Universidad Carlos III de Madrid
Escuela Politécnica Superior
Ingeniería en Informática

Autor: Víctor Galán Cortina
Tutora: Elena Castro Galán

Octubre 2015

PROYECTO FIN DE CARRERA

APLICACIÓN DE LA METODOLOGÍA CRISP-DM A UN PROYECTO DE MINERÍA DE DATOS EN EL ENTORNO UNIVERSITARIO

Agradecimientos

Me gustaría dedicar unas líneas de esta memoria para agradecer a todas aquellas personas que me han ayudado en alguna manera a llegar hasta este punto en mi carrera académica.

En primer lugar gracias a mi familia, en especial a mis padres, quienes han dado todo sin dudarlo para que crezca tanto como persona como en mi formación y educación, y que me han apoyado en todo momento, especialmente en los malos que es cuando más los he necesitado, sin ellos nada habría sido posible y siempre estaré en deuda con ellos. También a mis hermanas con las que he compartido tantos años de vida y que también han estado a mi lado en todo momento.

A todos y cada uno de los profesores que he tenido a lo largo de mi vida académica, porque he aprendido algo nuevo de cada uno de ellos. Me gustaría poner especial énfasis en mi tutora de este proyecto, Elena, que tanta paciencia ha tenido conmigo para ayudarme a llevarlo a cabo, que me ha mostrado siempre una disponibilidad absoluta y que ha sacrificado gran parte de su tiempo de manera altruista.

A mis amigos y compañeros de la universidad (algunos son ambas cosas para mí), porque he aprendido que en la facultad no sólo se adquieren conocimientos de cálculo, estadística, redes de ordenadores, programación, etc., sino que además uno puede aprender cosas que nunca se habría imaginado, como jugar al mus. Bromas aparte, de ellos he aprendido muchísimo en cuanto a valores humanos, y han sido un apoyo fundamental a lo largo de la carrera, no me imagino haber llegado hasta donde estoy hoy sin ellos. Gracias por las risas, las penas, alegrías, enfados, viajes y muchísimas memorias que quedarán siempre en el recuerdo como una de las mejores épocas de mi vida.

A mis amigos de fuera de la universidad que también han puesto de su parte para ayudarme a crecer como persona y me han ayudado a superar los baches y celebrar los éxitos.

Por último a la Universidad Carlos III de Madrid, en concreto al campus de Colmenarejo que ha sido como un segundo hogar durante estos años y que me ha dado la oportunidad de realizar mis estudios y me ha visto crecer como persona (y físicamente).

A todos ellos, gracias.

Índice

1. Introducción

En el mundo actual en que vivimos donde cada vez es más importante tener todo informatizado y cuantificado en las bases de datos de cada empresa u organización, surge la necesidad de encontrar alguna manera de sacar conclusiones a partir de estos datos, ya que de por sí solos los datos nada más que serían registros sin significado y que no darían ningún tipo de información valiosa que se pudiera explotar y sacar provecho ellos. Si bien es cierto que mediante consultas simples sobre estos datos se pueden obtener algunos resultados, a medida que crece la complejidad de la base de datos y el número de registros, estos resultados son cada vez más difícilmente interpretables para la persona u organización que desea utilizarlos con algún fin. De esta necesidad nace la minería de datos que es la ciencia que estudia patrones en grandes bases de datos y emplea técnicas de la inteligencia artificial, la estadística o el aprendizaje automático para extraer dicha información y traducirla a unos resultados interpretables por la persona o entidad que desea sacar partido a estos datos.

Así pues para comenzar el proceso de minería de datos es importante partir de una base de datos o data warehouse (almacén de datos) que contenga la información que se quiere analizar y que ésta información esté correctamente estructurada. La minería de datos trata de sacar toda la información posible de los almacenes de datos, no se conforma sólo con la visualización de estos datos como podría pasar con las consultas simples, si no que trata de obtener resultados en cuanto a la relación que existe entre los mismos y como podrían dar beneficios de algún modo al negocio.

Existen diversas técnicas y metodologías de minería de datos que se pueden utilizar y que pueden ser más o menos adecuadas para cada caso en concreto, pero en el presente proyecto se ha elegido la metodología CRISP-DM de la cual hablaremos detalladamente más adelante y trataremos de justificar el porqué de esta elección para el caso práctico que se nos planteó: la explotación de los datos contenidos en la base de datos de la Universidad Carlos III de Madrid.

Esta memoria está dividida en dos partes claramente diferenciadas. En la primera parte (la más teórica), se pretende poner en contexto al lector y darle una serie de conocimientos básicos acerca de la minería de datos y los distintos métodos que existen

así como las diferentes aplicaciones informáticas que existen actualmente en el mercado para facilitar el proceso de extracción de la información.

En la segunda parte es donde se aplica en la práctica las distintas etapas de la metodología escogida sobre los datos de los que disponemos para finalmente hacer una valoración lo más objetiva posible acerca de la viabilidad de este proyecto.

2. Objetivos

El objetivo de este proyecto está claramente delimitado: aplicar estrictamente cada una de las distintas etapas de la metodología CRISP-DM sobre los datos académicos almacenados por la universidad en sus sistemas informáticos. De esta forma se pretende sacar conclusiones que ayuden a mejorar los servicios que ofrece la universidad a sus alumnos. También se quiere demostrar que la metodología CRISP-DM es una metodología que funciona y que además es sencilla de usar, ya que solamente hay que seguir una serie de fases que están claramente delimitadas y está pensada para que cualquier persona con conocimientos de bases de datos y estadística pueda utilizarla.

El objetivo final del proyecto es por tanto aplicar la metodología CRISP-DM al ámbito académico de la universidad, mientras que el objetivo en sí de la metodología es el de sacar conclusiones y hacer predicciones lo más fiables posible partiendo de una serie de datos. Por tanto distinguimos entre los objetivos del presente proyecto que acabamos de mencionar, y los objetivos de la minería de datos que son explicados más adelante en el apartado de “objetivos del negocio” y “objetivos de la minería de datos”. Por este motivo, el hecho de no alcanzar los objetivos del negocio no implican necesariamente que no se cumplan los objetivos del proyecto, ya que en cualquier caso el objetivo quedará cubierto siempre y cuando hayamos conseguido aplicar por completo la metodología a nuestra problemática.

Parte I:

Fundamentos de la Minería de Datos

Esta primera parte del proyecto repasa la teoría detrás de la minería de datos, repasando algunos conceptos básicos y conocimientos previos que son necesarios para la práctica de la misma. Además, se introduce la metodología de minería de datos que se va a aplicar en la segunda parte de este proyecto, CRISP-DM, todo desde el punto de vista teórico listando y resumiendo cada una de sus fases. Por último se repasarán las distintas herramientas que hay disponibles y que se pueden emplear para llevar a cabo esta tarea.

1. Estado del Arte

1.1. Fases del Proceso de Extracción de Conocimiento

El proceso de extracción de conocimiento, en inglés KDD (*Knowledge Discovery from Databases*), es como se denomina al proceso que se encarga de sacar conclusiones o información "limpia" a partir de unos datos que generalmente están organizados en una base de datos. Este proceso es iterativo debido a que la salida de algunas de sus fases puede requerir que se vuelva a una fase anterior y también a que en ocasiones es necesario realizar varias iteraciones para poder extraer conocimiento de los datos.

El proceso de KDD consta de cinco fases diferenciadas [Hernández, Ramírez y Ferri, 2004], en este proyecto nos centraremos principalmente en una de ellas, la de minería de datos, pero también hablaremos de las demás ya que no se puede hablar de minería de datos sin tener unos conocimientos previos de los datos en sí, de cómo se seleccionan, limpian, transforman y almacenan estos datos, y por supuesto una vez concluida la minería será necesario saber cómo evaluar e interpretar las conclusiones sacadas del estudio. En la figura 1 [Hernández, Ramírez y Ferri, 2004], podemos observar el proceso de KDD global con cada una de sus fases.

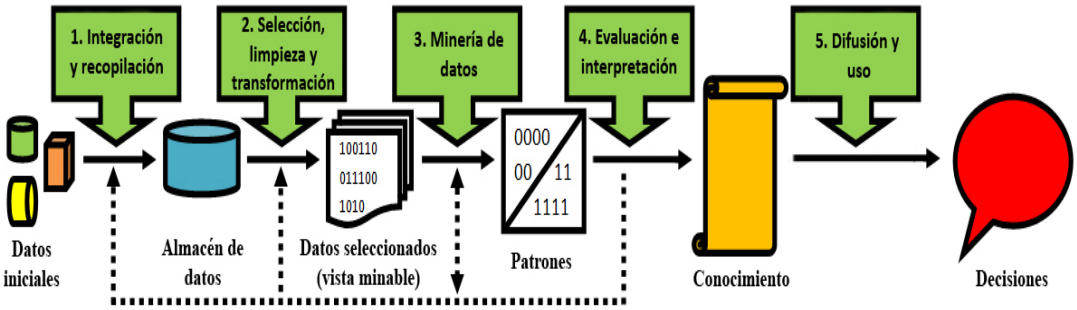


Figura 1 – Proceso KDD

Ahora veremos cada una de estas fases para tener una idea global del proceso de extracción de conocimiento:

1. **Fase de integración y recopilación de datos.**

   En esta primera fase del proceso de KDD se trata de decidir de dónde se van a sacar los datos que se utilizarán más adelante, es decir que fuentes de información nos van a resultar útiles. Después se transforman todos los datos a un formato en común, ya que como hemos dicho, estos pueden provenir de fuentes heterogéneas, generalmente esto se consigue usando un almacén de datos.

2. **Fase de selección, limpieza y transformación.**

   Los datos recopilados en el almacén pueden contener errores en sus valores, o incluso puede que a algunos de estos les falte algún valor, que estos sean erróneos. En esta fase se trata de corregir o incluso eliminar estos datos, y se decide qué hacer con aquellos datos que estén incompletos. También se realiza una selección de aquellos datos que son relevantes para el proceso de extracción de conocimiento que se desea realizar.

3. **Fase de minería de datos.**

   En esta fase, que es la fase principal que se trata en este proyecto, se debe decidir cuál es la tarea (agrupar, clasificar, etc.) que se va a realizar y se elige el método que se va a emplear para ello.

4. **Fase de evaluación e interpretación.**

   La fase de minería de datos debe dar unos resultados, por ejemplo unos patrones observados en los datos. En la fase de evaluación e interpretación se debe

evaluar e interpretar estos patrones, con el fin de poder entender el resultado obtenido.

5. **Fase de difusión y uso.**

   La última fase del proceso de KDD tiene como objetivo utilizar el nuevo conocimiento adquirido y hacer que dicho conocimiento sea empleado por todos los usuarios posibles.

1.2. El Almacenamiento de los Datos

Antes de comenzar a hablar del proyecto de minería de datos en sí es preciso tener una serie de conocimientos acerca de la forma en la que se almacenan los datos sobre los que se van a emplear las distintas técnicas de minería de datos. Esto no es un proceso tan trivial como podría parecer a simple vista, ya que en muchos casos la diversidad y el tamaño de las fuentes de los datos pueden convertir la tarea de la recopilación de dichos datos en algo muy complicado y tedioso.

Es realmente importante para que se pueda extraer toda la información posible que los datos de los cuales vamos a intentar sacar conocimiento estén lo mejor organizados posible. Pero, ¿a que nos referimos cuando se dice "lo mejor organizados posible"? esto es un concepto muy subjetivo, ya que depende de la información que se quiera obtener de los datos, y la técnica de minería de datos que se vaya a emplear para ello (veremos más adelante que algunas técnicas se ajustan mejor a un determinado tipo de datos que a otros), puede resultar más útil tener los datos almacenados de una manera o de otra. Existen muchísimas maneras de guardar los datos que escapan al propósito de este proyecto, pero podemos nombrar los más comunes:

- Los datos se pueden guardar en un fichero, esta es la forma más simple de almacenar datos, pero también es muy poco útil ya que si tenemos una cantidad de datos muy elevada, el proceso de extracción de conocimiento a partir de éste fichero se hace extremadamente complicado.
- Otra forma de almacenar datos y probablemente la más utilizada hoy en día son las bases de datos relacionales. En una base de datos relacional, los datos se encuentran agrupados en tablas, y dichas tablas están relacionadas con otras tablas a través de algún atributo de los datos. Una de las ventajas de tener los datos organizados en una base de datos relacional es que se puede tener un esquema asociado a la base de datos que permite tener una idea de cómo están estructurados los datos de una forma mucho más simple e intuitiva, es por ello que se dice que en este caso los datos están estructurados. Por si esto fuera poco, también existen lenguajes de consulta que están diseñados para obtener información de una base de datos relacional, como SQL, y que por tanto facilitan el proceso a quien desea realizar la minería de datos.

- Existe diversidad de bases de datos además de las relacionales, como las bases de datos espaciales, temporales, documentales, multimedia y objeto-relacionales.

No vamos a entrar a profundizar más acerca de las distintas formas de almacenar datos en este proyecto ya que a pesar de que es importante para el proceso de minería de datos la forma en la que se guardan y se presentan los datos, no es el objetivo principal, y además se parte de unos datos preestablecidos. Estos datos se encuentran almacenados en lo que se denomina un almacén de datos (*data warehouse).*

1.3. Los Almacenes de Datos

Un almacén de datos o *data warehouse* es "*un repositorio de información coleccionada de varias fuentes, almacenada bajo un esquema unificado que normalmente reside en un único emplazamiento"*. [Hernández, Ramírez y Ferri, 2004]. Los almacenes de datos no son imprescindibles para extraer conocimiento a partir de unos datos (también se podría hacer minería de datos sobre un simple fichero), sin embargo, las ventajas de tener los datos organizados en un almacén de datos son muchas a medio y largo plazo cuando tenemos un volumen de datos muy grande (o bien los datos van aumentando con el tiempo), o cuando los datos provienen de diferentes fuentes heterogéneas.

Generalmente, la información que se quiere investigar sobre un cierto dominio de una organización, se suele encontrar en bases de datos distintas (bases de datos distribuidas) y otras fuentes diversas que pueden ser tanto internas como externas. Lo que hace un almacén de datos es unificar en un solo lugar toda esta información que se considera importante para el estudio que se quiere realizar. Otra ventaja de tener la información organizada en un almacén de datos es que no se trabaja sobre los datos originales de las bases de datos de la organización, ya que el almacén de datos está separado de las bases de datos operacionales. Esta separación facilita el análisis de los datos en tiempo real, a este tipo de procesos se le denominan OLAP (On-Line Analytical Processing), y además de esta forma no se entorpecen los procesos OLTP (On-Line Transactional Processing) de las bases de datos originales.

En la figura 2 podemos ver un esquema de cómo se realiza un almacén de datos que integra datos heterogéneos y de distintas fuentes.

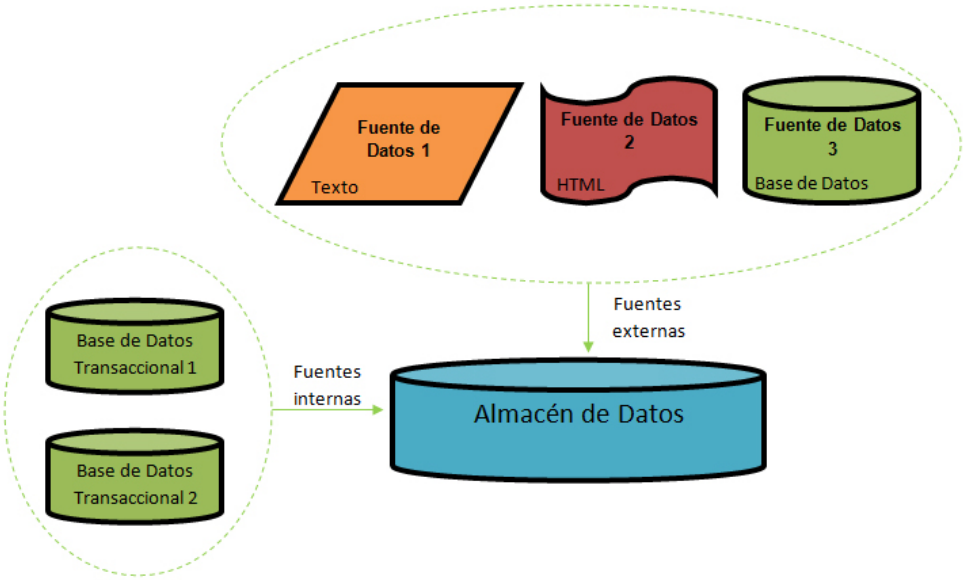


Figura 2 – Almacén de datos

En este proyecto los datos a explorar mediante la minería de datos provienen de un *Data Mart*, que se podría definir como una versión reducida de un almacén de datos. Generalmente los *Data Mart* contienen información específica de algún departamento concreto de la organización, como puede ser el departamento de marketing o el de recursos humanos. Idealmente estos *Data Mart* deberían ser un subconjunto del almacén de datos, a fin de mantener la consistencia de los datos corporativos y la seguridad e integridad de la información que se está usando. Debido a que el coste de desarrollo y de implantación de un almacén de datos es mucho mayor, hoy en día cada vez más organizaciones optan por un *Data Mart* para almacenar sus datos.

1.4. La Minería de Datos

Una vez que se han recopilado los datos necesarios y éstos están bien organizados y limpios, es la hora de aplicar sobre ellos el proceso de minería de datos cuyo objetivo es descubrir patrones que deben ser válidos, novedosos y por supuesto, comprensibles. Para ello existen diversas tareas de minería de datos y métodos (o técnicas) que permiten resolver dichas tareas.

- **Tareas**

Una tarea se puede definir como un tipo de problema a ser resuelto por un algoritmo de minería de datos. Por lo tanto esto implica que cada tarea tiene sus propios requisitos y que la información que se obtiene empleando una tarea en concreto puede ser muy distinta a la obtenida si se emplea otra tarea diferente.

Podemos dividir las tareas en dos tipos, predictivas o descriptivas. En las predictivas el objetivo es estimar valores futuros o desconocidos de algunas variables de interés a partir de otras variables independientes (variables predictivas). En el caso de las tareas descriptivas el objetivo es identificar patrones en los datos que los explican o resumen.

A continuación vamos a ver las tareas más importantes de la minería de datos para cada uno de los dos tipos anteriores:

- **Predictivas**
  - Clasificación o discriminación (en estadística)

    La clasificación asume que hay un conjunto de objetos caracterizados por algún atributo o rasgo que pertenece a distintas clases. La etiqueta de clase es un valor discreto y es conocido para cada objeto. El objetivo de esta tarea es asignar la etiqueta de clase correcta a objetos nuevos y sin etiqueta dados los valores de sus atributos. La clasificación es una las tareas más comunes en la minería de datos.

    Un ejemplo de clasificación puede ser clasificar un mensaje de correo electrónico como *spam* o no.
  - Clasificación suave

Esta tarea es igual que la clasificación pero se introduce una función que estima el grado de certeza de la predicción.

Un ejemplo sería clasificar un mensaje de correo electrónico como *spam* o no, proporcionando además la certeza de la clasificación.

- Estimación de probabilidad de clasificación

  Es una generalización de la clasificación suave. El problema a resolver es el mismo que para la clasificación y clasificación suave. La diferencia está en que en esta tarea el resultado es un conjunto de probabilidades de que el objeto pertenezca a una clase u otra.

  Por ejemplo, si se quisiera clasificar entre varios medicamentos cuál es el mejor para una determinada patología, esta tarea proporcionaría la probabilidad de que sea cada uno de los medicamentos escogidos.

- Categorización

  En esta tarea a diferencia de las clasificaciones donde a cada objeto le corresponde una y sólo una clase, a un objeto le puede corresponder n clases.

  Por ejemplo, dado un conjunto de documentos, asignar categorías de los temas que trata cada documento.

- Regresión

  Esta tarea es muy parecida a la clasificación ya que a cada elemento se le asigna únicamente un valor de salida, con la diferencia de que este valor de salida es un valor numérico, es decir, puede ser un valor entero o real.

  Un ejemplo muy sencillo sería estimar las ventas de un determinado producto para un determinado año.

- **Descriptivas**

  - Agrupamiento (clustering)

    El objetivo de esta tarea es agrupar los elementos de un conjunto de objetos, de tal manera que los elementos que formen parte de un mismo grupo tengan características similares. La diferencia con respecto a la clasificación es que en el agrupamiento son los grupos y la pertenencia a dichos grupos lo que se quiere determinar, y además a priori no se sabe cuántos grupos hay ni cómo son.

Un ejemplo podría ser agrupar clientes en grupos diferenciados para luego estudiar que grupos se comportan mejor ante determinados productos.

- Correlaciones y factorizaciones

  El objetivo es ver si dos o más atributos numéricos están correlacionados linealmente o relacionados de algún modo.

- Reglas de asociación

  Esta tarea es parecida a la de las correlaciones y factorizaciones pero en este caso se utilizan atributos nominales en lugar de numéricos.

  Se utilizan para descubrir hechos que ocurren en común dentro de un determinado conjunto de datos.

  Un ejemplo muy conocido sobre las reglas de asociación es el de la "cerveza y los pañales" que está en el comportamiento de los compradores en el supermercado. Se descubrió que muchos hombres compraban pañales por encargo de sus esposas, por lo que una cadena de supermercados decidió colocar la cerveza junto a los pañales y el resultado fue un aumento en la venta de cerveza.

- **Métodos**

  Para poder resolver cualquiera de las tareas anteriores se necesitan unas técnicas, algoritmos o métodos. Es importante saber que una tarea se puede resolver con diferentes métodos y no sólo uno, y que un método puede servir para resolver más de una tarea. A continuación veremos los distintos tipos de métodos que existen para poder resolver las tareas anteriormente mencionadas:

  - **Técnicas algebraicas y estadísticas**

    El objetivo de estas técnicas es expresar modelos y patrones mediantes fórmulas algebraicas, funciones lineales, funciones no lineales, distribuciones o valores agregados estadísticos tales como medias, varianzas, correlaciones, etc. Generalmente estas técnicas obtienen un patrón a partir de un modelo predeterminado del cual se estiman unos coeficientes o parámetros. Ejemplos de algoritmos dentro de estas técnicas son: regresión lineal (global o local), regresión logarítmica y regresión logística.

  - **Técnicas bayesianas**

El objetivo es estimar la probabilidad de pertenencia a una clase o grupo mediante la estimación de las probabilidades condicionales inversas o a priori, para lo cual utilizan el teorema de Bayes. Ejemplos de algoritmos dentro de este grupo son: el clasificador bayesiano naive, los métodos basados en máxima verisimilitud y el algoritmo EM.

- **Técnicas basadas en conteos de frecuencias y tablas de contingencia**

  Se trata de contar la frecuencia con la que dos o más sucesos se dan conjuntamente. Cuando el conjunto de sucesos posibles es muy grande, sólo se cuentan aquellos casos en los que las frecuencias conjuntas superan un determinado umbral.

- **Técnicas basadas en árboles de decisión y sistemas de aprendizaje de reglas**

  Estas técnicas además de sus representaciones en forma de reglas, también se basan en algoritmos del tipo "divide y vencerás" (por ejemplo el ID3/C4.5 o el CART) y "separa y vencerás" (por ejemplo el CN2).

- **Técnicas relacionales, declarativas y estructurales**

  Estas técnicas representan el modelo mediante lenguajes declarativos, como los lenguajes lógicos, funcionales o lógico-funcionales.

- **Técnicas basadas en redes neuronales artificiales**

  Son técnicas que aprenden un modelo mediante el entrenamiento de los pesos que conectan un conjunto de nodos o neuronas. La topología de la red y los pesos de las conexiones determinan el patrón aprendido. Existen multitud de variantes de organización con muchos algoritmos diferentes para cada organización, pero el más conocido probablemente sea el de retropropagación.

- **Técnicas basadas en núcleo y máquinas de soporte vectorial**

  Son técnicas que intentan maximizar el margen entre los grupos o las clases formadas. Para ello se basan en unas transformaciones que pueden aumentar la dimensionalidad, que se llaman núcleos o *kernels*.

- **Técnicas estocásticas y difusas**

  En este grupo se incluyen la mayoría de las técnicas que junto a las redes neuronales forman lo que se denomina computación flexible (*soft computing*). Son técnicas en las que o bien los componentes aleatorios son

fundamentales, como el *simulated annealing*, los métodos evolutivos y genéticos, o bien al utilizar funciones de pertenencia difusas (*fuzzy*).

- **Técnicas basadas en casos, en densidad o distancia**

  Son métodos que se basan en distancias al resto de elementos, ya sea directamente, como los vecinos más próximos, o de una manera más sofisticada, mediante la estimación de funciones de densidad. Algunos algoritmos muy conocidos son los jerárquicos, como Two-step o COBWEB, y los no jerárquicos como K medias.

1.5. La Metodología CRISP-DM

CRISP-DM (*Cross Industry Standard Process for Data Mining*), es un modelo de proceso de minería de datos que describe una manera en la que los expertos en esta materia abordan el problema.

Para implementar una tecnología en un negocio es necesaria una metodología. Estos métodos suelen venir de las experiencias propias y también de los procedimientos estándar más conocidos. En el caso de los proyectos de implementación de minería de datos una de las metodologías que ha tenido más apoyo de las empresas privadas y organismos públicos es CRISP-DM, como se puede observar en la siguiente gráfica (figura 3), publicada en kdnuggets.com, y que representa el grado de utilización de las principales guías de desarrollo de proyectos de minería de datos según las encuestas realizadas. Como se puede observar CRISP-DM ha experimentado un ligero descenso en los últimos años, pero sigue siendo la más empleada de las distintas metodologías.

