

UNIDAD 2: INTRODUCCIÓN AL ANÁLISIS DE GRANDES DATOS (BIG DATA) Y A LA ANALÍTICA AVANZADA

Módulo profesional:

Análisis de grandes datos (Big Data)

ÍNDICE

RESUMEN INTRODUCTORIO.....	6
INTRODUCCIÓN.....	6
CASO INTRODUCTORIO	7
1. BIG DATA ANALYTICS	8
1.1 La ciencia de datos.....	9
1.1.1 <i>El dato</i>	10
1.1.2 <i>El Big Data</i>	10
1.2 El flujo para la ciencia de datos	12
2. HERRAMIENTAS FUNDAMENTALES DEL BIG DATA ANALYTICS	13
2.1 La ciencia de datos en el mundo real.....	14
2.2 La ética, el big data y la ciencia de datos	15
2.3 Huella digital	15
2.4 Business intelligence.....	16
2.5 Minería de datos	17
3. FUTURO DEL BIG DATA	18
3.1 Machine learning.....	19
3.1.1 <i>El machine learning</i>	22
3.1.2 <i>Clasificación de algoritmos para ML</i>	23
3.1.3 <i>Machine learning supervisado</i>	24
3.1.4 <i>Machine learning sin supervisión</i>	25
3.1.5 <i>Algoritmos de machine learning reforzados</i>	26
3.2 Otros usos de ML	28
4. APLICACIONES DEL BUSINESS INTELLIGENCE Y EL BIG DATA	29
4.1 Data mining	32
4.2. Aplicaciones del data mining: Customer analytics	33
4.3 Proceso del data mining	36

4.4 Técnicas para data mining	39
4.4.1 <i>Modelado predictivo</i>	39
4.4.2 <i>Segmentación de la base de datos</i>	40
4.4.3 <i>Análisis de vínculos</i>	40
4.4.4 <i>Detección de desviación</i>	41
4.5 Diferencias entre ML y data mining	41
5. IMPLANTACIÓN DE UN PROYECTO DE BIG DATA.....	44
5.1 Consolidación de datos	45
5.2 Preprocesamiento de datos con WEKA	51
6. CUSTOMER ANALYTICS	54
6.1 Marketing	56
6.2 La calidad	58
6.3 Estadística descriptiva.....	59
6.4 Representaciones gráficas	62
6.5 Parámetros estadísticos	64
6.5.1 <i>Media</i>	66
6.5.2 <i>Mediana</i>	67
6.5.3 <i>Moda</i>	67
6.5.4 <i>Cuartiles</i>	68
6.5.5 <i>Quintiles</i>	68
6.5.6. <i>Deciles</i>	69
6.5.7 <i>Percentiles</i>	69
6.5.8 <i>Rangos</i>	70
6.5.9 <i>Desviación media</i>	70
6.5.10 <i>Varianza</i>	71
6.5.11. <i>Desviación típica</i>	71
6.5.12 <i>Coeficiente de Pearson</i>	71

<i>6.5.13 Análisis de tendencias</i>	72
6.6 Árboles de probabilidad (o árbol de decisión en análisis predictivo) .	72
6.7 Inferencia	79
7. SEGMENTACIÓN DE LOS DATOS	84
7.1 Clustering	85
7.2 Aplicaciones del análisis de clúster	85
7.3 Características deseadas en el clustering	87
7.4 K-medias clustering.....	89
7.5 Implementando K-medias en WEKA	89
7.6 Implementando K-medias en R	94
8. GESTIÓN DEL VALOR DEL CLIENTE	99
9. INTRODUCCIÓN A LENGUAJE R Y WEKA.....	100
9.1 Introducción a WEKA	101
9.2 Instalación.....	102
9.3 Uso.....	106
9.4 Introducción al lenguaje R	116
9.5 Instalación.....	116
9.6 Uso.....	119
9.7 Tipos de variables en R	120
9.8 Operadores	121
9.9 Instalando paquetes en R.....	123
9.10 Carga de datos	125
10. CLASIFICACIÓN DE DATOS CON WEKA Y R	131
10.1 Clasificación	132
10.2 Tipos de clasificación	133
10.3 Entrada y salida	133
10.4 Trabajando con clasificación	134

10.5 Árbol de decisión.....	137
10.6 Clasificador bayesiano ingenuo	138
10.7 Implementando la clasificación en WEKA mediante un árbol de decisión	139
10.8 Implementando clasificador bayesiano ingenuo en WEKA.....	145
10.9 Implementando clasificación en R mediante árboles de decisión ..	148
10.10 Clasificador bayesiano en R	150
11. REGRESIÓN CON R Y WEKA.....	154
11.1 Regresión Lineal.....	155
11.2 Regresión lineal simple.....	159
RESUMEN FINAL	163

RESUMEN INTRODUCTORIO

En esta unidad abordaremos los conceptos fundamentales sobre análisis de grandes datos, así como las principales herramientas disponibles para ello, como el data mining, el software estadístico y los lenguajes de programación especializados. Igualmente, conoceremos como están ligadas todas estas piezas al análisis de Big Data y a la ciencia de datos.

Aprenderemos los fundamentos para hacer uso de las principales técnicas de minería de datos y clasificación de estos, como son los árboles de decisiones, el particionado o segmentación de los datos, el clasificador Naive Bayes o clasificador bayesiano ingenuo y lo haremos mediante el uso de ejemplos prácticos que nos facilitará el entendimiento de estos conceptos.

Y, además estudiaremos herramientas como WEKA y lenguaje R.

INTRODUCCIÓN

Las mejoras en las infraestructuras tecnológicas están posibilitando unas capacidades inimaginables hace muy pocos años en materia de análisis avanzado de datos, lo que se suele denominar como Big Data Analytics.

Pero este cambio no es una cuestión exclusivamente informática, más bien la tecnología es un habilitador para un cambio de paradigma en la gestión empresarial, que se dirige hacia la empresa Data Driven, es decir, organizaciones donde toda planificación y toma de decisiones debe estar sustentada por un conjunto de evidencias en forma de datos, frente a la experiencia e intuición aplicada tradicionalmente.

Los proyectos de analítica avanzada, que persiguen un modelo Data Driven, se suelen enmarcar dentro de las estrategias de transformación digital. Esta visión es totalmente lógica, dado que el crecimiento exponencial de los datos se ha debido, en gran medida, a la propia digitalización de la empresa (redes globales, multicanal, internet de las cosas, etc.). Las compañías están poniendo tanto foco en dotarse de capacidades avanzadas de análisis de datos, que los presupuestos para ello continúan crecientes y las posiciones de nueva creación son más frecuentes que nunca. Empieza a ser difícil ya localizar una organización de cierta complejidad que no disponga de un chief data officer y un equipo de científicos de datos.

Dominar la gestión de los datos e interpretar sus beneficios requiere de múltiples habilidades y conocimientos: gestión empresarial, matemáticas, estadística, informática, psicología, etc. A lo largo de esta unidad vamos a

conocer muchos de esos conceptos, desde la minería de datos hasta los algoritmos que los convierten en inteligencia y conocimiento accionable.

Los perfiles de informáticos con conocimientos de análisis cada vez serán más frecuentes tanto en organizaciones públicas como privadas, es decir, profesionales con un alto conocimiento del negocio, que al mismo tiempo dominan, desde un punto de vista funcional, las tecnologías disponibles. Estos profesionales serán los responsables de diseñar e implementar las nuevas arquitecturas empresariales o, lo que es lo mismo, construir la organización Data Driven.

CASO INTRODUCTORIO

Eres el responsable de la informática de una pyme dedicada a la comercialización de productos electrónicos de bajo coste importados de China.

Vuestro modelo de negocio se basa en anticipar las tendencias del mercado para hacer una compra de volumen en China a bajo precio y, posteriormente, vender esos productos a las grandes cadenas de distribución comercial en España. Dado que los plazos logísticos desde China son de aproximadamente un mes más otro mes de antelación para que el fabricante pueda producir la cantidad demandada, la gestión de los tiempos es muy importante.

Ante el crecimiento de la competencia, la reducción de los márgenes de beneficio comercial y otras dificultades y amenazas que se vislumbran en el horizonte empresarial, el director general de la compañía te ha encargado que realices un análisis interno para determinar qué acciones se pueden poner en marcha que permitan mejorar el análisis realizado y, en general, la explotación de todos los datos corporativos.

Al finalizar esta unidad serás capaz de: distinguir los principales algoritmos para el procesamiento de los datos, determinar qué herramientas tecnológicas son más útiles para cada necesidad en materia de análisis, instalar e implementar algoritmos en Weka y desarrollar tus propios códigos de programación para análisis de datos en R.

1. BIG DATA ANALYTICS

Con el nuevo proyecto que te han encargado ya en tus manos, tienes por una parte una sensación de nerviosismo porque es algo completamente diferente de lo que has hecho hasta la fecha y, por otro lado, has generado una gran ilusión porque tienes la intuición de que, en esta ocasión, el departamento de informática que diriges va a poder hacer una gran aportación a la empresa que la ayudará a sobrevivir en este difícil mercado.

Tu primera decisión, dado que eres bastante novato en estas cuestiones, porque hasta el momento te habías ocupado de temas como el ERP o el CRM de la compañía, ha sido matricularte tanto tú, como los principales miembros de tu equipo en un curso de Big Data Analytics para comenzar a adquirir conocimientos que puedan ser puestos en práctica a muy corto plazo.

También has convocado una sesión de tormenta de ideas para, partiendo de una hoja en blanco, empezar a identificar puntos de trabajo.

Pero, ante todo, lo primero que quieras hacer, es verificar si realmente en la empresa estáis trabajando con Big Data o todavía no habéis llegado a esa fase. Es decir, si las V definitorias del Big Data se satisfacen o no.

A continuación, haremos un repaso por los conceptos más importantes que usaremos a lo largo de esta unidad. Es fundamental tener claro esto para poder entender en profundidad los temas siguientes.

- Ciencia de datos
- Datos
- Big Data



Big Data

Fuente: <https://www.pxfuel.com/en/free-photo-qfxuf>

1.1 La ciencia de datos

La ciencia de datos la podemos definir como la metadisciplina que se encarga del estudio de los datos a través de disciplinas como la **estadística**, la matemática, los procesos y los sistemas. Está altamente ligada a ramas de la informática, tales como el machine learning (aprendizaje automático), el data mining (minería de datos) y otros.

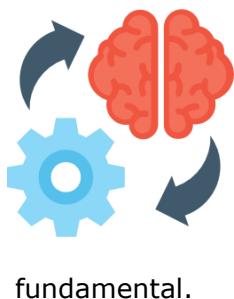
La ciencia de datos busca consolidarse en base a múltiples disciplinas, tales como la matemática, estadística y otras para lograr **predecir comportamientos** o **encontrar patrones** que pueden ser aplicados en distintas áreas, tales como: medicina, marketing, informática, biología, finanzas, robótica, educación, bioinformática, seguridad y muchas más.

Entre los usos más conocidos tenemos su aplicación en los motores de búsqueda, donde se usa para mostrar mejores resultados. En el campo de la salud se pueden prever comportamientos en la población comparando distintas variables como el clima y la cantidad de enfermos por gripe y se puede ir más allá con datos más detallados tales como temperatura, zona geográfica, historial médico de los pacientes....

Esta nueva metadisciplina ha generado un nuevo rol en nuestra sociedad, al que podríamos llamar el **científico de datos**, que vendría a tener un perfil avanzado mixto, tanto en el estudio de datos como de programación. Será

una persona que sabe más de estadística que un programador y más de programación que un estadístico.

También vemos que, por lógica, la importancia de los avances en estos campos genera una necesidad de capturar más y nuevos datos.



RECUERDA

En la ciencia de datos es bastante importante tener conocimientos de matemáticas y estadística para el análisis, pero es aún más importante tener una fuente de datos fiable, disponer una manera de capturar estos datos es fundamental.

1.1.1 El dato

Un dato no es más que un valor que puede ser cualitativo o cuantitativo, y describen una acción o un aspecto sobre algo.

Del latín *datum*, que significa "*lo que se da*", un dato puede ser, por ejemplo: la edad de una persona, su sexo, la temperatura del ambiente, el color del cielo, etc.

1.1.2 El Big Data

Es un término utilizado para grandes volúmenes de datos, ya sean estructurados o no, pero que poseen potencial para ser extraídos y obtener información y conocimiento de ellos. También se le conoce en español como macro datos. Se podría marcar una línea para decir cuando un conjunto de datos puede ser catalogado como Big Data. Esta línea se cruza cuando el tamaño de los datos supera la capacidad del software convencional para procesar y almacenar dichos datos en un tiempo razonable.

Cuando hablamos de Big Data, también nos referimos a cierto tipo de infraestructura, tecnologías y servicios asociados a los datos.

Algunos ejemplos claros de fuentes de Big Data son: internet y las redes sociales, transacciones financieras globales, datos sobre biometría y reconocimiento facial de una comunidad de personas...

Los cinco aspectos que debemos tener en cuenta cuando analizamos Big Data y que están muy difundidos en libros y páginas webs, es lo que se denominan las **5 v** del Big Data, que son las siguientes:

1. **Volumen:** nos referimos a que, por lo general, son cantidades tan masivas de datos que no pueden ser analizados por el software común. Se requiere por tanto de nuevas herramientas. Cabe destacar que, con la automatización y digitalización de muchos procesos a nivel global, la mayoría de esta información es generada de manera automática, ya sea por sistemas de información o por dispositivos. Es lo que se viene denominando internet de las cosas.
2. **Velocidad:** los datos se generan con gran velocidad, haciendo que sea muy difícil analizarlos en un tiempo que nos permita utilizarlos para la toma de decisiones. Es lo que se suele denominar **tiempo útil de toma de decisiones**.
3. **Variedad:** los datos provienen de distintas fuentes, pueden originarse en distintos dispositivos, diversos sistemas, por lo que lograr una integración de los datos se vuelve más difícil.
4. **Veracidad:** la gran cantidad de datos y las distintas fuentes, pueden provocar que nuestra información se torne inconsistente. Para evitarlo, se analizan los datos y se comprueba su veracidad con ciertas técnicas
5. **Valor:** generar información relevante después de procesar los datos será el mayor trabajo, pero este esfuerzo aportará soluciones a problemas.

En el amplio mundo del Big Data existe muchas nuevas tecnologías emergentes y algunas que ya tienen una trayectoria de tiempo en el mercado.

Entre las principales herramientas para el manejo de Big Data tenemos:

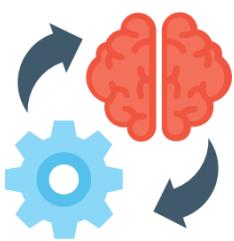
- **Lenguaje R:** Es uno de los lenguajes más utilizados en minería de datos y análisis estadístico, es bastante compatible con gran cantidad de gestores y repositorios de datos.
- **Hadoop:** Es una herramienta para el manejo de grandes volúmenes de datos. El mantenimiento de este producto es llevado a cabo por el proyecto Apache, bajo licencia de software libre. Además, ofrece una implementación de **MapReduce** que es un framework para la gestión de datos que nos permite manejar grandes cantidades de datos de manera paralela.

- **Spark:** Es un framework de computación en cluster y es de código abierto. Nos permite el uso de lenguajes como Java, Python y R. Es compatible con la mayoría de los sistemas gestores de bases de datos.
- **Storm:** Funciona en tiempo real, siendo también de código libre y permite utilizarlo con una gran cantidad de lenguajes.
- **Hive:** Consiste en una estructura para la implementación de un Data Warehousing, facilitando todas sus operaciones.
- **WEKA:** Es una colección de herramientas de visualización y algoritmos. Está realizado en java y es usado fundamentalmente para actividades docentes y de investigación. También es distribuido bajo la licencia GNU.

1.2 El flujo para la ciencia de datos

Para que dispongamos de un panorama general, diremos que el uso de la ciencia de datos pudiéramos dividirlo en tres grandes fases que serían:

1. **Captura:** No importan las herramientas utilizadas o la disponibilidad de software de última generación. Esto no servirá de nada si no existen datos para analizar. En este momento serán de mucha utilidad aquellas habilidades que poseamos sobre bases de datos y modelado de las mismas. Esto nos ayudará a ver como nuestros datos se conectan y se relacionan entre sí.
2. **Análisis:** Ya teniendo nuestros datos bien estructurados, funcionales y organizados, podremos ver la complejidad de estos y será posible comenzar a hacer estudios sobre dichos datos. Esto se consigue, entre otras, a través del uso de herramientas estadísticas. También la aplicación de **algoritmos** nos podrá ayudar a encontrar ciertos **patrones**.
3. **Presentación:** Buscar la mejor manera de mostrar los comportamientos de nuestros datos es importante para que personas que no están ligadas a las ramas de la estadística y matemática entiendan nuestro trabajo. El uso de gráficos será fundamental, así como de otras herramientas de visualización, storytelling, etc.



RECUERDA

El uso de la ciencia de datos no siempre es concluyente. Esto quiere decir que podemos encontrar un camino, pero que ese primer paso no nos permita continuar, o que debamos resolver previamente otra situación, para más adelante retomar el problema original.

2. HERRAMIENTAS FUNDAMENTALES DEL BIG DATA ANALYTICS

Una de las ideas que te propone uno de tus colaboradores es que comencéis a explotar de forma eficiente aquellos elementos que ya están presentes en vuestra infraestructura de sistemas, antes de introducirlos en la compra o construcción de nuevas herramientas.

Para este objetivo, os ha parecido muy interesante tener en cuenta el concepto de huella digital ya que desde hace aproximadamente diez años tenéis una web que les permite a vuestros clientes visualizar el catálogo de productos y hacer pedidos. La web puede ser una fuente tremadamente interesante de información histórica para comenzar a analizarla. Pero también sois conscientes de que debéis mejorar la capacidad analítica de la web para comenzar a realizar mediciones que hasta la fecha no se planteaban. Por ejemplo, ¿podría ser interesante analizar las búsquedas que efectúan los clientes en el buscador de productos de la web?

Creéis que, probablemente, pude ser un buen método para anticipar tendencias. Porque, si nos preocupamos de almacenar todos los datos que generamos en la empresa después podremos realizar análisis sencillos que nos devolverán respuestas muy interesantes.

Aparte de los productos de mercado ya construidos, unos de pago y otros de software libre, la programación será uno de nuestros principales ayudantes cuando desempeñemos el rol de analista o de científico de datos, así que es buena idea hacernos con un dominio de esta para facilitar nuestras tareas.

Para el análisis de datos vamos a requerir conocimientos de programación, pero estos van a depender de nuestras necesidades, o de lo que queramos

hacer y de donde estén nuestros datos. Por eso, es aconsejable ser flexible y prestar atención a los siguientes criterios en el momento de elegir una herramienta para trabajar:

- ¿Has trabajado anteriormente con ese lenguaje? Si es así, aunque no sea el lenguaje perfecto, tal vez tu conocimiento de este te haga decantarte por él.
- La disponibilidad de librerías y documentación sobre las mismas nos hará más fácil y rápido nuestro trabajo. Así que es una buena idea, antes de decantarnos por un lenguaje, explorar las librerías disponibles tanto gratuitas como de pago. En caso de tener un coste asociado será necesario valorar si el pago justifica la capacidad aportada.
- El tipo de análisis a realizar en nuestros datos de forma que el lenguaje elegido se adapte al análisis tanto desde un punto de vista funcional (que se pueda realizar) como desde el punto de vista de rendimiento (que se realice en un tiempo razonable)

2.1 La ciencia de datos en el mundo real

En la actualidad, el estudio de los datos se ha convertido en la principal herramienta en áreas como el marketing y publicidad. Pero también en el sector financiero y en el de los recursos humanos.

El machine learning domina la escena tecnología actual, permitiendo cosas que antes parecían imposibles, tales como el uso de publicidad personalizada basada en los gustos de un usuario.

¿Cómo es posible esto? Pues, por ejemplo, se consigue mediante el estudio de tipo histórico sobre las acciones tomadas en el pasado y con el uso de inteligencia artificial se pueden predecir las acciones que se tomaran en el futuro.

En base a la actividad usual y otras variables como el género, época del año, etc..., se pueden predecir futuras necesidades de las personas. O, en base al comportamiento de un trabajador en sus tareas profesionales, es posible predecir si va a abandonar la compañía.

Esto no se queda aquí. Por ejemplo, tenemos las aplicaciones en el área de la medicina, donde podríamos calcular las necesidades de ciertos medicamentos a lo largo de las temporadas de enfermedades recurrentes y en base a esto planificar una producción personalizada.

Estos son apenas algunas de las aplicaciones que puede tener la ciencia de datos en el mundo real. Pero, la verdad, es que son infinitas las posibilidades y, cada día, científicos y emprendedores en todo el mundo descubren una nueva.



SABÍAS QUE...

En el campo de los deportes se ha utilizado también la ciencia de datos. Un caso es en la NBA, donde han llegado a conocer la cantidad de aciertos que puede tener un jugador en determinadas jugadas. El uso de las cámaras en cada lance del partido permitió crear un gran big data el cual posteriormente fue estudiado por los equipos técnicos y deportivos de las distintas franquicias que componen la liga estadounidense.

2.2 La ética, el big data y la ciencia de datos

A medida que está más difundido el estudio de los datos, se plantea una duda, que es hasta donde resulta éticamente correcto recopilar información. Este es un debate que siempre va a estar presente entre la sociedad, políticos y medios de comunicación. En la actualidad, existe un monitoreo sobre las actividades que realizamos, tales como compras en internet, actividades en redes sociales públicas, transacciones financieras... Y todo va a depender del uso que se le dé a esta información... No es tanto el algoritmo el que será más o menos ético, sino que dependerá de las decisiones que se tomen en base a esos algoritmos.

2.3 Huella digital

Es un concepto que se escucha mucho actualmente y se refiere a un conjunto de acciones o actividades únicas realizadas en el internet rastreable. Es la información que queda como resultado de la navegación web de un usuario y es una gran fuente de datos para la aplicación de ciencia de datos sobre ella.

Existen dos tipos de huella digital. Están por un lado las pasivas, las cuales son recopiladas sin que el usuario lo sepa. Es verdad que, dependiendo del país, la legislación obliga a las páginas web a advertir de que se está generando una huella digital de cada persona, por ejemplo, mediante el uso de trackers y cookies. Aunque la realidad es que, por falta de tiempo o

desconocimiento, la gran parte de los usuarios no prestan atención a estas advertencias.

Las huellas activas son las que se crean cuando el usuario directamente divulga esta información. Es decir, cuando de forma voluntaria, en ocasiones de modo desinteresado y en otras circunstancias a cambio de una contraprestación, nosotros mismos aportamos información y datos a empresas y brokers de datos.

2.4 Business intelligence

A menudo vemos el término **business intelligence (BI)** acompañado del término **Big Data**. Podemos definir **BI** como un conjunto de herramientas, estrategias, aplicaciones, tecnologías y arquitecturas que están enfocadas a recolectar información sobre una organización. Toda esta información recopilada -por lo general- se puede recolectar y, haciendo uso de la ciencia de datos, explotarla. También es posible analizar un repositorio de **Big Data** creado por un tercero. Aunque no es el único campo de acción del **BI**, cuyo alcance también cubre a pequeños almacenes de datos. Por tanto, podemos decir que el business intelligence es aplicable sobre el Big Data, pero no todo el business intelligence tiene porque ser Big Data. En muchas ocasiones, con BI se puede obtener conocimiento muy valioso mediante la aplicación de sus herramientas a conjuntos de datos muy pequeños.

Algunas de las herramientas más conocidas de BI son desarrolladas por grandes empresas tecnológicas como Oracle, Microsoft o SAP. Otras son el resultado del trabajo de pequeñas empresas especializadas. Incluso muchas de ellas son tecnologías que se han creado en laboratorios o universidades y son de uso libre. También hay que reseñar que en ocasiones las herramientas de business intelligence son independientes y en otras circunstancias están integradas en plataformas de mayor calado como son los sistemas de CRM o de ERP.

Algunas de las más conocidas son:

- Style Intelligence
- Microsoft Dynamics
- SAP business intelligence
- Oracle Business Intelligence
- Tableau
- IBM Cognos Analytics
- Clear Analytics
- QlikView
- Gooddata

- Sisense

2.5 Minería de datos

Otro termino que está bastante relacionado con el Big Data es el de **minería de datos**. La minería de datos no es más que la transformación de datos en información y conocimiento accionable, es decir, el proceso de aplicar la ciencia de datos sobre un conjunto de Big Data, con el objetivo de crear o descubrir nuevo conocimiento, se denomina minería de datos.



Minería de datos

Fuente: <https://es.vecteezy.com/vectores-gratis/miner%C3%ADa-de-datos> > Minería De Datos Vectores por Vecteezy

3. FUTURO DEL BIG DATA

El machine learning está en boca de todo el sector. Cada vez que has representado a tu empresa en foros, congresos o que has visitado ferias tecnológicas has podido comprobar la cantidad de fabricantes de software que presentan tecnologías y soluciones de machine learning para tu sector. Incluso eres conocedor de lo rápido que han crecido empresas que son competencia de tus clientes, como es el caso de Amazon, gracias a una aplicación masiva del machine learning.

Por tanto, en tu agenda la palabra ML está marcada con letras rojas. Como tú no puedes abordar todas las líneas de trabajo y no quieres que el aspecto de machines learning quede descuidado por tu falta de tiempo, has encargado a dos personas del equipo que se centren exclusivamente en investigar que algoritmos de machine learning existen y cómo se están aplicando a la función de ventas y al sector de la distribución comercial.

La verdad es que son infinidad de aspectos en los que podríais mejorar con machine learning. Por ejemplo, una de las tareas que más tiempo y dinero consume en vuestra empresa, como en todas las compañías de distribución, es el servicio postventa. Con machine learning sería posible, dada la gran cantidad de productos que vendéis de muy diversos fabricantes, poder explotar los datos disponibles para identificar aquellos productos que presentan más fallos. De esta forma podrías decidir dejar de comercializar productos de determinados fabricantes para evitar que posteriormente las quejas de los clientes, los reemplazos, las indemnizaciones, etc. Os hagan perder tanto tiempo y dinero.

El futuro del Big Data, entre otras cuestiones, pasa principalmente por introducir la inteligencia artificial en la explotación de los datos. En el momento actual de la tecnología, hablar de Big Data y hablar de machine learning va de la mano.

En este apartado nos centraremos en entender cómo funciona el **machine learning**, conocido por sus siglas **ML**. Veremos de qué elementos está compuesto y cuáles son las principales funciones que se basan en él en la actualidad.

El ML también es conocido como aprendizaje automático y se ha convertido en una gran herramienta, indispensable en cualquier empresa.

Puede parecer futurista y podríamos considerar poco probable toparnos con esta tecnología, pero, la realidad es que ya está incorporada a múltiples usos e interactuamos con ella en nuestro día a día, a veces sin saberlo.

Cuando navegamos en internet, estamos constantemente expuestos a las capacidades de ML. Cuando utilizamos un buscador como Google o Bing, estos usan ML para optimizar que páginas mostrarnos, que publicidad será más eficaz...

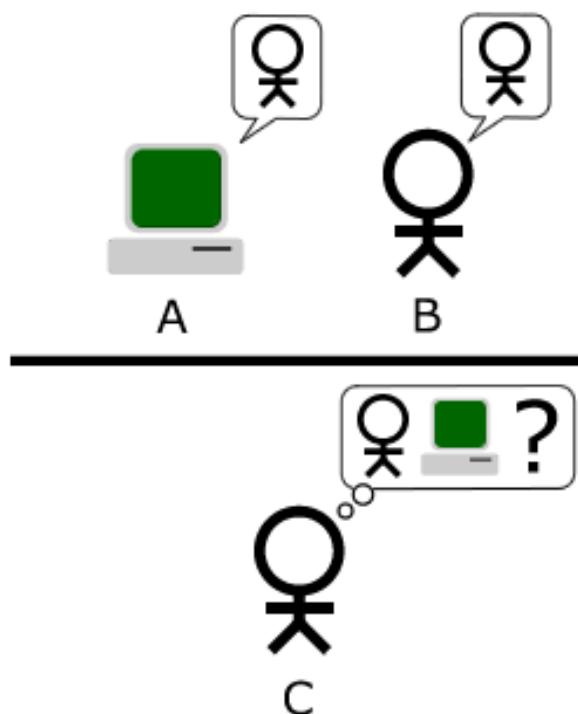
El ML es una rama de la inteligencia artificial y es lógico preguntarnos cómo se relaciona con la ciencia de datos y el data mining, puesto que cumple un papel importante en el análisis del Big Data, dado que su papel es encontrar patrones complejos de manera automatizada.

Como veremos a continuación, profundizaremos un poco para entender bien de qué trata el ML.

3.1 Machine learning

El ML y la inteligencia artificial en general ya tienen un recorrido histórico importante, desde el año 1950, cuando Alan Turing diseñó el "test de Turing", que determina si una máquina es inteligente.

Para que un sistema no humano pase el test, la máquina debe ser capaz de engañar a un ser humano haciéndole creer que está interactuando con un humano.



Esquema del test de Turing.

Fuente: <https://www.wikipedia.org>

Posteriormente, en 1952, Arthur Samuel escribe el primer programa de ordenador capaz de aprender. Este primitivo software jugaba a las damas y cada vez que lo hacía, mejoraba el nivel de su juego, partida tras partida.

En 1956, Marvin Minsky y John McCarthy, con la ayuda de Claude Shannon y Nathan Rochester, en una conferencia en Dartmouth, acuñaron el término inteligencia artificial.

En 1958 se crea el Perceptron, por Frank Rosenblatt, la primera red neuronal artificial.



Rosenblatt trabajando en el Perceptrón en los años 50

Fuente: cornell.edu

De igual manera, con el transcurso de los años, surgieron nuevos avances que conformaron la inteligencia artificial y permiten llegar al ML tal cual hoy lo entendemos. Algunos de los hitos más importantes que contribuyeron a esta evolución fueron:

- En 1959: Se crea el Standford Cart, un robot capaz de moverse por una habitación de manera automática.
- En 1967: Nace el algoritmo del “vecino más cercano” cuya invención marca el inicio de campo de **reconocimiento de patrones**.
- En 1981: Se crea el concepto de EBL (Explanation Based Design). Un modelo de prueba, donde se realizan entrenamientos con datos y se descartan los menos relevantes.
- En 1985: NetTalk, una red neuronal artificial, crearía un mecanismo de aprendizaje. Fue propuesto por Terrence Sejnowski y Charles Rosenberg y era capaz de aprender la pronunciación tal cual lo haría un niño

- En 1990: Empiezan los primeros estudios donde se contemplaba la función del análisis de grandes datos.
- En 1997: Un ordenador denominado **Deep Blue** vence al campeón de ajedrez Gary Kasparov.
- En 2006: Se comienza a utilizar el término deep learning (Aprendizaje profundo) para explicar nuevas tecnologías en el campo de la inteligencia artificial y el aprendizaje automático.
- En 2011: Un ordenador llamado Watson, vence a un grupo de humanos en un concurso de preguntas muy conocido en la televisión.
- En 2012: Nace el proyecto **Google Brain**, el cual hizo uso de los grandes recursos de Google, por lo que en el mismo año de su lanzamiento fue capaz de entrenarse a sí mismo para reconocer un gato basando diez millones de imágenes digitales tomadas de videos de Youtube, propiedad también de Google. Precisamente esto nos destaca la gran ventaja de la que disponen compañías masivas de internet al disponer de tantos datos que son capaces de entrenar a multitud de sistemas de machine learning, como, por ejemplo, los **traductores automáticos**, los **reconocedores faciales**, los sistemas **OCR** de transcripción de textos, los **chatbots**, etc.



Logo de Google Brain
Fuente: <https://research.google/teams/brain/>

- En 2014: Facebook desarrolla **DeepFace**, con la capacidad de reconocer rostros y que tiene una capacidad de acierto del 97%.
- En 2014: Google compra **DeepMind**, la cual ha creado una red neuronal capaz de imitar la memoria a corto plazo de un humano.
- En 2015: Amazon lanza su propia plataforma de ML.
- En 2015: Microsoft crea el Distributed Machine Learning Toolkit.
- En 2015: Se crea la organización OpenAI, que es una organización sin fines de lucro, aunque es verdad que recientemente ha desarrollado una filial comercial para poder abordar determinados proyectos. Esta organización se nutre de

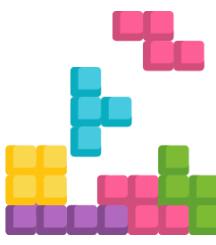
las donaciones millonarias que recibe de grandes empresas del sector.

- En 2016: Google DeepMind derrota a jugadores en el juego Go, considerado uno de los juegos de mesa más complicados, donde demostró la capacidad de creación de jugadas que resultaron sorprendentes e inexplicables, incluso para los mejores jugadores del mundo.



SABÍAS QUE...

En 1996 El ajedrecista Gary Kasparov venció al computador de IBM, pero después de introducir mejoras en el año siguiente, una nueva versión del software logra vencer al ruso.



EJEMPLO PRÁCTICO

El director de una empresa con muchos años de historia es consciente de que la burocracia está lastrando mucho los resultados de la compañía y ha decidido contratar a un consultor informático para que le recomiende que nuevas tecnologías aplicar

¿Qué recomendaciones podría ofrecer al respecto?

SOLUCIÓN

Al ser una empresa de larga tradición e historia, en muchos casos con estructuras de tipo funcional, este tipo de empresas se ven ralentizadas por la burocracia interna.

El uso de una herramienta como los chatbots, para resolver las consultas internas más frecuentes, podría aliviar parte de la pesada carga burocrática.

También serían una gran ayuda para reducir los costes derivados de procesos burocráticos de escaso valor añadido implantar tecnologías de las denominadas RPA (automatización robótica de procesos) para automatizar tareas como la elaboración de las nóminas.

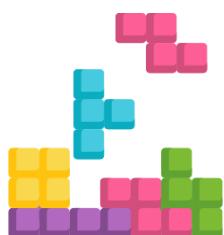
3.1.1 El machine learning

El machine learning es una rama de la inteligencia artificial, la cual se encarga de crear sistemas capaces de aprender por sí solos, buscando patrones dentro

de grandes cantidades de datos, es decir, los sistemas de IA tienen la capacidad de ir mejorando con el tiempo, haciendo uso de tecnologías como el deep learning, que consiste en el uso de enfoques de razonamiento humano para el aprendizaje.

En el ya lejano año de 1958, Arthur Samuel definió el ML como “el campo de estudio que da a las computadoras la habilidad de aprender sin ser explícitamente programadas”.

Este concepto que, como comprobamos, se creó ya hace más de medio siglo, no se pudo popularizar en aquel momento porque entonces no se tenían los recursos de hardware necesarios para poner en práctica muchas de las teorías sugeridas. En el año 1998, Tom Mitchel refinó el concepto de la siguiente manera, definiendo que un sistema informático aprende de la experiencia **[E]** respecto a la tarea **[T]** siendo su desempeño medido por **[P]**, si su desempeño **P**, al realizar la tarea **T**, mejora con la experiencia **E**.



EJEMPLO PRÁCTICO

Un compañero está teniendo problemas con el correo basura de la cuenta de correo del trabajo y, como sabe tus conocimientos en Big Data, te hace una pregunta que le está rondando la cabeza:

¿Cómo se podría describir la aplicación de machine learning a la detección del correo basura?

SOLUCIÓN

Imagina que tenemos un programa que rastrea aquellos correos que son spam y basado en esto, va mejorando su desempeño o eficacia aprendiendo a detectar cuales son spam y cuáles no.

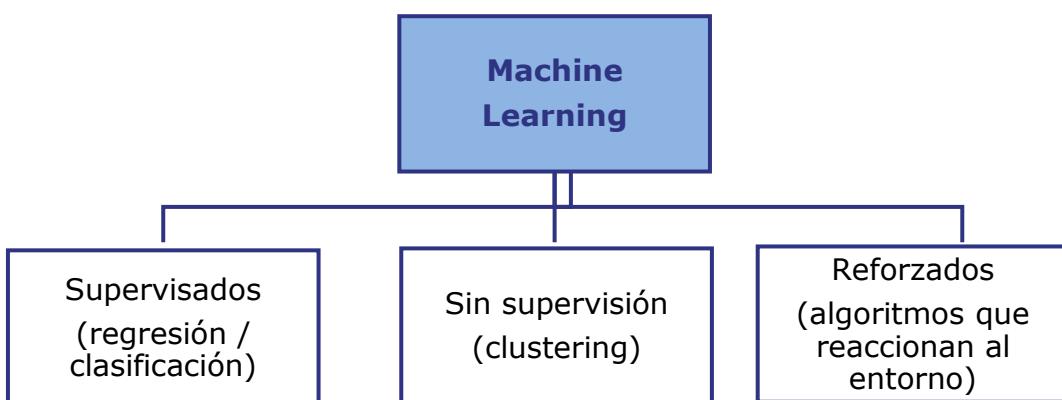
La clasificación de los correos que son spam sería la tarea **T**.

El rastreo de cuales correos el usuario marca como spam o no, será la Experiencia **E**.

El número de correos clasificados correctamente será la medida del desempeño **P**.

3.1.2 Clasificación de algoritmos para ML

En el siguiente esquema observamos la clasificación que se suele aceptar de manera generalizada entre los distintos algoritmos de machine learning, y que presentaremos con más detalle a continuación.



Taxonomía de algoritmos

Fuente: Elaboración propia

3.1.3 Machine learning supervisado

Supongamos que deseamos saber el precio estimado de un vehículo basándonos en su año de matriculación. Lo primero que debemos hacer es recopilar la información sobre el precio real de los vehículos clasificados por año, que podríamos presentar en un gráfico similar al siguiente:

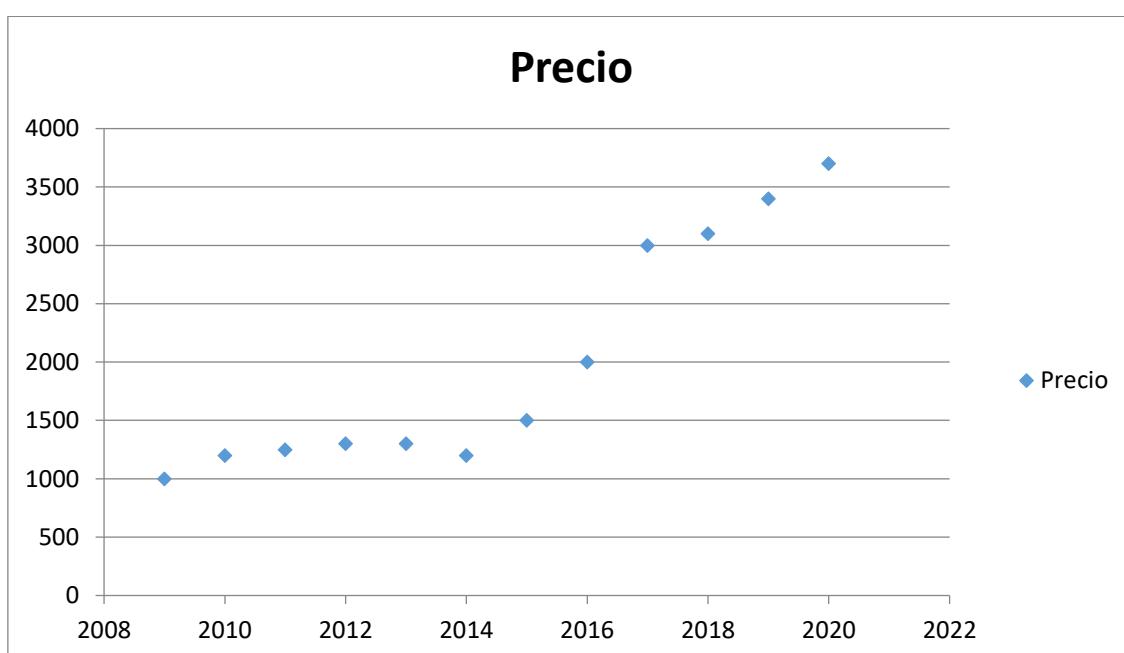


Gráfico de distribución de valores

Fuente: Elaboración propia

Ahora, quisiéramos saber o estimar el precio de un vehículo del año 2022. Con un algoritmo de ML podríamos hacer esto basándonos en los valores que ya tenemos porque este algoritmo podría dibujar una línea promedio y hacer una proyección de cuál sería el posible precio. Si, en vez de estimar en base a los datos disponibles, fuese posible realizar una segunda estimación utilizando un mayor número de datos, por ejemplo, tomando las estadísticas de compra venta de todo un país, se lograría una mejor estimación. Cuantos más datos tengamos, mejor será nuestra estimación, es una máxima que casi siempre se cumple en el machine learning.

Entonces ¿cuál es la parte supervisada en este ejemplo? Pues que debemos comprobar que al algoritmo se le aporten los valores correctos. Para que el algoritmo ofrezca respuestas correctas, debe recibir datos fiables.

A medida que añadimos información a nuestro modelo, este será más acertado.

3.1.4 Machine learning sin supervisión

Este tipo de algoritmo se usa cuando no se conoce como serán los resultados. Por lo general, se aplica cuando existen variables en el conjunto de datos de las cuales no sabemos qué efecto tienen. Solo tenemos atributos de entrada, los cuales se agruparán en grupos parecidos.

Por ejemplo, tenemos Google News, como un ejemplo de ML sin supervisión. Este algoritmo se encarga de agrupar noticias similares, es decir, cuando dos o más noticias de diferentes periódicos coinciden en el tema tratado, en vez de mostrarnos todas ellas por separado, se agrupan por similitud.

¿Cómo hace esto Google? Pues agrupando por los diferentes atributos de las noticias, como la frecuencia de una palabra, cantidad de páginas, longitud de frases, actores implicados y muchas más.

Entre otros muchos usos tenemos el análisis de las redes sociales, la segmentación de mercados e incluso para el uso astronómico detectando comportamientos de distintos cuerpos celestes.

Para que comprendamos mejor la diferencia entre ambos tipos de algoritmos, vamos a comparar las siguientes circunstancias. Tenemos una cesta llena de frutas y las queremos clasificar basándonos en nuestros conocimientos previos. Es decir, las frutas ya disponen de una etiqueta definida anteriormente (naranja, pera...). Seremos capaces de hacerlo fácilmente usando ML. En primer lugar, la tarea de aprender los tipos de frutas sería el entrenamiento y el resultado de este entrenamiento nos permitirá organizar las frutas por nombre. Por tanto, esto sería un ejemplo de **ML supervisado**.

Por lo contrario, en el **ML no supervisado** no se conocen los nombres de estas frutas, ni sus características para clasificarlas. En este caso, el algoritmo deberá tomar toda la información disponible que posee para organizar estos elementos y agruparlos según sus características. En definitiva, en el primer caso tenemos claro con antelación que salida queremos que presente el algoritmo y en el segundo caso, no.



PARA SABER MÁS

Compañías como la eléctrica EDP han abierto sus datos a cualquier interesado e, incluso, organizan hackatones online para recibir ideas. En algunas ocasiones incluso ofrecen un premio en dinero a los mejores algoritmos. Accede a este enlace para conocer más al respecto:

<https://opendata.edp.com/pages/homepage/>

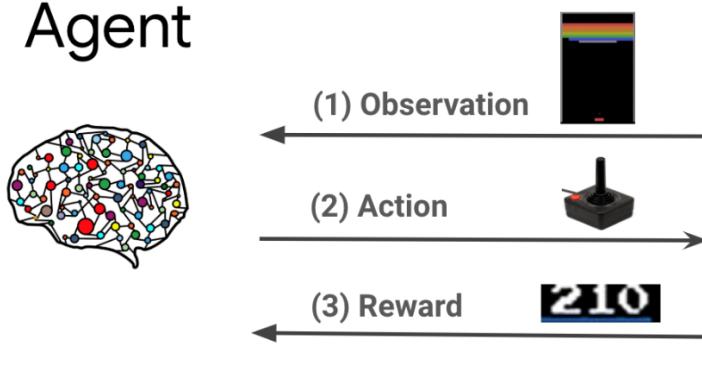
3.1.5 Algoritmos de machine learning reforzados

Este tipo de aprendizaje viene dado por el método de prueba y error, ya que el sistema tomará la decisión con la que obtenga mayor recompensa. Y, es muy usado en el desarrollo de juegos de ordenador o de navegación automatizada.

