GRADO EN ADMINISTRACIÓN Y DIRECCIÓN DE EMPRESAS

CURSO ACADÉMICO: 2018/19

TRABAJO FIN DE GRADO

Mención en Dirección General

Técnicas de Minería de Datos con la Herramienta Weka

Data Mining Techniques with the Weka Tool

Autor

Juan Manuel Farfán Tiscar

Director

Pedro Solana González

Fecha: septiembre de 2019

ÍNDICE

RESUMEN

El presente Trabajo de Fin de Grado se centra en la minería de datos, también conocida como *Data Mining*, que tiene una gran importancia en el mundo actual. La importancia de este tema se basa en que las personas son generadores natos de datos, que si son bien canalizados y estudiados dan a las compañías información valiosa que les puede servir ya sea para crear nuevos productos, para observar el comportamiento de clientes o clientes potenciales, entre otras utilidades.

Antiguamente las compañías simplemente podían almacenar los datos en archivos al comienzo en papel y más tarde en archivos digitales en ordenadores, que con la tecnología disponible era muy complicado poder analizarlos para transformarlos en información útil para la compañía.

Con este trabajo se pretende ser conscientes de los datos que se generan y como pueden ser utilizados por las compañías para adaptar los gustos y comportamientos a sus productos o servicios para fomentar su consumo. Para ello explicaremos en que consiste la minería de datos, así como las aplicaciones que actualmente se llevan a cabo con esta tecnología.

Se llevará a la práctica el *Data mining* con la utilización de la herramienta Weka, la cual se verá tres ejercicios básicos como son la regresión lineal, la red bayesiana y el árbol de decisión. Y, por último, un ejercicio más complejo, como es el *clustering.* Con ello, se busca comprender el funcionamiento de un programa de *Data mining* de manera más visual.

Palabras clave: Minería de datos, Inteligencia Artificial, Nodos.

ABSTRACT

This Final degree Project covers Data Mining, which is of extreme importance in the actual world. It´s importance is based on the idea that people are natural producers of data. This, if it is well systematized and studied, will give companies valuable information that will help them in numerous subjects, such as developing new products or analyze clients, or possible clients, behavior.

At first, companies could only stock the data in paper, moving on later to computers, by creating digital archives. Given the technologies of those times, analyzing them to create out of it useful information for the company was very complicated.

This thesis purpose is to make us realize the amount of data we produce and how it can be used by companies to adapt our likes or behaviors to their products or services to increase our consumption. With that in mind, we will explain in what consist as well as the applications it can be given.

With the help of Weka, we will take to practice de *Data mining*. This widget will be used in the three following exercises, which are linear regression, the Bayesian network and the decision tree. Lastly, a more complex exercise, such as the clustering. This will allow us to understand the behavior of the *Data mining* program in a more visual way.

Keywords: Data Mining, Artificial intelligence, Nodes.

INTRODUCCIÓN

Antes de entrar en materia, me gustaría comentar el motivo que me llevó a seleccionar este tema para mi Trabajo Final de Grado (TFG).

En primer lugar, vivimos en un mundo en el que cada vez hay más datos a nuestro alrededor, los cuales nos pueden dar un montón de información útil, lo cual muchas veces no nos es posible acceder a la información por varios motivos. Porque no somos capaces de ordenar los datos, ni de manejar las grandes bases de datos que se generan. Sin ser consciente de ello, todas las personas somos generadores de datos, muy importantes para las compañías. Por ello, creo que es de vital importancia conocer herramientas que nos faciliten esta tarea. Ya que, con estas herramientas no solo tratan de conocer las necesidades actuales que tenemos, sino que pueden ayudar a las compañías a predecir las necesidades que tendremos en un futuro. Con ellas, las empresas conseguirán ser líderes en la creación de esos productos o servicios que nos satisfagan esas necesidades nacientes.

En segundo lugar, hoy en día a la hora de encontrar trabajo se tiene en cuenta, entre otros muchos factores, el conocer distintas herramientas o sistemas informáticos que nos permitan movernos con mayor facilidad en el mundo global. Por eso, la elaboración de este TFG creo que me podrá abrir algunas salidas laborales que sin estos conocimientos no podría llegar a conseguir.

El objetivo principal por el que nace este trabajo es intentar definir de manera divulgativa la minería de datos, para que así cualquier persona a la que le llegue este documento pueda entenderlo. Así, los que estéis leyendo este documento entenderéis lo importante que son todos los datos que generamos continuamente para las compañías.

Para poder cumplir este objetivo empezaremos definiendo qué es la minería de datos y las distintas herramientas que hay a la hora de utilizar este tipo de tecnologías.

Este TFG se compone de cuatro partes a diferenciar.

En la primera parte se encuentra la parte teórica en la que se trata de explicar, en que consiste el Data Mining y sus características con el fin de que sea capaz de entender la utilidad que tiene para una organización ya sea pública o privada.

En la segunda parte, se trata de forma teórica de introducir al lector hacia el caso práctico de manera que sea capaz de comprender que se hace en cada momento y que puede aportar el software Weka al usuario.

En la tercera parte, se ven unos casos prácticos con los que se pretende guiar al lector dentro del software Weka, con el objetivo de que este pueda practicar y entender los datos que Weka dará al usuario.

Por último, se encuentra la conclusión con la experiencia que ha obtenido con su uso y una opinión sobre cómo va a evolucionar el Data Mining en el ámbito empresarial.

1. DATA MINING

Desde la creación de la minería de datos, numerosos autores han escrito distintas definiciones entre las que podemos destacar las siguientes:

Hand define el Data Mining como "*la extracción de patrones y modelos interesantes, potencialmente útiles y datos en base de datos de gran tamaño"* (Hand, 1998).

Hand, Mannila y Smyth lo enuncian como "*el análisis de grandes volúmenes de datos para encontrar relaciones no triviales, y para resumirlos de manera que sean entendibles y útiles*" (Hand, et al., 2001).

Se puede considerar ambas definiciones muy similares, ya que, en ambas se puede sacar en limpio que se analizan una gran cantidad de datos con el fin de encontrar relaciones que proporcionen una información útil. Es decir, como manejar una gran cantidad de datos para la obtención de información.

Fundamentalmente, nace para ayudar a entender el contenido del conjunto de datos, con el objetivo de utilizar estadística y, en ocasiones, algoritmos de búsqueda similares a la Inteligencia Artificial y a las redes neuronales (Academia Educación, 2017).

1.1. DIFERENCIAS ENTRE DATO, INFORMACIÓN Y CONOCIMIENTO

Las organizaciones dependen de los datos, por lo que una buena gestión de éstos es fundamental, ya que suceden millones de transacciones al cabo de un día. Las organizaciones tienden a almacenar cantidades inmensas de datos, por ello origina que haya dificultades para identificar los datos más importantes. Algunos sectores que dependen de ellos son el sector bancario, sector asegurador, entre otras.

La información es un mensaje que es enviado por el emisor al receptor, por lo que debe ser el receptor, el que considere si la información ha sido recibida correctamente o si el "ruido" ha afectado a la información.

Davenport y Prusak (1999) consideran que el conocimiento es un conjunto de valores, experiencias, información y ``saber hacer´´. El conocimiento se crea y aplica en la mente de los especialistas. Asimismo, en una organización, el conocimiento no solo se localiza en forma de documentos y almacenes de datos, sino que podemos localizarlo en procesos y normas. Cabe destacar, que el conocimiento no es un activo concreto y definible, como se suele pensar. El conocimiento es un activo difícil de controlar.

Para que la información se convierta en conocimiento debe pasar un proceso que contiene varias fases entre las que destacan la comparación, las consecuencias, las conexiones y las conversaciones. La creación de conocimiento se da entre personas (Universidad Nacional Autónoma de México, s.f.).

Los datos son cifras, letras o palabras, que por sí mismas puede que no tengan sentido. En el momento en el que un usuario les encuentra algún significado se convierten en información. Pero se puede ir un paso más allá, los especialistas crean un modelo, con el que consiguen una interpretación de este modelo unido a la información, añadiendo valor agregado, creando el siguiente escalón, el conocimiento (Academia Educación, 2017).

1.2. VENTAJAS Y DESVENTAJAS

La minería de datos es un campo multidisciplinar que cuenta con una serie de ventajas:

- Al mostrar datos verídicos y predecir los futuros resultados previene situaciones adversas.
- Al transformar datos en información contribuye en la toma de decisiones estratégicas de la empresa.
- Facilita la compresión de la información de la compañía, pudiendo comunicarla de manera más sencilla a los distintos usuarios.
- Al conocer mejor las necesidades y comportamientos de los clientes, se mejora la relación con ellos. Además, se consigue reducir las posibles pérdidas de clientes y atraer a clientes potenciales (Reporte Digital, 2018).

Aunque cuente con una serie de ventajas, como cualquier otra tecnología cuenta con una serie de desventajas que os enunciaré a continuación:

- Para poder llevar a cabo este tipo de prácticas es necesario contar con personal cualificado además de realizar una inversión.
- Sentimiento de inseguridad por parte de los usuarios estudiados que aportan los datos.
- Necesidad de mucho tiempo para poder procesar una gran cantidad de datos.
- Si no se cuenta con un sistema de seguridad apropiado, se pone en riesgo la información privada aportada por los usuarios.
- Si los datos no son exactos o correctos, la información y los resultados pueden verse afectados (Reporte Digital, 2018).

A la hora de decidir si implantar o no esta tecnología dentro de la compañía se debe tener en cuenta todas estas cuestiones para ver si la compañía puede conseguir minimizar las desventajas y sacar el máximo provecho de las ventajas.

1.3. CICLO DE VIDA DE UN PROYECTO DE MINERÍA DE DATOS

Los pasos que se siguen son siempre los mismos, sin importar la técnica que se utilice. Es por ello, que podemos decir que el ciclo de vida son los siguientes.

1.3.1. Entendimiento del negocio

Formular los problemas del negocio como previsión, segmentación de clientes, identificar los productos sustitutivos, las empresas competidoras, los proveedores, entre otros factores. De esta manera es más fácil poder detallar los objetivos del proyecto, y evaluar los riesgos que conlleva, por último, solo quedaría elaborar el proyecto (Sngular, s.f.).

1.3.2. Entendimiento de los datos

Una vez elaborado el proyecto, para completar esta fase hay que seguir el siguiente paso:

- Seleccionar los datos: elegir los datos a usar para el análisis, como variables independientes que sirven para el cálculo o las variables objetivo, que son aquellas que se quieren predecir o calcular (jpgarcia.cl, 2008).

1.3.3. Preparación de los datos

- Transformación de los datos

Normalmente no es posible utilizar ningún algoritmo, por ello que es necesario algún cambio.

- Filtrado de datos

Hay que depurar los datos con el fin de eliminar valores y datos erróneos, dependiendo del algoritmo que se vaya a utilizar y de las necesidades que haya.

