Facultad Ciencias Económicas y Empresariales

TRABAJO FIN DE CARRERA
Big Data. Técnicas de machine learning para la creación de modelos predictivos para empresas.

Autor: Alfonso Centeno Martín-Romero
Directora: María Jesús Giménez Abad

MADRID | Abril 2020

Resumen:

Este trabajo de fin de grado empieza describiendo qué es el Big Data para adentrarse en el Análisis Predictivo y en las distintas técnicas que nos ofrece el Machine Learning para crear modelos predictivos a partir de grandes volúmenes de datos. En este trabajo aparecerán explicadas las principales aplicaciones y funciones del Análisis Predictivo para el mundo empresarial, además de un resumen de los algoritmos más empleados por el Machine Learning para elaborar modelos predictivos. Finalmente, he decidido escoger un método de Machine Learning para desarrollar un ejemplo con datos reales a través de la herramienta RStudio. El método escogido es el método clúster o de agrupación.

Palabras Clave: Big Data, Análisis Predictivo, Machine Learning, Algoritmos, Clúster.

Abstract:

This dissertation begins by describing what Big Data is, and then moves on to Predictive Analysis and the various techniques offered by Machine Learning to create predictive models from large volumes of data. I n this paper, the main applications and functions of Predictive Analysis for the business world will be explained, as well as a summary of the most commonly used algorithms by Machine Learning to build predictive models. Finally, I have decided to choose a method of Machine Learning to develop an example with real data through the tool RStudio. The method I have selected is the cluster method.

Key Words: Big Data, Predictive Analysis, Machine Learning, Algorithms, Cluster.

ÍNDICE

1. INTRODUCCIÓN

Justificación y motivaciones: La razón por la que escogí este tema se debe al interés que siempre suscitó en mi el mundo del Big Data, IoT e Inteligencia Artificial. Desde que escuché hablar de ello supe que el futuro del mercado laboral se encontraba en estas disciplinas, aquel que supiera gestionar e interpretar los grandes volúmenes de datos e información que recibe de sus clientes, para sacarles valor, tendría una ventaja competitiva abismal frente a otras empresas. Sin embargo, mi decisión no fue final hasta que me decidí a participar en un Programa Ejecutivo en Business Analytics ofrecido por la Universidad Pontificia de Comillas ICADE, donde aprendí técnicas de Machine Learning y de Análisis Exploratorio de Datos utilizando la herramienta R. Así, una vez construida la base y teniendo en cuenta que el tema ya me interesaba de antes decidí aprovechar mi trabajo de fin de grado para seguir indagando y aprendiendo todo lo posible sobre una disciplina que considero será muy beneficiosa para mi futuro profesional.

Objetivos: Indagar en el análisis de grandes volúmenes de datos para aprender sobre una de sus aplicaciones más empleadas en el mundo empresarial, el Análisis Predictivo, y a su vez aprender sobre el Machine Learning, uno de los métodos que más se aplica para la creación de modelos predictivos, y los algoritmos que utiliza poniendo a prueba dos de ellos con datos reales.

Metodología: La metodología a seguir ha sido la lectura de libros, artículos académicos y manuales especializados en el uso de la herramienta R, herramienta con la que se ha llevado a cabo el análisis de los métodos clúster y de reglas de asociación y su posterior interpretación.

2. BIG DATA

2.1. ¿QUÉ ES?

Big Data es un término que se emplea para describir el gran volumen de datos que inunda los negocios, salud pública, economía y en general los distintos aspectos que componen la sociedad. Por ello, cuando hablamos de Big Data nos referimos a "conjuntos de datos o combinaciones de conjuntos de datos cuyo tamaño (volumen), complejidad (variedad) y ritmo de crecimiento (velocidad) dificultan su captura, gestión, análisis y procesamiento mediante tecnologías y herramientas convencionales". (powerdata.es, n.d.). No obstante, el tamaño para determinar si un conjunto de datos se considera Big Data no está definido y puede evolucionar con el tiempo. Actualmente, el consenso se encuentra en conjuntos de datos que oscilan entre los 30-50 Terabytes. (powerdata.es, n.d.).