Dentro de estos algoritmos, normalmente, se encuentran los siguientes elementos:

- **Agente (agent)**: Es usado para la toma de decisiones y el aprendizaje. El agente elige la mejor opción basado en una recompensa medida. Por ejemplo, en una cantidad de tiempo.
- **Entorno (environment)**: Es el ambiente con el cual el agente interactúa.
- **Acciones (actions)**: Son las tareas que un agente puede realizar.

Agent



Environment



Elementos de un algoritmo machine learning reforzado.

Fuente: tensorflow.org

Este tipo de algoritmo nos permite determinar cuál es el mejor comportamiento en un determinado contexto, con el propósito de maximizar el desempeño.

En la actualidad los sistemas de navegación automática de los automóviles de Tesla Motors y de los drones de Amazon Prime air delivery, están basados en este tipo de algoritmos.

En el caso de los videojuegos el sistema de aprendizaje puede enfrentarse a jugadores humanos, que es lo que utilizada para su aprendizaje. Cuando el ordenador realiza una acción con la que gana, optará por aplicar de nuevo la misma acción. Por lo contrario, una acción con la que de forma reiterativa el ordenador siempre pierde, es una acción que progresivamente dejará de aplicar.



Drone de Amazon

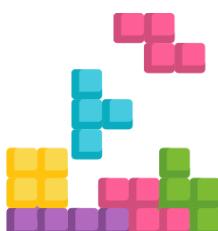
Fuente: amazon.com



ARTÍCULO DE INTERÉS

Para saber más de Amazon Prime Air lee este artículo:

<https://time.com/4185117/amazon-prime-air-drone-delivery/>



EJEMPLO PRÁCTICO

En una empresa de videojuegos de rol online se ha detectado una pérdida de usuarios porque en ciertos horarios es difícil para los jugadores encontrar contrincantes con los que enfrentarse.

Se desea buscar una alternativa para evitar esta pérdida de negocio.

SOLUCIÓN:

Para evitar perder ese tipo de jugadores que quieren conectarse a horas donde no hay contrincantes humanos, la estrategia pasa por permitirles jugar contra la máquina.

Lo ideal para crear ese jugador no humano, es aplicar un algoritmo de machine learning reforzado.

Esto es así porque en un juego de rol, los jugadores buscan la complejidad y la sorpresa. Un juego de rol donde las acciones por parte del contrincante están prediseñadas y son repetitivas no genera jugabilidad. Por este motivo, es necesario que el comportamiento de la máquina sea lo más parecido al humano, por lo que se recomienda programar un agente que irá aprendiendo por prueba y error y también aprenderá de las estrategias aplicadas por los jugadores humanos de forma que a medida que pase el tiempo será un jugador más cualificado y al que será más difícil vencer.

3.2 Otros usos de ML

En la actualidad, existen y se han popularizado múltiples asistentes personales automatizados como Siri o Alexa. Estos utilizan el procesamiento del lenguaje natural. También hay casos de empresas de transporte como UPS que usan ML para programar las rutas de sus entregas, minimizando el tiempo y los accidentes. Uber y Cabify también usan ML. Los bancos y el sector financiero en general es el campo más amplio de aplicación de ML en

el momento actual. Por ejemplo, a la hora de calcular los riesgos en la aprobación de préstamos y créditos.

Otro caso interesante podría ser Spotify, cuyas recomendaciones se hacen utilizando distintos modelos. El primero es uno de tipo colaborativo, donde se registran y analizan los comportamientos de los usuarios, sus opiniones y reproducciones de canciones. El otro es el que se encarga de buscar en blogs y comentarios en internet para revisar que canciones son tendencia y, por último, se analizan las canciones nuevas y se comparan con las más populares. Toda esta maquinaria de ML ofrece un sistema predictivo en base a los gustos de los usuarios y realiza sugerencias.

Ahora veamos el primer modelo de Spotify. Este debe usar algoritmos de tipo supervisado, pues se basan en la existencia de una información que es correcta y real (reproducción de canciones, opiniones, etc.). El de tipo no supervisado busca patrones comunes entre las nuevas canciones para predecir a qué usuarios podrán gustar.

4. APLICACIONES DEL BUSINESS INTELLIGENCE Y EL BIG DATA

El business intelligence es un concepto que en tu empresa resulta más familiar. No es algo tan novedoso como el Big Data. Es más, el ERP de la compañía, con el que desde hace años prestáis soporte a toda la administración corporativa: ventas, facturación, almacén, logística, etc. Dispone de un módulo de business intelligence que os ha permitido crear un cuadro de mando que es utilizado por el director general y los directores de área.

Pero también eres consciente de que pasar del análisis tradicional al análisis de Big Data va a requerir más trabajo y nuevas tecnologías.

En todo lo que tiene que con análisis de los clientes o customer analytics seguro que las herramientas de BI tienen mucho que aportar. Probablemente comenzar a segmentar clientes permitirá que desde el departamento de informática podáis dar un fuerte apoyo a la gente de marketing.

La decisión que has tomado aquí es convocar una reunión con los directores de las diversas áreas de la empresa: finanzas, recursos humanos, logística, marketing y comercial. La idea es identificar sus principales problemáticas, es decir, que tipo de información consideran que les falta para mejorar su toma de decisiones y para cada una de las circunstancias realizar el proceso de data mining habitual, comenzando por la definición del problema, el entendimiento de los datos, etc. Hasta ser capaces de entregar el resultado

esperado por el usuario de negocio sea este una segmentación, un análisis de relaciones, etc.

En la era de la información, las cantidades de datos que generan las industrias, empresas y entidades de todo tipo es increíblemente grande. La necesidad de convertir esos datos en información y conocimiento que sean de utilidad para la organización crea el ambiente propicio para el data mining.

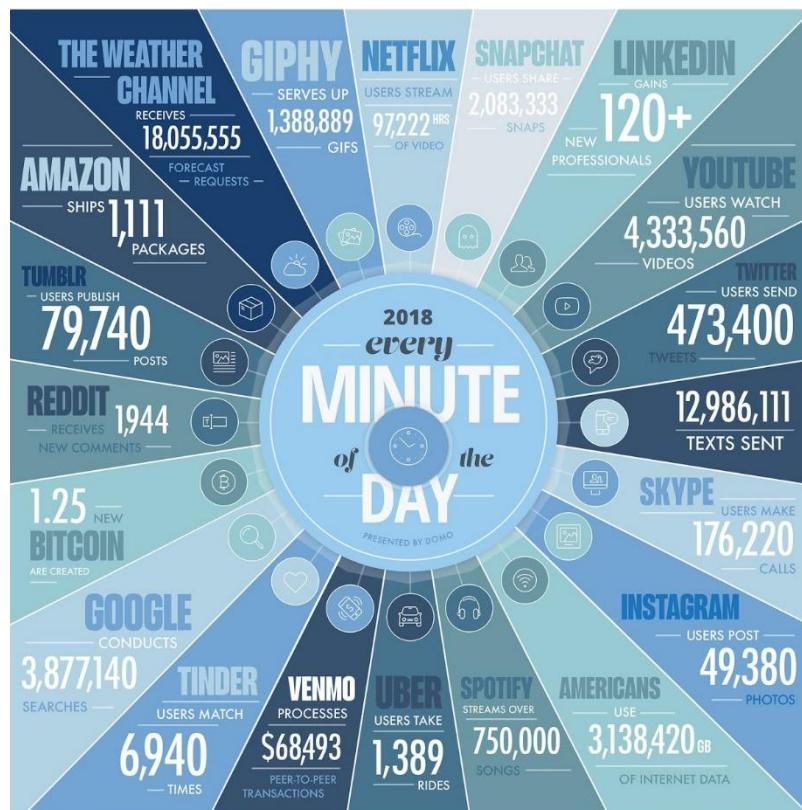
El data mining existe desde hace más de cincuenta años y ha renacido con un gran boom en la actualidad gracias, sobre todo, a la explosión de datos provocada por internet, aunque ya desde la década de los años sesenta el uso de bases de datos para almacenar información generaba enormes cantidades de datos.

En la actualidad, cada minuto que pasa, se generan muchísimos datos. El estudio de estos datos es una tarea importante y es realizada, en muchas ocasiones, a través del data mining. El analizar estos datos representa una gran oportunidad para mejorar el comportamiento de las empresas, porque analizar los datos les puede ayudar a encontrar valiosos conocimientos sobre los tipos de productos que sus clientes demandan, la cantidad que deben situar en cada punto de venta para que no haya roturas de stock, pero tampoco desperdicios, el lugar donde el cliente desea adquirir el producto y una gran cantidad más de variables de marketing, financieras, etc. que pueden marcar una gran diferencia.

Para que tengamos una referencia de la cantidad de datos que se generan a diario, podemos decir que el 90% de los datos del mundo han sido creados en los últimos 4 años.

Cerca de 2.5 quintillones de datos se producen cada día y se calcula que en la actualidad se generan 1.7 MB cada segundo, por cada persona en la tierra.

En el grafico siguiente tenemos un informe sobre la generacion de datos creado por Domo, una empresa de big data y análisis en la nube.



Indicadores sobre la creación de nuevos datos

Fuente: Domo.com



ENLACE DE INTERÉS

Para más información sobre Domo se puede visitar el siguiente enlace:

<https://www.domo.com/solution/data-never-sleeps-6>



PARA SABER MÁS

Algunos otros parámetros interesantes sobre la generación de datos son:

- Se suben unas quinientas horas de video cada minuto a YouTube.
- Más de quinientos nuevos sitios web son creados cada minuto en internet.
- Cien terabytes de datos son subidos a Facebook diariamente.

Estos son algunos de los números... ahora imaginemos lo difícil o, mejor dicho, el reto que significa el análisis de estos datos a nivel de infraestructura requerida. Los principales aliados para esto serán:

- La evolución de la tecnología de almacenamiento y comunicaciones con una reducción de costes asociada.
- La creciente valoración que se hace de los datos en las corporaciones.
- El uso de tecnologías modernas que permitan capturar datos cada vez de manera más rápida y confiable

Veremos con más detalle que es data mining y como se diferencia del ML, pero también cuáles son sus puntos en común y comprenderemos la necesidad de su existencia.

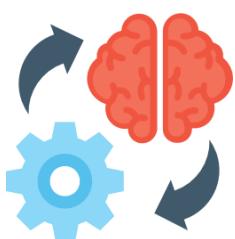
4.1 Data mining

Se define como un conjunto de técnicas para el descubrimiento (de manera automatizada y eficiente) de algo que se desconoce y que potencialmente será útil. El tipo de descubrimiento que se suele perseguir será un patrón entendible en grandes bases de datos y utilizable para la toma de decisiones.

Las principales características del data mining son:

- El data mining es un proceso automatizado para detectar patrones desconocidos que se encuentran ocultos en las bases de datos.
- Por lo general, estos datos están conformados por archivos históricos de la organización, conocidos como data warehouse. Esto no quiere decir, ni que el data warehouse obligatoriamente deba ser explotado mediante data mining, ni que el data mining solo se pueda aplicar sobre un data warehouse. Ambos conceptos son independientes, aunque tradicionalmente han estado asociados.
- El data mining debe lidiar con grandes volúmenes de datos, que van desde gigabytes, pasando por los terabytes hasta los zetabytes. A medida que han transcurrido los años, los sistemas de data mining se han adaptado a la explosión de los volúmenes de datos.
- Los patrones deben ser nuevos, útiles y entendibles. De no ser así, no tendría mucho valor el encontrar dichos patrones. Por este motivo no es suficiente con que el profesional del mining detecte un patrón con la aplicación de algún algoritmo, sino que alguien con conocimiento del negocio debe refrendar el valor o utilidad del patrón.

- El data mining debe determinar patrones en los registros históricos que permitan predecir posibles comportamientos, por lo tanto, extrapolar conocimiento del pasado al futuro.
- Aunque es posible aplicar data mining en bases de datos pequeñas, se suele cumplir la siguiente norma que dice que: cuanto mayor sea la cantidad de datos disponible, mejor será la precisión de la predicción que se pueda extraer de ellos.



RECUERDA

El data mining es una de las tecnologías que debemos observar con más atención en la actualidad, porque a pesar del mucho tiempo que lleva en el mercado, está viviendo una segunda juventud. No dejan de lanzarse nuevos softwares para explotar datos que aportan nuevas funcionalidades y enfoques y, sobre todo, se está alcanzando un caldo de cultivo muy prometedor con la fusión de inteligencia artificial y big data.



SABÍAS QUE...

El data mining es una poderosa herramienta capaz de detectar los patrones ocultos en los datos, pero no es capaz de determinar cuáles son importantes y cuáles no, de ahí que alguien en el equipo de datos deba tener un expertise solvente de la actividad de la empresa.

4.2. Aplicaciones del data mining: Customer analytics

En la actualidad, la utilidad del data mining alcanza a casi todos los campos de la empresa, ciencia y sector público desde las finanzas o las telecomunicaciones hasta los seguros, la detección de fraudes, la segmentación de marketing, el análisis de tendencias, la mejora del rendimiento comercial, el diseño web, los recursos humanos y la gestión del talento, etc.

Profundizaremos en algunas de estas aplicaciones a continuación:

- Aprobación de créditos y préstamos: Las instituciones bancarias necesitan saber si sus clientes les van a devolver el dinero prestado y para tener una predicción fiable de este extremo, se recolecta la

información de distintas fuentes, tales como historiales de transacciones bancarias. Con esto logran predecir las posibilidades de que este pague convenientemente.

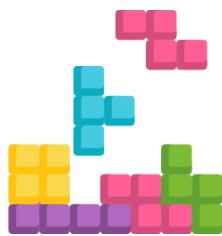
- Segmentación de mercado: Durante el proceso de compra los clientes, sin ser conscientes, generan muchos datos que pueden ser analizados. En sus históricos de compra existe mucha información sobre las preferencias de cada persona y sus patrones de consumo. Supongamos que una tienda quiere incrementar la venta de un libro determinado. En base al conocimiento de los historiales de compra de todos sus clientes podría segmentar el mercado y dirigirse a aquellos clientes que según su perfil presenten mayor probabilidad de compra, ya sea por sus gustos, género, edad, frecuencia de compra...
- Detección de fraude: La detección de fraude es una tarea sin duda complicada y puede ser realizada estudiando patrones que se desvén de lo esperado. Con la ayuda del data mining podemos identificar este tipo de patrones.
- Mejorar el rendimiento comercial: Tiendas como Amazon utilizan el historial de compra para sugerir productos a los clientes. Es lo que se denomina venta cruzada o cross selling. También se practica la estrategia del up selling que es convencer al cliente para que compre un producto más caro y rentable para el vendedor. Haciendo uso del data mining se pueden predecir los posibles productos que serán más atractivos para ciertos clientes, con lo cual las acciones de cross y up selling serán más eficaces.
- Análisis de tendencias: es posible analizar las tendencias para preparar mejores estrategias en las ventas, por ejemplo, según tendencias de estación o cíclicas, detectando cuando las personas compramos un producto repetitivamente cada año. Todo ello ayuda a mejorar las estrategias comerciales.
- Diseño web: en el diseño web, el data mining nos puede ser de utilidad para hacer que nuestros clientes encuentren más fácilmente la información que buscan. Por ejemplo, contabilizando la cantidad de clics en ciertas secciones de una página web o monitorizando los pasos previos que cada usuario da para realizar una acción determinada, nos puede ofrecer información importante para organizar la arquitectura de la información en la página.



ENLACE DE INTERÉS

Una herramienta de analytics muy usada y gratuita es la de Google:

<https://analytics.google.com/>



EJEMPLO PRÁCTICO

Un banco está interesado en mejorar la rentabilidad de sus clientes. Para conseguirlo, considera que debe fidelizar a los mejores clientes desde que son muy jóvenes, para que así no quieran cambiarse de banco a lo largo de su vida cuando, por ejemplo, comienzan a ganar una nómina o necesitan un préstamo para comprarse su primer coche o una hipoteca para su vivienda

SOLUCIÓN

La recomendación para este banco es aplicar data mining para la construcción de un modelo predictivo que se base en observaciones previas y se construye un modelo sobre un el concepto de rentabilidad y fidelidad.

Por ejemplo, si los datos históricos nos informan de que los clientes que han estudiado determinadas carreras son más rentables, entonces esto se puede utilizar posteriormente para predecir el futuro comportamiento de nuevos elementos, es decir, de los nuevos clientes.

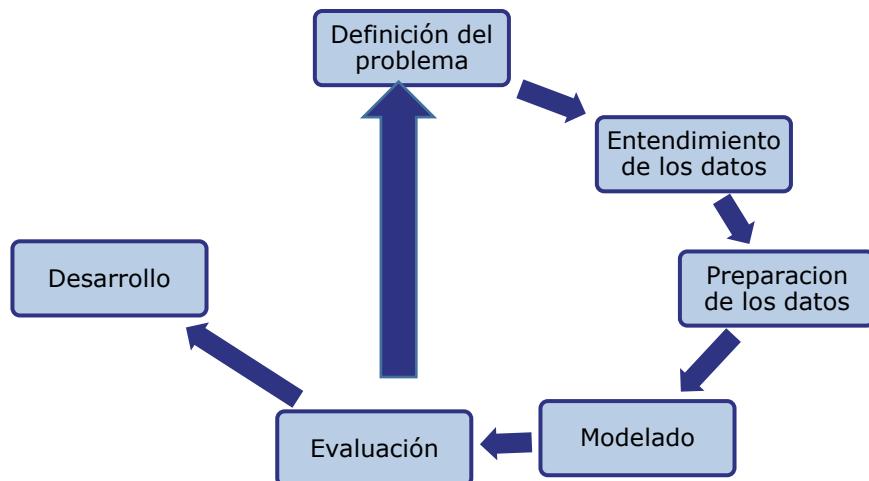
Por ejemplo, el modelo de cliente de alta rentabilidad nos ayudaría a determinar la política de captación de nuevos clientes al anticipar que clientes serán más rentables en el futuro.

De este modo, el banco podría tener comerciales específicos asignados a determinadas universidades o centros de formación para captar clientes que empiezan sus estudios ofreciéndoles ofertas para comprar un ordenador o un préstamo para pagar sus estudios.

Con este tipo de acciones se convertirán en clientes muy fieles que en el futuro cuando empiecen a trabajar e ingresar un dinero importante, seguirán fieles a la marca de nuestro banco.

4.3 Proceso del data mining

El proceso del data mining se realiza o se suele describir su aplicación en las siguientes seis fases debido a una metodología muy popular:



Proceso de data mining
Fuente: Elaboración propia

1. Definición del problema

La primera fase se enfoca en entender los requerimientos y el objetivo del proyecto. Después de que el proyecto esté definido, este se puede formular como un proyecto de data mining y se podrá hacer un planteamiento preliminar.



EJEMPLO PRÁCTICO

Supongamos que tenemos un problema que se pudiese describir con las siguientes palabras ¿cómo puedo vender más productos a mis clientes?

SOLUCIÓN

Pudiéramos traducir o convertir ese problema en un requerimiento de data mining que se puede expresar con las siguientes palabras ¿A cuáles de los clientes les gustaría comprar más cantidad de mi producto o más productos de mi catálogo?

Un modelo que prediga los compradores frecuentes de un producto puede ser construido utilizando los registros de las compras anteriores. Antes de esto, los datos deben ser organizados y estudiados para saber qué relaciones existen entre los clientes que compraron el producto y los que no, así como cuáles son las características que los hacen más proclives a comprar o no comprar. Las variables que se podrían estudiar serían edad, sexo, años de residencia, lugar de trabajo, número de hijos, etc.

2. Entendimiento de los datos

En la siguiente fase se comienza con la recolección de datos. En esta fase los datos son capturados desde distintas fuentes. Para la creación de un modelo exitoso, probablemente sea necesaria la carga de datos e integración de estos desde fuentes muy heterogéneas.

Tras ello se hace un análisis profundo para determinar qué datos son relevantes o no para resolver el problema planteado. De cualquier manera, sería posible añadir más datos o eliminar los que no sean de relevancia para el objetivo. En esta etapa también podríamos determinar los datos faltantes y proceder a completarlos. Ese proceso puede tomar diversas vías. Por ejemplo, si tenemos un atributo como fecha de nacimiento, pudiéramos cambiarlo al valor de edad simplemente realizando el cálculo. Todo ello con la idea de obtener más información enfocada en la resolución de nuestro planteamiento. En otros casos, la obtención de los datos no será ni mucho menos tan sencillo como la realización de una simple operación aritmética y deberemos buscar nuevas fuentes para ellos. En muchos casos, el coste económico de obtener estos datos puede volverse un porcentaje muy elevado de todo el proyecto.



ENLACE DE INTERÉS

Según Forbes, el blog más influyente en materia de data mining y análisis en general es el siguiente:

<https://blog.revolutionanalytics.com/>

3. Preparación de los datos

Esta fase nos ocupará, en un proyecto medio, el 90% de nuestro tiempo y consiste en la limpieza, formateado y selección de nuestros datos, para que estén listos para el análisis. Cuando se entrevista o se le pregunta a un analista o científico de datos es verdad que reconocen que su trabajo les apasiona y que es muy entretenido “buscar petróleo” en un océano de datos, pero igual de verdad es que reconocen que muchas veces no es una actividad tan emocionante como desde fuera la gente se imagina, dado que la mayor parte del tiempo se lo pasan realizando consultas de SQL a distintas bases de datos para obtener la materia prima sobre la cual trabajar posteriormente.

4. Modelado

En este momento aplicaremos los distintos métodos para construir modelos según sea el problema que se quiere resolver y poder ofrecer soluciones acordes al mismo.

5. Evaluación

En esta fase revisaremos los modelos generados para ver si satisfacen el propósito de nuestra investigación. Evaluaremos los modelos utilizando datos de entrenamiento y datos de prueba. Según como sea el resultado y su precisión, podríamos optar por volver a la fase previa para perfeccionar el modelo en caso de considerar que no es adecuado o confiable. Después de las iteraciones que sea necesario, si se demuestra cumplir con una precisión aceptable, podremos pasar a la siguiente fase.

6. Desarrollo

En esta fase se ofrece directamente la solución al problema, o se presenta el resultado del análisis que ayude a solucionarlo. Puede ser en forma de informe o algún curso de acción, tal como capturar nuevos datos que llevarían a un nuevo paso cada vez más cercano de la solución al planteamiento inicial.

4.4 Técnicas para data mining

Podemos clasificar las técnicas usadas en data mining en estas cuatro categorías:

1. Modelado predictivo
2. Segmentación de bases de datos
3. Análisis de vínculos
4. Detección de desviación



EJEMPLO PRÁCTICO

Una compañía de alta tecnología que requiere de trabajadores muy cualificados y especializados, está teniendo dificultades para localizar en el mercado candidatos a los que pueda contratar, lo que le está provocando pérdidas económicas y retrasos en los proyectos que tiene en marcha.

Como los perfiles que la empresa demanda están muy demandados tanto en España como en el extranjero, no se preocupan de apuntarse a las ofertas de empleo ni de enviar CVs por lo que es necesario ir a buscarlos de forma proactiva.

SOLUCIÓN

Desde el departamento de recursos humanos proponen disponer de una herramienta de minería de datos porque creen interesante poder capturar los miles de documentos que se publican cada día en revistas científicas, patentes, actas de congresos y restante literatura científica.

De este modo se podría analizar de forma masiva todos los conceptos y palabras clave que aparecen en ese material bibliográfico y localizar científicos a los que se podrían dirigir de forma directa para su contratación.

Sería un proceso de "head hunting data driven", dado que al requerir la compañía de perfiles muy concretos y de alta especialización, las redes tradicionales de contratación como portales de empleo o empresas de selección no resultan eficientes.

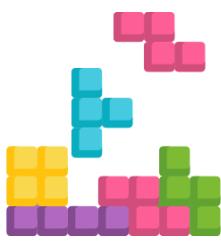
4.4.1 Modelado predictivo

Este modelado está enfocado a predecir la salida de un evento y está basado en un comportamiento similar al aprendizaje humano, mediante la observación de las características más importantes de una tarea. Es

desarrollado haciendo uso del aprendizaje supervisado. Para un modelo predictivo tendremos ya ciertos datos identificados en función de la predicción posible que pretende obtenerse. Puede hacerse de dos maneras, que es usando clasificación o aplicando regresión.

4.4.2 Segmentación de la base de datos

Esta es una técnica que se fundamenta en la aplicación del clustering, que es un tipo de aprendizaje no supervisado y que, como ya sabemos, produce una segmentación en grupos dadas unas características o relaciones comunes. Se usa bastante en la segmentación de grupos de clientes o para la realización de marketing y ventas directas, donde es muy interesante descubrir grupos de clientes que comparten determinados parámetros para dirigirnos a ellos con una oferta personalizada.



EJEMPLO PRÁCTICO

Unos grandes almacenes están viendo como la nueva competencia online le está robando gran parte de su cliente y quieren poner en marcha una iniciativa tecnológica para poder revertir la situación.

SOLUCIÓN

Una posibilidad sería tomar la base de datos de clientes históricos de la empresa, es decir, de todas aquellas personas que en algún momento han confiado en esos grandes almacenes y segmentar dicha base de datos.

Este proceso se podría realizar mediante aprendizaje no supervisado para encontrar grupos de clientes que comparten características comunes.

De este modo, el departamento de marketing podría decidir cuáles de esos grupos son más interesantes y desarrollar una estrategia de marketing personalizado para ellos, como podría ser ofrecerles financiación a aquellos clientes con menores ingresos y que por eso han dejado de comprar o una garantía adicional para aquellos clientes más conservadores.

4.4.3 Análisis de vínculos

Esta técnica consiste en el establecimiento de relaciones entre registros individuales o grupos de estos.

Esta técnica se divide en tres enfoques:

1. **Descubrimiento de asociaciones:** se trata de localizar elementos que impliquen la presencia de otros elementos. Un ejemplo de este tipo de asociación podría ser que, si un cliente alquila una propiedad por más de 2 años y es mayor de 25 años, en el 40% de los casos este mismo cliente comprará una propiedad. Está claro que un análisis de este tipo puede aportar buenos resultados a cualquier inmobiliaria, porque si es capaz de predecir cuales de los clientes de alquiler es más frecuente que se conviertan en clientes de compra, podrán realizar un marketing especializado.
2. **Descubrimiento de patrones secuenciales:** busca la presencia de eventos relacionados, es decir, una colección de elementos es seguida por otra en el transcurso de un periodo de tiempo.
3. **Descubrimiento de secuencias de tiempo:** esto es cuando estudiamos las relaciones entre ciertos elementos que se ven condicionadas por un periodo de tiempo. Por ejemplo, una persona que acaba de comprar una propiedad, durante los primeros tres meses tras la compra, adquirirá artículos como nevera, lavadora, etc.

4.4.4 Detección de desviación

Como su nombre indica, nos fijaremos en los datos que se salgan de la norma o de lo esperado. Nos valdremos de la estadística y de herramientas de visualización para lograr detectar todos estos valores que se acercan, anormalmente, a los márgenes extremos en cuanto a datos esperados.

Estas detecciones de desviaciones tienen aplicaciones múltiples. Por ejemplo, en calidad permiten descubrir partidas de productos defectuosos o en investigación del fraude detectar movimientos extraños o en prevención del blanqueo de capitales, identificar circunstancias irregulares.

4.5 Diferencias entre ML y data mining

El data mining se refiere a la extracción de conocimiento de una gran cantidad de datos. Es un proceso donde se descubren nuevos patrones relacionados con los datos y es repetitivo en dicha búsqueda de patrones nuevos para la toma de decisiones.

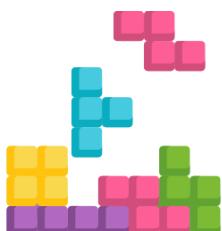
El data mining se podría decir que es una subclasificación del análisis de datos, similar a la investigación experimental. Dos elementos son necesarios para implementar el data mining, que son los datos y el ML.

El data mining es un proceso que incluye el entendimiento de los datos, su procesamiento y el modelado de los mismos, mientras que el ML toma los datos procesados como valor de entrada y formula predicciones. Por otra parte, el data mining conlleva un proceso de limpieza de los datos bajo la supervisión humana, mientras que en el machine learning la intervención humana se centra en definir el algoritmo.

	Data Mining	Machine Learning
Definición	Conlleva la extracción de información útil entre una gran cantidad de datos	Introduce nuevos algoritmos basados en los datos y la experiencia pasada
Historia	Creado en el año 1930. Inicialmente llamado descubrimiento en bases de datos	Fue creado en el año 1959
Responsabilidad	Es usado para determinar patrones en datos existentes. También puede ser utilizado para establecer reglas nuevas	ML enseña a los ordenadores como entender y aplicar reglas
Naturaleza	Involucra la interacción humana	Es automatizado cuando se implementa y requiere muy poca intervención humana

Tabla comparativa data mining y machine learning

Fuente: Elaboración propia



EJEMPLO PRÁCTICO

Una universidad quiere utilizar la tecnología para poder conseguir más financiación y fondos para poder desarrollarse, puesto que con la crisis económica y la bajada de natalidad cada vez es más difícil sobrevivir exclusivamente mediante el cobro de matrículas a los alumnos.

¿Cómo puede la tecnología ayudar a ello?

SOLUCIÓN

La búsqueda de fuentes de financiación alternativas en las universidades es una tendencia en alza en España. Cada vez son más los mecanismos que los centros universitarios ponen en marcha para conseguir fondos distintos a los cobros por matrículas.

Por este motivo, sería interesante realizar un análisis de big data sobre todos los C.V. y publicaciones de los profesores e investigadores en la universidad. La herramienta a aplicar debería ser minería de datos, específicamente minería de textos.

Hacerlo de manera manual sería inviable, pero automatizadamente se podría crear un mapa de las capacidades y competencias que suma todo el personal de la universidad para, posteriormente, identificar en el mercado a compañías potencialmente interesadas en pagar por esas capacidades, sea en forma de contratación profesional, encargos de I+D, asesorías, etc.

5. IMPLANTACIÓN DE UN PROYECTO DE BIG DATA

Para implantar un proyecto de Big Data, deberás decidir qué herramientas se van a utilizar. Pero aparte de la tecnología, no debes olvidar la parte metodológica que todo proyecto informático debe tener para que las herramientas sean usadas de forma coherente.

Haciendo esta reflexión te has dado cuenta de que las prisas os han hecho olvidaros en cierta medida de la materia prima, que son los datos. Es decir, aunque Weka o R u otro software que quieras elegir, puedan ser las herramientas más apropiadas para vuestros objetivos, si los datos disponibles no son suficientes o son incorrectos, o presentan déficits, los análisis no resultarán bien.

Así que os toca sumergiros en las distintas bases de datos, repositorios, logs y cualquier otra fuente de datos disponible dentro de la empresa o que esté a vuestro alcance en el exterior para comenzar a verificar la fiabilidad, idoneidad y calidad de los datos.

Por suerte, con un software como Weka, os será más sencillo desarrollar tareas de preprocesamiento de datos que los preparen para una mejor explotación posterior.

El procesamiento de datos antes de utilizarlos es imprescindible en casi todas las ocasiones. La consolidación e integración de distintas fuentes de datos, se podría decir que es una de las tareas que conllevan más trabajo, y si a esto le sumamos los fallos que puedan tener los datos, tales como errores de grabación o transcripción, incongruencias e inconsistencias, carencias de lógica, etc., provoca la necesidad de la limpieza de estos datos para que nuestro estudio y análisis sea coherente y útil.

Para hacer esto nos valdremos de herramientas y métodos que explicaremos a continuación, tales como la limpieza, integración, reducción y transformación de los datos.

Tras este preprocesamiento obtendremos datos confiables y entendibles para un exitoso proceso de data mining.



SABÍAS QUE...

La filial de datos del Banco BBVA es una de las empresas que más proyectos avanzados de Big Data desarrolla.

En su web comparten muchas de las experiencias reales que han tenido lo que es un gran modo de aprendizaje:

<https://www.bbvdadata.com/es/>

5.1 Consolidación de datos

Para entender esto de una manera correcta, estudiemos el siguiente ejemplo. Supongamos que tenemos los datos de una empresa que posee múltiples sucursales, la cual vende distintos tipos de productos a lo largo de todo el territorio. Esta compañía gestiona inventarios en cada sucursal y, debido a errores por parte de los empleados, existen diversos productos que son los mismos, pero presentan distintos códigos o nomenclaturas.

Se requiere hacer un minado efectivo de datos para lograr establecer cuando se deben reponer las existencias. Pero como vemos, esto es un error fatal, ya que no se puede confiar en los datos existentes. Vamos a analizar qué es lo que debemos hacer para limpiar estos datos. Estas recomendaciones también pueden servir para los registros de los clientes y, en general, podremos convertir cualquier dato en información realmente útil.

Veamos el siguiente registro:

Sucursal 1

Id	0005
Nombre	Abono
Marca	X
Existencia	50
Fecha	01/01/20

Sucursal 2

Id	00005
Nombre	Fertilizante
Marca	X
Existencia	2
Fecha	01/02/20

Sucursal central

Id	0005
Nombre	Abono
Marca	X
Existencia	50
Fecha	01/01/20

Id	00005
Nombre	Fertilizante
Marca	X
Existencia	2
Fecha	01/01/20

Como hemos podido observar, la sucursal 1 y 2 tienen el mismo producto con distinto nombre y con errores en el código. Por ende, en la sucursal central se genera una duplicidad de datos y se ven como si fueran dos productos distintos.

Sin embargo, en la sucursal central debería haber un registro así:

Id	0005
Nombre	Fertilizante o Abono
Marca	X
Existencia	52
Fecha	01/01/20

Para situaciones como esta, existe el preprocesamiento de los datos. Es un proceso complejo que forma parte de las tres conocidas fases ETL (Extraction, Transformation y Loading) en español extracción, transformación y carga de los datos.

Como hemos podido observar, los datos en bruto tienen gran cantidad de debilidades e inconsistencias que hay que eliminar o, por lo menos, minimizar. Para ello nos valdremos de métodos como el de limpieza de datos.

Este método se encarga de disminuir el ruido en los datos, en escenarios tales como:

- Falta de valores
- Ruido en los datos
- Inconsistencia en los datos
- Integración de datos

- Transformación de los datos
- Normalización
- Estandarización
- Reducción de datos
- Reducción de dimensión
- Reducción de volumen
- Compresión de datos

Falta de valores

Es muy común encontrarse con registros incompletos y a medio completar. La mejor manera para abordar estos casos sería:

1. Cumplimentar manualmente los datos faltantes. Pero esto solo será de utilidad cuando el volumen de los datos no sea muy grande.
2. Usar alguna constante global para llenar este valor. Aquí remplazaríamos el valor con una constante de tipo "desconocido", lo que nos permitirá estudiar también este factor. Este tipo de solución debe ser aplicada con precaución.
3. Usar la media global para llenar ese atributo: Puede ser una buena alternativa sacar una media global de ese atributo, lo que nos permitirá hacer uso del registro, suponiendo que se encuentra entre los valores normales.
4. Usar un valor pronosticado con alta probabilidad: Esta es otra buena alternativa y se puede lograr haciendo uso del formalismo bayesiano o de un árbol de decisión inducido. Consiste en predecir el valor con la ayuda de los datos que ya están.
5. Ignorar ese valor: Esto sería una buena opción cuando el registro posee más de un valor faltante, pero no es recomendable cuando se trata de un porcentaje alto de registros que tienen este problema.

En conclusión, podríamos decir que el uso de un valor promedio es la opción más usada de forma general. Pero también existen opciones probabilísticas, para lograr predecir el valor.

Ruido en los datos

El ruido en los datos viene determinado por esos valores que se comportan de manera aleatoria y no siguen ningún patrón. La reducción de este genera

un patrón más suave en las variables que queramos estudiar. Los métodos más comunes para la reducción del ruido son:

- **Métodos de agrupamiento:** Esto consiste en la conversión de los valores de los atributos en valores discretos, es decir, supongamos que tenemos varios valores numéricos, por ejemplo, las mediciones de la temperatura, los convertiríamos a valores nominales, tales como caliente, tibio, fresco, frío. Esto lo podemos realizar de dos maneras:
 - Calculando el tamaño de los intervalos donde entrarán los valores equitativamente.
 - La otra forma sería dividir los intervalos según la frecuencia de los valores. Este último modo es bueno, ya que nos garantiza que todos los intervalos tendrán una frecuencia de datos equivalente.

Veamos un ejemplo simple. Tenemos un grupo de 50 estudiantes con sus respectivas notas. Haciéndolo de la primera manera, podríamos dividir en 10 intervalos según las notas 0-10, 10-20, [...] ,90-100. La mayoría de los estudiantes estarían entre 60 y 80 y los otros intervalos solamente tendrían unos pocos y algunos, incluso, estarán vacíos.

En cambio, si dividimos basándonos en la frecuencia de estos, lo primero que haríamos sería ordenar los datos de manera descendiente o ascendente. Después revisaríamos cada registro y en función de la cantidad de elementos, crearíamos los intervalos.

- **Análisis atípico** o clustering: Aquí deberemos agrupar los valores que sean similares. Cada uno de estos grupos será denominado un cluster y los valores que queden fuera de estos serán llamados atípicos o ruido.
- **Regresión:** Este es otro método que nos será de bastante utilidad. Por lo general se usa la regresión lineal, que nos permitirá situar el valor apropiado entre dos valores.
- **Inspección humana** e informática: Aquí estaremos haciendo uso de la inspección manual y automatizada para detectar el ruido.

Inconsistencia en los datos

Esta se puede deber básicamente a un error durante el proceso de captura de datos, o a la presencia de datos duplicados. En este caso no existe otra manera de corregirlo más que hacerlo a mano, en la medida de lo posible.

Integración de datos

La integración de fuentes de datos es algo a lo que tendremos que habituarnos, ya que es muy frecuente a la hora de hacer un preprocesado de datos. La integración puede tener diversos orígenes, ya sean diversas bases de datos, archivos de texto plano o cubos de datos. Dentro del proceso de integración de datos, debemos contemplar las siguientes premisas:

- Evitar la redundancia de datos: esta ocurre cuando la misma información es referida por distintos registros en la base de datos, por ejemplo, productos repetidos con nombres similares.
- El uso de metadatos: que no son más que datos sobre los datos. Esto quiere decir que estudiaremos los datos y también cualquier otra información acerca de los mismos.

Transformación de los datos

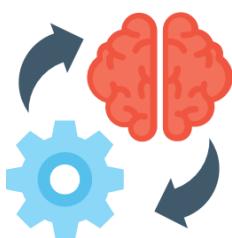
En general, nuestro trabajo será hacer que los datos sean de utilidad para el proceso de extracción de información o data mining. En muchas ocasiones tenemos datos que no aportan gran cosa, debido a la magnitud de su valor. Para manejar eso, tendremos dos procesos: la normalización y la estandarización.

Normalización

La normalización consiste en convertir los valores de cierto atributo en un rango o en un puntaje (0,1). Supongamos que tenemos el rango de los estudiantes con notas entre 35 y 45 sobre 100. El 35 será considerado 0 y un 45 será considerado 1. Y clasificaremos a los estudiantes con ese criterio de 1 y 0. Pero si un estudiante obtiene 90, sería un caso aislado y en ese caso, entonces 35 sería considerado 1 y 35 sería 0. Ello provocaría que muchos otros valores bajaran a 0. Para este tipo de casos sería mejor usar estandarización.

Estandarización

Cuando hablamos de estandarización, los valores serán extendidos con un valor de desviación estándar de 1. Es decir, serán los valores esperados probabilísticamente, pero si tenemos muchos casos aislados, lo mejor es usar normalización.



RECUERDA

La desviación estándar es un valor que nos indica la dispersión de los datos. La desviación estándar baja indica que la mayor parte de los datos de una muestra tienden a estar agrupados cerca de su media

Reducción de datos

La reducción de los datos viene al rescate cuando la cantidad de datos es tan grande que un análisis completo de los mismos tardaría demasiado tiempo en procesarse. Claro está que la reducción de datos se basa en técnicas para evitar la pérdida de precisión o integridad de nuestros datos.

La reducción de datos se refiere a la reducción de dimensiones o volumen.

Reducción de dimensión

En un almacén de datos, las dimensiones vienen a representar ciertos aspectos de nuestros datos. Estas no son necesarias para realizar todos los estudios, es decir, en algunos casos solo necesitaremos el uso de ciertas dimensiones. Con el uso de algoritmos podemos detectar información redundante o dimensiones innecesarias.

Reducción de volumen

Con esto nos referimos al estudio de los datos en formas de menor tamaño, que sean suficientemente representativas de los mismos datos.

Compresión de datos

Consiste en el uso de mecanismos tales como **Huffman coding** para reducir el tamaño de los datos.



ENLACE DE INTERÉS

Como se ha visto, la fiabilidad de los datos es fundamental.

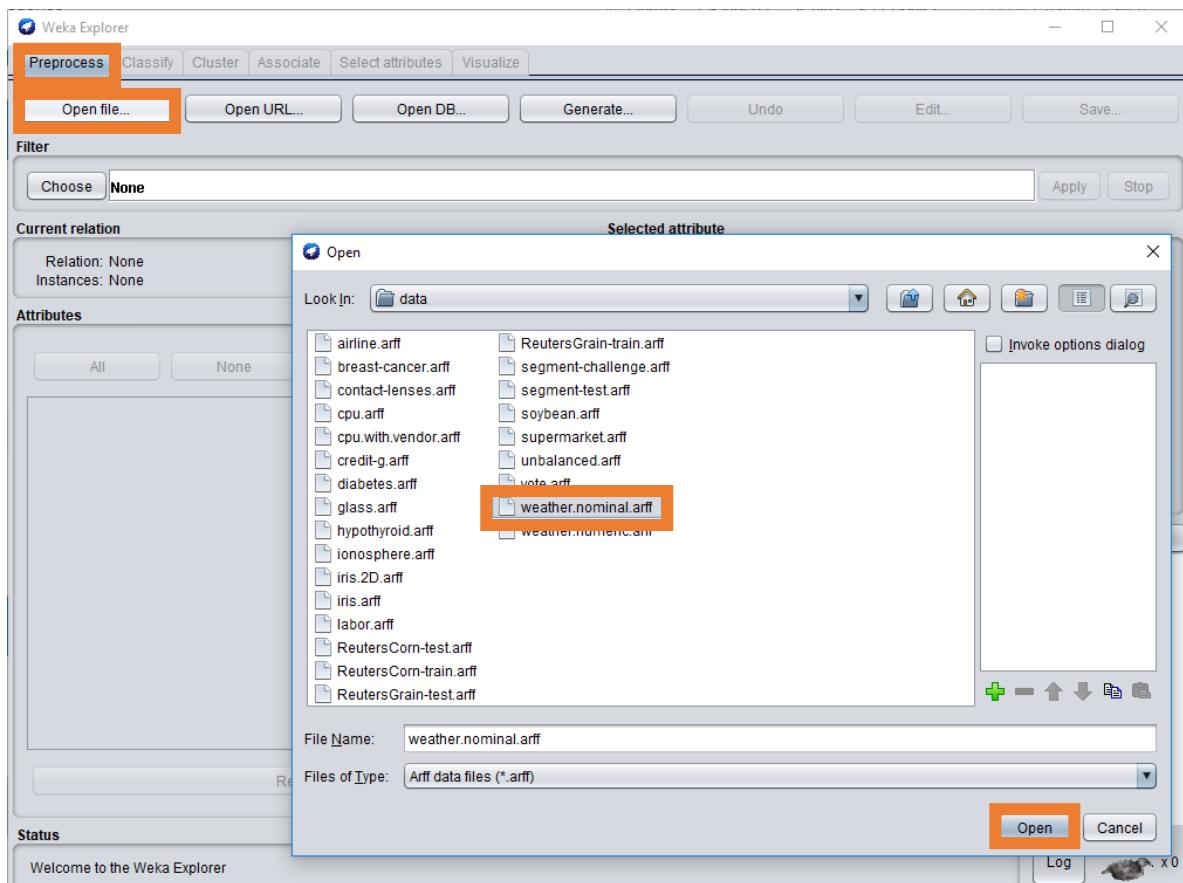
La siguiente web de IBM presenta sus avances para fomentar la fiabilidad de los datos

<https://www.ibm.com/es-es/analytics/dataops>

5.2 Preprocesamiento de datos con WEKA

Ahora mostraremos algunas de las posibilidades que nos aporta WEKA a la hora de hacer el preprocesado de datos.

