Facultad Ciencias Económicas y Empresariales

PROYECTO FIN DE CARRERA: Big Data. Análisis de grandes volúmenes de datos en organizaciones.

Alumno: Marta Ocaña del Llano
Tutora: María Jesús Giménez Abad

MADRID | Junio 2019

Resumen:

Este trabajo de fin de grado empieza describiendo qué es el big data para luego poder adentrarnos un poco en una de las disciplinas de la inteligencia artificial que esta muy relacionada con los grandes volúmenes de datos: el Machine Learning.
Como en el Machine Learning cada día aparecen nuevos algoritmos, en este trabajo también he hecho un pequeño resumen de los algoritmos más usados y conocidos, y luego escogí el algoritmo del clúster para enseñar la base matemática y estadística que tiene detrás, y también para desarrollar un ejemplo de clúster con datos reales con la intención de poder mostrar como se haría el clúster en el caso de tener bastantes datos.

Palabras clave: Big Data, Machine learning, Clúster, Algoritmos, Países UE

Abstract:

In this final degree project is described what big data is, with the aim of giving the reader a general knowledge before explaining one of the disciplines of the Artificial Intelligence which is related with the massive amounts of data: Machine Learning.
Every day new algorithms are created with the new amount of data, so I decided to resume in this project the main algorithms in Machine Learning, and then, to explain more in deep the mathematics and statistics of one of them: the cluster algorithm. I also decided to put an example of how a cluster would be created in the case of having an important volume of data.

Key words: Big Data, Machine Learning, Cluster, Algorithms, EU countries.

ÍNDICE

1 INTRODUCCIÓN

El **propósito** de este trabajo es explicar el Biga Data haciendo un análisis Top-down. Escogí este tema porque me quise informar acerca de en que consiste el mundo de los grandes datos, y para ello, me centré en un primer momento en que es el big data para luego explicar su relación con el Machine Learning (disciplina de la inteligencia artificial que crea sistemas que aprenden automáticamente gracias a algoritmos) para luego resumir los distintos tipos de algoritmos y desarrollar uno de ellos con datos reales. Básicamente es ir profundizando el tema de los grandes datos poco a poco e ir analizando de lo general a lo especifico.

Justificación y motivaciones: Este tema lo escogí porque desde que escuché hablar acerca del Big Data por primera vez, es algo que me llamó bastante la atención, y además, porque es un tema bastante actual y que está en boca de todo el mundo, ya que gracias a la gran cantidad de datos generados cada día, podemos aprender de ellos y mejorar los rendimientos que tienen las empresas o incluso la sociedad.
Me parecía un tema interesante de investigar, ya que pienso que este tema puede que esté muy presente en el mundo en el que tengo pensado moverme a partir del año que viene. Prácticamente todas las empresas tienen muy en cuenta los datos obtenidos de clientes, proveedores o incluso de la propia empresa a la hora de tomar decisiones, y creo que el hecho de que conozca un poco sobre Big Data y Machine Learning va a ser beneficioso en mi futuro profesional.

Objetivos: Hacer un análisis top-down para aprender acerca sobre los grandes datos, sobre el machine learning y los algoritmos que usa, y estudiar uno de esos algoritmos llegando a poner un ejemplo con datos reales.

Metodología: al ser un trabajo de investigación, la metodología a seguir ha sido la lectura de libros y artículos académicos. Luego para hacer el análisis del clúster busqué y organicé los datos para luego realizar el clúster e interpretarlo con el programa SPSS.

2 ¿QUÉ ES EL BIG DATA?

El big data (o grandes volúmenes de datos) ha ido creciendo de manera exponencial en los últimos años. Cada vez es más fácil y barato almacenar datos, y esto es un avance, ya que los data scientists y las personas que trabajan con estos datos van a tener información suficiente como para poder tener un mayor conocimiento de la empresa o negocio, y esto se va a traducir en una ventaja competitiva. El Big data, por ejemplo, nos va a ayudar a tener un mayor conocimiento de nuestros clientes, a darles un trato más personalizado, a mejorar las medidas antifraude del comercio electrónico, a evitar la perdida de clientes en manos de la competencia, y a ser más rápidos y eficientes en las decisiones que tomemos.

La definición del Big data, tal y como dice Luis Joyanes Aguilar en su libro, no está consensuada. Una definición que me gusta es la que usa McKinsey: “el Big Data se refiere a los conjuntos de datos cuyo tamaño está más allá de las capacidades de las herramientas típicas de software de bases de datos para capturar, almacenar, gestionar y analizar” (Mckinsey Global Institute 2011)
Lo que si que está claro es que la definición variará según los tipos de empresas: “para unas empresas prima el volumen; para otras, la velocidad; para otras, la variabilidad de las fuentes. Las empresas con muchos volúmenes van a estar de interesadas en capturar la información, guardarla, actualizarla e incorporarla en sus procesos de negocio; pero hay empresas que, aunque tengan mucho volumen, no necesitan almacenar, si no trabajar en tiempo real y a gran velocidad. Otras, por el contrario, pueden estar interesadas en gestionar diferentes tipos de datos” (Joyanes 2011, pg 3)

En un estudio de BSA se desveló que el 90% de los datos que tenemos se ha producido en los últimos dos años. “Se han creado 5 exabytes desde el nacimiento de la civilización hasta 2003. A día de hoy, esa información se crea cada dos días aproximadamente y ese crecimiento no deja de aumentar” (Eric Smichdt, CEO de Google). Y sin embargo, a pesar de tener datos más que suficientes, el estudio de IDC ha revelado que en el año 2012 sólo se usaba el 5% de los datos almacenados, lo que da a entender que el potencial del big data

es enorme. Se piensa que la razón por la que los datos no dejan de crecer es porque estamos en la 4º Revolución Industrial, y cada vez que hemos cambiado de etapa, ha habido cambios muy significativos en el estilo de vida de las personas:

- La primera Revolución Industrial sucedió en 1786 cuando se creó la máquina de vapor. Gracias a este invento se desarrollaron las industrias del transporte y la industria textil, y se introdujeron novedades en la sociedad como la calefacción a gas, el alcantarillado o las maquinas de coser.
- La segunda Revolución Industrial se sitúa a mediados del siglo XIX con la aparición de la electricidad y del motor que hizo posible el invento del automóvil. Además, en las viviendas se fueron introduciendo los teléfonos y el alumbramiento eléctrico.
- La Tercera Revolución data en 1920, época en la que la se descubren nuevos tipos de energía (por ejemplo, la atómica), se hacen avances importantes en la industria de la aviación, aparecen los primeros antibióticos, etc.
- La revolución informática o cuarta revolución industrial que según el World Economic Forum (WEF) no es una prolongación de la tercera revolución. Con esta revolución estamos tratando de llegar a la automatización completa, que gracias al internet de las cosas, la nube, Inteligencia Artifical (AI), neurotecnologías, etc. las máquinas son capaces de tomar decisiones.

Vamos hacia la automatización completa e inteligente. Los sistemas serán capaces de controlarse a si mismos durante todo el proceso de fabricación. Según un informe de Accenture, todas estas innovaciones van a ser capaces de crear 14,2 billones de dólares a la economía mundial en un periodo de 15 años. Pero lo que si que hay que tener muy en cuenta es que aquellas empresas que no innoven y que no se adapten se van a quedar atrás con respecto a sus competidores y no van a tener los mismos beneficios (hay quien llama a esto Darwinismo tecnológico). De momento, el mercado

asiático es el que más exhaustivamente está usando estas nuevas tecnologías.

2.1 CARACTERÍSTICAS DE LOS BIG DATA

Para considerar datos como big data tendremos que ver si se cumplen como mínimo las siguientes tres características:

a) volumen: la cantidad de datos disponibles en las empresas. Esta cifra es cada vez mayor debido a que hoy en día prácticamente todo produce datos, y además, cada vez en mayor medida. Según el informe de BSA, en el 2014 se crearon tantos datos como para poder formar una pila de DVDs que vaya desde la Tierra a la luna y vuelva.
b) velocidad: los flujos de datos son cada vez mayores y más rápidos. Además, gracias a la nube tenemos los datos en todos sitios.
c) variedad: representa los tipos de datos que se generan, ya sean:
    a. estructurados: datos fáciles de analizar porque tienen una forma o esquema fijo.
    b. no estructurados: No están definidos en una estructura. Como ejemplo de estos datos podemos poner fotografías, texto, audios, etc., que no suelen estar en las bases de datos de la empresa. Son los datos más difíciles de entender y de conseguir.
    c. semiestructurados: datos sin formato fijo, pero con etiquetas que determinan cómo proceder para analizar los datos. No son tan fáciles de comprender como los datos estructurados.

Aparte de estas tres características, se le puede añadir veracidad y valor como requisitos a los datos si quieren ser parte del Big Data. Las empresas van a buscar que sus datos sean fiables y eficientes.

2.2 TIPOS DE DATOS

Según la procedencia de los datos, estos van a ser:

datos web: Son aquellos datos que se generan al navegar en páginas web. Estos datos proporcionan información sobre las preferencias de los consumidores, con lo que se puede detectar motivaciones de compra o posibles acciones futuras. Ejemplos de estos datos son vistas de productos en internet, visualizaciones de videos, comentarios en blog…

datos de texto: es una de las fuentes de datos más comunes y más grandes de los big data. Ejemplo de estos datos son los tweets, los correos electrónicos, whatsapps… Para poder analizar estos datos y que nos sean útiles lo que se hace es crear datos estructurados a partir de los datos no estructurados que se recogen.

Datos de sensores: En los últimos años se han instalado sensores en la mayoría de las máquinas, ya sean teléfonos móviles, medios de transporte (aviones, coches, camiones), electrodomésticos, etiquetas, etc., con el propósito de obtener datos a tiempo real. Aunque estos datos sean estructurados tienen la dificultad de presentarse en volúmenes muy grandes que, sin las herramientas necesarias, puede dificultar mucho su análisis.

Datos de RFID y NFC: son las etiquetas de barras que presentan alimentos, medicamentos o ciertos artículos. Estas etiquetas nos facilitan la localización, la logística, los pagos, el seguimiento de paquetes, inventarios automáticos…

Datos de las redes sociales: la analítica de estos datos se reconoce como analítica social. El análisis de estas redes puede servir, por ejemplo, para mejorar la publicidad y para conocer mejor los gustos de nuestros clientes.

Datos de las operadoras de telecomunicaciones: son las llamadas y mensajes que una operadora puede registrar. Es considerado Big Data por la cantidad de volumen que se gestiona.

2.3 ÁREAS EN DONDE EL BIG DATA TENDRÁ CIERTA IMPORTANCIA

En el informe de McKinsey Global Institute del año 2011 se seleccionaron cinco dominios en donde el big data tendrá una importancia significante:

a) sector de la salud: la atención hospitalaria será más personalizada y eficiente, tendremos muchos más datos de enfermedades y medicinas, haciendo que las enfermedades sean mucho más fáciles gracias a todos los datos que se recogen como puede ser datos del ADN o genoma humano.
b) administración publica, ya que las autoridades se enfrentan a un aumento de datos que pueden abarcar temas tan distintos como puede ser la educación, la ciencia, la medicina…
c) comercio minorista: con el volumen de datos podríamos llegar a conocer a nuestros clientes, a desarrollar una atención personalizada y a gestionar mejor la imagen de la marca.
d) fabricación: datos a tiempo real van a ayudar a la empresa a tener una buena logística, inventario y detectar rápido los errores de fabricación.

