



FACULTAD DE
CIENCIAS ECONÓMICAS
Y EMPRESARIALES

**Business Intelligence para la toma de decisiones en
la empresa: Aplicación de métodos de minería de
datos en el sector comercial**

Business Intelligence to make decisions in the company:
Application of data mining methods in the commercial
sector

Trabajo de Fin de Grado
para acceder al

GRADO EN ADMINISTRACIÓN Y DIRECCIÓN
DE EMPRESAS

Autor: Jesús Sañudo Barquín
Tutora: Eliana Rocío Rocha Blanco
Santander, Septiembre de 2017

RESUMEN

El proyecto está desarrollado desde un punto de vista teórico-práctico, donde se trata el estudio de las principales características del concepto Business Intelligence (BI), un término que cada vez está más presente en las empresas y con el que se puede realizar diagnósticos que permiten obtener un conocimiento que ayuda a las empresas para la toma de decisiones y así obtener una ventaja competitiva, teniendo en cuenta la competencia entre las organizaciones hoy en día.

Se analiza los principales procesos, técnicas y herramientas de BI, como desarrollan, estructuran y analizan para obtener una ventaja competitiva en las organizaciones.

Por otro lado, se profundiza en una técnica del BI, concretamente en la técnica de minería de datos analizando su funcionamiento, características principales y puestas en práctica mediante el análisis de datos.

SUMMARY

The project is developed from a theoretical-practical point of view, which deals with the study of the main characteristics of the concept of Business Intelligence (BI), a term that is increase in companies and allow you to obtain a knowledge that helps the companies for the decision making and thus obtain a competitive advantage, taking into account the competition between the organizations today.

It analyses the main processes, techniques and tools of BI, as they develop, structure and analyse to obtain a competitive advantage in organizations.

On the other hand, it is deepened in a technique of the BI, specifically in the technique of data mining analysing its operation, main characteristics and put into practice by the analysis of data.

Tabla de contenido

1. Introducción	5
1.1 Motivación	5
1.2 Objetivos	5
2. Marco teórico: Business Intelligence.....	6
2.1 Definición de Business Intelligence.....	6
2.2 Ventajas del Business Intelligence.....	7
2.3 Componentes del Business Intelligence	8
2.4 Fases del Business Intelligence.....	9
2.5 Business Intelligence en el ámbito empresarial	10
2.6 KPI	11
2.7 Tendencia actual del Business Intelligence	12
3. Tecnologías y herramientas utilizadas	12
3.1 Herramientas de Business Intelligence	12
3.2 Minería de datos.....	13
3.3 Minería de datos vs análisis estadístico.....	14
3.4 Proceso de extracción de conocimiento.....	15
3.5 Técnicas para el proceso de minería de datos.....	15
3.6 Ejemplos de aplicación de minería de datos en las empresas	17
3.7 Herramientas de minería de datos.....	17
4. Aplicación práctica en el sector comercial.....	19
4.1 Herramienta de minería de datos: WEKA.....	19
4.2 Ventajas y desventajas de utilizar Weka.....	20
4.3 Objetivo del análisis.....	20
4.4 Fuente de datos.....	20
4.5 Pre-procesamiento	21
4.6 Clasificación	23
4.7 Clúster	24
4.8 Reglas de Asociación	25
4.9 Selección de atributos	26
4.10 Visualización.....	27
5. Análisis	28
6. Conclusiones	28
7. Bibliografía.....	29

Listado de figuras

Figura 1: Componentes del Business Intelligence.....	8
Figura 2: Fases del Business Intelligence.....	9
Figura 3: Transformación de datos en información	14
Figura 4: Mapa del proceso de minería de datos.	15
Figura 5: Ámbitos de trabajo en Weka	19
Figura 6: Atributos de la base de datos.....	21
Figura 7: Apartado “edit” en Weka	22
Figura 8: Pre-procesado de los datos	22
Figura 9: Resultados clasificación de Weka.....	23
Figura 10: Utilización del algoritmo J48	23
Figura 11: Cantidad de datos utilizada.....	24
Figura 12: Resultados del método de clasificación	24
Figura 13: Resultados análisis Clúster.....	25
Figura 14: Resultados método de asociación	25
Figura 15: Resultados selección de atributos.....	27
Figura 16: Método de visualización.....	27
Figura 17: Separación de cestas de consumo	28

Listado de tablas

Tabla 1: Minería de datos vs Análisis estadístico Fuente: (Larissa T. Moss, 2003).....	14
Tabla 2: Ventajas y desventajas de Weka. Fuente: (Fallas, 2011).....	20
Tabla 3: Fuente de datos	21
Tabla 4: Resultados algoritmo A Priori.....	26

1. Introducción

EL Business Intelligence, un concepto muy importante hoy día, donde la comunicación y las nuevas tecnologías es total, donde todo avanza muy rápido y que, sobre todo, que es de gran utilidad para el futuro tanto académico como laboral. El BI es fundamental hoy día para las empresas de modo de obtención de ventajas competitivas para el negocio, que transforma datos procedentes de diferentes fuentes (bases de datos de departamentos, CRM, ERP...), para convertirla mediante diferentes técnicas en información relevante y que abarca o presenta un número muy amplio de procesos de BI. Además, la minería de datos, una de las muchas técnicas que tiene el BI, resulta importante en todo el proyecto que se va a desarrollar junto con la herramienta "Weka".

1.1 Motivación

El Business Intelligence (BI) es el conjunto de metodologías, aplicaciones, prácticas y capacidades enfocadas a la creación y administración de información que permite tomar mejores decisiones a los usuarios de una organización (Caralt, Jordi Conesa ; Díaz, 2010).

Hoy en día el BI se ha convertido en una herramienta fundamental para el desarrollo y funcionamiento normal de la empresa. Ya en una organización pequeña es relevante, pero a medida que la empresa crece la cantidad de información que recopila (Proveedores, clientes, facturas, gastos, ventas...) es mayor por lo que si toda la información no está bien organizada, puede hacer perder volumen de negocio, o simplemente no poder actuar de manera rápida frente a algún problema. Es necesario para el objetivo final que tiene toda empresa, la de obtener beneficios, y más hoy día, donde la competencia en cualquier sector es máxima y donde se trata de maximizar los costes.

En cuanto a la motivación personal, el Business Intelligence, desde el punto de vista académico se puede decir que se ha abordado desde una perspectiva general, sin llegar a profundizar. Por ello con la realización de este proyecto, se quiere tener un conocimiento más profundo acerca del BI que sirva para un entorno laboral.

La realización y desarrollo del proyecto consiste en la creación de una base de datos que permita una visión única, conformada persistente y de calidad de toda la información de una cadena de supermercados que quiere entender el comportamiento y preferencias del consumidor para poder ofrecer ofertas especializadas y así responder a tiempo a cambios reales en el mercado. Además, entre otros, permitirá identificar los problemas que existan en la calidad de la información, soportar el crecimiento de la empresa, optimizar el stock o posibles correcciones presupuestarias.

1.2 Objetivos

EL objetivo general que se persigue es conocer ampliamente que es el BI y cuáles son las herramientas, procesos y métodos más importantes. Consiste en un desarrollo teórico, en analizar las distintas técnicas y procesos de BI, como se llevan a cabo, ver que empresas han llevado a cabo estos procesos, de qué manera y si han tenido éxito o no. A partir de saber en qué consiste el BI, los objetivos también se centran en una parte específica del BI para analizarla más ampliamente.

Por lo tanto, en cuanto a objetivos específicos, la finalidad en un principio era la de analizar mediante este proyecto una base de datos real a partir de la cual, desde un punto de vista empresarial se pueda trabajar para analizar, estudiar, ver y dar posibles soluciones a toda la información obtenida y poder dar soporte a una empresa mediante el análisis de los datos y así obtener una ventaja competitiva.

Ante la imposibilidad de obtener una información “rica en recursos”, el objetivo específico consiste en obtener un mayor conocimiento en una herramienta concreta del BI. Dicha herramienta concreta es la técnica de minería de datos. Esta rama de business Intelligence es muy importante ya que a partir de los volúmenes de datos que provienen de CRM, ERP, bases de datos... la empresa, con diferentes herramientas y mediante una serie de algoritmos reunidos en un único software, es capaz de procesar toda la información y extraer la parte importante de la misma para conseguir un mayor beneficio empresarial. Además de ser una herramienta con la que se puede perfilar y clasificar a los clientes, permite predecir posibles escenarios futuros, no solo en cuestiones de marketing o comunicación, también en los procesos de creación y mejora de los productos, entre otros...

Por último, se utilizará la herramienta Weka para la minería de datos, software muy completo, open source, que posee una serie de algoritmos que permite extraer información relevante de diferentes fuentes de datos y que puede resultar relevante para el análisis de los datos en cualquier organización o empresa.