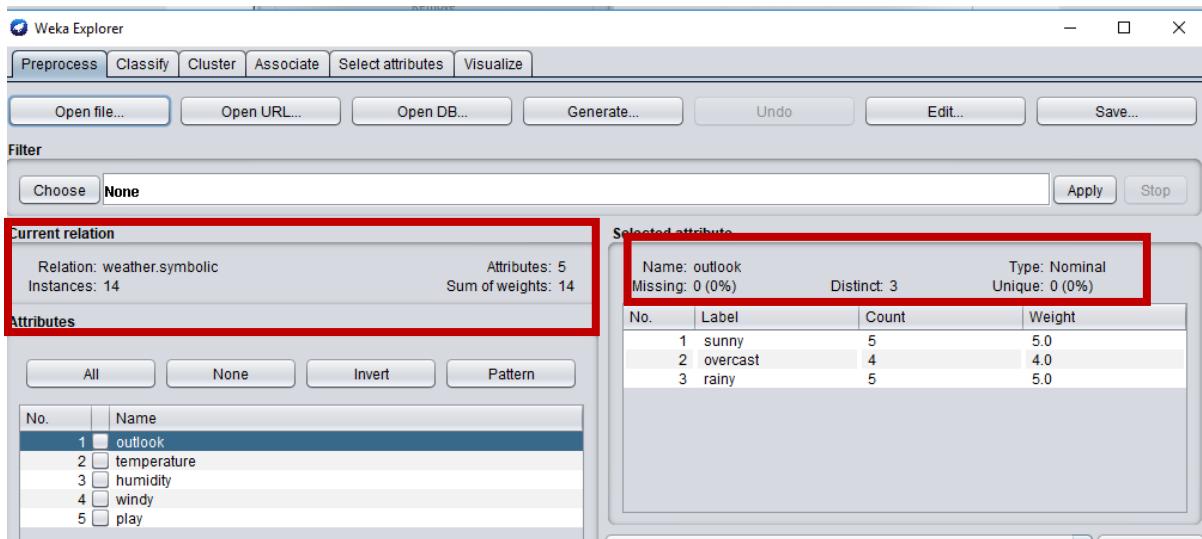
Abriremos WEKA y vamos a la pantalla de Explorer y en la pestaña de preprocess abriremos el archivo weather.nominal.arff (este se encuentra en la carpeta donde se instaló WEKA dentro de la carpeta data)



Explorer de Weka
Fuente: Elaboración propia

Tras esto veremos nuestros datos cargados. Ahora entendamos que en estos datos se nos muestran los siguientes elementos:

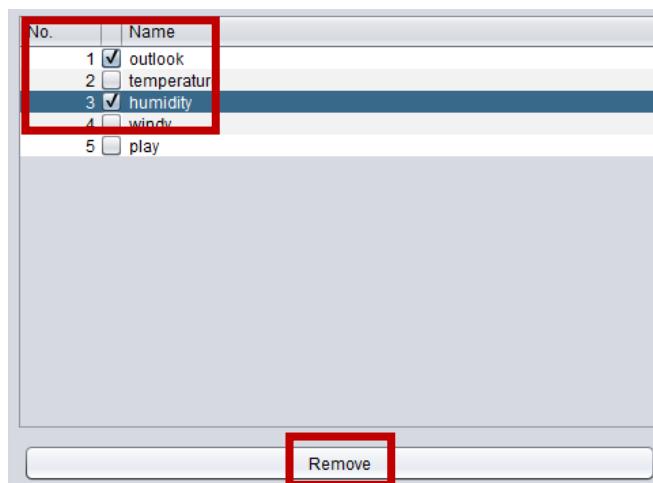
- Tenemos 5 atributos: Outlook, temperature, humidity, windy y play.
- Tenemos 14 instancias, no tenemos valores faltantes, ni valores únicos y las variables son tipo nominal.



Explorer de Weka

Fuente: Elaboración propia

Ahora, para este ejemplo, si no quisiéramos usar las variables outlook y humidity, las podríamos eliminar de la siguiente manera. Las seleccionamos y pulsamos el botón de **remove**:



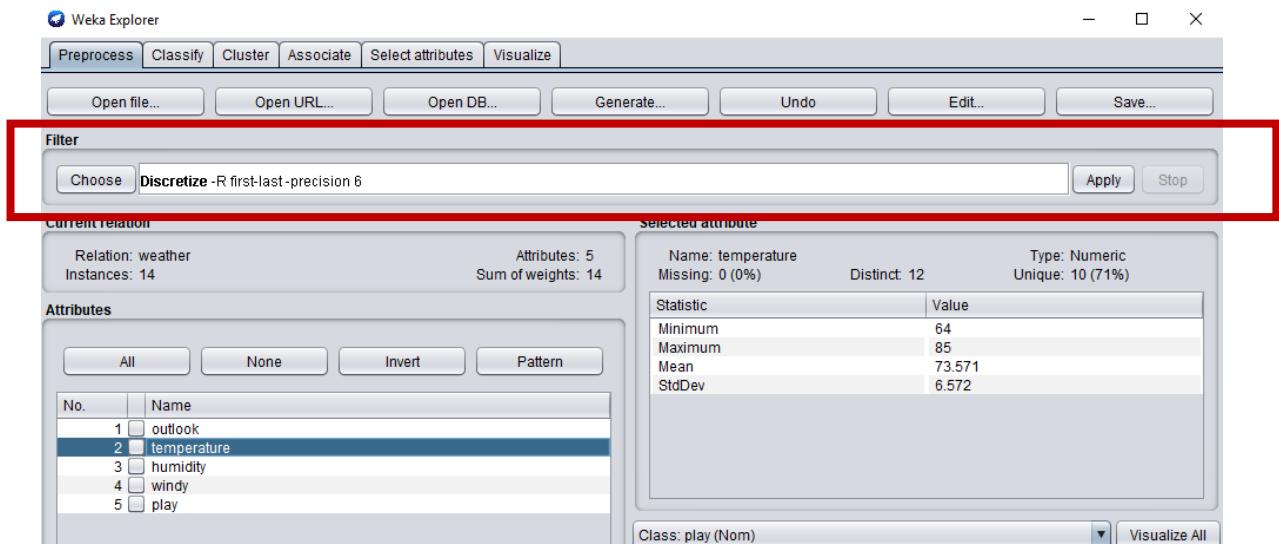
Explorer de Weka

Fuente: Elaboración propia

Otra manera de modificar los datos sería transformar datos numéricos en discretos, una estrategia sobre la cual ya hemos discutido bastante a lo largo de este apartado. Para hacer esto en WEKA tenemos los filtros, así que veamos cómo usarlos.

Vamos a la pantalla de Explorer y en la pestaña de preprocess abriremos el archivo weather.numeric.arff (este, de nuevo, se encuentra en la carpeta donde se instaló el WEKA, dentro de la carpeta data)

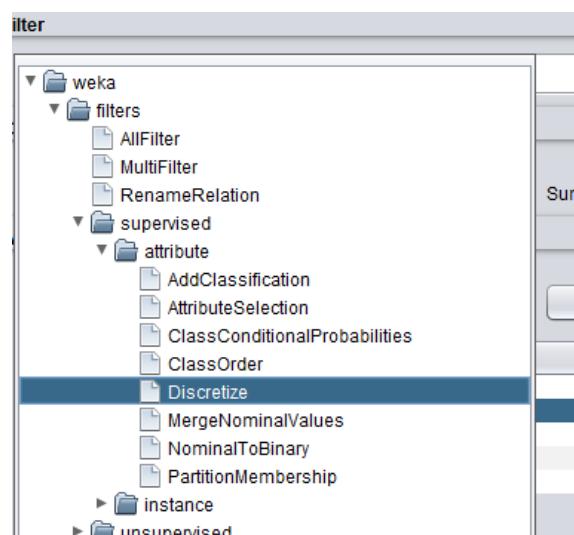
Como podemos observar, este es un poco distinto al otro porque tenemos variables de tipo numérico, que queremos convertir a nominales. Para ello haremos uso de la sección filter:



Explorer de Weka
Fuente: Elaboración propia

En el botón **choose** abriremos el menú y buscaremos la opción:

Weka->filter->supervised->attribute->discretize



Explorer de Weka
Fuente: Elaboración propia

Tras esto, hacemos clic en el botón **apply** y veremos que nuestros atributos numéricos se han vuelto nominales.

En los próximos apartados veremos con más detalle el uso de estas herramientas.

6. CUSTOMER ANALYTICS

No está de más, ahora que vais a comenzar una nueva era en el departamento de informática, con un enfoque hacia el analytics, que comencéis a dar importancia a todas las herramientas que nos ofrece la estadística.

Al fin y al cabo, por debajo de las herramientas como Weka o el machine learning en general, lo que subyace es una gran cantidad de instrumentos estadísticos.

Aquí se pueden dar varias circunstancias. Podría ser que la mejor opción, en caso de que el equipo de informáticos no disponga de ningún perfil adecuado, consista en contratar a un científico de datos que integre la visión del estadístico, con la del programador con la de experto en análisis de negocio.

Pero también es posible que, dado que las herramientas de software ayudan mucho a la aplicación de las técnicas de la estadística que no os resulte muy complejo a vosotros mismos mejorar vuestros conocimientos en la materia para estar más preparados a la hora de realizar los análisis en curso y poder interpretarlos adecuadamente.

Customer analytics es un área de trabajo tremadamente amplia, porque todo lo que tenga que ver con aplicar análisis e inteligencia a los clientes se engloba dentro de este epígrafe.

La piedra angular del customer analytics en una empresa suele ser el CRM (customer relationship management), puesto que es el repositorio que contiene todos los datos disponibles sobre los clientes para su análisis.

Identificar los principales tipos de clientes (segmentos) que resultan más representativos también es una actividad muy importante porque, al fin y al cabo, todos los esfuerzos de la compañía están destinados a captar su atención: es vital conocer quiénes consumen los productos y servicios del sector.

Vamos a introducirnos en los fundamentos de la estadística, porque es una herramienta muy poderosa e imprescindible para el desarrollo de análisis de todo tipo, incluyendo aquellos en el plano de los clientes y la empresa en general.

En primer lugar, conoceremos algunas de las **áreas de la empresa** donde la estadística ha apoyado de forma relevante el avance de las capacidades analíticas.

Algunas de las principales cuestiones que el customer analytics aspira a responder, son las siguientes:

- Segmentos de clientes, cuotas de mercado y tasas de crecimiento
- Actitud y fidelidad de los clientes, su posicionamiento en el mercado y quiénes son sus principales clientes
- Política de marcas y marcas principales, así como la valoración de las marcas por los clientes
- Política de fijación de precios, descuentos y promociones
- Política de canales de distribución (directos e indirectos), acuerdos de distribución y cobertura geográfica
- Estrategia de comunicación y publicidad
- Naturaleza y tamaño de la fuerza de ventas y modelo de retribución de los comerciales

Continuaremos con el estudio de los **principales parámetros y valores de la estadística descriptiva**, que permite explicar un fenómeno determinado.

Muy vinculada a la estadística se encuentra la **probabilidad**, que será otro campo de estudio y que permite calcular la certidumbre de un evento. Conoceremos también los árboles de probabilidad, una interesante herramienta a la hora de analizar situaciones complejas y para la realización de experimentos.

Finalmente, nos introducimos en la **inferencia**, que es el proceso por el cual, a partir de unas premisas dadas, se derivan unas conclusiones.

La estadística y la probabilidad se desarrolla a partir del siglo XVII de la mano de matemáticos como fueron Fermat y Pascal, que se preocuparon de resolver cuestiones llamativas sobre distintos juegos de azar.

Ya es en el reciente siglo XX que la matemática detrás de estos fenómenos se normaliza y se establecen los axiomas, las definiciones y los teoremas correspondientes. Muy pronto su aplicación práctica empieza a fructificar y la economía, la administración de empresas, la ingeniería y otras ciencias comienzan a utilizar este instrumental.

Bien es verdad que la primera estadística básica es anterior al siglo XVII, desde el punto de vista de una herramienta para recolectar, organizar y

presentar los datos en forma de tabla o bien en forma de gráfico. Pero también es verdad es esta visión simple, meramente descriptiva de la estadística, se potencia de forma espectacular con los avances en materia de probabilidad, porque gracias a ello la estadística tomó la capacidad de no solo describir, si no también realizar conclusiones y dar soporte a la toma de decisiones, por ejemplo, mediante el uso de la teoría del muestreo y de la predicción.

Prácticamente todas las áreas de la economía y de la empresa se benefician del uso de técnicas estadísticas. Además, en los últimos tiempos, gracias a la generalización de las técnicas de análisis de Big Data e inteligencia artificial, la estadística ha tomado más valor incluso.

Analizaremos, brevemente, a continuación, dos de los ámbitos donde más frecuente y tradicional ha sido su aplicación.



ENLACE DE INTERÉS

Probablemente la empresa que en la actualidad tiene una oferta de software para customer analytics más amplia y difundida es Zendesk.

<https://www.zendesk.es/>

6.1 Marketing

El análisis de los datos nos permitirá conocer mejor la evolución de las ventas, las necesidades de los clientes, posibles problemas en la distribución de los productos... Para poder interpretar la información necesitaremos tener ciertos conocimientos de estadística.

La estadística es una ciencia que, a partir de un conjunto de datos recogidos y organizados de una muestra concreta de una población, trata de sacar conclusiones para dicha población a partir del cálculo de probabilidades.

Dicho de otra manera, gracias a la estadística, por ejemplo, al estudiar los **hábitos de compra de un grupo de clientes** podremos inferir conclusiones acerca de sus **futuras compras** o de las futuras compras de otros clientes que formen parte de su **mismo segmento** (misma edad, sexo, ubicación...).

Cuando estudiamos a todos elementos y no solo a un subconjunto, se trata de **estadística descriptiva**. Por ejemplo, si elaboramos estadísticas con los datos de todos los clientes de una empresa para conocer qué franjas de edad

consumen más los productos comercializados, estaremos ante un caso de estadística descriptiva.

Mientras que si, a través de la estadística, sacamos conclusiones para toda la población a partir del estudio de un segmento, hablamos de **estadística inferencial**. Siguiendo con el ejemplo, si hacemos un sondeo entre algunos de los clientes para conocer su percepción sobre la calidad de los productos y servicios, estaremos ante un caso de estadística inferencial, porque de la percepción de algunos clientes extrapolamos conclusiones aplicables a todos ellos.

Las herramientas más utilizadas para el cálculo de valores estadísticos en la economía y la empresa son:

- IBM SPSS. Este programa es muy utilizado en el ámbito de la investigación de mercados. Tiene su origen a finales de los años 60 y su éxito se debe a que permite gestionar grandes cantidades de datos a través de una interfaz clara y sencilla. Para analizar los datos primero hay que introducirlos organizados en el programa, después se seleccionan las variables que interesen con las que se podrán calcular los valores medios, las desviaciones estándar, etc.
- Statgraphics. Se trata de un software de análisis estadístico creado a principios de los años 80 por Neil Polhemus y se utiliza habitualmente en las organizaciones que utilizan la metodología de mejora de procesos Seis Sigma. Este programa permite representar gráficamente todo tipo de estadísticas, así como facilita la estimación del futuro comportamiento de los valores.
- Minitab. Se trata de una versión más simple del programa de análisis creado por el Instituto Nacional de Estándares y Tecnología (NIST) de los Estados Unidos. Se usa habitualmente para analizar datos de encuestas. Es una herramienta muy visual que permite hacer predicciones a partir de tabulaciones cruzadas.
- SPLUS. Esta es una aplicación más avanzada basada en un lenguaje de programación orientado a objetos.
- Microsoft Excel. Permite almacenar toda clase de datos a los que después podemos aplicar diferentes funciones para el análisis de valores estadísticos. La primera versión de Excel apareció en 1985 solo para ordenadores MAC, mientras que la primera versión para Windows es de 1987. Es la aplicación más utilizada por las pequeñas y medianas empresas.

6.2 La calidad

En los años 20, Walter A. Shewhart, que trabajaba en los laboratorios de la compañía telefónica Bell, desarrolló las primeras teorías para el control estadístico de la calidad, mediante la utilización de métodos estadísticos y de probabilidad aplicados a los problemas de calidad que presentaban los procesos de fabricación. El control estadístico de la calidad en los procesos parte de dos supuestos:

- Lo variables que podían llegar a ser los productos resultantes de un mismo proceso.
- Lo variables que podían ser procesos diferentes.

Por lo tanto, la calidad se traduce en la aplicación de herramientas y técnicas estadísticas que facilitan la detección de productos defectuosos. El control estadístico de la calidad es practicado sobre muestras representativas de los lotes de producción y no sobre todos los productos. El uso de muestras para verificar la calidad ofrece las siguientes ventajas:

- La inspección dura menos tiempo.
- El coste es menor que cuando se evalúan el 100% de los productos.

Durante esta etapa se integra de forma generalizada la calidad dentro de los departamentos de ingeniería y producción de las empresas, lo que conlleva la aparición de personal experto en el control de la calidad.

A medida que avanza el siglo XX, progresivamente se mejoran los métodos estadísticos para medir y mejorar la estabilidad de la producción en las factorías, reduciendo el porcentaje de defectos y cumpliendo las especificaciones de los diseños. Sin embargo, esta visión cambia al considerarse la calidad como un factor estratégico. Ya no se trata de una actividad inspectora, sino preventiva: planificar, diseñar, fijar objetivos, educar e implementar un proceso de mejora continua. La gestión estratégica de la calidad hace de ésta una fuente de ventaja competitiva que requiere del esfuerzo colectivo de todas las áreas y miembros de la organización.

Los sistemas de aseguramiento de la calidad surgen en el momento en el cual las empresas necesitan proporcionar confianza acerca de los requisitos de calidad de sus productos. La norma ISO 8402 define el aseguramiento de la calidad como un conjunto de actividades preestablecidas y sistematizadas, aplicadas al sistema de calidad, que ha sido demostrado que son necesarias para dar confianza adecuada de que un producto o servicio satisfará los requisitos para la calidad.

Existen modelos que ayudan a asegurar la calidad como las normas ISO 9000, que permiten -además de adoptar un modelo que asegure la calidad-, la certificación del modelo de calidad ante un organismo reconocido.

La calidad y la orientación al cliente no siempre han ido a la par. El concepto calidad, en origen, hace referencia a la calidad de los productos, lo que suponía un mayor control e inspección (y por lo tanto un aumento del precio de los productos). La superación de esta concepción se produce con la extensión del aseguramiento de la calidad, que convierte a la calidad en responsabilidad de toda la organización. La superación del aseguramiento de la calidad se produce gracias a la orientación al cliente y la gestión de la calidad total.

La calidad en este momento es sinónimo de ofrecer a los consumidores lo que desean. Además, implantar un sistema de calidad permite mejorar la imagen corporativa, ayuda a la función de marketing y aumenta el espíritu de equipo.



ENLACE DE INTERÉS

Para conocer más sobre las tendencias en calidad en Europa debes visitar la web de la EFQM que es la Fundación Europea para la gestión de la calidad

<https://www.efqm.org/>

6.3 Estadística descriptiva

La estadística descriptiva son las técnicas matemáticas y de representación gráfica que permiten describir y analizar un conjunto de datos. A diferencia de lo que ocurre con la estadística inferencial, no buscan conclusiones sobre la población origen de los datos.

El análisis estadístico permite abordar problemas que de otro modo sería imposible. Frecuentemente no se puede abordar la toma de datos de un conjunto de población grande, por ejemplo, de todos los individuos de un país. Por este motivo, la toma de datos de esa población se simplificará en una muestra. Por tanto, los conceptos básicos de la estadística descriptiva son:

- **Población**, que es el total de individuos que se analizan y se espera obtener conclusiones.
- **Individuo**, que es cada miembro o representante de la población.

- **Muestra**, que es el subconjunto que se toma entre la población y que para que tenga valor estadístico debe ser representativo.

La aplicación de estadística descriptiva requiere de analizar la variable de estudio. Esta variable estadística corresponde a una característica de los individuos. Por ejemplo, su edad, si hablamos de personas físicas.

Las variables estadísticas pueden ser:

- **Cualitativas**, también llamadas categóricas, que significa que no son expresables en modo numérico.
- **Cuantitativas** cuando si se expresan numéricamente. Las cuales, se dividen en:
 - **Discretas**, cuando no puede tomar valores intermedios. Por ejemplo, un individuo puede tener 1 o 2 hijos, pero no 1,5 hijos.
 - **Continuas**, cuando los valores válidos intermedios no están limitados. Por ejemplo, un individuo puede pesar 80 kg u 81 kg, pero también podrá pesar 80,45 kg.

Otro concepto estadístico básico son las **distribuciones de frecuencias**, representadas mediante tablas de frecuencias, que facilitan resumir datos estadísticos obtenidos en una muestra. Para manejar distribuciones muy amplias será necesario el apoyo de herramientas como calculadoras u ordenadores.

Las tablas de frecuencias representan la información contenida en una muestra de tamaño n , tal que (x_1, \dots, x_n) .

Cada valor que puede tomar una variable (cualitativa o cuantitativa discreta) se expresa como c_i , $i = 1, \dots, k$.

La frecuencia absoluta es el número de veces que aparece un determinado valor en un estudio estadístico. Se representa por f_i .

La suma de las frecuencias absolutas es igual al número total de datos, que se representa por N .

$$f_1 + f_2 + f_3 + \dots + f_n = N$$

Para indicar resumidamente estas sumas se utiliza la letra griega Σ , que se lee como sumatorio.

Por ejemplo, si en una empresa, los años de experiencia de cada trabajador son los siguientes: 5, 5, 4, 5, 1, 2, 1, 2, 2, 2

La frecuencia absoluta se representaría así

xi	fi
1	2
2	4
4	1
5	3

La frecuencia relativa será expresada como el cociente entre la frecuencia absoluta y la cantidad total de datos. Se expresa en % y se representa por ni.

xi	fi	ni
1	2	0,2
2	4	0,4
4	1	0,1
5	3	0,3
	10	1

Con variables cualitativas y también con variables cuantitativas discretas con pocos valores es factible determinar las modalidades de la variable. Por lo contrario, con variables cuantitativas continuas y también con variables cuantitativas discretas, pero con muchos valores es necesario definir un tipo de modalidad que son los **intervalos de clase**, de modo que se agrupan los valores en intervalos, de amplitud análoga, que llamamos **clases** y tendrán su frecuencia calculada.

Cada clase tendrá unos límites, que serán el límite inferior y el límite superior de la clase. Dados esos límites, las clases tendrán una amplitud, calculada como la diferencia entre el límite superior de la clase y el límite inferior de la clase.

Cada clase también tendrá una marca, que será el punto medio. Este punto medio representa al intervalo y será usado para determinados cálculos estadísticos.

Supongamos que los datos de experiencia laboral de un grupo de trabajadores corresponde a una muestra mucho más extensa, que requiere de la definición de clases, siendo estos valores: 2, 3, 2, 3, 1, 1, 2, 2, 1, 3, 4, 3, 2, 5, 4, 7, 9, 1, 1, 2, 1, 1, 2, 1, 5, 8, 6, 1, 2, 1, 2, 1, 1, 1, 2, 1, 2, 2, 1, 23.

En primer lugar, se identifican los valores extremos, que será 1 año de experiencia y 23 años de experiencia.

Se calcula la diferencia entre ellos, que es 22 y ahora debemos establecer un valor superior a 22, que al tiempo sea divisible entre el número de intervalos que se pretende definir. Si, por ejemplo, nos interesan 6 intervalos, podemos tomar el valor 24. Y procedemos a definir los intervalos, teniendo en cuenta que el límite inferior será parte de la clase, pero el límite superior pertenece a la clase siguiente.

	fi	Fi	ni	Ni
[1,5)	33	33	0,825	0,825
[5,9)	5	38	0,125	0,950
[9,13)	1	39	0,025	0,975
[13,17)	0	39	0	0,975
[17,21)	0	39	0	0,975
[21,25)	1	40	0,025	1

6.4 Representaciones gráficas

Según se trate de variables cualitativas, cuantitativas discretas o cuantitativas continuas, la visualización gráfica más adecuada será diferente.

Para representar variables cualitativas es común el uso del diagrama de barras o del diagrama de sectores. El eje horizontal del diagrama de barra representa las modalidades de la variable a representar y las barras tendrán altura proporcional a la frecuencia de la modalidad absoluta o relativa.

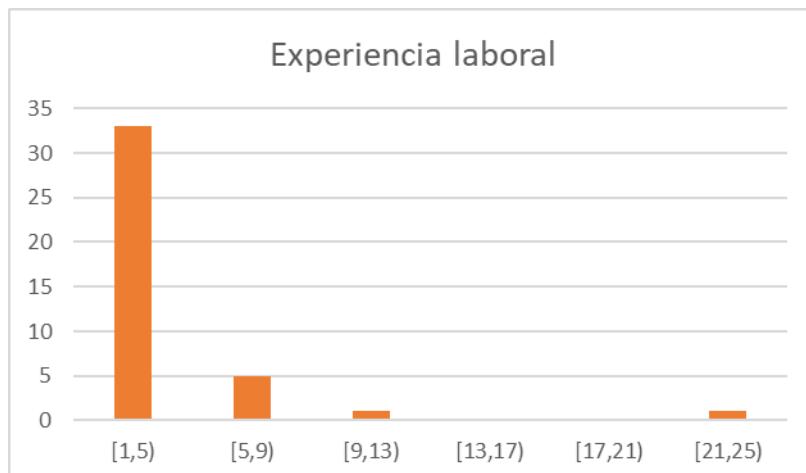


Diagrama de barra

Fuente: Elaboración propia

Los diagramas de sectores visualizan igualmente las distintas modalidades y la frecuencia, pero en forma circular, correspondiendo cada segmento de área a la frecuencia.



Diagrama de sectores

Fuente: Elaboración propia

Para variables cuantitativas discretas, además de los anteriores, también es común el uso de diagramas acumulativos de frecuencias. Si se utilizan frecuencias relativas acumuladas entonces el valor máximo del diagrama acumulativo será 1. Si se utilizan frecuencias absolutas acumuladas entonces el máximo coincide con el número de datos de la muestra.

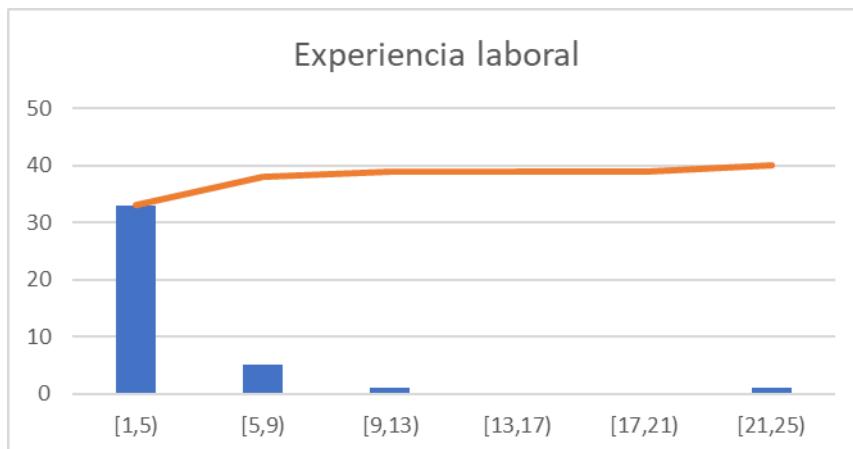


Diagrama acumulativo
Fuente: Elaboración propia



ENLACE DE INTERÉS

Knoema es una plataforma online gratuita donde encontrarás miles de datos y herramientas para llevar a cabo su visualización gráfica

<https://knoema.es/>

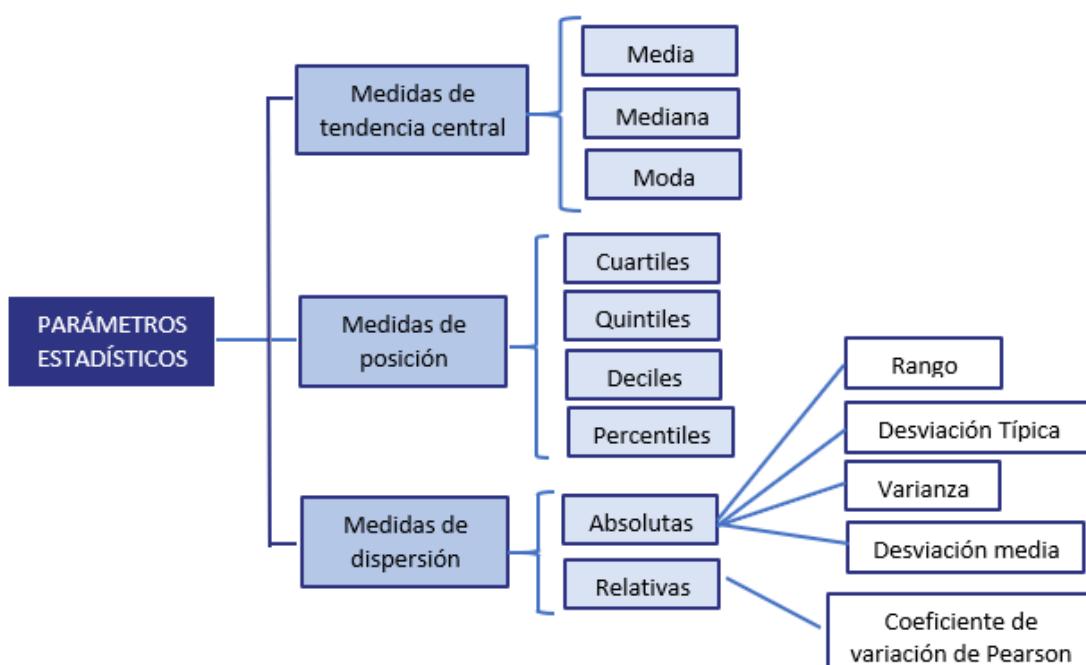
6.5 Parámetros estadísticos

Los parámetros estadísticos se obtienen de la distribución estadística. Son una **herramienta de síntesis**. Pueden ser:

- De **centralización**, también llamados medidas de tendencia central: es un valor respecto del cual se distribuyen los datos, siendo la media aritmética, la mediana y la moda los más utilizados.
- De **posición**, que dividen los datos en grupos con la misma cantidad de elementos, siendo los cuartiles (en cuatro partes), los quintiles (en cinco partes), los deciles (en diez partes) y los centiles (en cien partes) los más utilizados.
- De **dispersión**. Estos parámetros nos permiten saber cuánto se distancian del centro el conjunto de datos con los que trabajamos. Pueden ser medidas absolutas o relativas.

- Absolutas: Estas medidas se expresan en valores absolutos y serán los rangos, las desviaciones medias, la varianza y la desviación típica.
- Relativas: Estas medidas se expresan en forma de porcentajes. La más conocida es el coeficiente de variación de Pearson.

Resumimos en el siguiente esquema los principales parámetros de la estadística descriptiva.



Taxonomía de parámetros estadísticos

Fuente: Elaboración propia



ENLACE DE INTERÉS

Statista es un portal con millones de estadísticas en línea

<https://es.statista.com/>

6.5.1 Media

La media es el valor promedio de la distribución, es decir, la suma de todos los datos dividida entre el número total de datos.

En nuestro caso de ejemplo con la distribución 2, 3, 2, 3, 1, 1, 2, 2, 1, 3, 4, 3, 2, 5, 4, 7, 9, 1, 1, 2, 1, 1, 2, 1, 5, 8, 6, 1, 2, 1, 2, 1, 1, 2, 1, 2, 2, 1, 23, la media será 3,05

O si, por ejemplo, tenemos registradas las siguientes ventas en una empresa:

Transacción número	Código cliente	Código producto	Importe por unidad	Número de unidades	Importe total de la venta
1	11111	AAAA	30,00€	10	300,00€
2	22222	AAAA	30,00€	12	360,00€
3	33333	AAAA	30,00€	5	150,00€
4	44444	BBBB	40,00€	1	40,00€
5	55555	CCCC	50,00€	10	500,00€
6	66666	BBBB	40,00€	13	520,00€
7	77777	AAAA	30,00€	15	450,00€
8	88888	BBBB	40,00€	20	800,00€
9	99999	CCCC	50,00€	16	800,00€
					Importe total de ventas
					3.920,00€

Distribución de ventas y cálculo de media

Fuente: Elaboración propia

Si queremos calcular el importe medio de cada venta, tendremos que dividir el importe total de las ventas (3.920€) entre el número de transacciones (9), con lo que tendremos que el valor medio de las ventas que hemos realizado es 435,56€.

Cuando los valores que tenemos son muy extremos, la media no resulta representativa del conjunto. Esto lo vemos claramente en el ejemplo anterior en el que el importe de la transacción número 4 son 40€ y los importes de las transacciones 8 y 9 son 800€.

6.5.2 Mediana

La mediana es el valor que divide la distribución en dos partes iguales.

En nuestro caso de ejemplo con la distribución 2, 3, 2, 3, 1, 1, 2, 2, 1, 3, 4, 3, 2, 5, 4, 7, 9, 1, 1, 2, 1, 1, 2, 1, 5, 8, 6, 1, 2, 1, 2, 1, 1, 1, 2, 1, 2, 2, 1, 23 la mediana será 2

O si queremos calcular la mediana de las ventas realizadas, si ordenamos de menor a mayor tendremos la siguiente tabla:

Transacción número	Código cliente	Código producto	Importe por unidad	Número de unidades	Importe total de la venta
4	44444	BBBB	40,00€	1	40,00€
3	33333	AAAA	30,00€	5	150,00€
1	11111	AAAA	30,00€	10	300,00€
2	22222	AAAA	30,00€	12	360,00€
7	77777	AAAA	30,00€	15	450,00€
5	55555	CCCC	50,00€	10	500,00€
6	66666	BBBB	40,00€	13	520,00€
8	88888	BBBB	40,00€	20	800,00€
9	99999	CCCC	50,00€	16	800,00€
				Importe total de ventas	3.920,00€

Distribución de ventas y cálculo de mediana

Fuente: Elaboración propia

La posición central de los valores que se encuentran en la columna importe total de la venta es 450. Si el número de transacciones fuese par, para calcular la mediana debemos calcular la media de las dos posiciones centrales.

6.5.3 Moda

La moda es el valor con mayor frecuencia absoluta. Y es un parámetro calculable tanto en variables cualitativas como cuantitativas.

En nuestro caso de ejemplo, con la distribución 2, 3, 2, 3, 1, 1, 2, 2, 1, 3, 4, 3, 2, 5, 4, 7, 9, 1, 1, 2, 1, 1, 2, 1, 5, 8, 6, 1, 2, 1, 2, 1, 1, 1, 2, 1, 2, 2, 1, 23, la moda será 1.

Si dos valores resultasen ser la moda, entonces se trata de una distribución bimodal o multimodal si hubiese tres o más modas. Excepto en aquellas distribuciones donde existan dos modas correspondiendo a valores correlativos, que entonces se establecerá la moda como la media de ambas modas.

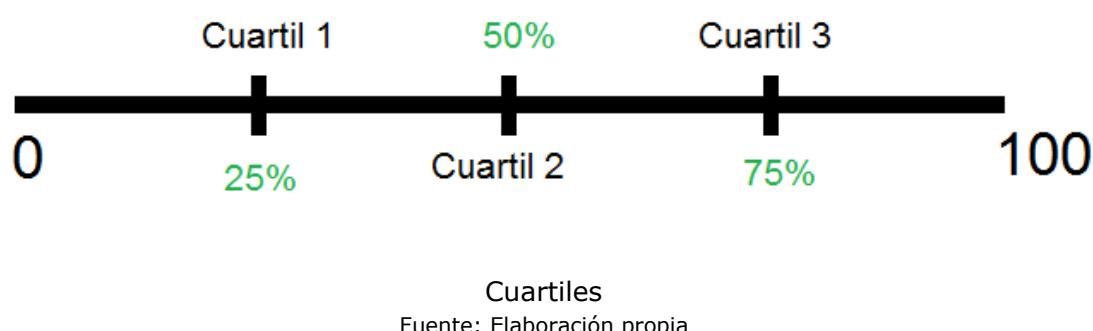
Si todos los valores tienen la misma frecuencia, entonces no habrá moda en dicha distribución.

En las transacciones de la tabla de ejemplo anterior, la moda del importe total de ventas es 800€, ya que es el único que se repite.

6.5.4 Cuartiles

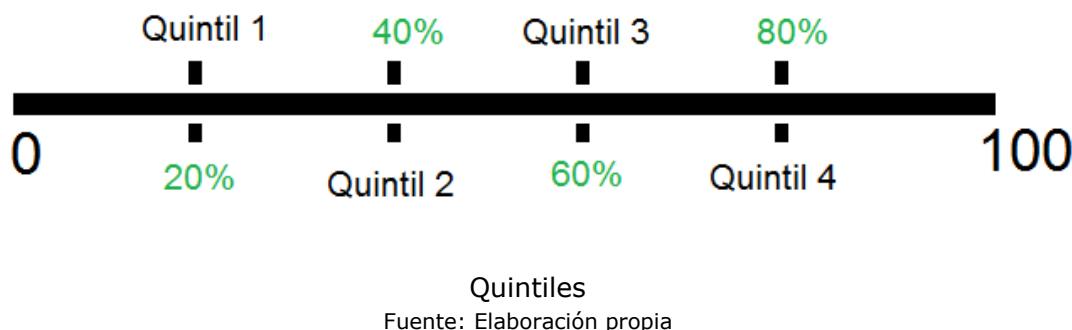
Para calcular los cuartiles, al igual que las restantes medidas de posición, los datos tienen que estar dispuestos en orden creciente.

Los cuartiles son tres valores Q1, Q2 y Q3 que dividen a los datos en cuatro partes iguales y representan por tanto al 25%, al 50% y al 75% respectivamente de los mismos. Además, se dará la coincidencia de que Q2 será, al tiempo, la mediana.



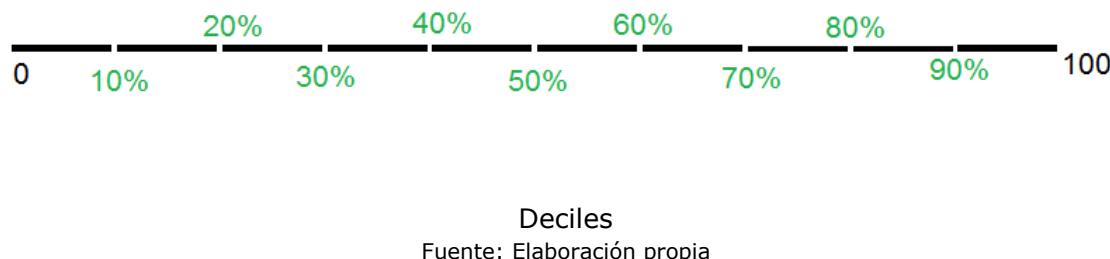
6.5.5 Quintiles

En este caso se trata de los cuatro valores que dividen los datos en cinco partes iguales. Por tanto, el primer quintil será el 20%, el segundo el 40%, el tercero el 60% y el cuarto el 80%.



6.5.6. Deciles

Son los nueve valores que resultan de la división de los datos en diez partes del mismo tamaño. Cada decil se corresponderá con el 10% del total de los datos, el 20%, etc.



6.5.7 Percentiles

Son los 99 valores que dividen los datos en 100 partes iguales. En este caso los valores serán el 1%, el 2%, etc.



ENLACE DE INTERÉS

Para saber más acerca de los percentiles, accede a este enlace:

<https://curiosoando.com/que-son-los-percentiles>

6.5.8 Rangos

Se trata de la diferencia entre el mayor y el menor valor que toman los datos. Por ejemplo, en la siguiente tabla, el rango de todas las ventas realizadas es: $800-40=760$.

Transacción número	Código cliente	Código producto	Importe por unidad	Número de unidades	Importe total de la venta
1	11111	AAAA	30,00€	10	300,00€
2	22222	AAAA	30,00€	12	360,00€
3	33333	AAAA	30,00€	5	150,00€
4	44444	BBBB	40,00€	1	40,00€
5	55555	CCCC	50,00€	10	500,00€
6	66666	BBBB	40,00€	13	520,00€
7	77777	AAAA	30,00€	15	450,00€
8	88888	BBBB	40,00€	20	800,00€
9	99999	CCCC	50,00€	16	800,00€
					Importe total de ventas
					3.920,00€

Cálculo de rangos

Fuente: Elaboración propia

6.5.9 Desviación media

Es la diferencia de cada valor con respecto a la media de todos los valores. Para calcularla primero calcularemos la media y después calcularemos la media de la diferencia de cada valor con respecto a la media. Por ejemplo, la desviación media de la tabla anterior se calculará de la siguiente manera:

$$\text{Media} = (300+360+150+40+500+520+450+800+800)/9 = 435,56$$

Diferencia de cada valor con respecto a la media:

Transacción número	Importe total de la venta	Diferencia de cada transacción con respecto a la media
1	300,00€	135,56
2	360,00€	75,56
3	150,00€	285,56
4	40,00€	395,56
5	500,00€	64,44
6	520,00€	84,44
7	450,00€	14,44
8	800,00€	64,44
9	800,00€	364,44

Cálculo de desviación media

Fuente: Elaboración propia

Desviación media =

$$(135.56+75.56+285.56+395.56+64.44+84.44+14.44+364.44+364.44)/9 \\ = 198.27$$

6.5.10 Varianza

Es la media de las desviaciones al cuadrado de cada valor con respecto a la media de todos los datos. Siguiendo con el ejemplo anterior, la varianza sería:

$$\text{Varianza} = ((\text{Transacción número 1-Media})^2 + \dots + (\text{Transacción número 9-Media})^2)/9 = 67402,77778$$

6.5.11. Desviación típica

Es la raíz cuadrada de la varianza. La raíz cuadrada de la varianza que acabamos de calcular es 259.62. Esto quiere decir que, con los datos anteriores, la media de cada pedido facturado es 435.53€, con la tendencia a variar 259.62€ (ya sea por encima o por debajo). En este caso, al ser mucha la diferencia entre los valores máximos y los mínimos, tanto la media, como la varianza o la desviación típica son poco representativas.

6.5.12 Coeficiente de Pearson

Como ya se ha indicado, el coeficiente de Pearson es el caso más común en medidas relativas, las cuales se expresan en forma de porcentajes. Para calcularlo dividimos la desviación típica entre la media. Siguiendo con el mismo ejemplo, tendremos que el coeficiente de variación será:

Coeficiente de variación=259.62/435.53*100=59.61%

Este parámetro nos permite comprobar en qué medida están dispersos los datos (en este caso la dispersión con respecto a la media es casi el 60%).

6.5.13 Análisis de tendencias

Una tendencia es la dirección hacia la que se mueven unos valores a lo largo del tiempo. Cuando los valores son cada vez mayores, hablamos de tendencia **alcista**, mientras que cuando los valores bajan, diremos que la tendencia es **bajista**. En el caso de que los valores no se dirijan hacia ninguna dirección, sino que se mantengan constantes a lo largo del tiempo, la tendencia es lateral. Es importante, en distintas investigaciones económico-empresariales, por ejemplo, en la bursátil, conocer cómo evolucionan los **valores del mercado** en el que nos encontramos, para poder adaptarnos con más rapidez y facilidad a cualquier nueva situación que se pueda producir.

