, this is a sector that is not temporary and that takes hold.

In order to bet on a particular sport we must do so through a bookmaker, who will present different rates. The format we will use in this project will be European. Through an example, if you bet 2€ for a tennis player who has a quote of 2 € in a match, these fees would benefit mb • b = (q - 1) = 2 • (2-1) = 2 €.

Having explained the format share with betting we use, we must explain its operation. After the new law of 2011 bookmakers are only using the bet system against the user. This method is that sports bookmakers calculate, by its own criteria, the probabilities of different results, thus establishing quotes to which the user must bet to try to make profits. In this operation, the bookmaker always reserves a profit margin of between 4% and 15% by calculating the probabilities of an event with a total that totals more than 1 for, if the user gets hit far the business is still thriving.

Knowing the betting system and the format of quotes, then we have to decide what bookmarker will choose to bet our money. After analyzing Bet365, Betfair, Sportium, William Hill, Bwin and Luckia, we choose the first of all for their safety, fluidity and simplicity of online gambling, their industry experience and good shares.

Finally, we must decide what kind of betting we are going to play between all modes that offer us the different bookmarkers. Since our proposed model will get the odds that other players win games, using simple betting is so consistent, being able to squeeze the most of our model.

In addition, tennis has special characteristics that influence the different results and that we should value when developing our prediction model, such as being an individual sport, having different surfaces that may influence the results or the peculiarity that there are no ties.

The strategy that finally we will follow to bet our money will be using the Kelly criterion, an application designed to optimize profits and avoid bankruptcy in binary long-term games that calculates the percentage of our bankroll we should play based on the probability to give to a certain result:

$$f = \frac{p \cdot (b + 1) - 1}{b}$$

Finally, we discuss the regulatory framework of sports betting sector whose strength has never been the law and whose laws have been applied in recent years in spite of centuries of history. In Spain, the law regulating the sector is 13/2011, which allowed legalize and develop the bookmaker, profit for the state of a growth industry and ensure users a safe game.

# 3. Probabilistic model design and machine learning

To calculate the probability of a player winning a match use the so-called Bayesian approach, characteristic for give high importance to a priori information through a prior function assigned. Here we describe the process that we will follow, starting with making a list of players who will award a quality factor considering all possible elements.

We use Bayes' rule for calculating the likelihood and prior to help us avoid possible overfitting, and finally allows us to infer the quality parameters maximizing players back.

los jugadores maximizando la posterior.

- Bayes:

$$p(\Theta|\,Y) = \frac{p(Y\,|\Theta\,)\cdot p(\Theta\,)}{p(Y)} = p(Y\mid\,\Theta)\cdot p(\Theta\,) \quad (3.1)\ (3.2)$$

- Likelihood:

$$\mathrm{p(Y\,|\Theta)} = \prod_{\mathrm{n=1}}^{\mathrm{N}} \mathrm{p(Y_n \mid \Theta)} = \prod_{\mathrm{n=1}}^{\mathrm{N}} \mathrm{p}_{\mathrm{i(n),j(n)}}^{\mathrm{y}}\ (1 - \mathrm{p_{i(n),j(n)}})^{1-\mathrm{y}} \quad (3.5)$$

- Prior maximized:

$$\begin{aligned} -\log p(\Theta) = -\log\left(\prod_{n=1}^{M} p(\Theta)\right) &= -\sum_{n=1}^{M} \log \frac{\Theta(n)^{k-1} e^{-\frac{\Theta(n)}{s}}}{s^k \gamma(k)} \\ &= (k-1)\sum_{n=1}^{M} ln\big(\Theta(n)\big) \\ &\quad - \sum_{n=1}^{M} \frac{\Theta(n)}{s} - M\cdot k(s) - M\cdot ln\,((k-1)!) \end{aligned} \quad (3.6)$$

- Posterior maximized:

$$L = -\log p(\Theta|Y) = -\log p(Y|\Theta)p(\Theta) \tag{3.7}$$

Wherein the probability function models used in the likelihood are:

$$f(k;p) = \begin{cases} p & if\ k = 1 \\ 1 - p & if\ k = 0 \end{cases} \tag{3.3}$$

$$p_{ij} = \frac{1}{1 + e^{\theta_j - \theta_i}} \tag{3.4}$$

This process is the same that has been used in the previous model and that we will keep for the proposed new model. The calculations are performed in the Matlab tool and are based on the function fmincon to maximize the final value of Θ

Once determined the process of obtaining the probability and the quality factor, we present our proposed model. This model will be based, in addition to the machine learning, on the cross-validation method applied to various factors that will introduce to improve results. The way forward will be the same for all: we try to assess what matches provide us more information and how to select them.