- Preprocesamiento

Se analizan las gráficas y se adquieren muestras con las que obtener mayor eficiencia y velocidad de los algoritmos o disminuyendo posibles valores mediante el redondeo, la agrupación y la agregación.

1.3.4. Modelado

Para crear un modelo, se seleccionan las variables, de esta manera, atribuyendo características podemos reducir el tamaño de los datos. Existen dos métodos con los que elegir las particularidades de los datos que tengan mayor dependencia respecto al problema. Estos se pueden diferenciar en la selección de aquellos con los atributos que mejor caractericen al problema y aquellos basados en variables independientes mediante algoritmos heurísticos.

La extracción de conocimiento es la parte fundamental de la minería de datos que con una técnica se consigue un modelo de conocimiento por el cual se manifiestan patrones de comportamiento de los valores del problema. Los modelos que se crean son denotados por varias formas como las *reglas*, *árboles*, y *redes neuronales*.

1.3.5. Evaluación

Tras lo mencionado, hay que validar si el modelo ofrece conclusiones válidas. Si se han usado modelos a través de diferentes técnicas, hay que comparar los modelos para observar cual se ajusta mejor al problema.

Si no se alcanza el resultado deseado habrá que cambiar alguna de las fases anteriores para ver si de esta manera se consigue un resultado satisfactorio, lo que implica que se deberá repetir esto último, las veces que sea necesario para que el modelo sea válido.

1.3.6. Implantación

Se establece en aplicaciones para resolver el problema de negocio (Monjas, s.f.).

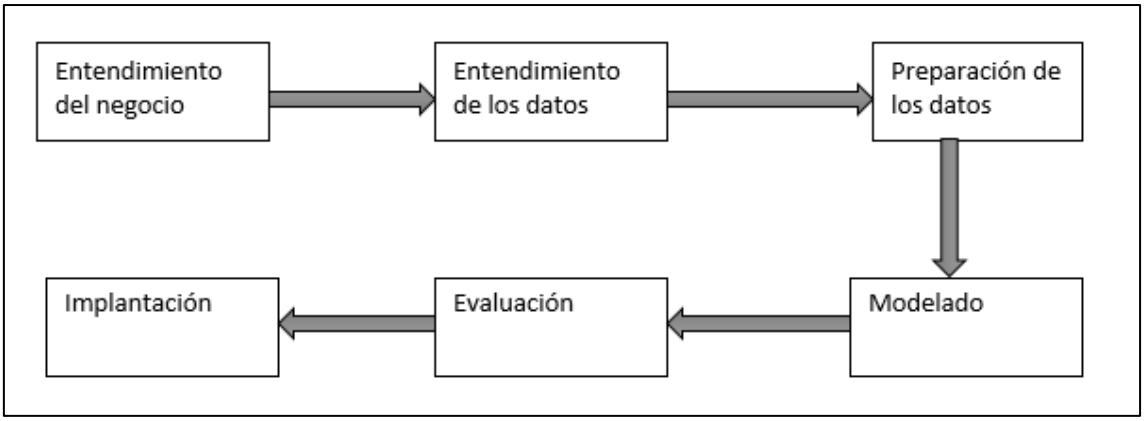


Figura 1.1.: Ciclo de vida de la minería de datos.

Fuente: Elaboración propia.

En la figura 1.1. podemos ver de manera gráfica el ciclo de vida de la minería de datos.

1.4. TÉCNICAS DE LA MINERÍA DE DATOS

Asimismo, las técnicas de la minería de datos consisten en datos de la Inteligencia Artificial y de la estadística lo que originan algoritmos que dependiendo del objetivo que conlleva el análisis de datos, se pueden diferenciar en:

Predicción: Observar una tendencia de los datos.

Supervisados: Predicen un dato que en principio es desconocido, con la ayuda de los datos conocidos.

No supervisados: Se observan las tendencias y patrones de los datos (Monjas, s.f.).

Las técnicas de la minería de datos más simbólicas son las siguientes:

1.4.1. Redes neuronales

Es un sistema de interconexión de neuronas en una red que colabora para crear un estímulo de salida. Puede ser desarrollada por software o hardware y con ello, se puede crear un sistema capaz de aprender, adaptarse, o predecir el estado futuro de algunos modelos.

Estas técnicas son capaces de enfrentarse a problemas que antes solo eran posible solucionarlas mediante el cerebro humano.

Algunos ejemplos de redes neuronales son las redes de Kohonen, el perceptrón, y el perceptrón multicapa (Monjas, s.f.).

Hay dos tipos de redes neuronales:

- **Redes monocapa**

Son redes con una sola capa que para que las neuronas puedan unirse, crean conexiones laterales con las que relacionarse a otras neuronas de esa misma capa (Aguilar, s.f.).

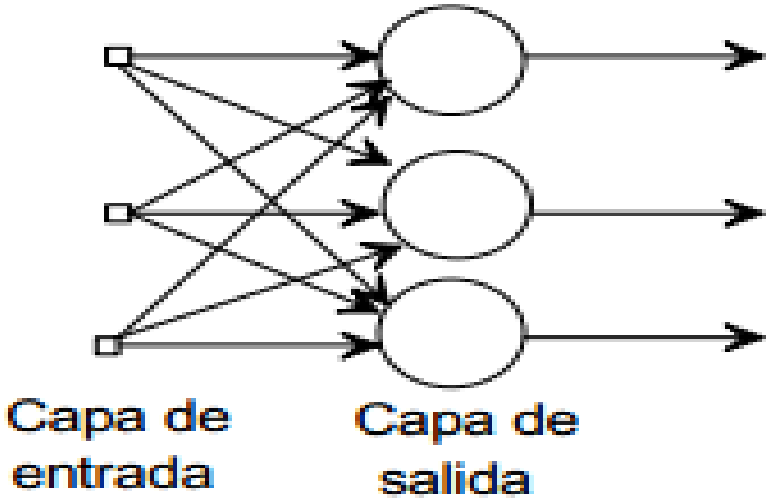


Figura 1.2.: Redes monocapa.

Fuente: Página web: Facultad de Ciencias Exactas de la Universidad Nacional de Tucumán por Gustavo E. Juárez (2017)

En la Figura 1.2. vemos un ejemplo de Redes monocapa.

- **Redes multicapa**

Las redes multicapa están compuestas por varias capas de neuronas, estas pueden clasificarse en varios tipos:

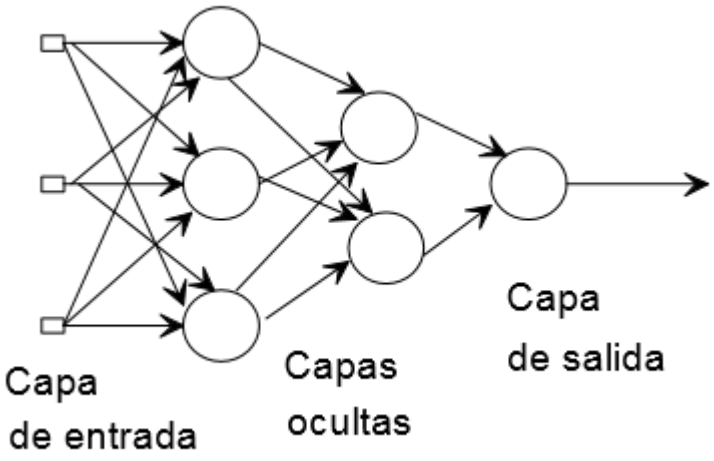


Figura 1.3.: Redes multicapa.

Fuente: Página web: ResearchGate (s.f.)

En la Figura 1.3. se puede ver un ejemplo de redes multicapa, en las que destacan las diferentes capas: capa de entrada, capas ocultas y capa de salida.

- **Redes con conexiones hacia delante o perceptrón multicapa**

Estas redes solo tienen conexiones con las capas de delante.

- **Redes con conexiones hacia atrás o perceptrón**

Son redes que, a diferencia de las anteriores, sí pueden tener una dinámica de la red hacia capas de atrás (Neural Networks Framework, s.f.).

1.4.2. Árboles de decisión

En las tareas del *Data mining* son uno de los algoritmos clasificadores que más se conocen, debido a que tienen una forma sencilla de clasificación (Monjas, s.f.). En esta herramienta se trata de formar grupos de población con el fin de encontrar grupos con características similares según alguna variable de respuesta. Con ello se representará gráficamente una serie de reglas sobre la decisión tomada en la asignación de un elemento a clase o valor de salida (Aguilar, s.f.). Es decir, viendo las características se tratará de unir a los individuos que puedan responder de igual manera. Suelen utilizarse en tareas de clasificación y en menor medida en la predicción (Aguilar, s.f.).

Dentro de un árbol de decisión se podrá encontrar nodos de decisión (interiores), nodos-respuesta (hojas) y arcos (Extracción y recuperación de información, s.f.).

- Nodo decisión: Se asocia a un determinado atributo, del que salen varias ramas que cada una de las ramas representa los posibles valores que puede tomar este atributo (Extracción y recuperación de información, s.f.).
- Nodos-respuesta: Vinculado a la clasificación que se quiere proporcionar y da la decisión del árbol con respecto al ejemplo de entrada (Extracción y recuperación de información, s.f.).
- Arcos: Los distintos valores del nodo (Extracción y recuperación de información, s.f.).

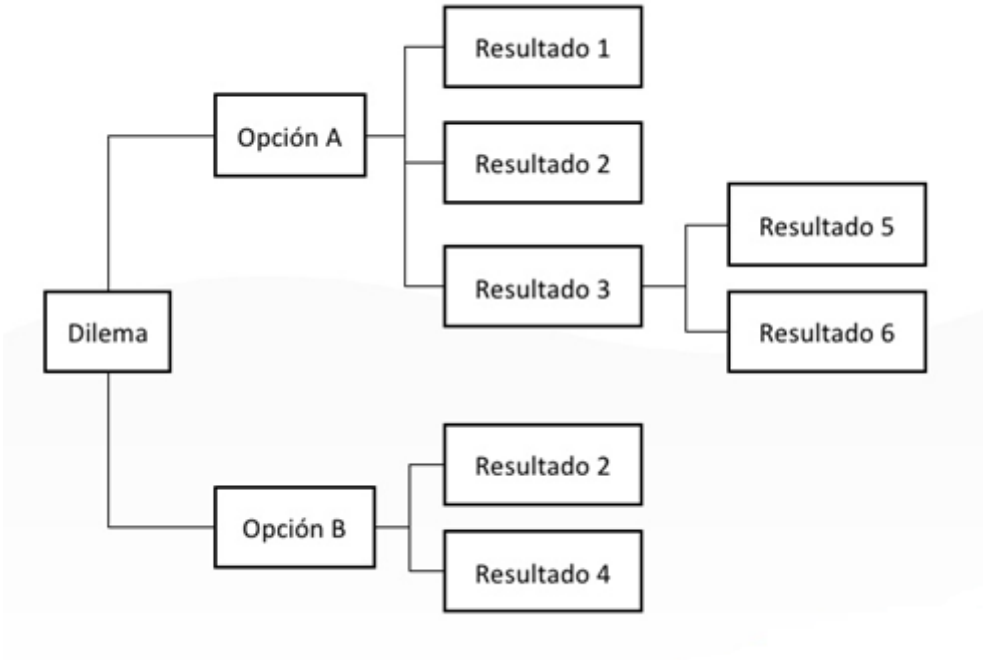


Figura 1.4.: Árbol de decisión

Fuente: Pagina web: Estilos de liderazgo (s.f.)