El mito que se comenta en el informe de BSA acerca de que los datos solo benefician al sector tecnológico es completamente falso, ya que los datos benefician a la economía en su conjunto. Se piensa que gracias a los datos, las empresas podrán aumentar entre un 5 y un 6 por ciento su productividad (ahorraría energía, reduciría costes, aumentaría la vida útil de los activos…) y además, el uso de los datos no solo crearía trabajo en el departamento de tecnología si no que se espera que por cada puesto en IT se creen tres puestos de trabajo externos a este departamento (mejora global para la economía)

2.4 SEGURIDAD DE LOS DATOS

Open data se refiere a los datos públicos y privados que deberían de estar a disposición de los ciudadanos y empresas, aunque algunos datos como pueden ser los datos personales deben de estar protegidos.
Los datos personales son aquellos que aportan información sobre una persona física y que permite identificar a esa persona. Los datos personales que hayan sido anonimizados, de forma que la persona no sea identificable o deje de serlo, dejarán de considerarse datos personales. Para que los datos se consideren verdaderamente anónimos, la anonimización debe ser irreversible.

En los últimos años se han ido aprobando leyes para proteger los datos de carácter personal, y desde el 25 de mayo del 2018, los países de la UE siguen la misma regulación a la que han llamado RGPD. Esta regulación lo que pide es que se respeten los derechos y libertades de las personas, ya que al final, los datos afectan de manera directa a la privacidad y seguridad. Desde esta fecha, es el propio ciudadano el que tiene que tiene que aceptar la recogida y uso de sus datos.

Esta ley afecta a todas las empresas que recojan datos en el desarrollo de sus actividades, ya sean datos a proveedores, a clientes, pacientes, empleados, etc y a todas las personas que tengan dominios en internet. Por ejemplo, si tienes un blog en el que la gente deja comentarios a través de su email, tienes que seguir esta nueva ley. Básicamente en lo que consiste esta ley es en informar que estamos recogiendo datos, explicar cómo se van a tratar, explicar cuales son los fines, el tiempo que se van a almacenar dentro de la empresa y cómo se van a proteger, y luego la persona tiene que aceptar esos términos.

2.5 GESTIÓN DE LOS DATOS PARA UN RENDIMIENTO SATISFACTORIO

Según los investigadores del MIT, McAfee y Brynjoffson, una buena gestión de los datos tiene que ser ejercida en 5 apartados:

1. Liderazgo: No solo basta con tener buenos datos, si no que tendremos que saber que es lo que tenemos que conseguir a partir de esos datos. Los líderes tendrán que usar estos datos para ver oportunidades, la evolución del mercado, etc.
2. Gestión del talento: El Big Data trae consigo nuevas profesiones, como los analistas y científicos de datos. Las empresas demandarán cada vez más este tipo de profesionales, que tendrán que tener aptitudes para la estadística y para organizar y visualizar información.

   El científico de datos se diferencia del analista en que estos se centran en la toma de decisiones además de en el análisis. Por tanto, las funciones que tienen son:

   1. Buscar las fuentes de datos que sean interesantes
   2. Limpiar aquellos datos que son prescindibles

3. Una vez juntados los datos los analizarán desde distintas perspectivas para encontrar valor
4. Finalmente aportar ideas que saquen provecho de los datos.

El científico de datos es una persona que tiene que entender el negocio y también tiene que entender la tecnología usada. Tiene que saber entender los datos para poder agregar valor.

3. Tecnología: Las herramientas para guardar y analizar los datos son cada vez más asequibles.
4. Toma de decisiones: Una vez que está definido el objetivo a alcanzar, tendremos que identificar la información necesaria. A mayor volumen, menos datos pasarán desapercibidos, pero tenemos que encontrar el punto óptimo para evitar perder eficiencia y que el modelo no sea excesivamente complejo.
5. Cultura corporativa: cuando los datos sean analizados, tienen que mostrar el resultado de manera simple y entendible para todo el mundo, y las herramientas para el análisis deben de ser simples para que la pueda usar el personal de la empresa.

2.6 CONCEPTOD SOBRE LA GESTIÓN DE DATOS:

"**La ciencia de los datos o data science** se refiere a las técnicas y teorías implicadas en el proceso de adquirir, limpiar, ordenar, procesar, mostrar, almacenar, los datos que nos pueden ayudar a detectar problemas en nuestro negocio o a optimizar y mejorar nuestros procesos" (Luis Joyanes, pag 95). Ejemplo de cómo las empresas usan el data science son Amazon, la cual recomienda productos basados en las búsquedas del mismo usuario y en las búsquedas de otros usuarios con gustos similares, o Facebook, red social que recomienda personas que puedes conocer basado en tu lista de amigos.

La **inteligencia de negocios en el big data o business intelligence** son las herramientas que sirven para controlar que los datos que tenemos y el análisis son correctos. (Luis Joyanes, pg 103)

La **analítica del Big Data** son los procesos que permiten analizar los datos rápidamente y sin un coste muy elevado. Hoy en día no vale con simplemente tener los datos almacenados, hay que saber qué información nos dan estos datos. Es una de las fases por las que pasan los datos.

2.7 FASES POR LAS QUE PASAN LOS DATOS:

2.7.1 ADQUISICIÓN:

Se buscan los datos necesarios, sean internos o externos a la empresa, estructurados o desestructurados.
Es necesario tener en mente la pregunta que tratamos de resolver, y a partir de ella buscaremos los datos adecuados. No tiene sentido recoger datos innecesarios, ya que además de ocupar espacio cuando son almacenados, es una perdida de tiempo que a la empresa le puede costar dinero.

2.7.2 ORGANIZACIÓN:

Se prepara la información para que puede ser analizada y poder sacar los mejores resultados posibles.
Por ejemplo, en el clúster que he realizado y que explicaré más adelante, al empezar he tenido que organizar todos los datos en un Excel para luego exportarlo al programa SPSS. Y no solo eso, si no que tienes que poner los datos de manera correcta para que el programa los lea adecuadamente y tipificar para poder comparar las variables de manera correcta.

2.7.3 ANALISIS DE DATOS:

Hay diferentes categorías dentro del análisis de datos:

- **La analítica del Big Data** "es el proceso de examinar grandes cantidades de datos de una variedad de tipos para descubrir patrones ocultos, correlaciones desconocidas y otra información útil. Dicha información puede proporcionar ventajas competitivas sobre organizaciones rivales y

brindar beneficios en los negocios tales como un marketing más eficiente y un aumento de los ingresos" (Luis Joyanes Aguilar. Big Data, Análisis de grandes volúmenes de datos en organizaciones pg 243)

- **La analítica web** se centra en los datos de las paginas web. La dificultad es encontrar los datos que son significativos, ya que hay una gran cantidad de datos que no aportan ninguna información. Para saber diferenciar la información relevante de la que no lo es, según Luis Joyanes, se usan métricas tales como:
    - Visitas: numero de veces que se accede a una pagina web.
    - Visitantes: personas que acceden. No es relevante porque si un visitante accede varias veces, solo se le cuenta una vez.
    - Visitante único: numero de personas diferentes que acceden a un sitio web
    - Tiempo en la página
    - Tasa de rebote: cuando las personas pasan poco tiempo en la página web y no hacen clics. Es decir, que no es de interés para el visitante.
    - Tasa de conversión: Es el porcentaje de resultados conseguidos por el numero de visitantes únicos. Por ejemplo, si de 100.000 visitantes, 29.000 leen el articulo de la pagina web, la tasa de conversión es de casi un 30%.
    - Compromiso: Razones por las que un visitante decide visitar una pagina web en vez de otra. Es de las métricas más difíciles.

Los KPI (Key Performance Indicator) o indicadores claves de desempeño se han definido por varios autores como las métricas que ayudan a conseguir los objetivos del negocio. Todos los KPI son métricas, pero no todas las métricas son KPI.

Una vez que están definidas las métricas y los KPI se realizan los informes que permiten analizar los datos, y el objetivo de estos informes es facilitar la toma de decisiones.

En Google Analytics hay tres tipos de informes: el estándar (datos sobre los visitantes, el origen de las visitas, conversiones…), el personalizado

(creados para el usuario con los datos que necesite) y los informes sociales (informes para saber el impacto de las redes sociales en los objetivos de la empresa y para saber el valor que aportan las conversiones realizadas). No nos podemos olvidar que los datos se pueden y se deben segmentar para hacer un mejor análisis. Google Analytics permite segmentar según usuarios nuevos, visitas recurrentes, visitas con conversiones…

- **Analítica Móvil** es un campo dentro de la analítica web y pretende recoger y analizar los datos de los smartphones. Estos datos nos darán información sobre quién usa la aplicación, qué demandan estos usuarios y cómo funciona el sitio web desde terminales móviles.

- **La analítica social** es la disciplina que analiza, mide y pone en contexto los datos no estructurados de correos electrónicos, blogs, redes sociales, mensajería instantánea, etc.

  A medida que aumentan los datos de este tipo, mas complejo es su análisis. Se debe de identificar los datos relevantes entre los miles y millones de datos, datos que en muchos casos son no estructurados. Y también hay que tener en cuenta que el número de datos crece cada día más, debido al crecimiento casi a diario de personas con móviles y tabletas a su disposición..

  Las métricas que utiliza la analítica social son los seguidores, las publicaciones, la audiencia potencial, el numero de visitas a la pagina web, trending topics…

  Las métricas que sirven para planear y cumplir los objetivos empresariales específicos son los indicadores clave de rendimiento.

  Gracias a las nuevas tecnologías es posible analizar automáticamente la opinión o sentimiento de los comentarios escritos en las redes sociales. Este análisis es comúnmente denominado análisis de sentimientos y lo que hace es traducir estos sentimientos a indicadores medibles que puedan aportar información a la empresa. Es muy útil para que la empresa tenga una idea sobre la opinión general que se tiene sobre una marca, conversaciones entre usuarios, detectar nuevas tendencias…

2.7.4 DECISIÓN:

Tomar decisiones a tiempo para poder aprovechar las ventajas que te da el análisis de los datos.

La dirección comunicará la decisión tomada al resto de miembros de la organización con el objetivo de que todos participen en conseguir los objetivos que se piensan posibles tras haber analizado los datos.

Esta decisión va enfocada a aumentar la competitividad de la empresa y a ser más eficientes. Todos los procesos anteriores nos sirven para entender mejor a los clientes y a la propia empresa, pero el buen resultado de la empresa dependerá de que una vez que tenemos la información y la hemos interpretado correctamente tomemos las decisiones correctas.

3 ¿QUÉ ES EL MACHINE LEARNIG?

