GRADO EN ADMINISTRACIÓN Y DIRECCIÓN DE EMPRESAS

CURSO 2018/2019

TRABAJO FIN DE GRADO

BUSINESS INTELLIGENCE:
MINERÍA DE DATOS APLICADA EN RECURSOS HUMANOS

AUTOR:

ANDRÉS RENILLA MERCHÁN

TUTORA:

ROCÍO ROCHA BLANCO

Septiembre 2019

DEGREE IN BUSINESS ADMINISTRATION AND MANAGEMENT

ACADEMIC COURSE 2018/2019

DEGREE'S FINAL PROJECT

BUSINESS INTELLIGENCE:
DATA MINING APPLIED IN HUMAN RESOURCES

AUTHOR:

ANDRÉS RENILLA MERCHÁN

DIRECTOR:

ROCÍO ROCHA BLANCO

September 2019

RESUMEN

El *Big Data* nació como consecuencia del descontrol de los datos almacenados en los sistemas. Las empresas, desbordadas por la situación, decidieron crear una estructura dentro de sus organizaciones que permitiera tanto el control de los datos como también su aprovechamiento, obteniendo información muy valiosa hasta el momento desconocida. Esta estructura se denomina *Business Intelligence* y, actualmente, es un pilar fundamental de las empresas para poder sobrevivir en un ambiente cada vez más competitivo. La implementación del *Business Intelligence* no solamente dependerá de los recursos con los que cuente la compañía ni de la involucración de sus altos ejecutivos, sino también de la capacidad de sus trabajadores para adaptarse a un ámbito nuevo.

Dentro del *Business Intelligence*, entre otros muchos factores, está el *Data Mining*, cuya función es encontrar patrones o modelos en los datos que se traduzcan en información para adoptar las decisiones necesarias.

En este trabajo se abarcarán todos los factores que intervienen en el *Business Intelligence*. Asimismo, se abordará un ejemplo práctico aplicando *Data Mining* sobre datos reales acerca de las ausencias de los trabajadores de una pequeña empresa de paquetería que, transformados en conocimiento, permitirán al departamento de recursos humanos tomar decisiones que mejoren el funcionamiento de la empresa. De este modo se demostrará que el *Business Intelligence* no solamente está al alcance de grandes corporaciones sino de aquellos que tienen la voluntad por desarrollar la empresa.

ABSTRACT

Big Data was born as the result of uncontrolled saved data in technologies. Companies decided to create structures inside their organizations to control data and, more important, to take advantage of data itself, getting useful information until the moment unknown. This structure is called *Business Intelligence*. Nowadays is an important factor between companies to survive in a more competitive environment. The introduction of *Business Intelligence* will not only depend on the resources a company has, or the high executives' involvement, it will also depend on the capabilities of his employees.

One important factor in *Business Intelligence* process is *Data Mining*. It finds patterns and models in datasets, which is transformed to information that is appropriated to choose the correct decisions.

This project includes every component *Business Intelligence* has. In addition, it will show a practical example of *Data Mining* working with real data about absenteeism in a small delivering company. It will extract knowledge very useful for human resources department who is the responsible to take the best decision to improve the company. This way, this project will prove *Business Intelligence* is not only for big companies. *Business Intelligence* is for those who want to grow the company.

ÍNDICE

1. INTRODUCCIÓN 1
1.1. EVOLUCIÓN DE LOS DATOS 1
1.2. NECESIDADES DE LAS EMPRESAS: FORMACIÓN DE EMPLEADOS EN ANÁLISIS DE DATO 4
2. OBJETIVOS DEL TRABAJO 5
2.1. OBJETIVO GENERAL 5
2.2. OBJETIVOS ESPECÍFICOS 5
3. MARCO TEÓRICO 6
3.1. BUSINESS INTELLIGENCE (BI) Y BUSINESS ANALYTICS (BA) 6
3.2. APLICACIONES DEL BUSINESS INTELLIGENCE EN EMPRESAS 6
3.2.1. Wal-Mart 6
3.2.2. Amazon 7
3.3. PROCESO BUSINESS INTELLIGENCE 7
3.4. TÉCNICAS BUSINESS INTELLIGENCE 9
3.4.1. Visualización de datos (Data Visualization) 9
3.4.2. OLAP 11
3.4.3. Data Mining: definición, componentes y fases 12
4. SITUACIÓN ACTUAL 15
4.1. HERRAMIENTAS BUSINESS INTELLIGENCE 15
4.2. ETAPAS BUSINES INTELLIGENCE: 1.0 – 2.0 – 3.0 16
5. METODOLOGÍA 17
6. DESARROLLO EMPÍRICO 17
6.1. PREPROCESAMIENTO 17
6.2. ANÁLISIS EXPLORATORIO DESCRIPTIVO 21
6.3. APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS 29
6.3.1. Técnicas de predicción mediante clasificación 29
6.3.1.1. Aplicación del modelo Naive Bayes 29
6.3.1.2. Modelo de clasificación: Árbol de decisiones J48 32
6.3.2. Clustering 34
6.4. CONCLUSIONES 35
7. VALORACIÓN FINAL 36
8. BIBLIOGRAFÍA 37

1. INTRODUCCIÓN

A lo largo de la historia se ha visto cómo los avances tecnológicos de cualquier ámbito han sido cada vez más rápidos. Un ejemplo claro es el modo en que las personas se han desplazado de un lugar a otro: desde la creación de la rueda hasta la llegada del hombre a la luna.

Aunque la rueda puede ser considerada como uno de los grandes inventos revolucionarios de la historia por el impacto que tuvo, en la Figura 1.1 se extrae que a medida que se avanza en el tiempo los cambios en la tecnología son cada vez más rápidos y más sorprendentes, como se observa con la creación de naves aeroespaciales que pueden volar fuera del planeta.

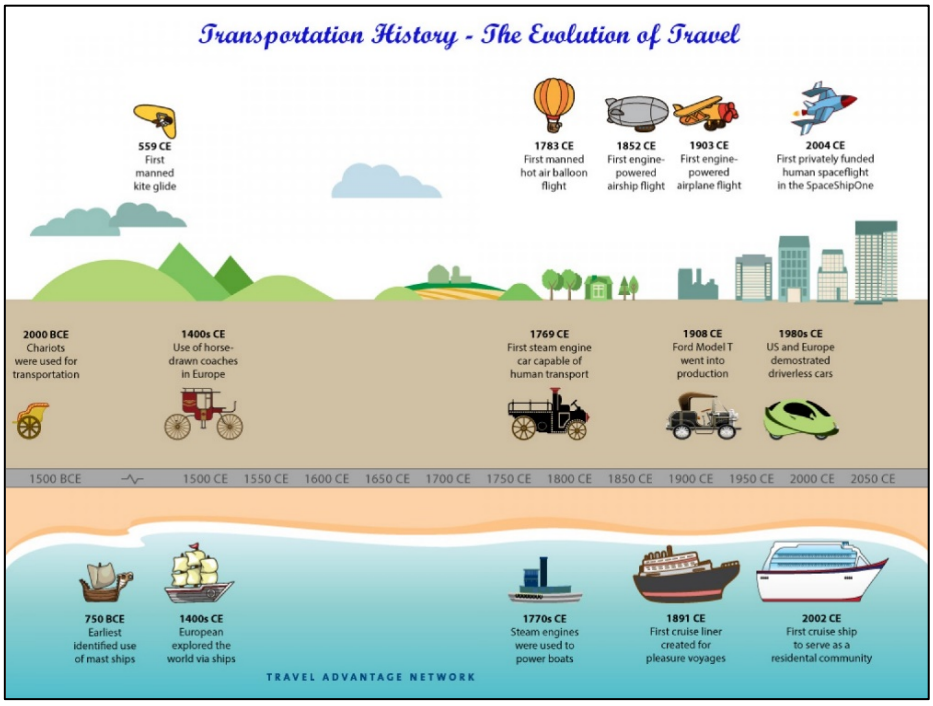


Figura 1.1 - Evolución del transporte. Fuente: (Visually, 2013)

El ejemplo anterior se puede extrapolar a otros muchos campos llegando todos a la misma conclusión: estamos en un mundo cada vez más cambiante y en el que avanzamos cada vez más deprisa a través de la mejora continua.

El ámbito en el que se centra este trabajo será el campo de los datos. Los datos sirven para crear información y, consecuentemente, conocimiento. A primera vista da la sensación de que los datos y la información no son tan trascendentales como una nave aeroespacial que es capaz de llegar a la luna. Sin embargo, “la información es poder”.

No tan lejos, en la Segunda Guerra Mundial, gracias a Alan Turing y su trabajo en el cifrado de los mensajes codificados que se transmitían los soldados alemanes, los británicos obtuvieron información de los movimientos germanos para poder acabar con sus barcos durante los primeros años de la década de los 40 (Central Intelligence Agency, 2015).

1.1. EVOLUCIÓN DE LOS DATOS

El área de los datos también ha sufrido una evolución, aunque ésta no comenzó hasta hace relativamente poco, en los años 80. No obstante, como ocurre en el mundo del

transporte, se verá que la transformación que sufre el campo de los datos es paralelamente trascendente y en constante cambio.

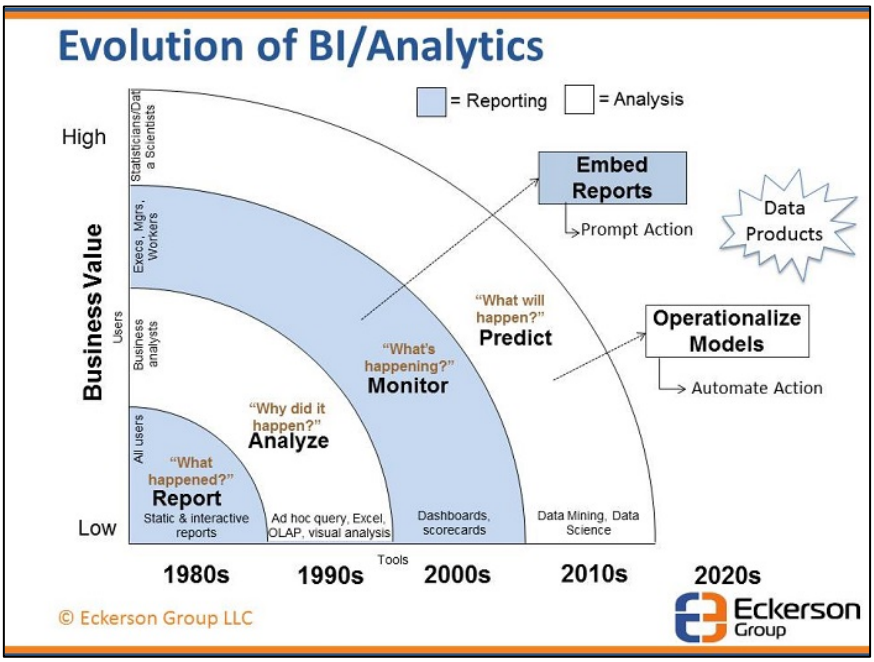


Figura 1.2 - Evolución del BI/BA. Fuente: (Eckerson Group, 2016)

La Figura 1.2, ilustra perfectamente la evolución del *Business Intelligence (BI)* a lo largo del tiempo en cuatro etapas, distinguidas por el valor que tiene el uso que se le da a los datos.

1) Década de los 80: los datos se utilizaban para realizar informes de los resultados obtenidos. Un informe lo puede leer y entender cualquiera que sepa leer y entenderlo, valga la redundancia.
2) Años 90: los datos se utilizaban para realizar análisis y poder dar explicaciones de los acontecimientos.
3) Comienzo del nuevo milenio: los datos se usaban para crear tablas y gráficos de manera que la información se mostrara de una manera visual y, por tanto, más rápida de entender.
4) Actualidad: los datos se usan para predecir lo que va a ocurrir.

Además, se observa en la Figura 1.2 que las etapas 1 y 3, coloreadas de azul, el principal objetivo es presentar los datos, mientras que las etapas 2 y 4, en blanco, son etapas en que los datos se analizan y exprimen.

Es importante destacar que hay diferentes maneras de situarse en la Figura 1.2: una en el tiempo (eje de abscisas) para hablar del contexto general, de la situación por la que está pasando el BI desde un punto de vista histórico, evolutivo; y otra como usuario (eje de ordenadas). Como usuario, la mayoría se encuentran "estancados en los años 90" debido a que solamente usan los datos para trabajarlos y presentarlos en Excel sin hacerse una idea de todas las funciones y fórmulas que contiene realmente el programa, sin haber mencionado qué es el BI.

Lo mismo que ha ido evolucionando el BI, ha ido aumentando el número de datos que se almacenan en los sistemas. En una fotografía realizada con el móvil se registra desde el momento y lugar en que se hizo la foto hasta el número de píxeles de esta. Estos son los conocidos metadatos, es decir, datos acerca de los datos.

Solamente en Facebook se suben casi 300 millones de fotos al día. (Betfy, s.f.) Entonces, ¿cómo se puede gestionar tanta cantidad de datos? A través del *Big Data* y el *Business Intelligence*. Empresas como Facebook y no necesariamente multinacionales tienen que ser capaces de controlar los datos, tanto externos como internos, y, lo más importante, darles utilidad. En el excelente artículo "*10 Charts That Will Change Your Perspective of Big Data's Growth*" (Columbus, 2018), se destacan tres puntos cruciales:

- Aquellas empresas que no trabajen en el *Big Data* perderán fuerza competitiva y podrían extinguirse.
- La inversión en *Big Data* crece continuamente, así como los ingresos derivados de su uso.
- El departamento que más ha evolucionado con la introducción del *Big Data* ha sido el departamento de ventas y marketing.

Además, el artículo apunta que un 59% de directivos piensa que el *Big Data* mejoraría con el uso de la inteligencia artificial en la empresa. La inteligencia artificial es un componente determinante en el BI a través del *Data Mining*, como se verá más adelante.

Todo ello conlleva una transformación en las empresas, empezando por la creación de un nuevo área o departamento dentro de sus organizaciones liderado por el *Chief Data Officer* (CDO). Esta nueva área está caracterizada, sobre todo, por el perfil de sus trabajadores: expertos en la administración y transformación de los datos en las empresas y su arquitectura, así como en el análisis y predicción de estos. Antiguamente, este papel lo desempeñaba el departamento *Information Technology* (IT) pero no eran capaces de abordar la transformación tecnológica que supone el *Big Data* (Gourévitch, et al., 2017).

Un caso sencillo: a medida que el tiempo transcurría, la capacidad de almacenamiento y procesamiento de los datos aumenta hasta llegar a límites que las empresas no pueden abordar, ya sea por falta de espacio o falta de recursos. Es así porque el departamento IT solamente se preocupaba de contratar más espacio hasta que surgió este problema. Esta acción se conoce como *scale-up*. Entonces, se crearon sistemas que conectan múltiples nodos de manera que se podían ingerir más datos sin necesidad de contar con más espacio y reduciendo el tiempo de procesamiento. Esto se conoce como *scale-out* y aquí entraron en escena los científicos de datos.

Sin embargo, con la creación de una nueva sección no es suficiente, tal como se indica en el artículo "*Data-Driven Transformation: Accelerate at Scale Now*" (Gourévitch, et al., 2017). La Figura 1.3, plantea las cuestiones más cruciales a la hora de implantar este nuevo sistema para los datos:

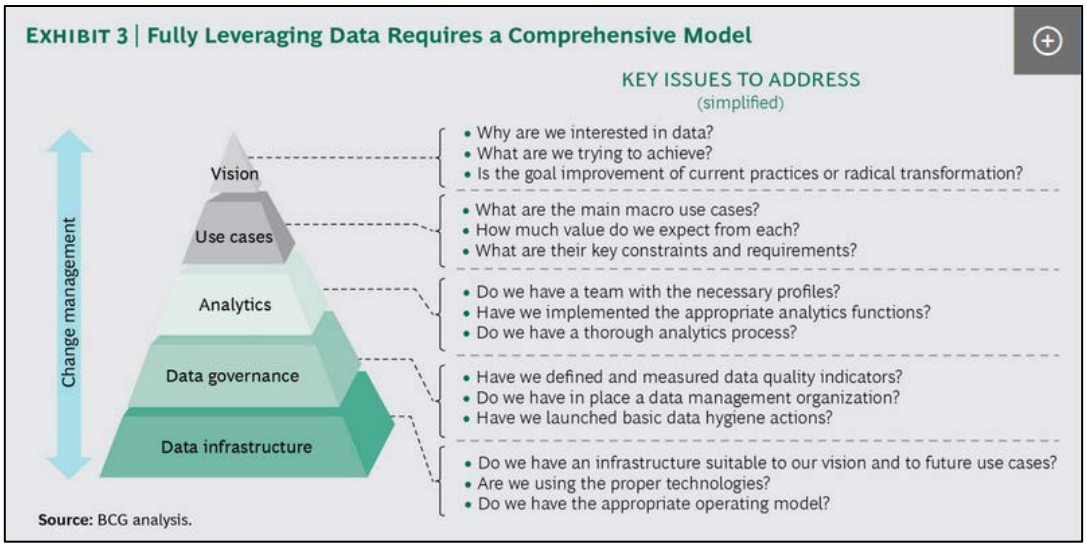


Figura 1.3 - Temas Clave de la Integración del Área de Datos. Fuente; (Gourévitch, et al., 2017)

En definitiva, esta transición conlleva un cambio de cultura organizacional que implica a todos los sectores y niveles de la empresa. Esto no significa que todos los empleados deban ser expertos en el campo de los datos, pero seguramente implique aprender nuevas maneras de trabajar.