En el análisis de tendencias y series de datos podemos establecer la relación que existe entre dos variables. Las medidas más utilizadas son:

- **Centro de gravedad.** Partiendo de dos variables, el centro de gravedad es el conjunto de la media de cada variable.
- **Covarianza.** Permite analizar si hay una relación directa entre dos variables. Cuando el resultado es positivo, quiere decir que la relación es directamente proporcional, es decir, cuando una variable aumenta, la otra también. Mientras que cuando la covarianza es negativa quiere decir que la relación es inversamente proporcional, de manera que al aumentar una de las variables la otra disminuye. Si la covarianza es cero, significa que no hay relación entre las variables.

La mayoría de los programas que permiten realizar análisis estadísticos también permiten analizar tendencias.

6.6 Árboles de probabilidad (o árbol de decisión en análisis predictivo)

Un árbol de probabilidad es una herramienta de análisis que:

- Permite analizar causas y efectos desde la perspectiva de la probabilidad.
- Estima qué escenario es el más probable y cuál es el menos probable.
- Determina decisiones y eventos en escenarios alternativos.

Lo primero que puede mostrarnos un árbol de probabilidad es en qué decisión o eventos debemos centrar nuestros esfuerzos si queremos incrementar la probabilidad en un escenario particular. Para ello hay tres reglas fundamentales:

- Los eventos tienen que ser mutuamente excluyentes.
- Los eventos tienen que ser colectivamente exhaustivos.
- La suma total de las probabilidades tiene que ser uno.

La utilidad de estos es que posibilitan aumentar nuestra habilidad para identificar los eventos determinantes de modo arbitrario, variando las probabilidades que les asignamos, haciendo lo que se llama **análisis sensitivo**.

Los tipos de probabilidad manejados pueden ser:

- Probabilidad mutuamente excluyente.
- Probabilidad condicionalmente dependiente.

Los pasos que seguir para aplicar esta técnica son:

1. Identificar el problema.
2. Identificar las decisiones principales y los eventos para ser analizados.
3. Construir un árbol de causa / efecto reflejando todos los escenarios alternativos importantes:
 - Hay que asegurar que las decisiones / eventos de cada rama son mutuamente excluyentes.
 - Asegurarse de que las decisiones / eventos de cada rama son colectivamente exhaustivos.
4. Asignar una probabilidad para cada decisión/evento. Las probabilidades de cada rama deben de ser iguales a 1.0.
5. Calcular la probabilidad condicional para cada escenario individual.
6. Calcular las respuestas para las preguntas de probabilidad relacionadas con las decisiones / eventos.

Las ventajas de esta técnica son que, al añadirse la probabilidad a nuestro análisis, este adquiere mayor fiabilidad. Y que un árbol de probabilidades, además, alumbría y nos permite enfocar nuestra atención en, una vez recogidas las evidencias, cambiar las asunciones, dado que estas decisiones son las que modifican la viabilidad de los escenarios que son importantes para nosotros.

Cuando hablamos de probabilidad estamos hablando de estimar, que es lo que hacemos cuando nos quedamos sin información. Podemos decir que el lenguaje de la estimación es la probabilidad.

Es verdad que la probabilidad permite todos los análisis. No obstante, la correcta aplicación de las leyes de la probabilidad no es algo intuitivo para los humanos.

La probabilidad se guía por juicios mucho más subjetivos que los cálculos matemáticos. Estos juicios subjetivos, como otros, están sujetos a prejuicios y otros problemas relacionados con las tendencias de la mente humana. Por ello, la mayoría de las personas puede ver intuitivamente la respuesta sólo para los problemas simples que envuelven la probabilidad.

Los tipos de probabilidad que hemos anticipado son los de:

- **Probabilidad mutuamente excluyente:** Dos o más eventos que ocurren como resultado de una única decisión o evento “esto...o...esto” por lo que sumamos las probabilidades individuales.
- **Probabilidad condicionalmente dependiente:** Dos o más eventos que ocurren en sucesión “esto...y...esto” por lo que multiplicamos las probabilidades individuales.

Un modo sencillo de contabilizar probabilidades mutuamente excluyentes es pensar en ellas en términos de porcentajes. ¿Qué porcentaje es X sobre Y? Dividimos X/Y

Para ello, es conveniente usar el término “es tanto sobre”, porque en ocasiones el contexto del problema hace difícil terminar que número esta dividido por cuál.

Si decidimos esto....excluimos lo otro.

Por ejemplo, el lanzamiento de una moneda (cara o cruz)

Calculamos la probabilidad condicionalmente dependiente multiplicando las probabilidades que están vinculadas condicionalmente.

Si decidimos A... ocurre B... y ocurre C...etc.

Veamos un ejemplo sencillo de un árbol de probabilidad para la siguiente situación: Un programa de televisión de “todo o nada”, donde un concursante está a punto de tomar una importante decisión: quedarse con el premio que ya ha ganado (un coche valorado en 6.000 €) o abrir la puerta roja, donde puede haber un premio de 1 millón de €.

Si abre la puerta y no hay nada, pierde el coche y otros premios menores que ya ha ganado:

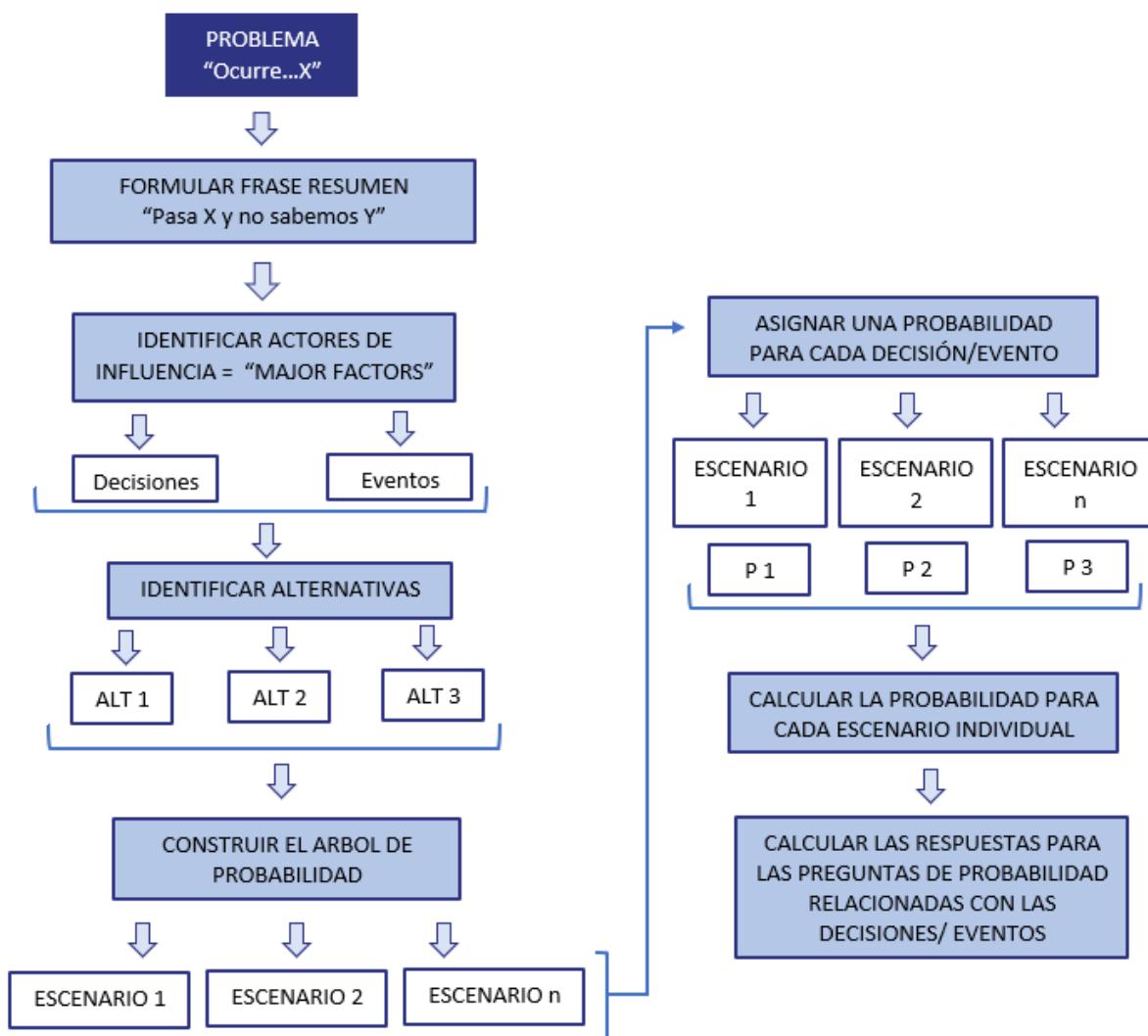


Ejemplo de árbol de probabilidad

Fuente: Elaboración propia

Realizar esta reflexión mediante el uso de árboles de probabilidad es una forma de reducir los sesgos. Nuestras mentes pueden ser fácilmente inducidas en las estimaciones de probabilidad, porque hay un juicio analítico por medio. Los sesgos cognitivos más habituales sin la técnica son asumir que la probabilidad es la misma para todos los resultados, sesgo de familiaridad, sesgo de optimismo, etc.

El proceso de aplicación del árbol de probabilidad queda reflejado en la siguiente figura.



Árbol de probabilidad

Fuente: Elaboración propia



EJEMPLO PRÁCTICO

El contexto es que MUM S.L., una importante empresa del sector de los lácteos, con una imagen de marca ya consolidada, sufre en mayo de 2020 un importante contratiempo: una partida de 300.000 cartones de leche distribuidos por más de 50 grandes centros comerciales españoles ha salido defectuosa, transmitiendo una peligrosa bacteria que afecta a la salud de los consumidores.

La noticia ha sido rápidamente difundida por blogs y medios de comunicación. La empresa teme que lo ocurrido afecte a la facturación de 2020 y a su imagen de marca.

SOLUCIÓN

Comenzamos identificando los actores y situaciones del problema:

- Empleados de Mum
- Equipo directivo
- Accionistas de Mum
- Competidores en el sector
- Asociaciones de consumidores
- Consumidores habituales
- Consumidores potenciales
- Medios de comunicación
- Autoridades sanitarias

Formulamos el problema en una frase corta y resumida:

Un producto en mal estado ha llegado a los consumidores y ha minado su confianza en la marca, desencadenando una crisis en la compañía.

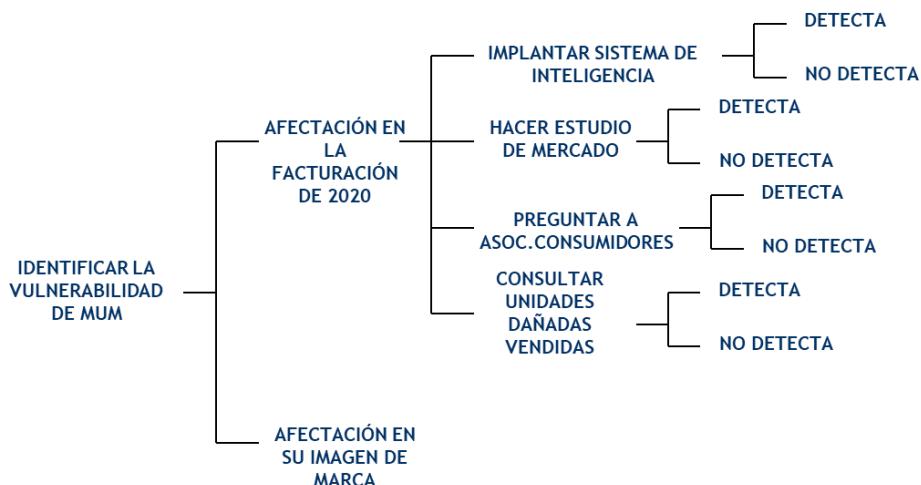
Identificamos los factores clave y las alternativas:

MUM S.L., una importante empresa del sector de los lácteos con una imagen de marca ya consolidada sufre en mayo de 2008 un importante contratiempo: una partida de 300.000 cartones de leche distribuidos por más de 50 grandes centros comerciales españoles ha salido defectuosa, transmitiendo una peligrosa bacteria que afecta a la salud de los consumidores. La noticia ha sido rápidamente difundida por blogs y medios de comunicación. La empresa teme que lo ocurrido afecte a la facturación de 2020 y a su imagen de marca.

Esquematizamos diciendo que:

- Problema: Vulnerabilidad de MUM por los efectos de la partida defectuosa.
- Actores:
 - Consumidores
 - MUM
 - Blogs y medios de comunicación
- Eventos:
 - Afectación a la facturación de 2020
 - Afectación a la imagen de marca de MUM
- Decisiones:
 - Implantar un sistema de inteligencia
 - Hacer un estudio de mercado
 - Preguntar a la asociación de consumidores
 - Preguntar cuántas unidades defectuosas se han vendido

Construimos el árbol de probabilidad:



Árbol de probabilidad

Fuente: Elaboración propia

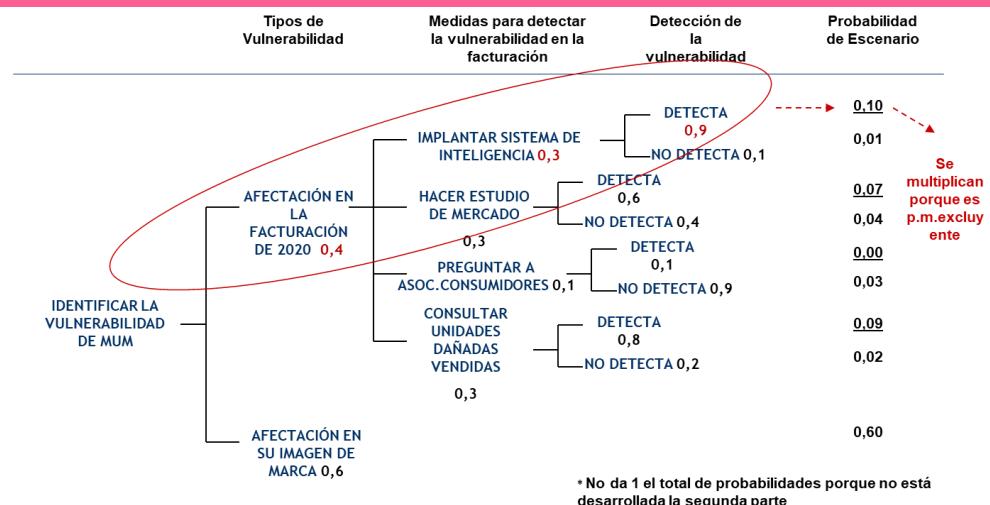
Asignamos una probabilidad para cada decisión / evento:

Tipos de Vulnerabilidad	Medidas para detectar la vulnerabilidad en la facturación	Detección de la vulnerabilidad	Probabilidad de Escenario
IDENTIFICAR LA VULNERABILIDAD DE MUM AFECTACIÓN EN LA FACTURACIÓN DE 2020 0,4 AFECTACIÓN EN SU IMAGEN DE MARCA 0,6	AFECTACIÓN EN LA FACTURACIÓN DE 2020 0,4 IMPLANTAR SISTEMA DE INTELIGENCIA 0,3 HACER ESTUDIO DE MERCADO 0,3 PREGUNTAR A ASOC.CONSUMIDORES 0,1 CONSULTAR UNIDADES DAÑADAS VENDIDAS 0,3	DETECTA 0,9 NO DETECTA 0,1 DETECTA 0,6 NO DETECTA 0,4 DETECTA 0,1 NO DETECTA 0,9 DETECTA 0,8 NO DETECTA 0,2	

Árbol de probabilidad

Fuente: Elaboración propia

Calculamos la probabilidad para cada escenario individual:



Árbol de probabilidad

Fuente: Elaboración propia

Calculamos las respuestas para las preguntas de probabilidad:

- ¿Qué probabilidad hay de que se detecte la vulnerabilidad de la empresa MUM?: $0,10 + 0,07 + 0,00 + 0,09 = 0,26$
- ¿Qué probabilidad hay de que no se detecte la vulnerabilidad de la empresa MUM?: $0,01 + 0,04 + 0,03 + 0,02 = 0,1$
- ¿Cuál es la mejor medida para detectar la vulnerabilidad en la facturación de 2020? Miramos los resultados para "detección de la vulnerabilidad" y aquella probabilidad más alta será la que corresponda a la "mejor medida"
 - Implantar un sistema de inteligencia = 0,90
 - Consultar unidades dañadas vendidas = 0,80



VIDEOTUTORIAL

En el siguiente videotutorial se explican los fundamentos de la estadística:

<https://vimeo.com/user64513894/review/447075301/dbb770bb66>

6.7 Inferencia

Murray R. Spiegel, en su obra probabilidad y estadística, se expresa en los siguientes términos al respecto de la inferencia estadística: "Con frecuencia

en la práctica estamos interesados en extraer conclusiones válidas respecto a un grupo grande de individuos u objetos. En vez de examinar un grupo entero, llamado la población, lo cual puede resultar difícil o imposible, puede llegarse a la idea de examinar solamente una parte pequeña de esta población, que se llama la muestra. Esto se hace con el propósito de inferir ciertos hechos respecto de la población de los resultados hallados en la muestra, un proceso conocido como inferencia estadística."

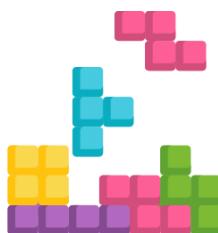
Pero la inferencia no siempre tiene una base estadística. En palabras de Alejandro Díaz-Bautista: "Otro elemento del proceder científico es el uso sistemático de la inferencia, o razonamiento deductivo. Inferir significa **sacar consecuencias de un principio o supuesto**. La inferencia opera durante la investigación y, por lo general, de la siguiente manera: una vez formulada una hipótesis se deducen de ella posibles consecuencias prácticas, que luego son sometidas, a su vez, a verificación"

A continuación, mostramos posibles **inferencias de utilidad en la inteligencia de negocio**:

- Inferencias especificativas
 - ¿Qué componentes conceptuales faltan, probablemente, en un grupo conceptual incompleto?
- Inferencias causales
 - ¿Cuáles fueron las causas probables de una acción o estado?
- Inferencias resultativas
 - ¿Cuáles son los resultados probables (efectos sobre el mundo) de una acción o estado?
- Inferencias motivacionales
 - ¿Por qué quiso (o quería) un actor realizar una acción?
 - ¿Cuáles fueron sus intenciones?
- Inferencias sobre capacidades
 - ¿Qué estados del mundo deben ser (deben haber sido) verdaderos para que tenga lugar una acción?
- Inferencias funcionales
 - ¿Por qué desea la gente poseer objetos?
- Inferencias sobre capacitación y predicción

- Si una persona quiere que exista un estado del mundo particular, ¿se debe ello a que ese estado posibilitara alguna acción predecible?
- Inferencias sobre incapacitación
 - Si una persona no puede realizar alguna acción que desea, ¿puede explicarse por algún estado del mundo que es un requisito previo y no tiene lugar?
- Inferencias sobre mediaciones
 - Si una acción está causando en el mundo (o causará) resultados indeseables.
- Inferencias predictivas de acciones
 - Conociendo las necesidades y deseos de una persona, ¿qué acciones realizará ésta, probablemente, para lograr esos deseos?
- Inferencias sobre propagación de conocimiento
 - Sabiendo que una persona sabe ciertas cosas, ¿qué otras cosas se pueden predecir que también sabe?
- Inferencias normativas
 - Por relación a un conocimiento de lo que es normal en el mundo, determinar con qué fuerza se ha de creer una noticia en ausencia de un conocimiento específico.
- Inferencias sobre permanencias de estado
 - ¿Cuánto tiempo, aproximadamente, se puede predecir que durará algún estado a acción prolongada?
- Inferencias de rasgos
 - Conociendo algunos rasgos de una entidad, y las situaciones en que aparece esa entidad, ¿qué cosas adicionales se pueden predecir acerca de esa entidad?
- Inferencias de situación
 - ¿Qué otra información se puede imaginar (inferir) de una situación familiar?
- Inferencias de intención de expresión

- ¿Qué se puede inferir de la forma en que se dice algo?
- ¿Por qué lo dijo el hablante?
- Inferencias relacionales
 - Diacrónicas: con quién o con qué se puede relacionar a un actor.
 - Sincrónicas: con quién o con qué se ha relacionado un actor en el período del desarrollo de una acción o suceso.
- Inferencias sobre propagación de relaciones
 - Sabiendo que un actor se relaciona con unas entidades, ¿qué otras entidades podemos decir que se relacionan con él?



EJEMPLO PRÁCTICO

Eres analista de datos en un banco que cada año invierte una cantidad de dinero importante para enviar a decenas de sus empleados a realizar los exámenes para obtener distintas certificaciones.

Por ejemplo, muchas entidades financieras envían cada año a un grupo de trabajadores a las pruebas CFA (Chartered Financial Analyst)

¿Cómo podría hacer el banco para tener una previsión de los trabajadores que van a aprobar la certificación?

SOLUCIÓN:

Un mecanismo para poder tener esa capacidad de inferencia sería aplicar la **regresión lineal para estimar la probabilidad de que un grupo de empleados aprueben una certificación**.

La regresión lineal simple la podemos definir como un modelo matemático donde tendremos una variable dependiente y con las variables independientes x, siendo b la influencia de estas variables independientes sobre la variable dependiente y:

$$y = a + bx$$

Al ser pruebas de tipo test con cuatro opciones a elegir, podemos decir que el valor de a en la regresión lineal sería de 25, porque hay una entre cuatro posibilidades de acertar la respuesta.

La otra parte de la ecuación bx es un poco más compleja de determinar, pero podríamos decir que x son las horas de estudio que cada trabajador

ha dedicado a preparar la certificación y b sería el incremento de probabilidad de obtener el certificado por cada hora extra de estudio.

La correlación es el nivel de dependencia que existe entre ambas variables. Cuando su valor es más próximo a 1, se trata de una dependencia más profunda. Mientras que cuando tiende a -1, su relación es inversa.

Imaginemos que este año los trabajadores de nuestra entidad financiera que se han presentado al examen CFA han declarado haber estudiado las siguientes horas y han obtenido las siguientes notas.

Trabajador	X (horas de estudio)	Y (nota obtenida siendo 100 la máxima puntuación posible)
José	0	20
Juan	2	50
María	5	80
Antonia	8	90
Luis	8	85
Alfredo	3	40
Cayetana	6	60
Ángel	5	50
Luis	7	75
Laura	1	20
Alex	2	20

Si utilizamos un software para el cálculo de la regresión lineal obtendremos que la fórmula que describe una recta que aproxima del mejor modo posible a los datos utilizados es $Y=16,6497+8,6565X$

El coeficiente de correlación, el cual indica el nivel de relación que tienen ambas variables es 0,9196, lo que quiere decir que la correlación entre horas de estudio y puntuación obtenida es muy alta.

7. SEGMENTACIÓN DE LOS DATOS

Otra de las líneas de trabajo que vas a abordar en tu proyecto apoyándote en el uso de Weka y de R, será aplicar y comprobar los resultados de distintos algoritmos de clustering.

De nuevo, marketing ha sido uno de los principales beneficiarios del esfuerzo que habéis hecho en el departamento de informática. Habéis comenzado a clusterizar la base de datos de clientes de la empresa para intentar encontrar conocimiento oculto en ella. Al director de marketing esta iniciativa también le ha parecido muy interesante porque le ha abierto los ojos y descubierto circunstancias que nunca había imaginado. Se ha descubierto que hay grupos de clientes que comparten unos patrones de compra muy determinado.

Por ejemplo, clientes de alta estacionalidad, que concentran el 80% de las compras que os realizan en los dos o tres meses previos a las campañas de navidad. De esta forma, ahora la planificación de ventas de vuestra empresa se va a poder hacer de mejor modo, distribuyendo los tiempos de los vendedores de modo que dirijan sus esfuerzos comerciales a cada cliente en el momento adecuado y no os ocurra lo que sucedía muchas veces en el pasado que era que se llegaba demasiado tarde y cuando se visitaba al director de una empresa cliente te informaba que ya había realizado una compra importante a vuestro principal competidor que se os había adelantado en el tiempo.

Muchas veces, cuando estudiamos grandes conjuntos de datos, estos datos no tienen ningún tipo de etiqueta o clasificación, ya que esto requiere una cantidad importante de esfuerzo humano. Los datos que no están etiquetados o clasificados pueden ser analizados haciendo uso de técnicas de clustering.

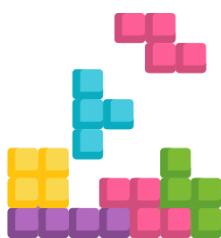
El clustering es una técnica de aprendizaje sin supervisión que no requiere que los datos estén previamente etiquetados. El clustering puede ser definido como el agrupamiento de un conjunto de objetos similares en clases o clústeres. En otras palabras, durante el análisis de clúster, los datos son agrupados en clases, es decir, los registros dentro de un clúster tendrán semejanzas, pero serán distintos en comparación con los de los otros clústeres.

Las semejanzas de los registros son identificadas en base a los valores que describen los objetos.

En este apartado nos centraremos en el clustering, en comprender cómo funciona y entenderlo plenamente para, posteriormente, poder implementar esta técnica con las ya conocidas herramientas WEKA y lenguaje R, aplicadas a ejemplos reales para un mejor entendimiento.

7.1 Clustering

El análisis en clúster es una actividad muy importante realizada desde hace mucho tiempo por la humanidad. Podríamos decir que el ser humano ha evolucionado aprendiendo mediante un proceso de clustering. El simple hecho de observar cada objeto, cuando no se conoce el nombre de este y basándonos en las similitudes con otros, es una forma de clusterizar.



EJEMPLO PRÁCTICO

Se acaba de incorporar un nuevo compañero al equipo y tu jefe te pide que le pongas un poco al día sobre el proceso de clustering que lleváis a cabo en vuestra tarea diaria.

SOLUCIÓN

Hemos definidos clústeres llamados árboles, otros llamados frutas y así sucesivamente. Después, esos clústeres pueden ser clasificados más en detalle por sus propiedades, como tamaño, color, sabor y forma, estableciendo conceptos como cítricos o leguminosas.

Finalmente, les colocamos etiquetas o nombres como manzana, plátano, naranja, piña y así hasta que, finalmente, todos los objetos acaban siendo etiquetados.

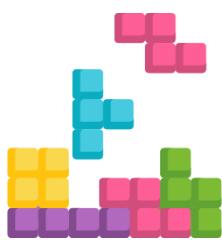
7.2 Aplicaciones del análisis de clúster

El análisis de clúster es utilizado con frecuencia en varios campos importantes de la actividad empresarial:

- Marketing: El clustering nos ayuda a encontrar los grupos distintivos dentro de las bases de datos de clientes.
- Geografía y sistemas de información geográficas, urbanismo, agricultura: reconocimiento del uso de tierras basáandonos en

fotografías aéreas, agrupándolas por uso. Por ejemplo: bosque, huerta, cereal, etc.

- Seguros: detección de grupos con mayor riesgo.
- Planeamiento y ciudades inteligentes: identificación de clústeres de viviendas por tipología. Por ejemplo: en altura, unifamiliares, adosados, etc.
- Estudio de terremotos en sismología: los registros y mediciones realizadas sobre los epicentros de los movimientos son agrupados a lo largo de las fallas en los continentes.
- Estudios biológicos: Clasificación de plantas y animales, identificación de genes con funcionalidades similares, etc.
- Descubrimiento web: para la categorización automática de documentos en la red, encontrando similitudes entre ellos, incluyendo la autoría.
- Detección de fraudes: útil para detectar aplicaciones con comportamientos extraños.



EJEMPLO PRÁCTICO

El departamento de recursos humanos de la empresa donde trabajas quiere tener una visión de cómo son los empleados de la compañía y para ello encargan al departamento de informática que realice un análisis de los mismos.

SOLUCIÓN

El departamento de informática decide realizar un análisis de cluster, buscando grupos homogéneos de trabajadores para que recursos humanos disponga de ese conocimiento y así mejorar la gestión que hace del personal.

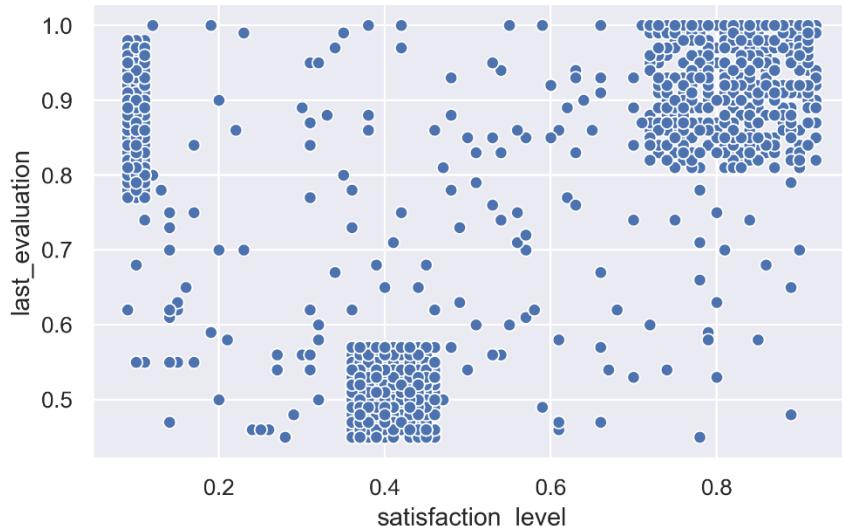
El resultado del proceso serán varios agrupamientos para un conjunto de elementos que representan a empleados de una compañía.

Los datos utilizados para clusterizar a estos individuos se ha decidido que sean el nivel de satisfacción declarado en las encuestas que rellenan los empleados y los resultados de la evaluación que realizan sus superiores jerárquicos, así como los compañeros.

Se han obtenido tres clústeres generados, que agrupan a un porcentaje muy elevado de todos los empleados y se observa la aparición de tres grandes grupos:

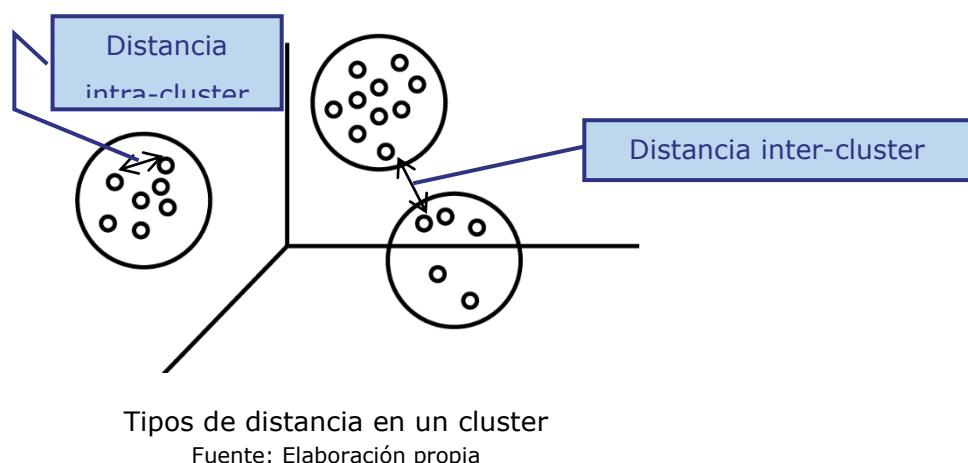
- Empleados muy valorados, pero muy insatisfechos. Se trata del grupo prioritario a gestionar, por el riesgo de abandono que presentan.

- Empleados altamente valorados, al tiempo que muy satisfechos. Podrían definir el perfil prototípico de empleado a contratar en el futuro por su buen rendimiento y adaptación a la cultura de la empresa.
- Empleados razonablemente satisfechos, pero con muy bajas valoraciones. Se trata de un grupo de empleados conformista.



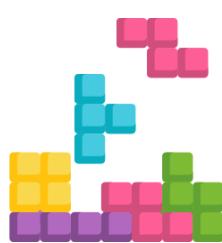
7.3 Características deseadas en el clustering

Las características deseadas o ideales al usar la técnica de clustering es que la distancia intra-cluster sea minimizada y la distancia inter-cluster sea maximizada. Es decir, que los clústeres sean muy coherentes y que, a su vez, sean muy diferentes de otros clústeres, evitando ambigüedades.



Otras características deseadas para que un proceso de clusterización sea considerado como óptimo, son:

- Escalabilidad: Los algoritmos de clustering deben ser capaces tanto de manejar grandes cantidades de información como pequeñas.
- Capacidad de gestionar distintos tipos de atributos: Los algoritmos de clustering deben poder manejar distintos tipos de datos como binarios, categóricos, numéricos.
- Independientes del orden de entrada de los datos.
- Identificación del clúster de distintas formas.
- Capacidad para manejar datos con interferencias o ruido.
- Alto rendimiento: de modo que el algoritmo debería ejecutarse completamente con un solo escaneo de la base de datos.
- Los resultados deben ser interpretables lógicos y usables.
- La posibilidad de detener el algoritmo y continuar con posterioridad. Esto es especialmente útil para cuando se estudian enormes cantidades de datos.



EJEMPLO PRÁCTICO

Un centro de investigación quiere fomentar la innovación entre todos los investigadores que lo componen y los datos y el análisis de los mismos pueden ser una gran ayuda para conseguirlo.

SOLUCIÓN

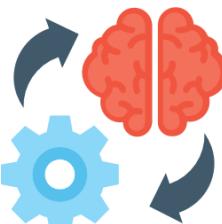
Una iniciativa a lanzar por parte de ese centro de investigación que puede aportar ideas muy interesantes, sería aplicar un algoritmo de clusterización a todos los investigadores que trabajan en él.

Para este proceso sería necesario disponer de datos sobre las áreas de investigación y de interés históricas y presentes de los investigadores. De este modo sería posible encontrar intersecciones en los intereses de profesores que entre ellos no se conocen, para proponerles e incluso facilitarles la creación de equipos interdisciplinares de investigación.

Para que los investigadores que se agrupan en cada clúster realmente sea coherentes entre ellos, los clústeres tienen que estar muy bien hechos, lo que significa que la distancia entre clústeres debe ser lo mayor posible (para que no se solapen los intereses entre grupos de investigadores lo que podría provocar conflictos) al tiempo que la distancia entre los elementos debe ser mínima (es decir, que los investigadores que se han agrupado en cada clúster realmente comparten intereses en común, porque de otro modo será para ellos muy difícil trabajar colaborativamente porque no se entenderían).

7.4 K-medias clustering

En el algoritmo k-medias, n objetos son agrupados dentro de K clústeres o particiones, esto en base a los atributos, donde $k < n$ es un numero positivo entero. Es decir, en el algoritmo k-medias, los objetos son agrupados en k clústeres basado en sus atributos o características. El agrupamiento de los objetos es realizado minimizando la suma de los cuadrados de la distancia, entendiendo como distancia la euclidiana entre los datos y el correspondiente centroide del clúster.


RECUERDA

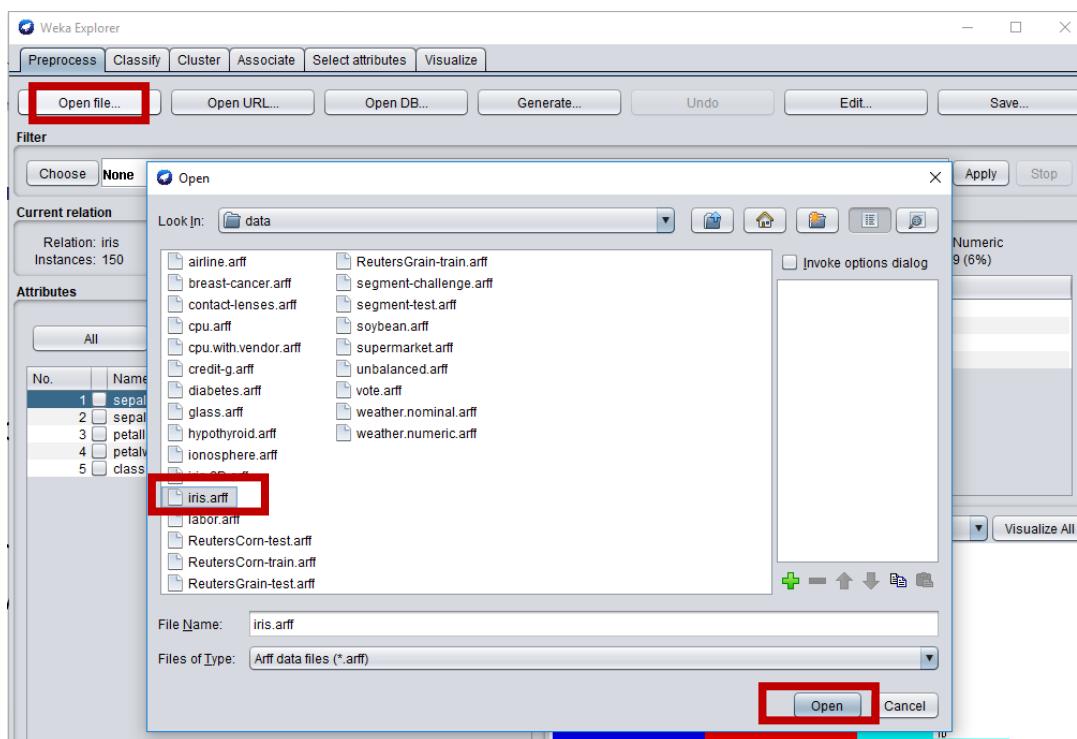
La distancia Euclidiana es la comúnmente usada para calcular la distancia entre dos puntos en el plano con coordenadas (x, y) y (a, b) . Para ello se usa la fórmula:

$$\text{dist}((x, y), (a, b)) = \sqrt{(x - a)^2 + (y - b)^2}$$

7.5 Implementando K-medias en WEKA

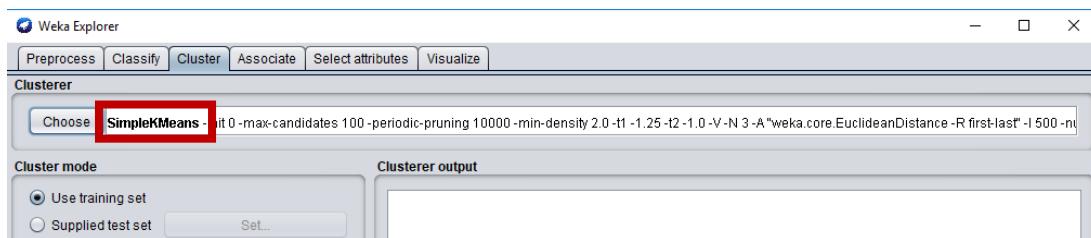
Ahora veamos cómo implementar este algoritmo en WEKA con el dataset de la flor de Iris:

1. Vayamos a la función de Explorer en WEKA y cargamos los datos desde el archivo `iris.arff`

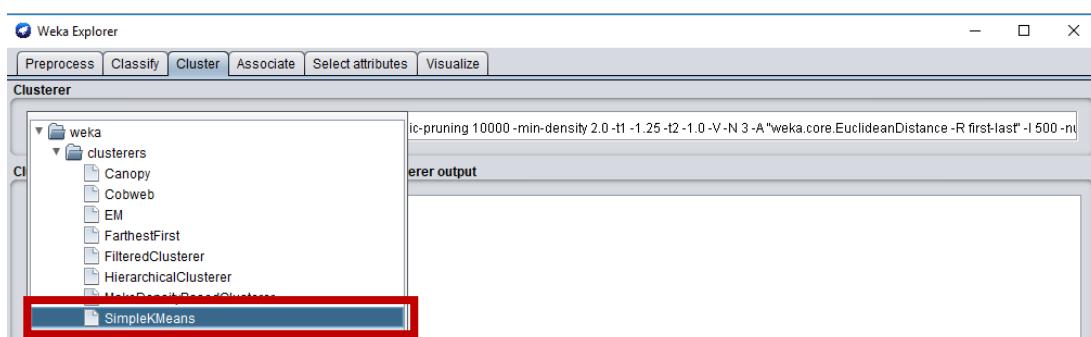


Carga de datos desde el archivo iris.arff
Fuente: Elaboración propia

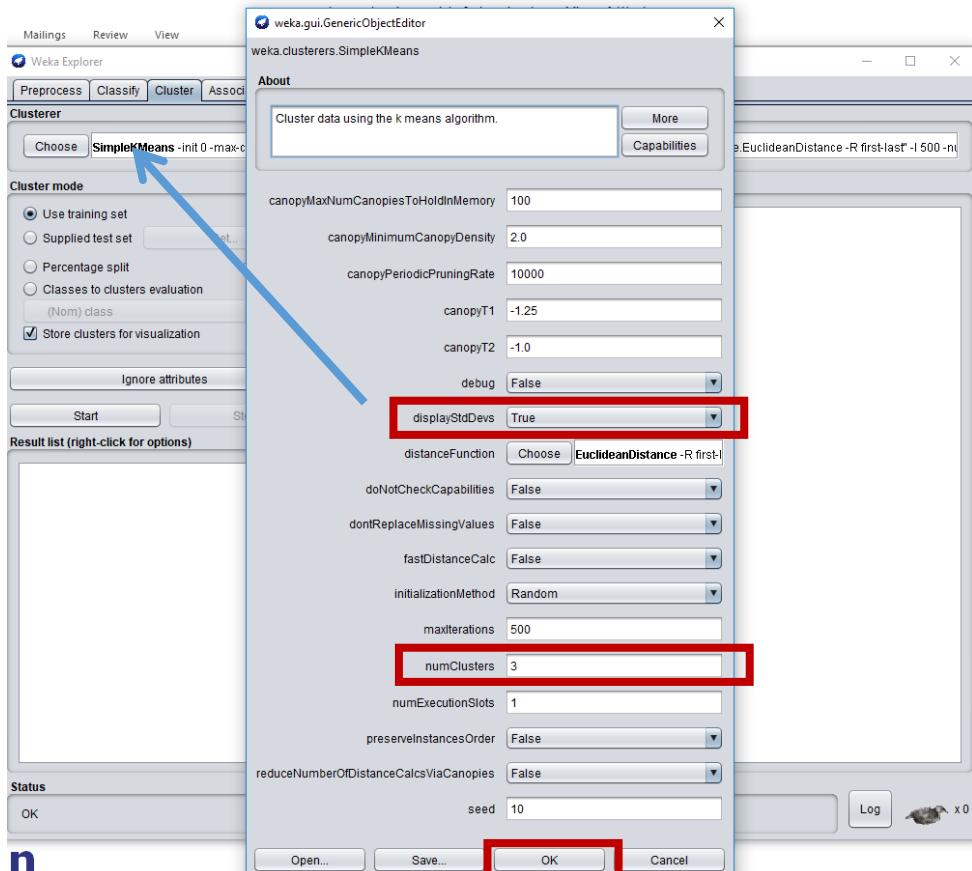
- Después nos dirigimos a la pestaña clúster y seleccionamos el algoritmo **SimpleKMeans**. A este algoritmo le configuraremos el número de clústeres a 3, y activaremos que nos muestre las desviaciones estándar con el parámetro **displayStdDevs** en true.