"El término Machine Learning se refiere a la detección automática de patrones significativos en los datos" (Shalev-Shwartz, S. and Ben-David, S, 2014), y para ello se utilizan distintos algoritmos. En otras palabras, cuando hablamos de Machine Learning, estamos hablando de la disciplina de la Inteligencia Artificial que se centra en aprender un modelo determinado basándose en unos datos.

El Big Data está relacionado con el Machine Learning porque cuantos más datos, más adecuado será el algoritmo usado (Normalmente, se forman primero los algoritmos y más tarde se añaden los datos, pero en el caso del Machine Learning es justo el proceso inverso: se recogen los datos y a partir de ellos se forma el algoritmo). Sin suficientes datos no vamos a poder hacer buenas predicciones, puesto que lo tendremos más difícil para detectar cualquier cambio.

Pero el hecho de que necesitemos muchos datos no significa que nos valga cualquiera: hay que comprobar que los datos sean adecuados, que estén bien contextualizados y que sean suficientes. No vale para nada tener unos datos si no los entendemos y no los sabemos poner en contexto.

En los últimos años se ha discutido la importancia de la correlación de datos. A pesar de reconocer su importancia, se llegó a la conclusión de que puede llevar a confusiones: "Puede haber una correlación entre la caída de las hojas de los árboles y la compra de abrigos. Las dos cosas pasan a la vez por la bajada de temperaturas, pero no hay ninguna relación entre hojas y abrigos" (Machine Learning for Dummies pg 21)

El principal reto es saber qué datos necesitamos y en qué cantidad para que el algoritmo de Machine Learning usado sea fiable y poder hacer unos buenos análisis. Si tenemos un buen análisis, podremos llegar a anticipar el futuro. Las predicciones del futuro están basadas en datos históricos y a partir de modelos predictivos puede llegar a reaccionar ante cambios en el ambiente de la empresa.

Los análisis que se hacen suelen ser:

- descriptivos: se usan para conocer la empresa, y para ello es necesario comprender los datos históricos.
- predictivos: sirven para anticipar cambios observando patrones en los datos. En estos análisis es necesario actualizar constantemente los datos para no pasar por alto ningún cambio de tendencia.

La estadística y la minería de datos son dos disciplinas imprescindibles dentro del Machine Learning:

- La estadística lo que hace es analizar los datos, y en un modelo estadístico lo que se hace es ver la validez de un algoritmo para saber si es adecuado para hacer predicciones.
- La minería de datos “es un proceso interactivo que extrae patrones predictivos ocultos en los datos usando tecnologías de Inteligencia Artifical y la estadística” (Mena 1999. Citado por Aluja)

“Los datos almacenados son un tesoro para las organizaciones. Es donde se guardan las interacciones pasadas con los clientes, la contabilidad de sus procesos internos, y representan la memoria de la organización. Pero con tener memoria no es suficiente, hay que pasar a la acción inteligente sobre los datos para extraer la información que almacenan. Este es el objetivo de la Minería de Datos” (Aluja T.)

El Machine Learning es una de las disciplinas de la Inteligencia Artificial, y este a su vez tiene distintas modalidades dependiendo del volumen y del tipo de datos:

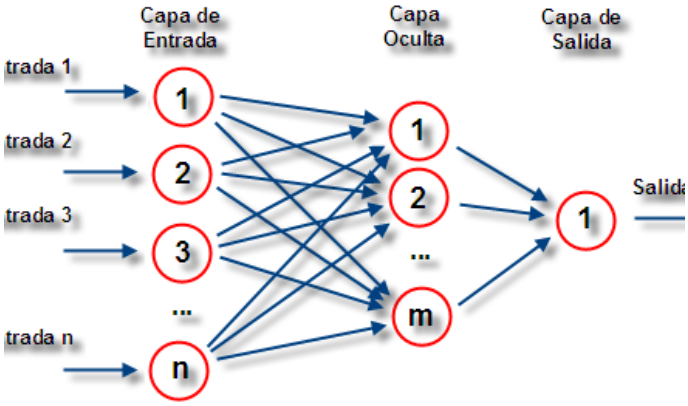
1. Aprendizaje supervisado: Permite buscar patrones en datos históricos que suelen estar etiquetados.

   Existe dos tipos de aprendizaje supervisado: de regresión y de clasificación. La regresión predice valores continuos, y la clasificación asigna distintas categorías dentro de un conjunto de datos.
2. Aprendizaje sin supervisión: Es la modalidad adecuada cuando tenemos muchos datos que no están etiquetados de ninguna manera, por ejemplo, datos sacados de las redes sociales. “Su función es

la agrupación, por lo que el algoritmo debería catalogar por similitud y poder crear grupos, sin tener la capacidad de definir cómo es cada individualidad de cada uno de los integrantes del grupo" (Zambrano J. online)

3. Aprendizaje por refuerzo: El algoritmo recibe feedback del análisis de los datos por lo que poco a poco va aprendiendo. Es un sistema de aprendizaje basado en prueba y error. Es el modelo usado en un montón de robots y coches automáticos.
4. Redes neurales y aprendizaje profundo: Las redes neuronales es un modelo basado en varias unidades conectadas (neuronas). Cada conexión puede transmitir información (señales) de una neurona a otra. Las neuronas que reciben la información, la procesan, y la mandan a otras neuronas que estas conectadas a ella. Como resultado de estas transmisiones, decimos que este modelo tiene como mínimo tres niveles: input (o entrada), nivel en el que se reciben los datos y sus probabilidades; nivel oculto, que es la capa en donde se procesan los datos, se asignan pesos a las probabilidades y se conectan los nodos de entrada y salida; y nivel de outputs (o salidas) en la que se extrae la información media que se ha ido transmitiendo de neurona a neurona. Representan los valores de la predicción.

   El aprendizaje profundo es un modelo de Machine Learning muy similar a las redes neuronales, pero con la diferencia que tiene varias capas ocultas.



*Fuente sacada de google imágenes (https://images.app.goo.gl/Pn8npDMaMNtQekSS9)

Los algoritmos son instrucciones a los ordenadores para que sepan cómo interactuar, manipular y transformar los datos. Los algoritmos del Machine learning son distintos del resto de algoritmos porque en este caso, los datos van creando el algoritmo, por lo que, a mayor número de datos, más exacto será el algoritmo. En el resto de algoritmos (aquellos que no pertenecen al Machine Learning), el programador crea el algoritmo en un primer momento para luego usarlo con distintos datos.

Los tipos de algoritmos que usa el Machine Learning son:

Modelo	Tipos de datos
Bayes	Este algoritmo hace suposiciones anticipadas sobre la posible distribución de la respuesta, y se basa en el modelo estadístico bayesiano. Es especialmente útil cuando no tenemos datos suficientes como para hacer un modelo fuerte y sin errores. Es clave tener información previa de los datos (información a priori), ya que a partir de aquí se calcula la probabilidad a posteriori. Debido a que tenemos que conocer y clasificar los datos con anterioridad, estamos hablando de **aprendizaje supervisado**
Clúster	Es un método que se usa con aquellos **datos que aun no han sido etiquetados, por lo que pertenece al aprendizaje sin supervisión**. Lo que hace este algoritmo es dividir los datos en distintos grupos, de manera que los datos que pertenecen a un mismo grupo son entre ellos lo más homogéneos posible, y consecuentemente, con respecto a otros grupos, son heterogéneos.
Árboles de toma de decisiones	Este algoritmo representa las decisiones de manera que podemos ver visualmente cual es la estrategia que debemos seguir para alcanzar un objetivo.

Modelo	Tipos de datos
	Cada condición (nodo) se divide en distintas ramas hasta que estas ramas no se dividen más (entonces se denominan hojas) Lo positivo de este algoritmo es que se puede usar en **datos numéricos y categóricos,** pero, sin embargo, es muy fácil tender al sobreajuste, o que pequeños cambios en los datos hagan que haya unas variaciones grandes en el árbol de toma de decisión.
Reducción de la dimensionalidad o Reducción de datos	Dentro del Machine Learning hay veces en los que tenemos demasiadas variables y demasiada información, y esto, en muchos casos nos dificulta poder analizar y trabajar con los datos. Lo que hace este logaritmo es reducir las variables al mínimo necesario, de manera que no tenemos datos redundantes. Este modelo es complejo y a su vez se divide en dos modelos matemáticos que son: - selección de variables → se trata de escoger las variables más significativas. Para ello se usa la asociación y la correlación) - análisis de componentes principales → se construye una variable nueva a partir de la fusión de dos ya existentes
Algoritmos basados en la instancia	Se crea un modelo a partir de una base de datos y luego se van añadiendo nuevos datos. El objetivo es hacer una predicción comparando la similitud de los nuevos datos con los datos de la base inicial.
Redes Neuronales y Aprendizaje profundo	Este modelo ha sido previamente mencionado, pero podemos añadir que es un modelo que se usa con datos **sin etiquetar y sin ninguna estructura**. La intención es copiar como la información pasa por las

Modelo	Tipos de datos
	neuronas del cerebro humano. Las neuronas que reciben la información, la procesan, y la mandan a otras neuronas que estas conectadas a ella. Este algoritmo se usa mucho en análisis de comercialización (ver si una campaña de publicidad tendrá éxito o no), analizar procesos industriales, predecir los movimientos de una acción, etc.
Regresión lineal	Es un algoritmo dentro del **aprendizaje supervisado** que mide la correlación entre datos para ver la relación que hay entre estos. Lo que hay que tener en cuenta es que las predicciones de este modelo se basan en datos históricos y que no vamos a tener unas buenas predicciones si no entendemos a la perfección el contexto en el que se encuentran los datos.
Regularización para evitar el sobreajuste	Este algoritmo se puede aplicar en c**ualquier modelo de machine learning**. Lo que pretende es modificar el modelo para que no haya un sobreajuste que de unas predicciones inadecuadas. El sobreajuste ocurre porque los modelos son creados para unos datos en concreto, por lo que no valen para todos los datos.
Sistemas basados en reglas	Este algoritmo es de los más fáciles de entender, ya que simplemente consiste en una regla formada por una premisa y por una conclusión (si x → entonces y). Las conclusiones se pueden sacar por modus ponens (Hay una regla: si X es cierto, Y es cierto. El hecho es que x sea cierto, implica Y también es cierto) o modus Tollems (La regla sería la siguiente:

Modelo	Tipos de datos
	Si X es cierto, Y es cierto. El hecho es que Y sea falso implica que X también es falsa)

4 CLÚSTER

El algoritmo del clúster permite trabajar con datos sin estar clasificados. Por norma general, no hay una regla que establezca si estamos ante un buen clúster o no. Va a depender mucho de usuario y de la necesidad que pretenda satisfacer: podemos estar interesados en buscar grupos de datos homogéneos, grupos naturales que nos ayuden a descubrir nuevas cosas acerca de los datos que tenemos, o podemos querer agrupar los datos en grupos según su utilidad.

Lo que va a hacer el algoritmo es crear distintos grupos con los datos que sean homogéneos y luego va a comprobar que estos grupos sean heterogéneos entre ellos.