1.2. NECESIDADES DE LAS EMPRESAS: FORMACIÓN DE EMPLEADOS EN ANÁLISIS DE DATO

En España, según el portal *InfoJobs*, los puestos de Científico de Datos y Analista de Datos son los puestos emergentes del futuro. “Se consideran puestos emergentes aquellos que apenas existían en el mundo laboral, suelen tener un bajo nivel de competencia y se convierten en buenas oportunidades de empleo con relación al salario.” (InfoJobs, 2018).

	2009	2010	2011	2012	2013	2014	2015	2016	2017
Back - Front end	124	198	380	614	997	2.230	4.450	7.921	**9.822**
Programador Móvil	213	316	1.123	1.688	2.302	3.861	5.373	6.237	**6.253**
Desarrollador soluciones Big Data	45	25	27	162	646	1.793	3.447	4.292	**5.494**
Especialista en Ciberseguridad	160	161	275	261	219	508	1.054	1.265	**1.795**
Especialista en Agile / Scrum	4	60	100	281	339	325	650	1.241	**1.635**
Business Analyst / Data Analyst	160	248	372	289	444	697	819	1.210	**1.417**
Cloud	1	66	88	166	277	461	892	1.135	**1.389**
Robótica	49	68	140	96	164	177	333	759	**904**
Especialista en User Experience	154	123	186	230	299	643	884	1.018	**859**
Data Scientist	8	0	372	289	444	695	168	314	**584**

Figura 1.4 - Evolución de puestos emergentes. Fuente: (InfoJobs, 2018)

En la Figura 1.4, resaltado en rojo, se observa la tendencia creciente lo que significa que es una buena oportunidad para apostar por estos puestos. Además, la competencia es baja ya que como mucho son 30 los inscritos por vacantes (InfoJobs, 2018), un indicador de que apenas hay gente formada en estos ámbitos. También se puede observar otro puesto muy demandado relacionado con el *Big Data*, el de desarrollador, pero si no se ha señalado es debido a que se inclina más al ámbito de la programación.

El BI es un área muy técnica y necesita de conocimientos muy cuantitativos, como tienen los estadísticos y los matemáticos, para identificar los patrones o las anomalías que ellos ven en los números. Por ello, son los perfiles que más buscan las empresas para este tipo de trabajos, incluso para los analistas de negocio.

La Figura 1.4 recoge un listado de los grados en ciencia de datos en España. El de la Universidad Carlos III de Madrid se inauguró en 2018 y solamente cuenta con 45 plazas (Universidad Carlos III de Madrid, 2019). Si se supone que hay 45 graduados cada año en cada centro, se obtienen un total de 540 graduados al año, bastante por debajo de las 2001 vacantes publicadas en 2017, sin mencionar que hoy en día no se graduarían los alumnos de la Universidad Carlos III puesto que se inauguró el curso pasado.

Sí que es cierto que este estudio no recoge la cantidad de másteres, tanto oficiales como propios relacionados con la ciencia de datos que existen en España, pero no cabe lugar a dudas que es una buena oportunidad para adentrarse en este campo.

Tabla 1.1 - Titulaciones en Ciencia de Datos. Fuente: (Min. Ciencia, Innovación y Universidades, 2018)

	Título	Universidad
1	Graduado/a en Ciencia de Datos Aplicada Applied Data Science	Universitat Oberta de Catalunya
2	Graduado/a en Ciencia de Datos	Universidad Pública de Navarra
3	Graduado/a en Ciencia de Datos	Universitat de València (Estudi General)
4	Graduado/a en Ciencia de Datos	Universitat Politècnica de València
5	Graduado/a en Ciencia e Ingeniería de Datos	Universidad Carlos III de Madrid
6	Graduado/a en Ciencia e Ingeniería de Datos	Universidad Politécnica de Catalunya
7	Graduado/a en Datos y Analítica de Negocio Bachelor in Data and Business Analytics	IE Universidad
8	Graduado/a en Gestión Empresarial Basada en el Análisis De Datos (Business Analytics)	Universidad Europea de Madrid
9	Graduado/a en Ingeniería de Datos	Universidad Autónoma de Barcelona
10	Graduado/a en Ingeniería en Matemática Aplicada al Análisis de Datos	Universidad Europea de Madrid
11	Graduado/a en Ingeniería Matemática en Ciencia de Datos Mathematical Engineering on Data Science	Universidad Pompeu Fabra
12	Graduado/a en Matemática Computacional y Analítica de Datos	Universidad Autónoma de Barcelona

2. OBJETIVOS DEL TRABAJO

2.1. OBJETIVO GENERAL

El objetivo general de este trabajo es la aplicación de dos de los principales motores del *Business Intelligence*: visualización de datos (*Data Visualization*) y minería de datos (*Data Mining*).

2.2. OBJETIVOS ESPECÍFICOS

El objetivo específico de este trabajo es la extracción de nueva información a partir de los datos de absentismo laboral de una empresa. Esta información se conseguirá a través de:

- La búsqueda de patrones o comportamientos entre los datos.
- La aplicación de algoritmos basados en la estadística y/o probabilidad.

Posteriormente se pretende generar conocimiento que permita a los responsables de la compañía tomar decisiones con el objetivo de mejorar y seguir creciendo.

3. MARCO TEÓRICO

3.1. BUSINESS INTELLIGENCE (BI) Y BUSINESS ANALYTICS (BA)

Según el blog de IBM (Plowden, 2018) "el BI es un conjunto sincronizado de programas, servicios y procesos de trabajo para ingerir datos y presentarlos de manera visual a través de informes, gráficos y mapas."

Otra interesante definición obtenida de la web CIO (Pratt, 2017): "el BI aprovecha los programas y los servicios que ofrecen para transformar los datos en inteligencia (conocimiento) que informa acerca de las decisiones tácticas y estratégicas que la empresa debe tomar."

Entre las dos definiciones se obtiene una buena idea de lo que es el BI y para qué sirve. Este término suele ir muy asociado al del *Business Analytics (BA)*. De hecho, da lugar a muchas confusiones porque los empresarios piensan que es lo mismo, pero existen pequeños matices entre los dos términos. Tal y como explica la web *SelectHub* (Adair, 2018) "el BI usa datos pasados y presentes para optimizar la situación actual con lo que está pasando mientras que el BA se encarga de usar datos históricos para analizar el presente y prepararse para el futuro". Por tanto, en la Figura 1.2 se afirma que las etapas blancas están caracterizadas por el BA mientras que las etapas azules son más propias del BI.

3.2. APLICACIONES DEL BUSINESS INTELLIGENCE EN EMPRESAS

Sin lugar a duda las empresas que más usan el BI son las empresas multinacionales que manejan millones de datos. Las que ponen el BI al servicio de los demás, son, principalmente, Microsoft, IBM, SAP, SAS y Oracle entre otros ejemplos menos conocidos (Predictive Analytics Today, s.f.). Pero las que más usan las herramientas BI son empresas que tienen millones de clientes/usuarios como son las redes sociales Facebook, Twitter, Instagram… Así lo indica Bernard Marr en su libro *"Big Data in Practice"* (Marr, 2016).
Para la creación de estas empresas se necesitaron de conocimientos más bien técnicos que administrativos. Los ejemplos que se verán a continuación son más prácticos de cara al empresario. Se tratan de empresas del sector *retail*, como Wal-Mart y Amazon, que han usado el BI para ser más eficientes y destacan por sus decisiones tomadas gracias al BI adoptando las estrategias adecuadas para desarrollar el negocio y crecer.

3.2.1. Wal-Mart

(Marr, 2016) pone como ejemplo la empresa Wal-Mart, entre muchos otros ejemplos de empresas que usan el BI. Wal-Mart es la empresa más grande de *retail* de los Estados Unidos (y del mundo). En España, el equivalente sería Carrefour (aunque la empresa es francesa). Estas empresas se encargan de vender todo tipo de cosas, principalmente para el hogar.

En 2011, Wal-Mart decidió crear un proyecto basado en el análisis de datos percatándose de la importancia que éstos tenían dentro de su negocio. Este proyecto fue denominado *Data Café* y consiguió una reducción en el tiempo de respuesta ante

una anomalía de 2-3 semanas a cuestión de 20 minutos, por ejemplo, creando alertas que saltaran cuando algo no iba bien.

- En una ocasión, la sección de ultramarinos de la empresa no entendía por qué la venta de un determinado producto cayó en picado. El equipo de analistas de *Data Café* detectó en poco tiempo que se debía a un error en el precio de venta.
- En Halloween, hicieron seguimiento de los productos nuevos para esa campaña. El equipo de *Data Café* pronto descubrió que en algunos lugares no se vendía. Resultó que en estos sitios el producto ni siquiera se había puesto en las estanterías.

El *Data Café*, además de descubrir los problemas, permiten dar con la solución rápidamente, principal objetivo del BI.

3.2.2. Amazon

Otro ejemplo de (Marr, 2016) es Amazon. El éxito de Amazon se ha basado fundamentalmente en la aplicación de las tecnologías en todos los aspectos de la empresa. No hace falta mencionar el uso de sus robots autodirigidos por sus almacenes.

Más allá de mejorar su eficiencia logística, Amazon ha conseguido perfiles de consumidores a través de su “motor de recomendación” basándose en los datos que recibe cuando se navega por su web.

Con esos datos, lo que Amazon pretende es saber todo lo posible acerca del usuario para que pueda predecir qué es lo que se quiere y cuándo se quiere. De esta manera, es la propia web la que sugiere productos sin que surja la necesidad de buscarlos.

Pero ¿cómo obtiene Amazon los datos? De múltiples maneras:

- Desde el instante en que se accede a su sitio web no pierde ni un momento para recopilar datos. Ahora que incluye *Amazon Prime* para ver series, se sabe incluso cuánto tiempo un usuario se pasa enganchado a ellas.
- Con el domicilio del perfil del usuario se sabe en qué vecindario vive y se puede deducir los ingresos del usuario.
- Cuando se compra o se mira un producto de Amazon, se conoce la hora de compra para intuir las costumbres y comportamientos del usuario.

Con estos datos, Amazon construye un perfil o lo asigna a uno ya creado que se asemejará a los que tiene modelados en sus bases de datos. Por tanto, Amazon puede encontrar personas con gustos similares e inquietudes para ofrecer lo que estas personas ya hayan comprado o mirado anteriormente.

3.3. PROCESO BUSINESS INTELLIGENCE

Ahora que se tiene una amplia visión del BI, es hora de saber cuál es la mecánica del BI:

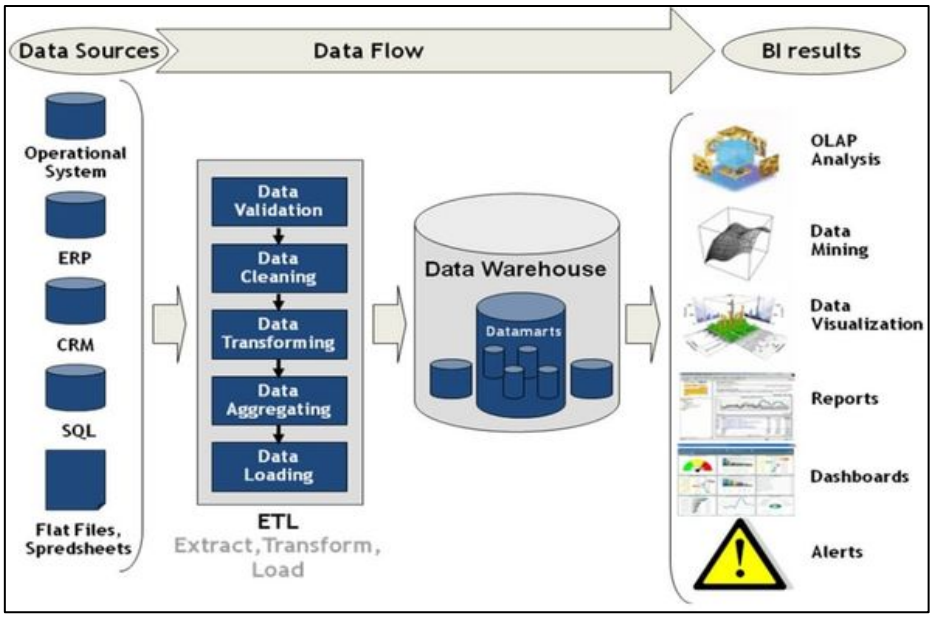


Figura 3.1. Esquema BI. Fuente: (Larion, s.f.)

La Figura 3.1 muestra todo el proceso de transformación de la materia prima, los datos, a información.

1) Se extraen los datos que se desean de las diferentes plataformas: *Operational Systems (OS), Enterprise Resource Planning (ERP), Customer Relationship Management (CRM), Structured Query Language (SQL)* y/o archivos comunes como plantillas Excel, por ejemplo.
2) **ETL (*Extract, Transform, Load*)**: proceso mediante el cual los datos escogidos de las diferentes fuentes se convierten a un formato que pueda ser analizado y guardado en un almacén de datos diseñado para su análisis. (SAS, s.f.)
3) **Almacenes de datos (*Data Warehouse*)**: pueden contener subsecciones (*Data Marts*) diseñados específicamente para que un equipo/departamento de una organización los pueda extraer sin necesidad de buscar y seleccionar los datos por todo el *Data Warehouse*. (Sisense, s.f.)
4) A través del almacén de datos aplicamos las distintas metodologías para conseguir la información que deseamos.

Un *Data Warehouse* no es lo mismo que una base de datos. Dale Sanders, en su artículo "*Wal-Mart and the Birth of the Data Warehouse*" (Sanders, 2013), explica las características de un *Data Warehouse*:

- **Estandariza los datos provenientes de múltiples fuentes**. En grandes empresas es bastante común recoger datos de, al menos, 50 fuentes diferentes tanto de fuentes internas como externas.
- **Análisis multi-organizativo**: están diseñados para permitir el análisis de datos de y entre las diferentes secciones de la organización.
- **Ayuda a identificar tendencias y modas entre los datos**, lo que permite agilizar el trabajo de los usuarios.
- **Gran capacidad**: puede contener miles de millones de datos que ocupan cientos de Terabytes. En tiempo, puede almacenar hasta 30 años de datos históricos si se refiere a empresas como Wal-Mart y Amazon.

El artículo, además de profundizar sobre el *Data Warehouse*, pone otro ejemplo de Wal-Mart usando el BI. En este caso, pasado el día de Acción de Gracias, bastante famoso en Estados Unidos, el gerente de la Costa Este puso un ordenador a la venta de una manera bastante llamativa lo que provocó unas ventas enormes. El equipo BI de Wal-Mart, seguramente el mismo equipo *Data Café* del ejemplo pasado, alertó sobre el éxito de las ventas, de manera que la organización mandó a que todos los gerentes pusieran a la venta el ordenador de manera similar al responsable que lo hizo en la Costa Este.

3.4. TÉCNICAS BUSINESS INTELLIGENCE

Una vez diseñado el *Data Warehouse*, se puede generar información en las distintas formas que aparece en la Figura 3.1. El último esquema "*BI Results*" no es del todo correcto pues los procesos BI se agrupan en *OLAP Analysis*, *Data Mining* y *Data Visualization*. "*Reports*", "*Dashboards*" y "*Alerts*" pertenecen al último grupo mencionado.

3.4.1. Visualización de datos (Data Visualization)

Es necesario distinguir qué es cada elemento mencionado anteriormente. Los únicos que pueden dar lugar a confusiones son "*Reports*" y "*Dashboards*" puesto que son términos similares:

Tabla 3.1 - Report vs.Balanced Scorecard vs. Dashboard.
Fuente: Elaboración propia a partir de la información obtenida de (Jody, 2009) y (Blitz, 2018)

	Report	**Dashboard**
Documento	Texto/Tablas/Gráficos	Gráficos
Interacción	Estática	Dinámica
Contenido	Complejo/Denso	Sencillo/Conciso
Elementos	Datos Históricos	KPIs
Línea Temporal	Diaria/Mensual/Trimestral/…	Tiempo Real

La Tabla 3.1 recoge los aspectos más determinantes entre un tipo de visualización y otro, recogidos de los artículos "*Dashboard vs. Reports – Which do you need?"* (Blitz, 2018) y *"The Difference Between a Dashboard and a Report"* (Jody, 2009).
No hay uno mejor que otro. Como bien indica Shelby Blitz en su artículo, depende de múltiples factores como pueden ser la necesidad, el tiempo disponible y/o las personas a las que van dirigidas. Normalmente, para altos cargos de organizaciones, es más práctico el uso de cuadros de mandos (*dashboards*) aunque si estas personas desean profundizar o conocer más sobre un aspecto en concreto, son mejores los informes (*reports*). Los informes son más utilizados para la identificación de patrones de comportamiento o tendencias que van tomando las variables de un determinado modelo (Blitz, 2018).