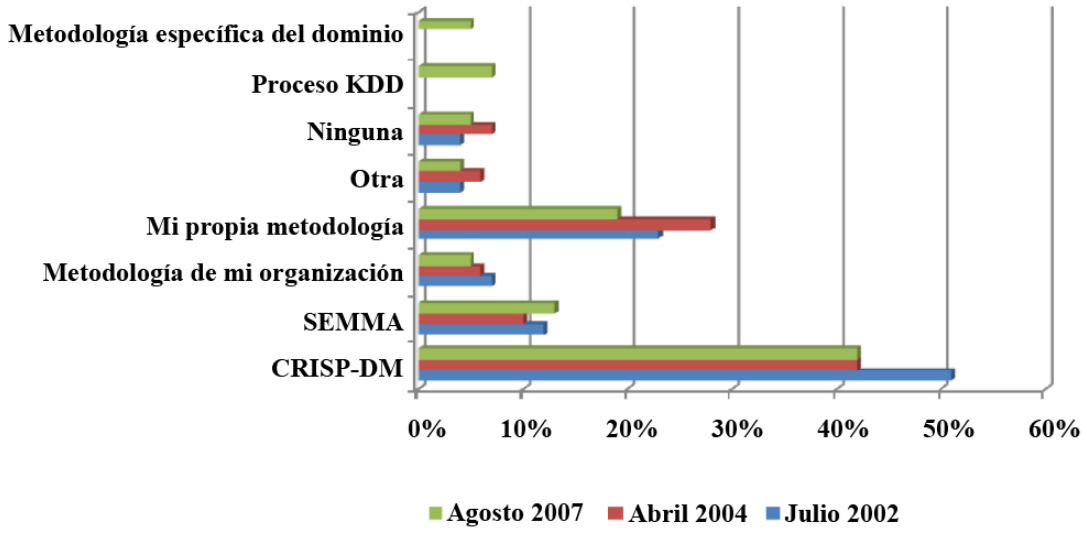


Figura 3 – Grado de utilización de las distintas metodologías de minería de datos

CRISP-DM incluye un modelo y una guía, estructurados en seis fases, algunas de las cuales son bidireccionales, es decir que de una fase en concreto se puede volver a una fase anterior para poder revisarla, por lo que la sucesión de fases no tiene porqué ser ordenada desde la primera hasta la última. En la figura 4 se puede observar las fases en las que se divide CRISP-DM y las posibles secuencias a seguir entre ellas.

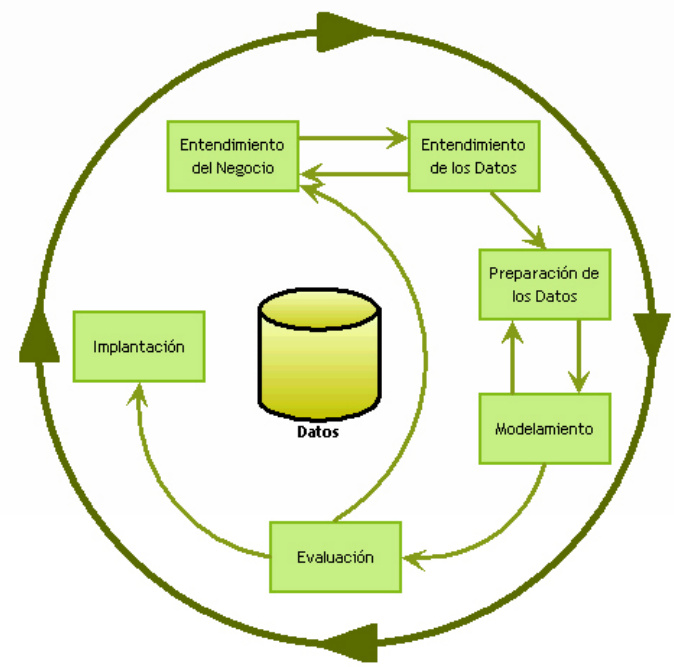


Figura 4 – Secuencia del proceso CRISP-DM

A continuación se explica cada una de estas fases [Rodríguez, 2010]:

1. **Comprensión del negocio.**

   Esta primera fase es probablemente la más importante y aglutina las tareas de comprensión de los objetivos y requisitos del proyecto desde una perspectiva de negocio, con el fin de convertirlos en objetivos técnicos y en un plan de proyecto. Sin lograr comprender dichos objetivos, ningún algoritmo por muy sofisticado que sea, permitirá obtener resultados fiables. Para obtener el mejor provecho de la minería de datos, es necesario entender de la manera más completa el problema que se desea resolver, esto permitirá recolectar los datos correctos e interpretar correctamente los resultados. En esta fase, es muy importante la capacidad de poder convertir el conocimiento adquirido del negocio en un problema de minería de datos y en un plan preliminar cuya meta sea el alcanzar los objetivos del negocio. A continuación vemos una descripción de cada una de las principales tareas que componen esta fase, figura 5 [CRISP-DM, 2000].

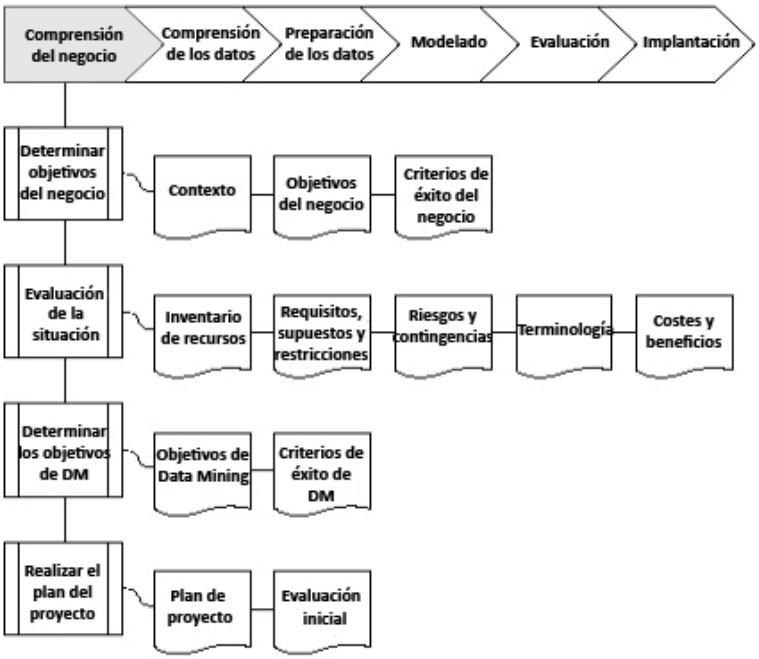


Figura 5 – Fase de comprensión del negocio

- **Determinar los objetivos del negocio.**

  Esta es la primera tarea a desarrollar y tiene como metas determinar cuál es el problema que se desea resolver, por qué la necesidad de utilizar la minería de datos y definir los criterios de éxito. Los problemas pueden ser diversos, como por ejemplo, detectar fraude en el uso de tarjetas de crédito, detección de intentos de ingreso indebido a un sistema, asegurar el éxito de una determinada campaña publicitaria, etc. En cuanto a los criterios de éxito, estos pueden ser de tipo cualitativo, en cuyo caso un experto en el área de dominio califica el resultado del proceso de minería de datos, o bien de tipo cuantitativo, por ejemplo, el número de detecciones de fraude o la respuesta de clientes ante una campaña publicitaria.

- **Evaluación de la situación.**

  En esta tarea se debe calificar el estado de la situación antes de iniciar el proceso de minería de datos, considerando aspectos tales como: ¿cuál es el conocimiento previo disponible acerca del problema?, ¿se cuenta con la cantidad de datos requerida para resolver el problema?, ¿cuál es la relación coste beneficio de la aplicación de minería de datos?, etc. En esta fase se definen los requisitos del problema, tanto en términos de negocio como en términos de minería de datos.

- **Determinar los objetivos de la minería de datos.**

  Esta tarea tiene como objetivo representar los objetivos del negocio en términos de las metas del proyecto de minería de datos, como por ejemplo, si el objetivo del negocio es el desarrollo de una campaña publicitaria para incrementar la asignación de créditos hipotecarios, la meta de minería de datos será por ejemplo determinar el perfil de los clientes respecto de su capacidad de endeudamiento.

- **Realizar el plan del proyecto.**

  Esta última tarea de la primera fase de CRISP-DM tiene como meta desarrollar un plan para el proyecto, que describa los pasos a seguir y las técnicas a emplear en cada uno de ellos.

2. **Comprensión de los datos.**

   Esta segunda fase comprende la recolección inicial de los datos con el objetivo de establecer un primer contacto con el problema, familiarizarse con ellos, identificar su calidad y establecer las relaciones más evidentes que permitan definir las primeras hipótesis. Esta fase junto a las dos siguientes fases son las que demandan el mayor esfuerzo y tiempo en un proyecto de minería de datos. Por lo general si la organización cuenta con una base de datos corporativa, es deseable crear una nueva base de datos específica para el proyecto de DM (*Data Mining*), ya que durante el desarrollo del proyecto es posible que se generen frecuentes y abundantes accesos a la base de datos con el fin de realizar consultas y probablemente se produzcan modificaciones, lo cual podría generar muchos problemas. Vemos las tareas que componen esta fase, figura 6 [CRISP-DM, 2000].

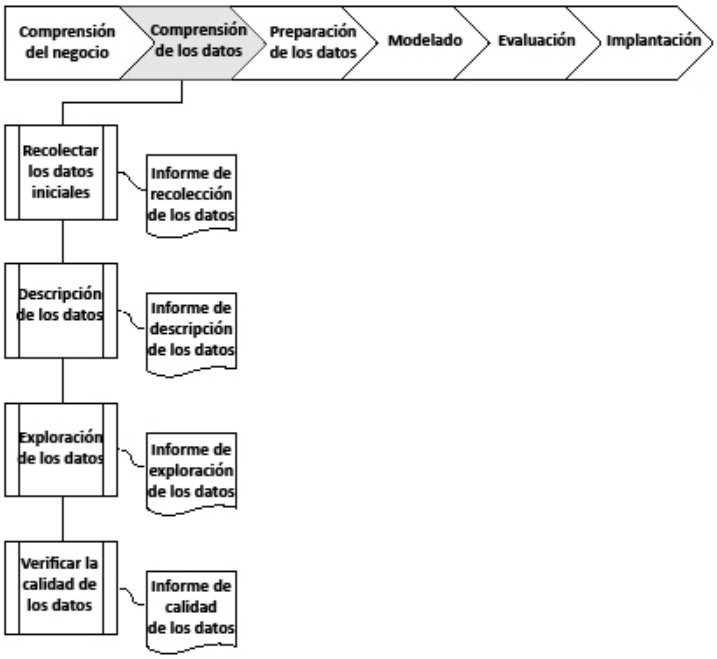


Figura 6 – Fase de comprensión de los datos

- **Recolectar los datos iniciales.**

  La primera tarea en esta segunda fase del proceso de CRISP-DM es la recolección de los datos iniciales y su adecuación para el futuro procesamiento. Esta tarea tiene como objetivo elaborar informes con una lista de los datos adquiridos, su localización, las técnicas utilizadas en su recolección y los problemas y soluciones inherentes a este proceso.

- **Descripción de los datos.**

  Después de adquiridos los datos iniciales, estos deben ser descritos. Este proceso implica establecer volúmenes de datos (número de registros y campos por registro), su identificación, el significado de cada campo y la descripción del formato inicial.

- **Exploración de los datos.**

  Una vez realizada la descripción de los datos, se procede a su exploración, cuyo fin es encontrar una estructura general para los datos. Esto implica la aplicación de pruebas estadísticas básicas que revelen propiedades en los datos recién adquiridos, se crean tablas de frecuencia y se construyen gráficos de distribución. La salida de esta tarea es un informe de exploración de los datos.

- **Verificar la calidad de los datos.**

  En esta tarea se efectúan verificaciones sobre los datos para determinar la consistencia de los valores individuales de los campos, la cantidad y distribución de los valores nulos, y para encontrar valores fuera de rango, los cuales pueden constituirse en ruido para el proceso. La idea una vez llegados a este punto es poder garantizar la completitud y corrección de los datos.

3. **Preparación de los datos.**

   En esta fase y una vez efectuada la recolección inicial de los datos, se procede a su preparación para adaptarlos a las técnicas de minería de datos que se van a utilizar posteriormente, éstas pueden ser técnicas de visualización de datos, de búsqueda de relaciones entre variables u otras medidas para explotación de los datos. La preparación de los datos incluye las tareas generales de selección de datos a los que se va a aplicar una determinada técnica de modelado, limpieza de datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato.

   Esta fase se encuentra relacionada con la fase de modelado, ya que en función de la técnica de modelado elegida, los datos requieren ser procesados de una manera o de otra, por esta razón las fases de preparación y de modelado interactúan de forma permanente. En la figura 7 [CRISP-DM, 2000] se pueden ver cada una de las tareas de las que se compone esta fase así como las salidas de cada una de ellas.

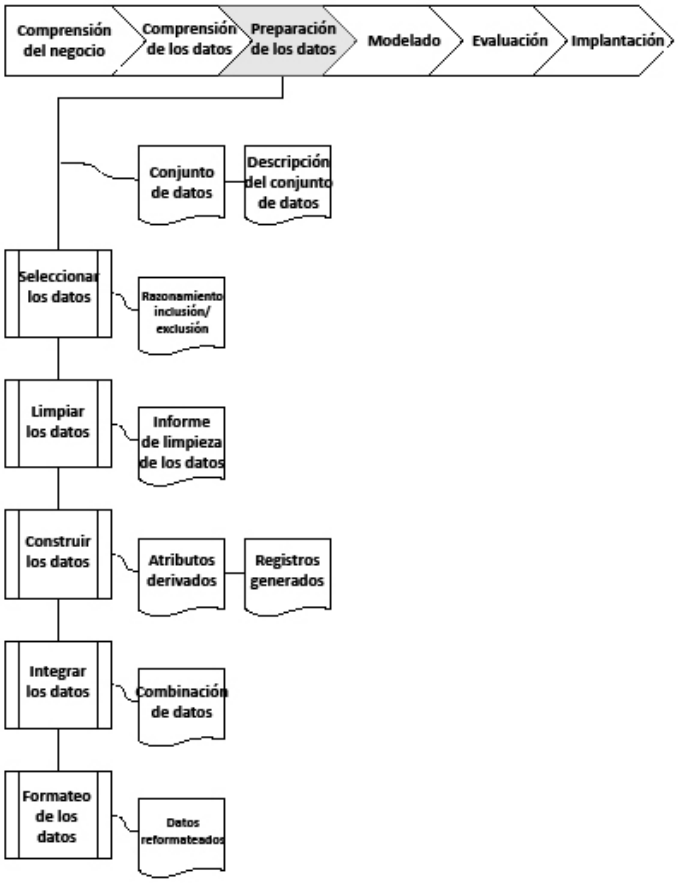


Figura 7 – Fase de preparación de los datos

- **Seleccionar los datos.**

  En esta etapa se selecciona un subconjunto de los datos adquiridos anteriormente apoyándose en criterios previamente definidos en las fases anteriores como la calidad de los datos en cuanto a su completitud, corrección de los datos y limitaciones en el volumen o en los tipos de datos que están relacionados con las técnicas de minería de datos seleccionadas.

- **Limpiar los datos.**

  Esta tarea complementa a la anterior y es una de las que más tiempo y esfuerzo consume debido a la diversidad de técnicas que pueden aplicarse para optimizar la calidad de los datos a objeto de prepararlos para la fase de modelación. Algunas de las técnicas a utilizar para este propósito son la normalización de los datos, discretización de campos numéricos, tratamiento de valores faltantes, reducción del volumen de datos, etc.

- **Construir los datos.**

Esta tarea incluye las operaciones de preparación de los datos tales como la generación de nuevos atributos a partir de atributos ya existentes, integración de nuevos registros o transformación de valores para atributos existentes.

- **Integrar los datos.**

  La integración de los datos implica la creación de nuevas estructuras a partir de los datos seleccionados, por ejemplo, generación de nuevos campos a partir de otros existentes, creación de nuevos registros, fusión de tablas campos o nuevas tablas donde se resumen características de múltiples registros o de otros campos en nuevas tablas de resumen.

- **Formateo de los datos.**

  Esta tarea consiste principalmente en la realización de transformaciones sintácticas de los datos sin modificar su significado de tal forma que se permita y se facilite utilizar alguna técnica de minería de datos en concreto, como por ejemplo la reordenación de los campos y/o de los registros de la tabla o el ajuste de los valores de los campos a las limitaciones de las herramientas de modelación (eliminar comas, tabuladores, caracteres especiales, máximos y mínimos para las cadenas de caracteres, etc.).

4. **Modelado.**

   En esta fase de CRISP-DM se seleccionan las técnicas de modelado más apropiadas para el proyecto de minería de datos específico. Las técnicas a utilizar en esta fase se eligen en función de los siguientes criterios:

   - Ser apropiada para el problema.
   - Disponer de los datos adecuados.
   - Cumplir los requisitos del problema.
   - Tiempo adecuado para obtener un modelo.
   - Conocimiento de la técnica.

   Previamente al modelado de los datos se debe determinar un método de evaluación de los modelos que permita establecer el grado de adecuación de cada uno de ellos. Después de concluir estas tareas genéricas se procede a la generación y evaluación del modelo. Los parámetros utilizados en la generación del modelo dependen de las características de los datos y de las características de precisión que se quieran lograr con el modelo. La figura 8 muestra las tareas y

las salidas que se obtienen en esta fase, a continuación describimos las tareas principales de esta fase.

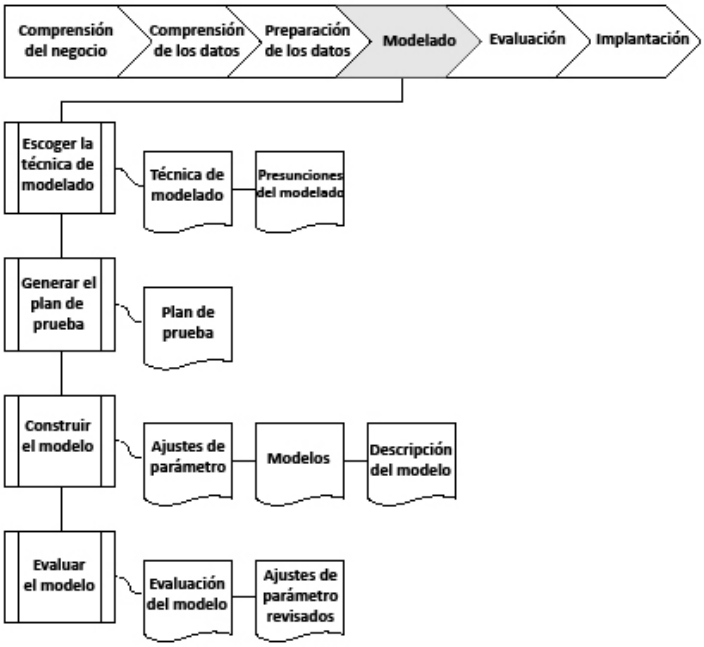


Figura 8 - Fase de modelado

- **Escoger la técnica de modelado.**

  Esta tarea consiste en la selección de la técnica de minería de datos más apropiada al tipo de problema que se quiere resolver. Para esta selección, se debe considerar el objetivo principal del proyecto y la relación con las herramientas de minería de datos existentes. Por ejemplo, si el problema es de clasificación, se podrá elegir de entre árboles de decisión, k-nearest neighbors o razonamiento basado en casos (CBR), si el problema es de predicción, análisis de regresión o redes neuronales, o si el problema es de segmentación, redes neuronales, técnicas de visualización, etc.

- **Generar el plan de prueba.**

  Se debe generar un procedimiento destinado a probar la calidad y validez del modelo elegido una vez que éste esté construido. Por ejemplo, en una tarea supervisada de minería de datos como la clasificación, es común usar la razón de error como medida de la calidad. Entonces, típicamente se separan los datos en dos conjuntos, uno de entrenamiento y otro de prueba, para

luego construir el modelo basado en el conjunto de entrenamiento y medir la calidad del modelo generado con el conjunto de prueba.

- **Construir el modelo.**

  A continuación se ejecuta la técnica seleccionada sobre los datos previamente preparados para generar uno o más modelos. Todas las técnicas de modelado tienen un conjunto de parámetros que determinan las características del modelo a generar. La selección de los mejores parámetros es un proceso iterativo y se basa exclusivamente en los resultados generados. Estos deben ser interpretados y su rendimiento justificado.

- **Evaluar el modelo.**

  En esta última tarea de esta fase de modelado los ingenieros de DM interpretan los modelos de acuerdo al conocimiento preexistente del dominio y los criterios de éxito preestablecidos. Expertos en el dominio del problema juzgan los modelos dentro del contexto del dominio y expertos en minería de datos aplican sus propios criterios (seguridad del conjunto de prueba, pérdida o ganancia de tablas, etc.).

5. **Evaluación.**

   En esta fase se evalúa el modelo, teniendo en cuenta el cumplimiento de los criterios de éxito del problema. Debe considerarse además que la fiabilidad calculada para el modelo se aplica solamente para los datos sobre los que se realizó el análisis. Es preciso revisar el proceso, teniendo en cuenta los resultados obtenidos, para poder repetir algún paso anterior, en el que se pueda haber cometido algún error. Considerar que se pueden emplear múltiples herramientas para la interpretación de los resultados. Si el modelo generado es válido en función de los criterios de éxito establecidos en la fase anterior, se procede a la explotación del modelo. La figura 9 detalla las tareas que componen esta fase y los resultados que se deben obtener. Las tareas involucradas en esta fase del proceso son las siguientes:

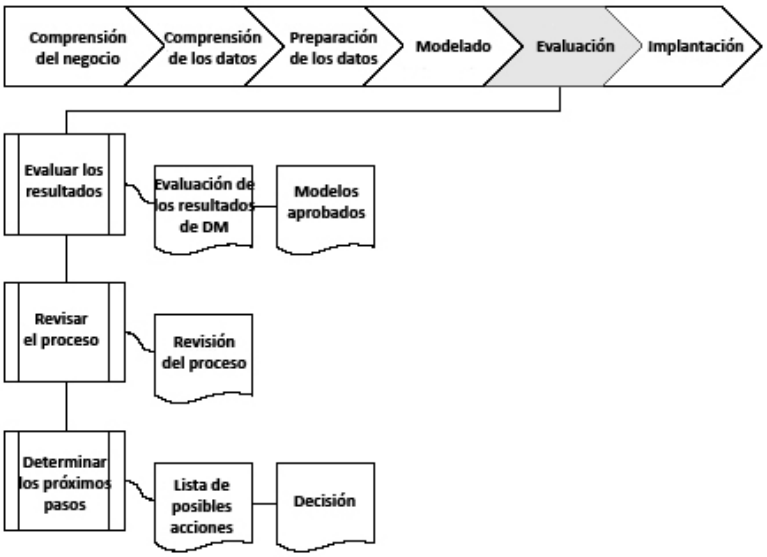


Figura 9 - Fase de evaluación

- **Evaluar los resultados.**

  En los pasos de evaluación anteriores se trataron factores tales como la exactitud y generalidad del modelo generado. Esta tarea involucra la evaluación del modelo en relación a los objetivos del negocio y busca determinar si hay alguna razón de negocio para la cual el modelo sea deficiente, o si es aconsejable probar el modelo en un problema real si el tiempo y las restricciones lo permiten. Además de los resultados directamente relacionados con el objetivo del proyecto, ¿es aconsejable evaluar el modelo en relación a otros objetivos distintos a los originales?, esto podría revelar información adicional.

- **Revisar el proceso.**

  Este proceso se refiere a calificar al proceso entero de minería de datos a objeto de identificar elementos que pudieran ser mejorados.

- **Determinar los próximos pasos.**

  Si se ha determinado que las fases hasta este momento han generado resultados satisfactorios podría pasarse a la siguiente fase, en caso contrario podría decidirse por hacer otra iteración desde la fase de preparación de los datos o de modelado con distintos parámetros. Podría incluso darse el caso de que en esta fase se decida empezar desde cero con un nuevo proyecto de minería de datos.

6. **Despliegue o implantación.**

   En esta fase, y una vez que el modelo ha sido construido y validado, se transforma el conocimiento obtenido en acciones dentro del proceso de negocio, esto puede hacerse por ejemplo cuando el analista recomienda acciones basadas en la observación del modelo y sus resultados, o por ejemplo aplicando el modelo a diferentes conjuntos de datos o como parte del proceso (en análisis de riesgo de créditos, detección de fraudes, etc.). Generalmente un proyecto de minería de datos no concluye en la implantación del modelo, ya que se deben documentar y presentar los resultados de manera comprensible para el usuario con el objetivo de lograr un incremento del conocimiento. Por otra parte, en la fase de explotación se debe asegurar el mantenimiento de la aplicación y la posible difusión de los resultados. Las tareas que componen esta fase (figura 10) son:

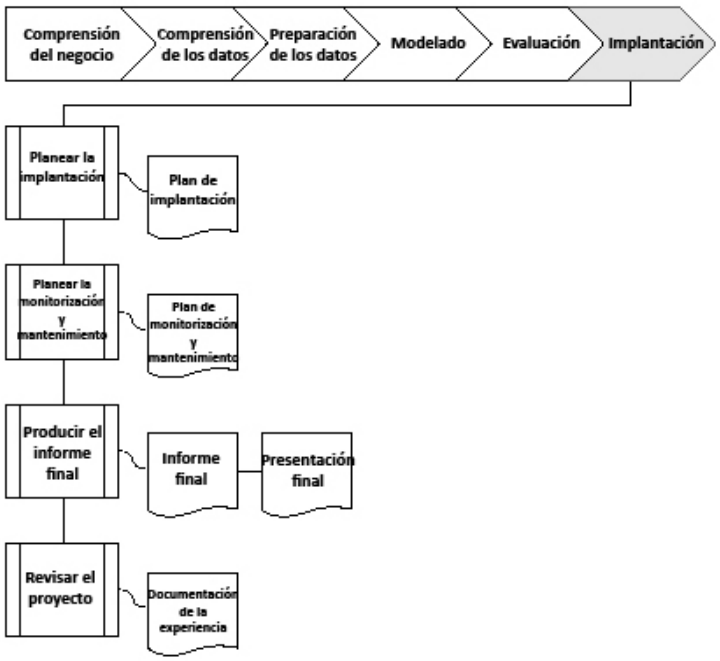


Figura 10 - Fase de implantación

- **Planear la implantación.**

  Para implementar el resultado de la minería de datos en la organización, esta tarea toma los resultados de la evaluación y concluye una estrategia para su implementación. Si un procedimiento general se ha identificado para crear el

modelo, este procedimiento debe ser documentado para su posterior implementación.