Ejemplos de aplicaciones del análisis clúster en diferentes campos:

1. **Marketing:** Pueden ser usado para descubrir nuevos segmentos de consumidores. Esta técnica se denomina segmentación dentro de esta disciplina.
2. **Biología:** Lo utilizan para clasificar a las diferentes especies de plantas y animales.
3. **Librerías:** Lo utilizan para diferenciar los libros según los temas o la información.
4. **Seguros:** Es usado para conocer a los clientes, sus pólizas y los fraudes.
5. **Planificación urbana:** Se crean distintos grupos para estudiar los distintos hogares.
6. **Estudios de terremotos:** Se puede determinar cuáles son las zonas de peligro basándonos en las zonas que han sufrido un terremoto.

4.1 ANALISIS CLÚSTER

El análisis clúster es una técnica multivalente que agrupa objetos, individuos, etc. y se basa en las características que forman estos objetos.
El objetivo buscado en la mayoría de los análisis clúster es identificar un pequeño número de grupos de manera que los elementos que pertenezcan a un grupo son más similares entre ellos que los que forman otro grupo. Lo que se hace es reducir información. Pasamos de tener un conjunto de n elementos a tener g grupos (siempre n>g).
La diferencia del análisis clúster con respecto a otras técnicas de agrupación es que en este caso no conocemos los grupos a priori.

4.2 DESCRIPCIÓN DEL PROCESO

El proceso habitualmente comienza de la siguiente manera:

1. tomamos n objetos de los cuales obtenemos p individuos de cada uno de ellos
2. La matriz de datos nxp va a transformarse en una matriz nxn de semejanzas/similitudes o distancias, donde las semejanzas o distancias se establecen entre bases de objetos a través de las p variables.
3. El siguiente paso es seleccionar un algoritmo de agrupación (clusterización). Este algoritmo define las reglas para agrupar objetos en diferentes grupos teniendo en cuenta la semejanza entre los objetos.
4. El objetivo es llegar a agrupar los objetos que muestran una elevada homogeneidad interna y una alta heterogeneidad externa.

4.3 MEDIDAS DE SIMILARIDAD

Las medidas de similaridad se pueden clasificar en dos clases según los datos disponibles. Si los datos son de naturaleza cualitativa, se usarán medidas de asociación, y si los datos se pueden medir de alguna manera, es decir, son de naturaleza cuantitativa, usaremos medidas de distancia.

4.3.1 MEDIDAS DE DISTANCIA (datos cuantitativos)

Asumimos que los datos han sido recogidos de n objetos. Cada objeto está representado por un vector → $X' = (x_1, x_2, \dots, x_p)$

Las medidas se suelen basar en la **métrica Minkowski**

$$dij = \left\{ \sum_{K=1}^{P} \left| x_{ik} - x_{jk} \right|^r \right\}^{1/r}$$

- **distancia euclídea (r=2 en la métrica Minkowski)**:

$$dij = \left\{ \sum_{K=1}^{P} \left| x_{ik} - x_{jk} \right|^2 \right\}^{1/2}$$

Ejemplo:

Personas	Pesos en libras	Altura en pies	Altura pulgadas
A	160	5,5	66
B	163	6,2	74,4
C	165	6,0	72

Utilizando la métrica euclídea, las distancias entre las personas son:

d_{AB} → $(3^2+0{,}7^2)^{1/2} = 3{,}08$

d_{AC} → $(5^2+0{,}5^2)^{1/2} = 5{,}02$ → menos similar

d_{BC} → $(2^2+0{,}2^2)^{1/2} = 2{,}01$ → más similar

Hay casos en los que esta distancia no es correcta por errores en la escala de medidas que usamos. Si utilizásemos, por ejemplo, pulgadas para medir la altura, ya no tendríamos las mismas conclusiones:

d_{AB} → $(3^2+8{,}4^2)^{1/2} = 8{,}92$ → menos similar

d_{AC} → $(5^2+6^2)^{1/2} = 7{,}81$

d_{BC} → $(2^2+2{,}4^2)^{1/2} = 3{,}12$ → más similar

Es por esto, que en muchas ocasiones (por no decir la mayoría), cuando no se tiene la misma escala se decide tipificar.

- **distancia absoluta o city black (r=1 en la métrica Minkowski)**

$$dij = \left\{ \sum_{k=1}^{p} |x_{ik} - x_{jk}| \right\}$$

4.3.2 MEDIDAS DE ASOCIACIÓN (datos cualitativos)

Este tipo de medidas toman valores del rango 0 - 1 y están basadas en el razonamiento que 2 individuos pueden ser vistos como similares si comparten ciertos atributos.

Cada variable va a reflejar la presencia o la ausencia del atributo:

P / A	1	2	3	4	5	6
A	0	1	1	0	1	1
B	1	0	1	0	0	1

Con los anteriores datos se construye una tabla de doble entrada

PERSONA B \ PERSONA A	+	-	
+	2 (grupo a)	1(grupo b)	3
-	2(grupo c)	1(grupo d)	3
	4	2	6

Algunas de las medidas que se definen para este tipo de datos son:

$1 \rightarrow \frac{a+d}{a+b+c+d} \rightarrow \frac{nº\ coincidencias}{total} = 3/6$ $\quad 2 \rightarrow \frac{a}{a+b+c} = 2/5$ $\quad 3 \rightarrow \frac{2a}{2a+b+c} = 4/7$

$4 \rightarrow \frac{2(a+d)}{2(a+b)+b+c} = 6/10$ $\quad 2 \rightarrow \frac{a}{a+2(b+c)} = 2/8$ $\ 3 \rightarrow \frac{a}{a+b+c+d} = 2/6$

Como podemos ver, las medidas difieren entre ellas, por lo que la elección es una cuestión a discutir, pues proporcionan valores distintos para el mismo conjunto de datos.

4.4 TÉCNICAS DE AGRUPACIÓN:

Hay métodos jerárquicos y no jerárquicos.

4.4.1 Métodos jerárquicos:

Estos modelos tratan de fusionar o dividir los datos. Dentro de estos métodos podemos hablar de:

- Métodos aglomerativos: fusionan el conjunto de n objetos en diferentes grupos.
- Métodos disociativos: parten de un conjunto de n objetos y van realizando subdivisiones.

Ambos métodos parten de un Dendograma que es un diagrama de árbol que muestra las fusiones o divisiones realizadas en cada nivel.

4.4.1.1 Métodos Aglomerativos:

Este modelo funciona de la siguiente manera: en el primer nivel, cada objeto pertenece a un propio clúster. En el siguiente nivel, los dos o más objetos que son mas cercanos se fusionan. En el tercer nivel se añade un nuevo objeto al clúster con dos objetos (pasamos a tener un clúster con tres objetos) o se crea un nuevo clúster de dos objetos. El proceso podría continuar hasta tener un solo clúster de n objetos.

Hay dos tipos de técnicas:

- **el vecino más cercano** (simple linkage): se basa en la distancia mínima. Selecciona a los dos individuos que tienen la distancia más corta entre ellos para que formen el primer clúster y luego, una vez formado este primer clúster se siguen juntado los individuos con la mínima distancia.
  Ejemplo:

$$\begin{matrix} & A & B & C & D & E & \\ \begin{pmatrix} 0 & 1 & 5 & 6 & 8 \\ 1 & 0 & 3 & 8 & 7 \\ 5 & 3 & 0 & 4 & 6 \\ 6 & 8 & 4 & 0 & 2 \\ 8 & 7 & 6 & 2 & 0 \end{pmatrix} & \begin{matrix} A \\ B \\ C \\ D \end{matrix} \end{matrix}$$

En el primer nivel, los individuos A y B forman el primer clúster por ser los más cercanos, y como ya hemos seleccionado el clúster, el siguiente paso es calcular la distancia entre el clúster y los demás individuos para formar otra matriz de distancias:

$d_{(A,B),C} \rightarrow \min (d_{AC}, d_{BC}) = d_{BC} = 3$

$d_{(A,B),D} \rightarrow \min (d_{AD}, d_{BD}) = d_{AC} = 6$

$d_{(A,B),E} \rightarrow \min (d_{AE}, d_{BE}) = d_{BE} = 7$

$$\begin{array}{c} \begin{matrix} AB & C & D & E \end{matrix} \\ \begin{pmatrix} 0 & 3 & 6 & 7 \\ 3 & 0 & 4 & 6 \\ 6 & 4 & 0 & 2 \\ 7 & 6 & 2 & 0 \end{pmatrix} \begin{matrix} AB \\ C \\ D \\ E \end{matrix} \end{array}$$

Con esta nueva matriz, la distancia más pequeña es d_{DE}, por tanto, los individuos D y E forman un segundo clúster:

$d_{(A,B),C} \rightarrow \min (d_{AC}, d_{BC}) = d_{BC} = 3$

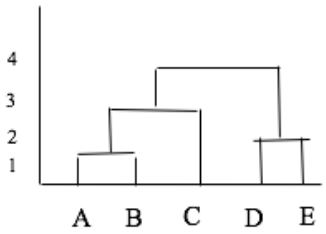
$d_{(A,B),(D,E)} \rightarrow \min (d_{AD}, d_{AE}, d_{BD}, d_{BE}) = d_{AB,D} = 6$

$d_{(D,E),C} \rightarrow \min (d_{DC}, d_{EC}) = d_{DC} = 4$

$$\begin{array}{c} \begin{matrix} AB & C & DE \end{matrix} \\ \begin{pmatrix} 0 & 3 & 6 \\ 3 & 0 & 4 \\ 6 & 4 & 0 \end{pmatrix} \begin{matrix} AB \\ C \\ DE \end{matrix} \end{array}$$

En esta nueva matriz, la distancia más pequeña es $d_{(AB),C}$, por tanto, un individuo C acompaña al primer clúster que contiene a los individuos A y B.

Por tanto el dendograma quedaría así:



- **El vecino más lejano**: Es exactamente igual al método anterior, pero con la diferencia de que en vez de buscar la distancia mínima se busca la máxima con el objetivo de separar a los individuos más heterogéneos.