The first factor to consider is the forgetting factor. By the same we give different weight to matches based on their proximity in time, considering the most important around the time played. It is a growing exponential:

$$\alpha = a_0 - (a_0 - b_0)e^{-\beta(t-1)}$$

In which we modify the parameter β to try to improve results.

On the other hand, we analyze the factor of surface. Depending on a player's game, faster, more physical and more aggressive, different surfaces they come better or worse. We will classify these surfaces in slow-fast with the order clay, hard outdoor, hard indoor and grass to establish a matrix of different weights to the surfaces depending on which one is being disputing a meeting to analyze the follows:

$$S = \begin{pmatrix} 1 & 0.6 & 0.4 & 0.2 \\ 0.6 & 1 & 0.2 & 0.4 \\ 0.4 & 0.2 & 1 & 0.6 \\ 0.2 & 0.4 & 0.6 & 1 \end{pmatrix}$$

This matrix S will be modified to perform different tests.

Finally we introduce a factor of direct confrontation to try to hit the games in which a player who has taken the measure to another almost always wins, regardless of their qualities. To do this we will set different weights between these clashes and other meetings that may be two, three or ten times higher.

# 4. Results and evaluation

The cost function we use to assess the results will be:

$$LogLoss = -\frac{1}{n}\sum_{i=1}^{n} y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)$$

We make an initial test with the previous model on different seasons of male and female tennis, for a later check of the improvement of our model. The results are:

| Season | Men's tennis | Women's tennis |
|---|---|---|
| 2013 | 0,6442 | 0,6423 |
| 2012 | 0,6291 | 0,6562 |
| 2011 | 0,6416 | 0,6397 |
| 2011-2012 | 0,6098 | 0,6344 |
| 2012-2013 | 0,6184 | 0,6342 |
| 2011-2013 | 0,6079 | 0,6275 |

Table 4.1: LogLoss previous model

We decided to apply our model for the 2012-2013 season for his margin improvement and his latest character, applying the training and validation processes of 2012 and the test process on 2013. The results for both circuits are as follows.

- **Forgetting factor:**

| Model | 2012 Male | 2012 Female |
|---|---|---|
| Previous model | 0,6291 | 0,6562 |
| Proposed model (with forgetting factor) | 0,6034 | 0,62 |

Table 4.18: Comparison evolution LogLoss with forgetting factor.

Where we can see that the improvement is remarkable, being higher in women's tennis, which means that the players are more irregular and more dependent on their spells.

- **Surface factor:**

| Model | 2012 Male | 2012 Female |
|---|---|---|
| Previous model | 0,6291 | 0,6562 |
| Proposed model (with surface factor) | 0,6223 | 0,6525 |

Table 4.19: Comparison evolution LogLoss with surface factor

We note that this factor does not have a very big effect, but still tells us that men's tennis is suffering more than the women's tennis with the changes of surfaces, where women are best suited.

- **Clash factor:**

| Matches | Probability Previous model | Probability proposed model (with clash factor) |
|---|---|---|
| Djokovic - Nadal | 0,4025 | 0,5012 |
| Murray - Federer | 0,2685 | 0,3593 |
| Kerber - Radwanska | 0,3983 | 0,5017 |
| Li - Azarenka | 0,2203 | 0,3178 |

Table 4.20: Comparison evolution probability for clash factor

We can see that the factor of confrontation gives us excellent results in both circuits.

- **Difference between players:**

| Model | Male | Female |
|---|---|---|
| 2012-2013 Previous model | 2,57 | 2,76 |
| 2012-2013 Proposed model (with all factors) | 2,53 | 2,47 |

Table 4.21: Comparison evolution differences between players

With this factor we can confirm that the model affects unequally to women's tennis, confirming its irregularity, while men's tennis remains stable.