- **Planear la monitorización y mantenimiento.**

  Si los modelos resultantes del proceso de minería de datos son implementados en el dominio del problema como parte de la rutina diaria, es aconsejable preparar estrategias de monitorización y mantenimiento para ser aplicadas sobre los modelos. La retroalimentación generada por la monitorización y mantenimiento pueden indicar si el modelo está siendo utilizado apropiadamente.

- **Producir el informe final.**

  Es la conclusión del proyecto de minería de datos realizado. Dependiendo del plan de implementación, este informe puede ser sólo un resumen de los puntos importantes del proyecto y la experiencia adquirida o puede ser una presentación final que incluya y explique los resultados logrados con el proyecto.

- **Revisar el proyecto.**

  En esta tarea se evalúa que cosas se hicieron correctamente y cuales fueron incorrectas, así como aquellos puntos que se podrían mejorar en el proyecto.

1.6. Herramientas

En este apartado comentaremos algunas de las herramientas, tanto de pago como gratuitas, que sirven de apoyo para la minería de datos, junto con una breve descripción de cada una de ellas. Clasificaremos estas herramientas en tres grupos para poder diferenciarlas de manera más directa, tal y como se hace en [Hernández, Ramírez y Ferri, 2004], estos tres grupos son: librerías, *suites*, y herramientas específicas.

1.6.1. Librerías

Las librerías de minería de datos son un conjunto de métodos que implementan las funcionalidades y utilidades básicas que se utilizan en la minería de datos, como por ejemplo el acceso a datos, inferencia de modelos, exportación y comprobación de resultados, etc. Estas librerías son en realidad una interfaz para que el desarrollador pueda utilizarlas, por lo que no son aptas para cualquier usuario que no tenga conocimientos de programación. De este grupo de herramientas veremos XELOPES y JDMP.

XELOPES

XELOPES (*eXtEnded Library fOr Prudsys Embedded Solutions*) [prudsys, 2011] es una librería con licencia pública para el desarrollo de aplicaciones de minería de datos basada en el estándar *Common Warehouse Metamodel* (CWM) del *Object Management Group* (OMG). Esta librería se puede utilizar prácticamente en cualquier plataforma y sobre la mayoría de fuentes de datos.

XELOPES permite exportar sus modelos de minería de datos en formato XML a otros entornos de minería de datos como por ejemplo el estándar PMML. XELOPES también permite hacer el proceso inverso e importar modelos PMML para poder ser utilizados como una nueva fuente de datos.

La librería XELOPES implementa algoritmos potentes de minería de datos que se integran fácilmente con cualquier aplicación, y además permite al usuario incorporar en la librería nuevos métodos de su propia creación.

En cuanto al acceso a datos, existe una clase especial que permite dar uniformidad a todos los modos de accesos de datos permitidos. Es decir, que el usuario podría accedes por ejemplo a archivos .log, archivos de bases de datos o incluso crear su propio formato de datos.

XELOPES además ofrece una gran variedad de modelos para la minería de datos, entre ellos están:

- Árboles de decisión lineales y no lineales.
- Máquinas de vectores soporte.
- Redes neuronales.
- Métodos de agrupamiento.
- Métodos de reglas de asociación.

Esta librería, implementada por *Prudsys AG* en colaboración con *Russian ZSoft Ltd.*, está disponible para C++, Java, C# y CORBA.

JDMP

JDMP (*Java Data Mining Package*) [JDMP, 2011] es una librería de código abierto en Java para el análisis de los datos y el aprendizaje automático. Esta librería facilita el acceso a las fuentes de datos y a los algoritmos de minería de datos (por ejemplo los de agrupamiento, regresión, clasificación, modelos gráficos u optimización) y además cuenta con módulos de visualización.

JDMP incluye una librería de matrices para el almacenamiento y procesamiento de cualquier tipo de datos, pudiendo manejar matrices muy grandes incluso si éstas no caben en la memoria. También dispone de interfaces de importación y exportación para las bases de datos de JDBC, TXT, CSV, Excel, Matlab, Latex, MTX, HTML, WAV, BMP y otros formatos de archivo. Otra característica de JDMP es que proporciona

interfaces para otros paquetes de minería de datos como por ejemplo WEKA, LibSVM, Mallet, Lucene u Octave.

JDMP se basa principalmente en un principio, una representación de los datos consistente. En JDMP muchos objetos se representan con una matriz, por ejemplo, se pueden combinar varias matrices para formar una variable, por ejemplo para una serie temporal. Se puede acceder a estas matrices de una en una o como una sola matriz más grande dependiendo de nuestros intereses. Varias variables se pueden combinar para formar una muestra, muchas muestras forman un conjunto de datos (*Data Set*). Este conjunto de datos se puede acceder muestra a muestra o como una gran matriz también. Se proporcionan algoritmos para la manipulación de variables, muestras y conjuntos de datos, además en JDMP los métodos de procesamiento de datos están separados de las fuentes de los datos, de tal manera que los algoritmos y los datos pueden encontrarse en ordenadores distintos y así se puede procesar de forma paralela.

1.6.2. *Suites*

La diferencia entre una *suite* y una librería es que en el caso de las *suites*, el usuario no necesita tener conocimientos de programación para poder usar las herramientas de minería de datos, ya que se proporciona una interfaz, generalmente gráfica que facilita y hace más intuitivo el uso de estas herramientas.

Una *suite* integra en un mismo entorno herramientas para el procesamiento de los datos, modelos de análisis, herramientas que facilitan el diseño de experimentos y una parte gráfica que hace más fácil la visualización de los resultados.

De este tipo de herramientas veremos IBM SPSS Modeler, WEKA, Oracle Data Mining, RapidMiner y STATISTICA Data Miner.

IBM SPSS Modeler

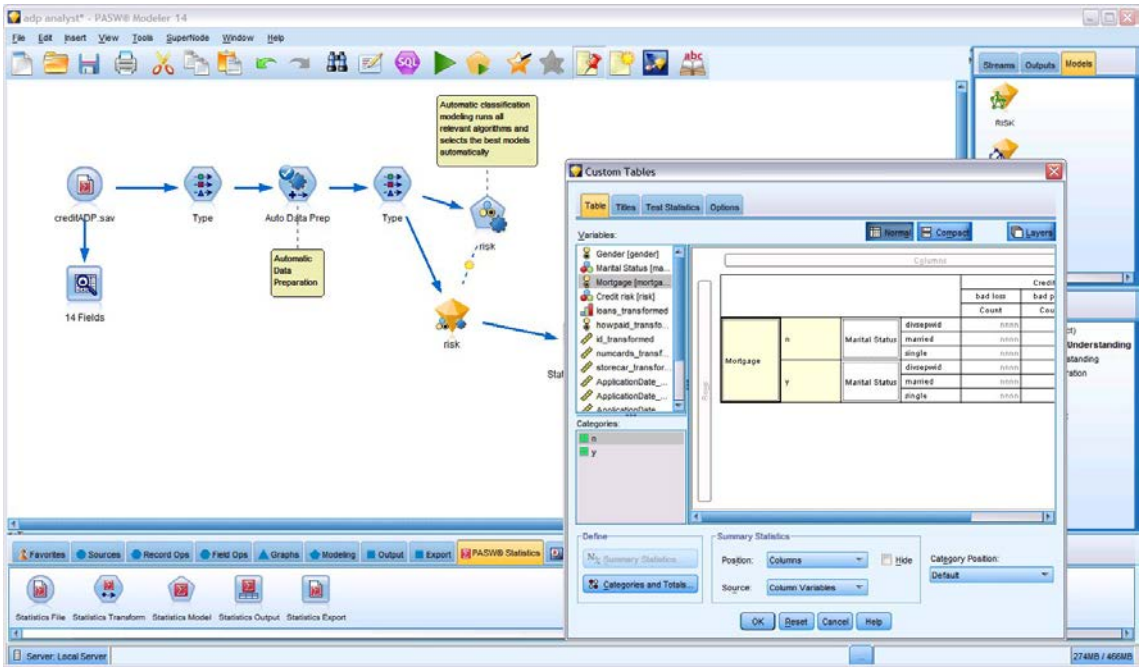
IBM SPSS Modeler [IBM, 2011] es una herramienta de software para la minería de datos desarrollada por SPSS Inc., una compañía de IBM. Se trata de una herramienta de

pago, originalmente se llamaba *SPSS Clementine*, pero en el año 2009 pasó a llamarse *PASW Modeler*. Actualmente se llama IBM SPSS Modeler después de que IBM adquiriera la compañía SPSS Inc.

El sistema se caracteriza por:

- Acceso a datos: fuentes de datos ODBC, tablas Excel, archivos planos ASCII y archivos SPSS.
- Pre-procesado de datos: *pick & mix*, muestreo, particiones, reordenación de campos, nuevas estrategias para la fusión de tablas, nuevas técnicas para recodificar intervalos numéricos, etc.
- Técnicas de aprendizaje: árboles de decisión, redes neuronales, agrupamiento, reglas de asociación, regresión lineal y logística, combinación de modelos.
- Técnicas para la evaluación de modelos guiadas por las condiciones especificadas por el experto.
- Visualización de resultados: ofrece un potente soporte gráfico que permite al usuario tener una visión global de todo el proceso, que comprende desde el análisis del problema hasta la imagen del modelo aprendido.
- Permite generar automáticamente informes en HTML y texto, volcar los resultados de la minería de datos obtenidos en bases de datos y exportar los modelos a distintos lenguajes como C, SPSS, HTML, PMML, SQL, etc.

Este software se encuentra disponible para diversas plataformas, como Windows, Linux, Sun Solaris, HP-UX o IBM-AIX.



WEKA

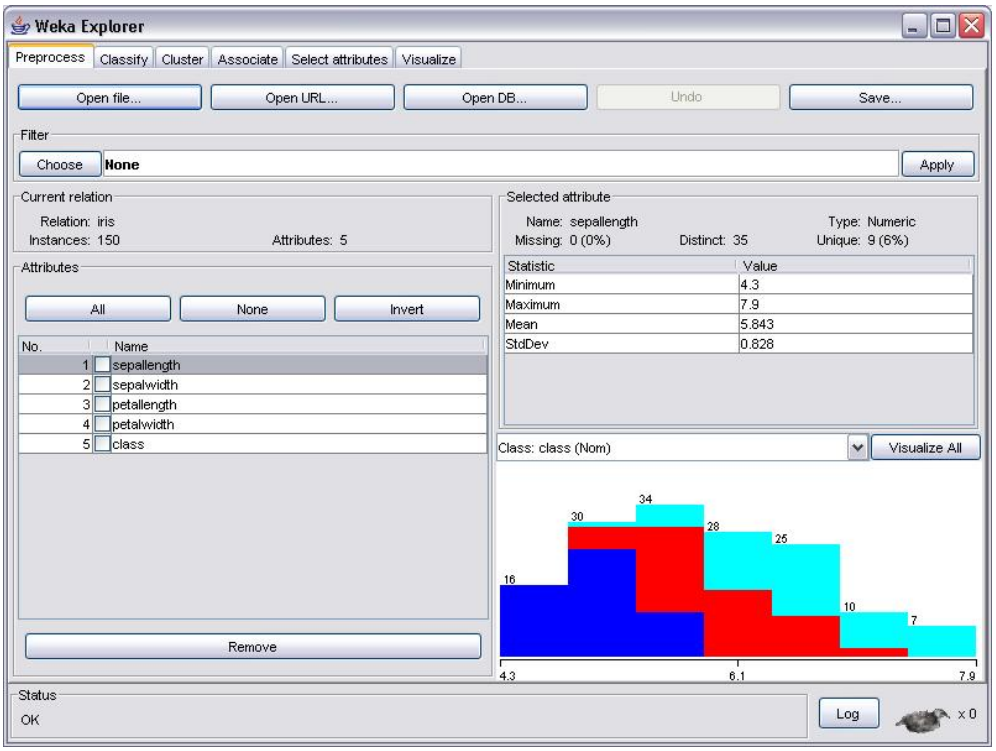
WEKA (*Waikato Environment for Knowledge Analysis*) [Waikato, 2011] es un software libre distribuido bajo licencia GNU-GPL, escrito en Java y desarrollado por la *Universidad de Waikato*, Nueva Zelanda.

WEKA contiene una colección de herramientas de visualización y algoritmos para el análisis de datos y modelado predictivo, junto con una interfaz gráfica para poder acceder fácilmente a sus funcionalidades.

Una de las ventajas de WEKA es que es altamente portable al estar completamente implementado en Java, por lo que se puede ejecutar en prácticamente cualquier plataforma. Además WEKA contiene una extensa colección de técnicas para pre-procesamiento de datos, como por ejemplo: selección de atributos, discretización, tratamiento de valores desconocidos y transformación de atributos numéricos. También proporciona una amplia gama de modelos de aprendizaje, concretamente: árboles de decisión, tablas de decisión, vecinos más próximos, máquinas de vectores soporte, reglas de asociación, métodos de agrupamiento y modelos combinados.

En cuanto a la interfaz gráfica, WEKA nos permite la opción de seleccionar entre cuatro posibles entornos para acceder a las funcionalidades del programa, éstos son "*Simple*

CLI", que es una consola que permite utilizar WEKA desde la línea de comandos, "*Explorer*", que permite controlar todas las operaciones que ofrece WEKA mediante una serie de paneles. El entorno "*Experimenter*" permite la comparación sistemática de una ejecución de los algoritmos predictivos de WEKA sobre una colección de conjuntos de datos. Por último, el entorno "*Knowledge Flow*" que es una interfaz que soporta básicamente las mismas funciones que el *Explorer* pero mediante una interfaz gráfica que permite "arrastrar y soltar" (*drag and drop*), además este entorno ofrece soporte para el aprendizaje incremental.



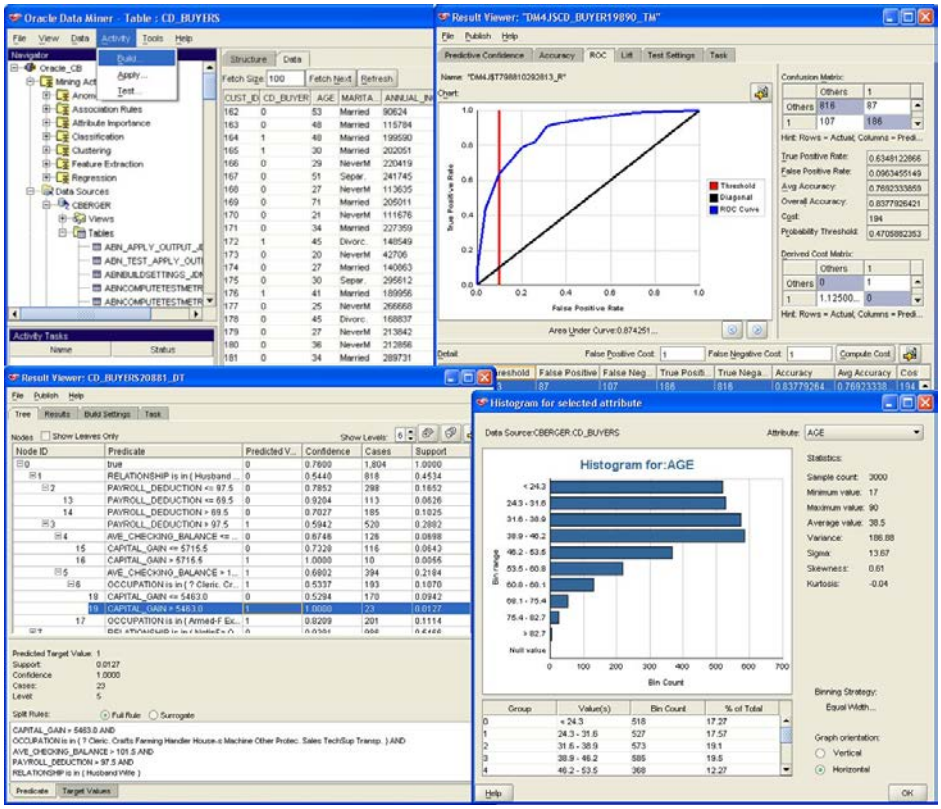
Oracle Data Mining

Oracle Data Mining [Oracle, 2011] fue originalmente desarrollado por *Thinking Machines Corporation* en los años 90 y distribuido con el nombre de Darwin. En el año 1999 Oracle adquirió la compañía y siguió distribuyendo el software bajo el mismo nombre hasta que en el año 2002 salió al mercado Oracle Data Mining, un rediseño casi completo del producto anterior. Oracle Data Mining es una opción del *Relational Database Management System* (RDBMS) *Enterprise Edition* (EE) de *Oracle Corporation*. Contiene una variedad de algoritmos de minería de datos para la

clasificación, predicción, regresión, agrupamiento, asociaciones, detección de anomalías y otras técnicas de análisis especializadas. Oracle Data Mining implementa estos algoritmos de minería de datos dentro de una base de datos relacional de Oracle, estas implementaciones se integran directamente en el núcleo de la base de datos de Oracle y opera con los datos almacenados en las tablas de la base de datos relacional, lo cual elimina la necesidad de extraer o transferir los datos a servidores específicos para la minería.

El sistema se organiza alrededor de unas pocas operaciones genéricas que proporcionan una interfaz general unificada con las funciones de minería de datos. Estas operaciones incluyen funciones para crear, aplicar, probar y manipular modelos de minería de datos. Los modelos se crean y se almacenan como objetos de la base de datos, y su gestión se hace desde dentro de la base de datos de manera parecida a como se gestionan las tablas y otros objetos de la base de datos. La interfaz gráfica de ODM (llamada Oracle Data Miner) guía al usuario paso a paso durante el proceso de la creación, evaluación e implantación de modelos de forma muy parecida a como lo hace la metodología CRISP-DM.

ODM ofrece una selección de modelos de aprendizaje automático como por ejemplo: árboles de decisión, aprendizaje bayesiano, máquinas de vectores de soporte, regresión lineal, reglas de asociación, técnicas de agrupamiento (K medias y O-agrupamiento), etc.



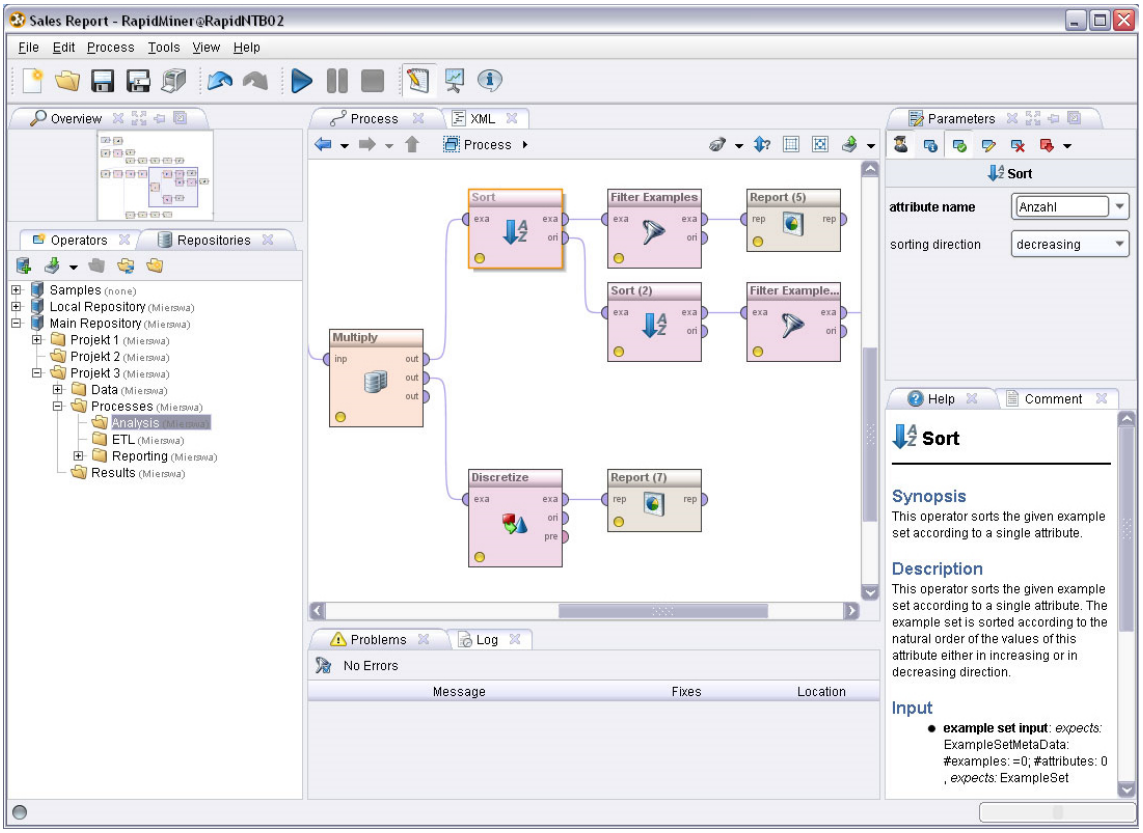
RapidMiner

RapidMiner [Rapid-I, 2011], conocido inicialmente como YALE (*Yet Another Learning Environment*) es otro software de minería de datos gratuito distribuido bajo licencia GPL e implementado en Java, por lo que es multiplataforma. Su versión inicial fue desarrollada por el departamento de inteligencia artificial de la *Universidad de Dortmund* en el año 2001. Este programa permite el desarrollo de procesos de análisis de datos mediante el encadenamiento de operadores a través de un entorno gráfico y se suele utilizar en investigación y en aplicaciones empresariales.

RapidMiner proporciona más de 500 operadores orientados al análisis de datos, incluyendo los necesarios para realizar operaciones de entrada y salida, pre-procesamiento de datos y visualización, además RapidMiner permite utilizar los algoritmos incluidos en WEKA.

Otras características de RapidMiner son:

- Utiliza una representación interna de los procesos de análisis de datos en ficheros XML.
- Permite el desarrollo de programas a través de un lenguaje de script.
- Puede usarse a través de una interfaz gráfica, línea de comandos, batch o incluso desde otros programas a través de llamadas a sus bibliotecas.
- Es extensible.
- Incluye gráficos y herramientas para la visualización de los datos.
- Dispone de un módulo de integración con el lenguaje de programación para análisis estadístico "R".



STATISTICA Data Miner

STATISTICA Data Miner [StatSoft, 2011] es un sistema visual desarrollado y comercializado por *StatSoft Ltd.*

De esta herramienta podemos destacar:

En cuanto al acceso a los datos el sistema está optimizado para trabajar con grandes volúmenes de datos de entrada, pudiendo importar datos en diversos formatos como Excel, tablas de Dbase, archivos de texto plano, Lotus, bases de datos de Oracle, Microsoft SQL Server, Sybase y un formato de archivo propio.

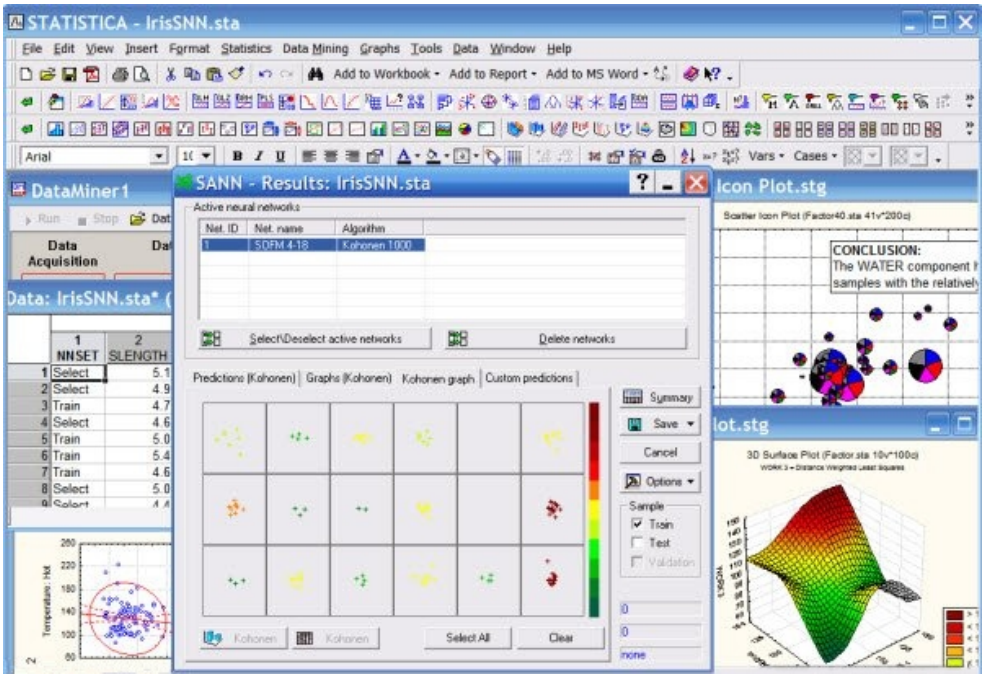
En cuanto al pre-procesamiento de los datos permite la selección de características, muestreo y operaciones estándar de filtrado, transformaciones de variables, tratamiento de valores desconocidos, etc.

Dispone de los siguientes modelos de análisis (entre otros):

- Reglas de asociación.
- Árboles de decisión.
- Métodos de agrupamiento (método K medias y EM).
- Redes neuronales.
- Utilidades estadísticas para la regresión de modelos lineales, no lineales, regresión múltiple, etc.

STATISTICA Data Miner dispone de una potente interfaz gráfica que facilita cualquier tarea que el usuario quiera ejecutar. Además, proporciona representaciones gráficas de los modelos con navegador de árboles de decisión, visualizadores de la topología de la red neuronal, visualizadores de reglas de asociación, etc. También permite la representación de gráficos estadísticos en dos y tres dimensiones (gráficos de barra, sectores, diagramas de líneas, diagramas de puntos, curvas de nivel, etc.) e incluso da la opción al usuario de poder especificar sus propias representaciones gráficas de los datos.

Esta herramienta sólo está disponible para el sistema operativo Microsoft Windows XP y sucesivos.



1.6.3. Herramientas Específicas

Este tipo de herramientas se caracterizan por ser específicas para un determinado modelo (redes neuronales, árboles de decisión, etc.) o una determinada tarea de minería de datos (clasificación, agrupamiento, etc.). Esto no quiere decir que estas herramientas no permitan realizar todo el proceso de la minería de datos, además para utilizarlas tampoco es necesario tener conocimientos de programación como pasa en el caso de las librerías. Vamos a ver dos ejemplos dentro de este tipo de herramientas, CART y See5/C5.0.