$d_{(A,B),C} \rightarrow \max(d_{AC}, d_{BC}) = d_{BC} = 5$

$d_{(A,B),D} \rightarrow \max(d_{AD}, d_{BD}) = d_{BC} = 8$

$d_{(A,B),E} \rightarrow \max(d_{AE}, d_{BE}) = d_{AE} = 8$

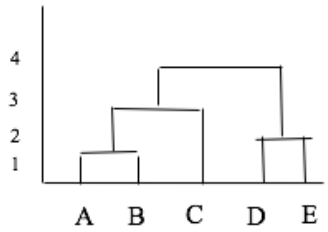
$$\begin{matrix} & \text{AB} & \text{C} & \text{D} & \text{E} \\ \text{AB} & 0 & 5 & 8 & 8 \\ \text{C} & 5 & 0 & 4 & 6 \\ \text{D} & 8 & 4 & 0 & 2 \\ \text{E} & 8 & 6 & 2 & 0 \end{matrix}$$

$d_{(A,B),C} \rightarrow \max(d_{AC}, d_{BC}) = d_{AC} = 5$

$d_{(A,B),(D,E)} \rightarrow \max(d_{AD}, d_{AE}, d_{BD}, d_{BE}) = 8$

$d_{(D,E),C} \rightarrow \max(d_{DC}, d_{EC}) = 6$

$$\begin{matrix} & \text{AB} & \text{C} \\ \text{AB} & 0 & 5 & 8 \\ \text{C} & 5 & 0 & 6 \\ \text{DE} & 8 & 6 & 0 \end{matrix}$$



- **Enlace promedio**: La distancia entre clústers es definida como el promedio de la distancia entre todos los componentes de los objetos. Dentro de los diferentes métodos propuestos para calcular el promedio de las distancias, este es el que vamos a proponer:

$$\frac{1}{n_i n_{j_i}} \sum_i \sum_j dij$$

dij → distancia entre i y j (cada uno pertenece a un clúster distinto)
ni nij →nº objetos en los clústers

Usando el ejemplo anterior tenemos:

$$d_{(AB)C} = \frac{1}{2*1} d_{AC} + d_{BC} = \frac{5+3}{2} = 4$$

$$d_{(AB)D} = \frac{1}{2*1} d_{AD} + d_{BD} = \frac{6+8}{2} = 7$$

$$d_{(AB)E} = \frac{1}{2*1} d_{AE} + d_{BE} = \frac{8+7}{2} = 7{,}5$$

$$\begin{matrix} & \text{AB} & \text{C} & \text{D} \\ \text{AB} & 0 & 4 & 7 & 7'5 \\ \text{C} & 4 & 0 & 4 & 6 \\ \text{D} & 7 & 4 & 0 & 2 \\ & 7{,}5 & 6 & 2 & 0 \end{matrix}$$

El siguiente clúster estará formado por los individuos D y E

$$d_{(AB)C} = \frac{1}{2*1} d_{AC} + d_{BC} = \frac{5+3}{2} = 4$$

$$d_{(AB)(DE)} = \frac{1}{2*2} d_{AD} + d_{AE} + d_{BD} + d_{BE} = \frac{6+8+8+7}{4} = \frac{29}{4}$$

$$d_{(DE)C} = \frac{1}{2*1} d_{DC} + d_{EC} = \frac{4+6}{2} = 5$$

$$\begin{matrix} \text{AB} & \text{C} & \text{DE} \end{matrix}$$
$$\begin{pmatrix} 0 & 4 & 29/4 \\ 4 & 0 & 5 \\ 29/4 & 5 & 0 \end{pmatrix} \begin{matrix} \text{AB} \\ \text{C} \\ \text{DE} \end{matrix}$$

- **Método de Ward**: Este método se basa en la pérdida de información resultante de la agrupación de individuos en un clúster, medida a través de la suma al cuadrado de las desviaciones de cada observación respecto a la medida del clúster a la que pertenece.

  La regla de la agrupación descansa en el aumento de la suma de los errores al cuadrado producido por la combinación de cada posible par de clúster. Este valor llamado EES es utilizado como una función objetivo.

  La **primera etapa** lo que hace es asignar un clúster a cada individuo:

  A= 2 B=5 C=9 D=10 E=15

  En la **segunda etapa** consiste en buscar que combinación de clusters produce el menor incremento en el valor de E.E.S. usando la siguiente fórmula:

$$\sum_{j=1}^{n} (\sum_{i=1}^{nj} x_{ij}^2 - \frac{1}{nj} \left(\sum_{i=1}^{nj} x_{ij} \right)^2)$$

  Luego:

  AB = $2^2 + 5^2 - \frac{1}{2}(2+5)^2$ = 4,5

  AD= $2^2 + 10^2 - \frac{1}{2}(2+10)^2$ = 32

  AC= $2^2 + 9^2 - \frac{1}{2}(2+9)^2$ = 24,5

  AE = $2^2 + 15^2 - \frac{1}{2}(2+15)^2$ = 98

BC = $5^2 + 9^2 - \frac{1}{2}(5+9)^2$ = 8

BD= $5^2 + 10^2 - \frac{1}{2}(5+10)^2$ = 12,5

BE = $5^2 + 15^2 - \frac{1}{2}(5+15)^2$ = 60,5

CD = $\mathbf{9^2 + 10^2 - \frac{1}{2}(9+10)^2}$ = 0,5

CE = $9^2 + 15^2 - \frac{1}{2}(9+15)^2$ = 24,5

DE = $10^2 + 15^2 - \frac{1}{2}(10+15)^2$ = 18

En la **tercera etapa** seguimos intentando hallar la combinación que produce el mínimo incremento de EES, pero teniendo en cuenta el clúster de la segunda etapa:

CDA = $2^2 + 9^2 + 10^2 - \frac{1}{3}(2+9+10)^2$ = 38

CDB = $5^2 + 9^2 + 10^2 - \frac{1}{3}(5+9+10)^2$ = 14

CDE = $15^2 + 9^2 + 10^2 - \frac{1}{3}(15+9+10)^2$ = 28,66

AB = 4,5

AE = 98

BE = 60,5

En la **cuarta etapa** tenemos formados dos grupos de clusters, por lo que para hallar el mínimo de incremento de EES quedaría algo así:

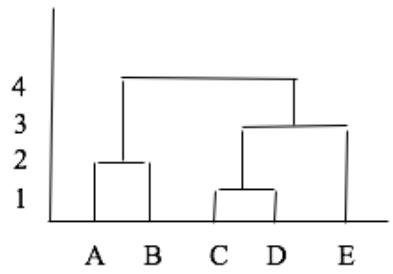
ABCD = $2^2 + 5^2 + 9^2 + 10^2 - \frac{1}{4}(2+5+9+10)^2$ = 41

ABE= $15^2 + 2^2 + 5^2 - \frac{1}{3}(15+2+5)^2$ = 108,66

CDE = 28,66

En este ejemplo, llegamos a la última combinación en la **quinta etapa:**

ABCDE = $2^2 + 5^2 + 15^2 + 9^2 + 10^2 - \frac{1}{5}(2+5+15+9+10)^2$ = 113,2



4.4.2 Métodos no jerárquicos**:**

Estos métodos no requieren que la localización de un objeto en el clúster sea definitiva. Estas técnicas están basadas en la optimización de algún criterio predefinido.

Las distintas técnicas que vamos a ver difieren de:

- cómo comienzan los clústers.
- Cómo se colocan los objetos en los clústers.
- Cómo alguno o todos los objetos son recolocados en los clústers

4.4.2.1 K-Medias:

Tomamos “n” individuos y sobre el analizamos “p” medidas de variable.

xij → valor del individuo i en la variable j

i= 1 ……. n

j= 1…….p

Asumiremos que las medidas analizadas permiten aplicar la distancia euclídea entre individuos.

Sea P(n,k) → la partición que resulta de que cada uno de los individuos sea colocado en uno de los k clúster.

Sea $\bar{x}$(l,j) media de la variable j en el clúster l

N(i) nº de individuos pertenecientes al clúster l

Utilizando esta notación podremos expresar la distancia entre el individuo i y el clúster l como:

$$D(i,l) = \left(\sum_{j=1}^{P} \left[x_{(i,j)} - \bar{X}_{(l,j)} \right]^2 \right)^{1/2}$$

Definiendo por:

$$E\big(P(n,k)\big) = \sum_{i=1}^{n} D[i, l(i)]^2$$

al error de los componentes de partición

l(i) → clúster que contiene al individuo i

D(i,l(i)) → distancia euclídea entre el individuo i y la media de clusters que contiene el individuo

El procedimiento es el siguiente: Buscar para una partición el menor componente de error (Por ejemplo, moviendo los individuos de uno a otro clúster hasta no trasladar un individuo que produzca una reducción en el sumatorio).

Una variedad de sugerencias se ha ofrecido para formar los k primeros puntos usados como estimadores iniciales de los centros de los clusters:

1. elegir los k primeros objetos en la muestra como los primeros k clusters
2. elegir los k objetos mutuamente más alejados
3. elegir los k clúster iniciales basado en un conocimiento "a priori"

Ejemplo:

Nutrientes en los tipos de pecado					
Tipo de pescado	Energía	Grasa	Calcio	Suma (i)	Δi
1	5	9	20	34	1,285
2	6	11	2	19	0,367
3	4	5	20	29	0,979
4	6	9	46	61	2,938
5	5	7	1	13	0
6	3	1	12	16	0,183

1 paso: Vamos a formar 3 grupos.

El criterio de agrupamiento es $\Delta i = K \frac{(suma(i) - Min)}{(Max - Min) + 1}$

Clúster 1 → ⟨2,5,6⟩
Clúster 2 → ⟨1,3⟩
Clúster 3 → ⟨4⟩

2 paso: Calculamos las medias de cada variable para cada clúster.

	Energía	Grasa	Calcio
Clúster 1	(6+5+3)/3 = 14/3	(11+7+1)/3 = 19/3	(2+1+12)/3 = 5
Clúster 2	(5+4)/2 = 9/2	(9+5)/2= 7	(20+20)/2 = 20
Clúster 3	6	9	46

3 paso: Calculamos las distancias de cada individuo a la media de su clúster:

$$D_{(2,1)} = \left[(6 - \frac{14}{3})^2 + (11 - \frac{19}{3})^2 + (2-5)^2\right]^{1/2}$$

Y a continuación calculamos el error de los componentes de la partición:

$E\big(P(6,3)\big) = D(2,1)^2 + D(5,1)^2 + D(6,1)^2 + D(1,2)^2 + D(3,2)^2 + D(4,3)^2$

$E\big(P(6,3)\big) = 137{,}805$

4 paso: Reubicación de objetos

Comprobamos si cualquier movimiento de un objeto a otro clúster produce una reducción en E. Para ello calculamos el siguiente valor:

$$Rl(i), l = \frac{n(i)D(i,l)^2}{n(l)+1} - \frac{n(l(i)D(I,L(i))^2}{n\big(l(I)\big)-1}$$

n(l) es el número de "casos" en el cluster l
l(i) es el clúster que tiene el caso i

Para el primer objeto los cálculos son:

$D(1,1)^2$= (5-14/3)^2 + (9-19/3)^2 + (20-5)^2 = 232, 22

$D(1,2)^2$= (5-9/2)^2 + (9-7)^2 + (20-20)^2 = 4,25

$D(1,3)^2$= (5-6)^2 + (9-9)^2 + (20-46)^2 = 4,25

$$R2(1), 1 = \frac{3}{3+1} * 232{,}22 - \frac{2}{2-1} * 4{,}25 > 0$$

$$R2(1), 3 = \frac{1}{1+1} * 677 - \frac{2}{2-1} * 4{,}25 > 0$$

Vemos que pasar el objeto 1 al clúster 1 o al clúster 3 produce un aumento en E. Esto mismo sucede para los objetos 2,3,4,5. Para el objeto 6

$D(6,1)^2$= (3-14/3)^2 + (1-19/3)^2 + (12-5)^2 = 80,21

$D(6,2)^2$= (3-9/2)^2 + (1-7)^2 + (12-20)^2 = 102,25

$D(6,3)^2$= (3-6)^2 + (1-9)^2 + (12-46)^2 = 1229

$$R1(6), 2 = \frac{2}{2+1} * 102{,}25 - \frac{3}{3-1} * 80{,}21 = -52{,}15$$

$$R1(6),3 = \frac{1}{1+1} * 1229 - \frac{3}{3-1} * 80{,}21 > 0$$

Pasar el objeto 6 del clúster 1 al 2 produce una reducción en E de -52,15

El nuevo error es 137,805 – 52,15 = 85,655

Los clusters resultantes son:

Clúster 1 → ⟨2,5⟩

Clúster 2 → ⟨1,3,6⟩

Clúster 3 → ⟨4⟩

5 paso: Volvemos a calcular las medias de cada variable para cada clúster:

	Energía	Grasa	Calcio
Clúster 1	11/3	9	3/2
Clúster 2	4	5	52/3
Clúster 3	6	9	46

Y repetimos todos los cálculos del 4 paso. Si no produce ninguna reducción en E, estos serán los clusters resultantes.