- **Función de coste: LogLoss:**

| Model | Male | Female |
|---|---|---|
| 2013 Previous model | 0,6442 | 0,6423 |
| 2012-2013 Previous model | 0,6184 | 0,6342 |
| 2012-2013 Proposed model (with all factors) | 0,6013 | 0,6132 |

Tabla 4.22: Comparison evolution LogLoss

Where a great result in men's tennis and a good result in women's tennis shown, if it is true that it does not meet the expectations we expected.

After analyzing the obtained model validation process thoroughly, we get the conclusion that, although the bookies have very good numbers and very difficult to overcome, likely results in 0.5-0.6 our model is better. Therefore we decided to finally bet on the results of this chance Gaza in 2013 and check whether the model works and gets benefits:

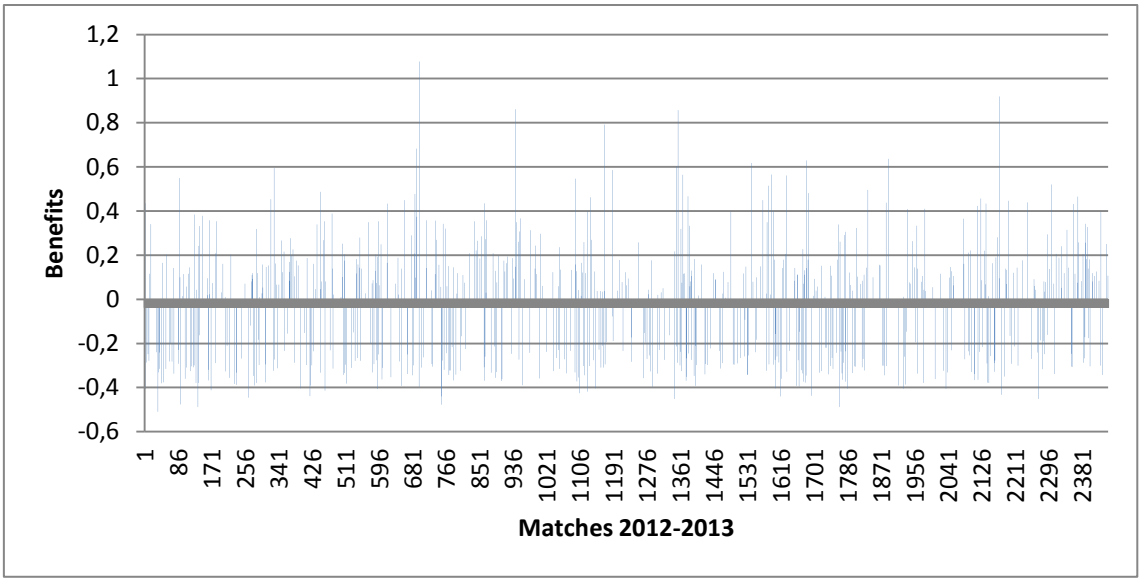


Figure 4.12: Money win / loss with threshold 2012-2013. Men's tennis.