CART

CART [Salford, 2011] es una herramienta gráfica desarrollada y comercializada por Salford Systems. Esta herramienta está especialmente orientada hacia la inferencia de árboles de decisión para tareas de clasificación o regresión. Esto se puede utilizar en aplicaciones para encontrar clientes potenciales, marketing, detección de fraudes en tarjetas de crédito, o la gestión de riesgos en créditos. CART está diseñada para usuarios tanto técnicos como no técnicos.

En cuanto a su accesibilidad se puede decir que CART puede trabajar con más de 70 formatos de archivos diferentes. Además, CART dispone de herramientas de visualización interactivas, pudiendo el usuario solicitar información detallada del modelo.

CART está disponible tanto para sistemas Windows como UNIX y Linux.

See5/C5.0

See5/C5.0 [RuleQuest, 2011] es una herramienta de fácil manejo desarrollada por la empresa *RuleQuest Research Ltd.*, fundada por Ross Quinlan, autor de algunos sistemas de minería de datos muy populares como ID3m C4.5 y FOIL.

El entorno See5/C5.0 se centra en la construcción de modelos de clasificación basados en árboles de decisión y conjuntos de reglas, y es una evolución del método C4.5 (es más exacto, más rápido, consume menos recursos y produce árboles de decisión más pequeños).

Esta herramienta ha sido diseñada para operar sobre grandes volúmenes de datos. Un inconveniente de esta herramienta es que trabaja con un formato de archivos predefinido (.data), aunque se puede utilizar una herramienta complementaria desarrollada también por *RuleQuest Research* llamada *ODBCHook* que permite traducir fuentes de datos accesibles vía ODBC en archivos .data. Además los modelos aprendidos se pueden exportar a código C para que se puedan incorporar como parte de otros sistemas de aprendizaje más complejos.

See5 está disponible para sistemas Windows mientras que C5.0 lo está para sistemas UNIX.

Para terminar este análisis de las herramientas de minería de datos más importantes en la actualidad, a continuación se muestra una gráfica (figura 11), que representa una encuesta realizada por el sitio web kdnuggets.com a cerca de mil usuarios sobre las herramientas de minería de datos más utilizadas durante el periodo de tiempo comprendido entre mayo de 2009 y mayo de 2010.

RapidMiner (345)	37.8%
R (272)	29.8%
Excel (222)	24.3%
KNIME (175)	19.2%
Your own code (168)	18.4%
Pentaho/Weka (131)	14.3%
SAS (110)	12.0%
MATLAB (84)	9.2%
IBM SPSS Statistics (72)	7.9%
Other free tools (67)	7.3%
IBM SPSS Modeler (former Clementine) (67)	7.3%
Microsoft SQL Server (63)	6.9%
Statsoft Statistica (57)	6.2%
Other commercial tools (56)	6.1%
SAS Enterprise Miner (50)	5.5%
Zementis (34)	3.7%
Orange (25)	2.7%
Oracle DM (19)	2.1%
KXEN (19)	2.1%
Salford CART Mars other (15)	1.6%
VisuaLinks (12)	1.3%
Viscovery (10)	1.1%
Angoss (8)	0.9%
TIBCO Insightful Miner (7)	0.8%
Miner3D (7)	0.8%
REvolution Computing (4)	0.4%
Megaputer Polyanalyst/TextAnalyst (3)	0.3%
Portrait Software (2)	0.2%
Data Applied (2)	0.2%
Centrifuge (2)	0.2%
PRSD Studio (1)	0.1%
Clario Analytics (1)	0.1%
Bayesia (1)	0.1%

Figura 11 – Herramientas de minería de datos más empleadas

1.7. ¿Por qué Oracle Data Mining?

Una vez que hemos visto las herramientas más importantes que hay en el mercado, en este apartado explicamos el porqué de la elección de la herramienta Oracle Data Mining que es la herramienta que se empleará para realizar la minería de datos sobre la base de datos de la que se dispone.

Este proyecto consiste en emplear la metodología CRISP-DM para hacer una minería de datos, por lo que lo lógico es que antes de escoger la herramienta adecuada nos aseguremos de que dicha herramienta se adapte a nuestra metodología. En este caso, Oracle Data Mining se adapta bien a esta metodología. Como se ha visto en el apartado 1.5, los pasos a seguir en el proceso de la minería de datos son:

1. Comprensión del negocio.
2. Comprensión de los datos.
3. Preparación de los datos.
4. Modelado.
5. Evaluación.
6. Implantación.

Oracle Data Mining da soporte a los pasos 4, 5 y 6 del proceso [Oracle DM, 2011]. El primer paso (comprensión del negocio) es único para cada negocio, el resto de pasos se realizan con una combinación de Oracle Data Mining y una base de datos Oracle, o un almacén de datos Oracle. Las bases de datos de Oracle proporcionan herramientas específicas para la comprensión y preparación de los datos, es por ello que se opta por utilizar un almacén de datos Oracle. Habiendo escogido como soporte de almacenamiento una base de datos (o almacén) Oracle, podemos afirmar que, al tratarse de una herramienta desarrollada por la misma compañía, Oracle Data Mining se integra a la perfección con estas bases de datos. Las herramientas que proporciona Oracle Data Mining están directamente integradas en el núcleo de la base de datos, por lo que operan de forma nativa sobre los datos almacenados en las bases de datos, de esta forma no es necesario tener que transferir la información desde la base de datos a cualquier otra herramienta para aplicar los algoritmos de minería de datos, es decir, los algoritmos se pueden ejecutar directamente sobre la base de datos Oracle.

Además, Oracle DM proporciona al usuario varios algoritmos distintos para diferentes tipos de análisis: árboles de decisión, clasificador bayesiano naive, SVM (Máquinas de Vectores de Soporte) y GLM (Modelo Lineal Generalizado), con lo que se pueden buscar resultados de varias formas distintas en el caso de que alguno de los algoritmos no produjera el resultado buscado.

Otra razón para la elección de esta herramienta es la facilidad de uso ya que Oracle Data Mining tiene una interfaz gráfica bastante intuitiva incluso para usuarios no expertos en el dominio de la minería de datos, esta interfaz gráfica de usuario (GUI) es conocida como Oracle Data Miner.

Como contrapartida cabe decir que Oracle Data Mining es una herramienta de pago.

Parte II:

Aplicación de la Metodología CRISP-DM al Problema

En esta segunda parte del proyecto pasamos a la parte más práctica, donde iremos aplicando cada una de las fases de la metodología CRISP-DM al problema práctico que nos planteamos, que es la extracción y explotación de datos del entorno universitario.

Iremos numerando cada una de las fases de la metodología tal y como están numeradas en el documento original.

1. Comprensión del Negocio

A continuación iremos siguiendo cada una de las tareas de las que consta esta primera fase en el proceso de la minería de datos, cuya finalidad es determinar los objetivos y requisitos del proyecto desde una perspectiva de negocio, para más adelante poder convertirlos en objetivos desde el punto de vista técnico y en un plan de proyecto.

1.1. Determinar los Objetivos del Negocio

El objetivo de la minería de datos que se va a aplicar en este proyecto es el de hacer predicciones lo más fiables posible a partir de los datos de los que ya se disponen de los alumnos en una universidad. El objetivo es proporcionar un mejor servicio de enseñanza a los alumnos y así poder captar más alumnos para que realicen sus estudios en la universidad.

Contexto

En referencia a la situación de negocio en la organización (universidad) al principio de este proyecto se puede decir que se cuenta con una base de datos de los alumnos actualmente cursando una titulación en la universidad y también de aquellos que ya han terminado sus estudios. Sin embargo no existe ningún estudio en profundidad sobre el comportamiento de los estudiantes de los que se puedan sacar conclusiones o patrones para hacer predicciones sobre los futuros estudiantes.

Objetivos del negocio

Los objetivos del negocio como ya se ha mencionado son la predicción de datos para los alumnos de nuevo ingreso de tal manera que se pueda hacer una estimación fiable partiendo de los datos que ya tenemos de dichos alumnos. Se podrían hacer muchas

predicciones según las necesidades de la universidad en cada momento, pero en este proyecto se han definido los siguientes objetivos:

- Hacer predicciones acerca del tiempo que los alumnos emplean para acabar sus titulaciones.
- Predecir las notas medias de los alumnos al acabar la carrera.
- Predecir notas de asignaturas problemáticas para los alumnos.

Estos informes pueden ser muy útiles para los alumnos a la hora de escoger la titulación que van a realizar en la universidad, así como para detectar aquellas asignaturas problemáticas para los alumnos y de esta forma intentar averiguar por qué ciertas asignaturas pueden resultar más complicadas, ya sea por falta de preparación por parte de los alumnos, del profesorado, etc. Todo esto permitirá a la universidad mejorar la calidad de los servicios ofrecidos a los estudiantes.

Criterios de éxito del negocio

Desde el punto de vista del negocio se establece como criterio de éxito la posibilidad de realizar predicciones sobre nuevos alumnos con un elevado porcentaje de fiabilidad, de tal forma que se puedan dar consejos útiles a los alumnos acerca de que titulación escoger antes de comenzar sus estudios en la universidad, y una vez escogida la titulación que asignaturas optativas elegir en función del nivel del alumno. Otro criterio de éxito del negocio sería elevar el porcentaje de aprobados en aquellas asignaturas que tengan un bajo porcentaje de aprobados por parte de los alumnos.

1.2. Evaluación de la Situación

Se cuenta con una base de datos Oracle 11g con información detallada de los alumnos que han cursado alguna de las titulaciones de la universidad desde el año 1997 hasta la actualidad, por lo que a priori se puede afirmar que se dispone de una cantidad de datos más que suficiente para poder resolver el problema. Esta información incluye la nota de acceso a la universidad obtenida durante el bachillerato, el centro de procedencia del alumno, provincia, tipo de bachillerato realizado y otros datos personales del alumno que nos pueden ser útiles a la hora de hacer la minería de datos.

Inventario de recursos

En cuanto a recursos de software disponemos del programa de minería de datos *Oracle Data Mining* que proporciona herramientas para realizar tareas de minería de datos sobre una base de datos Oracle 11g que es con la que contamos para el almacenamiento de los datos.

Los recursos de hardware de los que disponemos son un ordenador de sobremesa con las siguientes características:

- Marca: Packard Bell ©
- Modelo: iMedia X1617
- Procesador: Intel © Core 2 Quad Q6600 a 2.40 GHz
- Memoria RAM: 3,00 GB
- Capacidad de almacenamiento: 360 GB
- Tarjeta gráfica: NVIDIA © GeForce 9600 GT
- Sistema operativo: Microsoft Windows 7 Professional ©
- Monitor TFT: Philips © 190 S de 19’’

La fuente de datos es una base de datos Oracle con la información de los alumnos matriculados en la universidad desde el año 1997 hasta el 2011.

Requisitos, supuestos y restricciones

Al no poder utilizar los datos personales de alumnos reales debido a cuestiones legales, se ha tenido que utilizar una base de datos ficticia con datos no reales de alumnos inventados.

Terminología

Ver Anexo 1: Glosario de terminología de minería de datos.

Costes y beneficios

Los datos de este proyecto no suponen ningún coste adicional a la universidad ya que estos datos perteneces a la propia universidad desde el momento en el que el alumno se matricula en ella.

En cuanto a beneficios, no se puede decir que este proyecto genere algún beneficio económico para la universidad directamente, pero si que puede suponerlo indirectamente ya que el objetivo de este proyecto es mejorar la calidad de los servicios ofrecidos a los alumnos por parte de la universidad, y por tanto la satisfacción de los clientes (los alumnos), y esto se traduce en prestigio para la universidad, lo cual hará que más alumnos consideren cursar sus estudios en esta universidad a la hora de elegir una.

1.3. Determinar los Objetivos de la Minería de Datos

Los objetivos en términos de minería de datos son:

- Predecir el tiempo (en cursos académicos) que un alumno tardará en conseguir el título universitario en función de su nota de ingreso a la universidad, centro de procedencia, provincia y titulación escogida.
- Predecir la nota media de la carrera que un alumno obtendrá cuando termine sus estudios.
- Identificar aquellas asignaturas que parecen ser más complicadas para los alumnos en cada titulación y predecir la nota que un determinado alumno obtendrá en la asignatura.

Criterios de éxito de minería de datos

Desde el punto de vista de la minería de datos se establece como criterio de éxito la posibilidad de realizar predicciones sobre nuevos alumnos con un elevado porcentaje de fiabilidad, concretamente definimos este porcentaje en un 80%. El grado de fiabilidad lo determinará el algoritmo específico que se emplee a la hora de conseguir el modelo de la minería de datos, por lo que este tema se volverá a abordar más adelante en el paso 5 de la metodología (evaluación).

1.4. Realizar el Plan del Proyecto

El proyecto se dividirá en las siguientes etapas para facilitar su organización y estimar el tiempo de realización del mismo:

- Etapa 1: Análisis de la estructura de los datos y la información de la base de datos. Tiempo estimado: 2 semanas.
- Etapa 2: Ejecución de consultas para tener muestras representativas de los datos. Tiempo estimado: 1 semana.
- Etapa 3: Preparación de los datos (selección, limpieza, conversión y formateo, si fuera necesario) para facilitar la minería de datos sobre ellos. Tiempo estimado: 3 semanas.
- Etapa 4: Elección de las técnicas de modelado y ejecución de las mismas sobre los datos. Tiempo estimado: 1 semana.
- Etapa 5: Análisis de los resultados obtenidos en la etapa anterior, si fuera necesario repetir la etapa 4. Tiempo estimado: 1 semana.
- Etapa 6: Producción de informes con los resultados obtenidos en función de los objetivos de negocio y los criterios de éxito establecidos. Tiempo estimado: 1 semana.
- Etapa 7: Presentación de los resultados finales. Tiempo estimado: 1 semana.

Nota: en paralelo a la realización de cada una de estas etapas se irá construyendo el diccionario de terminología de minería de datos (Anexo 1).

Evaluación inicial de herramientas y técnicas

La herramienta que se va a utilizar para llevar a cabo este proyecto de minería de datos es Oracle Data Mining ya que como se comentó en el apartado 1.7, esta herramienta se adapta bien a la metodología que estamos empleando y sobre todo a la base de datos en la que están almacenados todos los datos de los estudiantes. Además gracias a esta herramienta no necesitamos pasar la información almacenada en la base de datos a otra base de datos o a una herramienta de minería de datos, ya que Oracle Data Mining opera directamente sobre la base de datos Oracle.

En cuanto a las técnicas que se van a emplear para la extracción de conocimiento, Oracle Data Mining nos ofrece los siguientes tipos de tareas de minería de datos:

- Predictivas
    - Clasificación
    - Regresión
- Descriptivas
    - Agrupamiento (clustering)
    - Reglas de asociación

Oracle Data Mining además utiliza los siguientes algoritmos para resolver los problemas: árboles de decisión, clasificador bayesiano naive, SVM (Máquinas de Vectores de Soporte) y GLM (Modelo Lineal Generalizado).

Los árboles de decisión, también llamados modelos basados en árboles, se fundamentan en el principio de "divide y vencerás". Los árboles se van construyendo con nodos, de tal forma que en cada nodo se establecen unas condiciones sobre uno o varios atributos, dividiendo de esta manera el conjunto de casos en subconjuntos que cumplen las condiciones. Estos subconjuntos a su vez se vuelven a dividir añadiendo más niveles al árbol hasta detenerse en el punto en el que se cumpla algún criterio.

El clasificador bayesiano naive es un clasificador probabilístico basado en el teorema de Bayes y algunas hipótesis simplificadoras adicionales. Es a causa de estas simplificaciones que se suelen resumir en la hipótesis de independencia entre las variables predictoras, que recibe el nombre de *naive* (ingenuo).

Las Máquinas de Vectores de Soporte o Support Vector Machines (SVM) son un conjunto de algoritmos de aprendizaje supervisado que están principalmente relacionados con problemas de clasificación y regresión. Dado un conjunto de ejemplos de entrenamiento (muestras) podemos etiquetar las clases y entrenar una SVM para construir un modelo que prediga la clase de una nueva muestra. Una SVM es un modelo que representa a los puntos de muestra en el espacio, separando las clases por un espacio lo más amplio posible. Las nuevas muestras se clasifican en una u otra clase en función de la proximidad con el modelo producido. Más formalmente, una SVM construye un hiperplano o conjunto de hiperplanos en un espacio de dimensionalidad

muy alta (o incluso infinita) que puede ser utilizado en problemas de clasificación o regresión. Una buena separación entre las clases permite una clasificación mejor.

El Modelo Lineal Generalizado o Generalized Linear Model (GLM) es una generalización de la regresión de mínimos cuadrados ordinaria. Relaciona la distribución aleatoria de la variable dependiente en el experimento (la función de distribución) con la parte sistemática (no aleatoria) a través de una función llamada la función de enlace. Los modelos lineales generalizados fueron formulados como una manera de unificar varios modelos estadísticos, incluyendo la regresión lineal, regresión logística y regresión de Poisson, bajo un solo marco teórico. Esto ha permitido desarrollar un algoritmo general para la estimación de máxima verosimilitud en todos estos modelos.

2. Comprensión de los Datos

En esta segunda fase de la metodología CRISP-DM se realiza la recolección inicial de los datos para poder establecer un primer contacto con el problema, familiarizarse con los datos y averiguar su calidad, así como identificar las relaciones más evidentes para formular las primeras hipótesis.

2.1. Recolectar los Datos Iniciales

Los datos utilizados en este proyecto son datos referentes a alumnos que incluyen información personal sobre ellos como puede ser, sus nombres y apellidos, DNI, fechas de nacimiento, etc., por lo que no hemos podido utilizar datos reales que estén en las bases de datos de la universidad debido a impedimentos legales como es lógico. Por lo tanto, hemos tenido que crear y utilizar datos ficticios de alumnos inexistentes lo cual conlleva una serie de problemas, ya que como el objetivo del proyecto es realizar predicciones y estudios lo más reales posible, estos datos no pueden ser datos aleatorios y debe existir algún tipo de relación entre los atributos de cada registro (por ejemplo los alumnos de ciertos centros de enseñanza sacan mejores notas en la prueba de acceso a la universidad, en este caso los atributos centro de procedencia y nota de acceso a la universidad están relacionados). Además de relaciones entre atributos, otros atributos numéricos como son las notas de los estudiantes tampoco se han generado aleatoriamente para que el proyecto sea más realista. En este caso se ha optado por generar las notas siguiendo una distribución normal (también llamada distribución gaussiana) [Wikipedia 1, 2011], tal y como sucede en el mundo real, además, los alumnos están divididos en tres tipos de estudiantes, buenos, malos y normales, según sus notas de acceso a la universidad. Los estudiantes del tipo “buen estudiante” sacarán generalmente mejores notas que aquellos del tipo “mal estudiante”, y tardarán menos cursos académicos en acabar sus carreras. Debido a la gran cantidad de registros que son necesarios para poder hacer un trabajo de minería de datos con éxito, la opción de insertar estos registros manualmente uno a uno en la base de datos no era viable, por lo que se optó por crear un programa en el lenguaje de programación Java, cuya salida fueran los distintos scripts de inserción de datos (uno por cada tabla) para la base de datos, estos scripts se pueden consultar en el Anexo 2.

A continuación listamos los datos adquiridos:

- **Asignaturas**

  Cada asignatura está identificada por un número. Toda asignatura está relacionada con una titulación a la cual pertenece.

- **Titulaciones**

  Cada titulación que la universidad oferta está identificada por un número.

- **Centros de enseñanza**

  Cada centro de enseñanza está también identificado por un número.

- **Alumnos**

  Cada alumno está identificado por su id de alumno que es un valor numérico. Todo alumno está relacionado con un centro de enseñanza y con una titulación que es la que el alumno cursará en la universidad.

- **Fechas**

  Las fechas son extraídas en formato numérico con el formato caaaa, donde c es el número del cuatrimestre en cuestión (1 para el primer cuatrimestre, 2 para el segundo, y 3 para la convocatoria extraordinaria), y aaaa son los cuatro dígitos del año al que se refiere. Así, 21998 se referiría al segundo cuatrimestre del año 1998.

Los atributos específicos que serán útiles a la hora de hacer la minería de datos son:

- Identificador de alumno
- Nota de acceso a la universidad del alumno
- Centro de procedencia del alumno
- Identificador de la titulación
- Nota media de cada curso terminado en la universidad
- Tiempo (en cursos académicos) de aprobado de cada curso terminado en la universidad
- Identificador de la asignatura

- Fecha en la que se cursó la asignatura
- Nota obtenida en la asignatura

Las tablas en las que se recogen los datos necesarios para la minería de datos son:

- Fecha
- Asignatura
- Alumno
- Titulación
- Seguimiento Académico

En cuanto a la transformación de datos, se ha tenido que codificar el campo de centro de procedencia ya que contenía caracteres alfanuméricos, de tal forma que a cada centro se le ha asignado un valor numérico para facilitar su integración con los algoritmos de minería de datos.

2.2. Descripción de los Datos

Los datos se encuentran almacenados en un almacén de datos con esquema dimensional en estrella [Wikipedia 2, 2011]. En la figura 12 podemos ver un esquema relacional de esta base de datos, para generar esta figura se ha utilizado la herramienta proporcionada por Oracle llamada *Oracle SQL Developer Data Modeler* que sirve para generar modelos de las bases de datos Oracle.

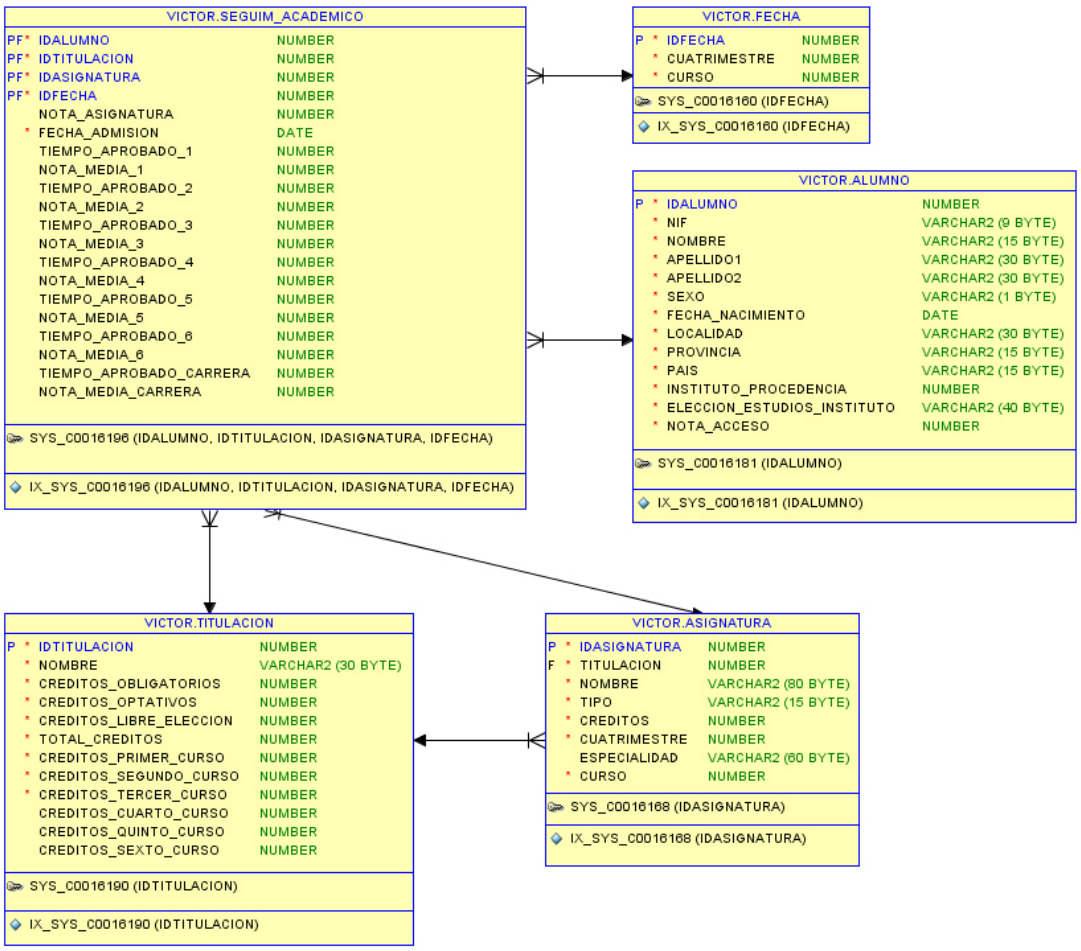


Figura 12 – Esquema relacional de la base de datos

En esta figura podemos ver claramente que el almacén de datos consta de cinco tablas: SEGUIM_ACADEMICO, FECHA, ALUMNO, ASIGNATURA y TITULACIÓN, a continuación describiremos cada una de ellas detallando cada uno de sus campos.

Tabla SEGUIM_ACADEMICO

Esta tabla es la tabla central del almacén de datos, también llamada "tabla de hechos", ya que es en esta tabla en la que se registra toda la información académica de cada alumno. Esta tabla al ser la tabla central tiene como clave primaria una combinación de cada una de las claves principales del resto de las tablas, llamadas tablas dimensionales, estas claves son: IDAlumno, IDTitulacion, IDAsignatura, e IDFecha. Estas claves son a la vez claves foráneas (*foreign keys*). Esta tabla tiene un total de 38.764 registros. Los campos de cada registro de esta tabla son:

- **IDAlumno.** Tipo numérico. Este campo es un número que identifica a cada alumno y que es único para cada alumno.
- **IDTitulacion.** Tipo numérico. Este campo es un número que identifica a cada titulación ofertada por la universidad y que es único para cada titulación.
- **IDAsignatura.** Tipo numérico. Este campo es un número que identifica a cada asignatura enseñada en la universidad y que es único para cada asignatura.
- **IDFecha.** Tipo numérico. Este campo es un número que identifica a cada fecha insertada en la tabla de fechas y que es único para cada fecha. El formato de este número es caaaa, donde c es el número del cuatrimestre que se quiere representar (1 para el primer cuatrimestre, 2 para el segundo y 3 para la convocatoria extraordinaria), y aaaa son los cuatro dígitos que representan al año en cuestión, por ejemplo, 12001 representaría al primer cuatrimestre del año 2001.
- **Nota_asignatura.** Tipo numérico. Este campo es un número que representa la nota obtenida por el alumno indicado en el campo IDAlumno para la asignatura representada en el campo IDAsignatura y para la fecha indicada en el campo IDFecha. Este valor tiene que estar comprendido entre 0 y 10, y tiene una precisión de dos dígitos decimales.
- **Fecha_admision.** Tipo fecha. Este campo indica la fecha en la que el alumno realizó su primera matricula en la universidad. El formato de la fecha es día-mes-año.
- **Tiempo_aprobado_1.** Tipo numérico. Es un número entero que representa el número de cursos académicos que un alumno ha necesitado para aprobar todas las asignaturas del primer curso de la titulación. Un valor nulo en este campo indica que el alumno aún no ha terminado este curso para la fecha indicada en el campo IDFecha.