2. Marco teórico: Business Intelligence

2.1 Definición de Business Intelligence.

La situación económica y la forma de comercio actual han llevado a las empresas y organizaciones a cambiar la forma de llevar sus negocios. En cualquier organización se generan datos como consecuencia de todas las transacciones que se llevan a cabo. Todos estos datos se almacenan mediante sistemas transaccionales en bases de datos relacionales. El objetivo de estos almacenes de datos es que se convierta en información fundamental para las decisiones de los ejecutivos, para así poder construir modelos que ayuden a predecir futuros eventos. Por otro lado, las compañías han tenido que realizar inversiones millonarias para hacer frente a la demanda de los clientes, así como mejorar su oferta y reducir costos para obtener una ventaja competitiva para permanecer a flote. Mediante las técnicas BI se puede contestar preguntas como: ¿Cuáles son los clientes que nos dan mayor ganancia y qué hacer para atraerlos más?, ¿Cómo es la proporción de nuestros gastos con respecto a nuestras ventas?, entre otras.

Además, las compañías deben centrarse en retener a sus clientes, como también la necesidad de adquirir otros nuevos. La retención de clientes es un factor muy importante para las mismas y, especialmente, entender la rentabilidad de un cliente.

Algunas definiciones de diferentes autores del Business Intelligence:

“EL BI podemos definirlo como el conjunto de estrategias enfocadas a la administración y creación del conocimiento sobre el medio a través del análisis de los datos existentes en una organización.” (Caralt, Jordi Conesa ; Díaz, 2010).

“EL BI es una de las pocas formas sustentables de ventaja competitiva que quedan. En el fondo dos competidores cualesquiera que estén bien establecidos en el mercado tendrán el mismo acceso a capital, tecnología, investigación de mercados, información de clientes y distribución. La gente y la calidad de las decisiones que ellos toman son los diferenciadores principales de la competencia en la era de la información.” (Vitt, Luckevich and Misner, 2002).

Por lo tanto, en una idea general podemos definir el Business Intelligence como un conjunto de herramientas o técnicas que emplean datos que las compañías tienen almacenados, los cuales se pueden transformar y adaptar de manera que facilite la toma de decisiones por parte del ejecutivo.

2.2 Ventajas del Business Intelligence

Desde el punto de vista de la actividad de la empresa, los beneficios que aportan los sistemas BI se pueden clasificar en:

Beneficios tangibles: generación de ingresos, reducción de costes, reducción de tiempo para las diferentes actividades. Los más importantes, entre muchos otros:

- Mejorar la adquisición de clientes
- Incrementar los ingresos por crecimiento de las ventas
- Conocer el entorno demográfico que rodea un negocio
- Permite analizar la competencia de una manera más detallada
- Reducir el tiempo de lanzamiento de nuevos productos
- Analizar la cesta de la compra y afinidad en la venta de productos
- Facilitar cambios en la estrategia competitiva de la empresas
- Identificación de clientes rentables en segmentos no rentables
- Control de costes y disminución de gastos (personal, materia prima, otros aprovisionamientos, acreedores, proveedores...)
- Detectar obsolescencia en maquinaria o equipos de trabajo

Beneficios intangibles: disponer de una información fiable y reconocida generara un uso adecuado de la misma:

- Optimizar la atención a los clientes
- Transformación de la información que genera un conocimiento que permite la toma de mejores decisiones
- Visión única conformada histórica persistente y de calidad de toda la información
- Permite aportar información tanto a nivel agregado como en detalle
- Mejora la comprensión y documentación de los sistemas de información en el contexto de una organización
- Mejora la competitividad de la organización, debido a, como ya se mencionó anteriormente, al acceso más rápido de la información, diferenciar lo relevante de lo superfluo o a una mayor agilidad en la toma de decisiones, entre otras.

Beneficios estratégicos: que permitirá conocer mejor a los clientes, mercados, productos..., mejorar la toma de decisiones, dar soporte a las estrategias y aumentar el valor del mercado:

- Identificar y nutrir a aquellos clientes con mayor potencial
- Mayor habilidad para analizar la estrategia de precios
- Mejora la toma de decisiones
- Mayor visibilidad en la gestión

- Dar soporte a las estrategias
- Aumentar el valor de mercado

A veces, resulta adecuada la implantación de un sistema BI, como puede ser la toma de decisiones de forma intuitiva en la organización, identificación de problemas, necesidad de cruzar información rápidamente entre los distintos departamentos de la organización, utilización de la información necesaria, ya que existe información que no es necesaria habitualmente.

2.3 Componentes del Business Intelligence

Los distintos componentes del BI quedan reflejados en la *Figura 1*:

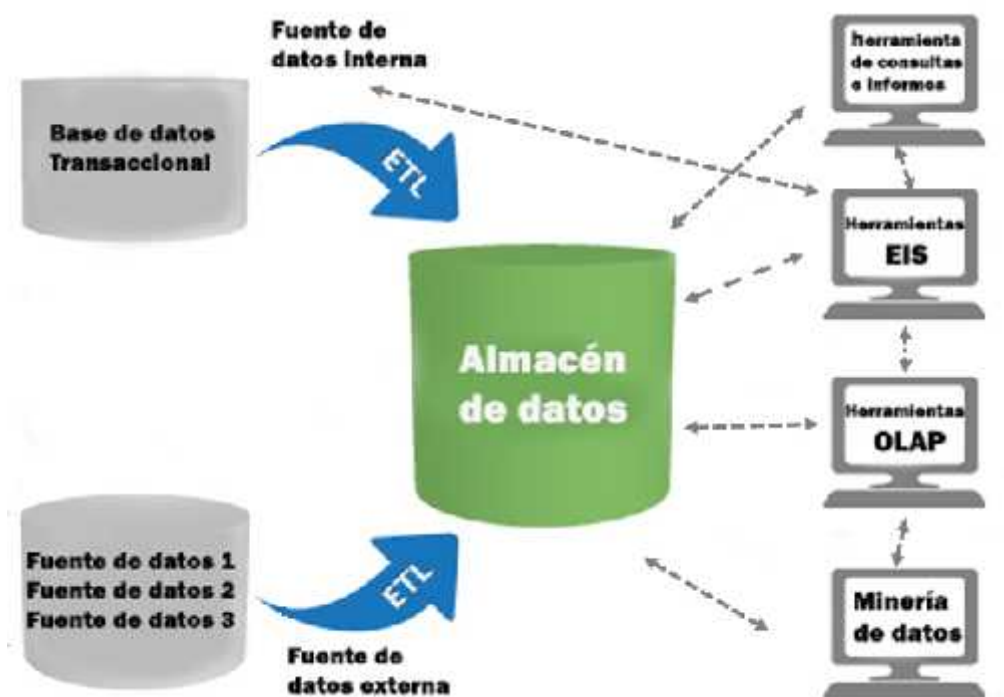


Figura 1: Componentes del Business Intelligence. Fuente: elaboración propia

- ✓ **Base de datos transaccional:** es una fuente de datos mediante la cual se mantiene el trabajo transaccional diario de los sistemas de información originales. Forman parte de las fuentes internas de una organización.
- ✓ **Fuentes externas de datos.**
- ✓ **Almacén de datos o Datawarehouse:** siempre está orientado hacia la información relevante de la organización. Se diseña para consultar eficientemente información relativa a las actividades (compras, ventas

producción...) básicas de la organización y no para soportar los procesos que se realizan en ella. EL almacén de datos está compuesto por los datos que provienen de los diferentes sistemas operacionales, incluyendo las bases de datos transaccionales y las fuentes externas de datos.

- ✓ Mediante las **herramientas ETL** (extraer, transformar y cargar) se extraen los datos de distintas fuentes para luego cargarlos en un almacén de datos. La información en el almacén de datos se recopila mediante los llamados datamarts, que son esquemas que contienen información al detalle que puede ser utilizada por los distintos departamentos o secciones de la organización.
- ✓ **EL proceso de visualización:** una vez que tenemos la información en el Datawarehouse, es necesario la utilización de aplicaciones para que el usuario final pueda ver la información que necesita. Las herramientas más comunes son las herramientas de informes, EIS (Executive Information System), OLAP y minería de datos.

El software que se crea mediante BI perdurará en el tiempo si conseguimos que una organización incremente su nivel financiero. En caso contrario, será sustituido por otro que aporte mejores y más precisos resultados.

2.4 Fases del Business Intelligence

Aunque cada proyecto de business Intelligence es diferente, se puede encontrar una serie de características comunes en cada uno de ellos (Ver *Figura 2*):



Figura 2: Fases del Business Intelligence. Fuente: (Dataprix, 2009).

- El primer paso para la realización de una aplicación o proyecto de BI es la **definición de los objetivos**, es decir, cuál es la finalidad de la realización de este proyecto.
- **Definición de los requerimientos**, donde se analiza y se trata de comprender las ventajas que puede ofrecer este tipo de sistemas. En este punto se detallan los pasos a seguir en un proyecto de este tipo.
- **Diseño y modelización**: Los requerimientos crearan la base para la creación del Datawarehouse. Se analizan las fuentes de datos y las transformaciones que serán necesarias para hacer útil la información de la que disponemos. En definitiva la creación del componente lógico que servirá para crear la estructura del Datawarehouse que permitirá adaptarnos al tipo de extracción de información que realizaremos en el mismo.
- **Implementación**: que conllevará la extracción de datos del sistema operacional los procesos de la carga de datos en el Datawarehouse y la explotación del mismo. El objetivo de esta fase es que el Datawarehouse esté operativo para el usuario final.
- **Revisión**: tratar de mejorar el programa de tratando de buscar errores o implementar nuevas características.