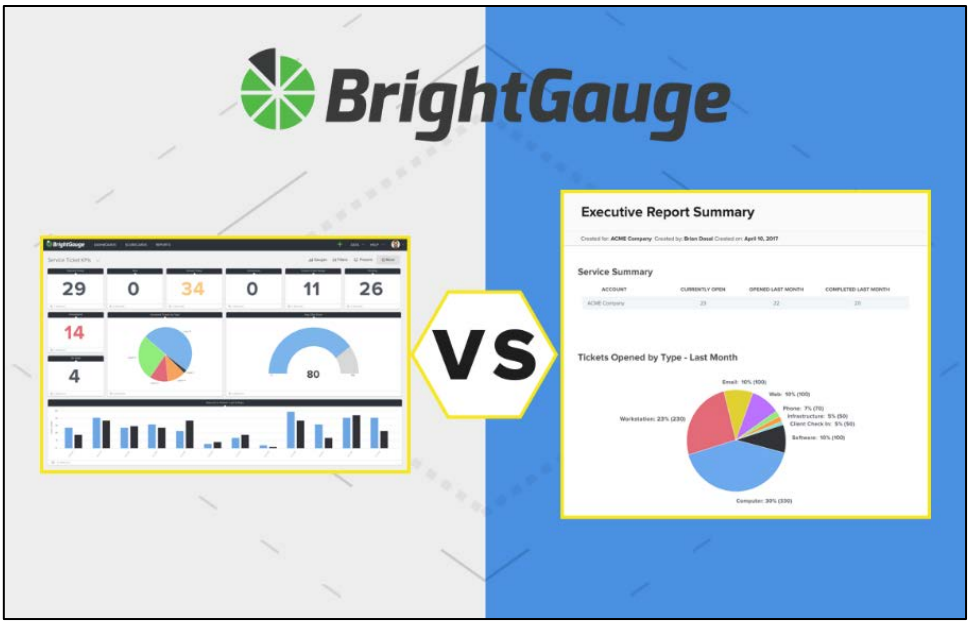


Figura 3.2 - Dashboard vs. Report. Fuente: (McCluney, 2017)

La Figura 3.2 muestra, de una manera simple pero ilustrativa, un cuadro de mandos (a la izquierda) y un informe (a la derecha).

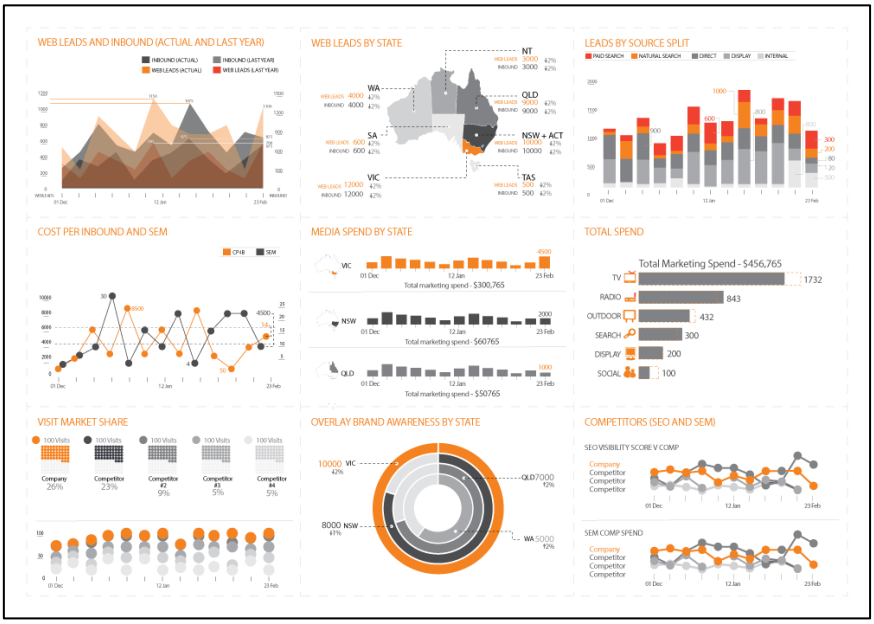


Figura 3.3 - Diseño Dashboard Tableau. Fuente: (DataLabs, s.f.)

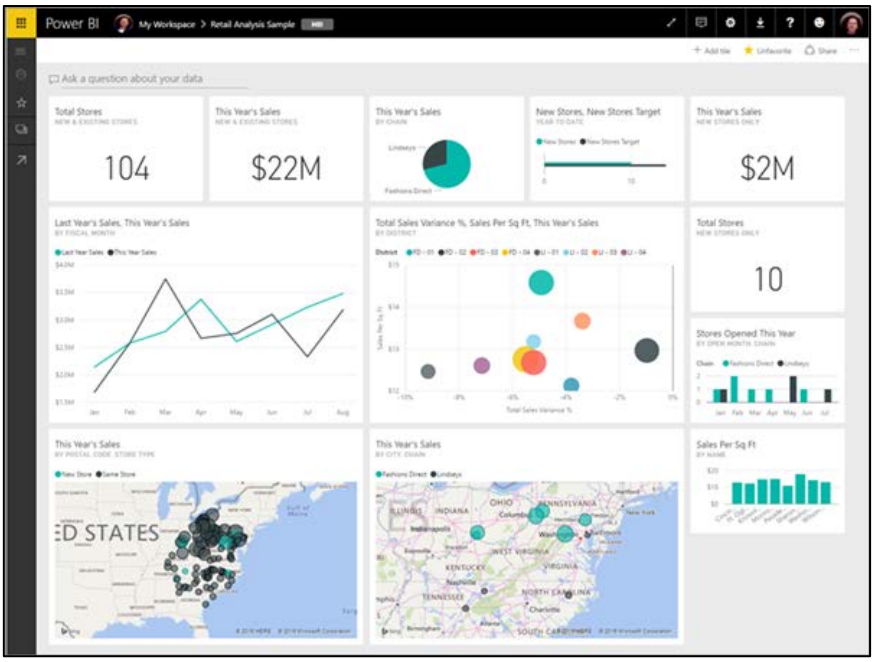


Figura 3.4 - Diseño Dashboard Power BI Mobile. Fuente: (Microsoft, 2018)

Hay otro tipo de visualización denominado marcador (*scoreboard)*, parecido al *dashboard.* El *scoreboard* compara los resultados de la empresa en cada intervalo de tiempo con los objetivos marcados. Por tanto, el *scoreboard* está diseñado especialmente para que los directivos puedan adoptar decisiones estratégicas según los resultados que aparecen en él (Savkin, s.f.).

Resumiendo, las técnicas de visualización de datos *(Data Visualization)* mejoran la eficiencia en las organizaciones transmitiendo la información de una manera rápida y sencilla para todos los usuarios.

3.4.2. OLAP

El acrónimo OLAP procede de *Online Analytical Processing*. “OLAP crea información a partir de análisis multidimensionales, además de cálculos complejos, descubrimiento de patrones y modelos de datos sofisticados.” (OLAP, s.f.).

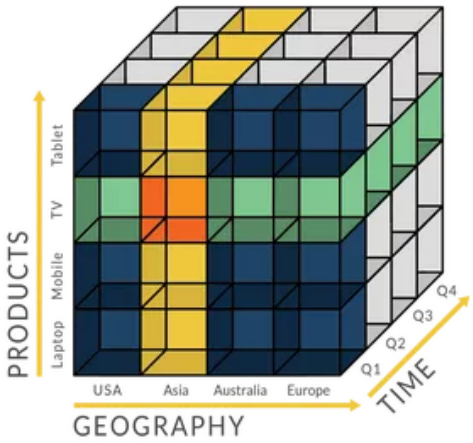


Figura 3.5 - Cubo OLAP. Fuente: (OLAP, s.f.)

Al contrario, como sucede con las tablas Excel, que usan solamente dos tipos de variables (eje x – eje y), los cubos OLAP trabajan con tres como se muestra en la Figura 3.5. En ese caso en concreto, se puede suponer por el cubo naranja

seccionado, que se desea conocer el consumo de los usuarios de televisión y asiáticos, durante el primer trimestre del año.

Buscando una mejor definición: “OLAP es una tecnología de bases de datos que está optimizada para consultar e informar en vez de procesar.” (Microsoft, s.f.). Además, el artículo indica que los datos son extraídos del *Data Warehouse* y estructurados en forma de cubo para permitir mayor agilidad al acceso y al análisis de los datos.

“Los cubos no tienen por qué ser exactamente cubos en el sentido matemático más estricto porque no tienen que ser lados iguales” (Microsoft, s.f.). El ejemplo más sencillo se puede aplicar a la Figura 3.5 si en vez de trimestres se usa cuatrimestres en el eje temporal o si se añade América del Sur en el eje geográfico. Esta transformación del cubo es posible porque los datos se pueden clasificar en diferentes grupos dentro de la propia variable, lo que permite mayor diversidad de análisis por subcategorías. De este modo se pueden plantear multitud de escenarios que serán mayores cuantas más variables (dimensiones) se planteen. OLAP no se limita solamente a tres dimensiones, aunque ya no formaría un cubo.

Además, Dan Vlamis, en su presentación “*Oracle OLAP in the Real World: Case Studies from the Trenches*.” (Vlamis, 2005) especifica: “OLAP tiene dos consecuencias inmediatas: la parte online que requiere que las respuestas a las consultas (*queries*) sean rápidas, y la parte analítica que indica que las consultas en sí son complejas.” En su presentación describe varios ejemplos prácticos de OLAP en empresas. Un ejemplo es de una empresa que tiene un *call center* en su propia organización y desean minimizar los tiempos de espera de las llamadas de clientes sin reducir el número de la plantilla. La solución pasó por OLAP, analizando tanto los tiempos de espera como el total de cada operador, además del tipo de llamada (resuelta, satisfacción del cliente, rellamada, …) y establecer unos bonus salariales que permitan obtener resultados eficientes.

3.4.3. Data Mining: definición, componentes y fases

Otra de las técnicas que se puede aplicar a un *Data Warehouse* es la minería de datos o *Data Mining*. También se usa el término *Knowledge Discovery*.

Data Mining “es la técnica de análisis de datos que permite la obtención de información valiosa y oculta entre la gran cantidad de datos que surgen durante el transcurso operativo del trabajo.” (Moss & Atre, 2003). “*Data Mining* también es capaz de encontrar anomalías, correlaciones y patrones entre los datos.” (SAS, s.f.)

Figura 3.6 - Componentes Data Mining (SAS, s.f.) g

En el motor del *Data Mining* intervienen tres factores (Figura 3.6): la estadística (*statistics*), que estudia numéricamente los datos; la inteligencia artificial (*artificial intelligence*), es decir, la inteligencia desarrollada por las personas aplicadas a programas o máquinas para que actúen como los humanos; y el aprendizaje máquina

(*machine learning*), algoritmos que son capaces de aprender de los datos para formular predicciones (SAS, s.f.).

Para saber cómo influye la estadística en el proceso del *Data Mining* es muy interesante ver la Figura 3.7 extraída del libro *"Business Intelligence Roadmap"* (Moss & Atre, 2003).

Statistical Analysis	***Data Mining***
• Statisticians usually start with a hypothesis.	• Data mining does not require a hypothesis.
• Statisticians have to develop their own equations to match their hypothesis.	• Data mining algorithms can automatically develop the equations.
• Statistical analysis uses only numerical data.	• Data mining can use different types of data (e.g., text, voice), not just numerical data.
• Statisticians can find and filter dirty data during their analysis.	• Data mining depends on clean, well-documented data.
• Statisticians interpret their own results and convey these results to the business managers and business executives.	• Data mining results are not easy to interpret. A statistician must still be involved in analyzing the data mining results and conveying the findings to the business managers and business executives.

Figura 3.7 - Estadística vs. Minería de Datos. Fuente: (Moss & Atre, 2003)

Cabe resaltar algunos aspectos de la Figura 3.7:

- Un ejemplo de datos de voz que se menciona es, por ejemplo, las voces de los asistentes para iPhone, Windows o Amazon: Siri, Cortana y Alexa, respectivamente, dotadas de inteligencia artificial.
- Una de las fases del *Data Mining* es el preprocesamiento de datos que se encarga de limpiar los datos inservibles para el proceso.

Para conocer qué es la inteligencia artificial y el aprendizaje máquina se puede poner como ejemplo a *AlphaGo*.

Go es el "ajedrez asiático". A pesar de que el ajedrez es complejo, nada comparado con Go. Independientemente de las reglas de cada juego, el número de posibilidades y/o situaciones de Go frente al Ajedrez son inmensamente mayores, debido principalmente a las dimensiones de cada tablero (19x19 del Go frente al 8x8 del Ajedrez) y, por tanto, el número de jugadas posibles.

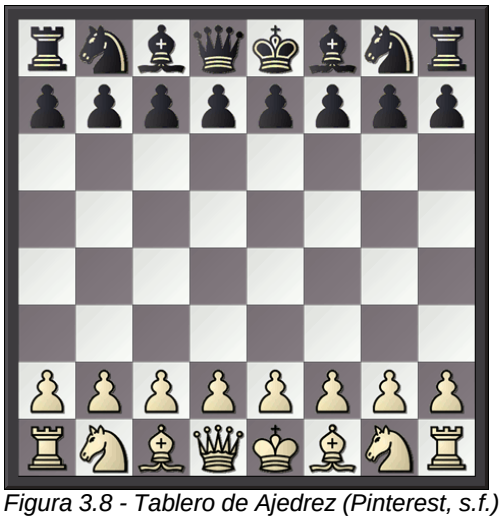


Figura 3.8 - Tablero de Ajedrez (Pinterest, s.f.)

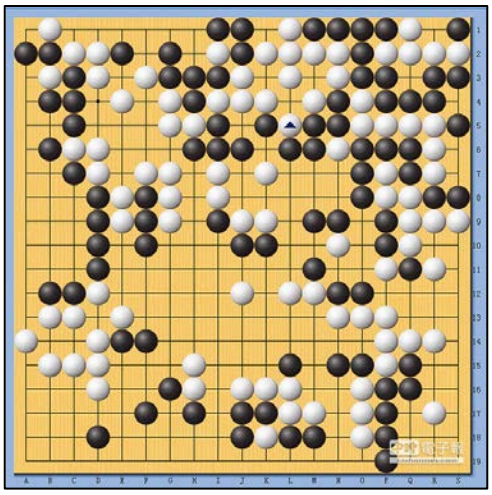


Figura 3.9 - Tablero de Go (ChinaTimes, 2017)

A través de la inteligencia artificial, se construyó *AlphaGo*, un motor a base de algoritmos tanto de búsqueda entre los millones de partidas de élite que tiene almacenados, como de cálculo de probabilidades de victoria en determinadas situaciones del juego de manera que derrotó en un evento organizado al mejor de 5 partidas al que era en su momento el mejor jugador del mundo de Go, el surcoreano Lee Sedol, por 4-1.

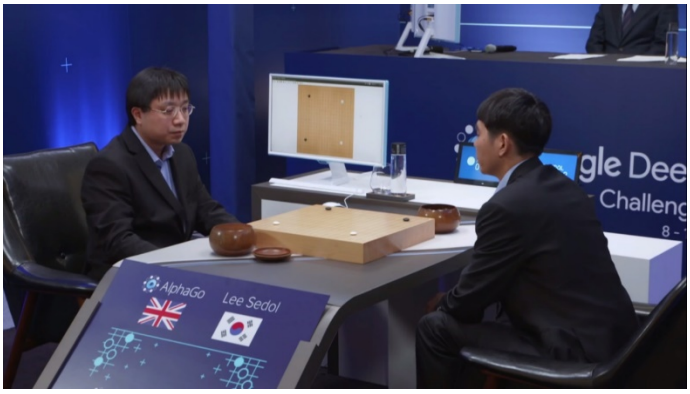


Figura 3.10 - AlphaGo vs. Lee Sedol. Fuente: (AlphaGo, 2017)

DeepMind, el equipo de desarrolladores de *AlphaGo* dio un paso más allá y creó *AlphaGo Zero* que cuenta con aprendizaje máquina, convirtiéndose en su propio prueba-error. El sistema cuenta con una red neural que solamente sabe las reglas básicas del juego. Empieza a jugar partidas contra sí mismo usando esta red y un potente algoritmo de búsqueda entre las partidas que va disputando. A medida que va jugando, la red neural va aprendiendo conceptos más complejos por su cuenta hasta predecir las jugadas futuras de la partida, convirtiéndose en un rival imbatible. Solamente le bastaron 3 días de auto entrenamiento para derrotar a su versión anterior *AlphaGo* por 100-0 (Hassabis & Silver, 2017).
“Esta técnica es más potente que las versiones anteriores debido a que carece de las limitaciones humanas.” (Hassabis & Silver, 2017). En otras palabras, se demostró que la inteligencia artificial es enormemente más eficiente que los humanos.