En cuanto a sus características, el consenso popular establece que los datos deben cumplir las siguientes propiedades para ser considerados como Big Data:

- Volumen: Se refiere a la cantidad de datos generados y disponibles a cada segundo, minuto, hora y día en nuestro entorno. La cantidad de información que generamos es masiva y no para de aumentar. Por ello, a medida que las bases de datos crecen las aplicaciones y la arquitectura construida para recoger y almacenar esos datos, también lo ha de hacer.

- Velocidad: Se refiere a la rapidez con la que los datos son creados, almacenados y procesados en tiempo real. En algunos procesos como la detección de fraude en una transacción bancaria o la monitorización de un evento en redes sociales, el tiempo resulta fundamental. Sí estos datos no son recibidos, almacenados y estudiados en tiempo real, corren el riesgo de que quedar obsoletos y de perder toda su utilidad.

- Variedad: Los datos que las empresas reciben son diversos. Pueden proceder de varias fuentes y se encuentran en distintos formatos. Por lo tanto, las empresas deben integrar tecnologías y aplicaciones que les permitan organizar, procesar e integrar datos de diferentes fuentes de información para que resulten útiles y para que se puedan sacar conclusiones efectivas o identificar patrones.

- Veracidad: Se refiere a la fiabilidad de la información recogida. la calidad de los datos obtenidos es fundamental para alcanzar conclusiones efectivas e incluso una ventaja competitiva. De nuevo, las empresas tienen que invertir en aplicaciones que sean capaces de identificar y eliminar datos imprevisibles o que causen incertidumbre.

Así, el Big Data permite a las empresas analizar grandes cantidades de conjuntos de datos y con ello obtener respuestas a preguntas, identificar problemas y mejorar procesos. La recopilación de grandes cantidades de datos y la búsqueda de tendencias dentro de los datos permiten a las empresas identificar oportunidades y crear valor de múltiples formas: reduciendo costes, diseñando nuevos productos y servicios, mejorando la eficiencia en la toma de decisiones, creando publicidad más dirigida a las necesidades del cliente etc.

2.1.1. Tipos de Datos

Según su procedencia y empleando la división propuesta por IBM, encontramos:

- Transacciones de datos: Todos los registros de facturación, llamadas y mensajes que una operadora puede registrar, telecomunicaciones, uso de tarjeta de crédito o débito etc. Estos datos se consideran Big Data debido a la cantidad de volumen que se gestiona.

- Redes sociales y páginas web: Hace referencia a toda aquella información que se genera a través de navegar y realizar transacciones en la red. Con esta información una empresa es capaz de conocer las preferencias y gustos de los consumidores.

- Machine-to-Machine: Se refiere a todas aquellas tecnologías que se conectan a otros dispositivos a modo de sensor para recoger grandes volúmenes de datos. Estos sensores pueden instalarse en todo tipo de dispositivos: teléfonos, medios de transporte, etiquetas, termómetros, parquímetros, sistemas de riego automático o incluso los contadores de electricidad en las viviendas, son algunos ejemplos.

- Biométricas: La información biométrica incluye el escaneo de huellas digitales, retina, reconocimiento facial y genético y en definitiva todos aquellos datos que faciliten el reconocimiento inequívoco de personas basado en rasgos físicos y de conducta.

- Generadas por los seres humanos: Hace referencia a toda la información que los humanos generan día a día con sus acciones cotidianas: llamar por teléfono, escribir un correo electrónico o un mensaje de texto, mandar notas de voz etc.

Según su estructura o formato:

- Datos estructurados: Datos perfectamente ordenados, etiquetados e identificados, dotados de una clara definición de longitud y formato que permite almacenarlos de forma específica en tablas y procesarlos fácilmente. (Instituto Europeo de Posgrado, n.d.).

- Datos no estructurados: Datos desorganizados, que carecen de una definición de longitud, formato y estructura interna identificable que permita almacenarlos de forma específica. Suelen ser datos en su forma original, es decir nada más ser recogidos. Estos datos no tendrán valor hasta que no se conviertan en datos estructurados, es decir hasta que no sean categorizados y organizados. Algunos ejemplos son: audios, imágenes, hojas de cálculo, publicaciones en redes sociales, correos electrónicos, archivos PDF o de procesador de texto etc. (Instituto Europeo de Posgrado, n.d.).

- Datos semi-estructurados: Datos que no están perfectamente estructurados, pero sí tienen una organización definida. Algunos ejemplos son: el HTML o lenguaje para la elaboración de páginas web. (Instituto Europeo de Posgrado, n.d.).