En la figura 1.4. se puede observar un árbol de decisión, en el que se es capaz de ver claramente cada una de las partes explicadas previamente. El dilema es el nodo de decisión, las opciones A y B son las distintas respuestas que puede dar este nodo, llamados nodos-respuestas. Los resultados son los distintos arcos que puede tomar los valores de este árbol de decisión.

1.4.3. Modelos estadísticos

La minería de datos utiliza modelos estadísticos con los que intenta explicar y predecir la conducta de los individuos, esto supone que la base de datos puede ser capaz de ejercer de agenda que notifique las tareas a realizar y el momento exacto de realizarlas (Neuronics, s.f.).

Un modelo estadístico podría definirse como una ecuación matemática que recrea las observaciones de la manera más exacta (Ocaña, s.f.).

1.4.4. Agrupamiento o *clustering*

Es un algoritmo no supervisado que tiene el objetivo de detectar posibles agrupamientos en el conjunto de los datos, identificando a los individuos parecidos en grupos homogéneos (Monjas, s.f.). Esta técnica es aplicada para agrupar a los clientes potenciales, según las visitas y comportamiento en una web, estableciéndoles unas características diferenciadoras entre agrupaciones y agrupaciones (Conceptos claros, s.f.). Algunos ejemplos de esta técnica son algoritmo *K-means* o el algoritmo *K-medoids* (Monjas, s.f.).

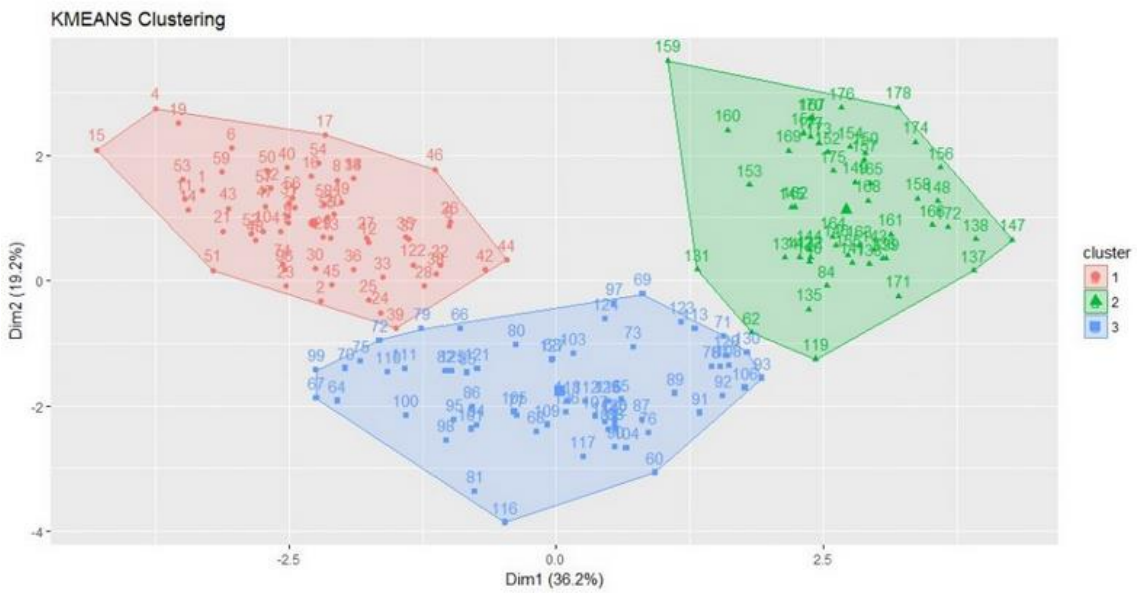


Figura 1.5.: Clustering con el algoritmo K-means

Fuente: Pág. Web: Conceptos Claros

En la figura 1.5. se puede observar un ejemplo de la técnica *Clustering* con el método *K-means*. De él cabe destacar que existen tres agrupaciones bien diferenciadas, cada una de ellas de un color distinto (rojo, verde y azul).

1.4.5. Algoritmos genéticos

Son unos algoritmos, inspirados en los procesos de evolución natural y evolución genética, de optimización búsqueda y aprendizaje (Ventura, 2013/14). Son métodos que se basan en la evolución que suministran nuevas maneras de trabajar con cierto tipo de problemas (Monjas, s.f.).

Para utilizarlos no es necesario conocer el sistema interno, pero si se debe saber el conocimiento de las salidas del sistema y efectos para poder elegir las mejores soluciones posibles. En ellos existe una "función de evaluación" que implica que en el momento que se alcanza el objetivo llegando a la solución el proceso queda congelado (Monjas, s.f.).

1.4.6. Machine learning

El *machine learning* es un procedimiento que utiliza el análisis de datos, con el que intenta mecanizar para crear modelos analíticos. El *machine learning* Es una rama de la Inteligencia Artificial basada en la creencia de que los sistemas pueden estudiar datos, localizar patrones, y tomar decisiones con la más mínima presencia de alguna persona.

El *machine learning* se creó a partir del reconocimiento de patrones y de la teoría con la que ordenadores pueden aprender sin necesidad de ser programadas en actividades específicas. Los investigadores querían averiguar si los ordenadores consiguieran aprender los datos gracias a la ayuda de la Inteligencia Artificial. Esto tiene una gran importancia debido a que a través de que los modelos van aprendiendo nuevos datos, estos se adaptan de manera autónoma. Aprenden cálculos anteriores para originar decisiones y resultados repetibles y verídicos.

La aplicación automática de cálculos matemáticos complejos en el *Big Data* es una posibilidad reciente pese a la existencia de algoritmos de aprendizaje desde hace tiempo.

La importancia del *machine learning* se debe a factores comunes con el *Data Mining* y el análisis Bayesianos. Gracias a ello, se pueden construir modelos veloz y mecánicamente con las que analizar datos muy grandes, complejos y obtener resultados más precisos y de manera más rápida.

Para obtener sistemas de *machine learning* con los que sacar el máximo provecho es necesario tener recursos con los que poder preparar los datos. Además, es importante poder utilizar algoritmos básicos y avanzados. Y, por último, poder automatizar los procesos repetitivos.

Las industrias que trabajan constantemente con una gran cantidad de datos reconocen la importancia que tiene este método, por ello que los siguientes sectores suelen utilizar el *machine learning* para conseguir ventajas respecto a competidores:

- **Servicios financieros:** Ayuda a prevenir el fraude, localizan las oportunidades de inversión e identifica los clientes de mayor riesgo.

- **Atención a la salud:** Permite evaluar la salud de los pacientes ayudando a analizar datos e identificando tendencias.

- **Gobierno:** Permite analizar datos con los que aumentar la eficiencia y ahorrar dinero. También, permite identificar fraudes y evitar el robo de identidad.

- **Marketing y ventas:** Analiza el historial de compras, permitiendo promocionar productos que podrían interesar más a cada cliente, y de esta manera personalizar su experiencia.

- **Petróleo y gas:** Otorga la capacidad de ser más eficiente en cuanto a la distribución del petróleo y permite encontrar nuevas fuentes de energía.

- **Transporte:** Identifica patrones y tendencias con las que localizar las rutas más eficientes y evitar futuros problemas con los que aumentar la rentabilidad (SAS, s.f.).

Los métodos más usados del *machine learning* son los siguientes:

- **Algoritmos de aprendizaje supervisado:** son utilizados en el etiquetado, donde con la entrada se sabe el resultado deseado. Este algoritmo se utiliza mandándole un conjunto de entradas junto con sus resultados correctos, el algoritmo puede encontrar errores comparando los resultados correctos con el real. Para corregir estos errores modifica el modelo. Este algoritmo utiliza métodos como la clasificación, regresión, predicción y aumento de gradiente, para usar patrones con el fin de predecir valores de la etiqueta en datos adicionales no etiquetados. En la actualizar es usado en aplicaciones en las cuales los datos históricos intuyen futuros acontecimientos. Un ejemplo de esto sería para localizar transacciones fraudulentas con tarjetas de crédito.

- **El aprendizaje semi-supervisado:** Se usan las aplicaciones que también se usa en el aprendizaje supervisado. La diferencia con el anterior punto se encuentra en que se utilizan datos etiquetados con una gran ración de datos no etiquetados. Esto se debe a que los datos no etiquetados son menos costosos y más fáciles de obtener. Este modo de aprendizaje se puede compaginar con métodos como la clasificación, regresión y predicción. El aprendizaje semi-supervisado es de gran utilidad si el coste agregado al etiquetado es muy alto como para impedir un proceso completamente etiquetado.

- **El aprendizaje no supervisado:** El objetivo principal consiste en analizar datos para localizar una estructura en su interior. Por lo que el aprendizaje no supervisado es ideal con datos de transacciones. Un ejemplo de esto sería identificar segmentos de clientes con características parecidas con las que poder tratarlos de manera similar en campañas de *marketing*. Varias de las técnicas más usadas contienen mapas con organización automática, *mapping, k-means, clustering* y descomposición de valores. Estos algoritmos también tienen capacidad para segmentar temas de texto, localizar valores inusuales y recomendar elementos.

- **El aprendizaje con refuerzo:** Se compone de tres partes primordiales: el agente (es el que aprende o también el que toma decisiones), el entorno (todo aquello con lo que interactúa el agente) y las acciones (lo que es capaz de hacer el agente). Suele ser utilizado en la robótica, juegos y navegación. Este algoritmo trata crear recompensas mediante ensayos y errores. Principalmente se busca que el agente realice acciones con las que maximizar la recompensa. El agente logrará el objetivo de manera más eficaz y rápida si aplica una correcta política.

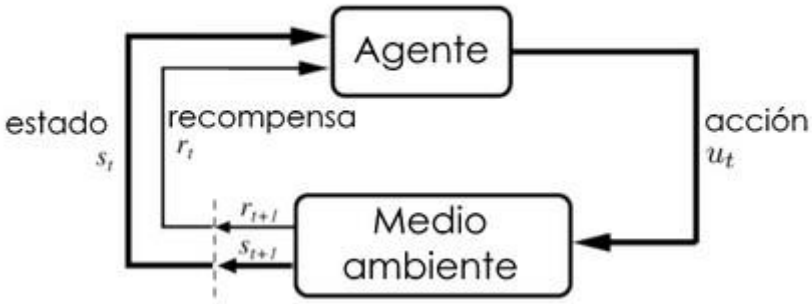


Figura 1.6.: Aprendizaje con refuerzo.
Fuente: Pág. Web: Sutton y Barto (1998)

En la Figura 1.6 podemos observar de manera resumida lo que equivaldría el aprendizaje con refuerzo.