La inspección de la última tabla nos lleva a la conclusión que el calcio es la variable dominante, y los tres grupos formados se caracterizan por su contenido en calcio: grupo 1 es bajo en calcio, grupo 2 es medio en calcio y grupo 3 es alto en calcio.

4.5 VALIDACIÓN DE LAS SOLUCIONES CLÚSTER

Hay una serie de cuestiones que deben ser tenidas en cuenta ante una solución de clúster dada:

1. cómo difieren los clusters.
2. cuál es el número óptimo de clúster.
3. Cómo de bueno es el ajuste de la solución.

Validar la calidad de la solución clúster es una cuestión particularmente problemática, pero sumamente importante. Es por ello que siempre hay que hacer los siguientes pasos:

a) Comparación entre clusters: Una vez que los objetos han sido agrupados, necesitamos comparar los diferentes clusters para tener una idea de cómo difieren. Una aproximación simple es comparar las medias y varianzas respecto a las p variables utilizadas para calcular la semejanza entre objetos o respecto a un conjunto externo de variables, cuya información está disponible para todos los miembros de los clusters, pero que no fue utilizado en el proceso de agrupamiento.
   Otra posibilidad radica en someter los clusters a un análisis discriminante. Este determina que variables han contribuido de manera más importante a los diferentes perfiles entre los clusters, además de ser una herramienta para predecir los futuros miembros de un clúster a partir de una muestra de objetos.

b) Número óptimo de clúster: El hecho de que se formen muchos o pocos clusters nos puede complicar mucho el análisis. En el caso de haber formado muchos clusters nos encontraremos con el problema de que probablemente no podamos interpretar la información, y en el caso de haber formado muy pocos clusters lo más probable es que tomemos una decisión incorrecta por una pérdida importante de información.
   Alguno de los métodos propuestos por Eduardo Morales y que se usan para encontrar el número adecuado de clusters es la visualización del conjunto de datos (funciona solo en el caso de tener dos dimensiones, ya que se puede

representar fácilmente en unos ejes y ver como se agrupan los datos en clusters) y también se puede calcular a través de establecer unas reglas que se basan en la estadística (por ejemplo, el error cuadrático, propiedades geopmétricas o estadística de los datos, la disimilaridad o similaridad, etc) y finalmente la optimización de alguna función de criterio bajo el modelo de mezcla de probabilidades.

Todos estos modelos llevan una complicada base estadística, por lo que no los explicaré en este trabajo.

c) Medidas de calidad: con estas medidas se estudia que agrupación es mejor. Eduardo Morales las divide en dos tipos de clases: evaluación interna (entre los grupos generados en el cluster) y evaluación externa (entre los grupos conocidos).

Los índices usados son:

- Índice Davies-Bouldin (interna): El algoritmo que nos de el valor menor, será el mejor

$$DB = \frac{1}{n}\sum_{i=1}^{n} max_{j\neq i}\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right)$$

Donde n = número de clusters

C = centroide de cada clúster

σ = distancia promedio de todos los elementos en el clúster.

- Índice Dunn (interna): busca identificar clusters densos y claramente separados. Cuanto mayor sea el valor del índice, mejor.

$$D = \frac{min_{\ 1\leq i<j\leq n}\ d(i.j)}{max_{1\leq k\leq n}\ d'(k)}$$

d(i,j) = distancia dentre los clusters i y j. Pueden usarse los centroides

d'(k) = distancia intra cluster k

- Coeficiente de Silueta (interna): Contrasta la distancia promedio de elementos en el mismo cluster con la distancia promedio de elementos en otros clusters. Elementos con alto valor se consideran bien agrupados, mientras que objetos con medidas bajas se consideran outliers.

- Purity (externa): mide en que medida los clusters contienen una sola clase:

  $$\frac{1}{N}\sum_{m\in M} max_{d\in D}|m \cap d|$$

  M = clusters
  D = clases
  N = datos

- Índice Rand (externa): Mide que tan parecidos son los clusters a las clases:

  $$RI = \frac{TP+TN}{TP+FP+TN+FN}$$

  TP = True Positive
  TN = True Negative
  FP = False Positive
  FN= False negative

- F-Measure (externo): Puede balancear los falsos negativos usando precisión (P) y recuento (R)

  $F_\beta = \frac{(\beta^2+1)*P*R}{\beta^2*P+R}$ → $P = \frac{TP}{TP+FP}$ y $R = \frac{TP}{TP+FN}$

- Índice de Jaccard (externo): mide la similaridad entre dos grupos. Los elementos comunes entre los dos grupos entre los elementos de los dos grupos:

  $$J(A,B) = \frac{|A \cap B|}{\lceil A \cup B \rceil} = \frac{TP}{TP+FP+FN}$$

4.6 RECOMENDACIONES FINALES

1. La mayor parte de técnicas clúster son muy sensibles a la presencia de valores atípicos, por lo que estos datos deben ser eliminados previamente de los análisis para evitar confusiones. Normalmente, son los valores extremos los que se eliminan.

2. La cuestión sobre qué medidas de distancia o similaridad es mejor sigue aun sin ser contestada, aunque la más usada es la distancia euclídea. No hay una distancia mejor que otra, así que en los análisis clúster se podrá usar aquella que resulte más cómoda o adecuada al caso.

3. Las técnicas de optimización siguen teniendo dificultades al utilizarse en muestras grandes por la cantidad de cálculos necesarios (tiempo de computación en el ordenador). La optimización, cuando es en más de dos dimensiones (y esto sucede en la gran mayoría de los casos), se complica bastante, incluso teniendo a nuestra disposición programas informáticos avanzados.

4. Siempre que sea posible los datos deben ser tipificados para asegurar la estabilidad de las soluciones clúster.

5 EJEMPLO DE CLÚSTER

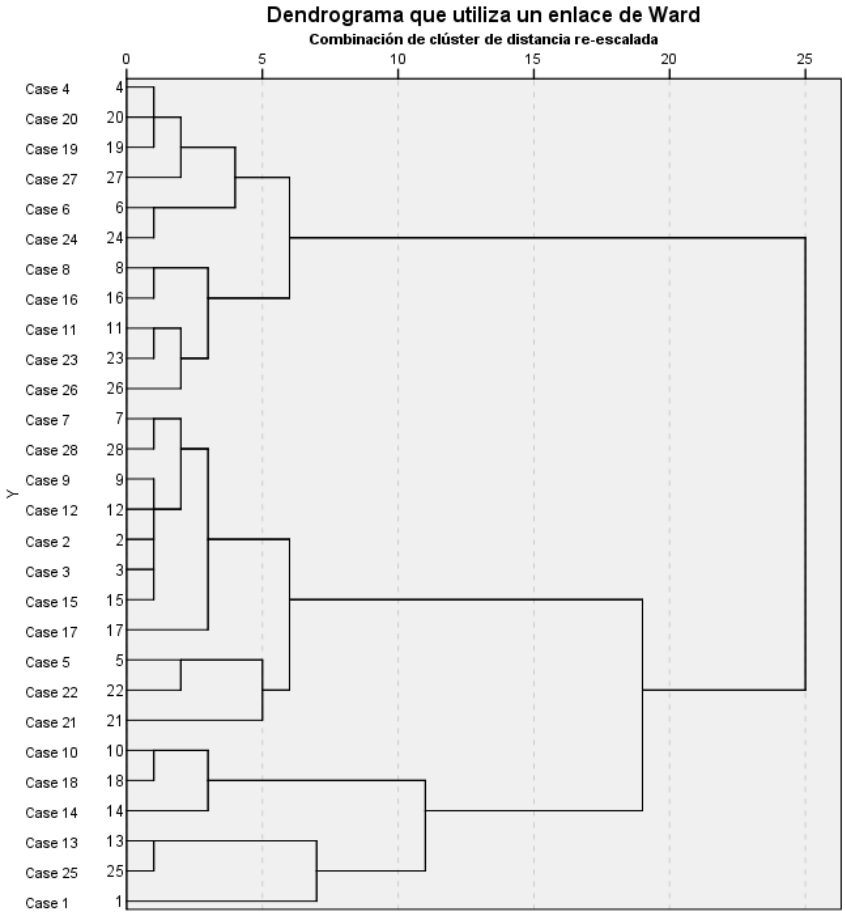
En el caso del big data, los clusters se suelen hacer con programas informáticos por la complejidad que tiene trabajar con grandes cantidades de datos. Antes hemos visto paso a paso como se forma un clúster con poca información.
Cuando se hacen clúster con grandes volúmenes de información en los ordenadores, en el fondo se están haciendo los mismos pasos y las mismas operaciones que si hacemos un clúster con pocos datos a mano. De hecho, se podría decir que las matemáticas que se usan actualmente en los ordenadores para resolver clusters no es nada novedoso. Si hasta ahora no se podrían hacer clusters con tantos datos era porque hacerlos llevaba mucho tiempo y porque no se disponía ni de memoria suficiente para almacenar los datos, ni de programas informáticos lo suficientemente potentes como para resolver de manera inmediata el clúster.

Para este trabajo, he escogido variables macroeconómicas de los países de la Unión Europea con el objetivo de agrupar a estos países en clúster según su similitud. Las variables escogidas fueron: PIB, deuda total, deuda exterior/PIB, gasto educación, gasto salud, tasa de desempleo, salario medio, vehículos cada 1000 habitantes, balanza comercial, tasa de inmigrantes, tasa de emigrantes, tasa de natalidad, tasa de mortalidad, porcentaje de población en riesgo de pobreza, esperanza de vida y número de habitantes.

El programa usado para el análisis clúster fue el SPSS por su simplicidad. Hay otros programas más complejos y especializados en este tipo de algoritmos, pero al no saber usarlos, me decante por SPSS, programa que ya había usado anteriormente en alguna asignatura.

Como bien he explicado antes, gracias a estos programas, podemos formar clusters de manera inmediata. Simplemente tenias que introducir los datos para formar el clúster y luego interpretarlo y validarlo.