2.1.2. Ciclo de Gestión de los Datos

Obtención de la información: ¿Dónde está la información que necesitamos y cómo podemos obtenerla? Como se ha mencionado actualmente, la cantidad de datos es masiva y pueden proceder de varías fuentes, por lo que se deberá hacer una distinción entre

aquella información útil y aquella desechable. Técnicas como el Web Scraping o el *Systems Network Architecture* (SNA) ayudan a extraer información de páginas web, además el uso de API u otros servicios ofrecen una enorme versatilidad para la integración y recopilación rápida de grandes volúmenes de datos.

Almacenamiento: Una vez recogida la información será necesario almacenarla. Los datos se preparan y se registran de manera organizada para su posterior análisis. Las hojas de cálculo o los sistemas NoSQL, son algunos ejemplos de métodos que permiten almacenar información de forma cómoda y flexible.

Análisis: El análisis de los datos tiene como propósito principal extraer conocimientos, establecer patrones, encontrar correlaciones desconocidas u obtener cualquier otra visión que pueda resultar útil de los datos recogidos y almacenados. En otras palabras, el análisis es la puesta en valor de los datos, que por sí solos carecen de valor. Al analizar los datos la empresa es capaz de construir conocimiento en múltiples ámbitos que abarcan casi todos los campos imaginables.

Decisión: Tras haber recogido, almacenado e interpretado la información, será necesario tomar una decisión. Es muy importante que esta decisión se tome a tiempo, ya que de no hacerse a tiempo aumentará la probabilidad de que los datos analizados queden obsoletos.

2.1.3. Seguridad y Anonimidad de los Datos

En palabras de la agencia española de protección de datos: "El tratamiento masivo de datos procedentes de los ciudadanos mediante el uso de técnicas basadas en Big Data, Inteligencia Artificial o Machine Learning obliga a la implementación de garantías o mecanismos para preservar la privacidad y el derecho a la protección de datos de carácter personal, entre ellas las basadas en la "anonimización" de los datos."

La digitalización ha permitido una mayor centralización, control y accesibilidad de los datos. Sin embargo, está mayor accesibilidad igualmente pone en riesgo la privacidad y protección de los datos personales de los ciudadanos. Para proteger los datos personales, se deberá anonimizar los datos, no obstante, para que estos datos sean verdaderamente anónimos, la "anonimización" deberá ser irreversible.

Atendiendo de nuevo a la Agencia Española de Protección de Datos, los datos personales pueden clasificarse como:

- Identificadores: datos que identifican unívocamente a los sujetos de los datos. Los procesos básicos de "anonimización" son capaces de disociar estos datos.
- Cuasi-Identificadores: datos que de manera aislada no identifican a un sujeto, pero si son convenientemente agrupados y cruzados con otras fuentes de información, pueden llegar a identificar a un individuo e incluso relacionarlo con categorías especiales de datos.

De pronto, eliminar los identificadores deja de ser suficiente para anonimizar los datos. La existencia de los datos Cuasi-Identificadores crea un riesgo de "desanonimización" de la información que se puede medir como la probabilidad de re-identificar a los sujetos a partir del conjunto de cuasi-identificadores. Todo ello, ha llevado a la creación de técnicas de SDC que buscan maximizar la privacidad de los datos sin que esto afecte a los objetivos de las empresas de explotación y extracción de información a partir de estos datos. Estas técnicas se pueden categorizar en dos grandes grupos:

- Técnicas que buscan perturbar o alterar los valores del conjunto de datos para crear incertidumbre sobre la veracidad de los datos.
- Técnicas que buscan reducir el nivel de detalle del conjunto de datos a través de generalización o la eliminación de ciertos valores, sin distorsionar o perturbar la estructura de los datos.