2.5 Business Intelligence en el ámbito empresarial

Como se menciona anteriormente, las empresas acumulan datos procedentes de diferentes departamentos y datos sobre clientes, producción inventarios, proveedores... además de datos del exterior como puede ser de los competidores, por lo que el BI puede aportar información importante a los distintos departamentos de una empresa:

- Departamento de compras: principalmente se analiza la relación entre coste-beneficio de cada producto o materia prima adquirida, además de analizar, si procede, la cadena de producción, por lo que puede ayudar a optimizar el volumen de compras.
- Departamento de producción: seguir y analizar el proceso de producción desde el control de calidad, el historial de producción e inventarios.
- Departamento de marketing: a partir de información de clientes estudiar su comportamiento, como puede ser el impacto que tienen en los mismos las promociones de productos o el cambio de precios.
- Departamento de ventas: El BI también permite adelantarse a las necesidades del cliente y responder a nuevas oportunidades del mercado y el estudio de patrones de compra, especialmente en la actualidad con un planeta globalizado.

- Departamento financiero: como es lógico, tener la capacidad de acceder fácilmente y mejorando operaciones de presupuestos, control de gestión, control de tesorería, cuentas anuales...
- Departamento de recursos humanos: puede ayudar a una contratación más eficiente, analizando la situación y satisfacción de los empleados, además de diferentes criterios como el absentismo laboral, rotación, beneficio hora/hombre...
- Departamento atención al cliente: puede obtener la satisfacción de los clientes y con ello ofrecer un mejor servicio y retener a los clientes más rentables.

Muchas veces las empresas presentan el desconocimiento en el mercado en el que se encuentran. Lo que trata de hacer el BI es eliminar ese desconocimiento aprovechando los datos que posee. Si la inteligencia de negocio es la estrategia dominante en la empresa, seguramente esta empresa obtenga unos rendimientos superiores. Por supuesto, hay que tener en cuenta que para que esto sea posible es necesario que exista el capital humano, si este es óptimo la velocidad de las funciones aumentará. Por lo tanto, el BI optimizará el rendimiento de los sistemas y de los diferentes departamentos, obtendrá información oculta de los datos (como ocurre en el proceso de minería de datos), mejora de contratos tanto con clientes como con proveedores, depurar o eliminar errores del pasado (a partir de los datos se puede obtener errores ocurridos en el pasado en cualquier departamento o función...), a partir de un cuadro de mando (*dashboards*) seguir los planes estratégicos... y poder establecer una parte importante para el BI: los KPI o indicadores de rendimiento.

2.6 KPI

Los KPI (*Key Performance Indicator*) o indicadores de rendimiento permiten identificar el rendimiento de una determinada estrategia, es decir, indican el nivel de desempeño respecto a los objetivos fijados anteriormente. Estos KPI se agrupan en *dashboards* de manera que los directivos tomen decisiones de una manera más rápida. Los KPI cambian de una empresa a otra en función de sus objetivos o necesidades y cada modelo de negocio tiene factores clave a medir diferentes. Por ejemplo una empresa que únicamente vende online tendrá distintos indicadores que un local tradicional. Ejemplos de KPI pueden ser financieros, de ventas, de clientes, de proveedores, gestión comercial.... Para establecer KPI's para una empresa determinada es necesario, lo primero de todo, contestar a una serie de preguntas: ¿qué se necesita? ¿Qué se quiere medir? ¿Por qué?, ¿Para qué? ¿Es realmente necesario?

Por ejemplo en cualquier comercio sería necesario establecer, entre otros:

- **KPI's financieros:** que reflejan los ingresos y gastos de la empresa, de todo el negocio o líneas de productos para poder evaluar la rentabilidad de cada producto, además de poder comparar como está la empresa respecto al plan inicial.
- **KPI's ventas:** saber las ventas diarias semanales mensuales... por familia de productos, comparar con periodos pasados, que clientes son los que más productos adquirieron.

- **KPI's de entrega a clientes**, como es el caso del e-commerce, saber si llego el producto a tiempo, envíos por región, productos que llegan más tarde...
- **KPI's de eficiencia de trabajo**, para saber si se cuenta con el número eficiente de trabajadores, causas de eficiencia eficacia de mano de obra...
- **KPI's de inventarios** para saber el costo de los mismos, la cantidad de stock precisa...
- **KPI's de proveedores** que refleje el importe de compras, si el proveedor entrega el producto a tiempo porcentaje de devoluciones a cada proveedor...

2.7 Tendencia actual del Business Intelligence

Para entender la tendencia actual del BI hay que tener en cuenta la evolución de sus herramientas, que las mismas soportan cada vez más funcionalidades, son más fáciles de usar, mejoras en la visualización de la información y sobre todo el aumento de las herramientas *open source*. Actualmente las empresas que trabaja con aplicaciones o programas BI se plantean proyectos a nivel corporativo más que departamental.

Cada vez más usuarios usan herramientas BI debido a la aparición de normas internacionales de tratamiento de la información, además de la mejora de calidad de la misma. Además, todas las herramientas que una empresa pudiera tener, hoy día tienen la posibilidad de agruparlas en un único software, permitiendo así la reducción del costo de mantenimiento, licencias o formación. La tendencia de este tipo de herramientas era con la vista puesta en el futuro a largo plazo, pero hoy día son usadas a corto plazo, para actividades de tipo operacional, del día a día.

Esta tendencia actual también se puede apreciar en el mercado:

- Los fabricantes están comercializando sus propios productos, como puede ser el Business Warehouse de SAP.
- Especialización de software en situaciones concretas.
- Los creadores de motores de bases de datos tienen sus propias soluciones, como el SQL de Microsoft.
- Reducción del coste de las licencias.
- Aparición de nuevas soluciones open source.

3. Tecnologías y herramientas utilizadas

3.1 Herramientas de Business Intelligence

El business Intelligence cuenta con una serie de estrategias o procedimientos principales entre los que destacan:

- **Análisis OLAP**, permite al usuario organizar datos y manipularles por medio de consultas o informes. Un ejemplo de herramienta OLAP sería que un usuario pueda analizar los datos para mostrar un archivo Excel que muestre todos los

Smartphone que se venden en Cantabria en el mes de Diciembre, para comprar comparar los gastos e ingresos que se producen respecto a Julio. A diferencia de otras herramientas, los datos se organizan en bases de datos multidimensionales, para poder considerar cada atributo de los datos (producto, lugar de ventas, gastos, periodo de tiempo...). Se suele utilizar OLAP debido a la rapidez y agilización de respuesta a las consultas. Jasper o Pentaho son dos software que permiten la utilización de herramientas OLAP.

- EIS (Executive Information System) son herramientas que permiten analizar los datos en la empresa, pensados principalmente para aportar información en tiempo real a la alta dirección para poder tomar decisiones de distinta índole principalmente cuando se producen tendencias indeseadas. Muestran gráficos de las diferentes áreas funcionales que componen la empresa. La dificultad de estos equipos se produce cuando hay que adaptar los datos que provienen de fuentes externas, al estar en diferentes formatos, por lo que el sistema ha de ser capaz de organizar y recopilar los datos.
- Análisis de flujo de clics, que analiza el comportamiento de visitantes de una web.
- Minería de datos, que extrae la información de datos que permanece oculta y que, como se verá a continuación, una herramienta BI de gran utilidad.
- Forecasting que permite entre otras cosas acceder a los datos de productos para estimar predicciones.
- Dashboards, que como se indica anteriormente, se utilizan para representar los principales KPI's de la empresa.
- Análisis de negocio.
- Herramientas de Visualización (Query y Reporting).

Como se comentó anteriormente, uno de los objetivos de este trabajo es el análisis de herramientas de Business Intelligence. En este caso, dicha herramienta o técnica que se va a analizar es la conocida minería de datos, un proceso de relevancia para una empresa que quiera implantar BI en su día a día.

3.2 Minería de datos

La minería de datos o Datamining es un proceso que consiste en la extracción de información de aporte significativo para una empresa, a partir de una o varias bases de datos y que pueden permitir al usuario realizar distintas predicciones o llevar a cabo diferentes pautas que puedan resolver problemas de negocio, lo cual puede generar una ventaja competitiva. Permite a partir de grandes volúmenes de información, separar o “quedarse” con la información más valiosa, o la información que necesita la empresa en un momento determinado. Las herramientas de minería de datos pueden permitir u obtener perspectivas futuras acerca de la situación futura de la empresa, descubriendo patrones o perfiles a través del análisis de los datos. Por lo tanto, brinda una estructura robusta en el almacenamiento de datos.