Una vez vistos los componentes de la minería de datos, en último lugar se verán las fases de esta técnica:

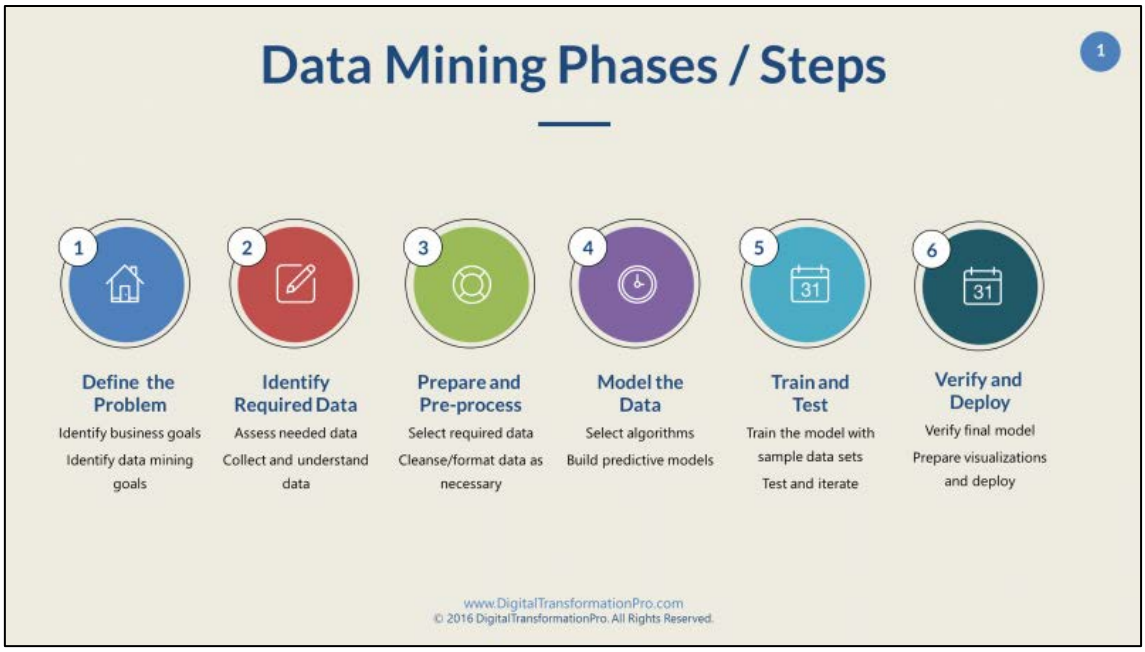


Figura 3.11 - Fases del Data Mining. Fuente: (Dontha, 2018)

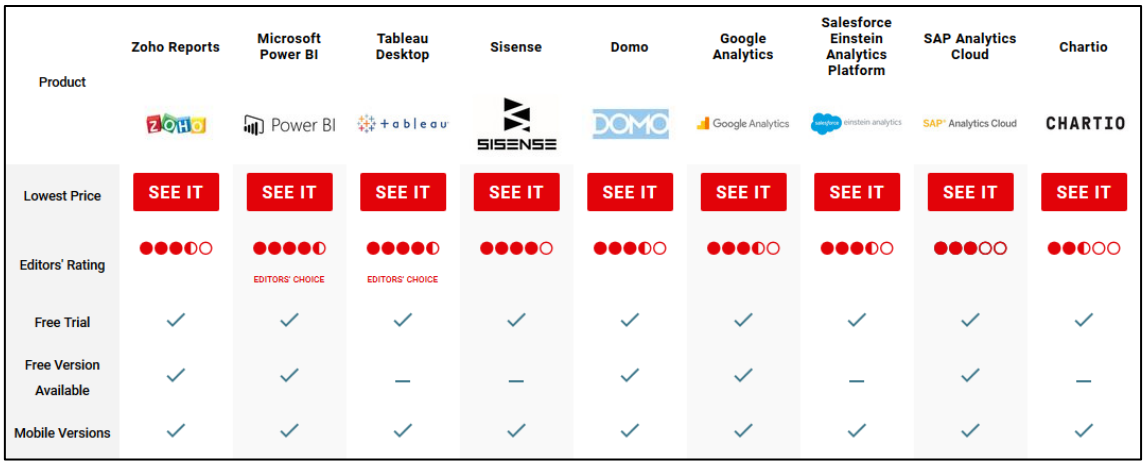
Los pasos descritos en la Figura 3.11 son un ejemplo del proceso *Data Mining*. Eso quiere decir que no necesariamente hay que seguir los mismos pasos cada vez que apliquemos *Data Mining*, como sucederá en la parte práctica del trabajo.
La definición del problema (paso 1), no es una fase explícita del *Data Mining* aunque sí elemental para sacar provecho de la técnica, tal y como indica Ramesh Dontha. Según explica, se debe identificar cuáles son los problemas de la empresa y, dependiendo de a lo que se enfrente la compañía, seleccionar los datos necesarios (paso 2) para analizarlos. Una vez obtenidos, el paso siguiente será preparar los datos para que se pueda trabajar con ellos en las demás fases del *Data Mining* (Dontha, 2018).

Como se puede apreciar la técnica de *Data Mining* es una pieza importante y muy poderosa del BI hasta el punto de poder predecir lo que va a pasar.

4. SITUACIÓN ACTUAL

4.1. HERRAMIENTAS BUSINESS INTELLIGENCE

Actualmente hay varias compañías que ofrecen BI no solamente a empresas, sino también a administraciones públicas. Existen versiones gratuitas que tienen restricciones en las funcionalidades del software, pero son suficientes para empresas pequeñas o que estén iniciando. De hecho, hay tutoriales online oficiales sobre cómo usar estos programas lo que facilita su uso, aunque requiere de ciertos conocimientos técnicos de informática por lo que no todo el mundo sería capaz de seguir los pasos de estos tutoriales ya que los mismos dan por hecho que se tienen unas competencias y habilidades adquiridas.



Product	Zoho Reports	Microsoft Power BI	Tableau Desktop	Sisense	Domo	Google Analytics	Salesforce Einstein Analytics Platform	SAP Analytics Cloud	Chartio
Lowest Price	SEE IT	SEE IT	SEE IT	SEE IT	SEE IT	SEE IT	SEE IT	SEE IT	SEE IT
Editors' Rating	●●●◐○	●●●●◐ EDITORS' CHOICE	●●●●◐ EDITORS' CHOICE	●●●●○	●●●◐○	●●●◐○	●●●◐○	●●◐○○	●●◐○○
Free Trial	✓	✓	✓	✓	✓	✓	✓	✓	✓
Free Version Available	✓	✓	–	–	✓	✓	–	✓	–
Mobile Versions	✓	✓	✓	✓	✓	✓	✓	✓	✓

Figura 4.1 - Comparativa PCMag Software BI. Fuente: (Baker, 2019)

La Figura 4.1, muestra varios ejemplos de herramientas BI. Todos los programas ofrecen una versión móvil para vincularse a la versión completa (y no gratuita), lo que conlleva al *Business Intelligence 3.0*.

4.2. ETAPAS BUSINES INTELLIGENCE: 1.0 – 2.0 – 3.0

La Figura 1.2, da una visión amplia acerca de la evolución del BI desde una perspectiva funcional. Existen otros elementos que marcan una nueva etapa del BI:

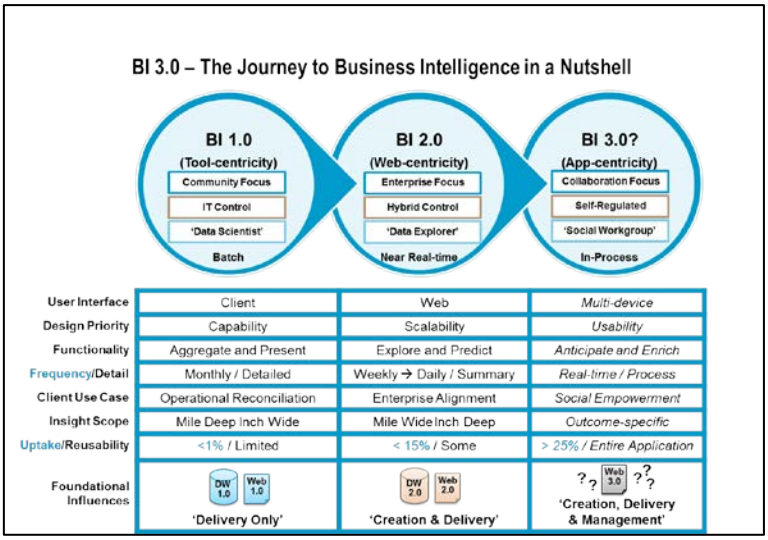


Figura 4.2 - The Journey to Business Intelligence in a Nutshell. Fuente: (Gratton, 2012)

La Figura 4.2, muestra que uno de los elementos diferenciadores de la transición del BI 1.0 al 2.0 y del que ya se ha hablado es la capacidad de predecir lo que va a ocurrir con los datos, entre otros.

Simon Gratton en su artículo *"BI 3.0 The Journey to Business Intelligence. What does it mean?"* (Gratton, 2012) describe perfectamente la transición de una etapa a otra del BI:

> *Business Intelligence 1.0*
>
> Nace la figura del científico de datos en la empresa, capaz de proporcionar los datos que se le pedía de una manera precisa, convirtiéndose rápidamente en un cuello de botella organizativo puesto que las exigencias eran cada vez

mayores tanto en la rapidez de respuesta como en la información requerida y las peticiones recibidas por toda la empresa.

Business Intelligence 2.0

Todo el mundo puede tener acceso a los datos y realizar los cálculos y las previsiones que quiera, incluso combinarlos, pero el acceso a estos datos está restringido bajo el departamento informático y el tiempo que transcurre hasta obtenerlos puede ser largo. En esta ocasión se crearon programas destinados a que la obtención de los datos fuera instantánea, efectiva y entendible por todo usuario.

Business Intelligence 3.0

Los programas BI son cada vez más potentes y te ofrecen gran variedad de opciones para combinar los datos… para los usuarios expertos, como los informáticos. Para usuarios básicos se requiere cierto nivel técnico para poder trabajar los datos.

En la actualidad el BI se encuentra en un proceso de transición al 3.0 enfocado a la accesibilidad del BI mediante cualquier dispositivo, en cualquier momento y de la manera más sencilla posible, aunque queda un gran camino por recorrer. El BI 3.0 requerirá, sobre todo, un gran consenso y trabajo en equipo por parte de todos los usuarios.

5. METODOLOGÍA

La metodología usada para la consecución de los objetivos ha sido la aplicación de técnicas del BI como *Data Visualization* y *Data Mining*. Las herramientas y los materiales utilizados han sido:

- WEKA, un *software open source* especializado en *Data Mining*, gracias al cual se han realizado los análisis necesarios (WEKA, s.f.).
- Un conjunto de datos (*dataset*) denominado “Abseentism_at_work” que contiene datos acerca de las ausencias laborales de los trabajadores y las características de estos en una empresa de paquetería de Brasil entre julio de 2007 a julio de 2010. Este *dataset* ha sido extraído del repositorio de la Universidad de California, Irvine (University of California Irvine Machine Learning Repository, s.f.), y fue elaborado por Andrea Martiniano, Ricardo Pinto Ferreira y Renato Jose Sassi para una investigación académica de la Universidad Nove de Julho. (Martiniano, et al., 2018)
- Los creadores de WEKA incluyen material de ayuda acerca de la utilización del programa y de la propia técnica *Data Mining*…
    - A través del libro “*The WEKA Workbench*” (Witten, et al., 2016)
    - En su perfil de *Youtube* WEKAMOOC se encuentran tres cursos de nivel básico, intermedio y avanzado, respectivamente (WEKA, s.f.).

6. DESARROLLO EMPÍRICO

6.1. PREPROCESAMIENTO

El *dataset* original “Absenteeism_at_work” está compuesto por 740x21 (filas por columnas). Las filas se denominan casos (*instances)* de las ausencias surgidas en la empresa mientras que las columnas se llaman atributos (*atributes*) que son las propiedades y características de cada caso.

Los atributos aparecen descritos en la Tabla 6.1:

Tabla 6.1 – Listado de atributos.
Fuente: Elaboración propia a partir de la información obtenida de (Martiniano, et al., 2018)

	Atributo	**Tipo**	**Descripción**
01	ID	Nominal	Número identificador del trabajador
02	Razón de Ausencia	Nominal	Razón por la cual el trabajador se ausentó
03	Mes de Ausencia	Nominal	Mes en el que el trabajador se ausentó
04	Día de Ausencia	Nominal	Día de la semana en el que el trabajador se ausentó
05	Estación del Año	Nominal	Estación del año en el que el trabajador se ausentó
06	Gasto Transporte	Numérico	Gasto en el que incurre el trabajador para ir al trabajo
07	Distancia Trabajo-Casa	Numérico	Kilómetros desde la casa del trabajador hasta la sede de trabajo
08	Hora de Servicio	Nominal	Hora a la que comienza a trabajar
09	Edad	Numérico	Edad del trabajador
10	Media Carga de Trabajo al día	Numérico	Media de la carga que un trabajador realiza ese día
11	Objetivo	Numérico	Porcentaje de objetivo por cumplir
12	Fallo Disciplinario	Nominal	Si la razón de la ausencia fue por fallo disciplinario
13	Educación	Nominal	Nivel de educación del trabajador
14	Hijos	Numérico	Número de hijos que tiene el trabajador
15	Bebedor	Nominal	Si el trabajador bebe alcohol ocasionalmente en su vida social
16	Fumador	Nominal	Si el trabajador fuma ocasionalmente en su vida social
17	Animales	Numérico	Número de animales que tiene el trabajador
18	Peso	Numérico	Peso en kilogramos del trabajador
19	Altura	Numérico	Altura en centímetros del trabajador
20	IMC	Numérico	Índice de Masa Corporal del trabajador
21	Horas Absentismo	Numérico	Número de horas laborales ausentadas desde el momento en el que comienza la ausencia

Como se puede ver en la Tabla 6.1. existen dos tipos de atributos:

- Los atributos **numéricos**: los datos de estos atributos permiten realizar estudios estadísticos como, por ejemplo, extraer la media de edad o de hijos de los trabajadores.
- Los atributos **nominales**: cada valor de un determinado atributo se corresponderá con una categoría. Sirven, por ejemplo, para clasificar los casos por agrupaciones si se desea.

Los valores de los atributos más relevantes nominales se pueden ver en las Tablas 6.2, 6.3, 6.4 y 6.5:

Tabla 6.2 - Clasificación Atributo Razón de Ausencia. Fuente: (Martiniano, et al., 2018)

		Atributo Razón de la Ausencia (CID = CÓDIGO OFICIAL DE ENFERMEDADES)	**Color**
	00	Fallo disciplinario	Azul
CID	**01**	Enfermedades infecciosas o parasitarias	
CID	**02**	Neoplasma	
CID	**03**	Enfermedades de la sangre	
CID	**04**	Enfermedades metabólicas y/o endocrinas	
CID	**05**	Enfermedades mentales y/o trastornos	
CID	**06**	Enfermedades del sistema nervioso	
CID	**07**	Enfermedades oculares	
CID	**08**	Enfermedades auditivas	
CID	**09**	Enfermedades del sistema circulatorio	
CID	**10**	Enfermedades del sistema respiratorio	

CID	**11**	Enfermedades del sistema digestivo	
CID	**12**	Enfermedades cutáneas	
CID	**13**	Enfermedades musculoesqueléticas	Gris
CID	**14**	Enfermedades genitales	
CID	**15**	Embarazo y lactancia	
CID	**16**	Ciertas condiciones originarias del sistema prenatal	
CID	**17**	Malformaciones congénitas y deformaciones	
CID	**18**	Síntomas de anormalidades clínicas sin clasificar	
CID	**19**	Lesiones y otras condiciones creadas por causas externas	Verde Oscuro
CID	**20**	Causas externas de obesidad (mórbida) y mortalidad	
CID	**21**	Factores que influencian en la condición de la salud	
	22	Seguimiento médico	
	23	Consulta médica	Turquesa
	24	Donación de sangre	
	25	Examen médico	
	26	Ausencia injustificada	
	27	Consulta fisioterapéutica	Marrón
	28	Consulta dental	Morado

Tabla 6.3 - Clasificación Atributo Educación. Fuente: (Martiniano, et al., 2018)

	Atributo Educación	**Color**
01	Bachiller	Azul
02	Grado Superior	Rojo
03	Posgrado	Celeste
04	Master/Doctor	Gris

Tabla 6.4 - Clasificación Atributos Fallo Disciplinario/Bebedor/Fumador. Fuente: (Martiniano, et al., 2018)

	Atributo Fallo Disciplinario **Atributo Bebedor** **Atributo Fumador**	**Color**
00	No	Azul
01	Sí	Rojo

Tabla 6.5 - Clasificación Atributo Edad 'Discretizado' y Atributo Hijos. Fuente: (Martiniano, et al., 2018)

Atributo Edad 'Discretizado'	**Color**
(-inf-37,33]	Azul
(37,33-47,67]	Rojo
(47,67-+inf)	Celeste

Atributo Hijos	**Color**
0	Azul
1	Rojo
2	Celeste
3	Gris
4	Rosa

Este *dataset* requiere eliminar algunos casos y atributos por las siguientes razones:

- Aparecen casos duplicados.
- Hay un caso en concreto en el que el valor de los atributos corresponde a otro ID. Se corrige.
- Hay varios casos cuyo valor de mes es 0 y no se corresponde con ningún mes. Se eliminan.
- Hay casos cuyo valor de estación es distinto teniendo el mismo valor de mes, por lo que se procede a eliminar el atributo “Estación de Ausencia”.