Cabe destacar que de estos dos grupos las técnicas no "perturbativas" se han coronado como el método preferido por empresas para implantar la anonimidad en un conjunto de datos. Con ello, las empresas aseguran la protección de los datos sin la introducción de información errónea en la fuente de datos original que podría llevar a confusiones dentro de la misma empresa. Así, es común el uso de técnicas de generalización y eliminación para aumentar la anonimidad de los datos. Por ejemplo, la información de tres individuos es la siguiente:

CODIGO POSTAL	EDAD	COLESTEROL
28230 (Las Rozas)	40	S
28931 (Móstoles)	44	S
08028 (Hospitalet)	13	S

Podemos hacer que los cuasi-identificadores se vuelvan menos precisos con técnicas de generalización y eliminación:

CODIGO POSTAL	EDAD	COLESTEROL
28***	40-49	S
28***	40-49	S

El atributo Edad se generaliza dentro de un rango numérico, mientras que el atributo Código Postal se generaliza como una jerarquía, en este caso Madrid. Por último, se elimina el último registro, ya que se trata de un valor disonante que poco o nada tiene que ver con los demás valores de la tabla y que al no poder generalizarse puede aumentar la probabilidad de re-identificar al sujeto.

2.2. MARCO LEGAL EN LA UE

El Reglamento General de Protección de Datos (GDPR) es un reglamento que tiene como objeto la protección de datos para todos los individuos dentro de la Unión Europea. El GDPR se redactó con la intención de otorgar a las personas más control sobre sus datos personales y cómo se utilizan y para dar a las empresas un entorno jurídico más simple y claro para operar. El reglamento afecta directamente al almacenamiento, procesamiento, acceso, transferencia y divulgación de los registros de datos de un individuo de la UE, con sanciones por incumplimiento que pueden alcanzar hasta el 4% de los ingresos globales de la compañía.

El GDPR se aplica a los “controladores” y “procesadores” de datos. Los controladores indican cómo y por qué se procesan los datos, estos pueden ser cualquier organización, desde una empresa hasta una organización benéfica o el gobierno de algún estado. Los procesadores en cambio llevan a cabo el procesamiento de los datos, estos pueden ser

cualquier empresa de TI especializada en el procesamiento de datos reales. Igualmente, el GDPR también se ocupa de la exportación de datos personales fuera de la UE. Por lo tanto, el GDPR se aplicará a nivel mundial a toda organización que se encuentre fuera de la UE, siempre y cuando procese datos personales de residentes de la UE.

3. ANÁLISIS PREDICTIVO

3.1. ¿QUÉ ES?

El término análisis predictivo describe la aplicación de una o varias técnicas estadísticas, o de aprendizaje automático para crear una predicción cuantitativa sobre el futuro. (MWa, n.d.) Esta predicción se hace en base a la información extraída de los datos recogidos y se fundamenta en la identificación de relaciones entre variables en eventos pasados. En otras palabras, "el análisis predictivo consiste en la tecnología que aprende de la experiencia para predecir el futuro comportamiento de individuos para tomar mejores decisiones" (Siegel, 2013).

"Para llevar a cabo el análisis predictivo es indispensable disponer de una considerable cantidad de datos, tanto actuales como pasados, para poder establecer correlaciones entre las variables y patrones de comportamiento." (Espino Timón, 2017). Por ejemplo, si una compañía de seguros quiere conocer que clientes son más propenso a sufrir un accidente, se cruzarán datos como la edad, sexo, tipo de vehículo, antecedentes o historial de conducción, para poder identificar que cliente es más probable que haga uso del seguro contratado en el futuro. (Rouse, n.d.). Este proceso se realiza gracias a técnicas de aprendizaje supervisado como las regresiones lineales, redes neuronales o los arboles de decisión.

Si bien establecer correlaciones entre variables es un aspecto imprescindible para que el análisis predictivo sea un éxito, es igual de importante saber interpretar correctamente estas correlaciones. La correlación no implica causalidad, es decir, la relación entre A y B no implica que una cause la otra (Espino Timón, 2017). Si existe una fuerte correlación, pero A no causa B, puede deberse o bien a que exista un tercer fenómeno C que provoque tanto A como B o por puro azar. Por ejemplo, si se identifica una correlación entre el aumento de reservas en hoteles de playa y el incremento en las ventas de helados, esto podría llevar a pensar que el hecho de reservar en un hotel de playa aumenta la probabilidad de que alguien se coma un helado. No obstante, lo cierto es que esta correlación sucede gracias a un tercer factor, el aumento de las temperaturas. Este aumento de las temperaturas hace que la gente aproveche para ir más a zonas de playa y a la vez para comerse algo refrescante como un helado.