Selección algoritmo SimpleKMeans
Fuente: Elaboración propia



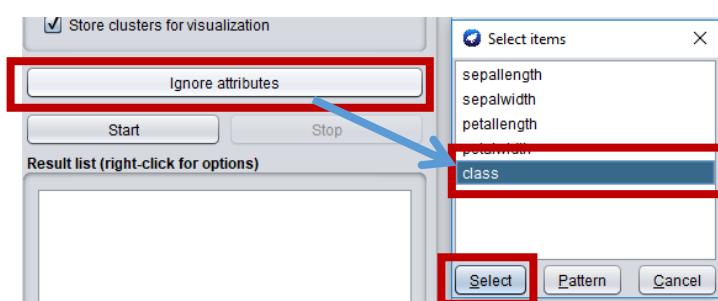
Selección algoritmo SimpleKMeans II
Fuente: Elaboración propia



Parámetro displayStdDevs en True

Fuente: Elaboración propia

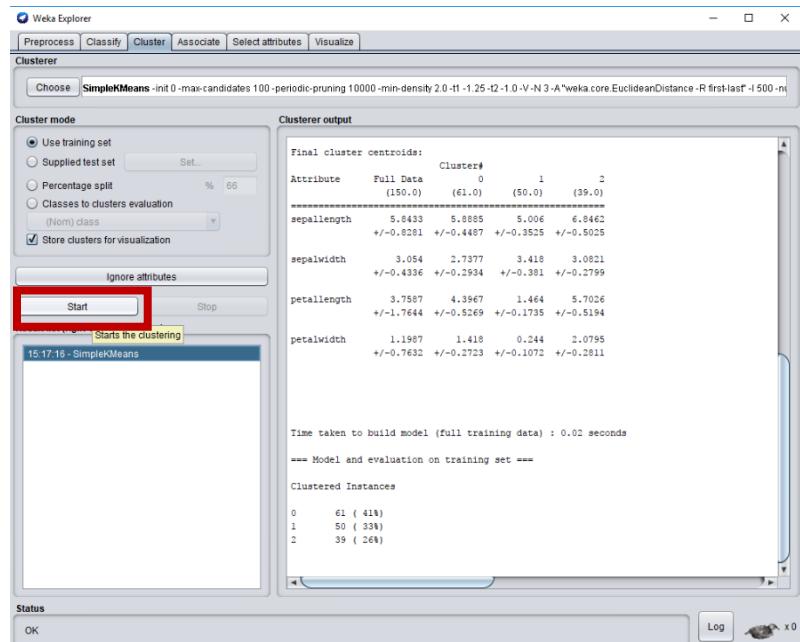
- Como los datos deben estar sin etiqueta para este tipo de análisis, ignoraremos el atributo clase de la siguiente manera:



Ignorar atributo clase

Fuente: Elaboración propia

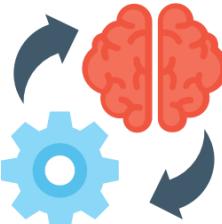
- Presionamos el botón **start**.



Botón start

Fuente: Elaboración propia

Analicemos ahora estos resultados. Como podemos ver, tenemos una matriz donde se nos muestran los centroides de los tres clústeres y sus desviaciones estándar.

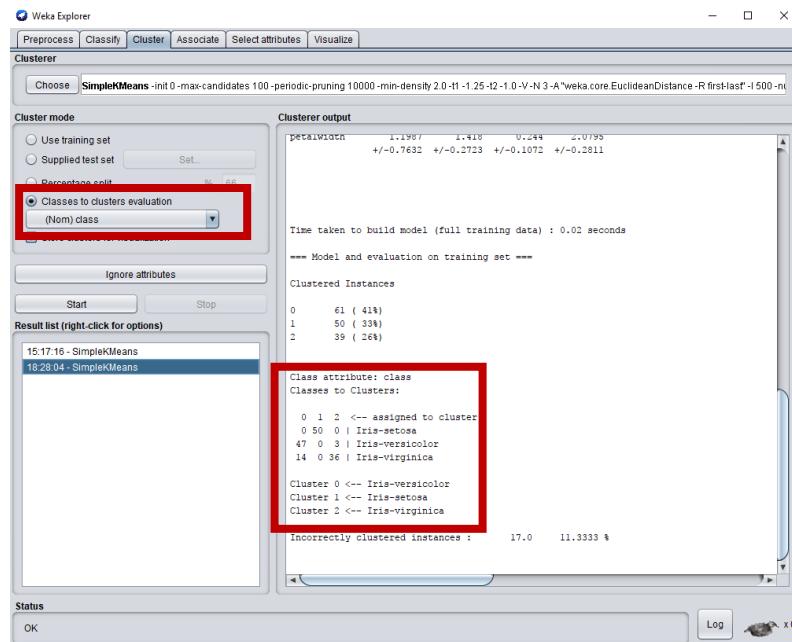


RECUERDA

El centroide de un clúster se define como el punto equidistante de los objetos pertenecientes a dicho clúster.

En estos resultados observamos que el algoritmo K-means encontró 3 clústeres, uno con 61, otro con 50 y otro con 39 miembros. Como estamos trabajando con los datos de la flor de Iris, los cuales nos empiezan a resultar familiares, ya sabemos que son tres clases con cincuenta instancias cada una y que, por tanto, existen 11 registros que están cambiados. Lo que implica que una clase tenga 11 en exceso y otra 11 en defecto respecto de los 50.

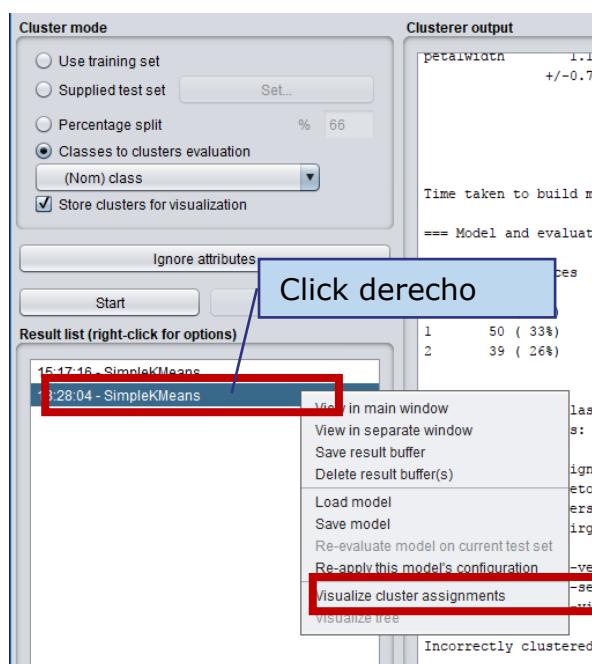
Para comparar estos resultados con las clases ya identificadas podemos seleccionar la opción **classes to clusters evaluation**.



Seleccionar la opción classes to clusters evaluation

Fuente: Elaboración propia

Ahora, podemos revisar como es la clasificación gráficamente con la siguiente opción:



SimpleKMeans. Visualize cluster assignments

Fuente: Elaboración propia

Esto nos mostrará el gráfico siguiente:

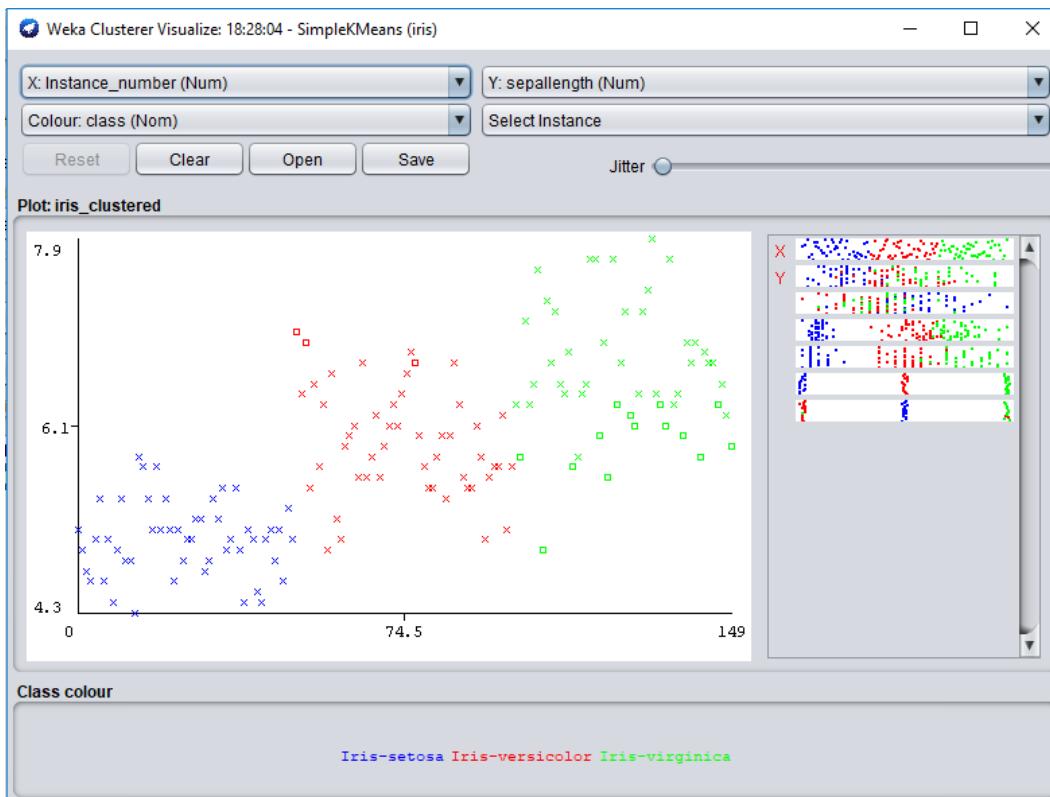


Gráfico SimpleKMeans. Weka clusterer visualize

Fuente: Elaboración propia

Aquí podríamos hacer más estudios y afinar nuestros resultados, revisando las múltiples opciones que nos proporciona WEKA.

7.6 Implementando K-medias en R

Al igual que con el ejemplo anterior, usaremos el dataset de la flor de Iris en R para aplicar el algoritmo de clusterización. Comprobamos como este dataset, al ser muy usado, ya está precargado por defecto en R. Veamos sus estadísticas con el comando **str(iris)**:

```
> str(iris)
'data.frame': 150 obs. of 5 variables:
 $ Sepal.Length: num 5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num 3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num 0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",...: 1 1 1 1 1 1 1 1 1 1 ...
```

Comando str(iris)

Fuente: Elaboración propia

Como podemos comprobar, tenemos 150 observaciones con 5 variables. Ahora usamos el siguiente comando para transformar el dataset en un dataframe:

```
iris_df <- iris
```

Y vemos su contenido mediante:

Iris_df

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3.0	1.4	0.1	setosa
14	4.3	3.0	1.1	0.1	setosa
15	5.8	4.0	1.2	0.2	setosa
16	5.7	4.4	1.5	0.4	setosa
17	5.4	3.9	1.3	0.4	setosa
18	5.1	3.5	1.4	0.3	setosa
19	5.7	3.8	1.7	0.3	setosa
20	5.1	3.8	1.5	0.3	setosa
21	5.4	3.4	1.7	0.2	setosa
22	5.1	3.7	1.5	0.4	setosa

Comando str(iris) II

Fuente: Elaboración propia

Ahora debemos eliminar la variable class de nuestros datos, porque haremos un análisis categórico y la guardaremos en una lista antes de eliminarla:

```
especies <- as.list(iris_df$Species)
especies <- unlist(especies)
iris_df <- iris_df[1:4]
```

Tras estos comandos, ya habremos eliminado la variable Species.

```

> especies <- as.list(iris_df$Species)
> especies <- unlist(especies)
> iris_df <- iris_df[1:4]
> iris_df
   Sepal.Length Sepal.Width Petal.Length Petal.Width
1          5.1         3.5          1.4         0.2
2          4.9         3.0          1.4         0.2
3          4.7         3.2          1.3         0.2
4          4.6         3.1          1.5         0.2
5          5.0         3.6          1.4         0.2
6          5.4         3.9          1.7         0.4
7          4.6         3.4          1.4         0.3
8          5.0         3.4          1.5         0.2
9          4.4         2.9          1.4         0.2
10         4.9         3.1          1.5         0.1
11         5.4         3.7          1.5         0.2
12         4.8         3.4          1.6         0.2
13         4.8         3.0          1.4         0.1
14         4.3         3.0          1.1         0.1
15         5.8         4.0          1.2         0.2
16         5.7         4.4          1.5         0.4
17         5.4         3.9          1.3         0.4
18         5.1         3.5          1.4         0.3
19         5.7         3.8          1.7         0.3

```

Variable especies
Fuente: Elaboración propia

Ahora, aplicaremos el algoritmo k-means. Aquí debemos especificar que son tres clústeres y qué serán veinte iteraciones:

```
irisCluster <- kmeans(iris[, df[, 3], nstart = 20])
```

Después, vemos el resultado :

irisCluster

Variable especies

En este resumen podemos observar la mayoría de los resultados. Vemos que se han clasificado en 3 clústeres ed 38, 62 y 50.

Podemos usar el comando: **table(irisCluster\$cluster)** para ver los tamaños de los clústeres:

```
> table(irisCluster$cluster)

 1  2  3
38 62 50
> |
```

Table(irisCluster\$cluster)
Fuente: Elaboración propia

También podemos revisar los centroides con el comando:
irisCluster\$centers:

```
> irisCluster$centers
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1     6.850000   3.073684    5.742105   2.071053
2     5.901613   2.748387    4.393548   1.433871
3     5.006000   3.428000    1.462000   0.246000
> |
```

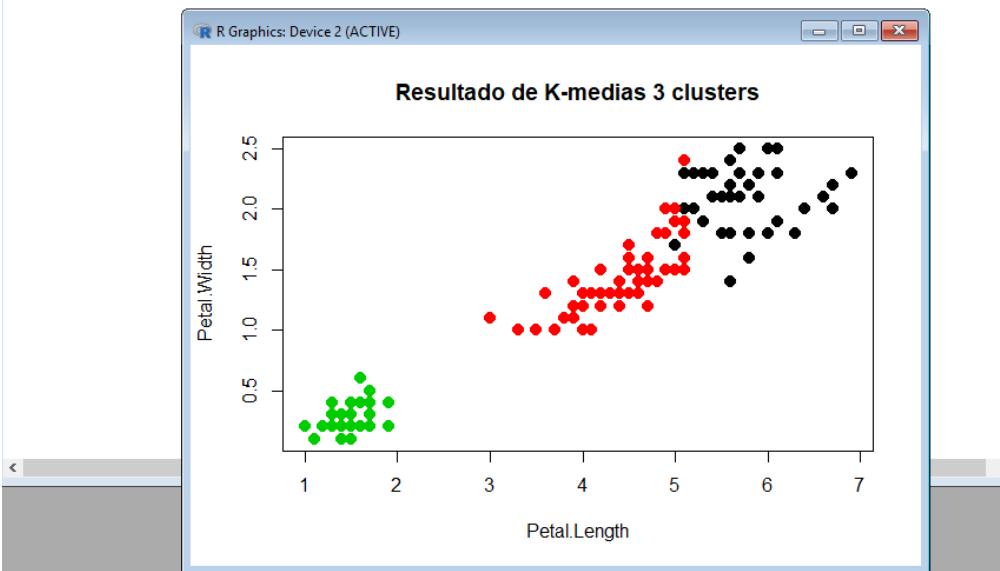
irisCluster\$centers
Fuente: Elaboración propia

Podemos también hacer un gráfico de la siguiente manera, para los atributos largo y ancho de pétalo:

```
irisCluster$cluster <-as.factor(irisCluster$cluster)

plot(iris_df[,3:4], col = (irisCluster$cluster), main="Resultado de K-medias
3 clusters", pch=20, cex =2)
```

```
> irisCluster$cluster <- as.factor(irisCluster$cluster)
> plot(iris_df[,3:4], col =(irisCluster$cluster), main="Resultado de K-medias 3 clusters", pch=20, cex =2)
> |
```



Clustering en R

Fuente: Elaboración propia

Comparemos los resultados del clustering versus los datos originales:

```
table(irisCluster$cluster,iris$Species)
```

```
> table(irisCluster$cluster, iris$Species)

    setosa versicolor virginica
1        0         2       36
2        0        48       14
3       50         0        0
> |
```

Clustering VS datos originales

Fuente: Elaboración propia

Como podemos observar, el algoritmo clasificó en versicolor 2, en el cluster 1 y las 36 restantes virginica, y el cluster 2 tiene 48 versicolor y 14 virginica.

8. GESTIÓN DEL VALOR DEL CLIENTE

En tu empresa, al igual que en cualquier otra compañía con gran actividad comercial, es necesario hacer una gestión adecuada de los clientes. No es lógico asignar los mismos recursos a un cliente que nos hace ganar mucho dinero, que a otro que incluso puede provocarnos pérdidas por retrasos o impagos, compras de poco importe, consumo de recursos comerciales, etc.

Sin duda, una de las aplicaciones más interesantes para una empresa es analizar los clientes y, en el caso de tu compañía es especialmente relevante, dado que como los márgenes son bajos, no podéis cometer los errores de asignar a recursos a un cliente que no es rentable.

La principal estrategia para gestionar el valor de los clientes es clasificarlos en función de la rentabilidad actual que generar para nuestra empresa y realizar las proyecciones en cuanto al crecimiento potencial de los clientes existentes.

La gestión del valor del cliente implica realizar análisis de crecimiento de clientes basados en las inversiones y en las previsiones de crecimiento por cliente, con análisis de los clientes con más posibilidades de crecimiento, análisis de rentabilidad por cliente, análisis de los clientes más rentables, y análisis y clasificaciones según la satisfacción del cliente.

Una de las formas más sencillas de hacer este análisis sobre los clientes es aplicar el Principio de Pareto o Principio del 20/80. A pesar de que se trata de un razonamiento muy sencillo, la realidad es que es eficaz.

Este Principio quiere decir que el 20% de mejores los clientes de la empresa, generan un 80% de los beneficios de la misma y, por tanto, debemos centrarnos en su gestión para maximizar el beneficio.

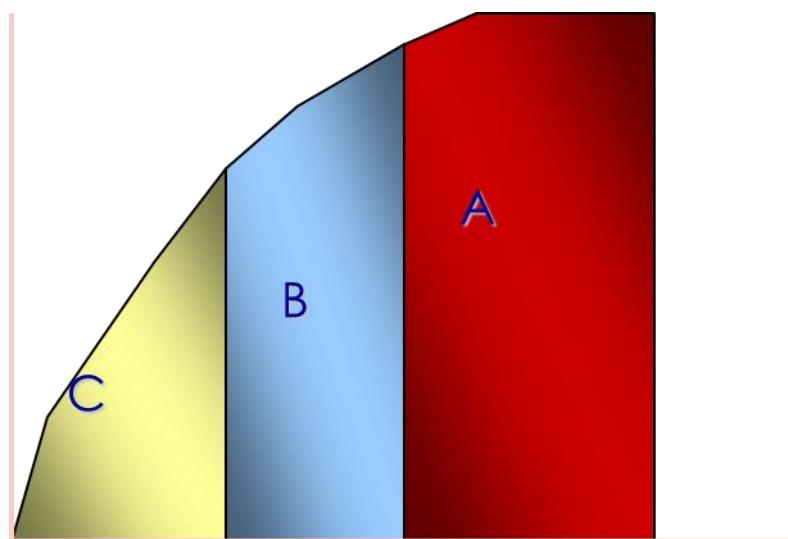
Un mecanismo un poco más avanzado que el de Pareto es el denominado ABC, donde en vez de haber dos grupos de clientes como ocurre en el Principio del 20/80, ahora tenemos tres grupos de clientes: A, B y C, de forma que:

- Clientes tipo A: Son los clientes a fidelizar, porque son los más rentables.
- Clientes de tipo B: Dentro de una escala ABC clásica de cartera de clientes, los clientes tipo B aportan valor al negocio, pero no suele resultar rentable procurar su fidelización a través de la atención personalizada de un gestor de cuentas.

Pero son clientes interesantes, por lo que la tecnología nos puede ayudar mucho a la gestión de su valor. Lo que se pretende es conseguir

la personalización de la relación utilizando tecnologías de la información para aumentar la frecuencia y la calidad de las interacciones con los clientes, intentando crear valor a través de la adaptación de la oferta a partir de las preferencias detectadas y contrastadas del cliente.

- Clientes tipo C: Son clientes a descartar.



Método ABC de clasificación de clientes por valor
Fuente: Elaboración propia

9. INTRODUCCIÓN A LENGUAJE R Y WEKA

Mientras tú has estado coordinando a todos los equipos de trabajo de tu departamento, al mismo tiempo que has mantenido contacto permanente con los responsables del resto de áreas de la compañía para pulsar y conocer de primera mano sus necesidades, tu equipo ha estado concentrado en mejorar sus conocimientos de machine learning y cómo aplicarlo.

Este estudio os permite estar en la mejor disposición de obtener valor aplicando las dos herramientas elegidas para comenzar a introducir capacidades de machine learning e incluso de otros tipos de análisis en la empresa que son Weka y R.

Lo mejor de todo es que son herramientas open source, así que por esa parte no os va a generar un coste financiero en forma de licencias, por lo que reservas parte de ese presupuesto para otro tipo de acciones.

Con Weka tendréis un entorno integrado de algoritmos que os permitirán implantar modelos predictivos con machine learning, hacer business intelligence avanzado, etc.

Con R, contaréis con una herramienta para que vuestros equipos de desarrolladores puedan hacer programas a medida para ofrecer la funcionalidad demanda por toda la empresa y que no pueda ser cubierta con un paquete como Weka.

La verdad es que estás bastante satisfecho con la recomendación que te han ofrecido y te parece que el proyecto de mejora que has puesto en marcha ahora va tomando mucha más forma y que es posible que en un plazo razonablemente corto de tiempo podrás ya comenzar a entregar los primeros resultados a los tomadores de decisiones en tu compañía.

Además del estudio de conceptos básicos, su origen, cómo nace la necesidad de estos, así como sus principales rasgos y elementos importantes, es imprescindible revisar las herramientas mediante las cuales pondremos en práctica estos conceptos.

Las herramientas seleccionadas serán WEKA y el lenguaje R. La primera es una herramienta de estudio estadístico, bastante difundida en modalidad de software libre, por lo que la podremos descargar sin ningún coste desde la página del proyecto. La segunda se trata del lenguaje R, que es una de las herramientas más usadas para el machine learning y data mining y que podremos también descargar fácilmente de la web de su proyecto.

9.1 Introducción a WEKA

WEKA, aunque tiene el nombre de un ave de origen neozelandés, sus siglas significan Waikato Environment for Knowledge Analysis, que vendría a traducirse como entorno Waikato para el análisis de conocimiento. Esta herramienta nace en 1993, en la Universidad de Waikato, en Nueva Zelanda.

Weka es un software para data mining desarrollado en Java, en el cual podremos aplicar un conjunto de algoritmos de machine learning directamente a un conjunto de datos, o podremos hacer uso de los nuestros propios programados en Java.



Logotipo Weka

Fuente: <https://www.cs.waikato.ac.nz/ml/weka/>

Weka contiene herramientas para las siguientes tareas:

- El procesamiento y la limpieza de los datos
- Clasificación
- Regresión
- Clustering
- Reglas de asociación
- Visualización y generación de informes
- Deep learning



SABÍAS QUE...

Los estudiantes de la universidad de Waikato, como parte de su curso de data mining, participaron de la creación de Weka. Originalmente estaba desarrollado en Tcl/Tk, después en C alrededor del año 1993, finalmente, en 1997, se decidieron a rehacerlo desde cero escribiendo el código en Java.

9.2 Instalación

Ahora veamos, paso a paso, como instalar WEKA. Los pasos aquí mencionados son equivalentes para todas las plataformas, sea Linux, Windows o IOS, pues WEKA está desarrollada en Java. Nosotros lo haremos sobre Windows 10.

← → C waikato.github.io/weka-wiki/downloading_weka/#windows

Aplicaciones Bookmarks Detected software... En Using jQuery to Co... ifixit - Buscar con G... iFixit: The free repai... Pasos para solicitar... CyanogenMod 7 fo...

Weka Wiki

Search docs

Home
Downloading and installing Weka
Snapshots
Stable version
Windows
Mac OS
Linux
Other platforms
Developer version
Windows
Mac OS
Linux
Other platforms
Old versions
Upgrading from Weka 3.7
Requirements
Documentation
Getting help
Citing Weka
Literature
Development
History
Resources

stable branch. Those who want the latest bug fixes before the next official release is made can download these [snapshots](#).

Stable version

Weka 3.8 is the latest stable version of Weka. This branch of Weka only receives bug fixes and upgrades that do not break compatibility with earlier 3.8 releases, although major new features may become available in packages. There are different options for downloading and installing it on your system:

Windows

- Click [here](#) to download a self-extracting executable for 64-bit Windows that includes Azul's 64-bit OpenJDK Java VM 11 (weka-3-8-4-azul-zulu-windows.exe; 118 MB)

This executable will install Weka in your Program Menu. Launching via the Program Menu or shortcuts will automatically use the included JVM to run Weka.

Mac OS

- Click [here](#) to download a disk image for Mac OS that contains a Mac application including Azul's 64-bit OpenJDK Java VM 11 (weka-3-8-4-azul-zulu-osx.dmg; 144 MB)

Linux

- Click [here](#) to download a zip archive for Linux that includes Azul's 64-bit OpenJDK Java VM 11 (weka-3-8-4-azul-zulu-linux.zip; 129 MB)

First unzip the the zip file. This will create a new directory called weka-3-8-4. To run Weka, change into that directory and type

```
./weka.sh
```

Instalación Weka

Fuente: Elaboración propia

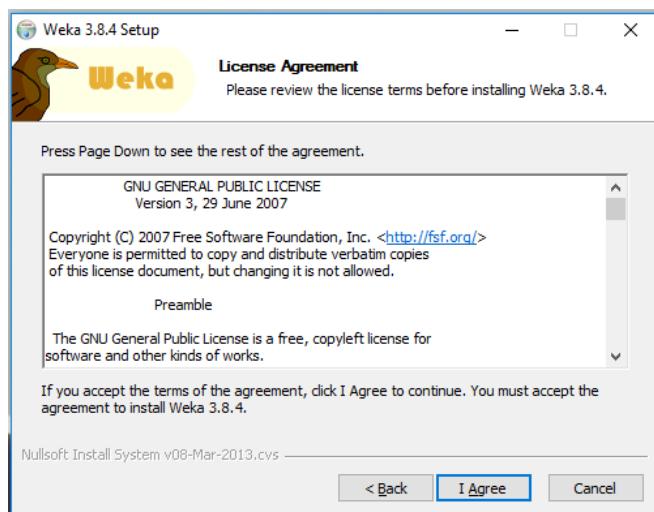
A continuación, debemos ejecutar el archivo descargado, que en este caso tiene unos 118 MB y se abrirá el siguiente diálogo con una asistente de instalación, donde debemos pulsar **Next**.



Instalación Weka. Next

Fuente: Elaboración propia

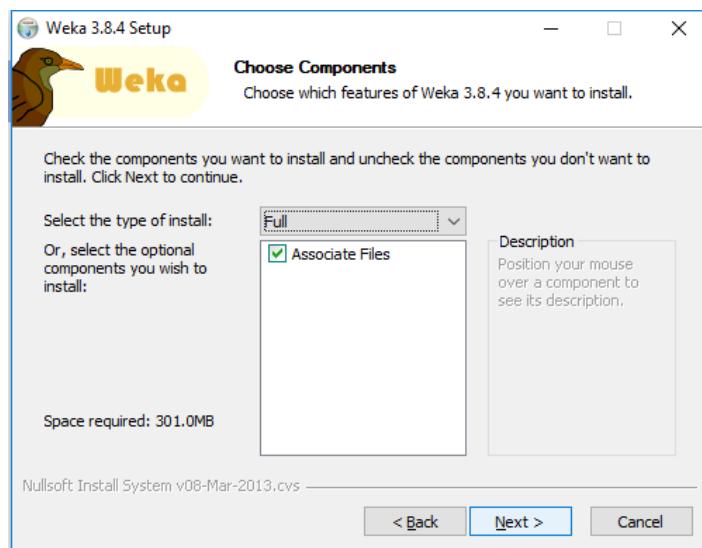
Debemos seguir con los pasos que nos muestra el asistente, aceptando los términos de uso, haciendo clic en el botón **I Agree**:



Instalación Weka. I Agree

Fuente: Elaboración propia

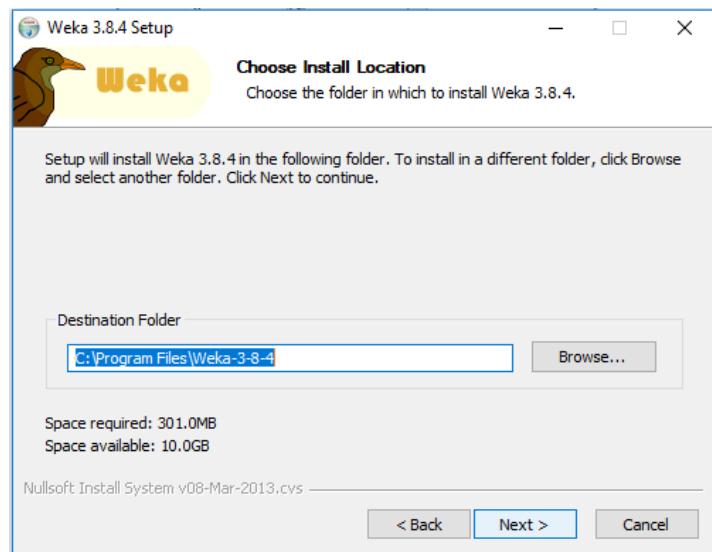
A continuación, seleccionamos el tipo de instalación Full y presionamos el botón **Next**.



Instalación Weka. Instalación full

Fuente: Elaboración propia

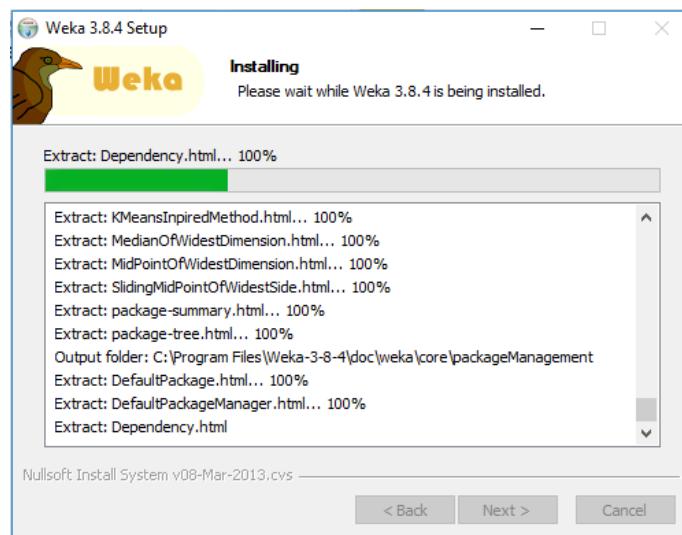
Elegiremos la carpeta por defecto y apretamos el botón **Next**.



Instalación Weka. Elección de carpeta

Fuente: Elaboración propia

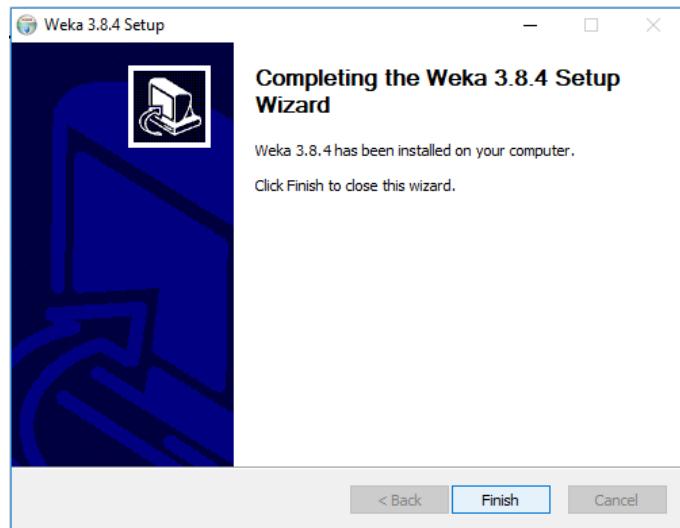
De aquí en adelante dejamos las opciones por defecto y presionamos **Next** en los siguientes diálogos para continuar la instalación.



Instalación Weka. Opciones por defecto

Fuente: Elaboración propia

Posteriormente, si todo ha salido bien, tendremos el siguiente diálogo:



Instalación Weka. Finish

Fuente: Elaboración propia

Ya tenemos WEKA instalado y listo para hacer nuestras pruebas.



ENLACE DE INTERÉS

Para poder instalar WEKA, debemos tener previamente instalado JVM (Java Virtual Machine). Accede a estos enlaces para la descarga de ambos:

- JVM (Java Virtual Machine) <https://www.java.com/es/download/>
- Weka (versión 3.8.4) https://waikato.github.io/weka-wiki/downloading_weka/

9.3 Uso

Cuando ya tengamos instalado WEKA, el siguiente paso es aprender a usarlo. Cuando abrimos la aplicación, se nos muestra el siguiente menú cuyas funciones principales se resumen a continuación:



Instalación Weka. Funciones principales

Fuente: Elaboración propia

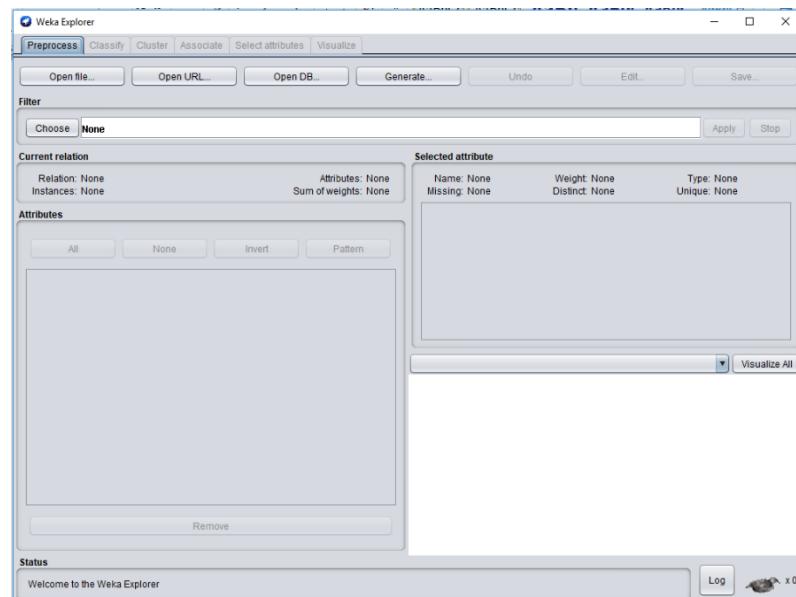
Acción	Descripción
Explorer	Nos abrirá un explorador, con el cual podremos navegar y buscar los datos en distintos formatos.
Experiment	Esta sección nos servirá para llevar a cabo experimentos con nuestros datos, tales como el uso de algoritmos y analizar los mismos.
Knowledge flow	Es una interfaz concebida para diseñar configuraciones para el procesamiento de datos en streaming.
Workbench	Es una interfaz que unifica las tres anteriores. Además de ello permite el uso de plugins una vez estén instalados.
Simple CLI	Nos muestra una consola de comandos donde podremos introducir ordenes sobre los datos directamente.

Funciones principales. Weka

Fuente: Elaboración propia

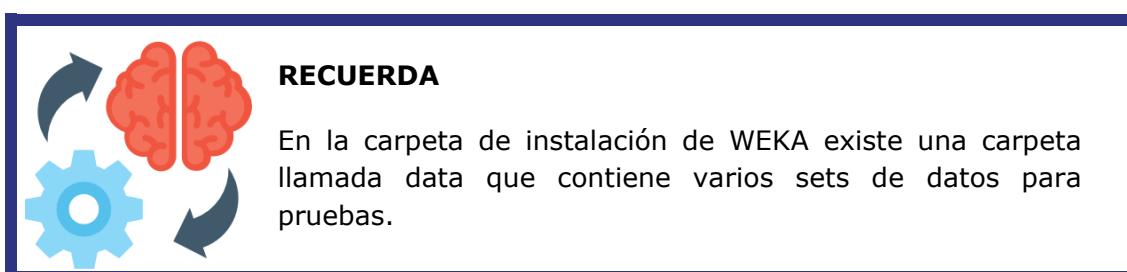
De momento, solo usaremos la opción de Explorer, ya que esta cubrirá la mayoría de nuestras necesidades para hacer data mining.

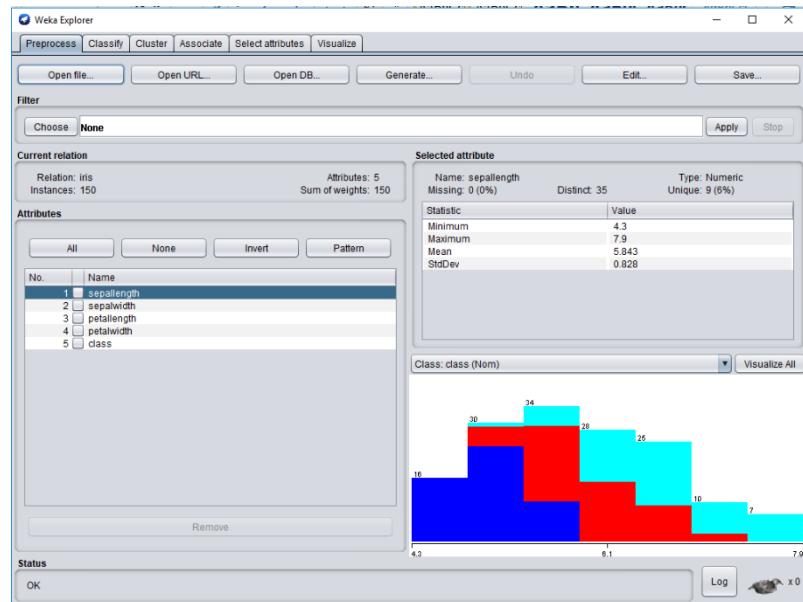
Seleccionaremos la opción de Explorer y nos mostrará una ventana como la siguiente:



Explorer de Weka
Fuente: Elaboración propia

En la pestaña de **Preprocess**, pulsamos en Open file para buscar un archivo llamado **iris.arff** y le indicamos Open.

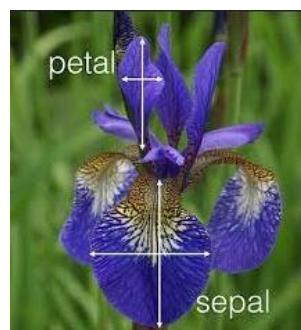




Explorer de Weka. Pestaña de Preprocess

Fuente: Elaboración propia

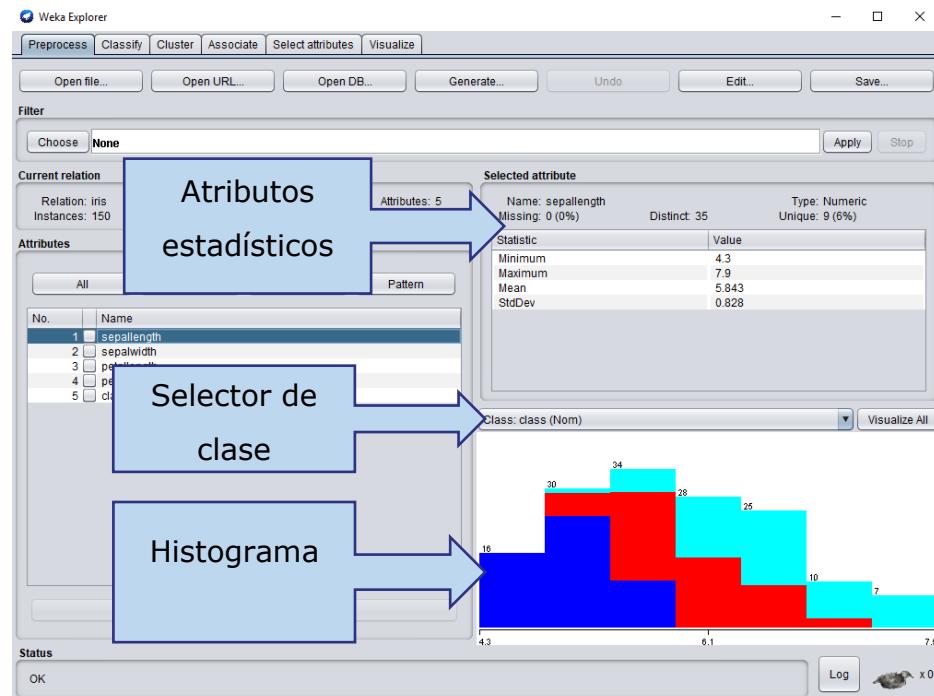
Este conjunto de datos que hemos abierto es muy conocido el mundo del data mining para hacer pruebas y estudios. Se trata de una colección de datos de cincuenta muestras de tres especies distintas de la flor iris de las cuales se midieron cuatro rasgos de las mismas, concretamente el largo y ancho tanto de sépalos como de los pétalos.



Flor iris. Colección de datos

Fuente: Elaboración propia

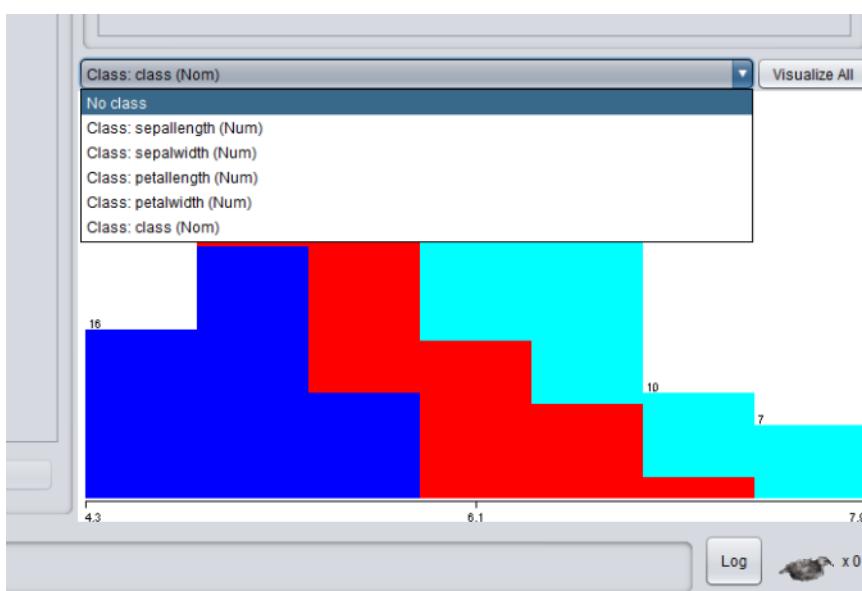
Veamos algunos elementos de la pantalla del explorador:



Elementos de la pantalla Explorer de Weka

Fuente: Elaboración propia

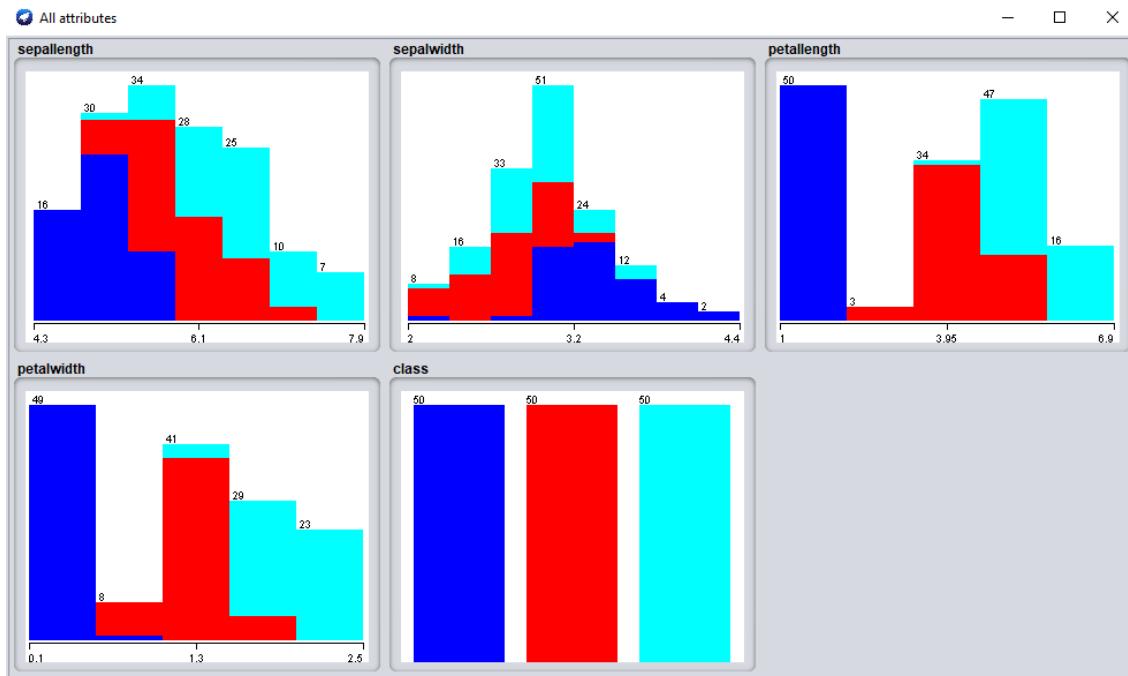
Como vemos en el selector de clase, podemos asignar cualquiera de los parámetros y WEKA selecciona por defecto el atributo class. Esto lo podemos cambiar en cualquier ocasión, según nuestros requerimientos en el momento de estudiar los datos.