- Hay varios casos cuyo valor de hora (valor igual a 29) no se corresponde al de una hora real en el atributo “Hora de Servicio”, por lo que se eliminan. Se procede a cambiar el valor 24 por el valor 0 (24h = 0h) para mejor representación de los datos.
- El atributo “Fallo Disciplinario” aparece repetido como valor 0 en el atributo “Razón de Ausencia”, por lo que sobra como atributo.
- El atributo “Índice de Masa Corporal” (IMC) es un indicador que depende del peso y la altura del trabajador, por lo que los dos atributos correspondientes a estas cualidades tampoco son necesarios. Además, se han añadido dos decimales al valor de cada caso para que su representación sea más real, realizando el cálculo del IMC con los datos de peso y altura:

$$IMC = \frac{Peso\ (kg)}{Altura^2\ (m)}$$

Según el Instituto Nacional de Salud de Estados Unidos (NIH), un índice saludable es aquel que está entre los valores 18,5 y 24,9. Más de 24,9 se considera sobrepeso y más de 30 obesidad. Cuando se alcanza el nivel de sobrepeso es más probable tener problemas cardiovasculares (Zelman, 2007).

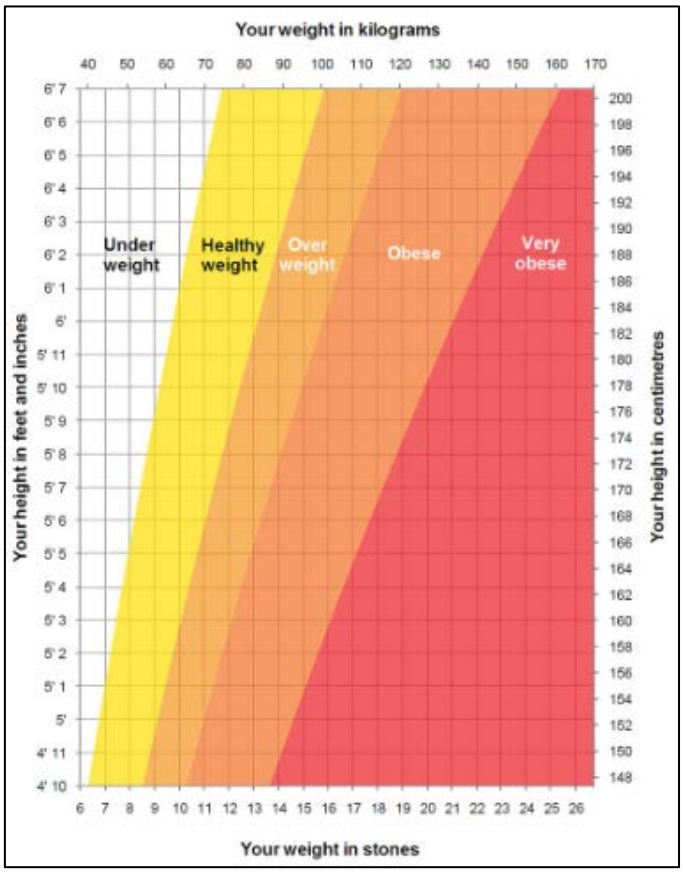


Figura 6.1 - Relación Peso-Altura con Obesidad. Fuente: (National Health Service, 2018)

El IMC no es una ciencia exacta, es decir, no por tener esos valores quiere decir que una persona sea obesa o flaca, siempre y cuando se cuide. Depende de otros factores como el sexo, la edad (menores de 18 años están excluidos en la Figura 6.1), etcétera. Por ejemplo, jugadores de baloncesto o de rugby pueden sobrepasar los valores de 25 de IMC porque necesitan mucha masa muscular debido a su rol en el juego (Zelman, 2007).

- Otros atributos como el atributo 10, “Media Carga de Trabajo al Día”, y el atributo 11, “Objetivo” no son interesantes para los análisis que se van a realizar.

6.2. ANÁLISIS EXPLORATORIO DESCRIPTIVO

Para empezar y a través de la herramienta de Excel análisis de datos, se realiza estadística descriptiva de los 33 trabajadores con ausencias del *dataset* con el que finalmente se trabajará (698 casos y 15 atributos). Los datos más relevantes se muestran en las Tablas 6.6 y 6.7:

Tabla 6.6 - Estadística descriptiva del dataset.
Fuente: Elaboración propia a partir del dataset “Abseenteim at work” (Martiniano, et al., 2018)

	Edad	**IMC**	**Hijos**	**Animales**
Media	37,52	25,91	1,18	1,06
Varianza	57,50	19,13	1,09	3,31
Mínimo	27	19,15	0	0
Máximo	58	38,01	4	8

Tabla 6.7 - Porcentaje de bebedores y fumadores.
Fuente: Elaboración propia a partir del dataset “Abseenteim at work” (Martiniano, et al., 2018)

	Bebedor	**Fumador**
Porcentaje	18/33 x 100 = 54,54%	7/33 x 100 = 21,21%

En definitiva, la mayor parte de la plantilla de la empresa tiene una media de edad de 37,52 lo que puede explicar que la media de hijos sea de 1,18. Además, tienen un IMC de 25,91, por lo que se deduce que es una plantilla que tiene sobrepeso. Puede que tenga relación con el desempeño del trabajo por dos razones: la primera que se necesita de gente medianamente fuerte para realizar las entregas de paquetes y, segundo, que posiblemente los trabajadores pasen la mayor parte sentados ya sea conduciendo o en labores de oficina.
Por el contrario, solamente 7 de los trabajadores fuman y un poco más de la mitad bebe ocasionalmente.

A continuación, se realiza un análisis exploratorio gracias a la visualización de datos que ofrece WEKA una vez que se importan los datos al programa. Solamente se exponen los aspectos más llamativos.

La Figura 6.2 muestra el histograma principal, en el que figuran los trabajadores con ausencias en la compañía. El trabajador con ID 3 es el que más ausencias tiene de toda la plantilla con un total de 97 ausencias. También se observa que el trabajador ID 4 no aparece en el listado lo que significa que no ha tenido ausencias durante los tres años. Tampoco los trabajadores con ID 32 y 35 han tenido ausencias (Figura 6.3).

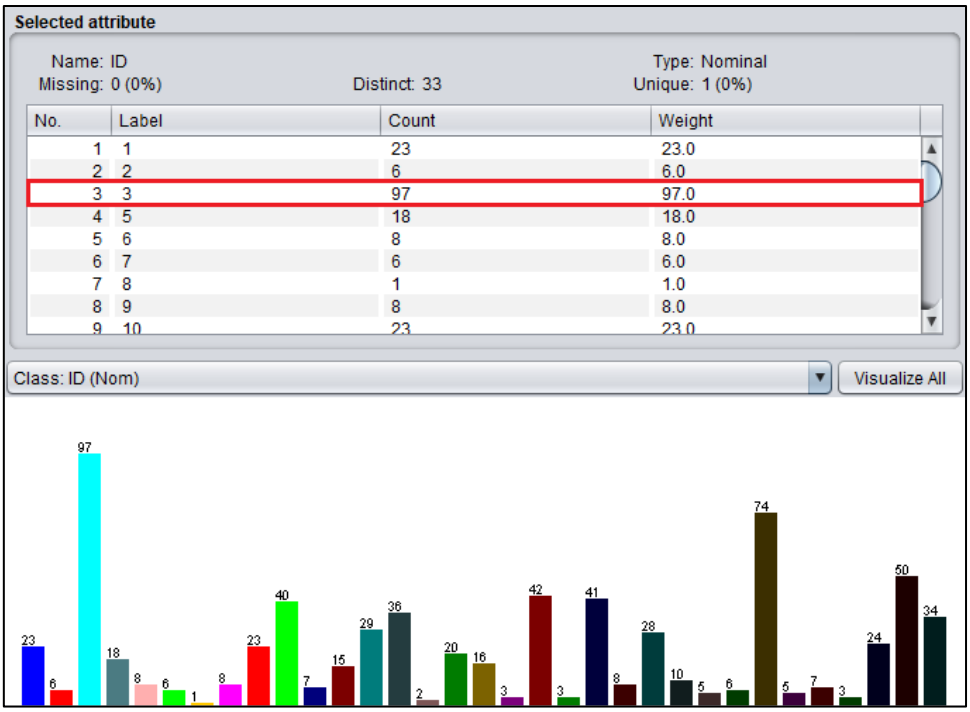


Figura 6.2 – Histograma Atributo 01 ID. Fuente: (WEKA, s.f.)

Otro de los trabajadores con más ausencias (Figura 6.3), es el trabajador ID 28, con un total de 75 ausencias.

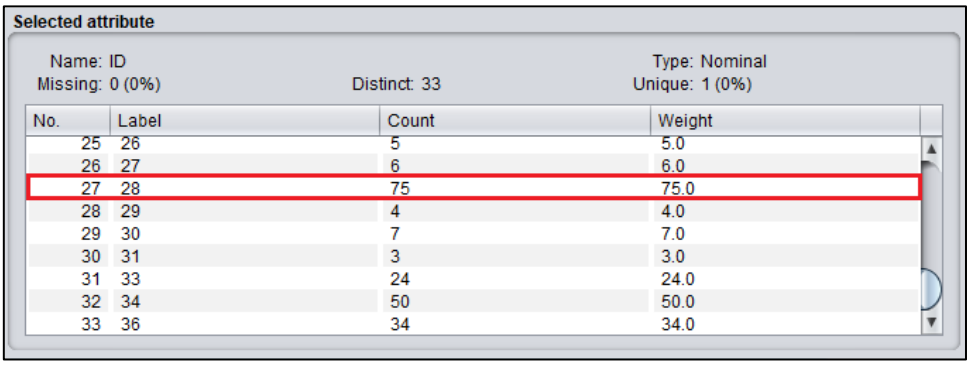


Figura 6.3 - Tabla Atributo ID. Fuente: (WEKA, s.f.)

El siguiente histograma de la Figura 6.4 representa la cantidad de veces por las que un trabajador ha faltado a su puesto de trabajo:

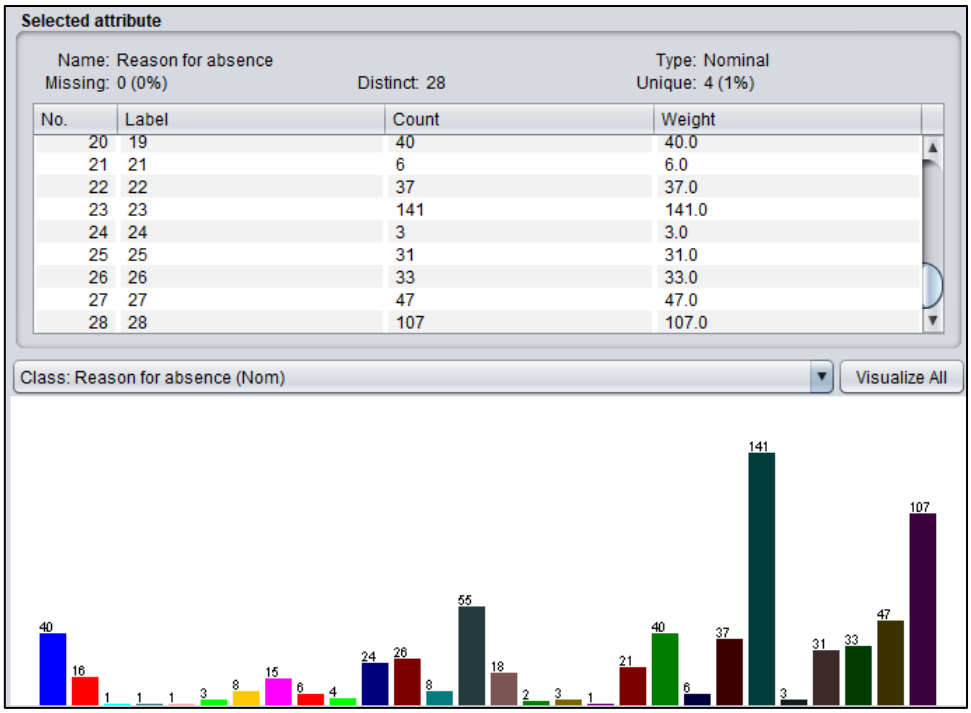


Figura 6.4 – Histograma Atributo Razón de Ausencia. Fuente: (WEKA, s.f.)

La visita al médico es la más numerosa entre todas las causas (141). Otras causas son: consulta dental (107), enfermedades musculoesqueléticas (55), consulta fisioterapéutica (47) y lesiones (40). Excepto la consulta dental, las otras causas pueden tener relación directa con el desempeño físico que se realiza en el trabajo puesto que se trata de una empresa de paquetería. Todas estas ausencias representan un 40,54% del total de ausencias en la empresa.

$$\frac{141 + 55 + 47 + 40}{698} = 40{,}54\%$$

Otra causa importante de ausencia es debido a un fallo disciplinario del trabajador (40) (columna azul Figura 6.4).

La Figura 6.5 es exactamente la misma que la Figura 6.2 con la diferencia de que los colores que se incluyen en cada columna de cada trabajador aparecen los colores que representan cada causa de ausencia de la Figura 6.4 tras seleccionarlo como atributo *class*. Es una de las ventajas de WEKA y su visualización de datos: permite combinar atributos de manera rápida y sencilla. El problema, en este caso, es que hay tantas causas de ausencia que la clasificación por colores no es apropiada para una presentación de los datos de manera ágil y sencilla.

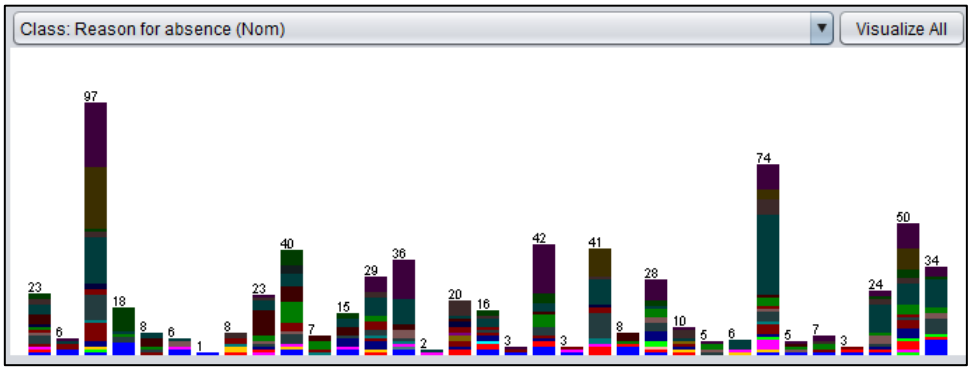


Figura 6.5 - Trabajador (ID) con clasificación Atributo Razón de Ausencia. Fuente: (WEKA, s.f.)

Más o menos se puede apreciar que los trabajadores con ID 5 (18) y con ID 36 (34) son problemáticos, aunque hay 18 trabajadores más que han sido expedientados.

No obstante, hay otros atributos que tienen menos categorías y son más visuales.

La mayoría de los trabajadores solamente tienen estudios primarios tal como se muestra en la Figura 6.6. Ocho empleados tienen estudios universitarios o superiores.