El último paso una vez encontradas e interpretadas correctamente las correlaciones, será crear el modelo predictivo. El modelo predictivo se utilizará para intentar predecir el comportamiento de las personas en situaciones particulares: si cambiarán de voto, si comprarán un producto o servicio determinado etc. (Espino, Timón, 2017). Se introducirán los datos de un individuo en el modelo y se obtendrá una clasificación que indicará la probabilidad de que se produzca la situación estudiada por el modelo (Espino Timón, 2017). Retomando el ejemplo anterior, el modelo predictivo de una compañía de seguros indicaría la probabilidad de que un cliente pueda tener un accidente, y en función del valor obtenido la compañía pondrá un precio u otro al seguro del cliente. Sin embargo, hay que tener en cuenta que un modelo predictivo por muy fiable que sea no acertará siempre. Esto es debido a que por mucho que se haya repetido un patrón de comportamiento esto no es un hecho seguro y no tiene por que repetirse (Espino Timón, 2017).

3.2. MODELOS APLICABLES AL ANÁLISIS PREDICTIVO

Cuando hablamos de análisis predictivo realmente nos referimos a la creación de modelos predictivos. (Espino Timón, 2017). No obstante, el auge que la analítica predictiva está experimentando ha hecho que el termino se utilice para referirse a todo lo relacionado con la disciplina del análisis de datos, incluyendo el modelado descriptivo y prescriptivo.

Como se ha mencionado en el apartado anterior, los modelos predictivos proporcionan una calificación dependiendo de los datos recibidos por parte de un individuo. Cuanto más alta sea esta calificación más probabilidades hay de que el individuo exhiba el comportamiento analizado por el modelo. (Espino Timón, 2017). Sin embargo, esta calificación por si sola no debe de ser tratada como un hecho real, sino que debe de ser tenida en cuenta con especial cuidado y debe complementarse con análisis adicionales de carácter descriptivo y prescriptivo. (Espino Timón, 2017). Por ello, para que el análisis de datos avanzado sea lo más completo y fiable posible es necesario incluir en mayor o menor medida los tres aspectos de la analítica de datos; cada cual más complejo que el anterior. Así, el modelado predictivo crea una estimación a partir del modelado descriptivo y a continuación el modelado prescriptivo indica cómo reaccionar de la mejor manera posible de acuerdo con la predicción.

3.2.1. Modelos Descriptivos

La analítica descriptiva examina los datos y analiza los sucesos pasados para entender el presente y saber cómo abordar el futuro. Se utilizan datos históricos para examinar el rendimiento pasado, entender ese rendimiento y encontrar las razones detrás del éxito o el fracaso del pasado. La mayoría de los informes de gestión, tales como ventas, marketing, operaciones y finanzas, utilizan este tipo de análisis. Por ejemplo, "la analítica descriptiva examina los datos históricos de uso de la electricidad para ayudar a planificar las necesidades de energía y permitir a las compañías eléctricas establecer precios óptimos." (analisicinetifico.com, 2017).

Los modelos descriptivos cuantifican las relaciones entre los datos de manera que es utilizado a menudo para clasificar clientes o contactos en grupos. (Espino Timo, 2017). A diferencia de los modelos predictivos que se centran en predecir el comportamiento de un cliente en particular (analisicinetifico.com, 2017), los modelos descriptivos identifican diferentes relaciones entre los clientes y los productos. Igualmente, "los modelos descriptivos no clasifican u ordenan a los clientes por su probabilidad de realizar una acción particular de la misma forma en la que lo hacen los modelos predictivos." (analisicinetifico.com, 2017). Sin embargo, los modelos descriptivos pueden ser utilizados por ejemplo para clasificar a los clientes según sus preferencias de producto, franja de edad, etc. (Espino Timón, 2017). Algunos ejemplos de modelado descriptivo son: simulaciones o técnicas de previsión.

3.2.2. Modelos Predictivos

Como se ha explicado el análisis predictivo utiliza datos para determinar el resultado futuro probable de un evento o la probabilidad de que se produzca una situación. Este tipo de análisis emplea una variedad de técnicas estadísticas de modelado, aprendizaje automático, Big Data y teoría de juegos que analizan los hechos actuales e históricos para hacer predicciones. (analisicinetifico.com, 2017 & MWa, n.d.)