Explorer de Weka. Atributo class

Fuente: Elaboración propia

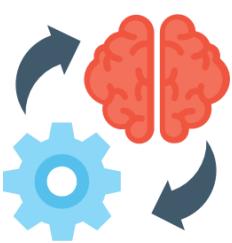
Después tenemos una sección que nos presenta un histograma, que cambiará según el selector de clases. Como podemos ver, en el histograma se representan los valores de las muestras y se marcan con colores según la clasificación de su clase. Si damos clic en el botón **Visualize All**, nos mostrará una ventana con todos los histogramas según los atributos.



Explorer de Weka. Histogramas

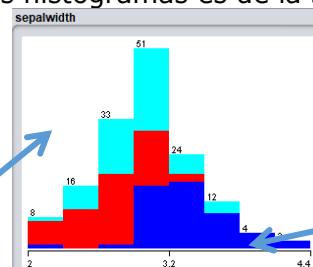
Fuente: Elaboración propia

Viendo estos histogramas podemos ver cuánto de diferentes son nuestras muestras. Si prestamos atención, podremos ver que da la impresión de que la especie Iris Setosa tiende a tener una longitud menor en el largo del **sépalo** y **pétalo**. En el ancho del **pétalo** también es de los menores. Por otra parte, el ancho del **sépalo** es bastante mayor y está representada por el color azul.



RECUERDA

La lectura de los histogramas es de la siguiente forma:



Siendo el color azul las muestras representadas por la clase Iris Setosa, podemos ver que la mayoría de las muestras están en valores cercanos de 3.2 y mayores.

A continuación, revisemos la sección de atributos estadísticos.

Name: sepalwidth	Missing: 0 (0%)	Distinct: 35	Type: Numeric	Unique: 9 (6%)
Statistic				
Minimum				
Maximum				
Mean				
StdDev				

Explorer de Weka. Atributos estadísticos

Fuente: Elaboración propia

Primero tenemos el nombre del atributo y a continuación vemos el valor **missing**: 0. Esto nos quiere decir que no hay ninguna muestra que no tenga ese valor, es decir, no hay datos faltantes.

Abajo tenemos los valores básicos de estadística, tales como mínimos (**minimun**) y máximos (**maximun**). También está la desviación estándar (**StdDev**). A continuación, en la parte superior, podemos ver dos valores que son **Distinct** y **Unique**.

El valor **Distinct** nos dice cuántos valores distintos fueron tomados por un cierto atributo, es decir, en nuestro ejemplo tenemos que el *sepallength* (largo del sépalo) tiene su valor a 35. Esto quiere decir que, en el total de las 150 muestras, existen 35 valores distintos.

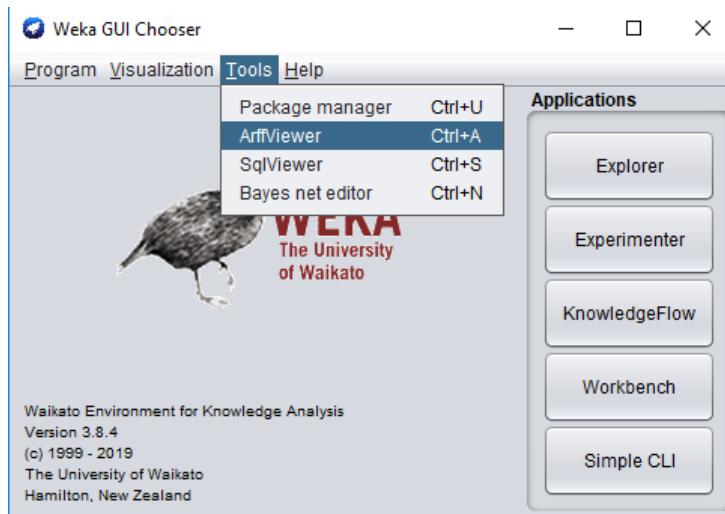
En el caso de **Unique** nos dice todos aquellos valores únicos. En nuestro caso son 9 los que poseen valores únicos en los 150 casos de muestra.

Veamos este concepto más claramente con un ejemplo:

Con los siguientes ocho valores: 26, 30, 50, 51, 26, 45, 30, 90. Los valores únicos y distintos son:

- Distinct: 5 (26,30,50,26,45,90)
- Unique: 4 (50,51,45,90)

Ahora, revisaremos el visor ARFF. Con este visor seremos capaces de revisar el conjunto de los datos sin cargarlos.



Visor ARFF de Weka

Fuente: Elaboración propia

También lo podemos invocar presionando la tecla **ctrl + a**. Esto nos abre una ventana y si le damos en File y abrimos nuevamente el archivo de los datos de la flor Iris, veremos la siguiente ventana:

ARFF-Viewer - C:\Program Files\Weka-3-8-4\data\iris.arff

File Edit View

iris.arff

Relation: iris

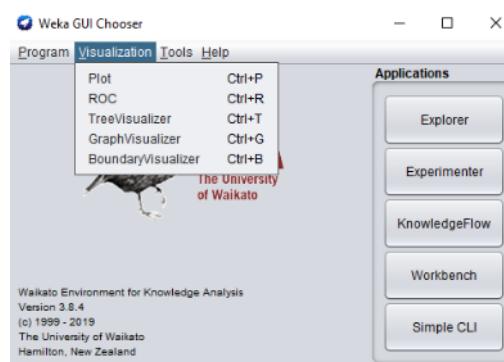
No.	1. sepal length	2. sepal width	3. petal length	4. petal width	5. class
	Numeric	Numeric	Numeric	Numeric	Nominal
1	5.1	3.5	1.4	0.2	Iris-S...
2	4.9	3.0	1.4	0.2	Iris-S...
3	4.7	3.2	1.3	0.2	Iris-S...
4	4.6	3.1	1.5	0.2	Iris-S...
5	5.0	3.6	1.4	0.2	Iris-S...
6	5.4	3.9	1.7	0.4	Iris-S...
7	4.6	3.4	1.4	0.3	Iris-S...
8	5.0	3.4	1.5	0.2	Iris-S...
9	4.4	2.9	1.4	0.2	Iris-S...
10	4.9	3.1	1.5	0.1	Iris-S...
11	5.4	3.7	1.5	0.2	Iris-S...
12	4.8	3.4	1.6	0.2	Iris-S...
13	4.8	3.0	1.4	0.1	Iris-S...
14	4.3	3.0	1.1	0.1	Iris-S...
15	5.8	4.0	1.2	0.2	Iris-S...
16	5.7	4.4	1.5	0.4	Iris-S...
17	5.4	3.9	1.3	0.4	Iris-S...
18	5.1	3.5	1.4	0.3	Iris-S...
19	5.7	3.8	1.7	0.3	Iris-S...
20	5.1	3.8	1.5	0.3	Iris-S...
21	5.4	3.4	1.7	0.2	Iris-S...
22	5.1	3.7	1.5	0.4	Iris-S...
23	4.6	3.6	1.0	0.2	Iris-S...
24	5.1	3.3	1.7	0.5	Iris-S...
25	4.8	3.4	1.9	0.2	Iris-S...
26	5.0	3.0	1.6	0.2	Iris-S...
27	5.0	3.4	1.6	0.4	Iris-S...
28	5.2	3.5	1.5	0.2	Iris-S...
29	5.2	3.4	1.4	0.2	Iris-S...
30	4.7	3.2	1.6	0.2	Iris-S...
31	4.8	3.1	1.6	0.2	Iris-S...
32	5.4	3.4	1.5	0.4	Iris-S...
33	5.2	4.1	1.5	0.1	Iris-S...
34	5.5	4.2	1.4	0.2	Iris-S...
35	4.9	3.1	1.5	0.1	Iris-S...
36	5.0	3.2	1.2	0.2	Iris-S...
37	5.5	3.5	1.3	0.2	Iris-S...
38	4.9	3.1	1.5	0.1	Iris-S...
39	4.4	3.0	1.3	0.2	Iris-S...

Visor ARFF de Weka. Flor de iris

Fuente: Elaboración propia

Entre los detalles que podemos visualizar está que los valores de la columna cinco son detectados como de tipo nominal y los otros de tipo numérico.

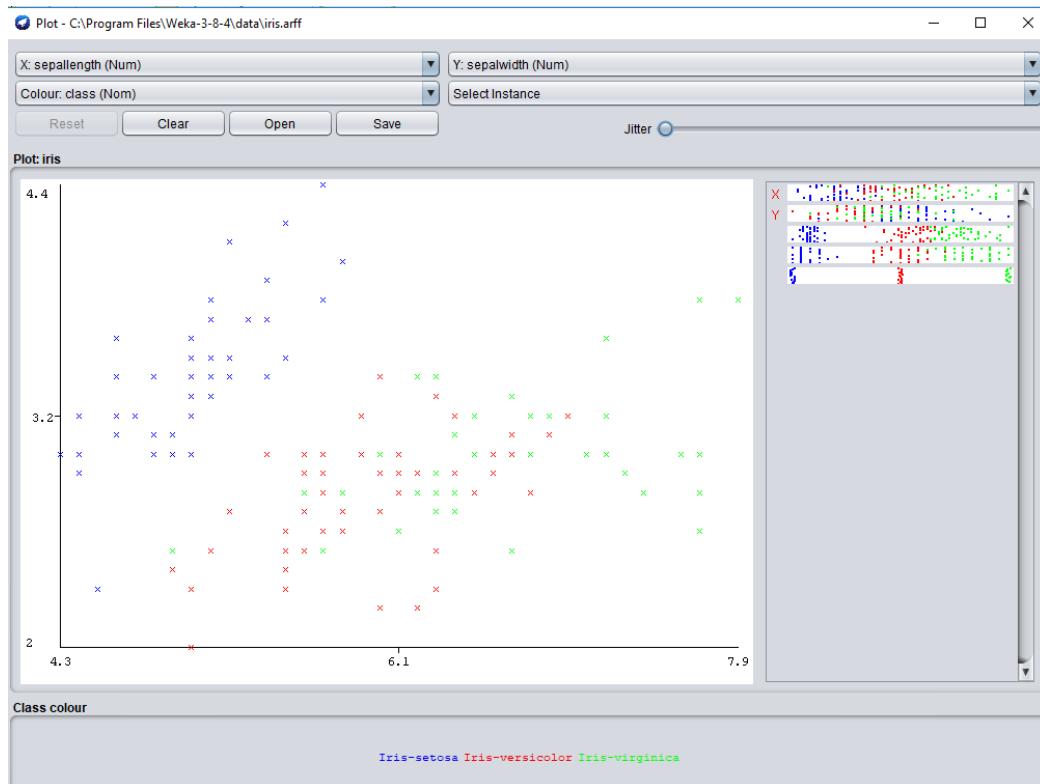
Otra de las herramientas interesantes está en el menú visualization (visualización):



Visualization Weka

Fuente: Elaboración propia

Aquí podemos graficar los datos usando la opción **Plot**, la cual a continuación nos abre una ventana de selección de archivo y buscamos el conjunto de datos de la flor de iris en la carpeta data:



Graficar datos. Opción Plot

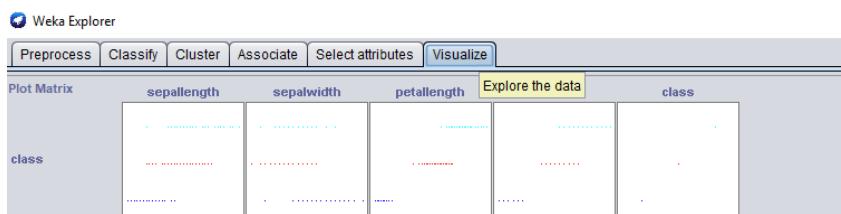
Fuente: Elaboración propia

En WEKA este tipo de gráficos es muy útil para ayudar a hacerse una idea de cómo se comportan los datos. Cuando se hace difícil ver los datos existe una opción llamada **jitter**. Esta opción genera cierto nivel de ruido en los datos aleatoriamente, para lograr que aquellos puntos que tal vez estaban ocultos por otros, sean visibles.



SABÍAS QUE...

El grafico de plot también lo puedes hacer desde la ventana de explorer:



9.4 Introducción al lenguaje R

Lenguaje R es un lenguaje de programación que se usa para crear código estadístico y fue creado en 1993 en la Universidad de Auckland en Nueva Zelanda. Su denominación viene dada por la primera letra de los nombres de sus creadores Robert Gentleman y Ross Ihaka.

Puede ser descargado gratuitamente de la página del proyecto y es distribuido bajo la licencia de software libre GNU, estando disponible para Windows, Apple y Linux.

Entre sus principales características están:

- Múltiples herramientas estadísticas, tales como modelos lineales y no lineales, test estadísticos, análisis de series temporales, algoritmos de agrupación y clasificación, etc...
- Es un lenguaje orientado a objetos.
- Posee alta capacidad de generar gráficos.
- Compatibilidad con numerosos gestores de bases de datos.
- Posee su propio formato para documentación basado en LaTex.
- Se puede usar como herramienta de cálculo numérico.
- Existe una interfaz llamada RWeka para interactuar con WEKA.



PARA SABER MÁS

LaTeX es un sistema de composición de textos orientado a la creación de documentos utilizado especialmente en los ámbitos académicos científicos.

<https://www.latex-project.org/get/>

9.5 Instalación

Para la puesta en marcha de R lo primero que debemos hacer es descargar los archivos de instalación.

Hay múltiples opciones para descargar. Existen muchos mirrors así que podemos elegir el que sea más conveniente para nosotros. Tras esto, nos llevará a la página del proyecto, donde podremos elegir la versión para nuestro sistema operativo.

The Comprehensive R Archive Network

Download and Install R

Precompiled binary distributions of the base system and contributed packages. **Windows and Mac** users most likely want one of these versions of R:

- [Download R for Linux](#)
- [Download R for \(Mac\) OS X](#)
- [Download R for Windows](#)

R is part of many Linux distributions, you should check with your Linux package management system in addition to the link above.

Source Code for all Platforms

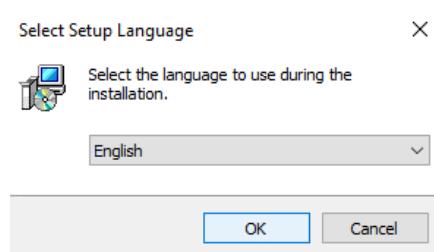
Windows and Mac users most likely want to download the precompiled binaries listed in the upper box, not the source code. The sources have to be compiled before you can use them. If you do not know what this means, you probably do not want to do it!

- The latest release (2020-02-29, Holding the Windsock) [R-3.6.3.tar.gz](#), read [what's new](#) in the latest version.
- Sources of [R alpha and beta releases](#) (daily snapshots, created only in time periods before a planned release).
- Daily snapshots of current patched and development versions are [available here](#). Please read about [new features and bug fixes](#) before filing corresponding feature requests or bug reports.
- Source code of older versions of R is [available here](#).
- Contributed extension [packages](#)

Web de R

Fuente: <https://cran.r-project.org/>

La versión 3.6.3, la más reciente en este momento, pesa unos 82.4 MB. Después de descargarla, la abrimos e instalamos:

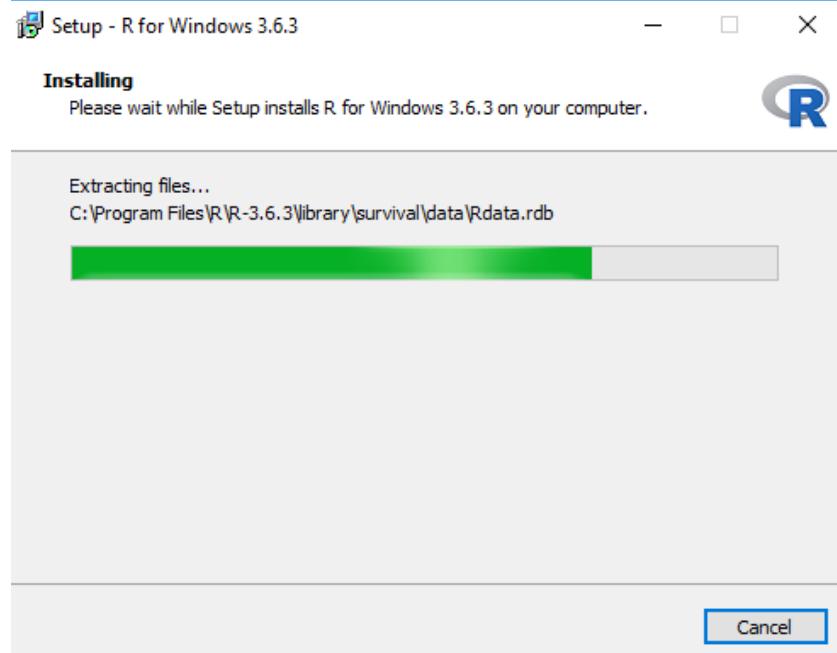


Instalación de R

Fuente: Elaboración propia

Elegimos el lenguaje de nuestra preferencia, aunque es altamente recomendable trabajar en inglés, puesto que mucha documentación y libros electrónicos para R solo están disponibles en este idioma.

Aceptamos todos los siguientes diálogos, con sus opciones por defecto, pulsando el botón **Next**.



Instalación de R II
Fuente: Elaboración propia

Tras esto finalizamos nuestra instalación y podremos utilizar la consola del lenguaje R.



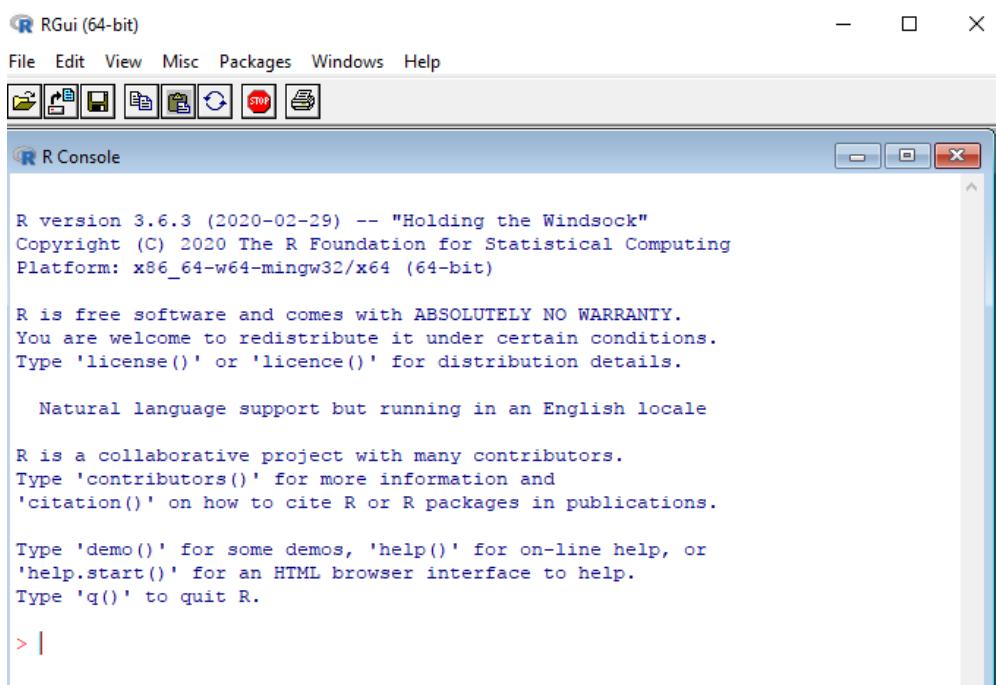
ENLACE DE INTERÉS

Puedes descargar los archivos de instalación de R en la siguiente url:

<http://cran.r-project.org/mirrors.html>

9.6 Uso

Abrimos la consola ubicada en el menú de inicio:



Consola de R
Fuente: Elaboración propia

Dependiendo de nuestras necesidades, podremos querer escribir un script en un archivo separado o hacer uso de la consola de comandos.

En lenguaje R existen varias maneras de asignar valores a las variables y se puede usar el igual '=' , '<- ' y '>'

Valor de Variable -> Nombre de variable

Nombre de variable <- Valor de variable

Nombre de variable = Valor de variable

R Gui (64-bit)

File Edit View Misc Packages Windows Help

R Console

```
R version 3.6.3 (2020-02-29) -- "Holding the Windsock"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> var1 = "hola"
> "mundo" -> var2
> var3 <- "!!"
> cat(var1, var2, var3)
hola mundo !!> |
```

Consola de R II
Fuente: Elaboración propia

Para imprimir el valor de una variable podemos usar la función **print** (variable) y para imprimir varias variables la función **cat** (var1,var2...)

9.7 Tipos de variables en R

En R no se asignan tipos de variables como en los otros lenguajes. El tipo de la variable viene dado por lo que le asignemos a esta, es decir, si a una variable le asigno un número, esta se volverá de tipo *numeric* y así sucesivamente.

Para determinar el tipo de una variable tenemos la función **class()**

```
> class(var1)
[1] "character"
```

Función class()
Fuente: Elaboración propia

Los tipos de Variables mas usados son:

Tipo	Descripción	Ejemplo
Character	Son las letras o símbolos	"hola","una frase"
Numeric	Almacena los números reales o decimales	25, 23.5
Integer	números enteros	1,2,3 ..99...
Logical	Se crea por la comparación de variables	TRUE, FALSE
Complex	números imaginarios, funciones, etc.	3+5i

Tipos de variables
Fuente: Elaboración propia

9.8 Operadores

En lenguaje R tenemos distintos tipos de operadores y están clasificados de la siguiente manera:

Tipo	Operadores
Aritmético	+, -, *, /, %%, ^
Relacional	<, >, <=, >=, !=, ==
Lógico	&, , &&, , !
Asignación	=, ->, <-

Tipos de Operadores
Fuente: Elaboración propia

Operadores aritméticos:

- + suma
- - resta
- * multiplicación
- / división
- %% modulo
- ^ exponenciación

```
> x<- 25
> y<- 7
> x+y
[1] 32
> x-y
[1] 18
> x%%y
[1] 4
> x^y
[1] 6103515625
> |
```

Operadores aritméticos

Fuente: Elaboración propia

Operadores relacionales:

- < menor que
- > mayor que
- <= menor igual que
- >= mayor igual
- != diferente que
- == igual que

```
> x<- 25
> y<- 7
> x<y
[1] FALSE
> x>y
[1] TRUE
> x==y
[1] FALSE
> |
```

Operadores relacionales

Fuente: Elaboración propia

Operadores lógicos:

- ! no
- && And
- || Or
- & es un and sabio
- | es un or sabio

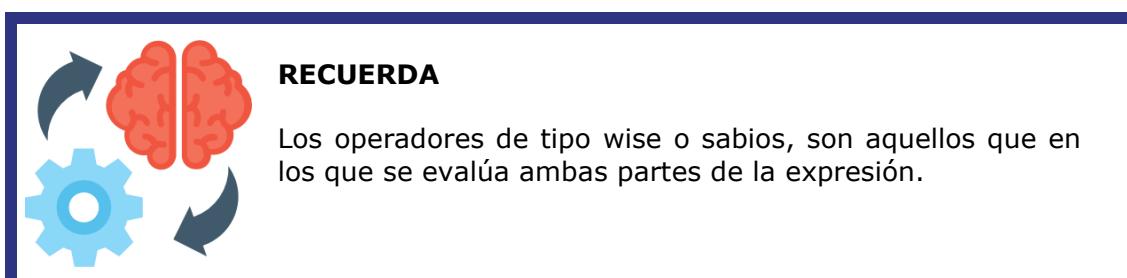
```

> a = 10
> b = 20
> a > b && b < a
[1] FALSE
> a > b & b < a
[1] FALSE
> a < b & b > a
[1] TRUE
>

```

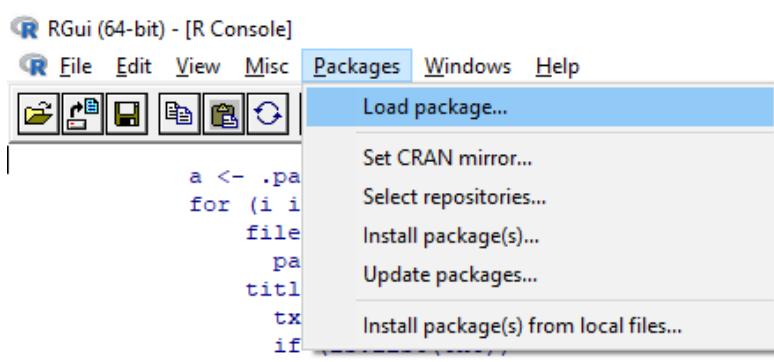
Operadores lógicos

Fuente: Elaboración propia



9.9 Instalando paquetes en R

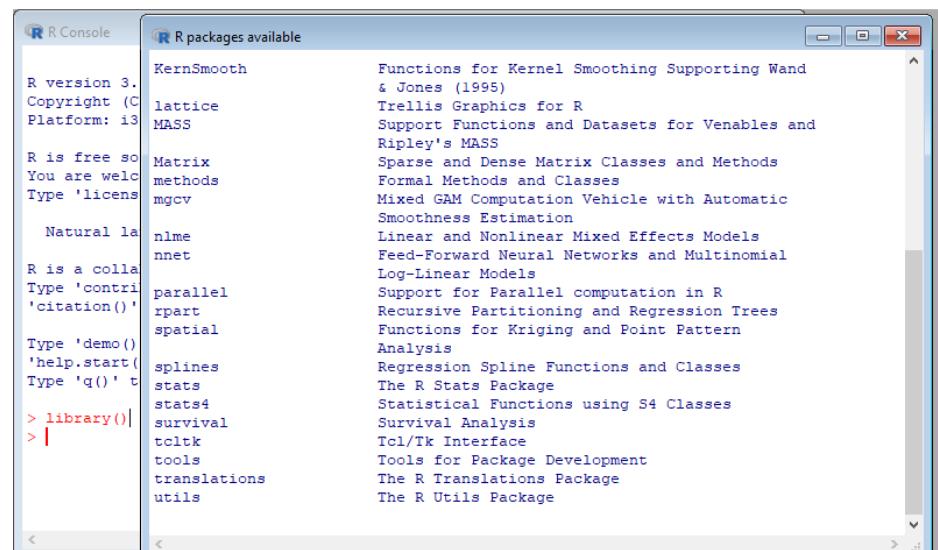
En lenguaje R los paquetes son un conjunto de librerías con herramientas que nos ayudarán en ciertas tareas. Éstas se guardan en la carpeta library del entorno de R. Para gestionar los paquetes instalados tenemos una opción en el menú:



Packages. Instalación paquetes R

Fuente: Elaboración propia

Para instalar paquetes nuevos damos clic en la opción **install packages**. También con el comando **library()** podremos ver los paquetes que actualmente están instalados:



Install packages

Fuente: Elaboración propia

Muchas veces necesitaremos instalar nuevas librerías para ir más allá de lo que nos ofrece R por defecto. Para ello haremos uso del comando **install.package("nombre del paquete")** o también podemos conseguirlo a través de la interfaz gráfica.

Para instalar, por ejemplo, el paquete **CORElearn** escribimos **install.package("CORElearn")**

Con ese comando nos aparecerá una ventana con los servidores mirrors. Simplemente elegimos uno cercano, tras lo cual se realizará la descarga del mismo.

RGui (32-bit)

File Edit View Misc Packages Windows Help

R Console

```
> install.packages('CORElearn')
Installing package into 'C:/Users/sebastian/Documents/R/win-library/3.6'
(as 'lib' is unspecified)
also installing the dependencies 'plotrix', 'rpart.plot'

trying URL 'https://www.icesi.edu.co/CRAN/bin/windows/contrib/3.6/plotrix_3.7-7.zip'
Content type 'application/zip' length 1132324 bytes (1.1 MB)
downloaded 1.1 MB

trying URL 'https://www.icesi.edu.co/CRAN/bin/windows/contrib/3.6/rpart.plot_3.0.8.zip'
Content type 'application/zip' length 1077736 bytes (1.0 MB)

26% downloaded
URL: ... esi.edu.co/CRAN/bin/windows/contrib/3.6/rpart.plot_3.0.8.zip
[Progress Bar]
```

install.package("CORElearn")
 Fuente: Elaboración propia

Cuando se termine de instalar, obtendremos el siguiente mensaje.

```
package 'rpart.plot' successfully unpacked and MD5 sums checked
package 'CORElearn' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
      C:\Users\sebastian\AppData\Local\Temp\RtmpABWvww\downloaded_packages
> |
```

install.package("CORElearn") II
 Fuente: Elaboración propia

9.10 Carga de datos

Para comenzar a trabajar con R, primero debemos aprender cómo cargar los datos que vamos a utilizar. Debemos destacar que R soporta múltiples tipos de archivos. Para este caso instalaríamos el paquete '**gdata**', que nos permitirá trabajar con archivos de tipo xls, csv y para archivos de tipo arff tenemos 'farff'.

Y su modo de empleo sería:

1. Instalar el paquete gdata: **install.package('gdata')**

```
> install.packages('gdata')
Installing package into 'C:/Users/sebastian/Documents/R/win-library/3.6'
(as 'lib' is unspecified)
also installing the dependency 'gtools'

trying URL 'https://ftp.acc.umu.se/mirror/CRAN/bin/windows/contrib/3.6/gtools_3$'
Content type 'application/zip' length 336453 bytes (328 KB)
downloaded 328 KB

trying URL 'https://ftp.acc.umu.se/mirror/CRAN/bin/windows/contrib/3.6/gdata_2.$'
Content type 'application/zip' length 1264087 bytes (1.2 MB)
downloaded 1.2 MB

package 'gtools' successfully unpacked and MD5 sums checked
package 'gdata' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
      C:\Users\sebastian\AppData\Local\Temp\RtmpoDmbxL\downloaded_packages
> |
```

Instalar el paquete gdata

Fuente: Elaboración propia

2. Cargamos la librería a utilizar: **library(gdata)**

```
> library(gdata)
gdata: read.xls support for 'XLS' (Excel 97-2004) files ENABLED.

gdata: read.xlsx support for 'XLSX' (Excel 2007+) files ENABLED.

Attaching package: 'gdata'

The following object is masked from 'package:stats':
  nobs

The following object is masked from 'package:utils':
  object.size

The following object is masked from 'package:base':
  startsWith

> |
```

library(gdata)

Fuente: Elaboración propia

3. Ahora, debemos cargar los datos con el método **read**:

```
datos = read.csv("Aquí indicamos la ruta al archivo")
```

4. Podemos verificar que se cargaron los datos haciendo uso de la función **summary(data)**:

```
> summary(datos)
  sepal_length    sepal_width     petal_length     petal_width
  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
  1st Qu.:5.100  1st Qu.:2.800  1st Qu.:1.600  1st Qu.:0.300
  Median :5.800  Median :3.000  Median :4.350  Median :1.300
  Mean   :5.843  Mean   :3.054  Mean   :4.375  Mean   :1.199
  3rd Qu.:6.400  3rd Qu.:3.300  3rd Qu.:5.100  3rd Qu.:1.800
  Max.   :7.900  Max.   :4.400  Max.   :6.900  Max.   :2.500
  species
  setosa    :50
  versicolor:50
  virginica:50
```

summary(data)
Fuente: Elaboración propia

Al igual que con WEKA, R también posee librerías para que trabajemos con datos de prueba de manera más cómoda:

1. Cargamos el paquete datasets: **library(datasets)**
2. Cargamos los datos de iris: **data(iris)**

```
> library(datasets)
> data(iris)
> names(iris)
[1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"  "Species"
> |
```

Librería R
Fuente: Elaboración propia

En la imagen, vemos que usamos la función **names** y esta nos devuelve los nombres de los atributos.

Siguiendo con los datos cargados:

Si queremos saber estadísticas sobre una columna específica, podemos usar el metodo **summary** de la siguiente forma:

summary(iris\$Sepal.Width)

```
> summary(iris$Sepal.Width)
  Min. 1st Qu. Median   Mean 3rd Qu.   Max.
  2.000  2.800  3.000  3.057  3.300  4.400
> |
```

Método summary
Fuente: Elaboración propia

Para revisar los datos en modo de tabla, tenemos el método **View()**:

`View(iris)`

```
> View(iris)
>
```

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa

Método view

Fuente: Elaboración propia

También podremos utilizar las funciones **is.na()** y **unique()** para encontrar los valores faltantes y valores únicos respectivamente.

`is.na(iris)`

```
> is.na(iris)
   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
[1,] FALSE      FALSE      FALSE      FALSE      FALSE
[2,] FALSE      FALSE      FALSE      FALSE      FALSE
[3,] FALSE      FALSE      FALSE      FALSE      FALSE
[4,] FALSE      FALSE      FALSE      FALSE      FALSE
[5,] FALSE      FALSE      FALSE      FALSE      FALSE
[6,] FALSE      FALSE      FALSE      FALSE      FALSE
[7,] FALSE      FALSE      FALSE      FALSE      FALSE
[8,] FALSE      FALSE      FALSE      FALSE      FALSE
[9,] FALSE      FALSE      FALSE      FALSE      FALSE
[10,] FALSE     FALSE      FALSE      FALSE      FALSE
[11,] FALSE     FALSE      FALSE      FALSE      FALSE
[12,] FALSE     FALSE      FALSE      FALSE      FALSE
[13,] FALSE     FALSE      FALSE      FALSE      FALSE
[14,] FALSE     FALSE      FALSE      FALSE      FALSE
[15,] FALSE     FALSE      FALSE      FALSE      FALSE
[16,] FALSE     FALSE      FALSE      FALSE      FALSE
[17,] FALSE     FALSE      FALSE      FALSE      FALSE
[18,] FALSE     FALSE      FALSE      FALSE      FALSE
[19,] FALSE     FALSE      FALSE      FALSE      FALSE
[20,] FALSE     FALSE      FALSE      FALSE      FALSE
```

Función `is.na()`

Fuente: Elaboración propia

No existen datos faltantes, por eso todos los valores devuelven FALSE.

Para los valores únicos usamos el siguiente comando:

`length(unique(iris$Sepal.Width))`

```
> length(unique(iris$Sepal.Width))
[1] 23
```

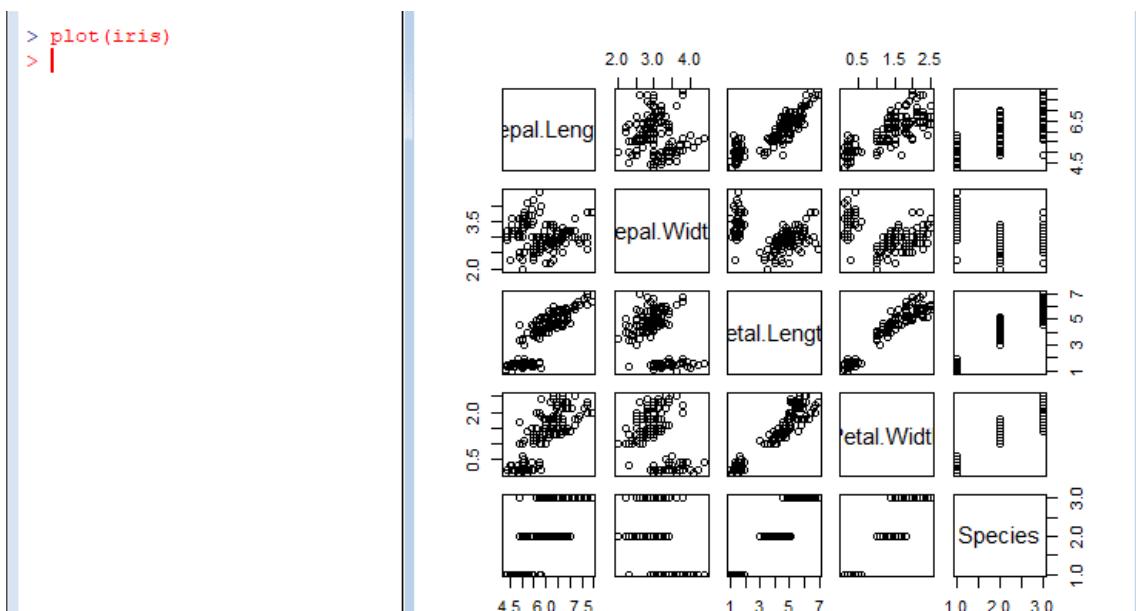
Comando valores únicos

Fuente: Elaboración propia

Vemos que existen 23 valores únicos para el atributo Sepal.Width. La función length se usa debido a que esta devuelve un vector con los valores únicos. Para saber la cantidad usamos length, que nos dice el número de elementos distintos en el vector.

Para los gráficos, usaremos la función **plot()**:

`plot(iris)`

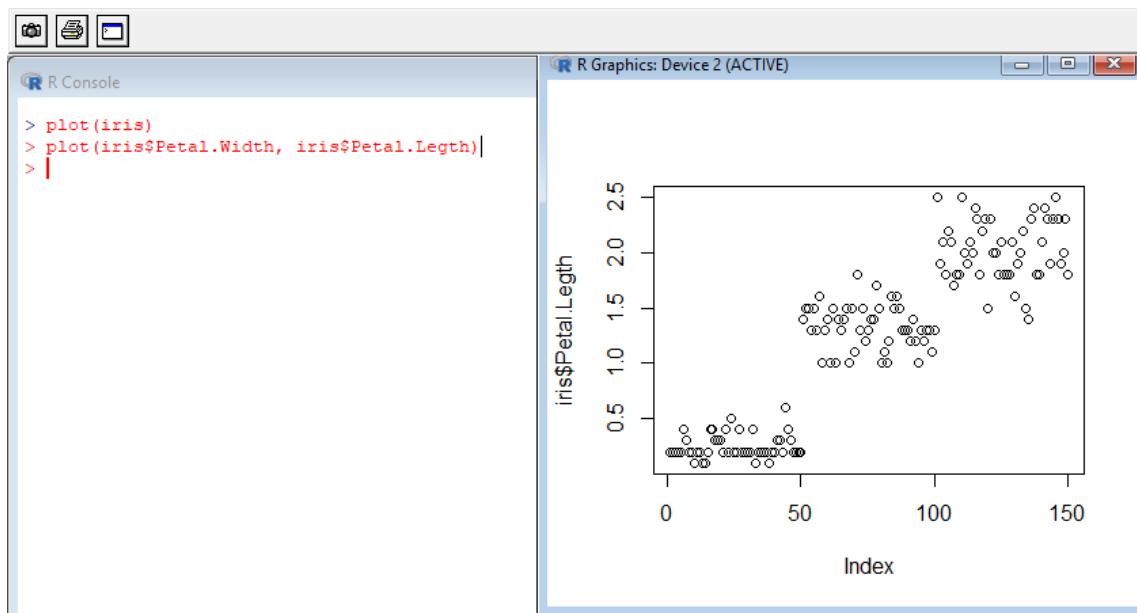


Gráficos de R

Fuente: Elaboración propia

El método **plot** acepta parámetros varios, es decir, podemos pasarle solo las columnas que queremos graficar. Por ejemplo:

`plot(iris$Petal.Width, iris$Petal.Length)`

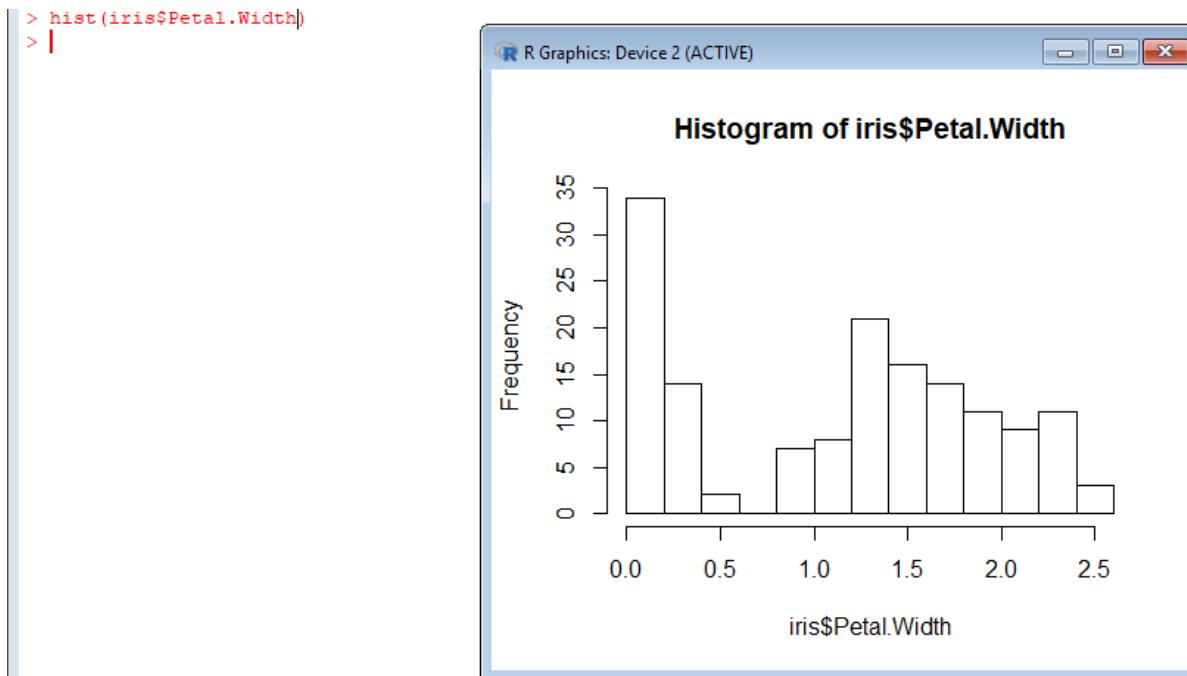


Gráficos de R II

Fuente: Elaboración propia

En cuanto a los histogramas, los podemos hacer con la función **hist()**. De igual manera usamos el comando:

```
hist(iris$Petal.Width)
```



Gráficos de R

Fuente: Elaboración propia



VIDEOTUTORIAL

En el siguiente videotutorial se muestra el uso de Weka

<https://vimeo.com/user64513894/review/447077193/7634d6934c>

10. CLASIFICACIÓN DE DATOS CON WEKA Y R

A medida que han pasado los días y los avances en el proyecto han sido claros y se han cumplido distintos hitos, va llegando el momento de entregar los primeros resultados.

De aquella reunión que mantuviste semanas atrás con los directores de departamento surgieron varias líneas de trabajo y ahora es el momento de presentar al director de marketing y ventas el análisis de compras que habéis realizado en base a toda la información histórica presente en el CRM de la compañía y en la página web. Para ellos habéis diseñado un formato de informe que resulte fácil de leer y sencillo de interpretar donde se muestra cual es la probabilidad de que cada cliente tenga interés en adquirir un determinado producto en el futuro.