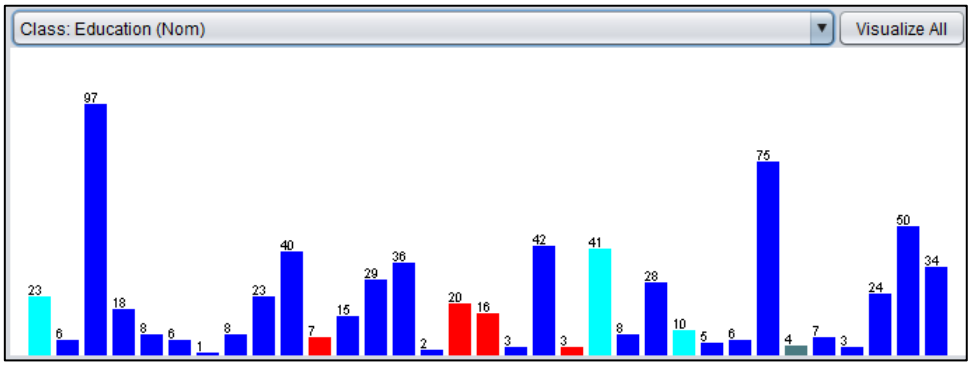


Figura 6.6 - Atributo ID con Clasificación Atributo Educación. Fuente: (WEKA, s.f.)

Solamente se identifican 4 trabajadores que beban y fuman (Figura 6.7 y Figura 6.8): los trabajadores con ID 7 (6), 16 (2), 26 (5) y 30 (7). Curiosamente son empleados que faltan poco al trabajo.

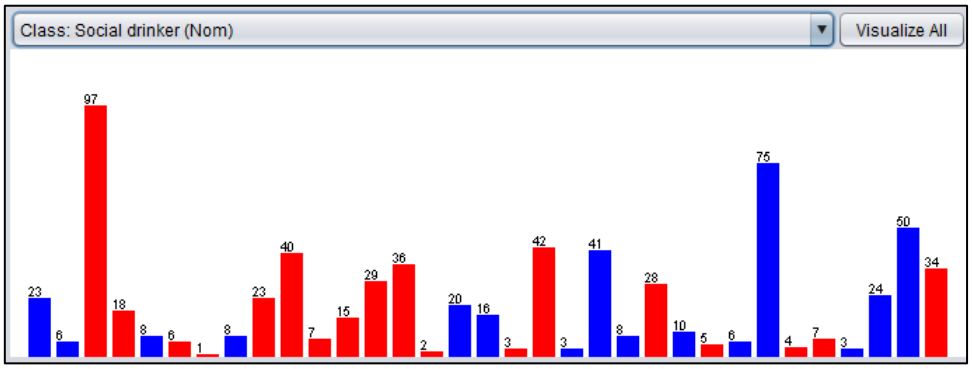


Figura 6.7 - Atributo ID con Clasificación Atributo Bebedor. Fuente: (WEKA, s.f.)

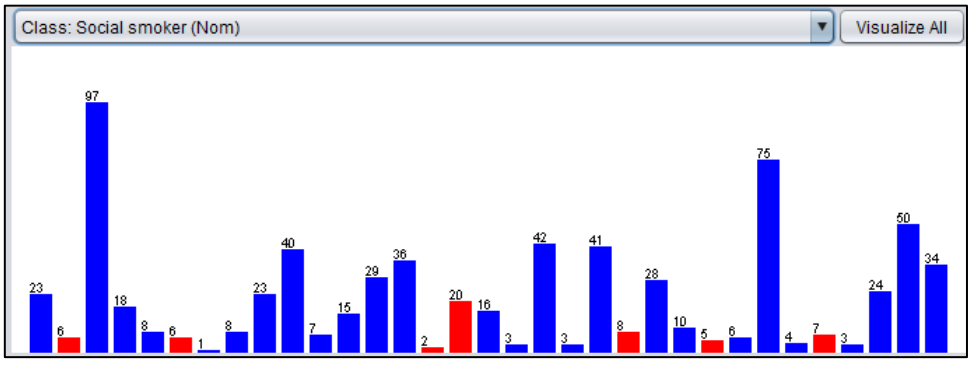


Figura 6.8 - Atributo ID con Clasificación Atributo Fumador. Fuente: (WEKA, s.f.)

Por el contrario, sí que existe una relación entre el sobrepeso y el número de ausencias:

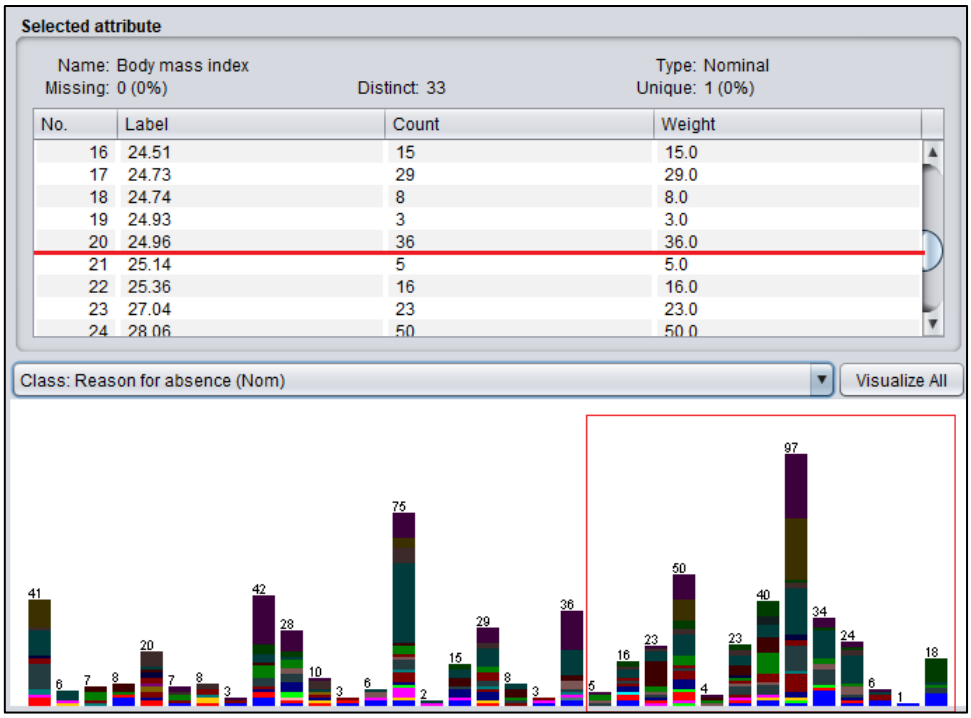


Figura 6.9 - Atributo IMC con Clasificación Atributo Razón de Ausencia. Fuente: (WEKA, s.f.)

En el histograma de la Figura 6.9, que está ordenado de menor a mayor IMC y que cada IMC se corresponde con un trabajador, hay 13 trabajadores con sobrepeso dentro del recuadro rojo, de los 33 trabajadores que han tenido ausencias. A simple vista se comprueba que hay bastantes trabajadores con un IMC adecuado que apenas tienen ausencias mientras que, por otra parte, la mayoría de los que tienen sobrepeso tienen ausencias de manera ocasional.

A continuación, se analizará el tiempo en horas de las ausencias producidas.

Las 41 ausencias reflejadas en primer lugar (Figura 6.10) se deben a fallos disciplinarios y se contabilizan como cero horas. En el recuadro rojo son las horas ausentadas en un día de trabajo, suponiendo que son 8 horas la jornada laboral ya que, a partir de ahí, los datos reflejan múltiplos de 8, como si fuesen días de ausencia. En este grupo, se observa que las principales razones de ausencia son debido a consultas, tanto médicas, dentales y fisioterapéuticas, que tienen los trabajadores. Cuando el trabajador ha faltado el día completo (8 horas), las razones son muy diversas, pero prácticamente todas ellas son por enfermedad.
También se muestra que las ausencias de más de un día son pocas teniendo en cuenta que los datos recogidos son en una línea temporal de 3 años.

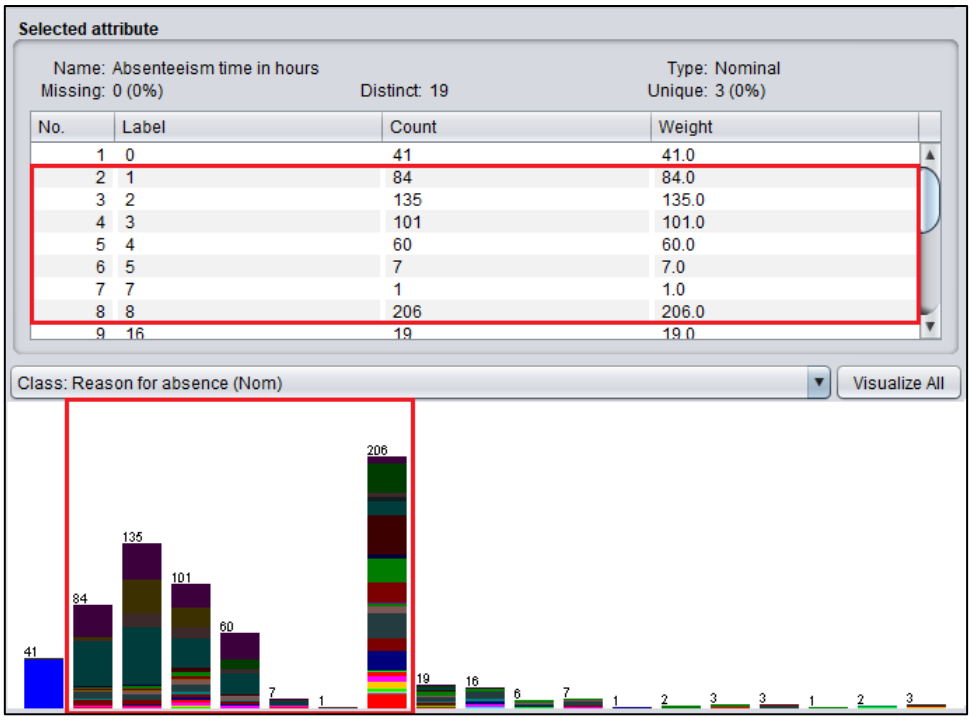


Figura 6.10 - Atributo Absentismo en Horas con Clasificación Atributo Razón de Ausencia. Fuente: (WEKA, s.f.)

En los dos últimos análisis (bebedor/fumador e IMC) se ha analizado si existe relación de las ausencias con aspectos asociados a la salud. En el *dataset* también existen otros atributos que pueden tener relación directa con las ausencias.

En la Figura 6.11 se observa que los trabajadores que entran entre las 0 y las 8 horas apenas se ausentan. En el histograma aparecen las ausencias y no el número de trabajadores, es decir, que si existen más ausencias en el horario de mañana que en el de noche posiblemente sea porque hay más gente trabajando a esas horas. Si se selecciona el atributo ID como atributo *class* así lo confirma (cada ID viene representado por un color), pero el número de ausencias de la mayoría de los trabajadores de este grupo es parecida o mayor que 23, es decir, mayor que el que más se ausenta en horario de noche (se puede apreciar por la altura de cada color), aunque hay una explicación. Como se vio anteriormente, las razones de ausencia más comunes son las consultas. Por tanto, se puede deducir que estas consultas sean en horario de mañana y el motivo por el que la gente de este horario falta más al trabajo.

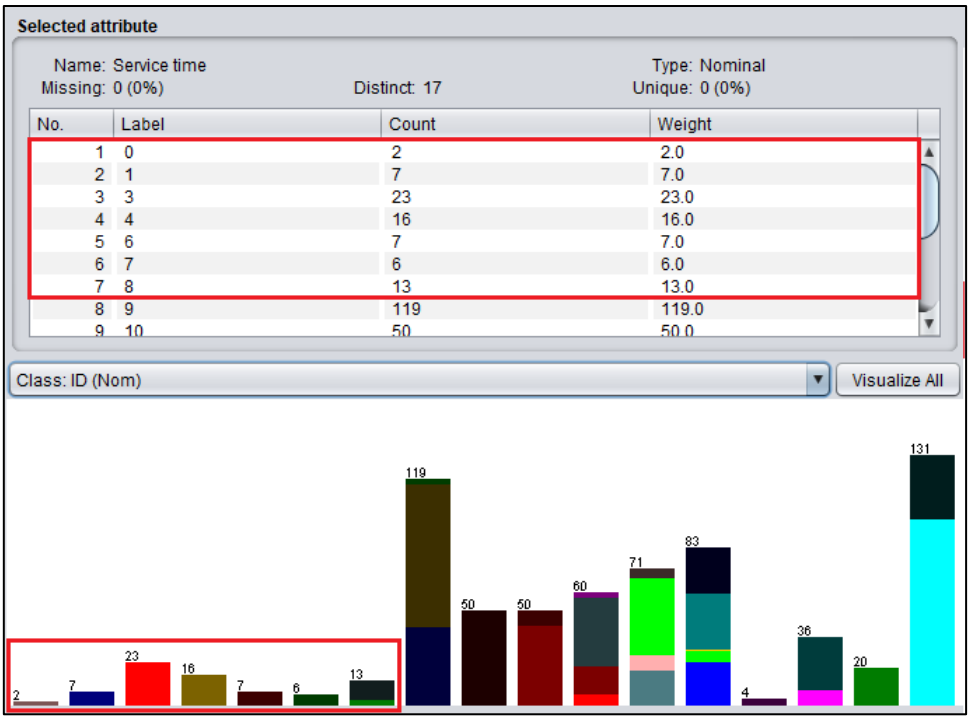


Figura 6.11 - Atributo Absentismo en Horas con Clasificación Atributo ID. Fuente: (WEKA, s.f.)

A continuación, se trabajará con el atributo edad. Para ello, se usará una técnica de la minería de datos dentro del preprocesamiento denominada “discretización” que WEKA tiene implementada en el apartado de filtros. Por discretizar se entiende a la transformación de un atributo numérico en grupos nominales establecidos por un rango de valores determinados. Como son muchos los valores de edades y en ocasiones repetidos para varios ID de trabajadores, se procede a discretizar este atributo.

La discretización creó tres rangos (Figura 6.12): desde menos infinito hasta 37,3 años; desde 37,3 años hasta 47,67 y desde 47,67 hasta infinito. En realidad, los valores de menos infinito a más infinito se pueden sustituir por los valores 27 y 58, respectivamente, que son los valores mínimo y máximo de edad de la Tabla 6.6. WEKA trata de realizar los grupos lo más homogéneos posibles, como se aprecia en este caso cuyo grupo abarca un rango de 10,33 años.

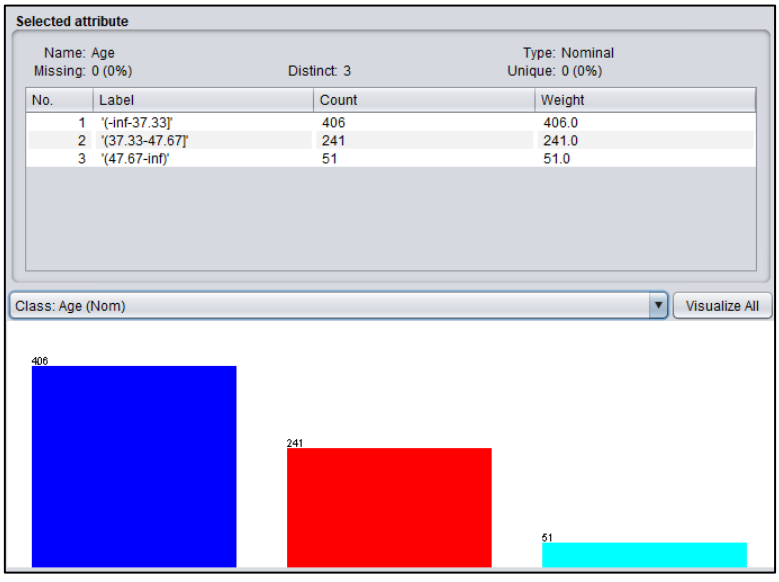


Figura 6.12 - Discretización Atributo Edad. Fuente: (WEKA, s.f.)

En la Figura 6.13 se observa que un número elevado de trabajadores pertenecen al grupo de los jóvenes, lo que justifica que la media de edad sea de 37,52. También se puede decir que son el grupo de edad que más se ausenta. Hasta 8 trabajadores se han ausentado más de 20 veces. Por el contrario, los más veteranos son los que menos ausencias tienen, exceptuando el último trabajador.

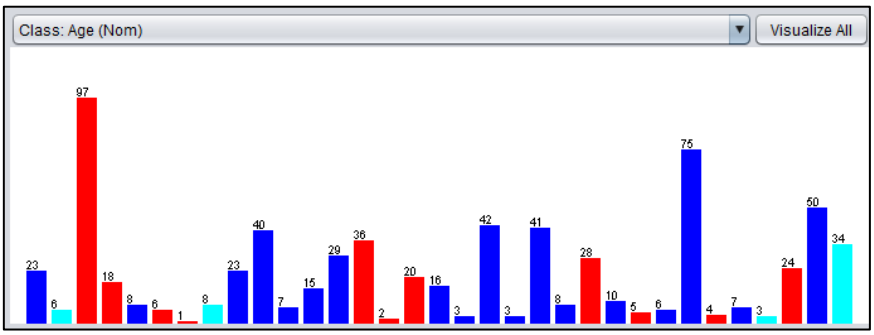


Figura 6.13 - Atributo ID con Clasificación Atributo Edad. Fuente: (WEKA, s.f.)