“Los modelos predictivos son modelos de la relación entre el rendimiento específico de un sujeto en una muestra y uno o más atributos o características del mismo sujeto.” (Espino Timón, 2017). El objetivo del modelo es evaluar la probabilidad de que un sujeto similar tenga el mismo rendimiento en una muestra diferente. Esta categoría engloba modelos en muchas áreas como el marketing, donde se buscan patrones de datos ocultos que respondan preguntas sobre el comportamiento de los clientes o modelos de detección de fraude. (Espino Timón, 2017). Gracias a los avances de ingeniería en el análisis de grandes volúmenes de datos estos modelos son capaces de simular el comportamiento humano frente a estímulos o situaciones específicas. (Espino Timón, 2017).

3.2.3. Modelos Prescriptivos

La analítica prescriptiva se trata de la forma de análisis de datos más compleja. “Esta forma de análisis va más allá de predecir los resultados futuros al sugerir también acciones para beneficiarse de las predicciones o incluso acciones para intentar modificarlas. Igualmente, el análisis prescriptivo no sólo anticipa lo que sucederá y cuándo ocurrirá, sino también por qué sucederá.” (analisicinetifico.com, 2017). Los modelos prescriptivos, combinan sinérgicamente datos (incluyendo los resultados de los modelos predictivos), reglas de negocios, técnicas de aprendizaje automático y modelos matemáticos complejos para mejorar la precisión de la predicción y proporcionar mejores opciones de decisión.

Estos modelos pueden ser utilizados en la optimización o maximización de determinados resultados al mismo tiempo que otros son minimizados. (Espino Timón, 2017). Los modelos de decisión son generalmente usados para el desarrollo de la decisión lógica o conjuntos de reglas de negocio que deberían producir el resultado deseado para cada cliente o circunstancia. (Espino Timón, 2017). Por ejemplo, dentro del sector energético los precios del gas natural fluctúan dependiendo de distintas variables como la oferta, la demanda, la economía, la geopolítica y las condiciones meteorológicas. Sectores como el de la producción de gas, el transporte o los servicios públicos, son altamente sensibles a las variaciones en el precio del gas, por lo que tienen un gran interés en predecir con mayor con exactitud los precios del gas para poder cubrirse de posibles riesgos a la baja y cuadrar cuentas de manera favorable. Así, un modelo prescriptivo es capaz no solo de predecir con un alto grado de fiabilidad los precios mediante el modelado de variables internas y externas simultáneamente, sino que también es capaz de proporcionar opciones

de decisión y de mostrar el impacto de cada opción de decisión. (analisicinetifico.com, 2017).

3.3. PRINCIPALES APLICACIONES DEL ANÁLISIS PREDICTIVO EN LA EMPRESA

En los negocios, los modelos predictivos explotan patrones encontrados en datos históricos y transaccionales para identificar riesgos y oportunidades. Los modelos captan las relaciones entre muchos factores para permitir la evaluación del riesgo o potencial asociado con un conjunto particular de condiciones, guiando la toma de decisiones para las transacciones candidatas. Una organización debe invertir en un equipo de expertos y crear algoritmos estadísticos para encontrar y acceder a los datos pertinentes. El equipo de análisis de datos trabaja con líderes empresariales para diseñar una estrategia para el uso de información predictiva. Empresas como IBM, SAS, Oracle o Microsoft se han coronado como los referentes del mercado predictivo, ofreciendo soluciones de análisis predictivo a otras empresas para mejorar su competitividad en el mercado. Las soluciones que ofrecen estas empresas son variadas:

- Soluciones de analítica de clientes para anticiparse a los grados de satisfacción o enfados de los clientes para retenerlos o incrementar sus ingresos. Con ellas las compañías son capaces de lanzar ofertas personalizadas en distintos canales, prever que clientes están a punto de cambiarse etc. (Cía, 2015).
- Soluciones de analítica operacional para evaluar costes operativos (Cía, 2015).
- Soluciones de analítica predictiva para Big Data. Estas soluciones ayudan a las compañías que disponen de una gran cantidad de datos a organizar, entender y extraer correlaciones de esos datos (Cía, 2015).
- Soluciones de amenazas y fraude para anticiparse a cualquier amenaza o fraude en la compañía (Cía, 2015).

Por otro lado, expertos como Thomas Shimanda y el doctor Fabián López distinguen en su artículo de la revista Inbound Logistic Latam, cinco áreas de aplicación de la analítica predictiva de cara a hacer más rentable la cartera de clientes:

1. Segmentación de clientes: permite adecuar las ofertas en función de variables como el nivel de ingresos, franja de edad, sexo o estudios realizados.