Otra definición del proceso de minería de datos: SAS Institute define el concepto de minería de datos como “El proceso de Seleccionar, Explorar, Modificar, Modelizar y Valorar grandes cantidades de datos con el objetivo de descubrir patrones desconocidos que puedan ser utilizados como ventaja comparativa respecto a los competidores” (Clinic-cloud, 2016).

En la *Figura 3* se puede resumir el concepto de la minería de datos, donde se tiene en un principio todos los datos en diferentes estancias o departamentos, que después “se trituran” o los hacemos uno solo, para posteriormente obtener información privilegiada:



Figura 3: Transformación de datos en información. Fuente: (Socialdoor, 2015).

3.3 Minería de datos vs análisis estadístico

La minería de datos no consiste simplemente en un análisis estadístico de datos, sino que profundiza en los mismos y obtiene información oculta. En la *Tabla 1* se puede ver una serie de diferencias entre ambas:

Análisis estadístico	Minería de datos
<ul style="list-style-type: none">• Normalmente comienzan con una hipótesis	<ul style="list-style-type: none">• No requiere de hipótesis
<ul style="list-style-type: none">• Tienen que desarrollar sus propias ecuaciones para adaptarse a la hipótesis	<ul style="list-style-type: none">• Los algoritmos de la minería de datos automáticamente desarrollan las ecuaciones
<ul style="list-style-type: none">• Únicamente datos numéricos	<ul style="list-style-type: none">• Datos numéricos y otros (texto, voz...)
<ul style="list-style-type: none">• Pueden encontrar datos incorrectos en el análisis y filtrarlos	<ul style="list-style-type: none">• Depende de los datos limpios y una buena documentación
<ul style="list-style-type: none">• Interpretan sus propios resultados	<ul style="list-style-type: none">• Los resultados no son fáciles de interpretar

Tabla 1: Minería de datos vs Análisis estadístico Fuente: (Larissa T. Moss, 2003)

3.4 Proceso de extracción de conocimiento

De una manera ejemplificada o esquematizada, se resume el proceso de minería de datos de la siguiente manera (Ver *Figura 4*):

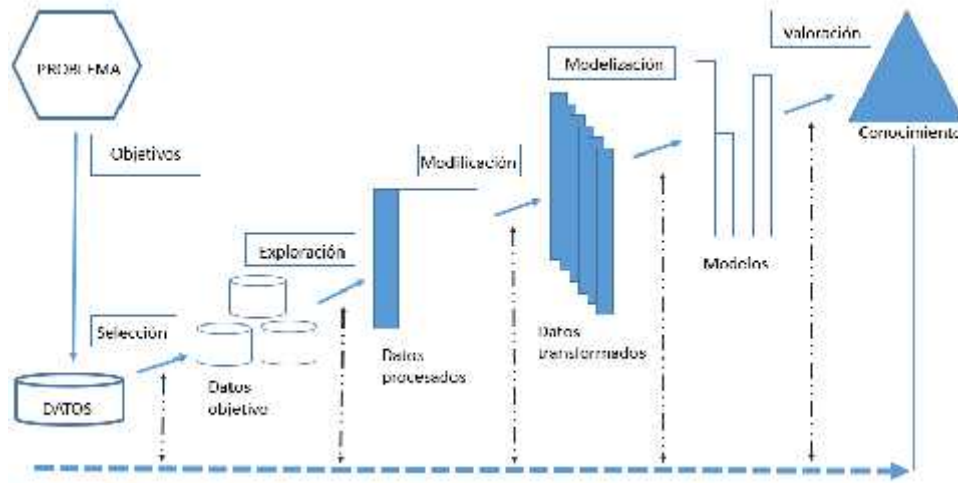


Figura 4: Mapa del proceso de minería de datos. Fuente: elaboración propia basado en: (Jojoa, 2012).

- **Selección:** cuales van a ser las variables objetivo y la selección de registros a utilizar.
- **Exploración** búsqueda de valores diferentes y usencia de otros.
- **Modificación:** transformación del conjunto de datos de entrada, es decir, normalizar los datos a una misma escala. Este proceso consume gran parte del tiempo total del proceso de minería de datos.
- **Modelización:** consiste en la realización, selección y aplicación de las técnicas de minería de datos que se van a emplear. Búsqueda de elementos semejantes y diferentes de los datos. Tras determinar los datos más importantes se comienzan a realizar análisis estadísticos y una visualización gráfica de los mismos para acercarse al objetivo final.
- **Valoración:** validar o comprobar que las conclusiones que hemos obtenido son válidas. Si hemos usado diferentes técnicas, se comparan los modelos en busca de aquel que nos solucione mejor el problema.

3.5 Técnicas para el proceso de minería de datos

Principalmente las técnicas para el análisis de minería de datos se pueden clasificar en:

- **Métodos descriptivos:** que buscan patrones interpretables para describir datos.

- **Métodos predictivos:** mediante el que se podrían predecir estimaciones o valores futuros o desconocidos de otras variables.

Algunas técnicas para conseguir los objetivos de estas tareas son:

- **Clasificación:** seguramente la tarea más familiar o popular de la minería de datos. Sirve para predecir a la clase a la que pertenecen nuevos objetos a partir de las variables restantes, permite ver el comportamiento de los atributos en determinados grupos. Encuentra modelos que describen y distinguen clases o conceptos para futuras predicciones. Esta tarea puede asignar clasificaciones a nuevos datos examinando los datos ya existentes. Un ejemplo de este tipo de técnica es conocer las características de los clientes que tienen cierta probabilidad de comprar un determinado producto y sabiendo esto, se pueden reducir costes de promociones, envíos...
- **Agrupamiento (clustering):** que consiste en obtener grupos naturales a partir de los datos. Es muy parecida a la de clasificación, con la diferencia de que con clúster se parten los datos en diferentes subdivisiones para así arrojar análisis de manera independiente. El objetivo de utilizar la técnica de clustering es el de agrupar elementos tratando de lograr la máxima homogeneidad en cada grupo y las mayores diferencias entre los grupos. Sin embargo, se podría decir que es una técnica exploratoria ya que en muchas ocasiones no utiliza ningún modelo estadístico para llevar a cabo el proceso de clasificación y adecuada para obtener información de datos sin restricciones, por lo que puede ser útil para elaborar hipótesis a diferentes problemas considerados.
- **Predicción (forecasting):** por un lado el método cuenta con el análisis de regresión con el que se puede predecir el valor que toma una variable en nuevos objetos a partir de la información proporcionada por el resto de variables. Por ejemplo una empresa de automóviles puede predecir los accesorios que se pueden vender de un coche mirando la cantidad de automóviles que se vendieron el mes pasado.
Por otro lado, el método de predicción cuenta con el *"Time sequence Discovery"* en la que se prevén valores de datos en función el tiempo. Por ejemplo, saber la tasa de accidentes en semana Santa teniendo en cuenta los accidentes que hubo en este mismo periodo de tiempo el año pasado.
- **Método de asociación:** para ver si existen relaciones que no están determinadas entre variables categóricas. Puede que sea la parte más importante del proceso de minería de datos ya que nos puede ofrecer información oculta, es decir, la que a veces no vemos a simple vista. Un ejemplo puede ser el descubrimiento de que los hombres que compran café de marca Premium tienen más probabilidades de comprar cigarrillos que aquellos que compran una marca blanca.
- **Asociaciones secuenciales:** igual que la regla de asociación pero buscan relaciones temporales entre los datos. Por ejemplo, se puede descubrir que la persona que compra un secador de pelo compra a las dos semanas un nuevo cepillo. Además para aumentar esta probabilidad el comercio puede ofrecer ofertas complementarias entre los diferentes ítem.

- **Correlaciones:** para ver el grado de similitud entre diferentes variables.

3.6 Ejemplos de aplicación de minería de datos en las empresas

La minería de datos puede permitir contactar con aquellos clientes que tienen una mayor probabilidad de aceptar de manera positiva a una determinada oferta. Además para una misma empresa, deberá crear distintos modelos predictivos dependiendo por ejemplo de la zona geográfica o que clientes van a ser rentables durante un cierto tiempo:

- Un típico ejemplo de minería de datos es el de los hábitos de compra en supermercados. El mejor ejemplo es el supermercado que realizó un estudio y detectó que los viernes había una cantidad inusual de ventas de pañales y cerveza de en una misma cesta de la compra. Detectaron que se debía a que acudían al supermercado padres jóvenes que pasaban el fin de semana cuidando a su hijo a la vez que tomaban una cerveza viendo la televisión. Lo que hizo el supermercado fue colocar las cervezas próximas a la sección de pañales con el objetivo de incrementar las ventas compulsivas.
- En industrias o sectores como la banca o telecomunicaciones surge la necesidad de las empresas en detectar la posibilidad de los clientes de rescindir sus contratos por lo que, previsiblemente, se pasarían a la competencia. Por lo tanto a estos clientes se les hacen ofertas personalizadas con el objetivo de mantener dichos contratos. Lo que permite la minería de datos en este caso es ver que clientes podría ser más factible su posibilidad de rescisión teniendo en cuenta su comportamiento y comparándolo con antiguas bajas de contratos.
- Puede ayudar a la contratación de personal a partir del análisis de las características de los empleados que tienen éxito en la empresa.
- Incluso en la práctica deportiva se utilizan este tipo de técnicas, como es el caso del equipo de fútbol AC Milán, que posee un sistema de redes neuronales para prevenir lesiones, donde el sistema tiene datos de cada jugador, con aproximadamente 5000 datos registrados y puede permitir al equipo ahorrar dinero en caso de fichar un jugador con una probabilidad elevada de sufrir lesiones. También en la NBA utilizan herramientas para monitorizar pases, encestes, rebotes... para después aplicando técnicas Minería de datos ver aquellas acciones que no se pueden observar en un momento real.
- Hoy en día resulta muy importante aplicar minería de datos en Internet, especialmente para averiguar los clientes potenciales que visitan las páginas de Internet, ofrecer una publicidad determinada o una vez vendido el producto ofrecer otros productos adicionales.