Los resultados están siendo prometedores y el director de marketing considera que va a ser una gran herramienta de trabajo. A partir de ahora, el equipo de vendedores, cada vez que hable con un cliente podrá sugerirle la compra de nuevos productos que en el pasado no había comprado.

En unos meses podréis tener la primera validación del mercado, es decir, los resultados de ventas dirán si el análisis ha aportado valor o no.

De forma análoga, desde el departamento de informática habéis creado un informe de tipo crediticio para uso del director financiero. Este informe analiza la trayectoria de pagos de cada cliente, así como otros datos de información disponible sobre cada empresa cliente, como es su facturación, cifra de beneficios, etc. Ahora el director de finanzas dispone de una orientación sobre si puede fiarse o no a la hora de vender a crédito a los respectivos clientes.

En la actualidad se usan las bases de datos con mucha frecuencia para mejorar el proceso de toma de decisiones. Para alcanzar este objetivo, es muy interesante realizar predicciones partiendo de los datos contenidos en la base de datos.

Las dos técnicas más utilizadas para ello son la clasificación y la regresión, métodos que utilizan los datos como punto de partida. Los modelos predictivos de clasificación usan valores discretos o de clase para estas predicciones. De forma análoga, también se usa esta estrategia para la categorización de objetos.

Los problemas de clasificación pueden ser de dos tipos: binarios o multiclase. En la clasificación binaria, los objetos solo pueden tomar uno entre dos valores o pertenecer a una de dos clases. Por ejemplo: ¿un juego se ganará o se perderá?, ¿el valor es positivo o negativo?, etc.

En el multiclase puede haber varios resultados, por ejemplo, una noticia puede ser clasificada en distintos géneros: financiero, social, deportes, entretenimiento etc.

Ahora nos centraremos en entender qué es la clasificación, cómo funciona y qué pasos debemos dar para implementarla en WEKA y en lenguaje R.

10.1 Clasificación

Algunos ejemplos donde pudiéramos aplicar los algoritmos de clasificación serían:

- El análisis del historial crediticio de los clientes de un banco, para calcular si es seguro o de riesgo hacer un préstamo a cada uno de ellos en concreto.
- El análisis de las compras de una tienda, para predecir cuando los clientes compraran cierto producto o no.

En el primer ejemplo, el sistema pudiera predecir haciendo uso de los valores discretos: "seguro" o "de riesgo". Mientras que el segundo, bastaría con un "si" o un "no".

Otros ejemplos podrían ser:

- Cuanto gasto realizará un cliente durante una oferta en nuestra tienda de comercio electrónico.
- A qué sueldo puede aspirar un estudiante recién graduado.

En estos últimos ejemplos, tendríamos como resultado un valor numérico. Por tanto, ambos son ejemplos de preguntas que se pueden responder con regresión.

10.2 Tipos de clasificación

Tenemos dos tipos de clasificación que son:

- **A posteriori:** en este caso, el resultado lo obtendremos del estudio de los hechos observados con anterioridad. Es un machine learning supervisado donde las clases ya son conocidas.
- **A priori:** en este caso estudiamos los hechos, pero existe una interpretación de los datos. Aquí no tenemos las clases identificadas previamente y se hace uso del clustering.

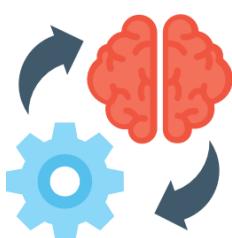
En nuestro caso, estudiaremos la clasificación **a posteriori**.

10.3 Entrada y salida

La entrada y salida de los datos es un tema que debemos tener bien claro. En nuestros datos tenemos dos tipos de atributos, que son los de **entrada** y los de **salida**. Los atributos de **salida** serán aquellos que son dependientes, mientras que los otros vendrán a ser los valores independientes. Veamos un ejemplo para que este concepto esté claro. En nuestro conjunto de datos de la flor de Iris tenemos los cuatro valores siguientes, ancho de pétalo, largo de pétalo, ancho de sépalo y largo de sépalo. Estos atributos serán los valores independientes o valores de **entrada**, mientras que la clase será nuestro valor de **salida**. Como funciona esto, recordemos, es que dependiendo de las características del pétalo y sépalo, la flor pertenecerá a una clase u otra.

	Entrada				Salida
No.	1: sepal length Numeric	2: sepal width Numeric	3: petal length Numeric	4: petal width Numeric	5: class Nominal
1	5.1	3.5	1.4	0.2	Iris-setosa
2	4.9	3.0	1.4	0.2	Iris-setosa
3	4.7	3.2	1.3	0.2	Iris-setosa
4	4.6	3.1	1.5	0.2	Iris-setosa
5	5.0	3.6	1.4	0.2	Iris-setosa
6	5.4	3.9	1.7	0.4	Iris-setosa
7	4.6	3.4	1.4	0.3	Iris-setosa
8	5.0	3.4	1.5	0.2	Iris-setosa
9	4.4	2.9	1.4	0.2	Iris-setosa
10	4.9	3.1	1.5	0.1	Iris-setosa
11	5.4	3.7	1.5	0.2	Iris-setosa
12	4.8	3.4	1.6	0.2	Iris-setosa
13	4.8	3.0	1.4	0.1	Iris-setosa
14	4.3	3.0	1.1	0.1	Iris-setosa
15	5.8	4.0	1.2	0.2	Iris-setosa
16	5.7	4.4	1.5	0.4	Iris-setosa
17	5.4	3.9	1.3	0.4	Iris-setosa
18	5.1	3.5	1.4	0.3	Iris-setosa
19	5.7	3.8	1.7	0.3	Iris-setosa
20	5.1	3.8	1.5	0.3	Iris-setosa
21	5.4	3.4	1.7	0.2	Iris-setosa

Entrada y salida
Fuente: Elaboración propia



RECUERDA

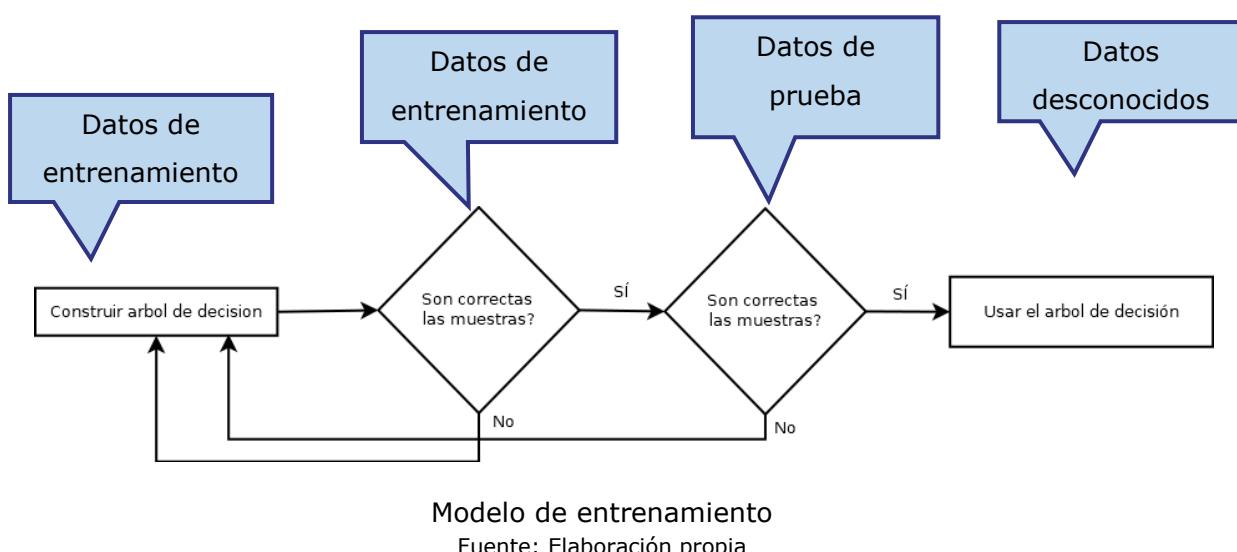
Durante la clasificación, es importante tener una base de datos lo suficientemente grande para entrenar el modelo adecuadamente.

10.4 Trabajando con clasificación

La clasificación es un proceso que se lleva a cabo en dos pasos, que son, en primer lugar, el entrenamiento del modelo y en segundo, la prueba del modelo para su validar su precisión.

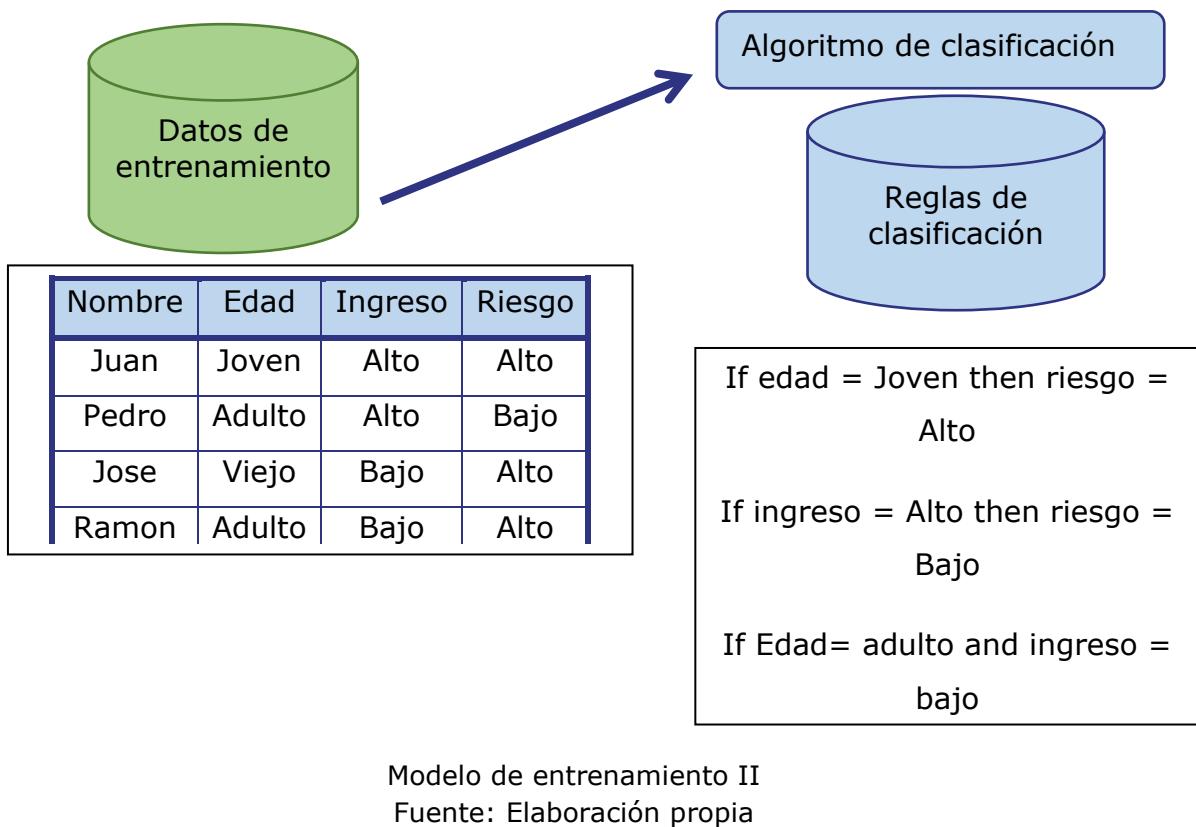
El primer paso crea un clasificador basado en el análisis de tuplas de la base de datos y sus etiquetas. Mediante el análisis de los **datos de entrenamiento** el sistema crea reglas para la predicción. En este paso las reglas se usan para obtener predicciones sobre los atributos de salida y también se calcula la exactitud de este clasificador.

En el segundo paso se prueba con datos aleatorios de la base de datos completa, que son independientes de los **datos de entrenamiento** de la fase anterior. Esto significa que el modelo no ha sido expuesto a estos **datos de prueba** durante la fase de entrenamiento. La exactitud del clasificador se calculará en base al porcentaje de clasificaciones correctas en los **datos de prueba**.

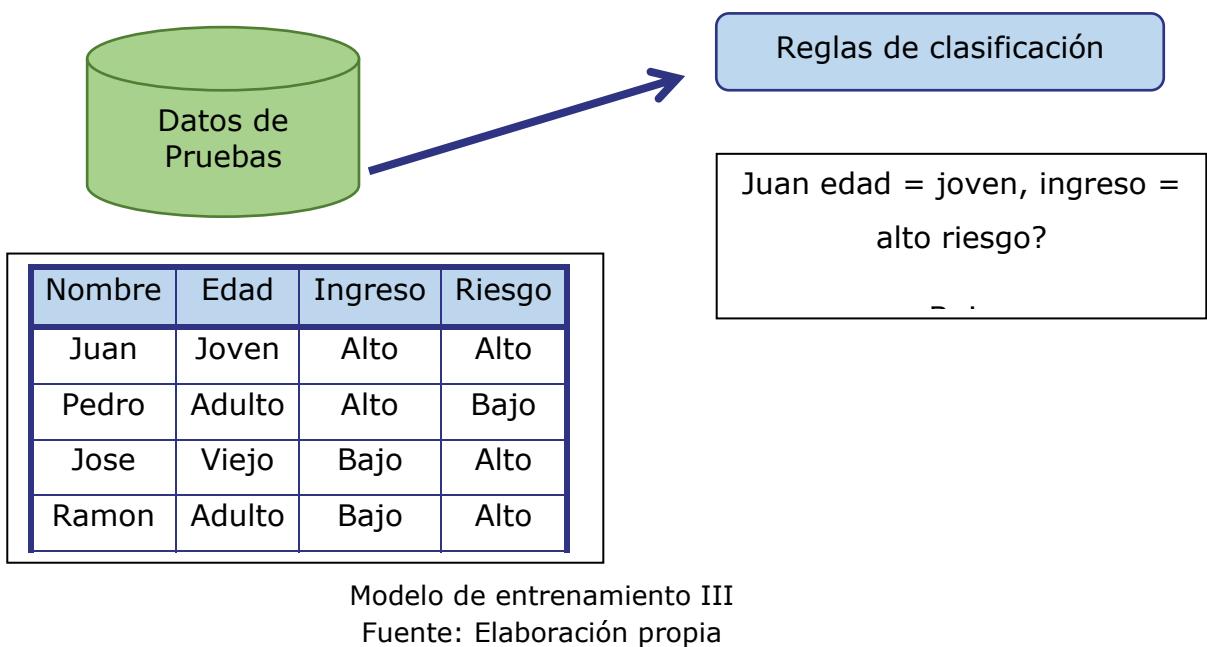


Este mismo proceso se puede aplicar, por ejemplo, para determinar el riesgo de hacer un préstamo. Veámoslo gráficamente:

1. Construyendo el clasificador



2. Prediciendo el riesgo usando los datos de prueba.



Entonces, como vemos, se trata de un proceso de entrenamiento y prueba el que nos permitirá hacer una predicción correcta.



EJEMPLO PRÁCTICO

Por el deterioro de la economía en general, una empresa se ve obligada por los acontecimientos a tomar decisiones bruscas, con la esperanza de que ello sirva para garantizar la supervivencia de la compañía. Entre ello se incluye realizar despidos.

A diferencia de otros despidos realizados en el pasado, esta empresa considera que en este momento la tecnología le puede ayudar con esta desagradable tarea.

SOLUCIÓN

La circunstancia donde probablemente más pueda ayudar el análisis de datos será a la hora de determinar qué trabajadores deben ser despedidos. Obviamente, todo el mundo desearía que la realidad fuera otra y que resultase innecesario acudir a la última opción que siempre es el despido.

Pero la realidad hay que enfrentarla sin autoengaños y toda situación como la descrita implica en la práctica totalidad de ocasiones un número mayor o menor de despidos.

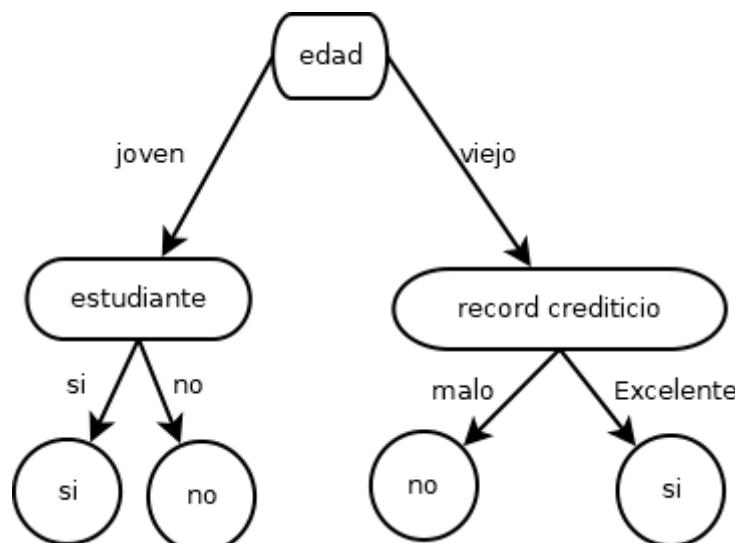
Por el bien de la empresa, por su propia subsistencia y, por tanto, por el bien de los trabajadores que permanecerán en la compañía, los candidatos al despido deben ser correctamente seleccionados. En demasiadas ocasiones, el proceso de toma de decisiones a la hora de afrontar un conjunto de despidos grande o, incluso masivo, se realiza de forma muy deficiente, lo que provoca que al tiempo que son despedidos trabajadores con un bajo impacto sobre el funcionamiento global de la compañía, también se despiden a otros empleados que aportaban un valor añadido superior.

Por ejemplo, se podría aplicar un algoritmo de clasificación para clasificar a los empleados más jóvenes de la empresa. De modo que se pudiesen utilizar los datos asociados al rendimiento de los trabajadores con mayor trayectoria cronológica en la compañía y una vez entrenado ese modelo de clasificación, aplicarlos a los datos disponibles de los trabajadores con menos tiempo de permanencia en la empresa.

De este modo, aquellos jóvenes profesionales con un potencial de crecimiento, desarrollo y rendimiento superior serían retenidos en la empresa y se despediría al resto.

10.5 Árbol de decisión

El árbol de decisión es una estructura donde las decisiones se hacen utilizando condicionales. Un árbol como estructura está compuesto por un nodo raíz, ramas y nodos hoja.



Árbol de entrenamiento

Fuente: Elaboración propia

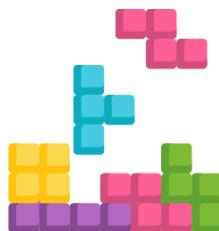
Un árbol de decisión puede ser fácilmente programado en reglas de clasificación con estructuras de tipo If. **El proceso que produce este árbol es conocido como inducción.**

Los árboles de decisiones son una herramienta de uso común en machine learning y han sido implementados por distintas tecnologías tales como WEKA, R, Python, Java, etc.



SABÍAS QUE...

J. Ross Quinlan, un investigador de ML, desarrolló el algoritmo de árbol de decisión conocido como ID3 entre 1970 y 1980 y posteriormente propuso otro como sucesor, que es el denominado C4.3.



EJEMPLO PRÁCTICO

Una empresa que deba realizar despidos, puede optar por crear un clasificador automático que le facilite la labor de terminar quien debe ser despedido y quien no.

Pero puede darse la circunstancia de que la compañía no disponga de datos suficientes para construir ese clasificador.

Por este motivo puede optar por crear un árbol de decisión que genere un resultado análogo.

SOLUCIÓN

En caso de no disponer de un volumen suficiente de datos como para entrenar un proceso de clasificación automática, se podría optar, alternativamente, por construir un modelo de árbol de decisión manejando los datos que se conozcan influyen en el rendimiento a lo largo de la carrera profesional de esa empresa y sector.

También se podrían aplicar herramientas de business intelligence para mostrar objetividad, es decir, poder presentar y demostrar ante terceras partes, como los sindicatos, informes que hagan patente que determinas áreas, departamentos, grupos de personas, etc. Presentan un rendimiento o valor inferior y que, por tanto, no sería justo una política de despidos indiscriminada o basada en factores que no tienen que ver con la calidad del profesional en concreto.

10.6 Clasificador bayesiano ingenuo

Está basado en el teorema de Bayes, que fue propuesto por Thomas Bayes en el siglo XVIII. Se fundamenta en la hipótesis de que los datos pertenecen a una clase específica. Para poder entender este teorema, debemos saber que utiliza el cálculo de probabilidades como herramienta matemática fundamental.

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)}$$

En términos generales, el teorema de Bayes vincula la probabilidad de A dado B y de B dado A. Por ejemplo, sabiendo la probabilidad de tener un dolor de cabeza, dado que se tenga gripe, contando con algunos datos más, se podría tener la probabilidad de tener gripe, si se tiene dolor de cabeza.

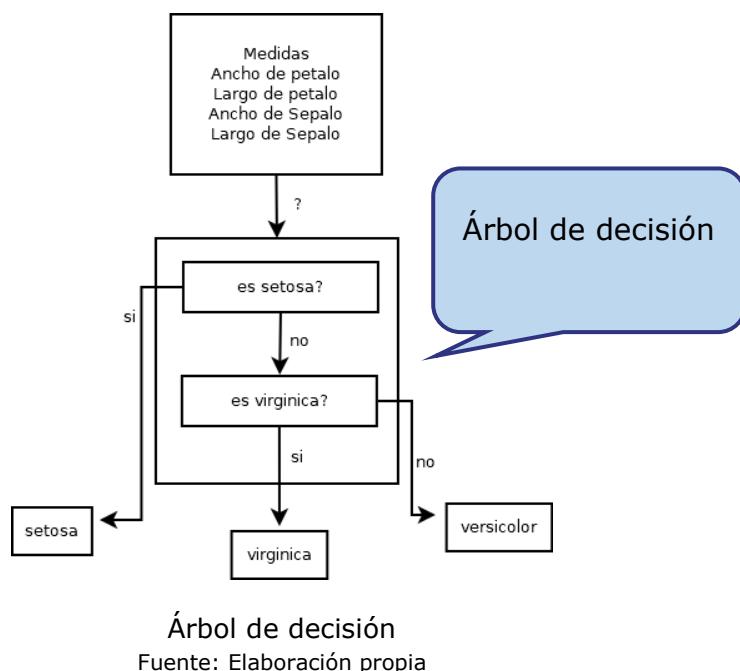
Otro ejemplo podría ser, digamos que, una manzana puede ser considerada como tal si es de color rojo, es redonda y tiene un radio de alrededor de 7 cm. Cada uno de estos atributos contribuye de manera independiente a que la fruta sea una manzana, independiente de si algunas características están presentes y otras no.

Una de las ventajas de usar este método para clasificar es que no requiere gran cantidad de datos para su entrenamiento.

10.7 Implementando la clasificación en WEKA mediante un árbol de decisión

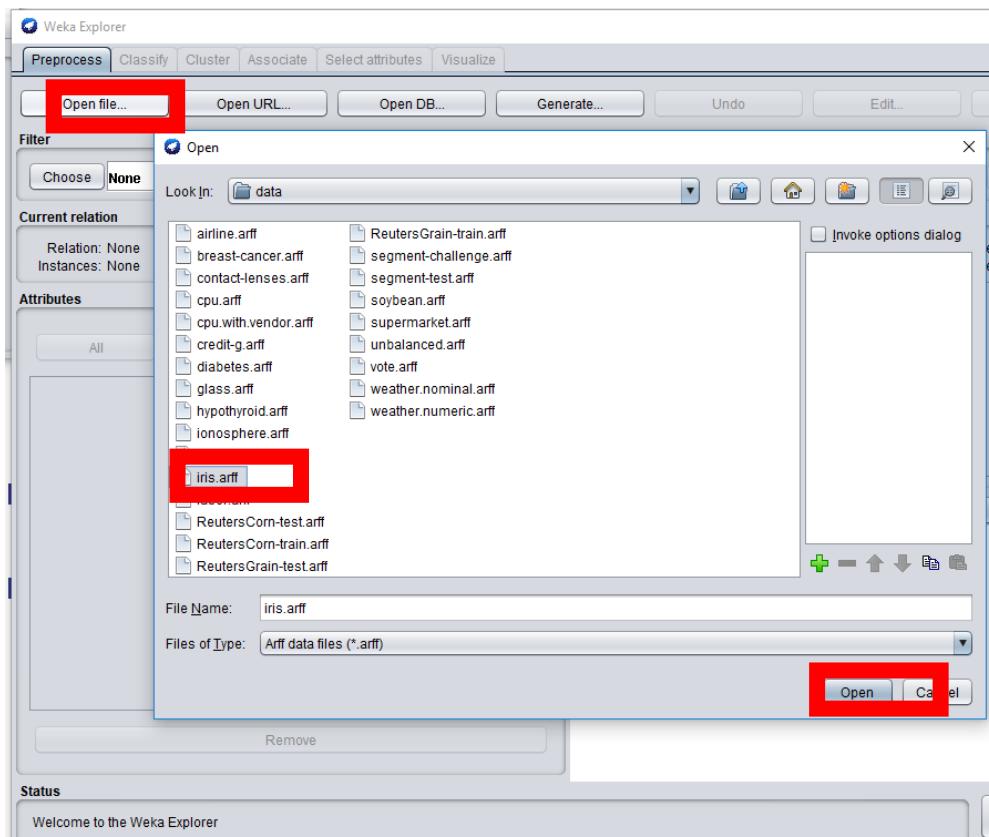
Veamos cómo aplicar la teoría de árbol de decisión en WEKA. Supongamos que queremos clasificar los datos de unas muestras desconocidas de la flor de Iris. Para ello crearemos un árbol de decisión, haciendo uso de las herramientas que nos proporciona WEKA y de los datos de Iris.arff.

Weka será de gran ayuda para clasificar los valores desconocidos con el enfoque que vemos en el siguiente esquema.



Los pasos para hacer esta implementación con WEKA son los que mostramos a continuación.

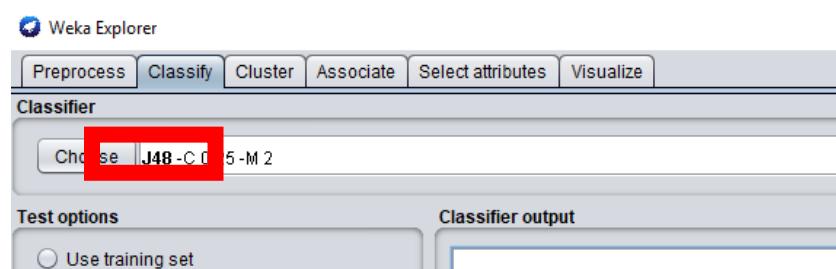
1. Abrimos WEKA, vamos al explorador y cargamos los datos de la flor de iris en la carpeta data, dentro del directorio de instalación de WEKA.



Explorer de weka. Cargar datos

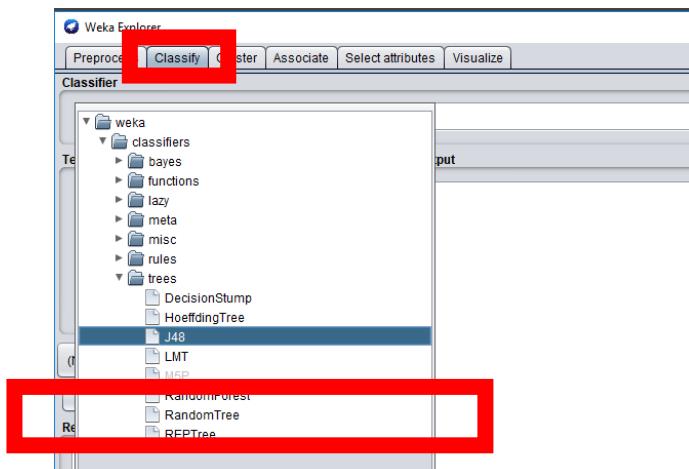
Fuente: Elaboración propia

2. Ahora, vamos a la pestaña de clasificación y seleccionamos el clasificador J48.



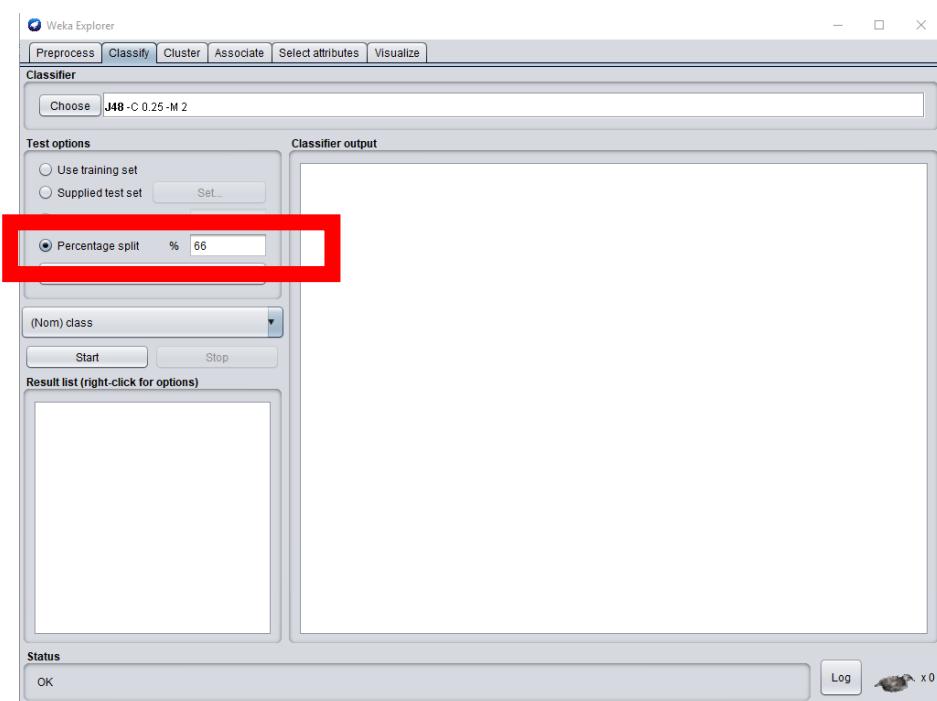
Clasificador J48 de weka

Fuente: Elaboración propia



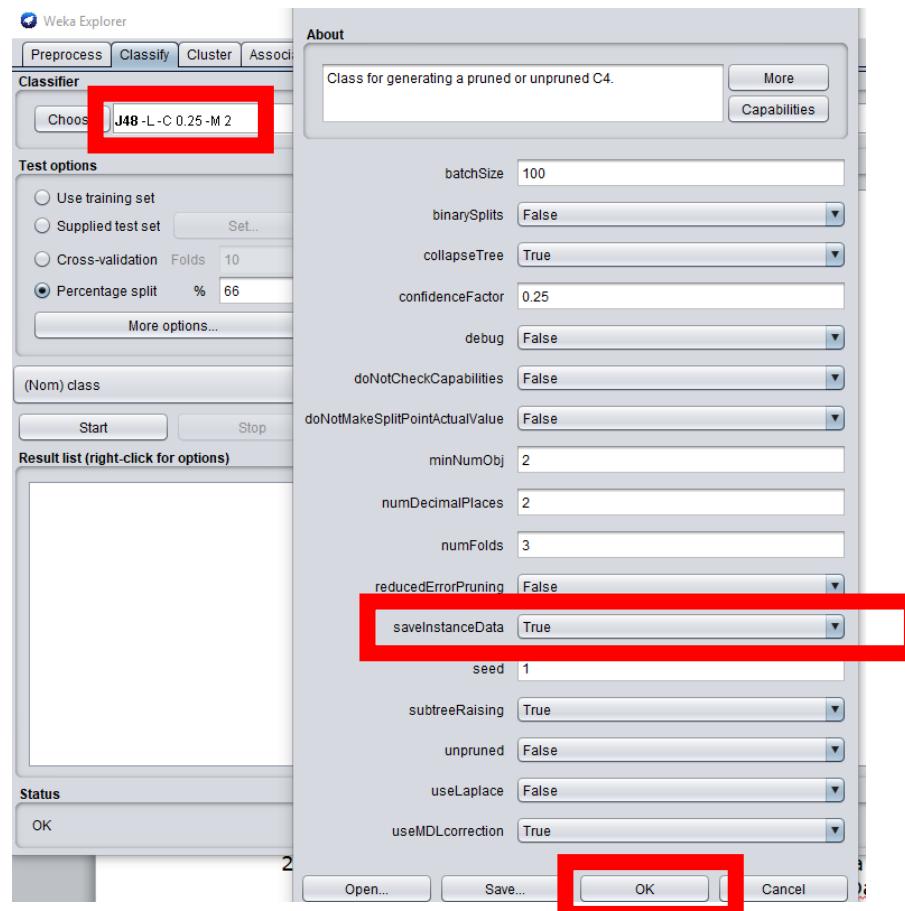
Clasificador de weka
Fuente: Elaboración propia

- Después de realizar esta selección revisaremos el **percentage split** en la sección de **test options**. Si aparece al 66%, lo dejamos así. Esto significa que 2/3 de los registros serán usados con el propósito de entrenar nuestro modelo y la otra parte será usada para hacer las pruebas.



Revisar el percentage split
Fuente: Elaboración propia

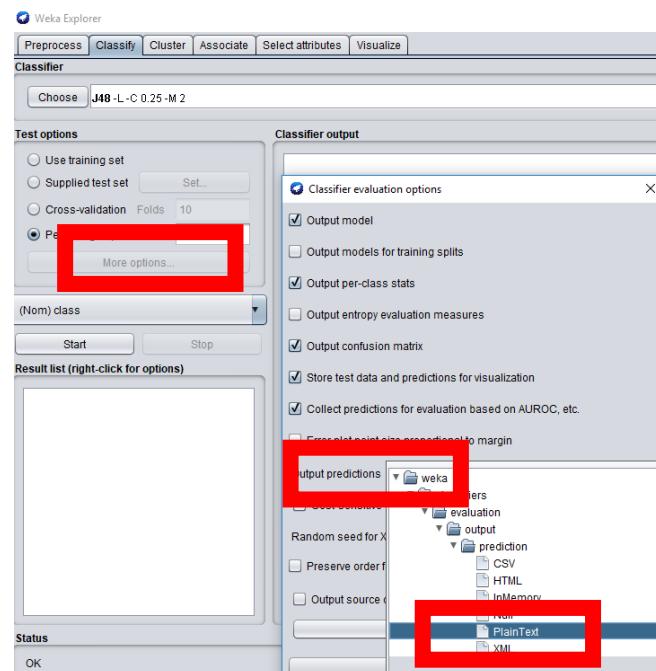
- Abrimos las opciones del algoritmo haciendo clic sobre su nombre y activamos la opción **saveInstance**, cambiando su valor a **true**.



Activar la opción saveInstance

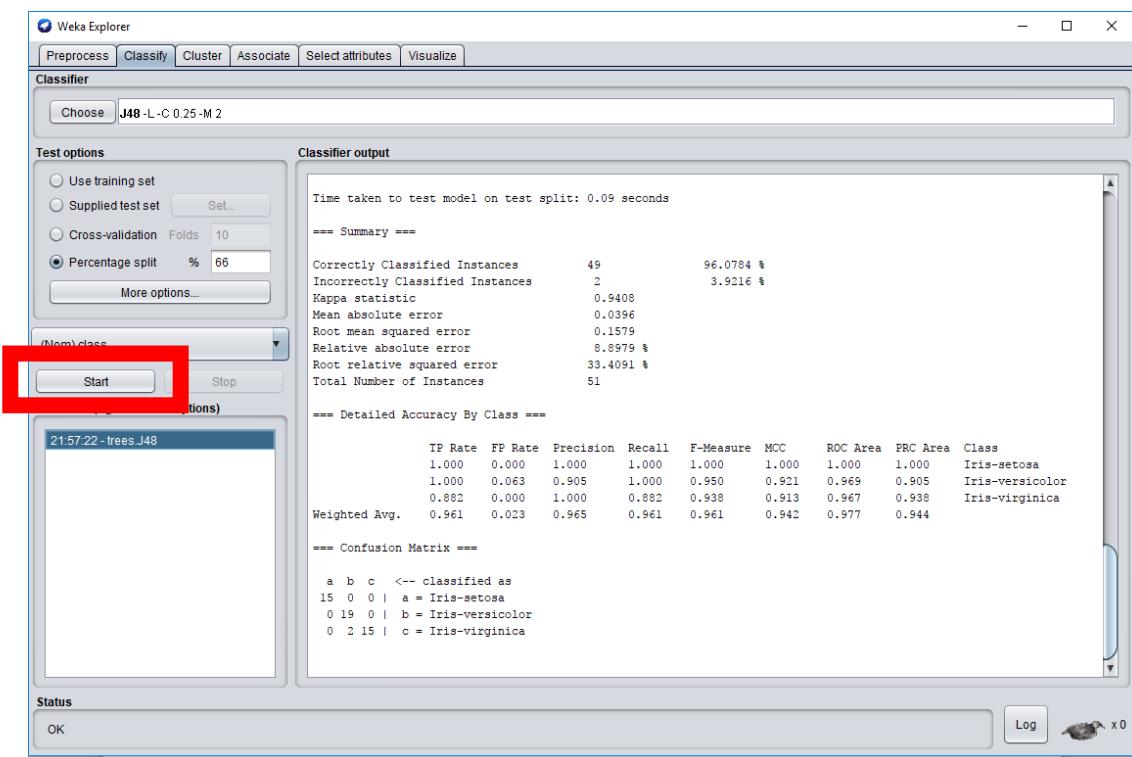
Fuente: Elaboración propia

5. Ahora también seleccionamos el formato de salida a texto plano en las opciones.



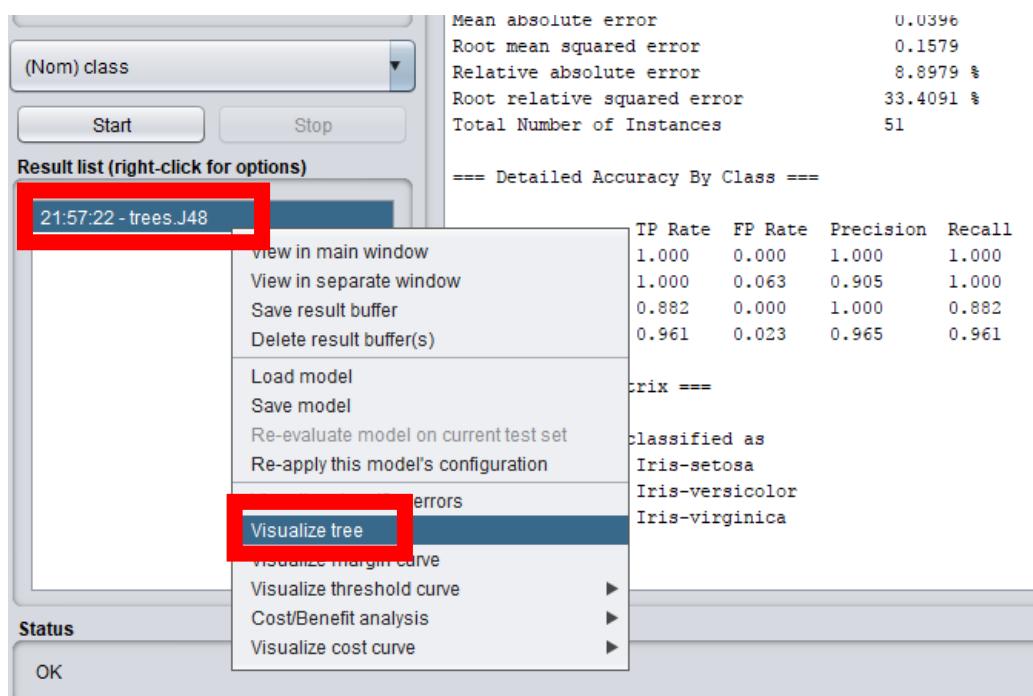
Formato de salida
Fuente: Elaboración propia

6. Ya estamos listos para lanzar nuestro algoritmo. Para ello hacemos clic en **start**.



Lanzar algoritmo. Botón start
Fuente: Elaboración propia

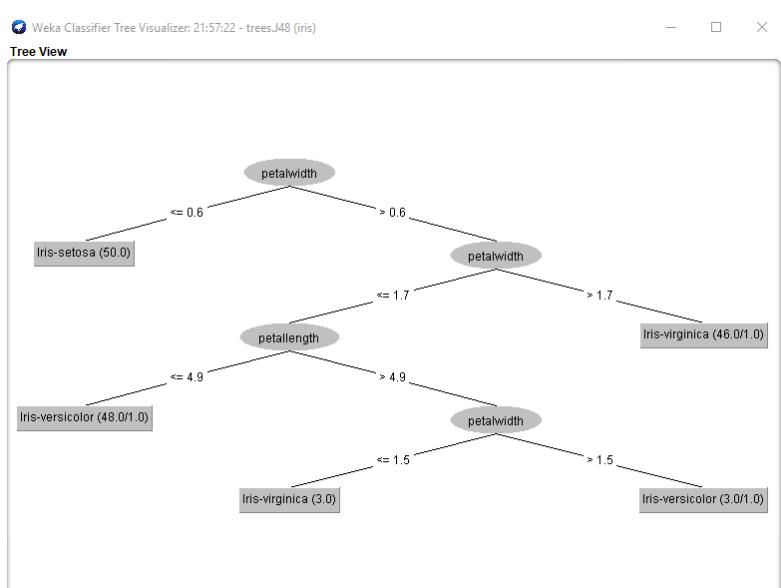
7. A la derecha tenemos el resultado del algoritmo. Para ver nuestro árbol, hacemos clic con el botón derecho sobre el resultado.



Resultado del algoritmo. Visualización

Fuente: Elaboración propia

8. Árbol de decisión generado por WEKA.



Árbol de decisión
 Fuente: Elaboración propia

Ahora, revisemos dos cuestiones importantes para entender los resultados:

La matriz de confusión:

==== Confusion Matrix ====			
	a	b	c
15	0	0	a = Iris-setosa
0	19	0	b = Iris-versicolor
0	2	5	c = Iris-virginica

Matriz de confusión

Fuente: Elaboración propia

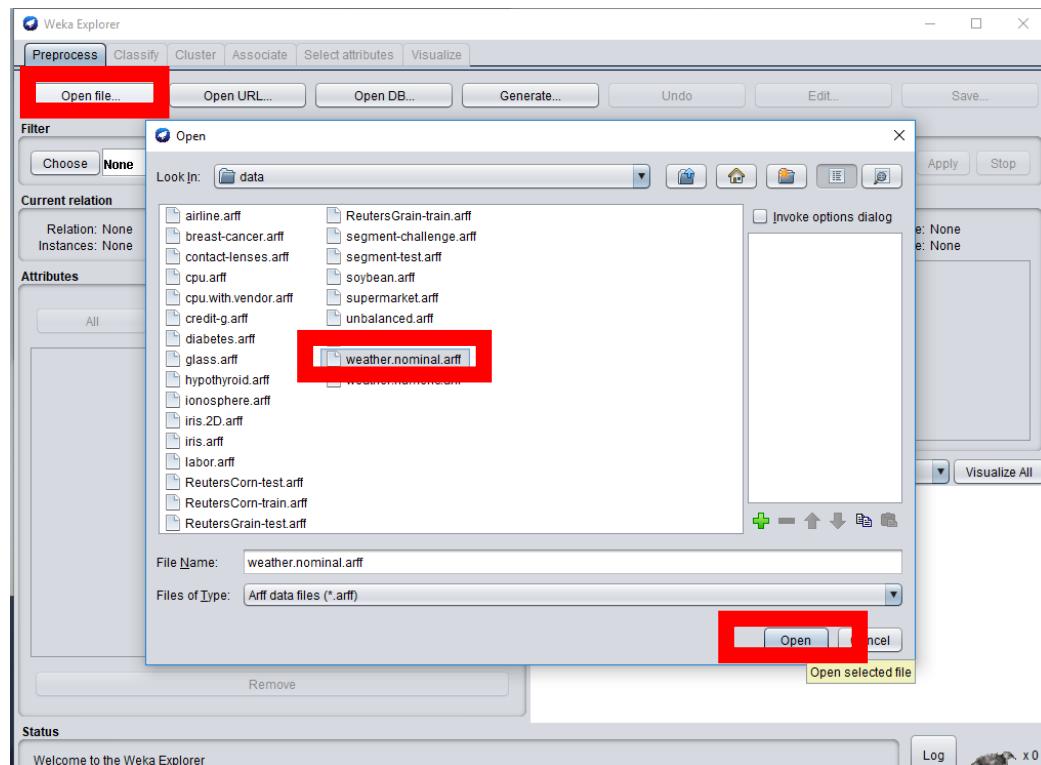
Esta matriz, lo que nos dice, es donde se equivocó nuestro modelo. Es decir, cuáles clasificó correctamente y cuáles no. Podemos comprobar que en este caso clasificó dos erróneamente.