La diferencia entre *Data Mining, Machine Learning y Deep Learning* son los siguientes:

El *Data Mining* usa métodos de bastantes áreas con las que localizar patrones de datos desconocidos. El *Machine Learning,* por el contrario, utiliza un proceso repetitivo con el que pretende entender los datos, el cual el proceso de aprendizaje puede ser automático. Por último, el *Deep Learning* mezcla avances informáticos con las redes neuronales para aprender y entender patrones en cantidades inmensas de datos (SAS, s.f.).

1.4.7. Regresión

La regresión se encarga de identificar una relación entre variables numéricas y variables dependientes. La relación es mostrada por un modelo funcional $y = f(x_1,\ldots,x_n)$. La forma más fácil es cuando solo hay una variable independiente X, lo que se conoce por regresión lineal simple (Aprende con Alf, s.f.).

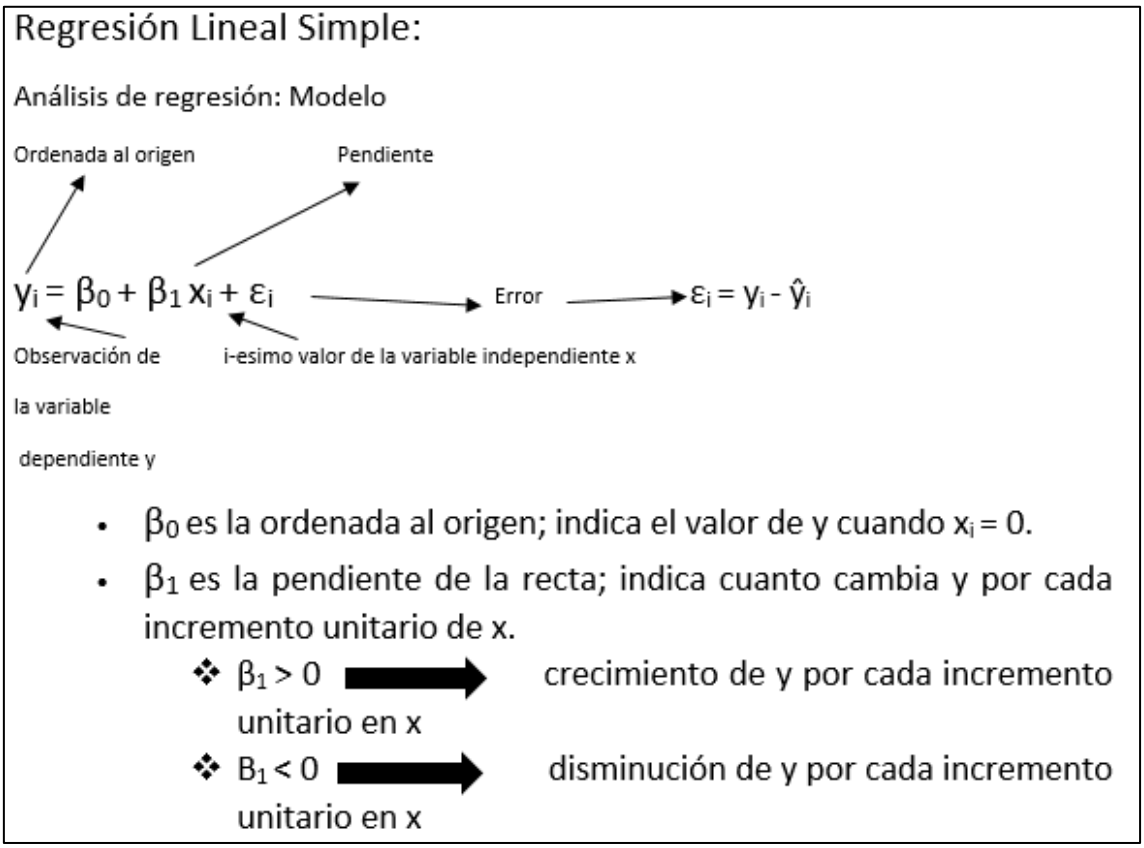


Figura 1.7.: Regresión Lineal Simple.

Fuente: Pág. Web: SlidePlayer.

En la Figura 1.7. se puede observar la Regresión Lineal Simple de manera visual con pequeñas definiciones que aclaran los conceptos.

La Regresión Lineal Múltiple también es una técnica con la que analizar hipótesis y relaciones causales, pero se diferencia respecto a la Regresión Lineal Simple en que la variable dependiente y las variables independientes tienen que ser ordinales o escalares, y las variables independientes no deben tener una alta correlación entre sí (Cardenas, 2014).

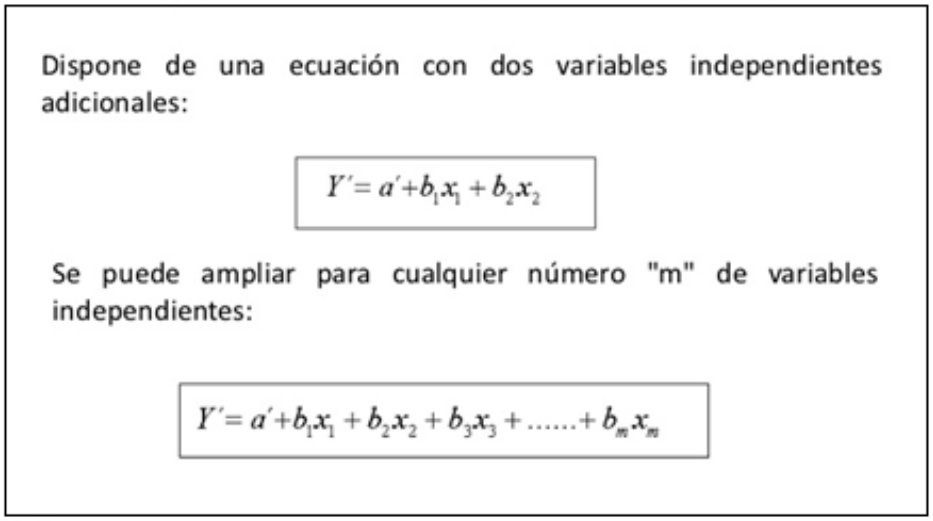


Figura: 1.8.: Regresión Lineal Múltiple.

Fuente: Pág. Web: SlideShare de Leonardo López C.

En la Figura 1.8. se puede observar un ejemplo visual de lo que sería la Regresión Lineal Múltiple.

1.5. APLICACIONES

En este apartado se centra en nombrar y explicar brevemente las distintas aplicaciones de la minería de datos que ayudan en la toma de decisiones de las compañías.

1.5.1. Detención de fraudes

Con la minería de datos se intenta acabar con las transacciones de blanqueo o fraudes con tarjetas de crédito y telefonía móvil. Este tipo de operaciones suelen regirse por una serie de patrones que permiten distinguirlas de las operaciones legítimas. Por ello las compañías emplean algoritmos para poder clasificar las operaciones en posibles fraudes y operaciones legítimas para poder tomar las decisiones pertinentes (Monjas, s.f.).

1.5.2. Recursos humanos

El departamento de RRHH dentro de una compañía es uno de sus pilares fundamentales, ya que, en él se selecciona al personal que trabaja en la misma, y una mala selección puede suponer una disminución de la producción y un coste tanto económico como una pérdida de tiempo.

Por ello, algunas empresas deciden implantar este tipo de tecnología para identificar las características y capacidades tanto de los empleados como de los solicitantes de empleo para así poder seleccionar a los empleados que mejor se adapten a las necesidades de la compañía y colocarles en el puesto en el que mejor resultados puedan obtener (Monjas, s.f.).

1.5.3. Terrorismo

El ejército de Estados Unidos utilizó esta técnica para identificar a posibles terroristas. Un año antes del atentado del 11 de septiembre consiguió identificar a 4 de los miembros de Al Qaeda (Monjas, s.f.).

1.5.4. Genética

Con el *Data Mining* se ha mejorado la forma de diagnóstico, prevención y tratamiento de algunas enfermedades, como el cáncer. A estas aplicaciones se las conoce como "reducción de dimensionalidad multifactorial" (Monjas, s.f.).

1.5.5. Ingeniería eléctrica

En este sector se utiliza este tipo de tecnologías para el análisis de anomalías como puede ser el análisis de cambios de carga en los trasformadores o el estado del aislamiento de los equipos eléctricos, entre otro tipo de aplicaciones (Monjas, s.f.).

1.5.6. Detección de hábitos de compras en supermercados

Analizando que productos se compran más en determinados días de la semana permite observar patrones en la compra de los clientes que cambiando la colocación de los productos permite el fomento por parte de los supermercados la compra compulsiva de sus clientes (Monjas, s.f.).

1.5.7. Bioinformática

Se podría definir la bioinformática como la mezcla entre la ciencia de la vida y la ciencia de la información. Es decir, aporta herramientas informáticas a la biomedicina. La minería de datos, en este campo, intenta acumular los datos de distintas investigaciones científicas de años para conseguir analizar el comportamiento de las células vivas, minimizando la acción del ser humano (Monjas, s.f.).

1.5.8. Web Mining

Trata de captar los datos que dejamos cuando entramos en distintas páginas web. Extrae datos acerca de las veces que visitan una web, que elementos de la web son los más visitados, tipo de software que emplean los usuarios o desde que enlace o web acceden (Monjas, s.f.). Esto permite analizar al propietario de la web si merece la pena pagar o no a distintos portales como Google para colocar su página en un lugar privilegiado dependiendo del tráfico que le genere.

2. METODOLOGÍA

Entre las aplicaciones que se usan en el Data Mining, hemos elegido el Weka para el desarrollo del caso práctico del Trabajo Fin de Grado debido a la importancia que ha adquirido esta aplicación para el ámbito empresarial.

2.1. HERRAMIENTA WEKA

Figura 2.1.: Logo de WEKA

Fuente: Pág. web Analytics Vidhya

Weka es un software libre y de código abierto, que fue desarrollado por la universidad de Waikato en 1993. Este programa contiene un conjunto de algoritmos, herramientas de visualización y modelos predictivos, con lo que solventar los problemas que se originan del Data Mining. Las siglas Weka representan al “Waikato Environment for Knowledge Analysis”.

El programa cuenta con una interfaz sencilla desarrollada por Java. Además, se inicia con GUI, permitiendo elegir cinco aplicaciones de las que cuenta el programa. Estas son las siguientes:

2.1.1. Explorer

Contiene una interfaz de pestañas para la carga y el preprocesado y filtrado de datos. La aplicación cuenta con distintos algoritmos de minería de datos lo que permite la clasificación, *clustering* y asociación. Para que esto sea más visible para los usuarios, cuenta con distintas herramientas que permiten ver los datos de manera visual, ya sean gráficos, diagramas, entre otros.