- **Nota_media_1.** Tipo numérico. Es un número que representa la media de las notas de todas las asignaturas del primer curso y que se calcula una vez que el alumno ha superado todas las asignaturas del primer curso. Puede tener valor nulo si el alumno aún no ha terminado el primer curso para la fecha indicada en el campo IDFecha. El valor de estará comprendido entre 5 y 10, ya que solo se computan las asignaturas superadas, y tiene una precisión de dos decimales.
- **Tiempo_aprobado_2.** Tipo numérico. Es un número entero que representa el número de cursos académicos que un alumno ha necesitado para aprobar todas las asignaturas del segundo curso de la titulación. Un valor nulo en este campo indica que el alumno aún no ha terminado este curso para la fecha indicada en el campo IDFecha.
- **Nota_media_2.** Tipo numérico. Es un número que representa la media de las notas de todas las asignaturas del segundo curso y que se calcula una vez que el alumno ha superado todas las asignaturas del segundo curso. Puede tener valor nulo si el alumno aún no ha terminado el segundo curso para la fecha indicada en el campo IDFecha. El valor de estará comprendido entre 5 y 10, ya que solo se computan las asignaturas superadas, y tiene una precisión de dos decimales.
- **Tiempo_aprobado_3.** Tipo numérico. Es un número entero que representa el número de cursos académicos que un alumno ha necesitado para aprobar todas las asignaturas del tercer curso de la titulación. Un valor nulo en este campo indica que el alumno aún no ha terminado este curso para la fecha indicada en el campo IDFecha.
- **Nota_media_3.** Tipo numérico. Es un número que representa la media de las notas de todas las asignaturas del tercer curso y que se calcula una vez que el alumno ha superado todas las asignaturas del tercer curso. Puede tener valor nulo si el alumno aún no ha terminado el tercer curso para la fecha indicada en el campo IDFecha. El valor de estará comprendido entre 5 y 10, ya que solo se computan las asignaturas superadas, y tiene una precisión de dos decimales.
- **Tiempo_aprobado_4.** Tipo numérico. Es un número entero que representa el número de cursos académicos que un alumno ha necesitado para aprobar todas las asignaturas del cuarto curso de la titulación. Un valor nulo en este campo indica que el alumno aún no ha terminado este curso para la fecha indicada en el campo IDFecha, o bien que la titulación cursada por el alumno no tiene cuatro cursos.

- **Nota_media_4.** Tipo numérico. Es un número que representa la media de las notas de todas las asignaturas del cuarto curso y que se calcula una vez que el alumno ha superado todas las asignaturas del cuarto curso. Puede tener valor nulo si el alumno aún no ha terminado el cuarto curso para la fecha indicada en el campo IDFecha, o su titulación tiene menos de cuatro cursos. El valor de estará comprendido entre 5 y 10, ya que solo se computan las asignaturas superadas, y tiene una precisión de dos decimales.
- **Tiempo_aprobado_5.** Tipo numérico. Es un número entero que representa el número de cursos académicos que un alumno ha necesitado para aprobar todas las asignaturas del quinto curso de la titulación. Un valor nulo en este campo indica que el alumno aún no ha terminado este curso para la fecha indicada en el campo IDFecha, o bien que la titulación cursada por el alumno no tiene cinco cursos.
- **Nota_media_5.** Tipo numérico. Es un número que representa la media de las notas de todas las asignaturas del quinto curso y que se calcula una vez que el alumno ha superado todas las asignaturas del quinto curso. Puede tener valor nulo si el alumno aún no ha terminado el quinto curso para la fecha indicada en el campo IDFecha, o su titulación tiene menos de cinco cursos. El valor de estará comprendido entre 5 y 10, ya que solo se computan las asignaturas superadas, y tiene una precisión de dos decimales.
- **Tiempo_aprobado_6.** Tipo numérico. Es un número entero que representa el número de cursos académicos que un alumno ha necesitado para aprobar todas las asignaturas del sexto curso de la titulación. Un valor nulo en este campo indica que el alumno aún no ha terminado este curso para la fecha indicada en el campo IDFecha, o bien que la titulación cursada por el alumno no tiene seis cursos.
- **Nota_media_6.** Tipo numérico. Es un número que representa la media de las notas de todas las asignaturas del sexto curso y que se calcula una vez que el alumno ha superado todas las asignaturas del sexto curso. Puede tener valor nulo si el alumno aún no ha terminado el sexto curso para la fecha indicada en el campo IDFecha, o su titulación tiene menos de seis cursos. El valor de estará comprendido entre 5 y 10, ya que solo se computan las asignaturas superadas, y tiene una precisión de dos decimales.
- **Tiempo_aprobado_carrera.** Tipo numérico. Es un número entero que representa el número de cursos académicos que un alumno ha necesitado para terminar todas las

asignaturas de la titulación. Un valor nulo en este campo indica que el alumno aún no ha terminado la titulación para la fecha indicada en el campo IDFecha.

- **Nota_media_carrera.** Tipo numérico. Es un número que representa la media de las notas de todas las asignaturas de la titulación y que se calcula una vez que el alumno ha terminado su titulación. Puede tener valor nulo si el alumno aún no ha terminado la titulación para la fecha indicada en el campo IDFecha. El valor de estará comprendido entre 5 y 10, ya que solo se computan las asignaturas superadas, y tiene una precisión de dos decimales.

Tabla FECHA

Esta tabla contiene la información acerca de las fechas. Su clave primaria es el campo IDFecha, y tiene un total de 45 registros que incluyen cada uno de los cuatrimestres y convocatorias extraordinarias desde el curso 1997 hasta el 2011. Los campos de cada registro de esta tabla son:

- **IDFecha.** Tipo numérico. Este campo es un número que identifica a cada fecha insertada en la tabla de fechas y que es único para cada fecha. El formato de este número es caaaa, donde c es el número del cuatrimestre que se quiere representar (1 para el primer cuatrimestre, 2 para el segundo y 3 para la convocatoria extraordinaria), y aaaa son los cuatro dígitos que representan al año en cuestión, por ejemplo, 12001 representaría al primer cuatrimestre del año 2001.
- **Cuatrimestre.** Tipo numérico. Identifica el cuatrimestre al que se refiere la fecha, sus valores posibles son 1 (primer cuatrimestre), 2 (segundo cuatrimestre) o 3 (convocatoria extraordinaria).
- **Curso.** Tipo numérico. Identifica el año al que se refiera la fecha. Los valores de este campo en la base de datos que se utilizará para la minería de datos van desde el número 1997 hasta el 2011.

Tabla ALUMNO

Esta tabla recoge toda la información personal y académica (previa a la universidad) necesaria acerca de los alumnos que cursan sus estudios en la universidad. Tiene como

clave primaria el campo IDAlumno y consta de un total de 4.000 registros (uno por cada alumno). Los campos de cada registro de esta tabla son:

- **IDAlumno.** Tipo numérico. Este campo es un número que identifica a cada alumno y que es único para cada alumno.
- **NIF.** Tipo alfanumérico. Representa el número de identificación fiscal de cada alumno, consta de ocho números y una letra al final.
- **Nombre.** Tipo alfanumérico. Es el nombre del alumno, este campo tiene una longitud máxima de 15 caracteres.
- **Apellido1.** Tipo alfanumérico. Representa el primer apellido del alumno. Tiene una longitud máxima de 30 caracteres.
- **Apellido2.** Tipo alfanumérico. Representa el segundo apellido del alumno. Tiene una longitud máxima de 30 caracteres.
- **Sexo.** Tipo alfanumérico. Representa el sexo del alumno. Puede tener el valor 'M' (mujer) o 'H' (hombre).
- **Fecha_nacimiento.** Tipo fecha. Indica la fecha de nacimiento del alumno. Tiene el formato día-mes-año.
- **Localidad.** Tipo alfanumérico. Representa la localidad de residencia del alumno. Tiene una longitud máxima de 30 caracteres.
- **Provincia.** Tipo alfanumérico. Representa la provincia en la que reside el alumno. Tiene una longitud máxima de 15 caracteres.
- **Pais.** Tipo alfanumérico. Representa el país de residencia del alumno. Tiene una longitud máxima de 15 caracteres.
- **Instituto_procedencia.** Tipo numérico. Es un número que representa el centro en el que el alumno cursó sus estudios previos a la universidad. Inicialmente este campo era de tipo alfanumérico y contenía el nombre completo del centro de estudios, pero se optó por codificarlo para facilitar las labores de la minería de datos.
- **Eleccion_estudios_instituto.** Tipo alfanumérico. Este campo contiene el plan de estudios elegido por el alumno durante sus estudios pre-universitarios. Tiene cuatro valores posibles: tecnológico, biosanitario, humanidades y ciencias sociales. Tiene una longitud máxima de 40 caracteres.

- **Nota_acceso.** Tipo numérico. Representa la nota final con la que el alumno accede a la universidad tras superar las pruebas de acceso a la universidad. Se trata de un valor entre 5 y 10 con una precisión de dos cifras decimales.

Tabla ASIGNATURA

Esta tabla contiene la información relativa a las asignaturas pertenecientes a una titulación. Esta tabla tiene una clave primaria que es el IDAsignatura y también una clave foránea, Titulacion (IDTitulacion en la tabla TITULACION), y contiene un total de 35 registros. Los campos de cada registro de esta tabla son:

- **IDAsignatura.** Tipo numérico. Es un número que identifica a cada asignatura. Al tratarse de una clave primaria, es único para cada registro.
- **Titulacion.** Tipo numérico. Es un número que identifica la titulación a la que pertenece la asignatura y es clave foránea ya que referencia a la clave primaria de la tabla TITULACION.
- **Nombre.** Tipo alfanumérico. Este campo representa el nombre de la asignatura y tiene una longitud máxima de 80 caracteres.
- **Tipo.** Tipo alfanumérico. Se refiere al tipo de asignatura que puede ser o bien "Obligatoria" u "Optativa". Tiene una longitud máxima de 15 caracteres.
- **Creditos.** Tipo numérico. Un número que representa el número de créditos que tiene la asignatura.
- **Cuatrimestre.** Tipo numérico. Es un número que representa el cuatrimestre al que pertenece la asignatura, puede tener dos valores: 1 (primer cuatrimestre) y 2 (segundo cuatrimestre).
- **Especialidad.** Tipo alfanumérico. Este campo representa la especialidad dentro de la titulación a la que pertenece una asignatura, si la asignatura es común a todas las especialidades en este campo aparecerá el valor "común", si no aparecerá el nombre de la especialidad. Tiene un máximo de 60 caracteres.
- **Curso.** Tipo numérico. Este campo contiene un número que representa el curso al que pertenece la asignatura.

Tabla TITULACION

Esta tabla contiene toda la información relativa a las titulaciones que se ofrecen en la universidad. Esta tabla tiene una clave primaria que es el IDTitulacion, y contiene un total de 3 registros. Los campos de cada registro de esta tabla son:

- **IDTitulacion.** Tipo numérico. Este campo es un número que identifica a cada titulación de forma única.
- **Nombre.** Tipo alfanumérico. Este campo representa el nombre de la titulación, tiene una longitud máxima de 30 caracteres.
- **Creditos_obligatorios.** Tipo numérico. Indica el número de créditos obligatorios en total de la titulación.
- **Creditos_optativos.** Tipo numérico. Indica el número de créditos optativos en total de la titulación.
- **Creditos_libre_eleccion.** Tipo numérico. Indica el número de créditos de libre elección en total de la titulación.
- **Total_creditos.** Tipo numérico. Indica el número total de créditos en la titulación, es decir la suma de los campos "Creditos_obligatorios", "Creditos_optativos" y "Creditos_libre_eleccion".
- **Creditos_primer_curso.** Tipo numérico. Indica la suma de créditos obligatorios y optativos del primer curso de la titulación.
- **Creditos_segundo_curso.** Tipo numérico. Indica la suma de créditos obligatorios y optativos del segundo curso de la titulación.
- **Creditos_tercer_curso.** Tipo numérico. Indica la suma de créditos obligatorios y optativos del tercer curso de la titulación.
- **Creditos_cuarto_curso.** Tipo numérico. Indica la suma de créditos obligatorios y optativos del cuarto curso de la titulación. Un valor nulo en este campo indica que la titulación no tiene cuarto curso.
- **Creditos_quinto_curso.** Tipo numérico. Indica la suma de créditos obligatorios y optativos del quinto curso de la titulación. Un valor nulo en este campo indica que la titulación no tiene quinto curso.

- **Creditos_sexto_curso.** Tipo numérico. Indica la suma de créditos obligatorios y optativos del sexto curso de la titulación. Un valor nulo en este campo indica que la titulación no tiene sexto curso.

2.3. Exploración de los Datos

Una vez que se han descrito los datos, se procede a explorarlos, esto implica aplicar pruebas estadísticas básicas que revelarán propiedades de los datos, y crear tablas de frecuencia y gráficos de distribución de los datos. Este informe sirve principalmente para determinar la consistencia y completitud de los datos. Mediante consultas SQL que se pueden encontrar en el Anexo 2, se han obtenido los datos necesarios para realizar las gráficas que se explican a continuación.

El gráfico 1 muestra la distribución de las notas de acceso a la universidad obtenidas por los alumnos, en este gráfico se aprecia la distribución normal que siguen las notas de acceso a la universidad

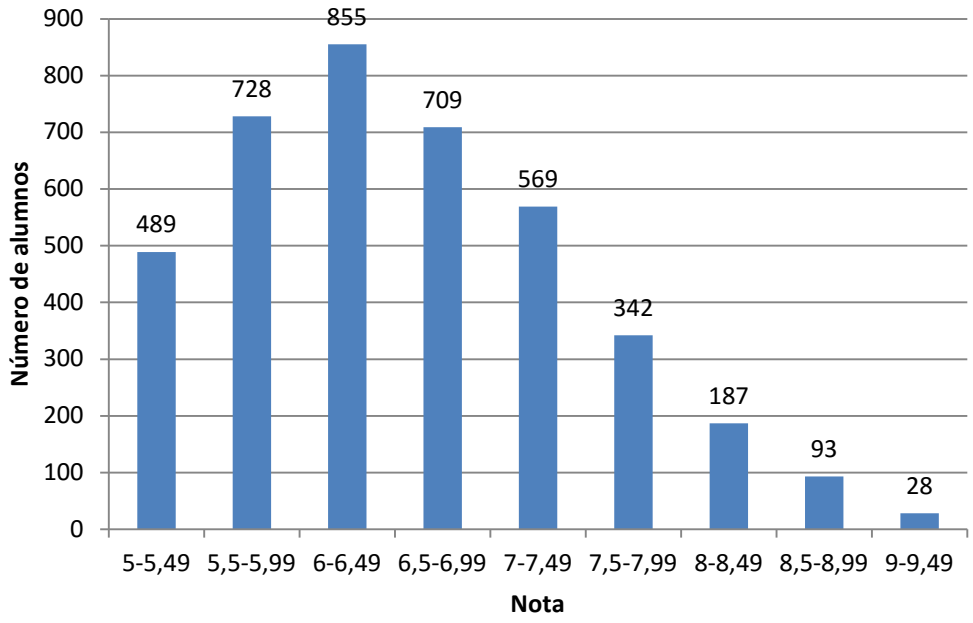


Gráfico 1: Número de alumnos por nota de acceso a la universidad

El gráfico 2 muestra el porcentaje de alumnos que han obtenido cada uno de los intervalos de las notas:

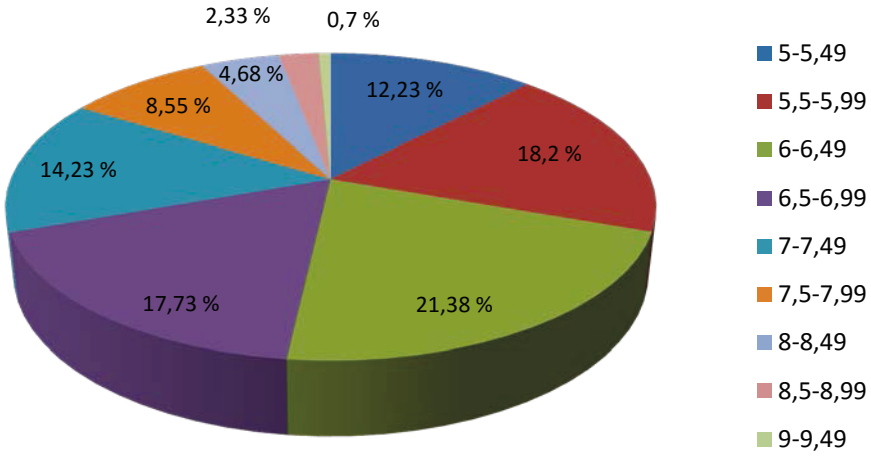


Gráfico 2: Porcentaje de alumnos por nota de acceso

En el gráfico 3 se muestra la distribución de las notas de acceso ordenadas por provincia.

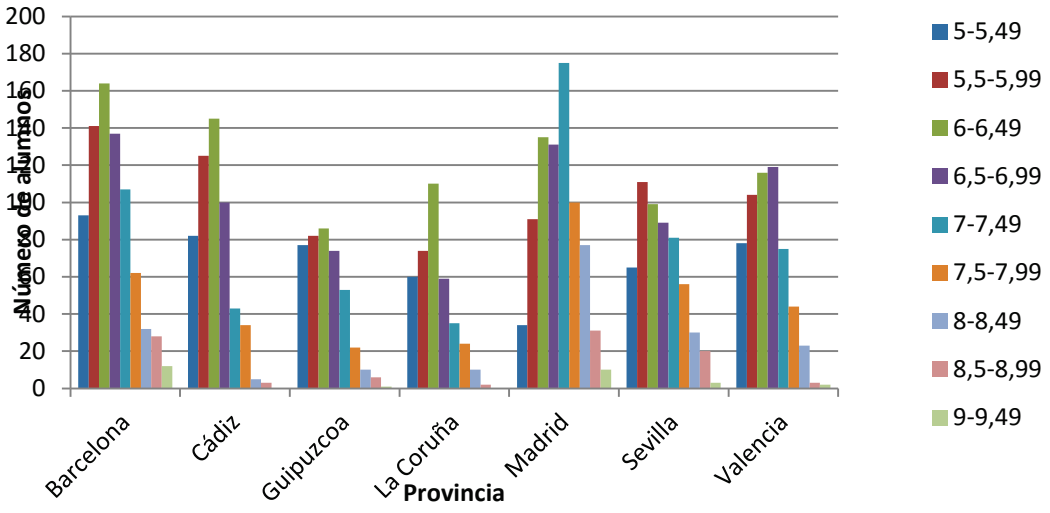


Gráfico 3: Notas de acceso por provincia

El gráfico 4 muestra el porcentaje de alumnos dentro de cada provincia que ha obtenido una determinada nota, así por ejemplo, si observamos el intervalo de notas entre 7 y 7,49 para la provincia de Madrid, se puede observar que un 22% de los alumnos de esta provincia han obtenido esta calificación.

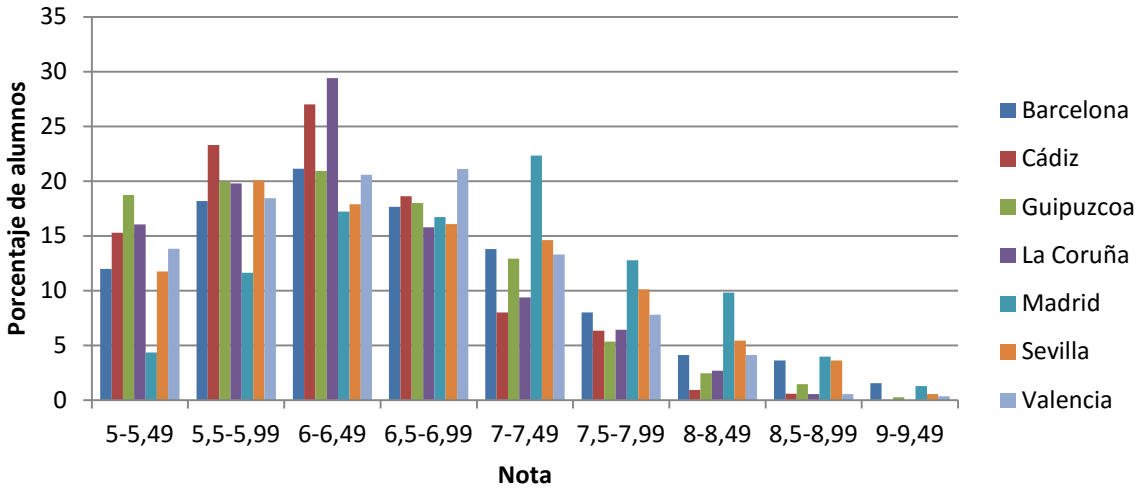


Gráfico 4: Porcentaje de alumnos de cada provincia ordenado por notas

El gráfico 5 representa la nota de acceso media de todos los alumnos de cada centro de enseñanza.

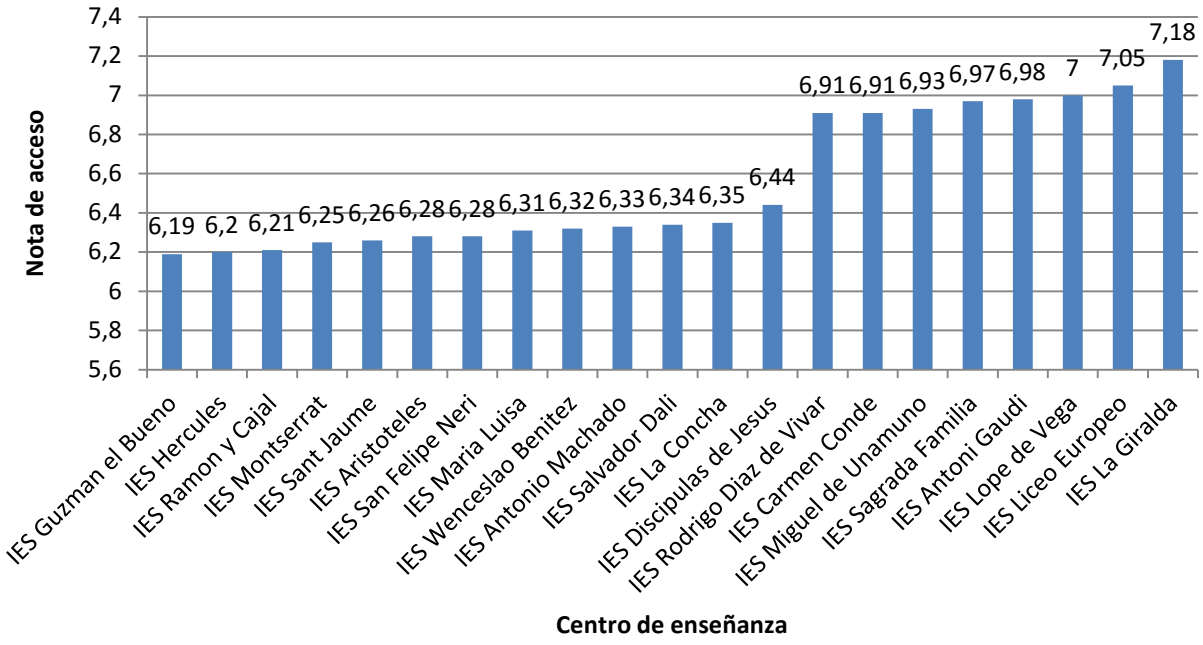


Gráfico 5: Nota de acceso media por centro de enseñanza

El gráfico 6 muestra la nota de acceso media de todos los alumnos por provincia de residencia.

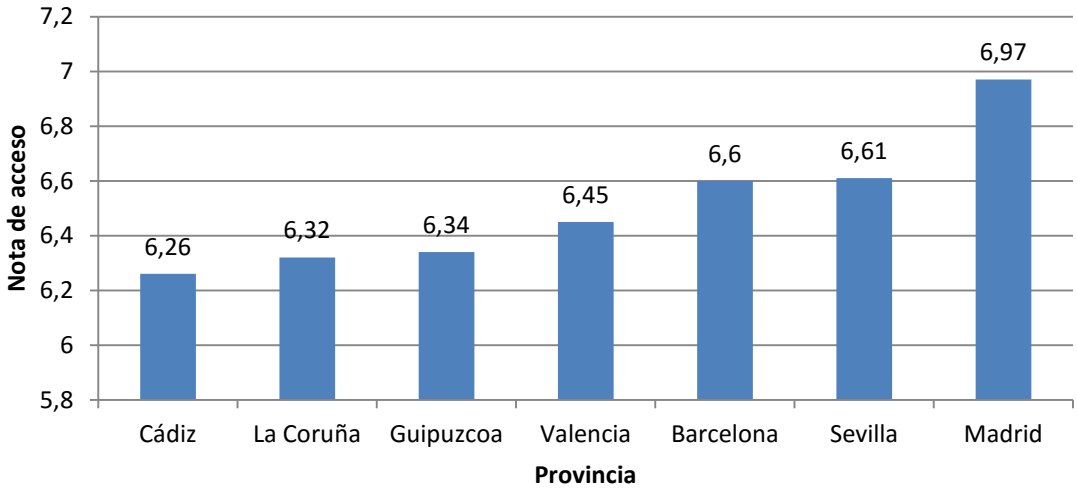


Gráfico 6: Nota de acceso media por provincia

El gráfico 7 muestra la nota media de cada curso para cada una de las titulaciones y la nota media al acabar cada titulación.