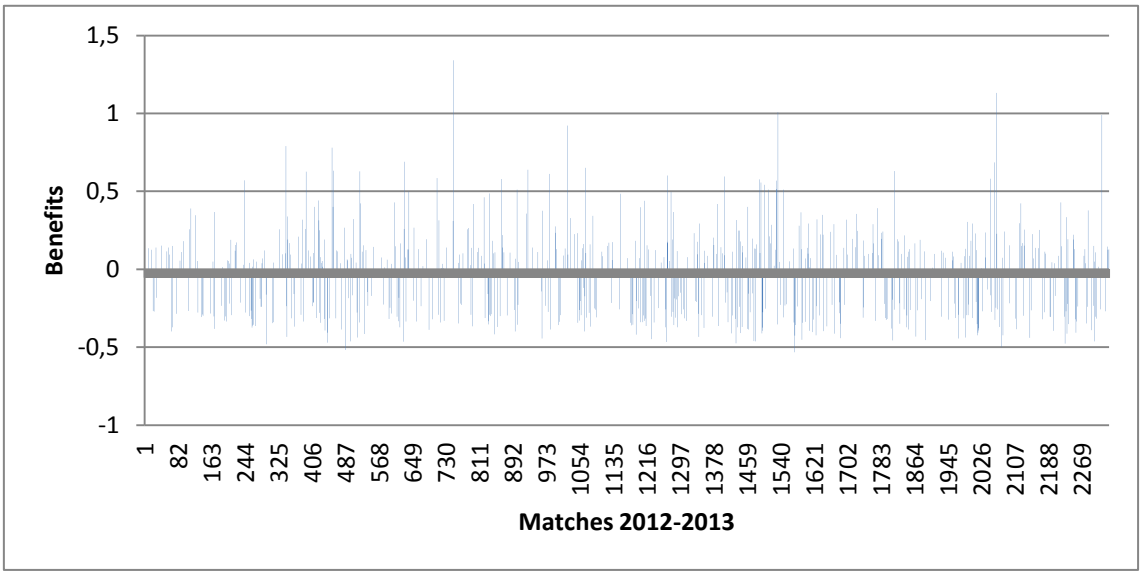


Figure 4.16: Money win / loss with threshold 2012-2013. Women's Tennis.

Where we observed that in men's tennis left the balance of benefit in almost 0, getting benefit of 1.2 € cents per game in the last 3 months and achieving excellent results, while in women's tennis got minimize losses but remain 0.5 cents in € per game.

# 5. Plan and budget.

Then we add to the budget and the Gantt chart obtained in the project:

## Presupuesto

| | |
|---|---|
| "Dirección" | Calle Río Duero |
| "Ciudad" | Leganés |
| "Provincia" | Madrid |
| "Telefóno" | 625959100 |
| "Codigo Post | 28913 |
| "DNI" | 53457478G |

| | | | |
|---|---|---|---|
| **Fecha Solicitud** | 15-12-14 | **Método de pago** | Transferencia bancaria |
| **Solicitado por:** | Cliente | | |

| Producto | Cantidad | Descripción | Precio / mes | Precio |
|---|---|---|---|---|
| 1 | | PROYECTO APUESTAS DEPORTIVAS | 2.144,00 € | 19.296,00 € |
| 2 | | | | |
| 3 | | | | |
| 4 | | | | |
| 5 | | | | |
| 6 | | | | |
| 7 | | | | |
| 8 | | | | |
| 9 | | | | |
| 10 | | | | |

| | | | |
|---|---|---|---|
| ***Si tiene alguna duda sobre este presupuesto no dude en comunicarse con nosotros*** | | **Subtotal** | 19.296,00 € |
| | **21,00%** | IVA | 4.052,16 € |
| | | Costes de Envio | |
| | | Seguro | |
| | | **Total** | **23.348,16 €** |

Figure 5.1: Budget

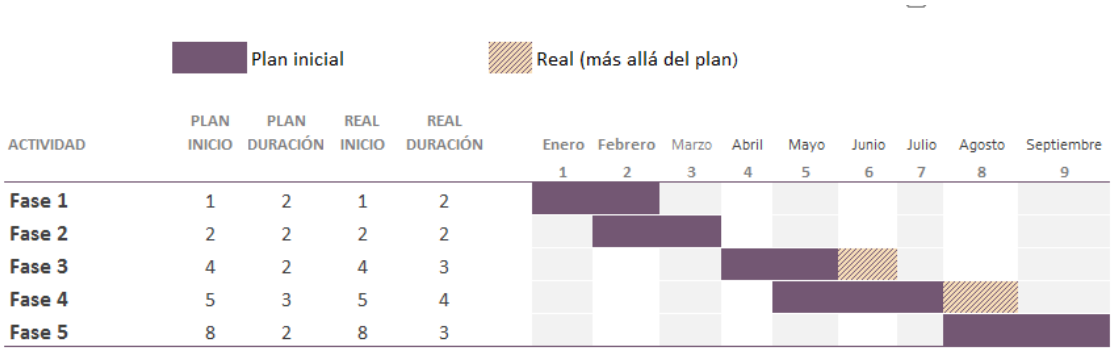


Figure 5.2: Gantt Chart

# 6. Conclusions and future work

We conclude that the results are significantly better in men's tennis, getting finally monetize. The model of women's tennis has become scarce us because the players will probably affect more factors that have analyzed with this model, and can be a line of future research to get more out. The aims to analyze the male and female circuits in depth to draw conclusions and to obtain acceptable results have been met, showing a more irregular and uneven in general male women's tour circuit.