Al hacer el clúster con el método Ward, el dendograma que nos salió fue el siguiente:



Esto nos dice que los países se agrupan en tres clusters:

- Grupo 1: España (10), Italia (18), Grecia (14), Francia (13), Reino Unido (25) y Alemania (1).
- Grupo 2: Dinamarca (7), Suecia (28), Eslovenia (9), Finlandia (12), Austria (2), Bélgica (3), Holanda (15), Irlanda (17), Chipre (5), Malta (22) y Luxemburgo (21)
- Grupo 3: Bulgaria (4), Lituania (20), Letonia (19), Rumania (27), Croacia (6), Portugal (24), Eslovaquia (8), Hungría (16), Estonia (11), Polonia (23) y la República Checa (26).

Aunque en un primer momento nos puede resultar un poco chocante este resultado de formación de clúster, tenemos que tener en cuenta que el clúster se va formando por etapas. En la primera etapa se juntan los países que son muy similares entre ellos, y luego, conforme van pasando las etapas, vemos que son menos exigentes con la similitud entre ellos.
Por ejemplo, en el grupo 1, en un primer momento se unen España e Italia por un lado, y Francia y Reino Unido por otro. Más tarde se incorpora Grecia a los países mediterráneos (formando ahora un grupo de 3) y Alemania se incorpora en la última etapa a Francia y Reino Unido. Finalmente, se agrupan los 6 países en un clúster debido a las similitudes que comentaré a continuación.

Si nos fijamos en la tabla del anexo, podemos ver que los países del grupo 1 son los que presentan un mayor PIB y que los países del grupo 3 son los que tienen un PIB menor. Además, también podemos ver que el porcentaje de deuda/PIB más bajo lo tienen los países del grupo 3, siendo el grupo 1 los que presentan un porcentaje mayor en esta variable (lo cual tiene sentido porque son países que se endeudan para poder prestar servicios sociales y estimular la economía). Sin embargo, si nos fijamos bien, es el grupo 2 el que más dinero se gasta en educación y en salud.
La balanza comercial es positiva en los países del grupo 2 y negativa en los del grupo 1 y 3. Esto se debe a que los países del grupo 2 son más exportadores que importadores.
En cuanto a la tasa de desempleo, vemos que el grupo que de media tiene más desempleo es el grupo 1 (con un 9% de media), y que los grupos 2 y 3 tienen niveles bastante parecidos (un 5%, que es básicamente el desempleo friccional o voluntario). Los salarios más altos los encontramos en el grupo 2 y los más bajos en el grupo 3. Esto puede que esté relacionado con que en el grupo 1, el 60% de la población pertenece a la clase media, en el grupo dos el 50%, y en el grupo 3 el 40%; y además puede que por ello las tasas de inmigración más altas estén en el grupo 1 y las de emigración en el grupo 3.
La tasa de natalidad es parecida en los países del grupo 1 y 3, pero la más alta está en el grupo 2, y en parte se debe a las ayudas y subvenciones que da el estado. Además, también coincide que en los países del grupo 1 y 3 es donde

se presenta la mayor tasa de mortalidad, siendo esta mayor que la de natalidad, lo que va a suponer un problema demográfico en un futuro cercano.

Un dato que he encontrado sorprendente en este estudio es que los países del grupo 1 y 3 tienen de media casi el mismo porcentaje de su población en riesgo de pobreza. Antes de ver los resultados nunca hubiese pensado que países como Italia o España (en teoría países desarrollados, con un sistema de bienestar y con ayudas) tuviésemos un riesgo de exclusión tan alto. Pero al investigar un poco más me he dado cuenta de que los datos recogidos son correctos. Los inmigrantes y la gente joven con un colectivo vulnerable en nuestro país.

Finalmente, ya solo me queda comentar que a la hora de formar los clusters, ha habido variables que no han tenido ninguna relevancia (balanza comercial y la tasa de natalidad). Esto se puede ver interpretando la segunda tabla del anexo. Además, a la hora de comparar grupos podemos ver que los países del grupo 3 y 2 tienen un PIB, una deuda, una tasa de desempleo y un numero de habitantes bastante parecido, y que los paises del grupo 1 y 2 tienen un gasto en salud, unos salarios, una tasa de mortalidad y una esperanza de vida que son parecidos a niveles significativos. El grupo 1 y 3 tienen un gasto en educación semejante.

6 Conclusiones

El big data es un tema muy amplio, tanto que no me extraña que se estén creando grados y postgrados relacionados con las grandes cantidades de datos. A lo que quiero llegar es que yo en este trabajo he pretendido comentar de manera general un poco sobre este tema, pero se que me he dejado muchas disciplinas sin comentar, sobre todo, aquellas que están relacionadas más con temas informáticos. Simplemente quería aprender un poco sobre un tema que me llamó la atención.

Decidí empezar a hacer el trabajo yendo desde lo más general, describiendo qué es el Big Data y sus posibles usos, para luego ir concretando cada vez un poco más en aquellos temas que más me interesaban.

De la primera parte, lo que me llevo es toda la información sobre el Big Data que he buscado y aprendido. El leer distintos libros e informes de este tema me ha gustado mucho, en especial ver todas las aplicaciones y todos los mitos alrededor de este tema.
Si nos ponemos a pensar, el mundo en el que vivimos es totalmente distinto al mundo de hace 40 años. Esto se debe a todos los avances tecnológicos que se han dado en las últimas décadas y años, y si hay algo que tengo claro, es que el mundo de hoy en día cambiará en los próximos años y esto va a ser gracias a los datos. Por ejemplo, gracias a los datos podremos montarnos en coches sin conductor, los trabajos serán distintos (se destruirán algunos trabajos, pero se crearán nuevos trabajos que hoy todavía no existen), habrá nuevas carreras (esto ya está pasando), nuevas formas de comercio, etc. Los datos, sin lugar a duda, van a ser el futuro

Con la segunda parte, que es el Machine Learning, me di cuenta de que puede ser muy compleja, y que aún hay mucho que profundizar, Si tuviese algunos conocimientos de informática, creo que es un tema que habría disfrutado incluso aún más.

En cuanto a la tercera parte, que es el clúster, he de comentar que fue un tema que al principio me asustaba porque pensé que iba a ser un algoritmo mucho más complicado, pero gracias a las múltiples tutorías que tuve, he logrado entender la base estadística que hay detrás, y he de decir que he disfrutado un mucho aprendiendo sobre este clúster.

También me gustaría agradecer al profesor Tomás Curto González por ayudarme con el algoritmo del clúster. Sin él, este trabajo habría tenido un enfoque totalmente distinto.

En cuanto al ejemplo del clúster de los países de la Unión Europea, soy plenamente consciente de que no he trabajado con grandes volúmenes de datos, pero el objetivo que tenia era demostrar que en el momento en el que tenemos un número importante de datos, es prácticamente imposible hacer el clúster a mano. Como bien dije, es mucho más cómodo usar programas que en el momento te calculan el algoritmo. Si hasta hace relativamente poco no se han podido usar es porque los ordenadores no eran potentes y no tenían memoria.

7 BIBLIOGRAFÍA:

Accenture.com. (n.d.). Gran éxito con Big Data [online] Available at: https://www.accenture.com/_acnmedia/Accenture/Conversion-Assets/DotCom/Documents/Local/es-la/PDF2/Accenture-Big-Data-POV-espanol.pdf [Accessed 21 Mar. 2019].

Afuente, A. (2018). Reducción de la dimensionalidad (o por qué más datos no siempre es mejor) - Aukera. [online] Aukera. Available at: https://aukera.es/blog/reduccion-dimensionalidad/ [Accessed 28 Jan. 2019].

Aluja, T. (n.d.). La minería de datos, entre la estadística y la Inteligencia Artificial. Universitat Politécnica de Catalunya.

Brynjolfsson, E. and McAfee, A. (2012). Big Data: The Management Revolution. [online] Harvard Business Review. Available at: https://hbr.org/2012/10/big-data-the-management-revolution [Accessed 21 Mar. 2019].

Caparrini, F. and Work, W. (2017). Sistemas Basados en Reglas - Fernando Sancho Caparrini. [online] Cs.us.es. Available at: http://www.cs.us.es/~fsancho/?e=103 [Accessed 29 Jan. 2019].

Comisión Europea - European Commission. (2019). Reforma de 2018 de las normas de protección de datos de la UE. [online] Available at: https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_es [Accessed 21 Mar. 2019].

Data.bsa.org. (2015). ¿Por qué son tan importantes los datos? [online] Available at: https://data.bsa.org/wp-content/uploads/2015/10/BSADataStudy_es.pdf [Accessed 15 Feb. 2019].

datosmacro.com. (2019). UE - Unión Europea 2019. [online] Available at: https://datosmacro.expansion.com/paises/grupos/union-europea [Accessed 21 Mar. 2019].

Docs.microsoft.com. (n.d.). Cómo elegir algoritmos - Azure Machine Learning Studio. [online] Available at: https://docs.microsoft.com/es-es/azure/machine-learning/studio/algorithm-choice [Accessed 21 Apr. 2019].

Enciclopedia.banrepcultural.org. (n.d.). Las revoluciones industriales - Enciclopedia | Banrepcultural. [online] Available at: http://enciclopedia.banrepcultural.org/index.php/Las_revoluciones_industriales [Accessed 15 Feb. 2019].

EY (2014). Big Data en el sector financiero español. [online] Available at: https://www.ey.com/es/es/home/ey-informe-sobre-big-data-y-analytics-en-el-sector-financiero-espanol#.XGaJmCNDnos [Accessed 15 Feb. 2019].

Gupta, P. (2017). Decision Trees in Machine Learning – Towards Data Science. [online] Towards Data Science. Available at: https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052 [Accessed 28 Jan. 2019].

Joyanes Aguilar, L. (2014). Big data. 1st ed. Barcelona: Marcombo.

Morales, E. and Escalante, H. (n.d.). Clustering [online] Ccc.inaoep.mx. Available at: https://ccc.inaoep.mx/~emorales/Cursos/NvoAprend/Acetatos/clustering.pdf [Accessed 11 Feb. 2019].

Mueller, J. and Massaron, L. (2016). Machine Learning for dummies.

Perasso, V. (2016). Qué es la cuarta revolución industrial (y por qué debería preocuparnos). [online] BBC News Mundo. Available at: https://www.bbc.com/mundo/noticias-37631834 [Accessed 15 Feb. 2019].

Priy, S. (2017). Clustering in Machine Learning. [online] GeeksforGeeks. Available at: https://www.geeksforgeeks.org/clustering-in-machine-learning/ [Accessed 28 Jan. 2019].

Profesor Tomás Curto Gonzáles, Apuntes sobre el clúster. Icade

Shalev-Shwartz, S. and Ben-David, S. (2014). Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press.

Siegler, M. (2010). Eric Schmidt: Every 2 Days We Create As Much Information As We Did Up To 2003. [online] TechCrunch. Available at: https://techcrunch.com/2010/08/04/schmidt-data/?guccounter=1 [Accessed 5 Feb. 2019].