Nota: La titulación de ingeniería en informática consta de 5 cursos, la de derecho y ADE consta de 6 cursos y la de turismo consta de 3 cursos.

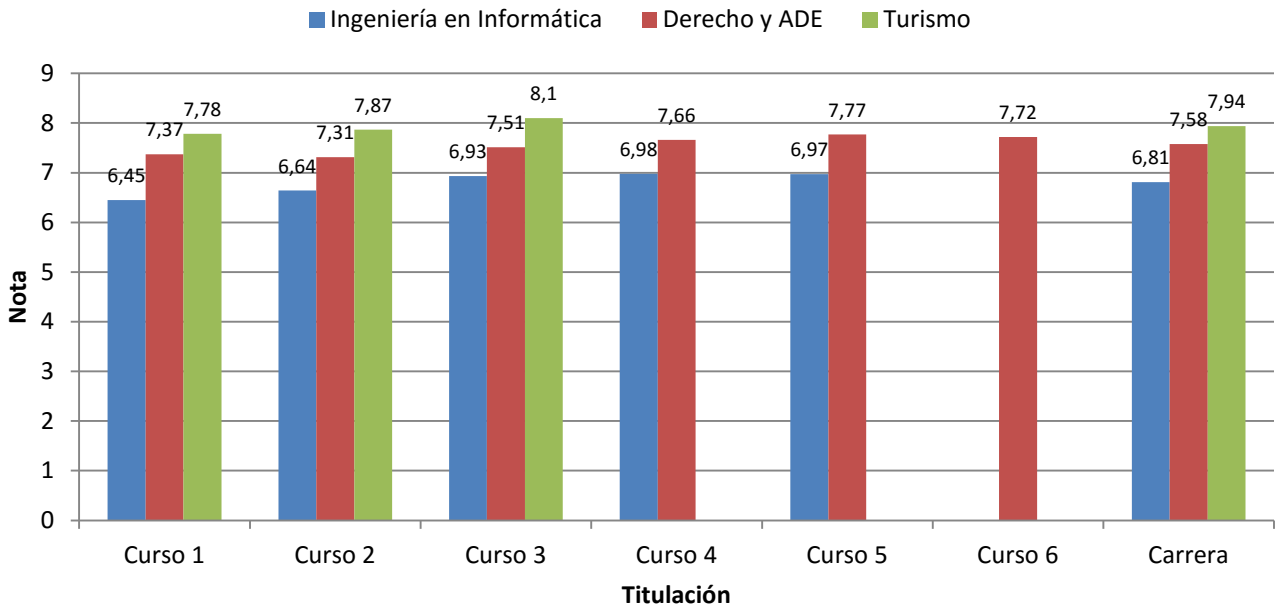


Gráfico 7: Notas medias de cada curso por titulación

El gráfico 8 muestra el número de años académicos de media que emplean los alumnos para completar cada uno de los cursos de la titulación, y la titulación al completo.

Nota: La titulación de ingeniería en informática consta de 5 cursos, la de derecho y ADE consta de 6 cursos y la de turismo consta de 3 cursos.

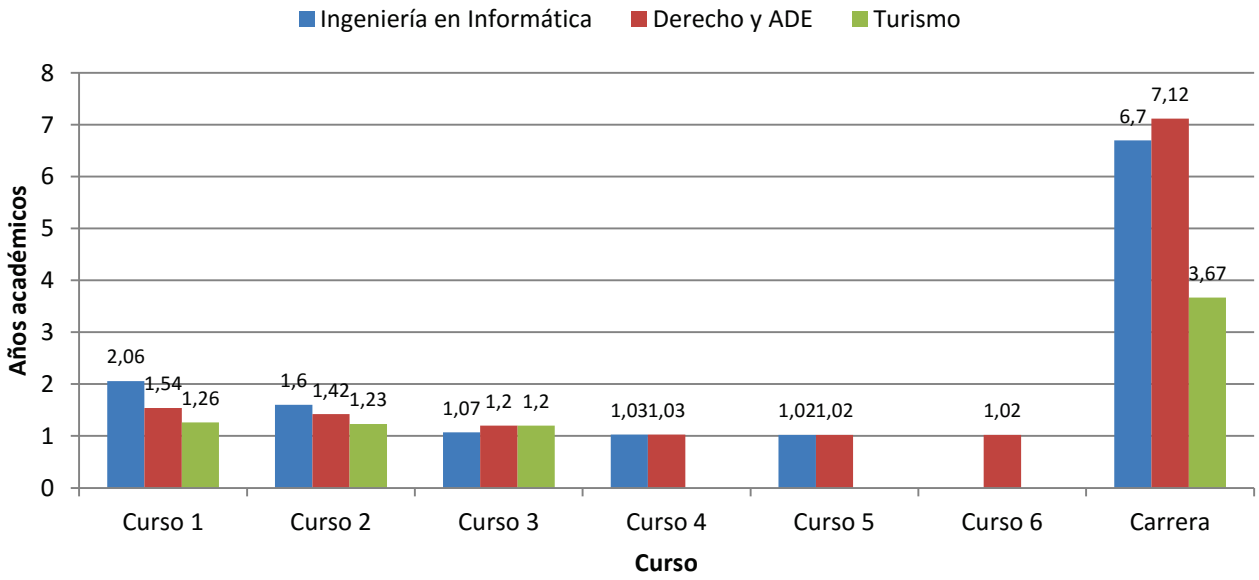


Gráfico 8: Número de años académicos empleados en terminar cada curso por titulación

Los gráficos 9, 10 y 11 nos muestran el porcentaje de alumnos de ingeniería en informática, derecho y ADE y turismo respectivamente, con respecto a la nota media del primer curso.

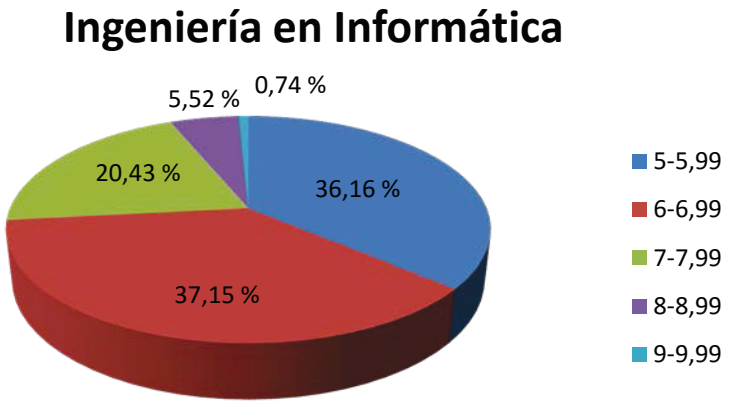


Gráfico 9: Porcentaje de alumnos de ingeniería informática con respecto a la nota media del primer curso

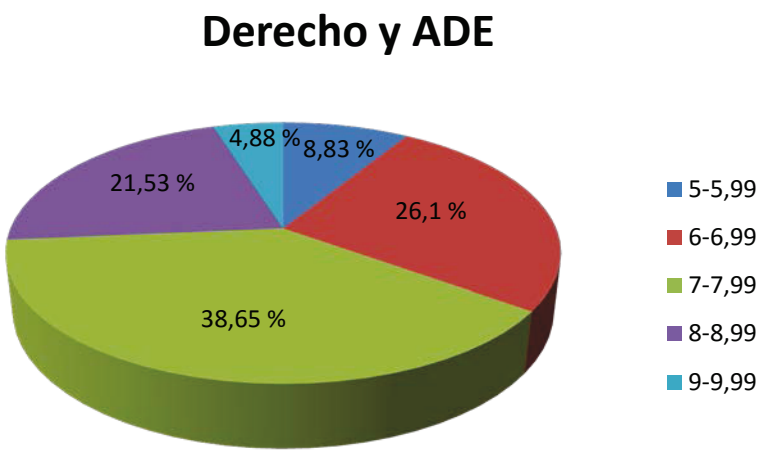


Gráfico 10: Porcentaje de alumnos de derecho y ADE con respecto a la nota media del primer curso

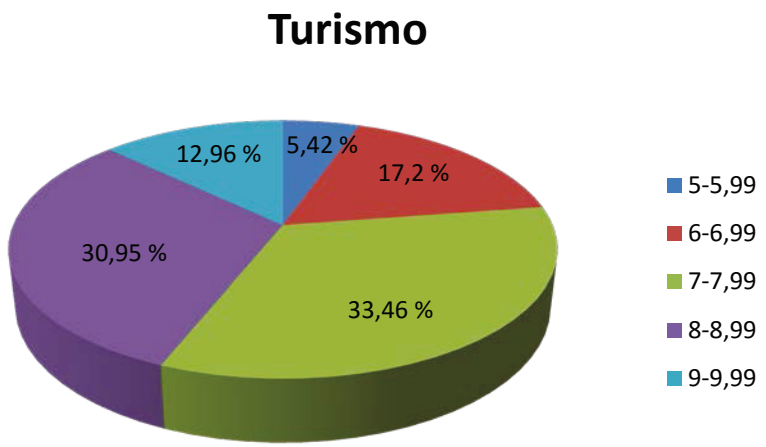


Gráfico 11: Porcentaje de alumnos de turismo con respecto a la nota media del primer curso

Mediante consultas SQL a la base de datos, obtenemos las asignaturas dentro de cada titulación cuya nota media de los alumnos son más bajas, es decir la asignatura más problemática dentro de cada una de las titulaciones. Estas asignaturas son:

- Para ingeniería en informática: Redes de ordenadores II, con una nota media de 5,77.
- Para derecho y ADE: Análisis de valores, con una nota media de 5,92.
- Para turismo: Recursos territoriales turísticos, con una nota media de 6,79.

A continuación, en los gráficos 12, 13 y 14, se muestra el porcentaje de alumnos con respecto a la nota obtenida en cada una de las tres asignaturas.

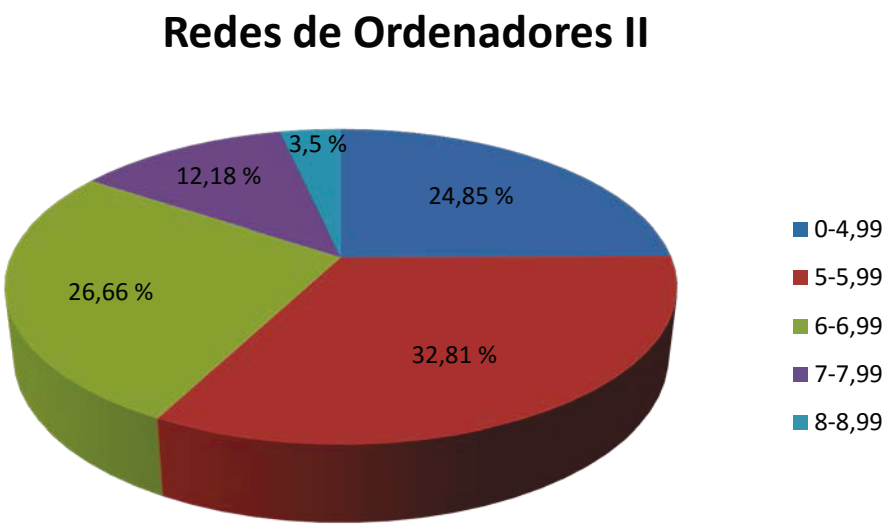


Gráfico 12: Porcentaje de alumnos de redes de ordenadores II con respecto a la nota obtenida

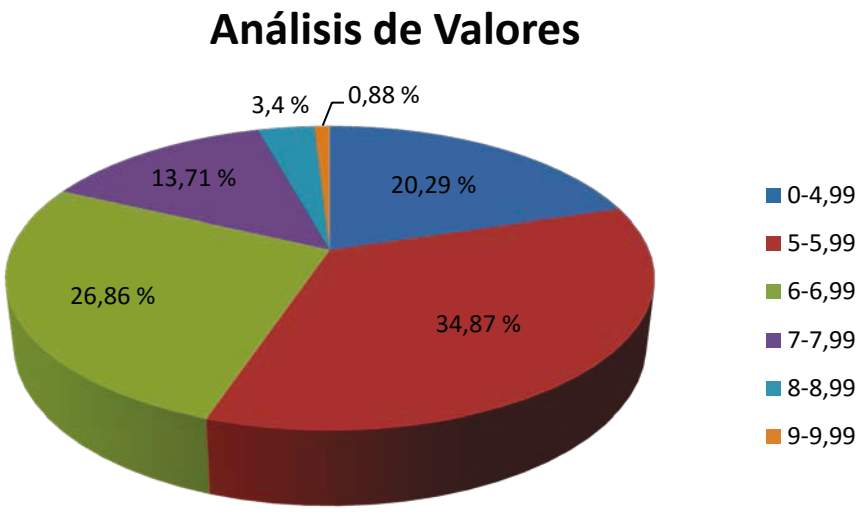


Gráfico 13: Porcentaje de alumnos análisis de valores con respecto a la nota obtenida

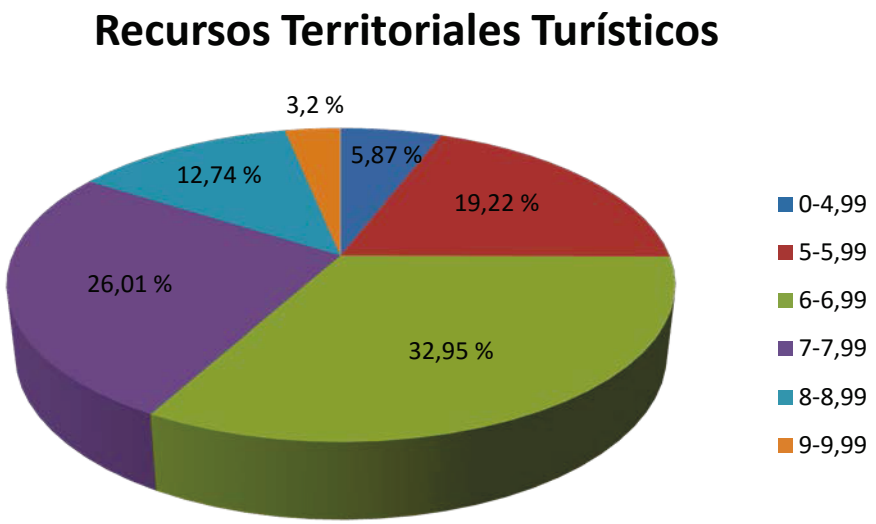


Gráfico 14: Porcentaje de alumnos de recursos territoriales turísticos con respecto a la nota obtenida

2.4. Verificar la Calidad de los Datos

Después de hacer la exploración inicial de los datos se puede afirmar que estos son completos. Los datos cubren los casos requeridos para la obtención de los resultados necesarios para poder cumplir los objetivos del proyecto. Los datos no contienen errores, ya que son datos generados automáticamente por el programa que crea los scripts de inserción de datos. Tampoco se encuentran valores fuera de rango, ya que los datos son controlados desde el mismo programa, por lo que no hay riesgo de ruido en el proceso de la minería de datos. En cuanto a los valores nulos, solo los encontramos en la tabla SEGUIM_ACADEMICO. En esta tabla hay un gran número de campos con valor nulo, concretamente en los campos que se refieren a los tiempos de aprobado y nota media de cada uno de los cursos de la titulación, ya que si un alumno aún no estuviera matriculado en un curso los valores para el tiempo de aprobado y la nota media de este curso serían nulos. Una posible solución a la hora de hacer minería de datos con estos campos sería ignorar aquellos que contengan valores nulos.

3. Preparación de los Datos

En esta fase de la metodología se trata de preparar los datos para adecuarlos a las técnicas de minería de datos que se van a emplear sobre ellos. Esto implica seleccionar el subconjunto de datos que se va a utilizar, limpiarlos para mejorar su calidad, añadir nuevos datos a partir de los existentes y darles el formato requerido por la herramienta de modelado.

3.1. Seleccionar los Datos

En términos de registros, se van a utilizar todos los registros dentro de cada tabla que compone la base de datos, ya que al ser ésta una base de datos específicamente creada para este proyecto, el número de registros que se han insertado ha sido elegido a propósito. Sin embargo, hay campos dentro de estos registros que no son necesarios para nuestros objetivos de minería de datos, por lo que se puede prescindir de algunos de ellos.

Los campos seleccionados para el análisis son los siguientes:

- **Tabla FECHA**
  - IDFecha
- **Tabla ASIGNATURA**
  - IDAsignatura
- **Tabla ALUMNO**
  - IDAlumno
  - Localidad
  - Provincia
  - Instituto_procedencia
  - Nota_acceso
- **Tabla TITULACION**
  - IDTitulacion
- **Tabla SEGUIM_ACADEMICO**
  - IDAlumno
  - IDTitulacion

- IDAsignatura
- IDFecha
- Nota_asignatura
- Tiempo_aprobado_1
- Nota_media_1
- Tiempo_aprobado_2
- Nota_media_2
- Tiempo_aprobado_3
- Nota_media_3
- Tiempo_aprobado_4
- Nota_media_4
- Tiempo_aprobado_5
- Nota_media_5
- Tiempo_aprobado_6
- Nota_media_6
- Tiempo_aprobado_carrera
- Nota_media_carrera

El motivo para la inclusión o exclusión de algunos campos es, como se ha mencionado antes, la importancia de dichos campos en relación con los objetivos de la minería de datos que se definieron en la fase 1 (comprensión del negocio) de la metodología.

3.2. Limpiar los Datos

La base de datos con la que se cuenta para el proyecto contiene toda la información necesaria para poder cumplir los objetivos de la minería de datos, además, estos datos al haber sido introducidos ex profeso para el caso práctico que se presenta, son datos limpios y por lo tanto no hay necesidad de hacer una limpieza más profunda sobre ellos.

Tampoco tenemos campos en los que falten valores, más allá de los valores nulos que aparecen cuando la información que se quiere representar no existe, y por lo tanto no se consideran como datos faltantes, por lo que no es necesario realizar ningún tipo de estimación de valores faltantes. Estos valores nulos se tratarán a la hora de hacer la minería de datos simplemente ignorándolos ya que no aportan ninguna información adicional al estudio.

Para generar el modelo relacionado con el primer objetivo de la minería de datos, es decir, el de la predicción del tiempo que tardará un alumno en terminar la carrera, será necesario utilizar un filtro que nos proporciona la herramienta Oracle Data Mining y de este modo sólo seleccionar aquellas filas cuyo campo "tiempo_aprobado_carrera" tenga un valor distinto de nulo, ya que en este caso solo nos servirán los datos de aquellos alumnos que ya han terminado la titulación. En la figura 13 se puede ver el sistema que se utiliza en Oracle Data Mining para filtrar filas.

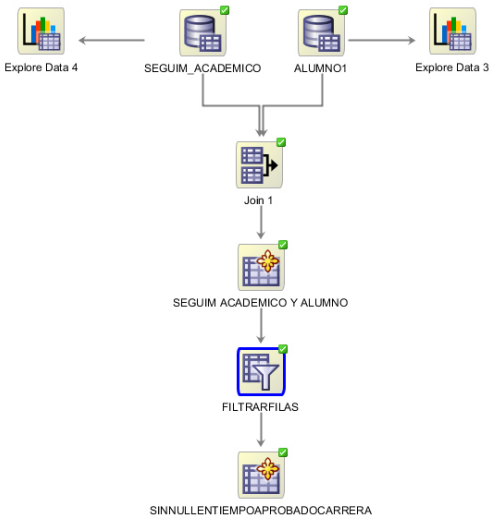


Figura 13 – Filtrado de filas

Y en la figura 14 podemos observar los parámetros que se han utilizado en el filtro.

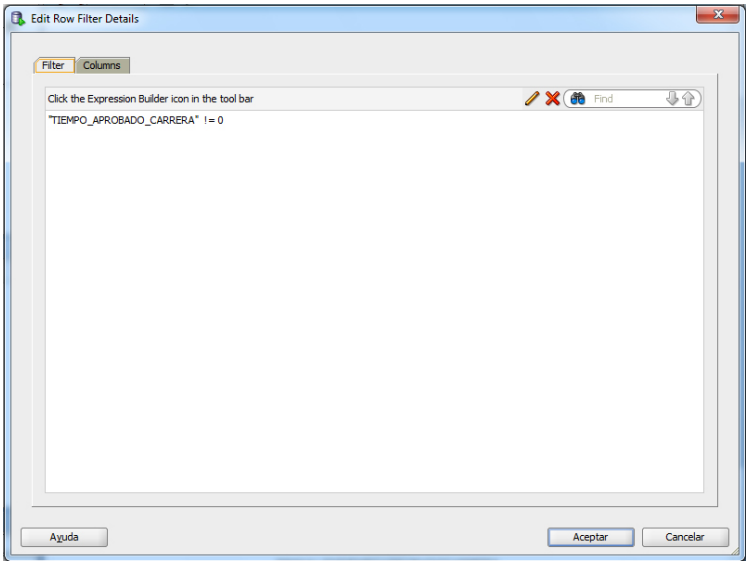


Figura 14 – Parámetros empleados para filtrar filas

3.3. Construir los Datos

Atributos derivados

En este apartado sólo se puede destacar la transformación del campo *Instituto_procedencia* de la tabla ALUMNO. Dicha transformación consistió en codificar numéricamente los valores del campo que inicialmente contenía caracteres alfanuméricos con el nombre del centro de procedencia del alumno.

Además de este campo, se crearon los campos referentes al tiempo que ha tardado un alumno en terminar cada curso (Tiempo_aprobado_1, Tiempo_aprobado_2, Tiempo_aprobado_3, Tiempo_aprobado_4, Tiempo_aprobado_5 y Tiempo_aprobado_6) y la carrera en total (Tiempo_aprobado_carrera) de la tabla SEGUIM_ACADEMICO, a partir de dos campos que se suprimieron y que contenían la fecha de inicio y de final de cada uno de los cursos y de la carrera en total. De esta forma se hace más útil la información ya que el programa de minería de datos necesita que estos campos sean numéricos y no fechas para poder generar el modelo.

Registros generados

Aparte de estas dos operaciones, no ha sido necesario generar nuevos atributos ni integrar nuevos registros a la base de datos ya que ésta está completa y ha sido creada específicamente para su uso en este proyecto.

3.4. Integrar los Datos

No ha sido necesaria la creación de nuevas estructuras (campos, registros, etc.), ni la fusión entre distintas tablas de la base de datos, ya que el programa Oracle Data Mining se encarga de realizar estas tareas automáticamente sin que el usuario tenga que crear nuevas tablas, registros o campos manualmente.

3.5. Formateo de los Datos

El campo con la información referente al centro de estudios de procedencia de los alumnos ha sido codificado con valores numéricos ya que la herramienta de minería de datos exige que los datos a estudiar sean numéricos. Inicialmente este campo contenía el nombre de cada centro escrito con caracteres alfanuméricos pero después se optó por asignar un número a cada centro. Los códigos para cada uno de los centros existentes en la base de datos quedaron así:

- 1 → I.E.S. Miguel de Unamuno
- 2 → I.E.S. Carmen Conde
- 3 → I.E.S. Liceo Europeo
- 4 → I.E.S. Sagrada Familia
- 5 → I.E.S. Sant Jaume
- 6 → I.E.S. Montserrat
- 7 → I.E.S. San Felipe Neri
- 8 → I.E.S. Wenceslao Benítez
- 9 → I.E.S. Guzmán el Bueno
- 10 → I.E.S. Hércules
- 11 → I.E.S. Discípulas de Jesús
- 12 → I.E.S. Rodrigo Díaz de Vivar
- 13 → I.E.S. Ramón y Cajal
- 14 → I.E.S. Aristóteles
- 15 → I.E.S. La Giralda
- 16 → I.E.S. María Luisa
- 17 → I.E.S. Antonio Machado

- 18 → I.E.S. Lope de Vega
- 19 → I.E.S. Antoni Gaudí
- 20 → I.E.S. La Concha
- 21 → I.E.S. Salvador Dalí

No es necesario cambiar el orden de ningún campo dentro de los registros, ni tampoco la reordenación de los registros dentro de las tablas. Tampoco es necesario cambiar el formato de ninguno de los campos que se van a utilizar para la minería de datos ya que el formato actual es admitido por la herramienta Oracle Data Mining.

4. Modelado

En esta fase de la metodología se escogerá la técnica (o técnicas) más apropiadas para los objetivos marcados de la minería de datos. A continuación y una vez realizado un plan de prueba para los modelos escogidos, se procederá a aplicar dichas técnicas sobre los datos para generar el modelo y por último se tendrá que evaluar si dicho modelo ha cumplido los criterios de éxito o no.

4.1. Escoger la Técnica de Modelado

Debido a que se va a utilizar el software *Oracle Data Mining* para realizar la minería de datos, deberemos utilizar alguna de las técnicas de modelado que nos ofrece esta herramienta de acuerdo con los objetivos de nuestro proyecto que están reflejados en el apartado 1.3 (Objetivos de la minería de datos).

De los modelos que nos ofrece *Oracle Data Mining*, el que mejor se adapta a nuestros objetivos sería un modelo de regresión, puesto que los problemas que queremos resolver son problemas de predicción y los campos que se quieren predecir contienen valores continuos.

4.2. Generar el Plan de Prueba

El procedimiento que se empleará para probar la calidad y validez del modelo será el de utilizar las medidas del error cuadrático medio (*root mean squared error*), el error absoluto medio (*mean absolute error*) y la "confianza predictiva" (*predictive confidence*). Estas medidas de error las calcula automáticamente Oracle Data Mining al ejecutar los modelos de regresión. Para entender mejor estos indicadores vamos a describirlos a continuación.

- El error cuadrático medio (RMSE) es la manera más habitual de evaluar un modelo de regresión. Mediante esta medida se calculan las diferencias entre los valores pronosticados por el modelo o un estimador y los valores reales a partir de los cuales se ha creado el modelo. El error cuadrático medio es una medida de la media de los

cuadrados de los errores. Por error se entiende la diferencia entre el valor estimado y el valor real. El error cuadrático medio se calcula de la siguiente manera:

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2}$$

- El error absoluto medio (MAE) es otra forma de evaluar la calidad en los modelos de regresión. Al igual que el error cuadrático medio, esta medida también sirve para calcular la diferencia entre las predicciones hechas por un estimador y los valores reales. La diferencia entre ambas surge del principal problema que tiene calcular el error cuadrático medio, y es que al elevar al cuadrado la diferencia se tiende a dar demasiado peso a los errores más extremos, afectando al resultado final, utilizando el error absoluto medio se puede limitar este problema. La fórmula para calcular el error absoluto medio es la siguiente:

$$MAE = \frac{1}{n}\sum_{j=1}^{n}\left|y_j - \hat{y}_j\right|$$

- La confianza predictiva (*predictive confidence*) es una medida de cómo el modelo mejora una predicción en comparación con el azar. Si el modelo fuera ingenuo (*naive*) y no realizara ningún tipo de análisis, simplemente utilizaría la media. La confianza predictiva representa el porcentaje de incremento de mejoría del modelo sobre el modelo ingenuo. Por ejemplo, una confianza predictiva de un 50% indicaría que el modelo es 50% mejor que un modelo ingenuo.