En definitiva, la minería de datos se aplica en cualquier campo o sector y resulta relevante para obtener una ventaja competitiva.

3.7 Herramientas de minería de datos

Existen una multitud de herramientas para el proceso de minería de datos, tanto de pago como *open source*. Algunas de las más significativas son:



Rapidminer es un programa open source para la minería de datos que permite el análisis de datos que pueden ser combinados mediante un entorno sencillo y entendible.

Un programa de minería de datos que se utiliza tanto para investigación como para el mundo real. Permite abrir multitud de formatos de archivos (CSV, XML, JSON...).

URL: <https://rapidminer.com/>



Pentaho es un software de inteligencia empresarial más completas que existen en los últimos años.

Está basado en lenguaje Java y supone una gran solución para los diferentes problemas o necesidades empresariales, desde las más comunes a otras más específicas. Integra ETL, Reporting, OLAP, Minería de datos y Dashboards.

URL: <http://www.pentaho.com/>



Escrito en lenguaje JAVA, Soporta diferentes tareas de minería de datos para realizar análisis de datos y modelado predictivo, además posee herramientas para la visualización de estos datos. Permite entre otros, hacer clasificaciones, clustering, asociación, visualización...

Como se trata de un proceso automático, únicamente son necesarios los datos almacenados en un determinado formato.

URL: <http://www.cs.waikato.ac.nz/ml/weka/>



Herramienta escrita en Java, fundamentalmente permite el proceso de extracción, transformación y carga de datos (ETL), la posibilidad de generar informes e integra componentes para la minería de datos. Permite analizar datos en origen. URL: <https://www.knime.com/>

4. Aplicación práctica en el sector comercial

En este apartado se emplean las principales técnicas de minería de datos. Concretamente se emplean una serie de datos de una empresa del sector comercial, un supermercado. Unos datos que pueden ser perfectamente extraídos de un comercio real, como se verá en los próximos apartados.

4.1 Herramienta de minería de datos: WEKA

Para poner en sentido o demostrar todo lo explicado anteriormente, hay que ponerlo en práctica. Para ello se utilizará uno de las herramientas mencionadas anteriormente, concretamente el Weka, una herramienta de minería de datos muy fácil de usar gracias a su sencilla interfaz gráfica. Un software creado en 1993 en la universidad de Waikato, que es capaz de procesar los datos gracias a que corre una serie de algoritmos que lo permiten

Weka cuenta con una serie de ámbitos de trabajo:



Figura 5: Ámbitos de trabajo en Weka

- **Explorer**, la función más importante de este software, ya que permite acceder a los componentes gráficos de trabajo de manera sencilla. Dividida en varios paneles, como veremos más adelante (pre-procesado, clasificación, clúster, asociación, selección de atributos y visualización)
- **Experimenter**: permite obtener diferentes índices estadísticos a partir del trabajo con información a gran escala.
- **KnowledgeFlow** permite crear proyectos de minería de datos a partir de flujos de información
- **Simple CLI**: se trata del trabajo en Weka mediante la línea de comandos de Java

- **Workbench**, que a partir de esta opción podemos utilizar todos los ámbitos de trabajo mencionados anteriormente

4.2 Ventajas y desventajas de utilizar Weka

Aunque Weka es una herramienta de gran utilidad para la técnica de minería de datos, posee una serie de desventajas:

VENTAJAS	DESVENTAJAS
<ul style="list-style-type: none">• Software libre y del sencilla interfaz gráfica• Numerosas técnicas de modelado y procesamiento de datos• Escrito en Java, lo que permite la compatibilidad con numerosas plataformas	<ul style="list-style-type: none">• Los resultados no dan la total confiabilidad de los mismos• Carece de algoritmos para el modelado de secuencias

Tabla 2: Ventajas y desventajas de Weka. Fuente: (Fallas, 2011)

4.3 Objetivo del análisis

Un primer objetivo de este estudio es el análisis de dicha base de datos de manera que se pueda obtener algún tipo de ventaja competitiva para la empresa, actuando en el establecimiento a partir de dicho estudio. Y como es lógico, otro objetivo a considerar es el aprendizaje o manejo de un software de minería de datos para poder aplicar las distintas técnicas de la misma.

Hay que seguir un proceso de extracción de conocimiento para la correcta aplicación de la minería de datos.

4.4 Fuente de datos

Para poder realizar esta práctica, es necesario lo primero de todo tener una base de datos. Este proyecto cuenta con una base de datos de un supermercado, que aunque los datos no sean reales, sí que puede ser una aproximación a la realidad empresarial. Esta base de datos ha sido obtenida de un repositorio de Gary M. Weiss, profesor de la universidad de Fordham especializado en ingeniería de software y aplicaciones Minería de datos para el tratamiento de datos en el ámbito empresarial. Las características de los datos se pueden resumir en la *Tabla 3*:

Sector	Comercial
Atributos	217
Registros	4627
Lugar	Universidad de Fordham, Nueva York
URL	http://storm.cis.fordham.edu/~gweiss/data-mining/weka-data/supermarket.arff

Tabla 3: Fuente de datos

Lo primero de todo es detectar el problema, es decir, que le lleva a la empresa a realizar este tipo de prácticas.

Un segundo paso sería el proceso de selección, y como se trata de un análisis completo de los datos, el objetivo será obtener todo tipo de información para conseguir una ventaja competitiva, además de un proceso de exploración, donde se debe apreciar que variables tenemos o cuales faltan.

El tercer paso y el más importante de todos, el proceso de modelización donde se llevan a cabo realmente las técnicas de Minería de datos y donde se transforman los datos en información para poder sacar conclusiones acerca de la información recopilada. Este paso es el más importante y el más extenso y por lo tanto, en el que se va a centrar este estudio, como se verá en los siguientes apartados.

4.5 Pre-procesamiento

El “pre-procesamiento”, en el que además de poder editar la base de datos y analizarla, también podemos remover atributos que no queremos considerar y aplicar otra serie de filtros. Permite eliminar datos que están incompletos o datos con ruido o inconsistentes. Seguramente sea este el proceso más importante de la minería de datos ya que para obtener conclusiones validas a partir de los datos que tenemos es importante realizar una buena preparación de los datos. Es el proceso que más tiempo consume cuando se aplica la técnica de minería de datos, ya que hay que recopilar los datos provenientes de diferentes fuentes

Lo primero que se tiene es un conjunto de datos una base de datos con 217 atributos y 4627 registros o instancias (*Figura 6*):

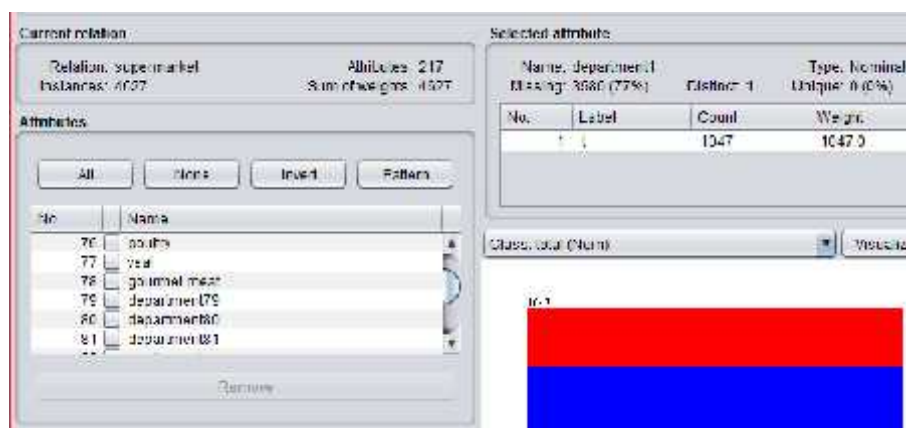


Figura 6: Atributos de la base de datos

Como se puede ver en la *Figura 6*, una vez que se cargan los datos en Weka, ya se puede ver como el programa nos indica información relevante. Se puede apreciar las variables fundamentales dentro de cada tributo de la base de datos.