Zambrano, J. (2018). ¿Aprendizaje supervisado o no supervisado? Conoce sus diferencias dentro del machine learning y la automatización inteligente. [online] Medium. Available at: https://medium.com/@juanzambrano/aprendizaje-supervisado-o-no-supervisado-39ccf1fd6e7b [Accessed 10 Jan. 2019].

8 Anexo:

Tabla 1:

Comparaciones múltiples

HSD Tukey

Variable dependiente	(I) Ward Method	(J) Ward Method	Diferencia de medias (I-J)	Error estándar	Sig.	Intervalo de confianza al 95% Límite inferior	Intervalo de confianza al 95% Límite superior
PIB2018	1	2	1603566,27300*	265645,40690	,000	941888,6914	2265243,8540
		3	1738750,72700*	265645,40690	,000	1077073,1460	2400428,3090
	2	1	-1603566,27300*	265645,40690	,000	-2265243,8540	-941888,6914
		3	135184,45450	223186,78360	,818	-420735,9090	691104,8181
	3	1	-1738750,72700*	265645,40690	,000	-2400428,3090	-1077073,1460
		2	-135184,45450	223186,78360	,818	-691104,8181	420735,9090
Deudatotal2017	1	2	1511939,68200*	188041,38900	,000	1043560,5260	1980318,8380
		3	1605917,77300*	188041,38900	,000	1137538,6170	2074296,9290
	2	1	-1511939,68200*	188041,38900	,000	-1980318,8380	-1043560,5260
		3	93978,09091	157986,36720	,824	-299539,1114	487495,2932
	3	1	-1605917,77300*	188041,38900	,000	-2074296,9290	-1137538,6170
		2	-93978,09091	157986,36720	,824	-487495,2932	299539,1114

Deuda exterior2017entrePIB	1	2	,46527*	,15624	,017	,0761	,8544
		3	,58236*	,15624	,003	,1932	,9715
	2	1	-,46527*	,15624	,017	-,8544	-,0761
		3	,11709	,13127	,650	-,2099	,4441
	3	1	-,58236*	,15624	,003	-,9715	-,1932
		2	-,11709	,13127	,650	-,4441	,2099
Gasto_educación2015	1	2	-,02525*	,00988	,043	-,0499	-,0006
		3	-,01131	,00988	,496	-,0359	,0133
	2	1	,02525*	,00988	,043	,0006	,0499
		3	,01395	,00830	,232	-,0067	,0346
	3	1	,01131	,00988	,496	-,0133	,0359
		2	-,01395	,00830	,232	-,0346	,0067
Gasto_salud2016	1	2	,00903	,01536	,828	-,0292	,0473
		3	,03934*	,01536	,043	,0011	,0776
	2	1	-,00903	,01536	,828	-,0473	,0292
		3	,03031	,01291	,067	-,0018	,0625
	3	1	-,03934*	,01536	,043	-,0776	-,0011
		2	-,03031	,01291	,067	-,0625	,0018
Tasadesempleo_enero_2019	1	2	,04439*	,01466	,015	,0079	,0809
		3	,04648*	,01466	,011	,0100	,0830
	2	1	-,04439*	,01466	,015	-,0809	-,0079
		3	,00209	,01232	,984	-,0286	,0328
	3	1	-,04648*	,01466	,011	-,0830	-,0100
		2	-,00209	,01232	,984	-,0328	,0286
Salario medio 2017	1	2	-5568,72727	5250,53866	,547	-18646,9271	7509,4726
		3	23587,00000*	5250,53866	,000	10508,8002	36665,1999
	2	1	5568,72727	5250,53866	,547	-7509,4726	18646,9271
		3	29155,72727*	4411,33483	,000	18167,8416	40143,6129
	3	1	-23587,00000*	5250,53866	,000	-36665,1999	-10508,8002
		2	-29155,72727*	4411,33483	,000	-40143,6129	-18167,8416
Vehículos_cada_1000_habitantes_2015	1	2	14,86591	45,88776	,944	-99,4327	129,1645
		3	125,96864*	45,88776	,029	11,6700	240,2673
	2	1	-14,86591	45,88776	,944	-129,1645	99,4327
		3	111,10273*	38,55343	,021	15,0727	207,1328

	3	1	-125,96864*	45,88776	,029	-240,2673	-11,6700
		2	-111,10273*	38,55343	,021	-207,1328	-15,0727
Balanza_comercial_2018	1	2	-13796,30303	31462,74519	,900	-92164,6565	64572,0504
		3	-318,55758	31462,74519	1,000	-78686,9110	78049,7959
	2	1	13796,30303	31462,74519	,900	-64572,0504	92164,6565
		3	13477,74545	26433,99328	,867	-52364,8413	79320,3322
	3	1	318,55758	31462,74519	1,000	-78049,7959	78686,9110
		2	-13477,74545	26433,99328	,867	-79320,3322	52364,8413
Tasa_inmigrantes_2017	1	2	-,04138	,03662	,505	-,1326	,0498
		3	,05647	,03662	,289	-,0347	,1477
	2	1	,04138	,03662	,505	-,0498	,1326
		3	,09785*	,03076	,010	,0212	,1745
	3	1	-,05647	,03662	,289	-,1477	,0347
		2	-,09785*	,03076	,010	-,1745	-,0212
Tasa_emigrantes_2017	1	2	-,04210	,02900	,331	-,1143	,0301
		3	-,10170*	,02900	,005	-,1739	-,0295
	2	1	,04210	,02900	,331	-,0301	,1143
		3	-,05961	,02436	,055	-,1203	,0011
	3	1	,10170*	,02900	,005	,0295	,1739
		2	,05961	,02436	,055	-,0011	,1203
Tasa_natalidad_2017_por_mil	1	2	-,99985	,57671	,213	-2,4363	,4366
		3	-,49985	,57671	,666	-1,9363	,9366
	2	1	,99985	,57671	,213	-,4366	2,4363
		3	,50000	,48453	,564	-,7069	1,7069
	3	1	,49985	,57671	,666	-,9366	1,9363
		2	-,50000	,48453	,564	-1,7069	,7069
Tasa_mortalidad_2017	1	2	1,60864	,79416	,127	-,3695	3,5867
		3	-2,36409*	,79416	,017	-4,3422	-,3860
	2	1	-1,60864	,79416	,127	-3,5867	,3695
		3	-3,97273*	,66722	,000	-5,6347	-2,3108
	3	1	2,36409*	,79416	,017	,3860	4,3422
		2	3,97273*	,66722	,000	2,3108	5,6347
	1	2	,03962	,01809	,093	-,0054	,0847

Porcentaje_riesgo_pobreza_2015		3	-,00538	,01809	,953	-,0504	,0397
	2	1	-,03962	,01809	,093	-,0847	,0054
		3	-,04500*	,01520	,018	-,0829	-,0071
	3	1	,00538	,01809	,953	-,0397	,0504
		2	,04500*	,01520	,018	,0071	,0829
Esperanza_de_vida_2017	1	2	,30303	,72314	,908	-1,4982	2,1042
		3	4,98485*	,72314	,000	3,1836	6,7861
	2	1	-,30303	,72314	,908	-2,1042	1,4982
		3	4,68182*	,60756	,000	3,1685	6,1951
	3	1	-4,98485*	,72314	,000	-6,7861	-3,1836
		2	-4,68182*	,60756	,000	-6,1951	-3,1685
Numero_Habitantes_2018	1	2	49558421,32000*	6822715,70800	,000	32564195,5200	66552647,1200
		3	45634533,23000*	6822715,70800	,000	28640307,4300	62628759,0300
	2	1	-49558421,32000*	6822715,70800	,000	-66552647,1200	-32564195,5200
		3	-3923888,09100	5732227,75500	,775	-18201893,6500	10354117,4700
	3	1	-45634533,23000*	6822715,70800	,000	-62628759,0300	-28640307,4300
		2	3923888,09100	5732227,75500	,775	-10354117,4700	18201893,6500

*. La diferencia de medias es significativa en el nivel 0.05.

Tabla 2:

		Suma de cuadrados	gl	Media cuadrática	F	Sig.
PIB2018	Entre grupos	1326643077000000,000	2	663321538400000,000	24,212	,000
	Dentro de grupos	684919680400000,000	25	273967872100000,000		
	Total	2011562757000000,000	27			

Deudatotal2017	Entre grupos	1150550958 0000,000	2	5752754791 000,000	41,906	,000
	Dentro de grupos	3431957681 000,000	25	1372783072 00,000		
	Total	1493746726 0000,000	27			
Deuda exterior2017entrePIB	Entre grupos	1,369	2	,684	7,222	,003
	Dentro de grupos	2,369	25	,095		
	Total	3,738	27			
Gasto_educación2015	Entre grupos	,003	2	,001	3,490	,046
	Dentro de grupos	,009	25	,000		
	Total	,012	27			
Gasto_salud2016	Entre grupos	,008	2	,004	4,261	,026
	Dentro de grupos	,023	25	,001		
	Total	,031	27			
Tasadesempleo_enero_2019	Entre grupos	,010	2	,005	5,844	,008
	Dentro de grupos	,021	25	,001		
	Total	,031	27			
Salario medio 2017	Entre grupos	5057943202, 000	2	2528971601, 000	23,629	,000
	Dentro de grupos	2675732808, 000	25	107029312,3 00		
	Total	7733676011, 000	27			
Vehículos_cada_1000_habitantes_2015	Entre grupos	91267,209	2	45633,604	5,582	,010
	Dentro de grupos	204375,481	25	8175,019		
	Total	295642,689	27			
Balanza_comercial_2018	Entre grupos	1233878873, 000	2	616939436,4 00	,161	,853
	Dentro de grupos	9607895013 0,000	25	3843158005, 000		
	Total	9731282900 0,000	27			
Tasa_inmigrantes_2017	Entre grupos	,053	2	,026	5,084	,014
	Dentro de grupos	,130	25	,005		
	Total	,183	27			

Tasa_emigrantes_2017	Entre grupos	,044	2	,022	6,725	,005
	Dentro de grupos	,082	25	,003		
	Total	,126	27			
Tasa_natalidad_2017_por_mil	Entre grupos	4,026	2	2,013	1,559	,230
	Dentro de grupos	32,281	25	1,291		
	Total	36,307	27			
Tasa_mortalidad_2017	Entre grupos	87,477	2	43,738	17,863	,000
	Dentro de grupos	61,213	25	2,449		
	Total	148,690	27			
Porcentaje_riesgo_pobreza_2015	Entre grupos	,013	2	,006	4,926	,016
	Dentro de grupos	,032	25	,001		
	Total	,044	27			
Esperanza_de_vida_2017	Entre grupos	153,512	2	76,756	37,807	,000
	Dentro de grupos	50,755	25	2,030		
	Total	204,267	27			
Numero_Habitantes_2018	Entre grupos	1076454200 0000000,000	2	5382270998 000000,000	29,782	,000
	Dentro de grupos	4518034818 000000,000	25	1807213927 00000,000		
	Total	1528257681 0000000,000	27			