2.1.2. Experimenter

Tiene unas funcionalidades similares a la aplicación anteriormente explicada, aunque se centra en tareas en un horizonte temporal a más largo plazo, ya que, desde ella se pueden guardar la sesión en la que se está trabajando para retomarla en otro momento. Esta aplicación se usa principalmente para la comparación en la precisión de distintos algoritmos trabajando simultáneamente con ellos.

2.1.3. Knowledge Flow

El usuario puede acceder para seleccionar varios componentes y arrastrarlos a un canvas con el que se tiene la oportunidad de diseñar distintos procesos o flujos mediante diagramas, lo que conlleva una ayuda para entender sobre el proceso del análisis de datos de manera más visual.

2.1.4. Workbench

Coordina en una sola interfaz las distintas interfaces de Weka, mejorando la experiencia del usuario, dejándoles acceder a todas las aplicaciones desde una única ventana.

2.1.5. Simple CLI

Da la posibilidad de acceder a las funcionalidades de Weka desde la interfaz de línea de comandos igual a la que se usaría para efectuar el código Java (Sainz, 2019).

2.2. EXCEL

Excel es un programa informático creado y distribuido por la empresa Microsoft Corp., constituida por Bill Gates. Con este programa realizaremos tareas contables y financieras a través del uso de hojas de cálculo.

Se creó en 1985 y desde entonces se han publicado distintas versiones y actualizaciones con el fin de dar al usuario un mejor servicio (Concepto.de, 2019).

2.3. FUENTES DE INFORMACIÓN

Con el objetivo de hacer una demostración práctica del programa Weka, se utilizará distintos manuales del programa, como:

- Manual Weka Versión 3-6-15 de la Universidad de Waikato, que son los creadores de esta herramienta.
    - Tutorial Weka Universidad Carlos III de Madrid.
    - Introducción al WEKA del curso de Doctorado Extracción Automática de Conocimiento en Bases de Datos e Ingeniería del Software de la Universitat Politècnica de València.
    - Manual Weka de Diego García Morate (Morate, s.f.).

Además de los distintos manuales para el empleo del programa, se utilizará distintas bases de datos para poder trabajar con el programa. Las bases de datos son las siguientes:

- Base de datos del precio medio del m^2 de San Sebastián procedentes de Eustat.

3. CASOS PRÁCTICOS

3.1. REGRESIÓN LINEAL

3.1.1. Estimar el precio de la vivienda para el año 2025

En este primer caso práctico se ha utilizado una base de datos acerca del precio por m^2 de un piso nuevo de San Sebastián. Esta base de datos pertenece a Eustat. Este caso busca predecir su precio por m^2 en el año 2025, teniendo en cuenta los datos desde el año 1994 hasta el 2019. Con el fin de obtener una ecuación que de la capacidad de predecir el precio por m^2, se exporta la base de datos al programa Weka. Para ello es necesario cambiar la extensión a CSV. Antes de exportarlo, también es obligatorio abrir el documento mediante un blog de notas con el objetivo de cambiar los “;” por comas y las comas que haya por puntos. De esta forma, en el momento de exportar el Excel a

Weka, se obtiene esta vez, dos atributos que en nuestro caso serán año y precio. Una vez exportado el archivo, vamos al apartado *classify* (debido a que la regresión lineal es un algoritmo de minería de datos de tipo clasificatorio). Además, en la opción *choose, y en el apartado functions* se seleccionará la regresión lineal. Por último, en la tabla que pone *Test options*, seleccionaremos la opción que se titula *use training set* y finalmente se le dará a la opción *start.* Tras haber ejecutado las instrucciones aparece la información que se muestra en la imagen de abajo, la cual se utilizará el modelo de regresión lineal que ha calculado Weka.

```
Linear Regression Model

precio =

     98.6003 * anyo +
-194911.5859

Time taken to build model: 0.26 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0.02 seconds

=== Summary ===

Correlation coefficient                  0.7717
Mean absolute error                    508.3766
Root mean squared error                609.5425
Relative absolute error                 59.8036 %
Root relative squared error             63.6043 %
Total Number of Instances               26
```

Figura 3.1.: Resultado de la opción de regresión lineal.

Fuente: Weka.

En la figura 3.1. se puede observar el resultado de los pasos explicados anteriormente para la obtención de una ecuación que permite predecir el precio futuro del m^2 en la ciudad de San Sebastián. De esta manera se destaca la ecuación "Precio = 98,6003 * año –194911,5859", como se puede observar, es una fracción que es ascendente, cuanto más tiempo pase, el valor del m^2 en esta ciudad ascenderá. Si por ejemplo se calculara el precio estimado para el año 2025 como se menciona en un principio, se escribe la formula en una hoja de cálculo de Excel y este dará un resultado de 4.754 €/m^2.

3.1.2. Comparativa del precio real y precio estimado para los años 1994-2019

En este apartado se usa la misma ecuación hallada en el apartado anterior con el programa Weka y se crea una tabla en una hoja de cálculo en Excel en la que se podrá comparar el precio real y el estimado (calculado con la ecuación). Esto da lugar a la siguiente tabla:

Año	Precio real	Precio estimado
1994	1.400	1.697,4123
1995	1.668	1.796,0126
1996	1.514	1.894,6129
1997	1.526	1.993,2132
1998	1.418	2.091,8135
1999	1.597	2.190,4138
2000	2.043	2.289,0141
2001	2.375	2.387,6144
2002	2.340	2.486,2147
2003	2.437	2.584,815
2004	2.861	2.683,4153
2005	3.455	2.782,0156
2006	3.594	2.880,6159
2007	4.194	2.979,2162
2008	4.243	3.077,8165
2009	4.372	3.176,4168
2010	4.033	3.275,0171
2011	3.876	3.373,6174
2012	3.682	3.472,2177
2013	3.523	3.570,818
2014	3.405	3.669,4183
2015	3.281	3.768,0186
2016	3.148	3.866,6189
2017	3.309	3.965,2192
2018	3.408	4.063,8195
2019	3.478	4.162,4198

Tabla 3.1.: Comparativa de precios.

Fuente: Elaboración propia, a partir de la base de datos Eustat.

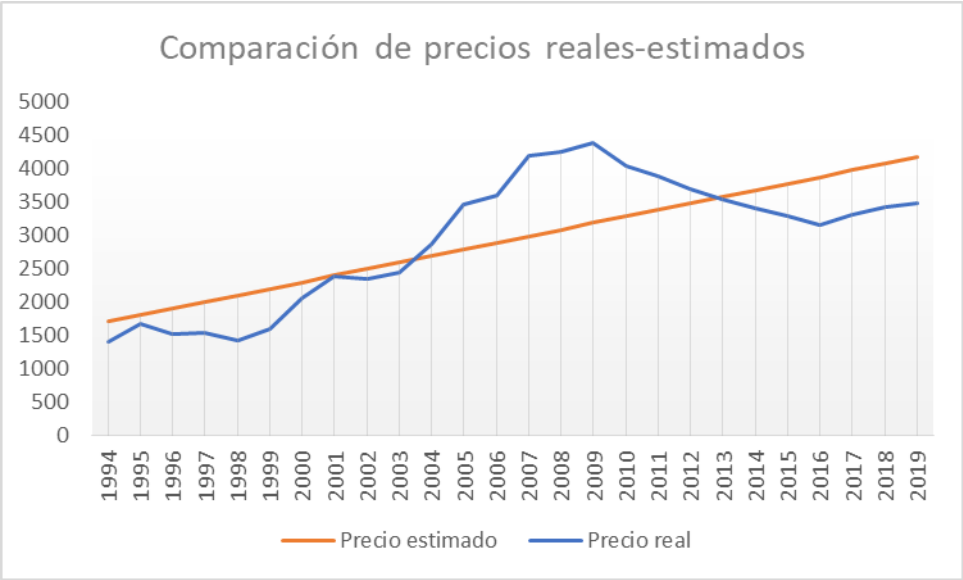


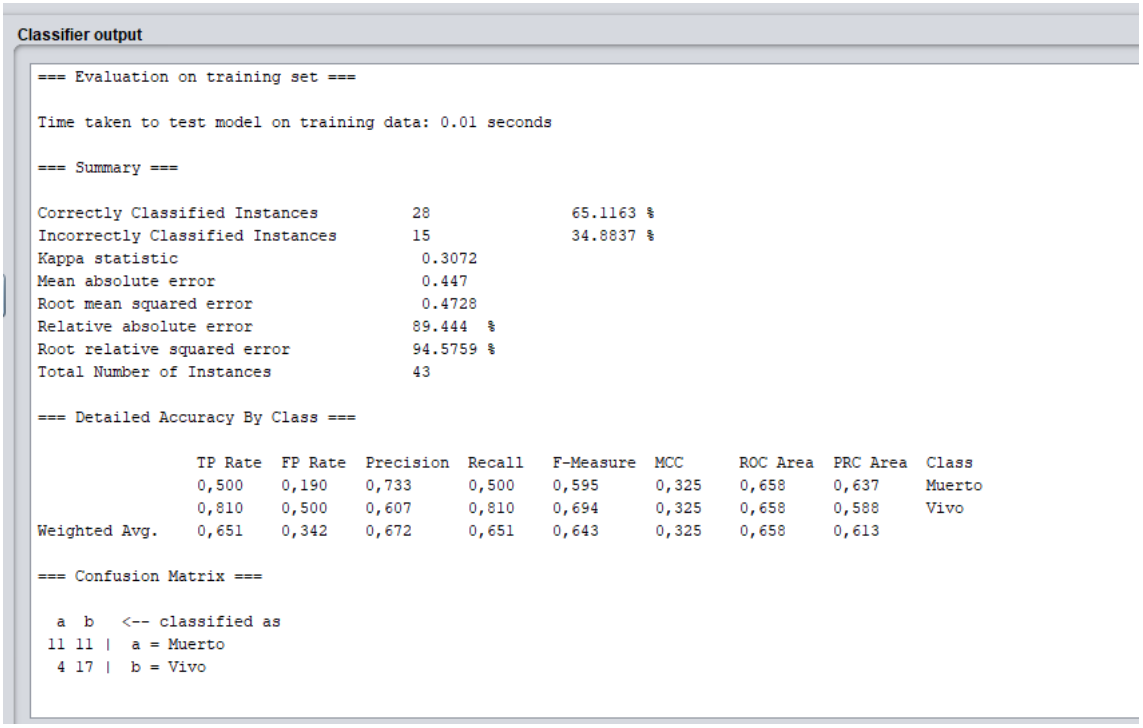
Gráfico 3.1. Comparativa de precios.

Fuente: Elaboración propia.