Hay que ir al apartado “edit” para ver cómo están estructurados los datos, ante de empezar a trabajar con los mismos:



Figura 7: Apartado “edit” en Weka

Analizando este apartado, aparece el número de clientes que ha comprado en el supermercado, además de los productos que ha comprado.

El producto que ha comprado aparece como *true* “t” y si no ha comprado el producto aparece en blanco. Por lo tanto, ha habido un total de 4625 clientes y se pueden apreciar un total de 217 atributos. Por un lado estos atributos se dividen en departamentos y por otro en productos.

En este caso los atributos “Departamentos” no aportan ningún tipo de información útil tras realizar una comprobación inicial aplicando diferentes algoritmos, por lo que se procede a eliminar estos atributos, y se quedan los atributos que corresponden a los productos, por lo que el pre-procesado es necesario ser aplicado en estos datos.

Por lo tanto, se va a proceder a trabajar o aplicar algoritmos a los datos con un total de 106 atributos y 4627 registros.



Figura 8: Pre-procesado de los datos

4.6 Clasificación

Permite ver el comportamiento de los atributos en determinados grupos, es decir, cuáles son las variables más importantes:

Classifier output

Time taken to build model: 0 seconds

Classified cross-validation

==== Summary =====

Correctly Classified Instances	2848	85.71% %
Incorrectly Classified Instances	474	14.29% %
Kappa statistic	0	
Mean absolute error	0.4174	
Root mean squared error	0.4558	
Relative absolute error	0	%
Root relative squared error	100	%
Total Number of Instances	4322	

==== Detailed Accuracy By Class: =====

	To Class	FP Rate	Precision	Recall	F-Measure	ROC	ROC Area	PRC Area	Class
	1,000	1,000	0,000	0,000	0,000	0,000	0,000	0,000	True
	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	False
Weighted Avg.	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	

==== Confusion Matrix: =====

	Actual \ Predicted	True	False
Predicted \ Actual	True	2848	474
False	False	0	0

Figura 9: Resultados clasificación de Weka

En este caso, el método de clasificación que se emplea el algoritmo J48, que generalmente ofrece datos precisos y mediante los cuales se puede obtener un árbol de decisión y permite conocer las relaciones existentes entre los datos o hacer una clasificación general de los mismos. El filtro está realizado sobre el total de atributos de la base de datos.

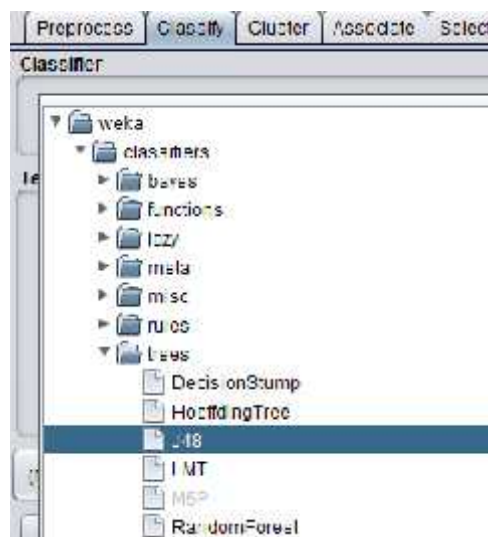


Figura 10: Utilización del algoritmo J48

Se Utiliza la opción “Use training set”, donde el aprendizaje que hace Weka se realiza a partir de todos los registros realizados en el mismo. También se puede utilizar el “Supplied training set”, el cual selecciona un archivo .arff que se encargará del

aprendizaje de la minería de datos para ser implementado en el .arff cargado en Weka, La opción “Cross validation” que será a partir de una cantidad de campos de los archivos cargados o el “Porcentaje” que consiste en elegir el porcentaje de registros serán utilizados para el aprendizaje.

Lo que se va a determinar en este caso es saber qué variables son las más determinantes.

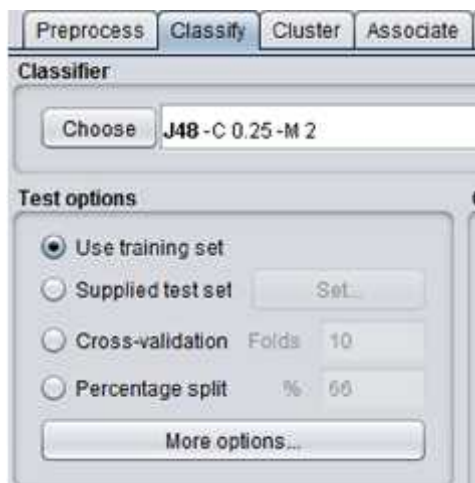


Figura 11: Cantidad de datos utilizada

Tras la aplicación del filtro, se llega a la conclusión que mediante el método de clasificación no podemos ver cuáles son las variables más importantes, ya que únicamente la aplicación nos indica los atributos y variables existentes. Además Weka no es capaz de crear un árbol de decisión, debido principalmente a que cada atributo (productos) son independientes entre sí o que Weka no es capaz de realizar una asociación entre los mismos. Clasifica correctamente un 63% de las instancias:

```

=== Summary ===
Correctly Classified Instances      2948      63.713 %
Incorrectly Classified Instances    1679      36.287 %
Kappa statistic                    0
Mean absolute error                 0.4624
Root mean squared error             0.4808
Relative absolute error             99.9965 %
Root relative squared error         100 %
Total Number of Instances          4627

=== Detailed Accuracy By Class ===

```

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area
1,000	1,000	0,637	1,000	0,778	0,900	0,500	
0,000	0,000	0,000	0,000	0,000	0,000	0,500	
Weighted Avg.	0,637	0,637	0,496	0,637	0,496	0,000	0,500

Figura 12: Resultados del método de clasificación

4.7 Clúster

Para comprobar su funcionamiento, en este caso selecciono un algoritmo para manejar el clúster. El más eficiente y preciso es el “SimpleKMeans”. Se vuelve a usar todos los datos de la base (Use training set), teniendo siempre en cuenta que los atributos “departamentos” se eliminaron en el pre-procesamiento. Se puede apreciar que Weka crea clúster con una cesta de consumo “High” (>20 productos) y otro clúster con la cesta de consumo “low” (<20 productos):


```
Number of iterations: *  
Within cluster sum of squares reduced: 0.0  
  
Initial starting points [random]:  
Cluster 1: 1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,  
Cluster 2: 1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,  
  
Missing values globally replaced with missing.  
  
Final cluster centroids:  


|          | Eucl Dist | Centroid  | n        |
|----------|-----------|-----------|----------|
| AgeGroup | (6827.0)  | (-6970.0) | [2048.0] |


|                    | L | I | T |
|--------------------|---|---|---|
| growing child      | 1 | 1 | 1 |
| baby needs         | r | r | r |
| cotton one word    | 1 | 1 | 1 |
| singing words      | 1 | 1 | 1 |
| wigwag             | r | r | r |
| jacks out from car | 1 | 1 | 1 |
| tax                | r | r | r |
| mashed             | 1 | 1 | 1 |
| popped fresh-wrap  | r | r | r |


```

Figura 13: Resultados análisis Clúster

Sin embargo, ya se conocía de antemano la relación entre estos dos atributos. Lo que se trataría en este caso de conseguir con la técnica de clustering es el de poder realizar grupos con los diferentes alimentos de la base de datos, y como Weka no ha sido capaz de establecer grupos entre los mismos, se considera que este proceso no es relevante para aplicar en este caso

Por lo tanto, la única conclusión que se obtiene es la que ya se sabía, que hay 1679 personas que compran más de 20 productos y 2948 que compran menos de 20 productos.

4.8 Reglas de Asociación

Generalmente la tarea más relevante de la minería de datos. Mediante las reglas de asociación se puede saber que patrones o transacciones ocurrieron anteriormente en un mismo conjunto de datos.

Associator output

```

Size of set of large sequences [4]: 40
Size of set of large sequences [5]: 400
Size of set of large sequences [6]: 1000
Size of set of large sequences [7]: 2000
Size of set of large sequences [8]: 1000
Size of set of large sequences [9]: 1000
Size of set of large sequences [10]: 500
Size of set of large sequences [11]: 100
Size of set of large sequences [12]: 100

```

Stack index output

```

1. bread and cake > 1000
2. bread and cake > 1000
3. bread and cake > 1000
4. bread and cake > 1000
5. bread and cake > 1000
6. bread and cake > 1000
7. bread and cake > 1000
8. bread and cake > 1000
9. bread and cake > 1000
10. bread and cake > 1000
11. bread and cake > 1000
12. bread and cake > 1000

```

Associator Output	Stack Index Output
Size of set of large sequences [4]: 40	1000
Size of set of large sequences [5]: 400	1000
Size of set of large sequences [6]: 1000	1000
Size of set of large sequences [7]: 2000	1000
Size of set of large sequences [8]: 1000	1000
Size of set of large sequences [9]: 1000	1000
Size of set of large sequences [10]: 500	1000
Size of set of large sequences [11]: 100	1000
Size of set of large sequences [12]: 100	1000

Figura 14: Resultados método de asociación

Por ejemplo, en el caso de un supermercado, si es habitual comprar pan y queso de manera conjunta, ambos ítem deben ser frecuentes.