Uno de los motivos por los que los jóvenes se ausenten puede ser por la tenencia de hijos (Figura 6.14). Excepto los ID 22 (41) y 34 (50), los demás trabajadores tienen al menos un hijo.En aquellos casos que no tienen hijos existen diferentes situaciones: trabajadores con muchas ausencias y trabajadores con pocas ausencias.

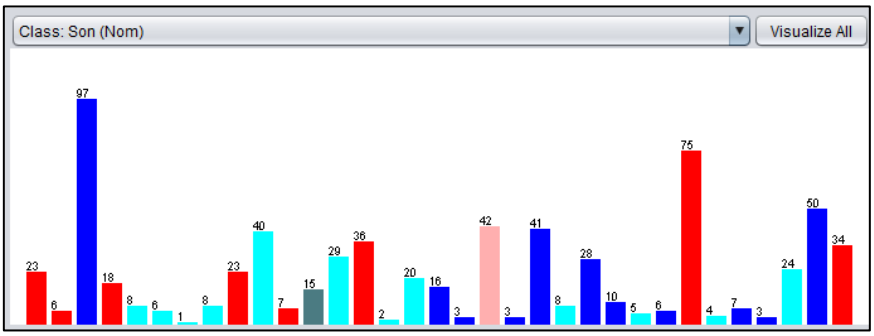


Figura 6.14 - Atributo ID con Clasificación Atributo Hijos. Fuente: (WEKA, s.f.)

Por último, cabe mencionar si las ausencias están relacionadas de alguna forma con los días o épocas del año. En la Figura 6.15 se aprecia que los meses de junio y julio son los que menos ausencias tienen, seguramente porque son meses de verano, aquellos donde los empleados se suelen ir de vacaciones. En septiembre puede haber más ausencias debido al estrés post-vacacional.

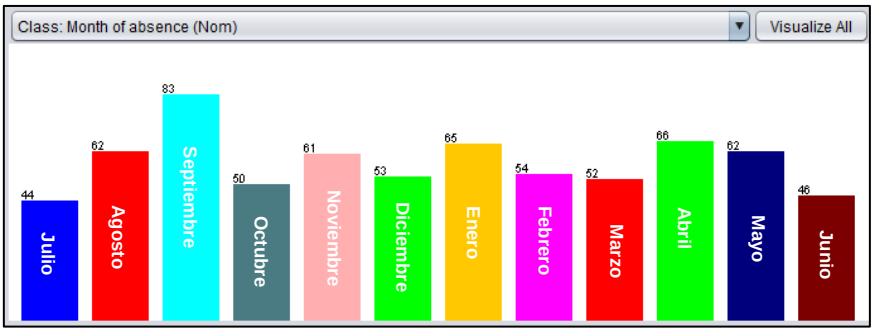


Figura 6.15 - Atributo Mes de Ausencia. Fuente: (WEKA, s.f.)

En cuanto a los días de la semana (Figura 6.16), las ausencias se reparten entre todos los días excepto el jueves donde hay un ligero descenso.

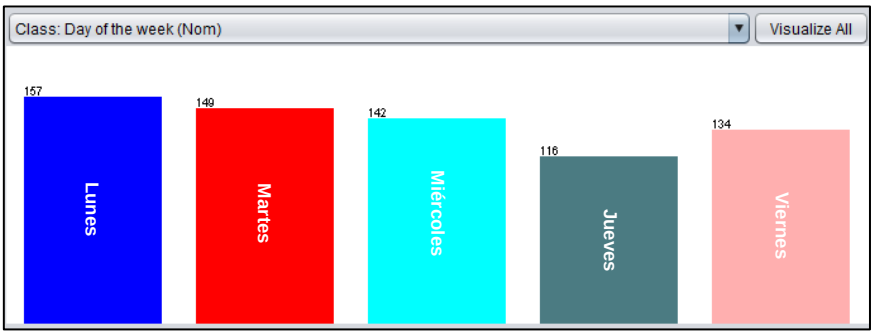


Figura 6.16 - Atributo Día de la Semana de Ausencia. Fuente: (WEKA, s.f.)

6.3. APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS

Hasta el momento se ha usado la visualización de datos para realizar un análisis preliminar. Ahora es turno de poner a funcionar las técnicas propias de la minería de datos que tiene implementada WEKA y sus algoritmos complejos con los datos importados al sistema.

6.3.1. Técnicas de predicción mediante clasificación

Los modelos de predicción mediante clasificación requieren obligatoriamente de atributos nominales, es decir, atributos categóricos. Estos modelos son capaces de discretizar por sí solos aquellos atributos numéricos con los que se trabaje. De hecho, la edad y el IMC se consideran atributos fundamentales para los análisis por lo que se incluirán en las diferentes técnicas.
Hay que tener en cuenta que no son necesarios todos los atributos para aplicar las técnicas. Por ejemplo, a nadie le interesa que una técnica prediga el número ID del trabajador o el día de la semana, por lo que solamente se contará con aquellos realmente necesarios.

6.3.1.1. Aplicación del modelo Naive Bayes

El modelo *Naive Bayes* se basa en los siguientes aspectos:

- Los atributos tienen la misma importancia.
- Los atributos son estadísticamente independientes.

Por tanto, el valor de uno de los atributos no dice nada del valor del otro.
El supuesto de independencia nunca es correcto porque en la práctica no suele ser así, aunque suele dar buenos resultados (WEKA, s.f.).

En primer lugar, se seleccionará como atributo principal la educación. Lo primero que hay que saber al construir un modelo es si es bueno para la predicción realizada. Se puede consultar por las clasificaciones correctas que ha realizado. En este caso, ha dado un resultado del 83,09% (Figura 6.17), por lo que se puede asegurar que la predicción es bastante buena.

```
Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         580               83.0946 %
Incorrectly Classified Instances       118               16.9054 %
Kappa statistic                          0.5208
Mean absolute error                      0.0936
Root mean squared error                  0.2434
Relative absolute error                 60.2235 %
Root relative squared error             87.6755 %
Total Number of Instances              698

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.875    0.226    0.947      0.875    0.909      0.580    0.910     0.969     1
               0.543    0.020    0.658      0.543    0.595      0.573    0.877     0.514     2
               0.662    0.123    0.389      0.662    0.490      0.431    0.912     0.731     3
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     4
Weighted Avg.  0.831    0.200    0.869      0.831    0.845      0.566    0.909     0.914

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 502  13  59   0 |   a = 1
   3  25  18   0 |   b = 2
  25   0  49   0 |   c = 3
   0   0   0   4 |   d = 4
```

Figura 6.17 - Resumen Modelo Naive Bayes. Fuente: (WEKA, s.f.)

En la Figura 6.18 aparece una predicción del número de ausencias clasificados por motivo basada en el nivel de educación del trabajador. Los números que aparecen entre paréntesis en el encabezado del atributo es el peso que tiene cada grupo de educación. Simplemente se basa en datos históricos, en los datos que se ha importado al sistema.
A simple vista, se puede resaltar que los trabajadores con estudios primarios faltarían más al trabajo por fallo disciplinario con respecto a personas con una educación superior. Las consultas seguirían siendo las ausencias más destacadas. El aspecto más llamativo son las causas más habituales por las que faltarían los trabajadores con posgrado, por lesiones físicas. Estas personas suelen ser trabajadores de oficina. Habría que estudiar la situación.

	Class			
Attribute	1	2	3	4
	(0.82)	(0.07)	(0.11)	(0.01)
Reason for absence				
0	37.0	4.0	2.0	1.0
1	10.0	3.0	6.0	1.0
2	1.0	2.0	1.0	1.0
3	2.0	1.0	1.0	1.0
4	2.0	1.0	1.0	1.0
5	4.0	1.0	1.0	1.0
6	8.0	1.0	2.0	1.0
7	13.0	2.0	3.0	1.0
8	5.0	3.0	1.0	1.0
9	5.0	1.0	1.0	1.0
10	21.0	4.0	2.0	1.0
11	21.0	5.0	3.0	1.0
12	6.0	2.0	3.0	1.0
13	42.0	2.0	14.0	1.0
14	17.0	1.0	2.0	2.0
15	3.0	1.0	1.0	1.0
16	1.0	3.0	2.0	1.0
17	1.0	2.0	1.0	1.0
18	15.0	5.0	4.0	1.0
19	36.0	4.0	2.0	2.0
21	3.0	3.0	3.0	1.0
22	30.0	4.0	5.0	2.0
23	123.0	5.0	16.0	1.0
24	4.0	1.0	1.0	1.0
25	19.0	8.0	7.0	1.0
26	30.0	3.0	3.0	1.0
27	37.0	1.0	12.0	1.0
28	106.0	1.0	2.0	2.0
[total]	602.0	74.0	102.0	32.0

Figura 6.18 - Predicción Educación basada en el Modelo Naive Bayes (1). Fuente: (WEKA, s.f.)

La Figura 6.19 muestra los resultados de los demás atributos según el nivel de educación. Se pueden deducir las siguientes conclusiones:

- Los trabajadores con posgrado no tienen hijos, posiblemente porque son los más jóvenes.
- Hay más trabajadores con sobrepeso en los grupos de educación básica y de máster/doctor, posiblemente porque beben ocasionalmente.
- El grupo doctorado tiene más hijos y animales que el resto, posiblemente por su nivel socioeconómico.

Age				
mean	36.9976	34.2277	32.345	40.7895
std. dev.	6.825	5.9267	3.642	0.2719
weight sum	574	46	74	4
precision	1.6316	1.6316	1.6316	1.6316
Son				
mean	1.162	1.0217	0.3108	2
std. dev.	1.1415	0.9205	0.4628	0.1667
weight sum	574	46	74	4
precision	1	1	1	1
Social drinker				
0	189.0	40.0	75.0	1.0
1	387.0	8.0	1.0	5.0
[total]	576.0	48.0	76.0	6.0
Social smoker				
0	541.0	27.0	75.0	5.0
1	35.0	21.0	1.0	1.0
[total]	576.0	48.0	76.0	6.0
Pet				
mean	0.7721	1.2174	0.4973	1.6
std. dev.	1.0374	2.8735	0.7405	0.2667
weight sum	574	46	74	4
precision	1.6	1.6	1.6	1.6
Body mass index				
mean	27.2427	23.1394	22.7945	28.29
std. dev.	3.8401	1.787	4.7439	0.0982
weight sum	574	46	74	4
precision	0.5894	0.5894	0.5894	0.5894
Absenteeism time in hours				
mean	6.4808	5.0725	4.1441	3.3333
std. dev.	15.0969	7.1133	8.7476	3.3333
weight sum	574	46	74	4
precision	6.6667	6.6667	6.6667	6.6667

Figura 6.19 - Predicción Educación basada en el Modelo Naive Bayes (2). Fuente: (WEKA, s.f.)

6.3.1.2. Modelo de clasificación: Árbol de decisiones J48

El modelo J48 se basa en un algoritmo denominado C4.5 que genera un árbol que va adoptando diferentes caminos (ramas) según las posibilidades que existen. Se elegirá la educación como atributo principal, como en el método anterior. También se eliminará el atributo animal ya que no interesa para este análisis.

Como se ha mencionado anteriormente, lo primero que hay que observar es la precisión del modelo de la Figura 6.20. Prácticamente se puede decir que el modelo es perfecto (99,14%) pero hay un pequeño inconveniente: el tamaño del árbol.

```
Number of Leaves  :     15

Size of the tree :      29


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         692               99.1404 %
Incorrectly Classified Instances         6                0.8596 %
Kappa statistic                          0.9715
Mean absolute error                      0.0053
Root mean squared error                  0.0649
Relative absolute error                  3.3889 %
Root relative squared error             23.385  %
Total Number of Instances              698

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               1.000    0.048    0.990      1.000    0.995      0.970    0.995     0.999     1
               0.957    0.000    1.000      0.957    0.978      0.977    1.000     0.994     2
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     3
               0.000    0.000    ?          0.000    ?          ?        0.884     0.024     4
Weighted Avg.  0.991    0.040    ?          0.991    ?          ?        0.995     0.993

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 574   0   0   0 |   a = 1
   2  44   0   0 |   b = 2
   0   0  74   0 |   c = 3
   4   0   0   0 |   d = 4
```

Figura 6.20 - Resumen Modelo Árbol J48. Fuente: (WEKA, s.f.)

En la Figura 6.22 se puede ver cómo queda ese árbol:

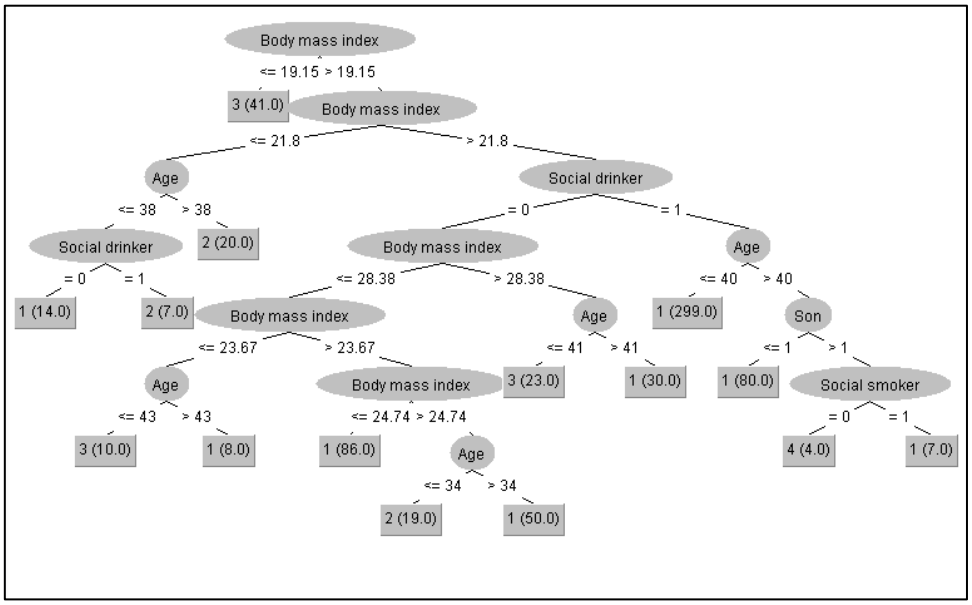


Figura 6.21 - Árbol J48. Fuente: (WEKA, s.f.)

Otra opción posible es aumentar el mínimo de casos para que se cree una rama (por defecto este valor es 2). De esta manera lo que se consigue es un árbol más pequeño, pero más visual a cambio de un peor modelo.

En la Figura 6.22, se aumentó el mínimo de casos a 20 y el modelo es preciso al 93,26%. La lectura de este árbol es bastante sencilla. Se va leyendo de arriba abajo y de izquierda a derecha. En este ejemplo, se indica qué tipo de educación tendrá el trabajador conforme a los valores de los otros atributos. De entrada, indica que, si se

contratara a una persona con un IMC menor o igual que 19,15, seguramente sea una persona con posgrado. Si esta persona está entre 19,15 y 21,8 de IMC y tiene una edad menor o igual que 38, se hablaría de una persona solamente con estudios primarios. Si fuera mayor, sería universitario. Y la misma lectura seguiría para el resto del árbol.

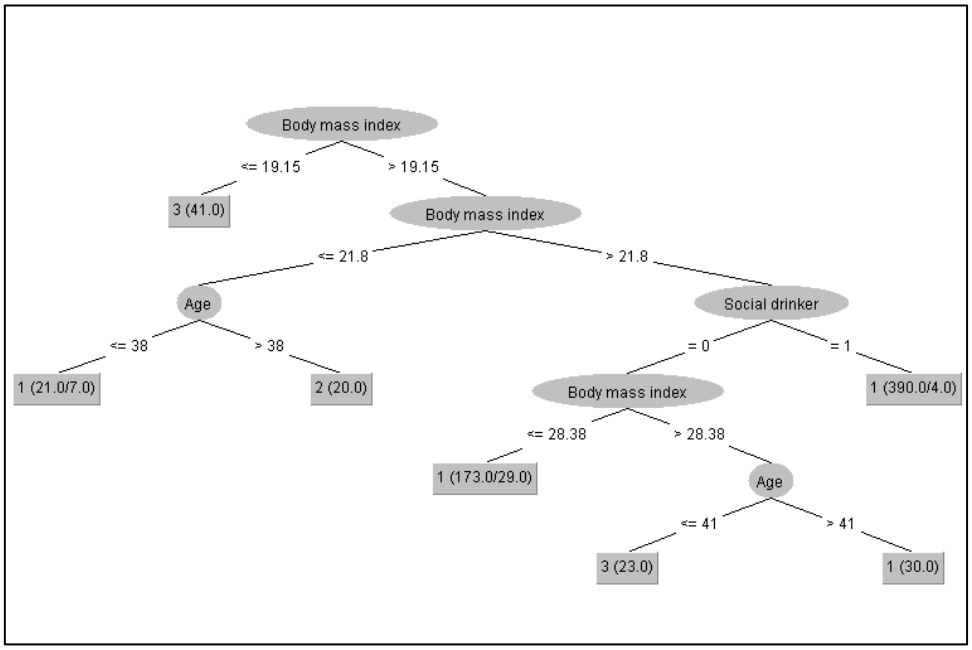


Figura 6.22 - Árbol J48 Alternativo 1. Fuente: (WEKA, s.f.)