En el gráfico 3.1. se puede observar de una manera más visual la evolución de los precios. Como se ve, hay en varios momentos que las líneas de precios se unen, y es en esos momentos en los que el precio real es el mismo que el precio estimado, por lo que en ese momento el error es 0. En cambio, en otros momentos el precio real se sitúa tanto por debajo como por encima, esto se debe a que el modelo de regresión lineal solo tiene dos betas. El modelo será más realista cuantas más variables tengamos en cuenta a la hora de estimar el precio, pero a la vez sería más complicado de desarrollar.

3.2. ÁRBOLES DE DECISIÓN

Para realizar el caso práctico dedicado a los árboles de decisión, se decide utilizar una base de datos de un accidente aéreo. En la base de datos se puede clasificar a las personas en el sexo (Hombre o Mujer), Estado civil (Soltero, casado o Viudo) y en su salud (Muerto o Vivo). Para exportar el documento se vuelve a cambiar la extensión del archivo a CSV. Además, se deberá abrir el archivo mediante el blog de notas, y cambiar los “;” a comas. De este modo, se puede comprobar que se tiene tres atributos, los cuales son Sexo, Estado civil y Salud. A continuación, se va a *classify,* en el apartado *choose* se elegirá el apartado *Trees* y a su vez, se selecciona el apartado j48. Tras lo anterior, se selecciona en la tabla *Test options* el apartado llamado *Use training set.* Y finalmente, se le dará a *start.* Lo que el programa Weka devolverá la siguiente información en forma de tabla.



Classifier output

```
=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances          28               65.1163 %
Incorrectly Classified Instances        15               34.8837 %
Kappa statistic                          0.3072
Mean absolute error                      0.447
Root mean squared error                  0.4728
Relative absolute error                 89.444  %
Root relative squared error             94.5759 %
Total Number of Instances               43

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,500    0,190    0,733      0,500    0,595      0,325    0,658     0,637     Muerto
                 0,810    0,500    0,607      0,810    0,694      0,325    0,658     0,588     Vivo
Weighted Avg.    0,651    0,342    0,672      0,651    0,643      0,325    0,658     0,613

=== Confusion Matrix ===

  a  b   <-- classified as
 11 11 |  a = Muerto
  4 17 |  b = Vivo
```

Figura 3.2.1.: Resultado del árbol de decisión.

Fuente: Weka.

En la Figura 3.2.1. se puede observar que el 65,1163% son datos confiables.

La siguiente imagen muestra el árbol de decisión de una manera más visual.

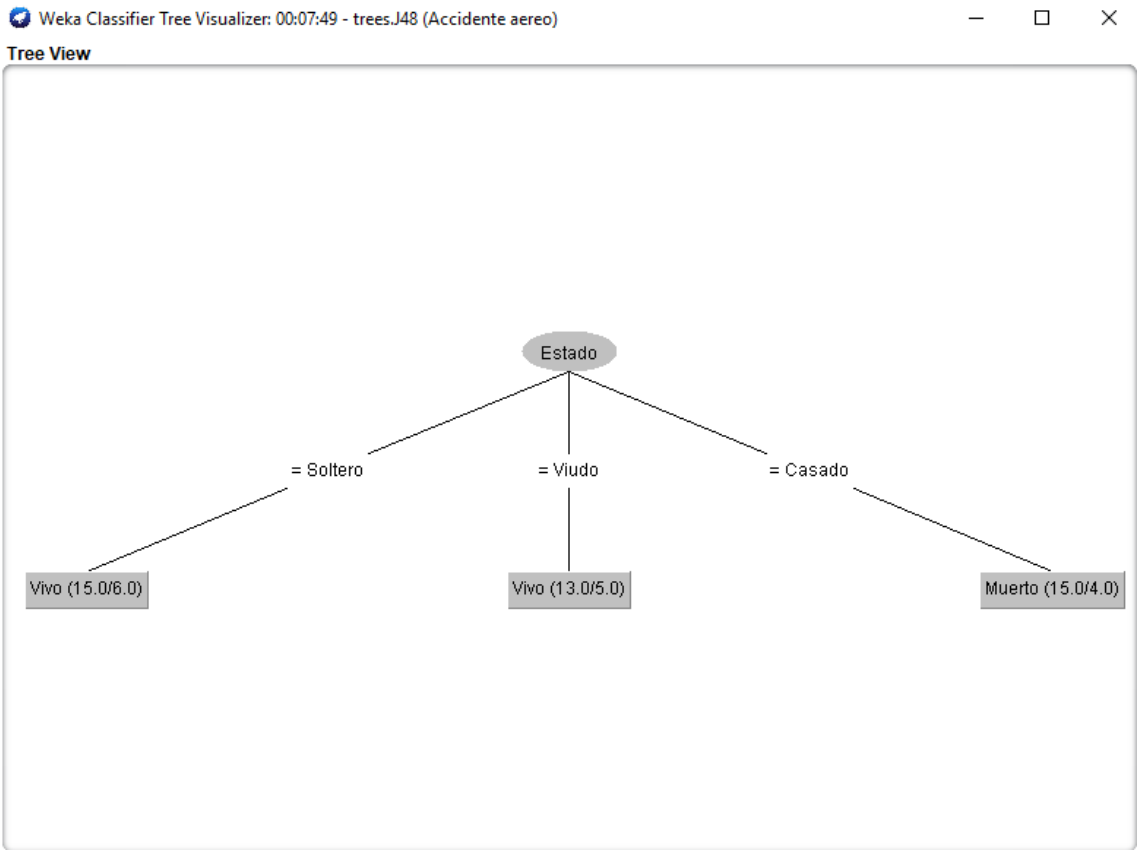


Figura 3.2.2.: Árbol de decisión del accidente aéreo.

Fuente: Weka.

En la Figura 3.2.2. se puede observar que de las personas que sobrevivieron al accidente aéreo, el sexo de la persona no era un factor determinante. Sin embargo, el estado civil de la persona si es el factor más determinante. En el caso de que la persona estuviera soltera, sobrevivieron 6 personas de las 15 personas solteras. De las personas viudas, sobrevivieron 5 personas del total de viudos que eran 13. Y de Casados murieron 4 personas de un total de 15 casados.

3.3. REDES BAYESIANAS

Para la tercera práctica se utilizará las redes Bayesianas del Weka. En este caso, se va a utilizar la misma base de datos que se creó con relación al accidente aéreo del apartado anterior. La única diferencia para analizar la base de datos respecto al apartado anterior consiste en que en vez de entrar en el apartado *Trees,* se va a seleccionar el apartado BayesNet.

Tras analizar la base de datos, el programa Weka, facilita la siguiente información:

Classifier output

```
=== Evaluation on training set ===

Time taken to test model on training data: 0.01 seconds

=== Summary ===

Correctly Classified Instances          28               65.1163 %
Incorrectly Classified Instances        15               34.8837 %
Kappa statistic                          0.3072
Mean absolute error                      0.4503
Root mean squared error                  0.4737
Relative absolute error                 90.0975 %
Root relative squared error             94.7587 %
Total Number of Instances               43

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,500    0,190    0,733      0,500    0,595      0,325    0,632     0,654     Muerto
                 0,810    0,500    0,607      0,810    0,694      0,325    0,632     0,555     Vivo
Weighted Avg.    0,651    0,342    0,672      0,651    0,643      0,325    0,632     0,606

=== Confusion Matrix ===

  a  b   <-- classified as
 11 11 |  a = Muerto
  4 17 |  b = Vivo
```

Figura 3.3.1.: Resultado de la red bayesiana.

Fuente: Weka.

En la Figura 3.3.1. se puede observar que el 65,1163% de los datos son confiables.

A continuación, para ver la gráfica, se clicará con el botón derecho del ratón sobre el resultado que se tendrá en la tabla titulada *Result list* y se le dará a visualizar gráfica.

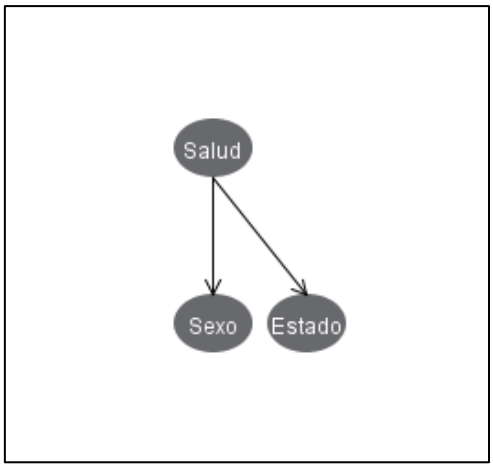


Figura: 3.3.2: Grafica red bayesiana.

Fuente: Weka.

En la figura 3.3.2. se puede observar la gráfica de la base de datos creada sobre un accidente aéreo.

Probability Distribution Table For Salud

Muerto	Vivo
0,511	0,489

Figura 3.3.3.: Probabilidad de sobrevivir al accidente aéreo.
Fuente: Weka.

En la figura 3.3.3. se puede observar que la probabilidad de sobrevivir al accidente aéreo es del 48,9%, mientras que la de fallecer es del 51,1%.

Probability Distribution Table For Sexo

Salud	Mujer	Hombre
Muerto	0,457	0,543
Vivo	0,432	0,568

Figura: 3.3.4.: Probabilidad de sobrevivir o de morir al accidente aéreo según el sexo.
Fuente: Weka.

Según la figura 3.3.4. se observa la probabilidad de ser hombre o mujer en el accidente aéreo según el estado de salud. Si se encuentra un individuo muerto es más probable que se trate de un hombre (54,3%) que una mujer (45,7%). Si se trata de un individuo vivo también tiene más posibilidad de ser hombre (56,8%) que de ser mujer (43,2%).

Probability Distribution Table For Estado

Salud	Soltero	Viudo	Casado
Muerto	0,277	0,234	0,489
Vivo	0,422	0,378	0,2

Figura: 3.3.5.: Probabilidad de sobrevivir y de fallecer según el estado civil.

Fuente: Weka.

Según la figura 3.3.5. es similar a la anterior, pero en vez de analizarlo por sexo, se analiza la posibilidad de su estado civil según su estado de salud. Si se encuentra ante un individuo muerto es más probable que se trate de una persona casada (48,9%) que, en cualquiera de los otros dos estados civiles, siendo menos probable que sea una persona viuda (23,4%).

En cambio, si se localiza una persona viva tiene mayor posibilidad de ser una persona soltera (42,2%) que, en los otros estados, siendo la que menos posibilidades tiene una persona casada (20%).