Oracle Data Mining ofrece al usuario la opción de dividir los datos en dos grupos automáticamente antes de generar el modelo: por un lado está el conjunto de datos que se van a utilizar para generar el modelo, llamados datos de entrenamiento, y un segundo conjunto de datos que se empleará para realizar las pruebas y medir la calidad del modelo, llamados datos de prueba o de evaluación. Normalmente se suele utilizar un 60% de los datos para los datos de entrenamiento y el 40% restante para los datos de

prueba, pero esta cantidad se puede modificar desde el propio programa para utilizar el porcentaje que el usuario quiera.

4.3. Construir el Modelo

A continuación se procederá a ejecutar el modelo elegido sobre los datos de entrenamiento. En este apartado se describirán los ajustes de parámetros del modelo que se eligen en la herramienta de minería de datos, así como la salida de dicho modelo y su descripción.

Ajustes de parámetros

Puesto que se han definido tres objetivos para la minería de datos, vamos a dividir esta sección en tres partes, una por cada objetivo, ya que los parámetros para el modelo variarán según el objetivo que deseamos conseguir.

- **Objetivo 1**. Predicción del tiempo que un alumno tardará en acabar la carrera.

En este caso el campo objetivo, es decir aquel sobre el cual queremos hacer la predicción es "tiempo_aprobado_carrera" y el *case ID* será el "idalumno". En cuanto a los parámetros empleados para el algoritmo de GLM, se utilizan los parámetros que vienen por defecto en Oracle Data Mining, que se pueden ver en la figura 15.

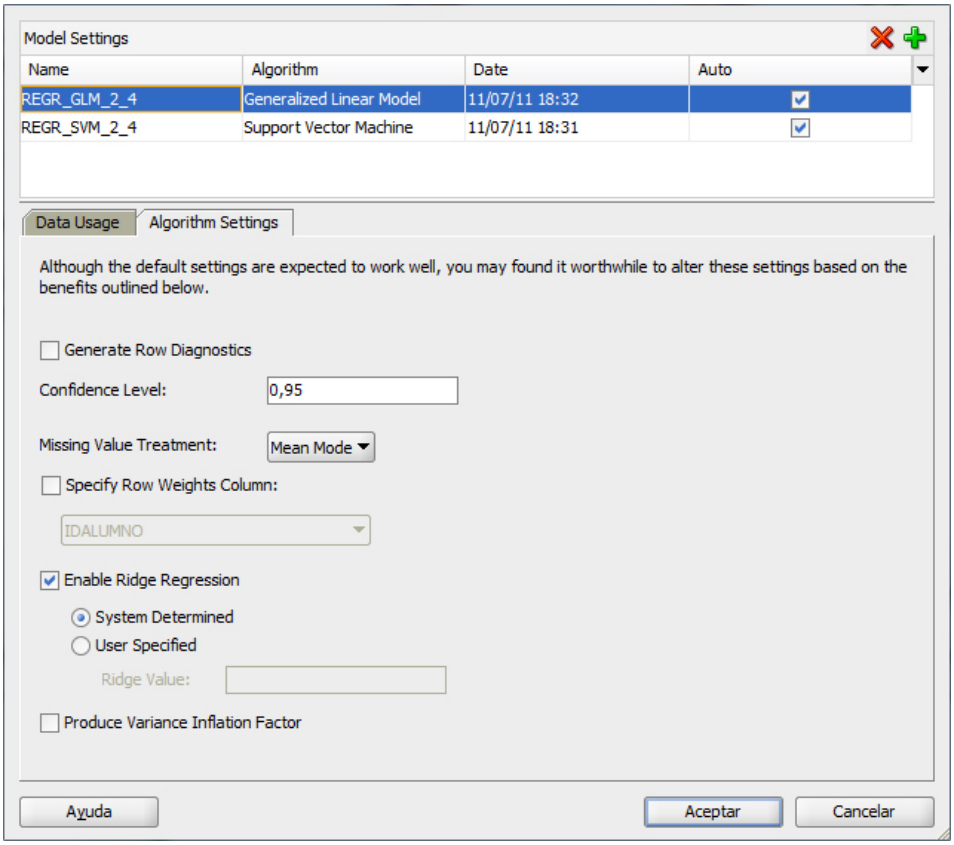


Figura 15 – Parámetros algoritmo GLM

Para el algoritmo SVM la configuración de los parámetros empleados se puede ver en la figura 16.

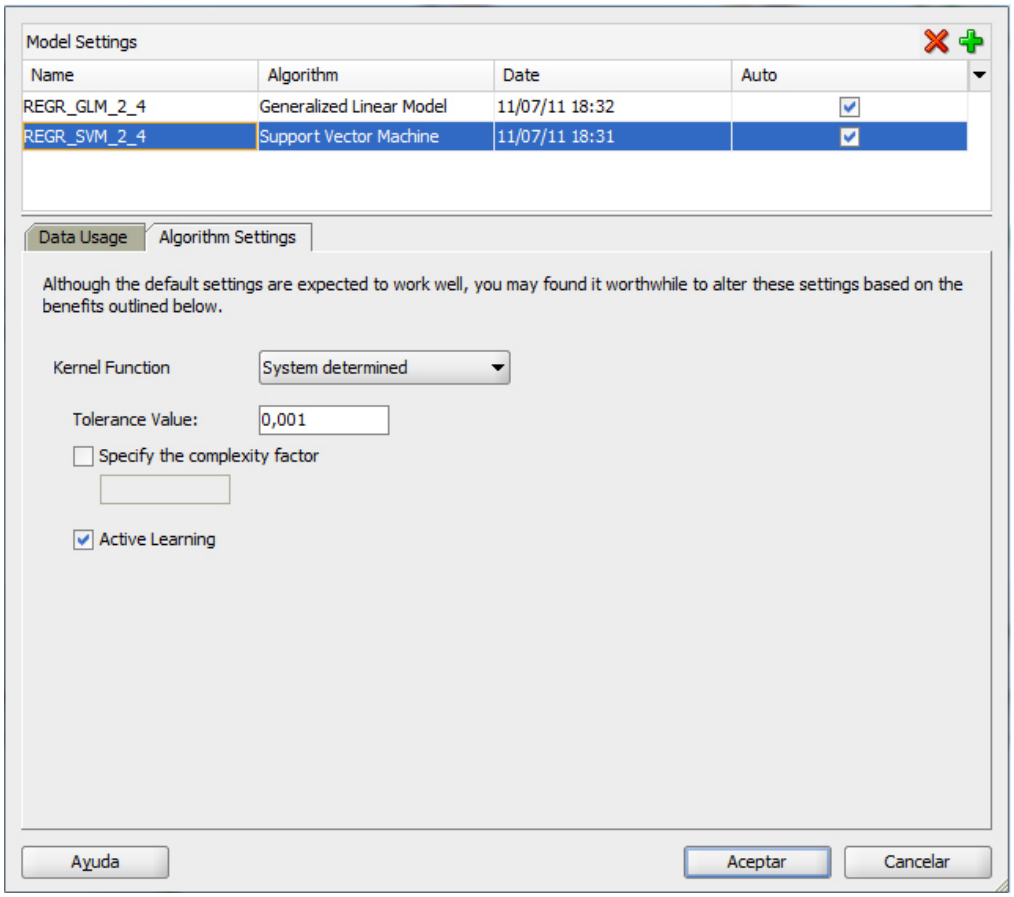


Figura 16 – Parámetros algoritmo SVM

Además de los parámetros empleados para la construcción del modelo, es necesario filtrar mediante la función de transformación que ofrece Oracle Data Mining algunas filas de las tablas originales ya que en muchos de los registros el campo objetivo (tiempo_aprobado_carrera) contiene el valor nulo. Además, es necesario filtrar también ciertas columnas de las tablas, para dejar solamente aquellos atributos que nos interesan para la consecución de este objetivo. Concretamente dejamos los atributos: idalumno, idtitulacion, instituto_procedencia y nota_acceso que son los que hemos escogido en el objetivo 1 de la minería de datos.

- **Objetivo 2**. Predicción de la nota media de la carrera que un alumno obtendrá al terminar sus estudios.

Para este problema el campo a predecir (objetivo) es "nota_media_carrera" y el *case ID* también será el "idalumno". En cuanto a los parámetros empleados para el algoritmo de GLM, y el de SVM se utilizan los parámetros que vienen por defecto en Oracle Data Mining, que son los mismos empleados para el objetivo 1. También se utilizan los mismos filtros de columnas y filas empleados para el objetivo 1.

- **Objetivo 3**. Predicción de la nota obtenida en una determinada asignatura.

En este tercer objetivo el campo que se desea predecir es "nota_asignatura" y el *case ID* también será el "idalumno". Los parámetros empleados para el algoritmo de GLM, y el de SVM serán una vez más los parámetros que vienen por defecto en Oracle Data Mining.

Modelos

Se ejecutan los tres modelos, uno por cada objetivo de la minería de datos, sobre un conjunto de datos de entrenamiento del 60%, con lo cual se deja el 40% de datos para el conjunto de prueba. Los detalles de la ejecución de cada modelo se pueden ver a continuación.

- **Modelo para el objetivo 1**

En la figura 17 se muestran los detalles de la ejecución del algoritmo GLM.

Name	Value
ADJUSTED_R_SQUARE	0,6428702
AIC	211,59662993
COEFF_VAR	19,27969307
CORRECTED_TOTAL_DF	1444
CORRECTED_TOT_SS	4642,42491349
DEPENDENT_MEAN	5,55778547
ERROR_DF	1433
ERROR_MEAN_SQUARE	1,14816364
ERROR_SUM_SQUARES	1645,31849418
F_VALUE	237,30434475
GMSEP	1,15778401
HOCKING_SP	0,00080179
J_P	1,15769856
MODEL_CONVERGED	Yes
MODEL_DF	11
MODEL_F_P_VALUE	0
MODEL_MEAN_SQUARE	272,46421994
MODEL_SUM_SQUARES	2997,10641932
NUM_PARAMS	12
NUM_ROWS	1445
ROOT_MEAN_SQ	1,07152398
R_SQ	0,64559072
SBIC	274,90700514
VALID_COVARIANCE_MATRIX	Yes

Figura 17 – Resultados algoritmo GLM para el objetivo 1

A continuación en la figura 18 se muestran los detalles para la ejecución del algoritmo SVM.

Name	Value
General	
Type	Regression
Owner	VICTOR
Model Name	REGR_SVM_2_4
Target Attribute	TIEMPO_APROBADO_CARRERA
Creation Date	26/07/11
Duration(Minutes)	0,05
Size(MB)	0,2035
Comment	
Algorithm	
Active Learning	Yes
Algorithm Name	Support Vector Machine
Automatic Preparation	On
Complexity Factor	0.896518
Kernel Cache Size	50000000
Kernel Function	Gaussian
Standard Deviation	1.026524
Tolerance	.001

Figura 18 – Resultados algoritmo SVM para el objetivo 1

- **Modelo para el objetivo 2**

En la figura 19 se muestran los detalles de la ejecución del algoritmo GLM.

Name	Value
ADJUSTED_R_SQUARE	0,63592985
AIC	-2238,3395074
COEFF_VAR	6,11590976
CORRECTED_TOTAL_DF	1444
CORRECTED_TOT_SS	835,14162381
DEPENDENT_MEAN	7,50287889
ERROR_DF	1432
ERROR_MEAN_SQUARE	0,21056104
ERROR_SUM_SQUARES	301,52340462
F_VALUE	211,18904817
GMSEP	0,21247369
HOCKING_SP	0,00014714
J_P	0,21245536
MODEL_CONVERGED	Yes
MODEL_DF	12
MODEL_F_P_VALUE	0
MODEL_MEAN_SQUARE	44,46818493
MODEL_SUM_SQUARES	533,61821919
NUM_PARAMS	13
NUM_ROWS	1445
ROOT_MEAN_SQ	0,4588693
R_SQ	0,63895536
SBIC	-2169,75326759
VALID_COVARIANCE_MATRIX	Yes

Figura 19 - Resultados algoritmo GLM para el objetivo 2

A continuación en la figura 20 se muestran los detalles para la ejecución del algoritmo SVM.

Name	Value
General	
Type	Regression
Owner	VICTOR
Model Name	REGR_SVM_3_4
Target Attribute	NOTA_MEDIA_CARRERA
Creation Date	26/07/11
Duration(Minutes)	0,0833
Size(MB)	0,2121
Comment	
Algorithm	
Active Learning	Yes
Algorithm Name	Support Vector Machine
Automatic Preparation	On
Complexity Factor	0.545810
Kernel Cache Size	50000000
Kernel Function	Gaussian
Standard Deviation	1.140772
Tolerance	.001

Figura 20 - Resultados algoritmo SVM para el objetivo 2

- **Modelo para el objetivo 3**

En la figura 21 se muestran los detalles de la ejecución del algoritmo GLM.

Name	Value
ADJUSTED_R_SQUARE	0,23701651
AIC	4180,03729593
COEFF_VAR	16,40571063
CORRECTED_TOTAL_DF	23110
CORRECTED_TOT_SS	36228,15258418
DEPENDENT_MEAN	6,66631215
ERROR_DF	23069
ERROR_MEAN_SQUARE	1,19608318
ERROR_SUM_SQUARES	27592,44296847
F_VALUE	173,8791184
GMSEP	1,19826089
HOCKING_SP	0,00005185
J_P	1,19825685
MODEL_CONVERGED	Yes
MODEL_DF	41
MODEL_F_P_VALUE	0
MODEL_MEAN_SQUARE	207,97388954
MODEL_SUM_SQUARES	8526,92947108
NUM_PARAMS	42
NUM_ROWS	23111
ROOT_MEAN_SQ	1,09365588
R_SQ	0,23837013
SBIC	4518,05598283
VALID_COVARIANCE_MATRIX	No

Figura 21 - Resultados algoritmo GLM para el objetivo 3

A continuación en la figura 22 se muestran los detalles para la ejecución del algoritmo SVM.

Name	Value
General	
Type	Regression
Owner	VICTOR
Model Name	REGR_SVM_4_4
Target Attribute	NOTA_ASIGNATURA
Creation Date	29/08/11
Duration(Minutes)	0,5167
Size(MB)	0,6791
Comment	
Algorithm	
Active Learning	Yes
Algorithm Name	Support Vector Machine
Automatic Preparation	On
Complexity Factor	0.427809
Kernel Cache Size	50000000
Kernel Function	Gaussian
Standard Deviation	1.850141
Tolerance	.001

Figura 22 - Resultados algoritmo SVM para el objetivo 3

Descripción del modelo

A continuación vamos a describir el resultado de la ejecución de cada uno de los modelos para cada objetivo, estos resultados se estudiarán más a fondo en la etapa de evaluación.

- **Modelo para el objetivo 1**

Este modelo ha devuelto los siguientes resultados:

Confianza predictiva (*predictive confidence*) para el algoritmo SVM tiene un valor del 68,05% y para el algoritmo GLM un valor del 40,87%.

Error absoluto medio (*mean absolute error*) para el algoritmo SVM tiene un valor de 0,44 y para el algoritmo GLM un valor de 0,89.

Error cuadrático medio (*root mean square error*) para el algoritmo SVM tiene un valor de 0,58 y para el algoritmo GLM un valor de 1,08.

El valor predicho medio (*mean predicted value*) para el algoritmo SVM tiene un valor de 5,46 y para el algoritmo GLM un valor de 5,55.

- **Modelo para el objetivo 2**

Este modelo ha devuelto los siguientes resultados:

Confianza predictiva (*predictive confidence*) para el algoritmo SVM tiene un valor del 47,44% y para el algoritmo GLM un valor del 44,78%.

Error absoluto medio (*mean absolute error*) para el algoritmo SVM tiene un valor de 0,34 y para el algoritmo GLM un valor de 0,35.

Error cuadrático medio (*root mean square error*) para el algoritmo SVM tiene un valor de 0,41 y para el algoritmo GLM un valor de 0,44.

El valor predicho medio (*mean predicted value*) para el algoritmo SVM tiene un valor de 7,5 y para el algoritmo GLM un valor de 7,49.

- **Modelo para el objetivo 3**

Este modelo ha devuelto los siguientes resultados:

Confianza predictiva (*predictive confidence*) para el algoritmo SVM tiene un valor del 0% y para el algoritmo GLM un valor del 12,19%.

Error absoluto medio (*mean absolute error*) para el algoritmo SVM tiene un valor de 1,04 y para el algoritmo GLM un valor de 0,88.

Error cuadrático medio (*root mean square error*) para el algoritmo SVM tiene un valor de 1,30 y para el algoritmo GLM un valor de 1,09.

El valor predicho medio (*mean predicted value*) para el algoritmo SVM tiene un valor de 6,56 y para el algoritmo GLM un valor de 6,68.

4.4. Evaluar el Modelo

Aunque en el paso 5 de la metodología CRISP-DM (evaluación) también se haga una evaluación de los modelos generados, en este apartado dicha evaluación está más orientada a los objetivos de la minería de datos, mientras que en el siguiente paso de la metodología, la evaluación se orienta más a los objetivos de negocio, si bien ambos objetivos están muy relacionados entre sí en este proyecto.

En términos de minería de datos, una buena manera de evaluar la efectividad de los modelos es utilizando los dos indicadores que se establecieron en el plan de pruebas de este documento, dichos indicadores son el error cuadrático medio (*root mean squared error*) y el error absoluto medio (*mean absolute error*). Además de estos dos indicadores, la herramienta Oracle Data Mining nos da más información acerca de los modelos que nos puede ser útil a la hora de evaluarlos, como son el indicador de confianza predictiva (*predictive confidence*) y el valor predicho medio junto al valor medio real. Utilizando estos valores que están reflejados en el apartado 4.3 procedemos a hacer una primera evaluación de cada modelo.

El primer modelo, es decir el que se ajusta al objetivo 1 tiene un valor de un 68,05% de confianza predictiva para el algoritmo SVM, este algoritmo da un mejor resultado que el GLM para este objetivo, ya que con éste último el valor de la confianza predictiva es tan solo de 40,87%. Además tanto el valor del error absoluto medio (0,44) como el del error cuadrático medio (0,58) es menor para el algoritmo SVM que para el algoritmo GLM, que presenta los valores 0,89 y 1,08 respectivamente, por lo que se emplearía este algoritmo para resolver el objetivo 1.

Para el segundo modelo, el que se ha utilizado para el objetivo 2, también es mejor emplear el algoritmo SVM ya que el valor para la confianza predictiva es mejor, 47,44% frente al 44,78% del algoritmo GLM, el valor para el error absoluto medio es ligeramente menor (0,34 frente a 0,35 del algoritmo GLM) y por último también el valor para error cuadrático medio es menor (0,41 frente a 0,44 del algoritmo GLM).

Por último tenemos el tercer modelo, para el que se han obtenido unos valores bastante bajos de confianza predictiva, tan sólo un 12,19% para el algoritmo GLM y un 0% para el SVM, y unos valores demasiado altos para el error absoluto medio (1,04 para SVM y

0,88 para GLM) y el error cuadrático medio (1,30 para SVM y 1,09 para GLM). Estos valores no son suficientes para garantizar una buena predicción para el valor de las notas en las asignaturas y es muy probable que se deba a que dicho valor de este campo no tiene relación directa con el resto de campos que podemos encontrar en la base de datos, por lo que para aumentar la fiabilidad de la predicción necesitaríamos obtener más datos acerca de las distintas asignaturas y encontrar relaciones que puedan darse entre las notas de estas asignaturas y otros factores. Por lo tanto, este último objetivo queda descartado por no ser un objetivo viable para los datos de los que disponemos.

En la siguiente tabla se pueden observar los valores para los distintos indicadores para hacer una mejor comparativa:

	Confianza Predictiva (SVM/GLM)	**Error Absoluto Medio (SVM/GLM)**	**Error Cuadrático Medio (SVM/GLM)**
Modelo 1	✓68,05/40,87	✓0,44/0,89	✓0,58/1,08
Modelo 2	✓47,44/44,78	✓0,34/0,35	✓0,41/0,44
Modelo 3	0/12,19✓	1,04/0,88✓	1,30/1,09✓

5. Evaluación

En esta fase de la metodología se intentan evaluar los modelos generados pero en esta ocasión la evaluación se hace desde el punto de vista de los objetivos de negocio en lugar de los objetivos de minería de datos. Una vez realizada esta evaluación, se debe decidir si los objetivos han sido cumplidos y de ser así se puede avanzar a la fase de implantación, de lo contrario se tendría que identificar cualquier factor que se haya podido pasar por alto y hacer una revisión del proceso.

5.1. Evaluar los Resultados

Desde el punto de vista del negocio, se había establecido como criterio de éxito principal el poder realizar predicciones con un porcentaje de fiabilidad "aceptable", este criterio puede ser algo subjetivo, por lo que es inevitable apoyarse principalmente en los criterios de éxito desde el punto de vista de la minería de datos que son mucho más específicos y precisos. Además, para poder calificar como aceptable o no las predicciones que se van a realizar es necesario tener una base objetiva, como lo son los indicadores estadísticos que se han obtenido al ejecutar los modelos. También sería conveniente la evaluación de los resultados por parte de un grupo de expertos en la minería de datos, si se contara con ellos. En cualquier caso, basándonos en los indicadores obtenidos mediante la herramienta de minería de datos, a continuación podemos hacer una evaluación de cada modelo para así descartar aquel que no cumpla con unos requisitos mínimos.

- **Modelo para el objetivo 1**

  Este modelo es factible ya que se pueden hacer predicciones acerca de cuánto tiempo va a tardar un alumno en terminar la carrera con un porcentaje de fiabilidad de un 68%, el cual consideramos aceptable desde el punto de vista de los objetivos de negocio.

- **Modelo para el objetivo 2**

  Este segundo modelo también es aceptable desde el punto de vista de los objetivos de negocio ya que podríamos hacer predicciones para la nota media que un alumno obtendrá una vez terminada su titulación con un porcentaje de fiabilidad de casi un 50%.

- **Modelo para el objetivo 3**

  Este modelo no es viable ya que no nos ofrece suficientes garantías con tan sólo un 12% de confianza predictiva para poder realizar predicciones fiables acerca de las notas que los alumnos obtendrán en una determinada asignatura. Por este motivo, este modelo debería ser descartado o revisado.

  Como nota aparte para este objetivo, más allá del modelo en sí, y como es conveniente identificar cualquier tipo de hallazgo que provenga de los datos originales, se puede destacar que a través de algunas consultas SQL realizadas sobre las bases de datos (dichas consultas se encuentran en el anexo 2) se han identificado aquellas asignaturas que son más problemáticas para los alumnos de cada titulación, concretamente "Redes de Ordenadores II", "Análisis de Valores" y "Recursos Territoriales Turísticos". En los gráficos 12, 13 y 14 se exponen estos datos en forma de porcentajes de notas obtenidas por los alumnos.

Modelos aprobados

Por las razones explicadas en este apartado y en el apartado 4.4 (evaluar el modelo) los modelos aprobados son el modelo 1 y el modelo 2 que cumplen con los criterios de éxito de negocio, mientras que el modelo 3 será descartado por no cumplir con los requisitos de negocio ni de minería de datos.

5.2. Revisar el Proceso

El proceso hasta este punto se ha ejecutado tal y como estaba previsto, si bien ha habido complicaciones a la hora de realizar el modelo para el objetivo 3 ya que se han obtenido valores muy deficientes para la confianza predictiva, el error absoluto medio y el error cuadrático medio. La causa de estos malos valores posiblemente se encuentre en la base de datos utilizada, ya que al no ser ésta una base de datos real, no disponemos de todos los datos que se podrían necesitar para hacer una predicción fiable sobre las notas de ciertas asignaturas. Esto en un escenario real, es decir, utilizando la base de datos de la que dispone la universidad posiblemente se podría subsanar. En cualquier caso, para el presente proyecto se ha decidido descartar este objetivo.

5.3. Determinar los Próximos Pasos

El siguiente paso a realizar en el proyecto es el de ejecutar la etapa de implantación para los objetivos 1 y 2.

6. Implantación

Esta es la última fase de la metodología CRISP-DM y el objetivo de la misma es el de explicar al cliente como poner en funcionamiento el proyecto que se ha construido en las fases anteriores, así como exponer los resultados obtenidos al cliente de forma que lo pueda entender fácilmente. Otro objetivo de esta fase es el de crear una estrategia para el mantenimiento del proyecto y producir un informe en el que se incluyan posibles mejoras para el futuro y un listado de las dificultades encontradas a la hora de realizarlo.

6.1. Planear la Implantación

Para poder implantar este proyecto en el negocio real sería necesario en primer lugar tener acceso a la base de datos real del negocio, es decir la base de datos que contiene toda la información relativa a los alumnos de la universidad. A partir de ahí, los pasos a seguir serían los mismos que se han seguido en este documento desde la comprensión del negocio hasta la implantación. Si bien, cabe decir que habrá algunas fases, como la de comprensión y preparación de los datos, que en el negocio real probablemente sean más complejas y llevarán más tiempo que en este proyecto ya que se puede esperar que en la base de datos real se tengan muchos más registros y estos mismos contengan más ruido que en nuestra base de datos ficticia creada específicamente para este uso.

En segundo lugar sería necesario que en el negocio (la universidad) se use una base de datos Oracle, de no ser así se tendrían dos opciones, la primera sería exportar la base de datos actual a una base de datos Oracle, y la segunda sería utilizar otro software de minería de datos distinto al utilizado en este proyecto (Oracle Data Mining), para este propósito se podría utilizar alguna de las herramientas listadas en el apartado 1.6 (Herramientas) que mejor se adapte a la base de datos original, para esto sería necesario hacer un estudio previo que determine que herramienta es la más apropiada.