Este modelo se puede llevar al extremo. Básicamente la Figura 6.23 muestra que, con una certeza del 86,10%, si el trabajador tiene un IMC menor o igual a 21,8, es posgraduado y, al mismo tiempo, si es mayor a 21,8 el trabajador solamente tendría estudios primarios.

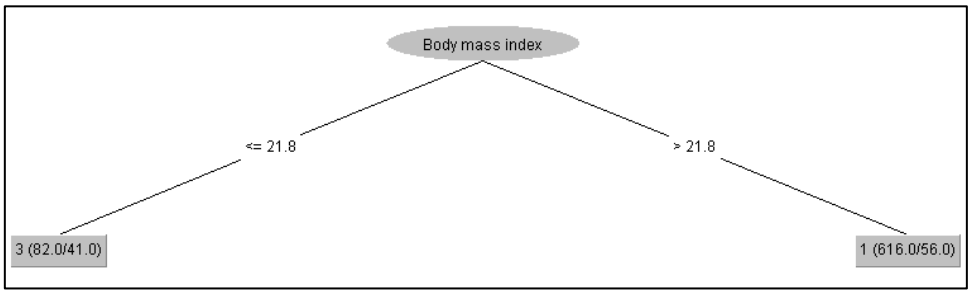


Figura 6.23 - Árbol J48 Alternativo 2. Fuente: (WEKA, s.f.)

6.3.2. Clustering

Por *Clustering* se entiende al proceso que genera "*cluster"* o grupos. Un "*cluster*" es un grupo de individuos que tienen características similares. En este proceso no existirá un atributo principal. De este modo, si es capaz de reproducir los grupos existentes en el *dataset* como, por ejemplo, los trabajadores según educación, el *clustering* resultará efectivo. Aunque hay diferentes técnicas de *clustering* que trabajan con diferentes algoritmos, solamente se aplicará el algoritmo *KMeans*.

Esta técnica de *clustering* se basa en distancias. Consiste en decir al algoritmo el número de *clusters* que se desea. En este ejemplo el número de grupos a crear es 4 para ver si es capaz de representar, los grupos de educación. Después escoge cuatro "puntos" o "ubicaciones" al azar que serán los centros de cada *cluster*. El objetivo será

escoger los casos más cercanos puesto que son los que comparten más características con ese centro.

Lo primero que salta a la vista (Figura 6.24) es que no ha sido capaz de reproducir un nivel de educación asociado a cada *cluster* por los valores que refleja el nivel de educación. Fundamentalmente es porque existen muchas ausencias de trabajadores que solamente tiene estudios primarios. Aun así, el nivel del modelo se puede medir por el error suma cuadrados que aparece en la parte superior de la Figura 6.24, el cual es muy elevado, es decir, que hay mucha "distancia entre los puntos y los centros de cada *cluster*", lo que indica que el *clustering* no ha tenido éxito.

```
Number of iterations: 5
Within cluster sum of squared errors: 739.4743845420574

Initial starting points (random):

Cluster 0: 25,40,2,2,0,1,21.8,8
Cluster 1: 19,33,1,2,1,0,30.42,32
Cluster 2: 11,32,1,0,1,0,22.76,4
Cluster 3: 0,47,1,2,0,0,31.59,0

Missing values globally replaced with mean/mode

Final cluster centroids:
                                Cluster#
Attribute                  Full Data          0          1          2          3
                             (698.0)     (29.0)    (136.0)    (271.0)    (262.0)
=============================================================================
Reason for absence               23         25         19         28         23
Age                          36.3567    38.4828    34.7132      39.31    33.9198
Education                        1          2          1          1          1
Son                          1.0673      1.931     2.2794      0.738     0.6832
Social drinker                   1          0          1          1          0
Social smoker                    0          1          0          0          0
Body mass index              26.5584    21.8317    26.1989    28.5603    25.1974
Absenteeism time in hours      7.202          6    12.1765     5.8155      6.187
```

Figura 6.25 - Resultados KMeans Clustering. Fuente: (WEKA, s.f.)

6.4. CONCLUSIONES

A la vista de los resultados surgidos tanto en el análisis exploratorio como en la aplicación de técnicas, se extraen las siguientes conclusiones de este *dataset*:

- La plantilla está comprometida con el trabajo debido a que no ha habido muchas ausencias de larga duración durante los tres años.
- La plantilla tiende a tener sobrepeso, provocando lesiones y enfermedades relacionadas con el desempeño laboral.
- No son elevados los casos de fallo disciplinario, pero sí están repartidos entre varios trabajadores, incluso de diferentes niveles educacionales.

En base a las siguientes conclusiones, la empresa o bien el equipo de recursos humanos puede adoptar las siguientes decisiones:

- Incentivar de manera global o individualizada el poco absentismo producido.

- Si se desea reducir aún más el número de ausencias, se puede plantear introducir una enfermería o un doctor dentro de la empresa. En el peor de los casos integrar cursos de ergonomía.
- Celebrar competiciones o eventos deportivos entre los empleados de la empresa, así como jornadas de comida saludable con el objetivo de adelgazar a los empleados y no de lesionarlos.
- También se puede revisar la normativa de la empresa para reducir los casos por fallo disciplinario. En su defecto, dar pequeñas charlas sobre cómo se debe trabajar.

En un artículo sobre absentismo laboral por enfermedad en trabajadores sedentarios elaborado por (López Bueno, et al., 2018) "se han encontrado estudios con riesgos relativos que apuntan a una probabilidad cuatro veces mayor de ausentarse del trabajo por motivos de enfermedad cuando se comparaban practicantes de actividades físicas con no practicantes y una diferencia de 5 días anuales entre grupos de intervención y control a favor de un menor absentismo de los primeros."

Las conclusiones se han elaborado basándose en la información obtenida gracias a la visualización y minería de datos. No obstante, determinados algoritmos no dan los resultados esperados, como se ha comprobado con el algoritmo *KMeans*, por lo que no todos los algoritmos o técnicas de minería de datos se pueden usar en cualquier *dataset*.

7. VALORACIÓN FINAL

El mundo del *Big Data* y *Business Intelligence* puede ser muy amplio y muy complejo, pero está al alcance de cualquiera. No se necesita de conocimientos muy técnicos ni de recursos muy grandes para ponerlo en marcha. Si bien es cierto que este ejemplo se ha reducido simplemente a las ausencias del personal, la realidad es que se puede extrapolar a cualquier ámbito de una empresa y prácticamente en cualquier magnitud.

Tal y como se ha citado anteriormente, aquellas empresas que no hagan uso del BI corren el riesgo de extinguirse. No se está hablando de una mejora o de un extra a la empresa sino de una necesidad. Las empresas deberán valorar en qué medida deben invertir sus recursos para implantar el BI no solamente como parte de su organización sino también de su cultura.

Este trabajo es la demostración de que con poco se puede conseguir muchas cosas. Se ha conseguido elaborar y obtener una información a partir de simples datos numéricos que pueden mejorar los resultados de una empresa si se adoptan las decisiones adecuadas.

8. BIBLIOGRAFÍA

Adair, B., 2018. *SelectHub.* [En línea]
Disponible en: https://selecthub.com/business-intelligence/business-intelligence-vs-business-analytics/
[Último acceso: 08 Junio 2019].

AlphaGo. 2017. [Película] Dirigido por Greg Kohs. Estados Unidos: Netflix.

Baker, P., 2019. *PCMag.* [En línea]
Disponible en: https://www.pcmag.com/roundup/338081/the-best-self-service-business-intelligence-bi-tools
[Último acceso: 15 Junio 2019].

Betfy, s.f. *Betfy.* [En línea]
Disponible en: https://www.betfy.co.uk/internet-realtime/#
[Último acceso: 11 Julio 2019].

Blitz, S., 2018. *Sisense.* [En línea]
Disponible en: https://www.sisense.com/blog/dashboards-vs-reports-need/
[Último acceso: 23 Junio 2019].

Central Intelligence Agency, 2015. *Central Intelligence Agency.* [En línea]
Disponible en: https://www.cia.gov/news-information/featured-story-archive/2015-featured-story-archive/the-enigma-of-alan-turing.html
[Último acceso: 06 Junio 2019].

ChinaTimes, 2017. *ChinaTimes.* [En línea]
Disponible en: https://www.chinatimes.com/newspapers/20170320000646-260309?chdtv
[Último acceso: 26 Junio 2019].

Columbus, L., 2018. *Forbes.* [En línea]
Disponible en: https://www.forbes.com/sites/louiscolumbus/2018/05/23/10-charts-that-will-change-your-perspective-of-big-datas-growth/#1d407e2c2926
[Último acceso: 22 Julio 2019].

DataLabs, s.f. *DataLabs Agency.* [En línea]
Disponible en: http://www.datalabsagency.com/tableau-business-intelligence-dashboard-designer/
[Último acceso: 23 Junio 2019].

Dontha, R., 2018. *Digital Transformation Pro.* [En línea]
Disponible en: https://digitaltransformationpro.com/data-mining-steps/
[Último acceso: 7 Agosto 2019].

Eckerson Group, 2016. *TDWI.* [En línea]
Disponible en: https://tdwi.org/Articles/2016/09/08/Embedded-Analytics-Future-of-BI.aspx?Page=1
[Último acceso: 2019 Junio 06].

Gourévitch, A., Faeste, L., Baltassis, E. & Marx, J., 2017. *Boston Consulting Group.* [En línea]
Disponible en: https://www.bcg.com/publications/2017/digital-transformation-transformation-data-driven-transformation.aspx
[Último acceso: 22 Julio 2019].

Gratton, S., 2012. *Capgemini.* [En línea]
Disponible en: https://www.capgemini.com/2012/07/bi-30-the-journey-to-business-intelligence-what-does-it-mean/
[Último acceso: 15 Junio 2019].

Hassabis, D. & Silver, D., 2017. *DeepMind.* [En línea]
Disponible en: https://deepmind.com/blog/alphago-zero-learning-scratch/
[Último acceso: 28 Junio 2019].

InfoJobs, 2018. *InfoJobs.* [En línea]
Disponible en: https://orientacion-laboral.infojobs.net/los-puestos-emergentes-mas-demandados
[Último acceso: 29 Junio 2019].

Jody, 2009. *Marketing Jive.* [En línea]
Disponible en: https://www.marketing-jive.com/2009/07/difference-between-dashboard-and-report.html
[Último acceso: 23 Junio 2019].

Larion, s.f. *Larion.* [En línea]
Disponible en: https://larion.com/category/resources/trends/bi/
[Último acceso: 16 Junio 2019].

López Bueno, R., Casajús Mallén, J. A. & Garatachea Vallejo, N., 2018. *Ministerio de Sanidad, Consumo y Bienestar Social.* [En línea]
Disponible en: https://www.mscbs.gob.es/biblioPublic/publicaciones/recursos_propios/resp/revista_cdrom/VOL92/REVISIONES/RS92C_201810071.pdf
[Último acceso: 07 Septiembre 2019].

Marr, B., 2016. *Big Data in Practice.* First ed. s.l.:Wiley.

Martiniano, A., Ferreira, R. P. & Sassi, R. J., 2018. *University of California, Irvine.* [En línea]
Disponible en: https://archive.ics.uci.edu/ml/datasets/Absenteeism+at+work
[Último acceso: 05 Julio 2019].

McCluney, A., 2017. *BrightGauge.* [En línea]
Disponible en: https://blog.brightgauge.com/dashboards-vs-reports
[Último acceso: 23 Junio 2019].

Microsoft, 2018. *Power BI.* [En línea]
Disponible en: https://docs.microsoft.com/en-us/power-bi/consumer/mobile/mobile-apps-view-dashboard
[Último acceso: 23 Junio 2019].

Microsoft, s.f. *Microsoft Office.* [En línea]
Disponible en: https://support.office.com/en-us/article/overview-of-online-analytical-processing-olap-15d2cdde-f70b-4277-b009-ed732b75fdd6
[Último acceso: 24 Junio 2019].

Min. Ciencia, Innovación y Universidades, 2018. *Min. Ciencia, Innovación y Universidades.* [En línea]
Disponible
en: https://www.educacion.gob.es/ruct/consultaestudios.action?actual=estudios
[Último acceso: 2019 Junio 30].

Moss, L. T. & Atre, S., 2003. *Business Intelligence Roadmap.* Primera ed. s.l.:Addison Wesley.

National Health Service, 2018. *National Health Service.* [En línea]
Disponible en: https://www.nhs.uk/live-well/healthy-weight/height-weight-chart/
[Último acceso: 17 Agosto 2019].

OLAP, s.f. *OLAP.com.* [En línea]
Disponible en: http://olap.com/olap-definition/
[Último acceso: 24 Junio 2019].

Pinterest, s.f. *Pinterest.* [En línea]
Disponible en: https://www.pinterest.com/pin/375487687661664112/
[Último acceso: 26 Junio 2019].

Plowden, N., 2018. *International Business Machines Corporation.* [En línea]
Disponible en: https://www.ibm.com/blogs/business-analytics/what-is-business-intelligence/
[Último acceso: 08 Junio 2019].

Pratt, M. K., 2017. *Chieff Information Officer.* [En línea]
Disponible en: https://www.cio.com/article/2439504/business-intelligence-definition-and-solutions.html
[Último acceso: 08 Junio 2019].

Predictive Analytics Today, s.f. *Predictive Analytics Today.* [En línea]
Disponible en: https://www.predictiveanalyticstoday.com/top-business-intelligence-companies/
[Último acceso: 13 Junio 2019].

Sanders, D., 2013. *HealthCatalyst.* [En línea]
Disponible en: https://www.healthcatalyst.com/wal-mart-birth-of-data-warehouse/
[Último acceso: 2019 Junio 16].

SAS, s.f. *SAS.* [En línea]
Disponible en: https://www.sas.com/en_us/insights/data-management/what-is-etl.html
[Último acceso: 16 Junio 2019].

SAS, s.f. *SAS.* [En línea]
Disponible en: https://www.sas.com/en_us/insights/analytics/data-mining.html
[Último acceso: 25 Junio 2019].

Savkin, A., s.f. *BSC Designer.* [En línea]
Disponible en: https://bscdesigner.com/dashboard-vs-balanced-scorecard.htm
[Último acceso: 23 Julio 2019].

Sisense, s.f. *Sisense.* [En línea]
Disponible en: https://www.sisense.com/glossary/data-mart/
[Último acceso: 16 Juno 2019].

Universidad Carlos III de Madrid, 2019. *Universidad Carlos III de Madrid.* [En línea]
Disponible
en: http://www.uc3m.eu/ss/Satellite/Grado/es/Detalle/Estudio_C/1371241688824/1371212987094/Grado_en_Ciencia_e_Ingenieria_de_Datos#profesoradoyplandocente
[Último acceso: 30 Junio 2019].

University of California Irvine Machine Learning Repository, s.f. *UCI Machine Learning Repository.* [En línea]
Disponible en: http://archive.ics.uci.edu/
[Último acceso: 15 Julio 2019].

Visually, 2013. *Visually.* [En línea]
Disponible en: https://visual.ly/community/infographic/travel/transportation-history-%E2%80%93-evolution-travel
[Último acceso: 06 Junio 2019].

Vlamis, D., 2005. *Vlamis Software Solutions.* [En línea]
Disponible en: http://vlamiscdn.com/papers/oow2005-presentation2.pdf
[Último acceso: 24 Junio 2019].

WEKA, s.f. *University of Waikato.* [En línea]
Disponible en: https://www.cs.waikato.ac.nz/ml/weka/
[Último acceso: 15 Julio 2019].

WEKA, s.f. *Youtube.* [En línea]
Disponible en: https://www.youtube.com/user/WekaMOOC
[Último acceso: 15 Julio 2019].

Witten, I. H., Frank, E. & Hall, M. A., 2016. *University of Waikato.* Cuarta ed. s.l.:Morgan Kaufmann.

Zelman, K. M., 2007. *WebMD.* [En línea]
Disponible en: https://www.webmd.com/diet/features/how-accurate-body-mass-index-bmi#1
[Último acceso: 17 Agosto 2019].