3.4. *CLUSTERING*

Por último, se verá un caso práctico sobre el *clustering* o agrupamiento. En este apartado se continuará usando la base de datos correspondiente al accidente aéreo. Tras exportar la base de datos al programa Weka, como se hizo anteriormente, se va al apartado *Cluster* y en el apartado *Choose* se elige la opción EM. Tras elegir en el *Cluster mode* la opción *Use training set*, se le dará a *start.*

Tras analizar la base de datos, Weka se mostrará la siguiente información:

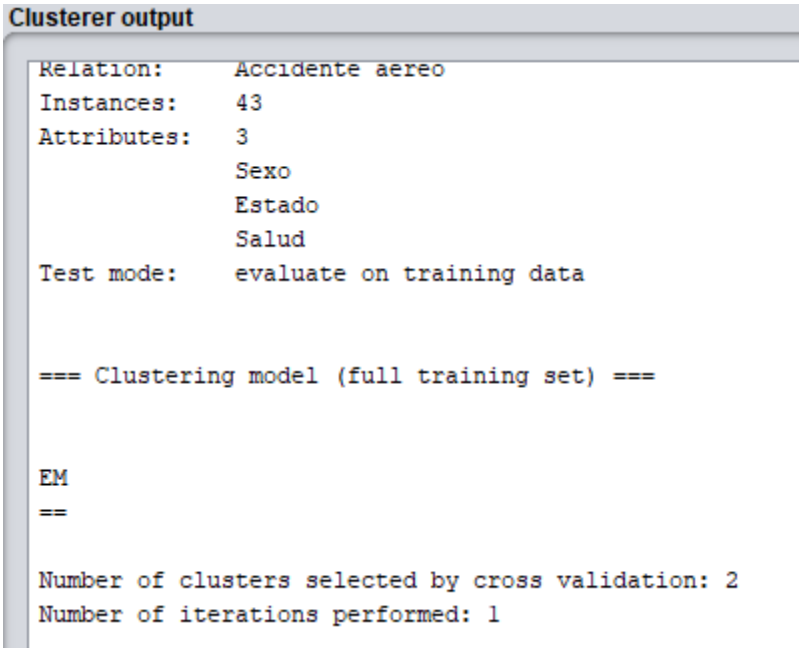


Figura 3.4.1: Resultado del Clustering.

Fuente: Weka.

En la figura 3.4.1. se puede observar en *Number of clusters selected by cross validation* que los datos se han agrupado en 2 asociaciones.

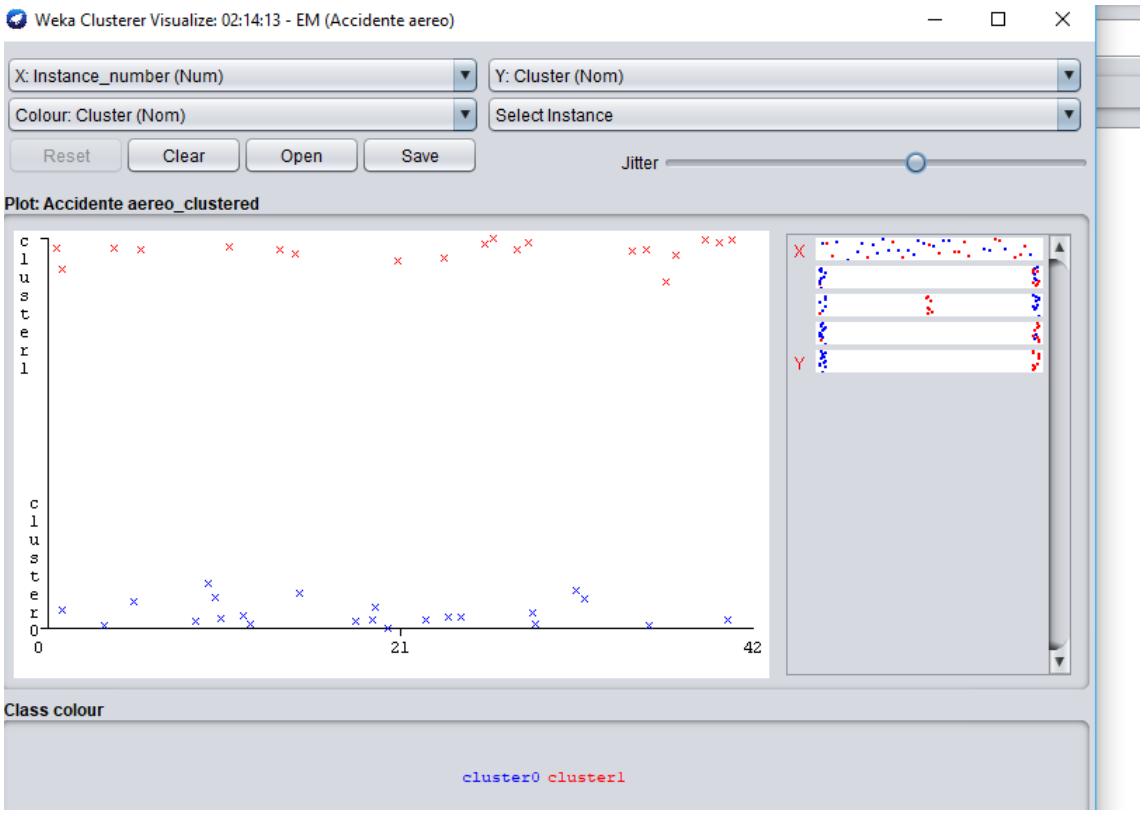


Figura: 3.4.2.: Grafica de clustering.

Fuente: Weka.

En la figura 3.4.2. se ve una representación gráfica de las dos agrupaciones. Estando la agrupación de hombres arriba de la gráfica y la de las mujeres en la parte inferior de la gráfica. Además, se observa hasta el puesto 21 del *cluster* de mujeres, que está más condensado que el de los hombres, lo que significa que hasta el puesto 21, hubo más usuarios de sexo femenino.

4. CONCLUSIONES

Mediante el transcurso de este Trabajo Fin de Grado se han aplicado conocimientos proporcionados durante el Grado en Administración y Dirección de Empresas. Aunque principalmente usamos conocimientos adquiridos en Econometría y en Sistemas de Información.

Teniendo en cuenta el objetivo por el que se ha hecho este trabajo, podemos mencionar que se ha enseñado la importancia que tiene la minería de datos en el ámbito empresarial. Y tras ver los casos prácticos, con el programa Weka, ya somos capaces de utilizar algunas funciones del programa Weka por nuestra propia cuenta. Asimismo, destaca la accesibilidad, y la interfaz sencilla que contiene este programa.

La minería de datos se va a ir introduciendo en todas las empresas de cualquier sector ya sea a gran escala o a pequeña escala, debido a que la minería de datos facilita la actividad de cualquier sector, y hace a la empresa más eficaz. Además, el uso de la minería de datos permite ahorrar en costes a las empresas haciéndolas más eficientes.

Cabe destacar que las predicciones que se pueden realizar en minería de datos no son 100% exactas, pero si nos da una orientación de los pasos a tomar. Y por norma general, cuantas más variables se introduzcan a la hora de predecir, mayor será la precisión de los resultados.

Un factor para señalar en cuanto al futuro de la minería de datos es que las empresas cada vez almacenan más datos, por ello, el uso de la minería de datos se va a convertir en fundamental, para realizar una actividad, y marcará la diferencia entre las empresas que lo usen y las que no, evitando en muchas empresas el fracaso debido a su uso.

Por último, me gustaría destacar una frase de Cesar Alierta: "*Los datos son el petróleo del siglo XXI. El despliegue de sensores y el incremento de la capacidad del procesamiento son claves en la transformación de muchos sectores y en la creación de un mundo más medible y programable".*

5. LIMITACIONES Y LINEAS FUTURAS DE INVESTIGACIÓN

Las limitaciones que se hayan en este Trabajo Fin de Grado son varias. Por ejemplo, que no se explique todas las funciones que tiene el programa Weka, ya que se han explicado los principales métodos, pero no se ha explicado los métodos de ensemble (métodos combinados), o las SVMs ("*Support Vector Machines")*. Además, en cuanto a la regresión, se ha explicado la regresión lineal, pero no se menciona la regresión logística o la regresión por mínimos cuadrados.

En cuanto a las líneas futuras de investigación, cabe destacar que completando las explicaciones de los métodos que ofrece el programa Weka, ayudaría al lector a ampliar sus conocimientos. Asimismo, otras líneas futuras de investigación pueden dedicarse a explicar otros softwares de minería de datos de código libre y abierto, como son Orange, RapidMiner, JHepWork y KNIME.

6. GLOSARIO

- Algoritmo: Se conoce por algoritmo al conjunto de reglas, que al aplicarse de manera sistemáticas son capaces de analizar un conjunto de datos y en unos pasos finitos dar la solución al problema (Fanjul, 2018).

- Big Data: Se refiere al conjunto de datos de gran tamaño que la capacidad de un software no es capaz de capturar, almacenar, administrar y analizar (Malvicino & Yoguel, 2016).

- Campo multidisciplinar: Se conoce como campo multidisciplinar al campo que afecta a varias disciplinas distintas (Di Rae, s.f.).

- Código abierto: Se refiere a cualquier programa que utiliza un código de fuente que permite el uso o modificación por otros usuarios o desarrolladores que lo deseen. Normalmente se desarrolla una colaboración pública y disponible de manera gratuita (Rouse, 2016).

- Conocimiento: Se conoce como conocimiento a la posesión de multitud de datos interrelacionados, que por sí solos no tienen tanto valor (Pérez Porto, 2008).

- Dato: es la representación de una variable cuantitativa o cualitativa. Pueden ser generados automáticamente y almacenados por sistemas informáticos o tienen integrarse siempre para poder formar parte de la base de datos (Concepto, s.f.).

- Deep Learning: es un algoritmo automático jerárquico que imita el aprendizaje humano para obtener conocimiento (Smart Panel, s.f.).

- Eficacia: es la capacidad para conseguir alcanzar las metas establecidas previamente (DIferenciador, s.f.).

- Eficiencia: es la capacidad de alcanzar los objetivos establecidos consumiendo la menos cantidad de recursos (DIferenciador, s.f.).

- Hardware: Se define como Hardware a todos los componentes físicos de un ordenador, es decir, lo que se puede tocar (Roble, s.f.).

- Implantación de un sistema: Es un proceso de inserción del sistema en la institución (Docirs, s.f.).

- Información: conjunto de datos supervisados y coordinados, que nos permite construir un mensaje (Definicion, s.f.).

- Inteligencia Artificial: Se conoce con este nombre al conjunto de algoritmos combinados entre sí para crear máquinas capaces de poseer capacidades similares a las de los seres humanos (Iberdrola, s.f.).

- Interfaz: Conexión física y a nivel de utilidad entre dos o más dispositivos o sistemas (Definicion, s.f.).

- Modelos analíticos: Método de investigación, que descompone cada parte del todo para poder observar las causas y efectos (Enciclopedia virtual, s.f.).

- Necesidad: carencia o escasez de algo unido al sentimiento de qué es imprescindible (Significados, s.f.).

- Nodos: Se conoce como nodo la unión o intersección entre varios elementos que se encuentran en el mismo lugar. Un ejemplo es la red de internet, que cuenta con cada servidor como un nodo (Definición, s.f.).
- Segmentación de clientes: Proceso que consiste en dividir a los clientes en grupos diferenciados con características comunes que permita darles un servicio o producto que responda mejor a sus necesidades específicas (Shopify, s.f.).