Se utiliza el algoritmo “A Priori”, que permite minar conjuntos de elementos frecuentes. Además en este apartado se elimina el atributo “total”, para trabajar únicamente con los productos de manera de conseguir un resultado más limpio. El resultado que arroja Weka tras aplicar el algoritmo es el siguiente (Ver *Tabla 4*):

1. biscuits=t frozen foods=t pet foods=t milk-cream=t vegetables=t 516 ==> bread and cake=t 475 <conf:(0.92)> lift:(1.28) lev:(0.02) [103] conv:(3.44)
2. baking needs=t biscuits=t milk-cream=t margarine=t fruit=t vegetables=t 505 ==> bread and cake=t 464 <conf:(0.92)> lift:(1.28) lev:(0.02) [100] conv:(3.37)
3. biscuits=t frozen foods=t milk-cream=t margarine=t vegetables=t 585 ==> bread and cake=t 537 <conf:(0.92)> lift:(1.28) lev:(0.03) [115] conv:(3.35)
4. biscuits=t canned vegetables=t frozen foods=t fruit=t vegetables=t 536 ==> bread and cake=t 492 <conf:(0.92)> lift:(1.28) lev:(0.02) [106] conv:(3.34)
5. baking needs=t frozen foods=t milk-cream=t margarine=t fruit=t vegetables=t 517 ==> bread and cake=t 474 <conf:(0.92)> lift:(1.27) lev:(0.02) [101] conv:(3.29)
6. biscuits=t frozen foods=t pet foods=t milk-cream=t fruit=t 511 ==> bread and cake=t 468 <conf:(0.92)> lift:(1.27) lev:(0.02) [100] conv:(3.26)
7. biscuits=t frozen foods=t tissues-paper prd=t milk-cream=t vegetables=t 575 ==> bread and cake=t 526 <conf:(0.91)> lift:(1.27) lev:(0.02) [112] conv:(3.22)
8. biscuits=t frozen foods=t beef=t fruit=t vegetables=t 536 ==> bread and cake=t 490 <conf:(0.91)> lift:(1.27) lev:(0.02) [104] conv:(3.2)
9. baking needs=t biscuits=t frozen foods=t cheese=t fruit=t 538 ==> bread and cake=t 491 <conf:(0.91)> lift:(1.27) lev:(0.02) [103] conv:(3.14)
10. biscuits=t frozen foods=t milk-cream=t margarine=t fruit=t 592 ==> bread and cake=t 540 <conf:(0.91)> lift:(1.27) lev:(0.02) [113] conv:(3.13)

Tabla 4: Resultados algoritmo A Priori.

Weka, mediante este algoritmo ha utilizado los elementos más frecuentes para obtener reglas de asociación. Se elige el filtro a priori y a su vez mediante ese filtro que las reglas se ordenen por confianza (“confidence”). Además, se aplica una confiabilidad mínima del 90%, para así obtener una mayor calidad de información

Se pueden apreciar los elementos que más se repiten en las cestas de consumo. Por ejemplo es común que el que compra galletas también compra comida congelada con una confiabilidad del 92%, lo mismo que en la mayoría de cestas de consumo, donde aparecen pan y pasteles constantemente.

Por lo tanto, las reglas de asociación no buscan la causalidad, si no los patrones de co-ocurrencia, es decir, donde se repiten los datos que se procesan y de qué manera.

4.9 Selección de atributos

En este caso no se puede demostrar cuales son los atributos más importantes que determinan la cesta de consumo, ya que cada producto es diferente y el consumo o compra de cada individuo es distinto.

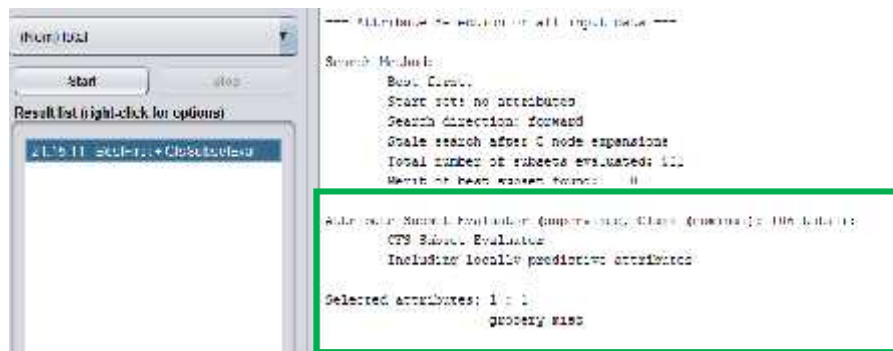


Figura 15: Resultados selección de atributos

4.10 Visualización

Muestra gráficamente la distribución de todos los atributos mostrando graficas en dos dimensiones en las que va representado en los ejes todos los posibles pares de combinación de los atributos. Las posibles opciones que permite este modo mediante las barras inferiores.

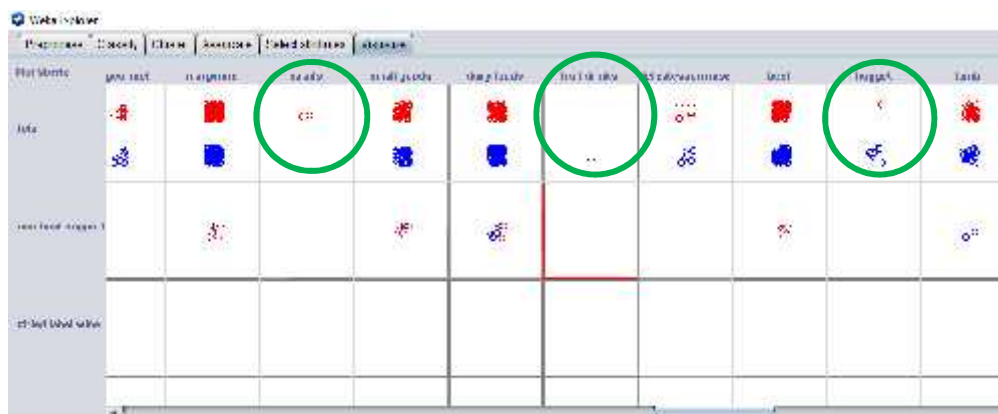


Figura 16: Método de visualización

En este caso se aprecia si los atributos forman parte de una cesta de consumo “high” o “low”. Como se aprecia en la *Figura 16*, el atributo “fruit drinks” únicamente aparece en la cesta de consumo “low”, el atributo “hogget” se vende mucho más en una cesta de consumo pequeña, o que la variable “salads” únicamente se vende en cestas de consumo “high”, por ejemplo. Por tanto indica los productos que forman tanto las cestas grandes como pequeñas de consumo.

Además, en la opción visualización, podemos separar las cestas según el consumo “high” o “low” y a partir de aquí trabajar con ellas creando un nuevo archivo .arff de manera que se puedan estudiar las cestas de consumo de manera independiente, con lo que se puede averiguar con más facilidad, por ejemplo, cuales son los productos que más se venden en cada cesta “low” o “high”.

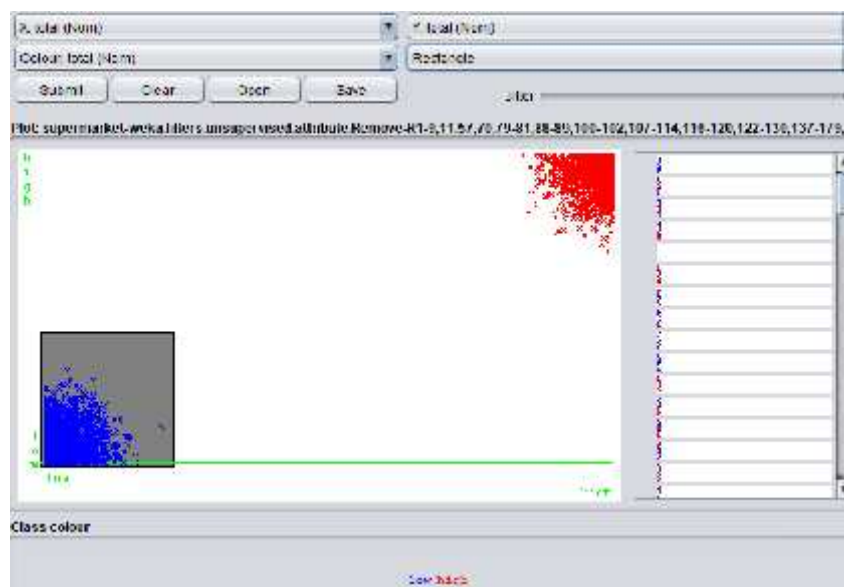


Figura 17: Separación de cestas de consumo

5. Análisis

Se puede obtener una serie de conclusiones sobre la práctica realizada anteriormente:

En este caso, aplicar una serie de metodologías de minería de datos, los únicos procesos que nos han resultado útil para obtener una ventaja competitiva de este supermercado es el método de asociación. Se ha podido ver, por ejemplo que las personas que compran pan también compran pasteles con una confiabilidad del 92%, o que la comida congelada se encuentra prácticamente en todas las cestas de consumo. A partir de esta simple información, el supermercado puede ya actuar para mejorar su negocio. En este caso podría colocar la estantería de las galletas cerca de los productos congelados, ya que prácticamente la persona que compra galletas también va a comprar productos congelados. También podría realizar nuevas ofertas y asociaciones entre productos. Además se puede apreciar que aproximadamente el 64% de los consumidores cuando acuden a la tienda compran menos de 20 productos. A partir de aquí también podría actuar el supermercado ya que sabiendo que hay más consumidores que compran un menor número de productos se podrían colocar menos cajas para cobrar, ahorrando tiempo tanto para el consumidor como para el trabajador, poner cestas o carritos más pequeños, disminuir el coste humano...