Es posible que el objetivo 3 de minería de datos, que en este proyecto ha sido descartado por no ser viable, utilizando los datos de la base de datos real fuera factible, por lo que se tendría que repasar la fase de evaluación prestando especial atención a este objetivo en particular a fin de determinar dicha viabilidad.

6.2. Planear la Monitorización y Mantenimiento

La supervisión y mantenimiento de la implementación del presente proyecto es una fase importante del mismo debido a que los datos que se procesan con mucha frecuencia pueden ser modificados por el personal de la universidad. Los datos pueden ser modificados por diferentes motivos como haber realizado una codificación incorrecta, haber asignado una nota incorrecta al alumno, etc. El volumen de estos datos en movimiento es grande motivo por el cual la extracción de las muestras debe ser realizada cuidadosamente y realizando siempre backups de los datos explotados en cada proceso. La minería de datos debería ser realizada en periodos de cuatro meses (cuatrimestres) ya que esta es la medida de tiempo utilizada en la universidad para realizar los exámenes y asignar las notas finales a los alumnos, sin embargo esta medida podría variar en cualquier momento en función del plan de estudios que esté vigente en cada momento.

Como plan de supervisión y mantenimiento se podría establecer los siguientes procesos:

- Extracción y almacenamiento cuatrimestral de los datos guardando la información obtenida en formato de hoja de cálculo
- Distribución de los datos en función de los modelos de software de minería de datos a trabajar.
- Los archivos de la explotación de datos deberán ser guardados en soporte magnético en la propia universidad, almacenándolos por ejemplo en carpetas ordenadas por procesos cuatrimestrales.
- Los resultados obtenidos en cada explotación de datos deberán ser llevados a formato de hoja de cálculo y generar gráficas de distintos tipos para una mejor visualización e interpretación de los resultados obtenidos en cada periodo.

6.3. Producir el Informe Final

En este paso se debe presentar un informe resumiendo los puntos importantes del proyecto y la experiencia adquirida durante su desarrollo. El público al que va dirigido este informe sería el personal de la universidad encargado de la docencia (profesores, directores de departamento, etc.) de tal manera que se pueda estudiar la situación actual y tomar medidas correctivas para la mejora del servicio académico. Cabe decir que parte de este informe final será presentado de manera oral con una presentación, por lo que en este apartado solamente haremos un breve resumen.

El uso de la metodología CRISP-DM en este proyecto ha permitido encontrar un comportamiento predictivo a la hora de estimar la duración de la carrera de los alumnos y la nota media de los mismos. Se ha podido encontrar un plan de extracción, normalización, y codificación de datos para la realización de procesos de minería de datos cuatrimestrales.

De los tres objetivos de minería de datos iniciales que se habían fijado se han podido alcanzar dos de ellos (objetivos 1 y 2). Además, al margen de estos objetivos, se han sacado otras conclusiones a partir de los datos estudiados, concretamente se han identificado las asignaturas más problemáticas para los alumnos de cada una de las titulaciones estudiadas.

Repasando las diferentes etapas que hemos seguido para llegar al objetivo:

La primera etapa ha sido una de las más laboriosas por no tener una base de datos de la que partir. Esto ha supuesto que tengamos que generar nosotros mismos un conjunto de datos sobre el que trabajar. Para poder hacer una simulación lo más real posible, no valía con generar datos aleatorios, si no que se ha tenido que desarrollar un pequeño programa en Java que generase estos datos de manera automática, debido a la gran cantidad de datos que necesitábamos manejar para hacer una estimación lo más precisa posible.

Cuando ya disponíamos de la base de datos sobre la que ejercer la minería de datos, se hizo un análisis de la estructura de los datos y la información contenida.

Se realizaron consultas significativas en SQL (Anexo 2) para tener muestras representativas de los datos, y sacar más conclusiones al margen de los objetivos iniciales de la minería.

El lado positivo de haber creado nosotros mismos la base de datos a nuestro antojo es que la fase de preparación de los datos fue mucho más sencilla ya que no hizo falta apenas hacer una limpieza de los datos, conversiones o formateo de los mismos. Esto redujo significantemente la duración estimada de la etapa 3 definida en el apartado 1.4 (Realización del Plan de Proyecto).

A continuación se realizó la elección de las técnicas de modelado y la ejecución de dichas técnicas sobre los datos empleando la herramienta escogida para ello (Oracle Data Miner). Esta herramienta facilitó por completo la aplicación de los modelos ya que nos permitió ver de manera muy intuitiva y visual cuales eran las técnicas más adecuadas para nuestra base de datos.

Por último, una vez obtenidos los modelos, se analizaron para determinar la adecuación o no de los mismos. En este caso determinamos que los modelos 1 y 2 podrían ser válidos para nuestros objetivos y se descartó el 3 por no ser lo suficientemente fiable.

Realizados todos estos pasos se presentan los resultados alcanzados al público que es el objetivo de este apartado.

6.4. Revisar el Proyecto

En esta última etapa de la metodología se debe hacer una evaluación de aquellas cosas que se hicieron correctamente y aquellas que no, así como posibles mejoras para que en las futuras ejecuciones de la minería de datos se vayan puliendo los fallos y se obtengan mejores resultados.

En primer lugar, y como ya se ha comentado anteriormente en otros apartados, el mayor lastre que se ha ido arrastrando a lo largo de este proyecto es el de no disponer de una base de datos real sobre la que actuar ya que esto condiciona en gran medida los resultados obtenidos. A pesar de haber intentado generar unos datos lo más veraces posible, no cabe duda que existen multitud de factores que no podemos manejar y que disponer de los datos reales con las notas y demás características de los alumnos incrementaría aún más la fiabilidad de los modelos de minería de datos elegidos en este proyecto. Esto se puede interpretar como algo positivo ya que si hemos dado por válidos dos de los tres modelos empleados, y sin disponer de la cantidad y veracidad de los datos que se manejan en la base de datos de la universidad, esto quiere decir que si el proyecto saliera en real los resultados mejorarían aún más.

Bibliografía

[Hernández, Ramírez y Ferri, 2004] José Hernández Orallo, Mª José Ramírez Quintana, César Ferri Ramírez. Introducción a la Minería de Datos. Ed. Pearson Educación, S.A. 2004.

[CRISP-DM, 2000] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, Rüdiger Wirth. CRISP-DM 1.0, Step-by-step Data Mining Guide, 2000.

[Rodríguez, 2010] Dr. Oldemar Rodríguez Rojas. Metodología para el Desarrollo de Proyectos en Minería de Datos CRISP-DM, 2010.

[prudsys, 2011] prudsys XELOPES.

http://www.prudsys.de/en/products/prudsys-xelopes/ [consulta: 23 febrero 2011]

[JDMP, 2011] Java Data Mining Package.

http://www.jdmp.org/ [consulta: 23 febrero 2011]

[IBM, 2011] IBM SPSS Modeler.

http://www.spss.com/es/software/modeling/modeler-pro/ [consulta: 23 febrero 2011]

[Waikato, 2011] WEKA (*Waikato Environment for Knowledge Analysis*). University of Waikato.

http://www.cs.waikato.ac.nz/ml/weka/ [consulta: 25 febrero 2011]

[Oracle, 2011] Oracle Data Mining.

http://www.oracle.com/technetwork/database/options/odm/index.html [consulta: 25 febrero 2011]

[Oracle DM, 2011] Oracle Data Mining Concepts (apartado 5.1)

http://download.oracle.com/docs/html/B14339_01/5dmtasks.htm [consulta: 1 marzo 2011]

[Rapid-I, 2011] RapidMiner.

http://rapid-i.com/content/view/181/196/lang,en/ [consulta: 25 febrero 2011]

[StatSoft, 2011] STATISTICA Data Miner

http://www.statsoft.com/products/statistica-data-miner/ [consulta: 25 febrero 2011]

[Salford, 2011] CART

http://salford-systems.com/cart.php [consulta: 25 febrero 2011]

[RuleQuest, 2011] See5/C5.0

http://www.rulequest.com/see5-info.html [consulta: 25 febrero 2011]

[Wikipedia 1, 2011] Distribución normal – Wikipedia

http://es.wikipedia.org/wiki/Distribuci%C3%B3n_normal [consulta: 7 julio 2011]

[Wikipedia 2, 2011] Esquema en estrella – Wikipedia

http://es.wikipedia.org/wiki/Esquema_en_estrella [consulta: 17 junio 2011]

Anexo 1: Glosario de Terminología de Minería de Datos

• Algoritmos genéticos: Técnicas de optimización que usan procesos tales como combinación genética, mutación y selección natural en un diseño basado en los conceptos de evolución natural.

• Análisis de series de tiempo (time-series): Análisis de una secuencia de medidas hechas a intervalos específicos. El tiempo es usualmente la dimensión dominante de los datos.

• Análisis prospectivo de datos: Análisis de datos que predice futuras tendencias, comportamientos o eventos basado en datos históricos.

• Análisis exploratorio de datos: Uso de técnicas estadísticas tanto gráficas como descriptivas para aprender acerca de la estructura de un conjunto de datos.

• Análisis retrospectivo de datos: Análisis de datos que provee una visión de las tendencias, comportamientos o eventos basado en datos históricos.

• Árbol de decisión: Estructura en forma de árbol que representa un conjunto de decisiones. Estas decisiones generan reglas para la clasificación de un conjunto de datos.

• Base de datos multidimensional: Base de datos diseñada para procesamiento analítico on-line (OLAP). Estructurada como un hipercubo con un eje por dimensión.

• CART (Árboles de clasificación y regresión): Una técnica de árbol de decisión usada para la clasificación de un conjunto da datos. Provee un conjunto de reglas que se pueden aplicar a un nuevo (sin clasificar) conjunto de datos para predecir cuáles registros darán un cierto resultado. Segmenta un conjunto de datos creando 2 divisiones. Requiere menos preparación de datos que CHAID.

• CHAID (Detección de interacción automática de Chi cuadrado): Una técnica de árbol de decisión usada para la clasificación de un conjunto da datos. Provee un conjunto de reglas que se pueden aplicar a un nuevo (sin clasificar) conjunto de datos para predecir cuáles registros darán un cierto resultado. Segmenta un conjunto de datos utilizando tests de chi cuadrado para crear múltiples divisiones. Antecede, y requiere más preparación de datos, que CART.

• Clasificación: Proceso de dividir un conjunto de datos en grupos mutuamente excluyentes de tal manera que cada miembro de un grupo esté lo "más cercano" posible a otro, y grupos diferentes estén lo "más lejos" posible uno del otro, donde la distancia está medida con respecto a variable(s) específica(s) las cuales se están tratando de predecir. Por ejemplo, un problema típico de clasificación es el de dividir una base de

datos de compañías en grupos que son lo más homogéneos posibles con respecto a variables como "posibilidades de crédito" con valores tales como "Bueno" y "Malo".

• Clustering (agrupamiento): Proceso de dividir un conjunto de datos en grupos mutuamente excluyentes de tal manera que cada miembro de un grupo esté lo "más cercano" posible a otro, y grupos diferentes estén lo "más lejos" posible uno del otro, donde la distancia está medida con respecto a todas las variables disponibles.

• Computadoras con multiprocesadores: Una computadora que incluye múltiples procesadores conectados por una red. Ver procesamiento paralelo.

• Data cleansing: Proceso de asegurar que todos los valores en un conjunto de datos sean consistentes y correctamente registrados.

• Data Mining: La extracción de información predecible escondida en grandes bases de datos.

• Data Warehouse: Sistema para el almacenamiento y distribución de cantidades masivas de datos.

• Datos anormales: Datos que resultan de errores o que representan eventos inusuales.

• Dimensión: En una base de datos relacional o plana, cada campo en un registro representa una dimensión. En una base de datos multidimensional, una dimensión es un conjunto de entidades similares; por ejemplo una base de datos multidimensional de ventas podría incluir las dimensiones Producto, Tiempo y Ciudad.

• Modelo analítico: Una estructura y proceso para analizar un conjunto de datos. Por ejemplo, un árbol de decisión es un modelo para la clasificación de un conjunto de datos.

• Modelo lineal: Un modelo analítico que asume relaciones lineales entre una variable seleccionada (dependiente) y sus predictores (variables independientes).

• Modelo no lineal: Un modelo analítico que no asume una relación lineal en los coeficientes de las variables que son estudiadas.

• Modelo predictivo: Estructura y proceso para predecir valores de variables especificadas en un conjunto de datos.

• Navegación de datos: Proceso de visualizar diferentes dimensiones, "fetas" y niveles de una base de datos multidimensional.

• OLAP Procesamiento analítico on-line (On Line Analitic Processing): Se refiere a aplicaciones de bases de datos orientadas a *arrays* que permite a los usuarios ver, navegar, manipular y analizar bases de datos multidimensionales.

• Outlier: Un ítem de datos cuyo valor cae fuera de los límites que encierran a la mayoría del resto de los valores correspondientes de la muestra. Puede indicar datos anormales. Deberían ser examinados detenidamente, pueden dar importante información.

• Procesamiento paralelo: Uso coordinado de múltiples procesadores para realizar tareas computacionales. El procesamiento paralelo puede ocurrir en una computadora con múltiples procesadores o en una red de estaciones de trabajo o PCs.

• Regresión lineal: Técnica estadística utilizada para encontrar la mejor relación lineal que encaja entre una variable seleccionada (dependiente) y sus predicados (variables independientes).

• Regresión logística: Una regresión lineal que predice las proporciones de una variable seleccionada categórica, tal como Tipo de Consumidor, en una población.

• SMP Multiprocesador simétrico (Symmetric Multiprocessor): Tipo de computadora con multiprocesadores en la cual la memoria es compartida entre los procesadores.

Anexo 2: Scripts de Creación de Tablas y Consultas SQL

- Script empleado para la creación del modelo de datos:

```
DROP TABLE SEGUIM_ACADEMICO;
DROP TABLE FECHA;
DROP TABLE ASIGNATURA;
DROP TABLE ALUMNO;
DROP TABLE TITULACION;

CREATE TABLE FECHA(
IDFecha number,
Cuatrimestre number not null,
Curso number not null,
Primary key (IDFecha)
);

CREATE TABLE ASIGNATURA(
IDAsignatura number,
Titulacion number not null,
Nombre varchar2(80) not null,
Tipo varchar2(15) not null,
Creditos number not null,
Cuatrimestre number not null,
Especialidad varchar2(60),
Curso number not null,
```

```
Primary key (IDAsignatura)
);

CREATE TABLE ALUMNO(
IDAlumno number,
NIF varchar2(9) not null,
Nombre varchar2(15) not null,
Apellido1 varchar2(30) not null,
Apellido2 varchar2(30) not null,
Sexo varchar2(1) not null,
Fecha_nacimiento date not null,
Localidad varchar2(30) not null,
Provincia varchar2(15) not null,
Pais varchar2(15) not null,
Instituto_procedencia number not null,
Eleccion_estudios_instituto varchar2(40) not null,
Nota_acceso number not null,
primary key (IDAlumno)
);

CREATE TABLE TITULACION(
IDTitulacion number,
Nombre varchar2(30) not null,
Creditos_obligatorios number not null,
Creditos_optativos number not null,
Creditos_libre_eleccion number not null,
```

```
Total_creditos number not null,
Creditos_primer_curso number not null,
Creditos_segundo_curso number not null,
Creditos_tercer_curso number not null,
Creditos_cuarto_curso number,
Creditos_quinto_curso number,
Creditos_sexto_curso number,
primary key (IDTitulacion)
);

CREATE TABLE SEGUIM_ACADEMICO(
IDAlumno number not null,
IDTitulacion number not null,
IDAsignatura number not null,
IDFecha number not null,
Nota_asignatura number,
Fecha_admision date not null,
Tiempo_aprobado_1 number,
Nota_media_1 number,
Tiempo_aprobado_2 number,
Nota_media_2 number,
Tiempo_aprobado_3 number,
Nota_media_3 number,
Tiempo_aprobado_4 number,
Nota_media_4 number,
Tiempo_aprobado_5 number,
```

```
Nota_media_5 number,
Tiempo_aprobado_6 number,
Nota_media_6 number,
Tiempo_aprobado_carrera number,
Nota_media_carrera number,
primary key (IDAlumno, IDTitulacion, IDAsignatura, IDFecha)
);

ALTER TABLE ASIGNATURA
ADD CONSTRAINT IDTitulacion FOREIGN KEY (Titulacion)
REFERENCES TITULACION (IDTitulacion);

ALTER TABLE SEGUIM_ACADEMICO
ADD CONSTRAINT IDAlumno FOREIGN KEY (IDAlumno)
REFERENCES ALUMNO (IDAlumno);

ALTER TABLE SEGUIM_ACADEMICO
ADD CONSTRAINT IDTitulacionSeguim FOREIGN KEY (IDTitulacion)
REFERENCES TITULACION (IDTitulacion);

ALTER TABLE SEGUIM_ACADEMICO
ADD CONSTRAINT IDAsignatura FOREIGN KEY (IDAsignatura)
REFERENCES ASIGNATURA (IDAsignatura);

ALTER TABLE SEGUIM_ACADEMICO
ADD CONSTRAINT IDFecha FOREIGN KEY (IDFecha)
```

REFERENCES FECHA (IDFecha);

- Consultas empleadas durante la realización del proyecto:

DEVUELVE EL ALUMNO CON MAYOR NOTA MEDIA AGRUPADO POR TITULACIÓN:

```
SELECT MAX (Nota), IDAlumno, IDTitulacion FROM
    (SELECT h.IDAlumno, tit.IDTitulacion,Nota_media_carrera Nota
    FROM alumno al, titulacion tit, Seguim_Academico h
    WHERE al.IDAlumno = h.IDAlumno AND tit.IDTitulacion = h.IDTitulacion
    AND tit.IDTitulacion IN
        (SELECT DISTINCT IDTitulacion FROM Seguim_Academico))
GROUP BY IDAlumno, IDtitulacion;
```

DEVUELVE LA NOTA MEDIA DE LA CARRERA DE TODOS LOS ALUMNOS QUE HAN TERMINADO POR TITULACIÓN:

```
SELECT Avg(NOTA_MEDIA_CARRERA),IDTITULACION
FROM SEGUIM_ACADEMICO
WHERE NOTA_MEDIA_CARRERA <> 0
GROUP BY IDTITULACION;
```

DEVUELVE LA NOTA MEDIA MÁS ALTA POR CADA TITULACIÓN:

```
SELECT MAX(NOTA_MEDIA_CARRERA) AS NOTA_MAS_ALTA,IDTITULACION
FROM SEGUIM_ACADEMICO
WHERE NOTA_MEDIA_CARRERA <> 0
GROUP BY IDTITULACION;
```

DEVUELVE LA NOTA MEDIA MÁS BAJA POR CADA TITULACIÓN:

```
SELECT MIN(NOTA_MEDIA_CARRERA) AS NOTA_MAS_BAJA,IDTITULACION
FROM SEGUIM_ACADEMICO
WHERE NOTA_MEDIA_CARRERA <> 0
GROUP BY IDTITULACION;
```

DEVUELVE LA NOTA MEDIA DE ACCESO DE LOS ALUMNOS DE CADA INSTITUTO ORDENADA DESCENDENTEMENTE:

```
SELECT Avg(NOTA_ACCESO) NOTA_MEDIA,INSTITUTO_PROCEDENCIA
FROM ALUMNO
GROUP BY INSTITUTO_PROCEDENCIA
ORDER BY NOTA_MEDIA DESC;
```

DEVUELVE LA NOTA MEDIA DE ACCESO DE LOS ALUMNOS DE CADA PROVINCIA ORDENADA DESCENDENTEMENTE:

```
SELECT Avg(NOTA_ACCESO) NOTA_MEDIA,PROVINCIA
FROM ALUMNO
GROUP BY PROVINCIA
ORDER BY NOTA_MEDIA DESC;
```

DEVUELVE LA NOTA MEDIA DE CADA CURSO DE CADA TITULACIÓN DE LOS ALUMNOS QUE HAN TERMINADO LA CARRERA:

```
SELECT
Avg(NOTA_MEDIA_1) NOTA_MEDIA_1,
Avg(NOTA_MEDIA_2) NOTA_MEDIA_2,
Avg(NOTA_MEDIA_3) NOTA_MEDIA_3,
Avg(NOTA_MEDIA_4) NOTA_MEDIA_4,
Avg(NOTA_MEDIA_5) NOTA_MEDIA_5,
Avg(NOTA_MEDIA_6) NOTA_MEDIA_6,
IDTITULACION
FROM SEGUIM_ACADEMICO
WHERE NOTA_MEDIA_CARRERA <> 0
GROUP BY IDTITULACION
ORDER BY IDTITULACION ASC;
```

DEVUELVE LA NOTA MEDIA DEL PRIMER CURSO DE CADA TITULACIÓN DE TODOS LOS ALUMNOS (INCLUYE LOS QUE NO HAN TERMINADO AÚN):

```
SELECT Avg(NOTA_MEDIA_1) NOTA_MEDIA_1,IDTITULACION
FROM
    (SELECT DISTINCT IDALUMNO,IDTITULACION,NOTA_MEDIA_1
    FROM SEGUIM_ACADEMICO
    WHERE NOTA_MEDIA_1 <> 0)
GROUP BY IDTITULACION
ORDER BY IDTITULACION ASC;
```

DEVUELVE LA NOTA MEDIA DE LA CARRERA DE CADA TITULACIÓN DE TODOS LOS ALUMNOS (INCLUYE LOS QUE NO HAN TERMINADO AÚN):

```
SELECT Avg(NOTA_MEDIA_CARRERA) NOTA_MEDIA_CARRERA,IDTITULACION
FROM
    (SELECT DISTINCT IDALUMNO,IDTITULACION,NTA_MEDIA_CARRERA
    FROM SEGUIM_ACADEMICO
    WHERE NOTA_MEDIA_CARRERA <> 0)
GROUP BY IDTITULACION
ORDER BY IDTITULACION ASC;
```

DEVUELVE LA MEDIA DE TIEMPO DE APROBADO DE CADA CURSO PARA CADA TITULACIÓN (SÓLO TIENE EN CUENTA ALUMNOS QUE HAN TERMINADO):

```
SELECT
Avg(TIEMPO_APROBADO_1),
Avg(TIEMPO_APROBADO_2),
Avg(TIEMPO_APROBADO_3),
Avg(TIEMPO_APROBADO_4),
Avg(TIEMPO_APROBADO_5),
Avg(TIEMPO_APROBADO_6),
IDTITULACION
FROM SEGUIM_ACADEMICO
WHERE TIEMPO_APROBADO_CARRERA <> 0
GROUP BY IDTITULACION
ORDER BY IDTITULACION ASC;
```

DEVUELVE LA MEDIA DEL TIEMPO DE APROBADO DEL PRIMER CURSO PARA TODOS LOS ALUMNOS DE CADA TITULACIÓN:

```
SELECT Avg(TIEMPO_APROBADO_1),IDTITULACION
FROM
    (SELECT DISTINCT IDALUMNO,IDTITULACION,TIEMPO_APROBADO_1
    FROM SEGUIM_ACADEMICO
    WHERE TIEMPO_APROBADO_1 <> 0)
GROUP BY IDTITULACION
ORDER BY IDTITULACION ASC;
```

DEVUELVE LA MEDIA DEL TIEMPO DE APROBADO DE LA CARRERA PARA TODOS LOS ALUMNOS DE CADA TITULACIÓN:

```
SELECT Avg(TIEMPO_APROBADO_CARRERA),IDTITULACION
FROM
    (SELECT
    DISTINCT IDALUMNO,IDTITULACION,TIEMPO_APROBADO_CARRERA
    FROM SEGUIM_ACADEMICO
    WHERE TIEMPO_APROBADO_CARRERA <> 0)
GROUP BY IDTITULACION
ORDER BY IDTITULACION ASC;
```

DEVUELVE UNA LISTA DE ALUMNOS DE LA TITULACIÓN 1 (INFORMATICA) ORDENADOS POR ID ASCENDENTE CUYAS NOTAS MEDIAS ESTÉN ENTRE 5 Y 6:

```
SELECT IDALUMNO,NOTA_MEDIA_1
FROM
    (SELECT DISTINCT IDALUMNO,IDTITULACION,NOTA_MEDIA_1
    FROM SEGUIM_ACADEMICO
    WHERE NOTA_MEDIA_1 <> 0)
WHERE IDTITULACION = 1 AND NOTA_MEDIA_1 >= 5 AND NOTA_MEDIA_1 < 6
ORDER BY IDALUMNO ASC;
```

DEVUELVE EL NÚMERO TOTAL DE ALUMNOS DE LA TITULACIÓN 1 (INFORMÁTICA):

```
SELECT COUNT(IDALUMNO) AS NUMERO_ALUMNOS
FROM
    (SELECT DISTINCT IDALUMNO
    FROM SEGUIM_ACADEMICO
    WHERE IDTITULACION = 1);
```

DEVUELVE LA LISTA DE ASIGNATURAS DE TODAS LAS TITULACIONES JUNTO CON LA MEDIA DE LAS NOTAS OBTENIDAS POR LOS ALUMNOS EN ORDEN ASCENDENTE:

```
SELECT Avg(NOTA_ASIGNATURA) NOTA_MEDIA,IDASIGNATURA
FROM
    (SELECT DISTINCT IDALUMNO,IDASIGNATURA,NOTA_ASIGNATURA
    FROM SEGUIM_ACADEMICO)
GROUP BY IDASIGNATURA
ORDER BY NOTA_MEDIA ASC;
```

DEVUELVE LA LISTA DE ALUMNOS DE LA ASIGNATURA 9 (REDES II) QUE HAN OBTENIDO UNA NOTA ENTRE 0 Y 5 ORDENADA POR IDALUMNO ASCENDENTE:

```
SELECT IDALUMNO,NOTA_ASIGNATURA
FROM
    (SELECT DISTINCT IDALUMNO,IDASIGNATURA,NOTA_ASIGNATURA
    FROM SEGUIM_ACADEMICO)
WHERE IDASIGNATURA = 9 AND NOTA_ASIGNATURA >= 0 AND NOTA_ASIGNATURA < 5
ORDER BY IDALUMNO ASC;
```