- Sistema: módulo de elementos organizados que se encuentran relacionados y que interactúan entre ellos (Definición, s.f.).

- Software: Se define como Software al conjunto de instrucciones que necesita un ordenador para poder funcionar, no se puede ni ver ni tocar (Roble, s.f.).

- Tecnología: EL conjunto de conocimientos y técnicas aplicadas de manera ordenada para alcanzar un objetivo o resolver un problema (Economipedia, s.f.)

- Web Mining: Se conoce como Web Mining al proceso del uso de técnicas de minería de datos y algoritmos para extraer información directamente desde una Web o a través de documentos Web o servicios Web, etc.

7. REFERENCIAS

Academia Educación, 2017. *Academia Educación.* [En línea]
Available at:
https://s3.amazonaws.com/academia.edu.documents/30126388/vigilancia_tecnologica_aenor-iale_4-3-08.pdf?response-content-disposition=inline%3B%20filename%3DDe_la_vigilancia_tecnologica_a_la_inteli.pdf&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Credential=AKIAIW
[Último acceso: 6 julio 2019].

Aguilar, J., s.f. *Universidad de Los Andes.* [En línea]
Available at: http://www.ing.ula.ve/~aguilar/actividad-docente/IN/transparencias/clase40.pdf
[Último acceso: 7 Julio 2019].

Aprende con Alf, s.f. *Aprende con Alf.* [En línea]
Available at: http://aprendeconalf.es/estadistica/spss/capitulo_regresion.pdf
[Último acceso: 20 Julio 2019].

Cardenas, J., 2014. *Networkianos.* [En línea]
Available at: http://networkianos.com/regresion-lineal-multiple/
[Último acceso: 20 Julio 2019].

Concepto.de, 2019. *Concepto.de.* [En línea]
Available at: https://concepto.de/excel/
[Último acceso: 2 Agosto 2019].

Concepto, s.f. *Concepto.* [En línea]
Available at: https://concepto.de/dato/
[Último acceso: 18 agosto 2019].

Conceptos claros, s.f. *Conceptos claros.* [En línea]
Available at: https://conceptosclaros.com/que-es-clustering/
[Último acceso: 10 julio 2019].

Definicion, s.f. *Definicion.* [En línea]
Available at: https://definicion.de/informacion/
[Último acceso: 18 agosto 2019].

Definicion, s.f. *Definicion.* [En línea]
Available at: https://definicion.de/interfaz/
[Último acceso: 18 agosto 2019].

Definición, s.f. *Definicion.de.* [En línea]
Available at: https://definicion.de/sistema/
[Último acceso: 10 julio 2019].

Definición, s.f. *Definicion.de.* [En línea]
Available at: https://definicion.de/nodo/
[Último acceso: 10 julio 2019].

Di Rae, s.f. *Di Rae.* [En línea]
Available at: https://dirae.es/palabras/multidisciplinar
[Último acceso: 2019 agosto 18].

DIferenciador, s.f. *Diferenciador.* [En línea]
Available at: https://www.diferenciador.com/diferencia-entre-eficacia-y-eficiencia/
[Último acceso: 18 agosto 2019].

Docirs, s.f. *Docirs.* [En línea]
Available at: https://www.docirs.cl/implantacion_sistema.htm
[Último acceso: 18 agosto 2019].

Economipedia, s.f. *Economipedia.* [En línea]
Available at: https://economipedia.com/definiciones/tecnologia.html
[Último acceso: 18 agosto 2019].

Enciclopedia virtual, s.f. *Enciclopedia virtual.* [En línea]
Available at: http://www.eumed.net/libros-gratis/2007a/257/7.1.htm
[Último acceso: 18 agosto 2019].

Extracción y recuperación de información, s.f. *Extracción y recuperación de información.* [En línea]
Available at:
https://sites.google.com/site/extraccionyrecuperaciondeinfo/home/arbolesdedecision
[Último acceso: 8 julio 2019].

Fanjul, S., 2018. *El Pais.* [En línea]
Available at:
https://retina.elpais.com/retina/2018/03/22/tendencias/1521745909_941081.html
[Último acceso: 10 julio 2019].

Hand, 1998. [En línea]
Available at: http://www.ing.ula.ve/~aguilar/actividad-docente/IN/transparencias/clase40.pdf
[Último acceso: 6 julio 2019].

Hand, Mannila & Smyth, 2001. [En línea]
Available at: http://www.ing.ula.ve/~aguilar/actividad-docente/IN/transparencias/clase40.pdf
[Último acceso: 6 julio 2019].

Iberdrola, s.f. *Iberdrola.* [En línea]
Available at: https://www.iberdrola.com/innovacion/que-es-inteligencia-artificial
[Último acceso: 11 julio 2019].

jpgarcia.cl, 2008. *jpgarcia.cl.* [En línea]
Available at: https://jpgarcia.cl/2008/07/25/metodologia-para-proyectos-de-mineria-de-datos/
[Último acceso: 30 Agosto 2019].

Malvicino, F. & Yoguel, G., 2016. *BIG DATA. AVANCES RECIENTES,* s.l.: s.n.

Monjas, Y. B., s.f. *Course Hero.* [En línea]
Available at: https://www.coursehero.com/file/41205894/MINER%C3%8DA-DE-DATOSpdf/
[Último acceso: 7 Julio 2019].

Morate, D. G., s.f. [En línea]
Available at:
https://knowledgesociety.usal.es/sites/default/files/MANUAL%20WEKA.pdf
[Último acceso: 14 Julio 2019].

Neural Networks Framework, s.f. *Neural Networks Framework.* [En línea]
Available at: http://www.redes-neuronales.com.es/tutorial-redes-neuronales/Las-redes-neuronales-multicapa.htm
[Último acceso: 7 Julio 2019].

Neuronics, s.f. *Neuronics.* [En línea]
Available at: http://www.neuronics.es/servicios_implementacion.html
[Último acceso: 9 7 2019].

Ocaña, R., s.f. *Divestadística.* [En línea]
Available at: http://www.divestadistica.es/es/que_es_un_modelo_estadistico.html
[Último acceso: 9 Julio 2019].

Pérez Porto, J., 2008. *Definicion.* [En línea]
Available at: https://definicion.de/conocimiento/
[Último acceso: 18 agosto 2019].

Reporte Digital, 2018. *Reporte Digital.* [En línea]
Available at: https://reportedigital.com/cloud/data-mining-dentro-empresa/
[Último acceso: 8 julio 2019].

Roble, s.f. *Roble.* [En línea]
Available at:
http://roble.pntic.mec.es/jprp0006/tecnologia/1eso_recursos/unidad02_componentes_ordenador/teoria/teoria1.htm
[Último acceso: 18 agosto 2019].

Rouse, M., 2016. *Search Data Center.* [En línea]
Available at: https://searchdatacenter.techtarget.com/es/definicion/Fuente-abierta-o-codigo-abierto-open-source
[Último acceso: 18 agosto 2019].

Sainz, A. M., 2019. *Ucrea.* [En línea]
Available at: https://repositorio.unican.es/xmlui/handle/10902/16108
[Último acceso: 12 Julio 2019].

SAS, s.f. *SAS.* [En línea]
Available at: https://www.sas.com/es_es/insights/analytics/machine-learning.html
[Último acceso: 19 Julio 2019].

Shopify, s.f. *Shopify.* [En línea]
Available at: https://es.shopify.com/enciclopedia/segmentacion-de-clientes
[Último acceso: 18 agosto 2019].

Significados, s.f. *Significados.* [En línea]
Available at: https://www.significados.com/necesidad/
[Último acceso: 18 agosto 2019].

Smart Panel, s.f. *Smart Panel.* [En línea]
Available at: https://www.smartpanel.com/que-es-deep-learning/
[Último acceso: 18 agosto 2019].

Sngular, s.f. *Sngular.* [En línea]
Available at: http://blog.sngular.team/crisp-dm-fase-i-comprension-del-negocio-business-understanding
[Último acceso: 30 Agosto 2019].

Universidad Nacional Autónoma de México, s.f. *Universidad Nacional Autónoma de México.* [En línea]
Available at:
http://iibi.unam.mx/voutssasmt/documentos/dato%20informacion%20conocimiento.pdf
[Último acceso: 11 Julio 2019].

Ventura, S., 2013/14. *Pág. web Universidad de Córdoba.* [En línea]
Available at:
https://sci2s.ugr.es/sites/default/files/files/Teaching/GraduatesCourses/Bioinformatica/Tema%2006%20-%20AGs%20I.pdf
[Último acceso: 10 julio 2019].

ANEXO 1: BASE DE DATOS: PRECIO DE M^2 DE VIVIENDA NUEVA EN SAN SEBASTIÁN

Anyo	Precio
1994	1400
1995	1668
1996	1514
1997	1526
1998	1418
1999	1597
2000	2043
2001	2375
2002	2340
2003	2437
2004	2861
2005	3455
2006	3594
2007	4194
2008	4243
2009	4372
2010	4033
2011	3876
2012	3682
2013	3523
2014	3405
2015	3281
2016	3148
2017	3309
2018	3408
2019	3478

Tabla A1.1: Base de datos precio m^2.

Fuente: Eustat.

ANEXO 2: BASE DE DATOS: ACCIDENTE AÉREO

Sexo	Estado	Salud
Hombre	Soltero	Muerto
Mujer	Viudo	Vivo
Hombre	Casado	Muerto
Hombre	Soltero	Muerto
Mujer	Soltero	Vivo
Hombre	Casado	Muerto
Hombre	Casado	Vivo
Hombre	Casado	Muerto
Hombre	Casado	Muerto
Mujer	Viudo	Muerto
Mujer	Soltero	Vivo
Mujer	Viudo	Vivo
Mujer	Casado	Muerto
Mujer	Viudo	Vivo
Mujer	Soltero	Muerto
Hombre	Viudo	Vivo
Mujer	Casado	Muerto
Mujer	Soltero	Vivo
Mujer	Casado	Vivo
Mujer	Casado	Vivo
Mujer	Soltero	Muerto
Hombre	Viudo	Vivo
Hombre	Viudo	Vivo
Mujer	Soltero	Vivo
Mujer	Casado	Vivo
Hombre	Casado	Vivo
Mujer	Soltero	Muerto
Hombre	Viudo	Muerto
Hombre	Soltero	Muerto
Hombre	Soltero	Vivo
Hombre	Casado	Vivo
Mujer	Soltero	Muerto
Mujer	Viudo	Vivo
Hombre	Soltero	Muerto
Mujer	Casado	Vivo
Mujer	Viudo	Muerto
Hombre	Viudo	Vivo
Hombre	Casado	Vivo
Hombre	Soltero	Muerto
Mujer	Viudo	Vivo
Hombre	Casado	Muerto
Mujer	Soltero	Muerto
Mujer	Soltero	Vivo

Tabla A2.1.: Base de datos accidente aéreo.

Fuente: Accidente aéreo.