Si revisamos el árbol de decisión, podremos observar que se basa en valores que pueden ser llevados fácilmente a cualquier lenguaje de programación para su uso y posterior realización de predicciones.

10.8 Implementando clasificador bayesiano ingenuo en WEKA

Los pasos son:

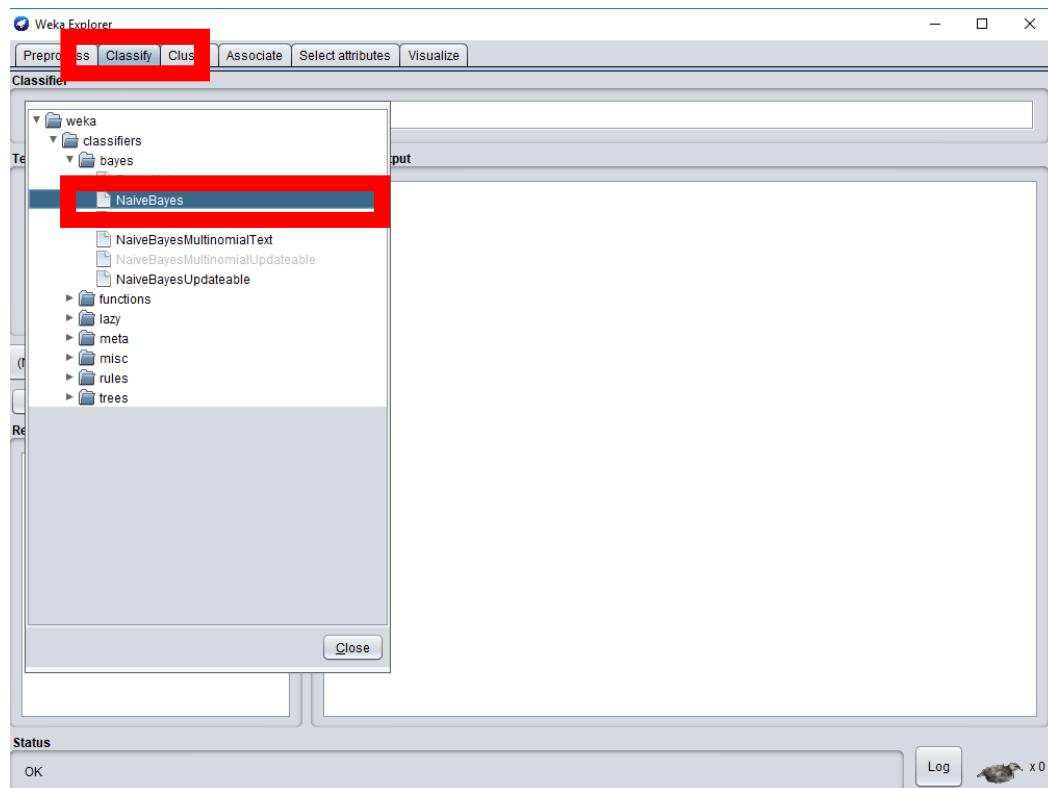
1. Primero debemos abrir WEKA e ir a la ventana del explorador y abrir los datos de prueba weather.nominal.arff.



Abrir datos de prueba

Fuente: Elaboración propia

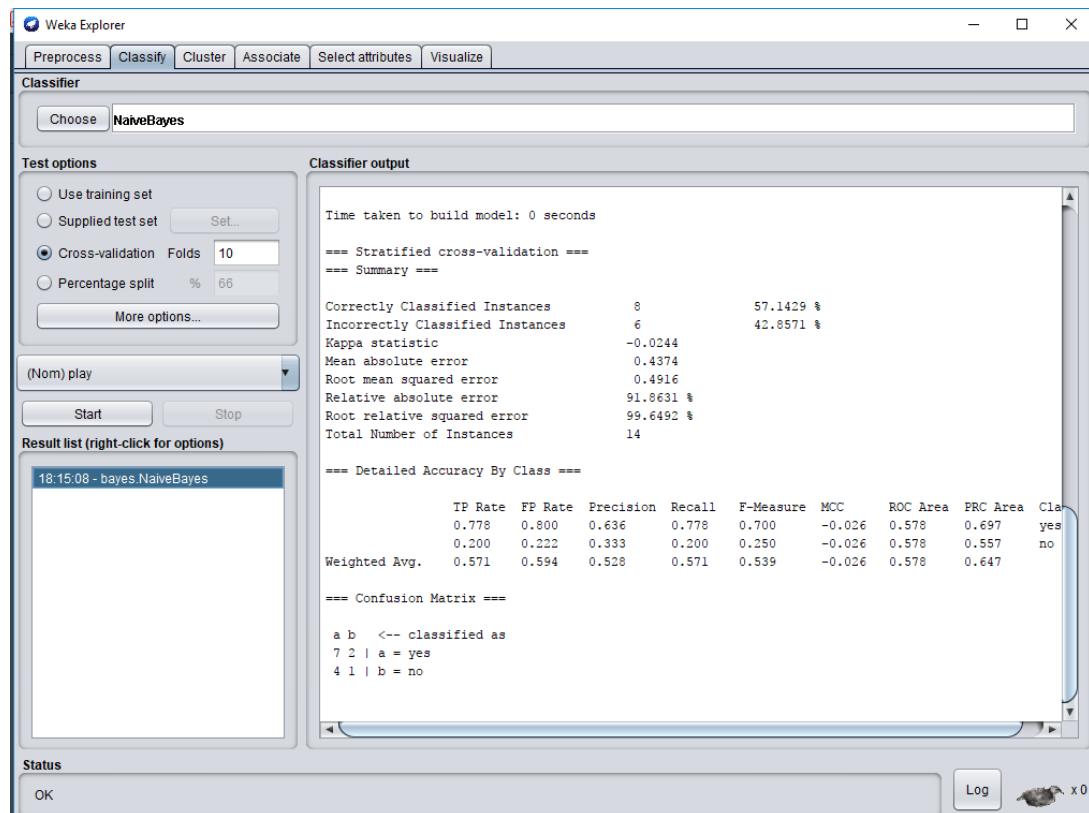
2. Vamos a la ventana de classify y seleccionamos NaiveBayes.



Selección NaiveBayes

Fuente: Elaboración propia

3. Ahora podemos ver en la ventana de resultados la matriz de confusión, al igual que ocurría con el árbol de decisión.



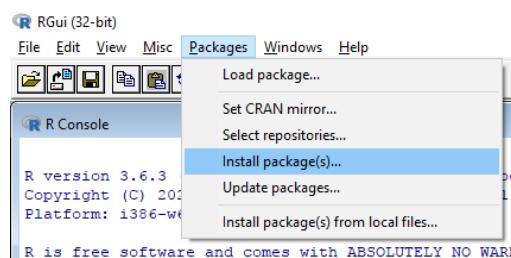
Resultados matriz confusión

Fuente: Elaboración propia

10.9 Implementando clasificación en R mediante árboles de decisión

En R existen varias librerías para hacer uso de los árboles. Nosotros usaremos una llamada 'party' para operar el árbol y 'ctree' para construirlo.

1. Instalamos los paquetes mencionados **party** y **ctree**.



Install package

Fuente: Elaboración propia

2. Despues de instalados los paquetes, debemos crear la variable target, que especifica qué especie es la clase y las otras variables que son independientes.

```
> target = Species ~ Sepal.Length + Sepal.Width + Petal.Length + Petal.Width
```

3. Luego, usamos la función ctree(), donde le pasamos la variable target y el dataset de iris.

```
> cdt <- ctree(target, iris)
```

4. Posteriormente, usaremos la función predict() para comprobar la predicción en los datos de entrenamiento. Con la función table() creamos la matriz de confusión y errores de la clasificación.

```
> table(predict(cdt), iris$Species)
```

5. El resultado:

```
setosa    versicolor   virginica
setosa      50          0          0
versicolor     0         49          5
virginica     0          1         45
> |
```

Resultados árbol

Fuente: Elaboración propia

Ahora ¿cómo podemos analizar estos resultados?

Bien, según la matriz de confusión, podemos ver que el árbol clasificó correctamente 50 setosa, 49 versicolor y 45 virginica. Por otro lado, clasificó 1 versicolor como virginica y 5 virginicas como versicolor.

En esta matriz, los valores horizontales representan la clasificación real y la vertical, es la clasificación realizada por el sistema.

Ahora, comprobemos cómo podemos ver las reglas del árbol de decisión con el comando **cdt**:

```
> cdt

Conditional inference tree with 4 terminal nodes

Response: Species
Inputs: Sepal.Length, Sepal.Width, Petal.Length, Petal.Width
Number of observations: 150

1) Petal.Length <= 1.9; criterion = 1, statistic = 140.264
   2)* weights = 50
1) Petal.Length > 1.9
   3) Petal.Width <= 1.7; criterion = 1, statistic = 67.894
      4) Petal.Length <= 4.8; criterion = 0.999, statistic = 13.865
         5)* weights = 46
      4) Petal.Length > 4.8
         6)* weights = 8
      3) Petal.Width > 1.7
         7)* weights = 46
> |
```

Comando cdt

Fuente: Elaboración propia

Podemos ver de manera visual el árbol con el comando `plot(cdt,type="simple")`:

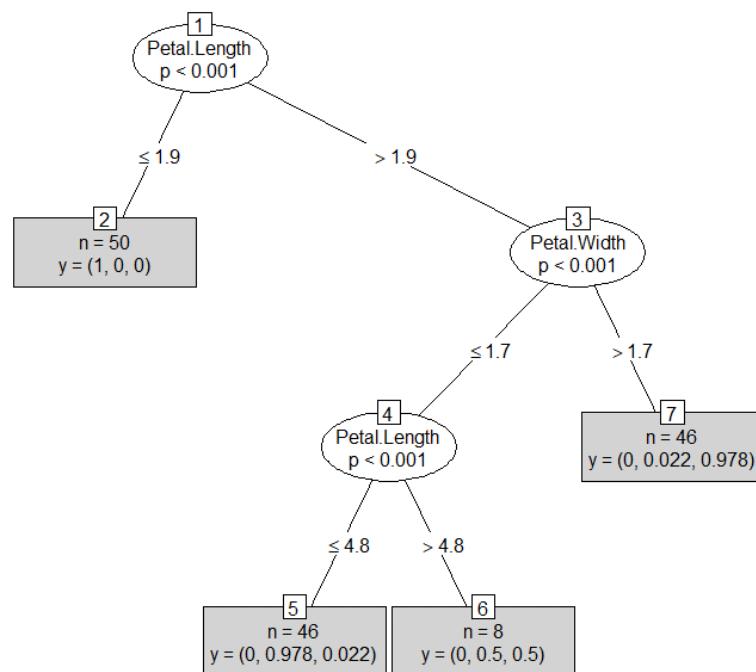
```

setosa      50
versicolor  0
virginica   0
> cdt

Conditional inference tree by class

response: Species
inputs: Sepal.Length, Sepal.Width, Petal.Length, Petal.Width
Number of observations: 150

.) Petal.Length <= 1.9; c(1, 0, 0)
 2)* weights = 50
.) Petal.Length > 1.9
 3) Petal.Width <= 1.7; c(0, 0.022, 0.978)
    4) Petal.Length <= 4.8
      5) weights = 46
      6) Petal.Length > 4.8
        7) weights = 8
        3) Petal.Width > 1.7
          7)* weights = 46
> plot(cdt,type="simple")
  
```



Visualizador de árbol en Weka

Fuente: Elaboración propia

10.10 Clasificador bayesiano en R

Para aplicar Naive Bayes en R usaremos el siguiente paquete: e1071

```
install.packages('e1071')
```

```

> install.packages('e1071')
Installing package into 'C:/Users/sebastian/Documents/R/win-library/3.6'
(as 'lib' is unspecified)
trying URL 'https://cloud.r-project.org/bin/windows/contrib/3.6/e1071_1.7-3.zip'
Content type 'application/zip' length 1022507 bytes (998 KB)
downloaded 998 KB

package 'e1071' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
  C:\Users\sebastian\AppData\Local\Temp\Rtmp8o2CYJ\downloaded_packages
  
```

```
install.packages('e1071')
```

Fuente: Elaboración propia

Ahora, revisamos si los datos de la flor de Iris están cargados mediante el comando: **summary(iris)**

```
> summary(iris)
   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
   Min. :4.300    Min. :2.000    Min. :1.000    Min. :0.100
   1st Qu.:5.100  1st Qu.:2.800  1st Qu.:1.600  1st Qu.:0.300
   Median :5.800  Median :3.000  Median :4.350  Median :1.300
   Mean   :5.843  Mean   :3.057  Mean   :4.358  Mean   :1.399
   3rd Qu.:6.400  3rd Qu.:3.300  3rd Qu.:5.100  3rd Qu.:1.800
   Max.   :7.900  Max.   :4.400  Max.   :6.900  Max.   :2.500
   Species
   setosa      :50
   versicolor  :50
   virginica   :50
```

Revisión carga de datos

Fuente: Elaboración propia

Revisamos también la estructura de los datos con str:

str(iris)

```
> str(iris)
'data.frame': 150 obs. of 5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",...: 1 1 1 1 1 1 1 1 1 1 ...
```

Revisión de estructura de datos

Fuente: Elaboración propia

Creamos una tabla para las clases con **table(iris\$Species)**

```
> table(iris$Species)

setosa versicolor virginica
 50       50       50
```

Tabla para clases

Fuente: Elaboración propia

Ahora, creamos unas muestras aleatorias con el comando **sample**:

muestras_iris = sample(150,110,replace=FALSE)

```
> muestras_iris=sample(150,110,replace = FALSE)
```

Creamos los datos de entrenamiento y para las pruebas:

iris_entrenamiento=iris[muestras_iris,]

```
iris_pruebas=iris[-muestras_iris,]
```

También debemos crear las etiquetas para las tablas:

```
iris_etiquetas_entrenamiento=iris[muestras_iris,]$Species
```

```
iris_etiquetas_pruebas=iris[-muestras_iris,]$Species
```

Luego, creamos las dos tablas:

```
table(iris_entrenamiento$Species)
```

```
table(iris_pruebas$Species)
```

```
> iris_entrenamiento=iris[muestras_iris,]
> iris_pruebas=iris[-muestras_iris,]
> iris_etiquetas_entrenamiento=iris[muestras_iris,]$Species
> iris_etiquetas_pruebas=iris[-muestras_iris,]$Species
> table(iris_entrenamiento$Species)

  setosa versicolor  virginica
  36       38       36
> table(iris_pruebas$Species)

  setosa versicolor  virginica
  14       12       14
```

Creación de tablas

Fuente: Elaboración propia

Ahora, debemos crear nuestro modelo clasificador. Para ello usaremos el método de **naiveBayes**:

```
clasificador_iris=naiveBayes(iris_entrenamiento,iris_etiquetas_entrenamiento)
```

Como vemos, le pasamos los datos de entrenamiento y las etiquetas.

Posteriormente, evaluamos la exactitud del modelo, haciendo uso de los datos de prueba:

```
iris_predicciones=predict(clasificador_iris,iris_pruebas)
```

```
> clasificador_iris=naiveBayes(iris_entrenamiento,iris_etiquetas_entrenamiento)
> iris_predicciones=predict(clasificador_iris,iris_pruebas)
> iris_predicciones
[1] setosa      setosa      setosa      setosa      setosa      setosa
[7] setosa      setosa      setosa      setosa      setosa      setosa
[13] setosa      setosa      versicolor  versicolor  versicolor  versicolor
[19] versicolor versicolor versicolor  versicolor  versicolor  versicolor
[25] versicolor versicolor virginica virginica virginica virginica
[31] virginica  virginica  virginica  virginica  virginica  virginica
[37] virginica  virginica  virginica  virginica  virginica  virginica
Levels: setosa versicolor virginica
```

Evaluador de exactitud

Fuente: Elaboración propia

Ahora, dibujamos una tabla instalando la librería gmodels:
install.packages('gmodels')

Y la cargamos:

```
library(gmodels)
```

Ahora, dibujamos nuestra tabla

```
CrossTable(iris_predicciones,iris_etiquetas_pruebas,prop.chisq = FALSE,
           prop.t= FALSE, prop.r = FALSE, dnn = c('predicted','actual'))
```

```
> CrossTable(iris_predicciones,iris_etiquetas_pruebas,prop.chisq = FALSE, prop.t= FALSE, prop.r = FALSE, dnn = c('predicted','actual'))

Cell Contents
-----|
|           N |
|   N / Col Total |
-----|
```

Total Observations in Table: 40

		actual			Row Total
predicted	setosa	versicolor	virginica		
setosa	14	0	0	14	
	1.000	0.000	0.000		
versicolor	0	12	0	12	
	0.000	1.000	0.000		
virginica	0	0	14	14	
	0.000	0.000	1.000		
Column Total	14	12	14	40	
	0.350	0.300	0.350		

CrossTable

Fuente: Elaboración propia

Con esto, hemos aplicado Naive Bayes en lenguaje R.

11. REGRESIÓN CON R Y WEKA

Al director financiero de la empresa le preocupa mucho los rápidos cambios en los precios que os ofrecen vuestros proveedores. Para él, este es un auténtico dolor de cabeza.

Por ejemplo, un fabricante chino, el año pasado os vendió cada unidad de un modelo de teléfono móvil a 80 €, pero de repente, cuando le vais a hacer una nueva compra, os dice que el precio ha subido a 92 €. Esa subida de precio hace que os quedéis sin margen de beneficio y, lo peor, ya no estáis a tiempo de encontrar un proveedor alternativo más económico.

Para reducir esta incertidumbre habéis implementado un modelo de regresión lineal que os permitan anticiparos y conocer el precio estimado que querrá aplicaros cada proveedor.

De este modo, si se anticipa que un proveedor, por distintos factores, como el cambio de divisa, la fecha del año, etc. Intentará negociar un precio demasiado alto para vuestros intereses, ahora estáis en disposición de buscar otro proveedor para poder negociar con él y utilizar esa capacidad de presión para conseguir una rebaja en las expectativas del fabricante chino.

Para muchas personas, la regresión lineal es un tema confuso. En este apartado lo aclararemos. Tanto la regresión lineal simple como la regresión lineal múltiple pueden ser aplicadas al estudio de los datos, pero para ello debemos comprender el uso de las variables.

Una variable se podría describir como un contenedor de información. En este caso veremos cómo hacer uso de las variables para aplicar la regresión lineal y como estas varían en el momento de implementar estos algoritmos.

A través de la regresión podremos llegar a usos complejos tales como la predicción y machine learning.

Cuando estudiamos los datos, debemos preguntarnos cuál es el significado de estos. Es decir, tenemos unos datos y estos son la respuesta a cierta pregunta. La cuestión está en determinar cuál es la pregunta que responden esos datos. Para ello, contamos con la herramienta de la regresión que define la pregunta. En esta ocasión, la pregunta es una ecuación que debemos resolver.

Después de tener nuestra ecuación, la podríamos alimentar con distintas fuentes de datos para lograr hacer predicciones. De todo esto es de lo que vamos a hablar en los siguientes apartados.

11.1 Regresión Lineal

La regresión lineal nos provee de una ecuación para una simple línea.

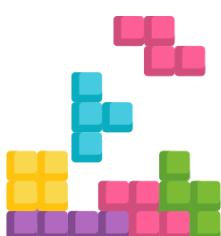
Escribiremos una ecuación que dibujará una línea recta entre un conjunto de puntos que representan datos. Conociendo la ecuación que mejor se adapta a nuestro conjunto de datos, podremos predecir otros puntos alrededor de esa línea. De tal modo que entonces podríamos decir que la regresión lineal comienza no con la respuesta, sino con la búsqueda de la pregunta en forma de ecuación.



SABÍAS QUE...

Carl Friedrich Gauss, en el año 1795, descubrió el método de los mínimos cuadrados, que significó un gran avance en los métodos de navegación marítima, basados en los movimientos de los planetas y estrellas.

Una de las cosas más importantes que debemos concluir sobre la regresión, es que ella -por si misma- representa la respuesta a un problema. En cada caso debemos comenzar con un conjunto de datos y luego responderemos una pregunta, usando una ecuación que expresa la correlación entre los datos. Entendiendo esta correlación, podremos hacer predicciones y como esos datos se comportarán en el futuro.



EJEMPLO PRÁCTICO

Veamos con un ejemplo práctico como implementamos la regresión lineal en Weka.

Este será el conjunto de datos que tenemos:

- **X:** serán las horas de estudio y supondremos que su valor es de (0-8)
- **Y:** son los puntos obtenidos en el examen tipo test (van de 0 a 100)

x	y
0	20
2	50
5	80
8	90
8	85
3	40
6	60
5	50
7	75

1	20
2	20

Teniendo estos datos, debemos determinar la relación entre las variables **X** e **Y**.

El primer paso será cargar nuestros datos en un archivo csv, para lo cual se puede usar **Excel** o cualquier otro programa de este tipo.

The screenshot shows a Microsoft Excel spreadsheet titled "Book1 - Microsoft Excel". The data is organized into two columns: "Horas" (Hours) and "Puntos" (Points). The data points are:

	Horas	Puntos
2	0	20
3	2	50
4	5	80
5	8	90
6	8	85
7	3	40
8	6	60
9	5	50
10	7	75
11	1	20
12	2	20

Después guardamos este archivo como csv.

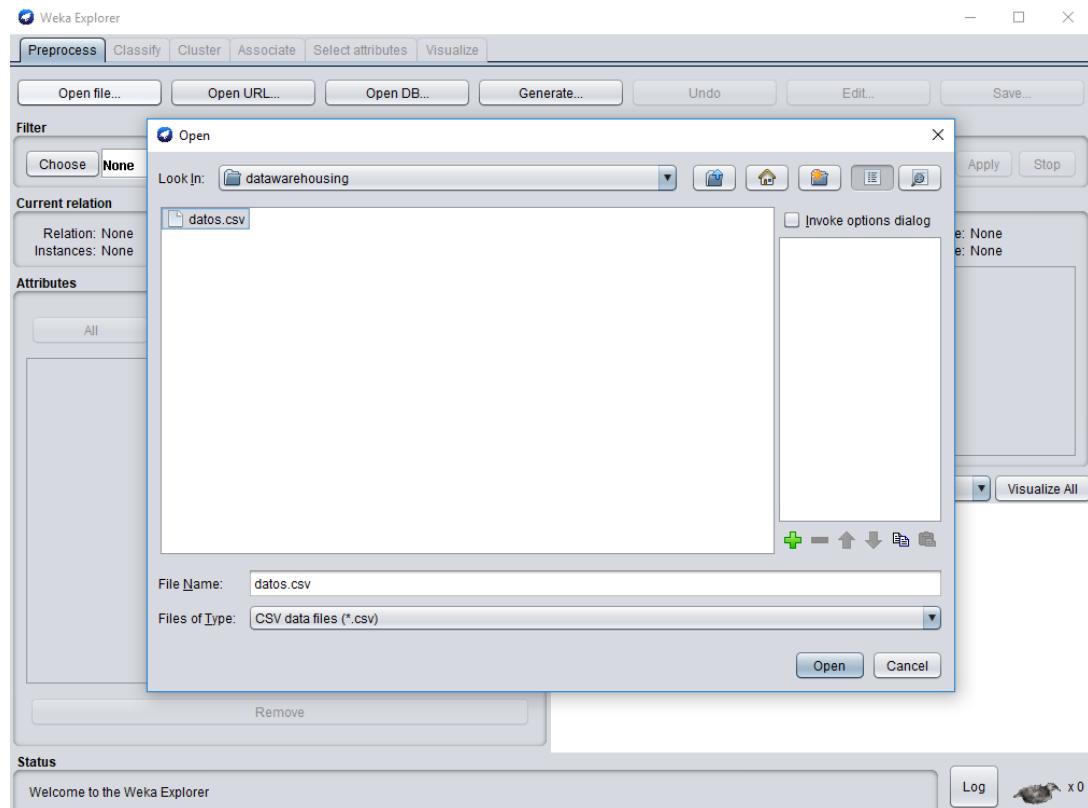
The screenshot shows the "Save As" dialog box in Microsoft Excel. The "File name:" field contains "datos" and the "Save as type:" dropdown menu is open, showing various file formats. The "Excel Workbook (*.xlsx)" option is selected.

File name: datos

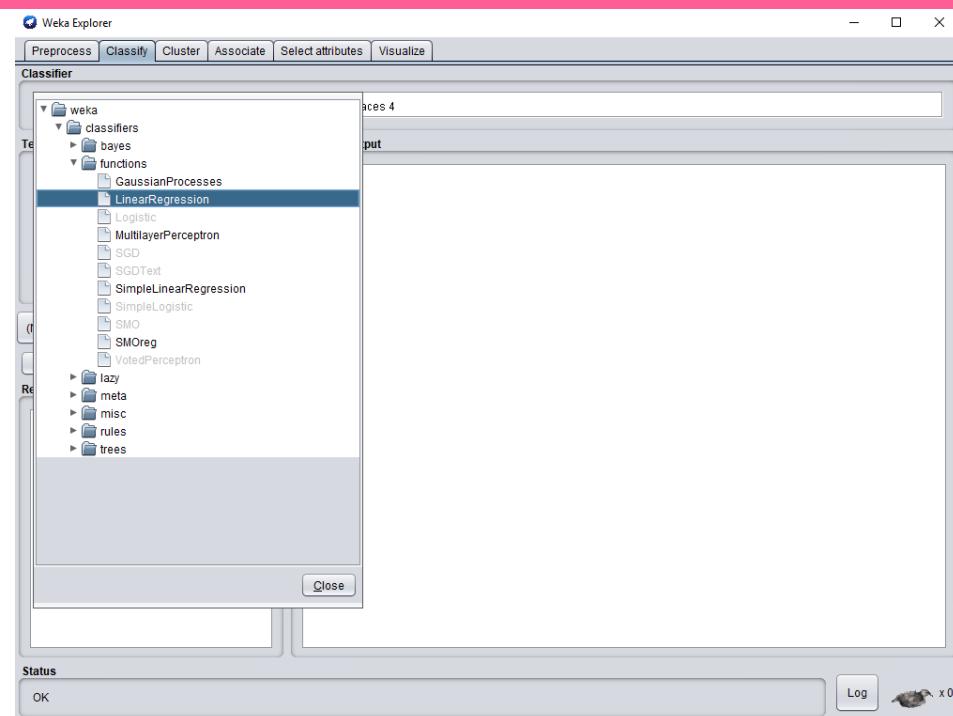
Save as type: Excel Workbook (*.xlsx)

- Excel Workbook (*.xlsx)
- Excel Macro-Enabled Workbook (*.xlsm)
- Excel Binary Workbook (*.xlsb)
- Excel 97-2003 Workbook (*.xls)
- XML Data (*.xml)
- Single File Web Page (*.mht; *.mhtml)
- Web Page (*.htm; *.html)
- Excel Template (*.xltx)
- Excel Macro-Enabled Template (*.xltm)
- Excel 97-2003 Template (*.xlt)
- Text (Tab delimited) (*.txt)
- Unicode Text (*.txt)
- XML Spreadsheet 2003 (*.xml)
- Microsoft Excel 5.0/95 Workbook (*.xls)
- CSV (Comma delimited) (*.csv)
- Formatted Text (Space delimited) (*.prn)

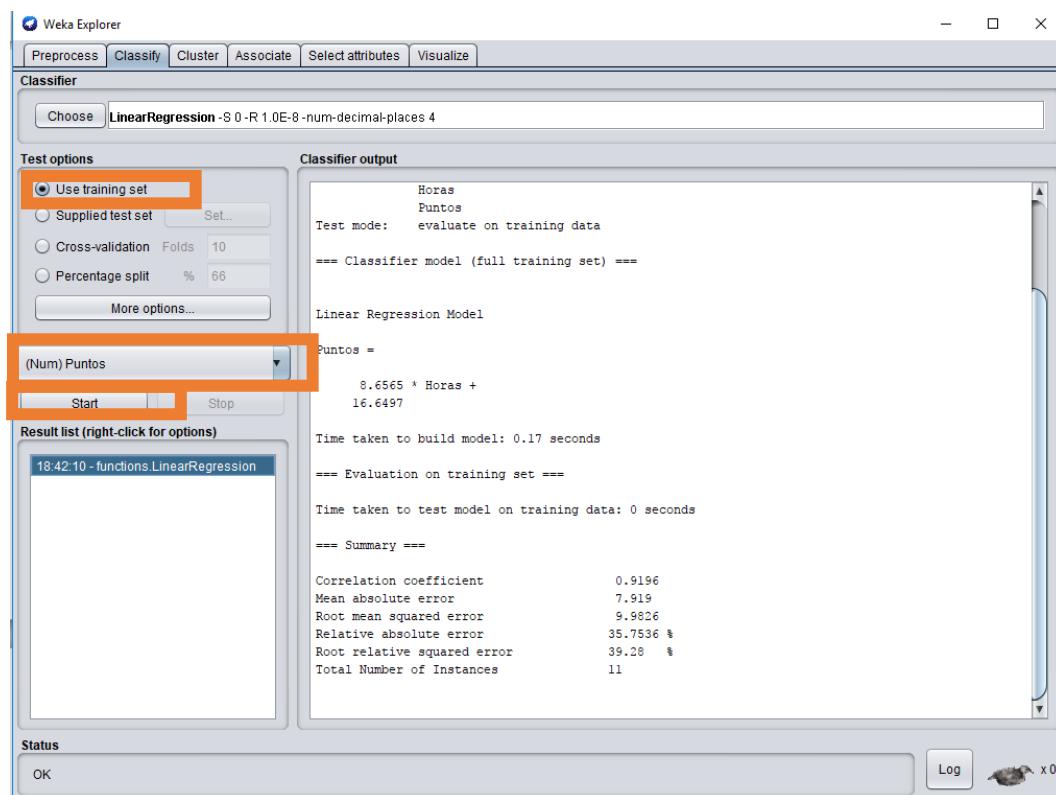
Ahora abriremos este archivo en WEKA, desde la ventana de Explorer y le damos clic a open file, seleccionando el tipo de archivo csv en la ruta donde hemos guardado nuestro archivo.



Ahora tenemos cargados nuestros datos en WEKA y debemos ir a la pestaña de classify y elegimos.



Ahora seleccionamos en las opciones "Use Training set" y elegimos la variable puntos que ya viene por defecto seleccionada y pulsamos el botón start.



Como podemos observar, en los resultados se nos muestra una fórmula:

```
Puntos =  
  
    8.6565 * Horas +  
    16.6497
```

Esta fórmula es nuestro modelo matemático, el cuál dibuja una línea cercana a todos los puntos, con la que podremos hacer predicciones.

Los otros valores son :

```
==== Summary ====  
  
Correlation coefficient          0.9196  
Mean absolute error            7.919  
Root mean squared error        9.9826  
Relative absolute error        35.7536 %  
Root relative squared error   39.28 %  
Total Number of Instances      11
```

El coeficiente de correlación, el cuál nos indica el nivel de relación que tienen ambas variables, es 0.9196, por lo que se puede concluir que es un nivel de relación muy alto.

11.2 Regresión lineal simple

La regresión lineal simple la podemos definir como un modelo matemático donde tendremos una variable dependiente Y, así como las variables independientes X, siendo b la influencia de estas variables independientes sobre Y (la variable dependiente):

$$y = a + bx$$

Veamos... todas las ecuaciones, en un principio, pueden parecer algo complicado, así que **y** diremos que es la puntuación; la constante **a** mucha gente la ve en términos de probabilidad. Por ejemplo, si tenemos una pregunta en una prueba de examen de opciones múltiples, con cuatro posibilidades de respuesta, cada posible elección nos da un 25% de oportunidad de éxito. Por tanto, en este ejemplo, usaríamos 25 para el valor de **a**.

Con esto ya tenemos construida la mitad de la ecuación. Ahora la parte **bx** es un poco más compleja de determinar. **x** vendrían a ser las horas de estudio, haciendo que 0 sea el valor mínimo en este caso. Entonces tendremos que x será nuestra variable independiente.

La parte final **b**, en este ejemplo, vendrá a ser cuanto se incrementa la nota debido a las horas de estudio.



EJEMPLO PRÁCTICO

Veamos como implementar la regresión lineal simple en R.

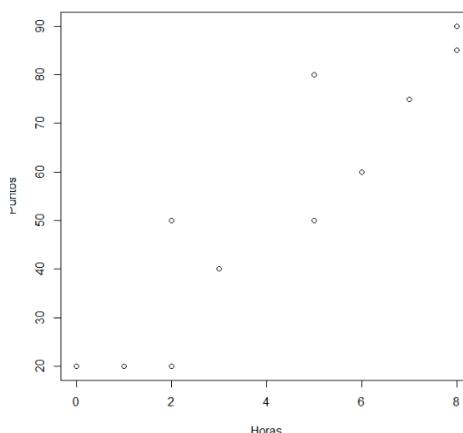
Lo primero que debemos hacer es cargar nuestros datos en R.
Lo hacemos de la siguiente manera:

Escribimos la dirección donde esta nuestro archivo y usamos la función **read.csv('ruta al archivo')**

```
> data = read.csv('/Users/sebastian/Desktop/Tutoriales/datawarehousing/datos.csv')
```

Posteriormente, podemos hacer uso de la función **plot**. Con ella es posible hacernos una idea de cómo se comportan los valores y cómo podemos observar, queda claro que existe una relación lineal positiva.

```
| > plot(data)
```



Ahora debemos utilizar la función **lm()**, que nos va a permitir ajustar nuestro modelo lineal y debemos decirle cuál es nuestra variable dependiente y cuál la independiente. Eso se consigue a través de la siguiente notación, en la cual usaremos el símbolo **~** de modo que a la izquierda se ubica la variable dependiente y a la derecha la(s) independiente(s):

```
lm(Puntos ~ Horas, data=data)
```

Esto lo guardaremos en una variable llamada modelo.

```

> modelo = lm(Puntos ~ Horas, data)
> summary(modelo)

Call:
lm(formula = Puntos ~ Horas, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-13.963 -6.947 -2.245  3.724 20.068 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 16.650     6.229   2.673  0.0255 *  
Horas        8.656     1.232   7.024 6.16e-05 *** 
---
Signif. codes:  0 '****' 0.001 '***' 0.01 '**' 0.05 '*' 0.1 '.' 1 

Residual standard error: 11.04 on 9 degrees of freedom
Multiple R-squared:  0.8457, Adjusted R-squared:  0.8286 
F-statistic: 49.33 on 1 and 9 DF,  p-value: 6.162e-05

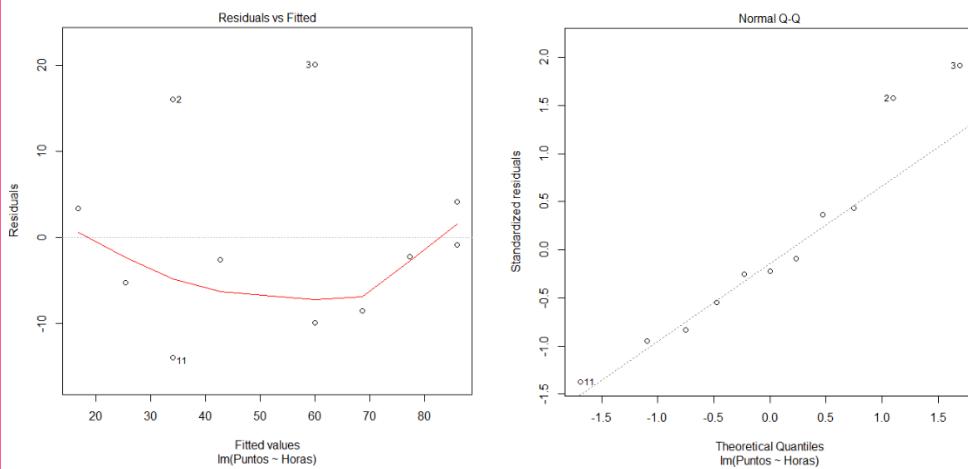
```

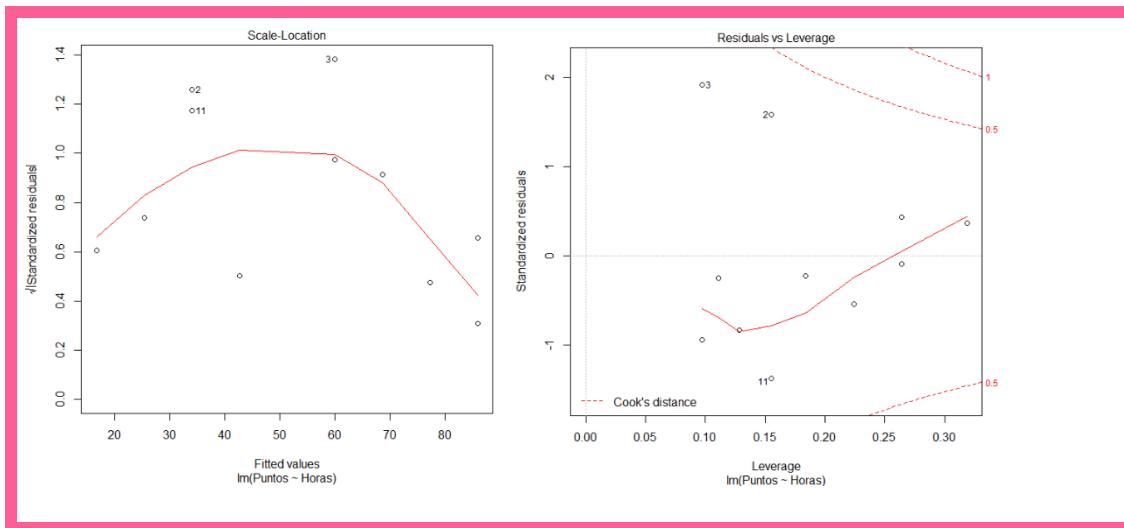
Ahora, vemos los valores para obtener nuestra fórmula en la sección de coeficientes. La fórmula sería:

$$\text{Puntos} = 8.656 * \text{Horas} + 16.650$$

A continuación, podemos graficar nuestro modelo para ver cómo se comporta:

```
plot(modelo)
```





Como podemos ver, es bastante sencillo implementar modelos lineales simples con WEKA y R. Por supuesto, estos softwares también tienen capacidades para implementar modelos más complejos.

Trata de medir el nivel de dependencia que existe entre ambas variables. Cuando su valor sea más próximo a 1, es una dependencia más profunda. Mientras que cuando tiende a -1, es más inversa su relación.



VIDEOTUTORIAL

En el siguiente vídeo se muestra la instalación y uso de R:

<https://vimeo.com/user64513894/review/447427650/06062b5d61>

RESUMEN FINAL

En esta unidad hemos revisado los conceptos básicos en la **ciencia de datos**, dado que estos son importantes para el entendimiento de conceptos posteriores.

Hemos conocido las principales características del big data, que recordemos vienen definidas por las 5 V de **volumen, velocidad, variedad, valor, veracidad** y que nos ayudaran a analizar nuestros datos de una manera eficaz.

Como hemos visto, el **machine learning** presenta un campo de acción amplio y está presente en productos y servicios desde hace ya unos cuantos años. Continúa creciendo en la actualidad y es importante saber el rol del ML en el data mining y en la ciencia de datos, porque son herramientas y tecnologías que en casi todas las ocasiones van de la mano.

También hemos revisado en este apartado los tipos de algoritmos usados en ML y su evolución histórica, desde la invención del test de Turing hasta la creación de Google Deep Mind.

Esta unidad nos ha servido para comprender los conceptos fundamentales de la doble perspectiva de la **estadística**: descriptiva e inferencial. Igualmente hemos valorado la importancia de los instrumentos estadísticos a la hora de realizar investigaciones empresariales y hemos aprendido a aplicar herramientas como los árboles de probabilidad.

WEKA y **R** son dos tecnologías potentes y con un gran número de funcionalidades que nos permitirán hacer, de una forma eficaz, data mining.

R es un potente lenguaje de programación, bastante usado en el mundo del ML y data mining. Hemos estudiado como cargar los datos y hacer un diagnóstico sencillo con la función summary, al igual que utilizar los gráficos e histogramas.

WEKA tiene años de historia en este sector y nos ofrece una interfaz sencilla y poderosa a su vez. Hemos aprendido varios aspectos importantes de su interfaz y también como instalarla. También hemos repasado unos conceptos básicos de data mining tales como valores faltantes, únicos y distintos.

En este módulo hemos revisado algunos aspectos importantes sobre el **preprocesamiento de los datos**. Hemos visto que son una parte fundamental para nuestros análisis y el hacerlos de una manera correcta nos ahorrará tiempo y aumentará la calidad de nuestros análisis y estudios.

Como hemos estudiado, la clasificación de los datos es una herramienta de vital importancia para el data mining y el estudio analítico. Y, a la vez, es un

tema complejo que requiere de conocimientos estadísticos y de informática, especialmente unas buenas nociones de algoritmia y programación. También es importante estar al día de los distintos softwares existentes.

En estas páginas también hemos podido revisar los tipos de clasificación existentes y hemos aprendido como implementarlos en WEKA y en R de manera fácil y sencilla.

Hemos comprendido que las estructuras de árbol son importantes para el proceso de clasificación y pueden llegar a ser bastante complejas cuando la cantidad de datos es grande. También hemos contextualizado y estudiado el conocido **teorema de Naive Bayes**.

A pesar de todo esto, queda mucho por estudiar en el campo de la clasificación. Es un área tremadamente grande y que necesitaría de años y la lectura de docenas de libros para dominarla.

Hemos entendido la importancia y la potencia del **clustering** mediante ejemplos prácticos y haciendo uso de WEKA y R. El actual desarrollo de la tecnología nos permite el estudio con una eficiencia bastante buena de los **datos sin etiquetar** o desconocidos, permitiendo la extracción de información importante para la toma de decisiones en las organizaciones.

Hemos estudiado con varios ejemplos el alcance de este tipo de analíticas aplicables a las finanzas, la geografía, la biología, etc.

También se ha estudiado el uso del algoritmo K-medias donde vimos los factores importantes que debe contemplar un buen análisis de clustering como son la importancia de las distancias intra e inter clúster. Igualmente conocimos como nos valemos de la distancia euclíadiana para calcularlo.

Sin dudas, el clustering es un campo muy denso, el cual requiere mucho estudio. Hay gran cantidad y variedad de algoritmos, tales como DBScan, MeanShift, Agnes, etc. que se pueden y deben estudiar para saber en qué circunstancias es mejor aplicar uno u otro. Aquí hemos marcado el punto de partida que ahora podemos continuar con estas mismas herramientas de WEKA y R.

Esta unidad, además, nos ha ofrecido una introducción básica al trabajo con modelos matemáticos, en concreto, implementando la **regresión lineal**. Haciendo uso de estas herramientas hemos comprobado que podemos estudiar la relación entre distintas variables, para con ello lograr realizar predicciones de comportamientos.