6. Conclusiones

Este es un claro ejemplo de minería de datos. A partir de una cantidad determinada de datos, mediante una aplicación de minería de datos, en este caso Weka, se ha podido ver esta circunstancia que quizá no se podría ver si no se realiza esta operación. Las empresas generalmente tienen unas cantidades exageradas de datos, por lo que este tipo de análisis son fundamentales para mejorar el negocio. Y no solo una empresa grande necesita este tipo de operaciones, también como en este caso empresas pequeñas les puede resultar útil utilizarlo. Si hubiera sido una cadena de supermercados

muy grande, habría que aplicar este tipo de análisis seguramente dependiendo de la región en la que se encuentre. Por poner un ejemplo, no es lo mismo las ventas en un lugar caluroso donde se compran más aires acondicionados que zonas con un clima más frío.

Sin embargo, y como se ha podido ver anteriormente, la minería de datos no siempre va a funcionar de una manera “total”, ya que a veces por mucha cantidad de datos que poseas no es posible establecer patrones o relaciones entre los atributos que lo componen, como en este caso, en el que los métodos de clasificación, selección de atributos o análisis clúster no han servido.

Una recomendación que podría hacer a este supermercado suponiendo que únicamente de su actividad diaria solo ha podido obtener estos datos, sería establecer un KPI para el control del supermercado: mínimo los KPI de finanzas, ventas y marketing, manufactura y de inventarios y compras.

Por tanto, si nos centramos en responder a las preguntas de BI ¿Qué paso?, ¿Qué pasa ahora?, ¿Por qué pasó?, ¿Qué pasará?, mediante la minería de datos podemos responder, en mayor medida a esta serie de preguntas.

En cuanto a la definición de Business Intelligence, es un concepto que cada vez más las empresas ha de tener en cuenta, que operan en un entorno donde la competencia es cada vez mayor y globalizada, donde los precios cada vez son más ajustados, por lo que resultará necesario obtener una ventaja competitiva, ya sea en ahorro de costes de fabricación, logística, personal..., como el aumento de ventas que se podría producir, aunque hay que tener en cuenta que este concepto implica muchas técnicas y herramientas, por lo que la empresa que necesite aplicar este tipo de procesos tiene que elegir el método adecuado que cumpla con los objetivos que busca la compañía.

7. Bibliografía

Definición de Business Intelligence:

- [1] Caralt, Jordi Conesa ; Díaz, J. C. (2010). *Introducción al Business Inteligence*. Disponible en: https://blogvirtualtec.files.wordpress.com/2015/02/introduccion_al_business_intelligence.pdf
- [2] Vitt, E., Luckevich, M., & Misner, S. (2002). *BUSINESS INTELLIGENCE Técnicas de análisis para la toma de decisiones estratégicas*.
- [3] Perez Marques, M. (2015). *Business intelligence : técnicas, heramientas y aplicaciones*. RC Libros. Disponible en: <http://catalogo.unican.es/cgi-bin/abnetopac/O7131/ID787a6bcc/NT1>

Ventajas del Business Intelligence

- [4] Pechuán, I. G., & Collado, J. (n.d.). *Business Intelligence*. Disponible en: http://mugi.webs.upv.es/wp-content/uploads/2016/06/1_webinar_BI.pdf

Componentes del Business Intelligence

- [5] Perez Marques, M. (2015). *Business intelligence : técnicas, heramientas y aplicaciones*. RC Libros. Disponible en: <http://catalogo.unican.es/cgi-bin/abnetopac/O7131/ID787a6bcc/NT1>

Fases del Business Intelligence

- [6] Dataprix (2009) *Fase de implantación de un Datawarehouse*. Disponible en: <http://bi-businessintelligence.blogspot.com.es/2009/01/fases-de-implantacin-de-un-data.html> (Accessed: 9 May 2017).

Business Intelligence en el ámbito empresarial:

- [7] Espinosa, R. (2016). KPI's. Disponible en: <http://robertoespinosa.es/2016/09/08/indicadores-de-gestion-que-es-kpi/>

KPI's

- [8] Espinosa, R. (2016) 'KPI's'. Disponible en: <http://robertoespinosa.es/2016/09/08/indicadores-de-gestion-que-es-kpi/>.

Tendencia actual en el Business Intelligence

- [9] Lluís Cano, J. (2007). *Business Intelligence Competir con información*. Disponible en: http://www.dataprix.com/files/BI-Business-Intelligence-Competir-Con-Informacion_Dataprix.pdf

Tecnologías y herramientas utilizadas

- [10] By Larissa T. Moss, S. A. (2003) *Business Intelligence Roadmap: The Complete Project Lifecycle for Decision*.

- [11] Caralt, Jordi Conesa ; Díaz, J. C. (2010) *Introducción al Business Intelligence*. Disponible en: https://blogvirtualtec.files.wordpress.com/2015/02/introduccion_al_business_intelligencie.pdf

- [12] Sistemas, B. G. de (2014) *Herramientas EIS*. Disponible en: <http://www.gestiondesistemas.es/sistemas-de-informacion-ejecutiva-eis/>.

Minería de datos:

- [13] Vieira Braga, L. P. (2009). *Introducción a la minería de Datos*.

- [14] Pérez López, C. (2007). *Minería de datos: técnicas y herramientas*.

- [15] Socialdoor (2015) 'Minería de datos'. Available at: <https://socialdoor.es/data-mining-y-turismo/>.

Proceso de extracción de conocimiento:

- [16] Jojooa. (2012). Proceso Minería de datos. Disponible en: <https://sites.google.com/site/jojooa/inteligencia-artificial/proceso-de-data-mining-mineria-de-datos>

Análisis estadístico vs minería de datos

- [17] By Larissa T. Moss, S. A. (2003) *Business Intelligence Roadmap: The Complete Project Lifecycle for Decision*.

Técnicas para el proceso de Minería de datos

- [18] ¿Qué es el Minería de datos? (2016). Disponible en: <https://clinic-cloud.com/blog/data-mining-que-es-definicion-mineria-de-datos/>
- [19] Augusto Osorio, G. (2016). Técnicas de Minería de Datos. Disponible en: http://www.bdigital.unal.edu.co/2037/2/germanaugustoosoriozuluaga_Parte2.pdf

Herramientas Minería de datos

- [20] Redacción. (2016). Herramientas Minería de datos. Disponible en: <http://culturacrm.com/data-mining/cinco-herramientas-data-mining/>
- [21] El rincón de Jmacoe. (2017). Herramientas Minería de datos. Disponible en: http://blog.jmacoe.com/gestion_ti/base_de_datos/5-mejores-software-mineria-datos-codigo-libre-abierto/

Ámbitos de trabajo en Weka

- [22] Fallas, L. C. (2011). Ámbitos de trabajo de Weka. Disponible en: <http://cor-mineriadatos.blogspot.com.es/2011/06/weka.html>

Ventajas y desventajas de Weka

- [23] Fallas, L. C. (2011) *Ventajas y desventajas de Weka*. Disponible en: <http://cor-mineriadatos.blogspot.com.es/2011/06/weka.html>.

Weka

- [24] Morate, D. G. (n.d.). Manual Weka. Disponible en: <http://sci2s.ugr.es/sites/default/files/files/Teaching/GraduatesCourses/InteligenciaDeNegocio/weka.pdf>
- [25] Figueras, M. S. (2011). Análisis “Clúster.” Disponible en: <https://es.scribd.com/doc/31635337/Ejemplos-Reales-de-Mineria-de-Datos>
- [26] Costas, B. J. S. (no date) *Reglas de asociación y secuencias*.

Base de datos para Weka

- [27] Gary M. Weiss (2004) ‘Base de datos para Weka’. Disponible en: <http://storm.cis.fordham.edu/~gweiss/>.