2. Personalización de la oferta: permite conocer cuál es la siguiente mejor oferta que se le puede hacer a un cliente a partir de su comportamiento histórico.

3. Detectar el riesgo de que el cliente abandone la relación comercial en función del ritmo de pedidos o contactos que realiza o de las incidencias que registra.

4. Conocer cuáles son los clientes más propensos a responder a las iniciativas de comunicación publicitaria, para sacar el mayor provecho a la inversión hecha.

5. Conocer la tasa de deserción, es decir, predecir en forma anticipada y proactiva cuáles son los clientes que están buscando otras ofertas para evitar que estos desvíen su atención hacia la competencia. A través de esta aplicación separo los clientes rentables de los que no lo son.

Las soluciones y aplicaciones propuestas se centran solo en el sector empresarial, sin embargo, el análisis predictivo también está presente otros sectores como sanidad, farmacéutico, automoción, aeroespacial y fabricación. A continuación, nos centraremos cómo los sectores de marketing, financiero y empresarial se benefician del análisis predictivo más en detalle.

3.3.1. Sector Financiero

Algunas de las principales aplicaciones del análisis predictivo en el sector financiero son:

- Gestión del riesgo crediticio: La autorización de créditos a personas y empresas es una parte muy importante del negocio de las instituciones financieras. Por ello, es necesario analizar la solvencia del individuo antes de poder concederle un préstamo. Datos como los patrones de compra, donde realizan las compras, cuales son los últimos acuerdos cerrados o el histórico de transacciones, son de gran utilidad para el sector financiero a la hora de determinar los préstamos que puede conceder a distintos individuos o empresas. (Ladrero, 2018).

- Compraventa de acciones: Existen modelos predictivos capaces de predecir si el valor de una acción subirá o bajará. Con esta información, el usuario del activo podrá anticiparse a las fluctuaciones en el valor del activo comprando o vendiendo cuando sea necesario. (Espino Timón, 2017 & Ladrero, 2018).

- Comercialización de mercancías o materias primas: Un comerciante que trabaje con los precios del aceite puede obtener flujos de datos directamente desde las granjas en tiempo real gracias al uso de sensores que le informan de las condiciones meteorológicas, tiempo de recolección, etc. Estos datos permiten a los comerciantes hacer apuestas apoyadas en datos y no simplemente apostar al azar para obtener mejor rendimiento. (Ladrero, 2018). Esto podría aplicarse con cualquier tipo de mercancía o materia prima incluyendo el barril de petróleo o el valor del oro.

- Asesoramiento de inversión: Los asesores de banca de inversión o de gestión de patrimonio son capaces de utilizar modelos predictivos para ofrecer nuevas posibilidades e ideas de inversión a sus clientes. Estas predicciones pueden llegar a incluir la probabilidad de éxito de la inversión en función de datos como: solvencia del cliente que va a llevar a cabo la operación, tamaño de la inversión, situación macroeconómica del país donde se va a realizar la inversión, tipo de cambio o fluctuación de los tipos de interés. (Ladrero, 2018).

3.3.2. Sector Empresarial

Algunas de las principales aplicaciones del análisis predictivo en el sector empresarial son:

- Retención de clientes: Los modelos predictivos son capaces de predecir que clientes tienen mayor probabilidad de abandonar la empresa en función de indicadores que puedan mostrar una pérdida de interés en su oferta, como, por ejemplo, no haber visitado la página web en un tiempo considerable, haber dado «me gusta» a un producto de la competencia en redes sociales etc. Esto ayudará a la empresa a identificar a los clientes más propensos a marcharse y a orientar sus esfuerzos a retener a dichos clientes. (Espino Timón, 2017 & Ladrero, 2018).

- Aumentar las ventas: La analítica predictiva es capaz de predecir el grado de aceptación que un nuevo producto o servicio puede tener basándose en datos recopilados de eventos pasados, como, por ejemplo, otros lanzamientos de productos o servicios anteriores, promociones antiguas, alteraciones de la demanda frente a modificaciones de productos ya existentes, etc. Con esta información, la empresa es capaz de asimilar y detallar mejor la demanda de sus consumidores entendiendo que productos y servicios tendrán más éxito entre sus consumidores, aumentando así sus ventas.

- Oferta más personalizada: La analítica predictiva otorga a la empresa la capacidad de prever cual será la demanda de los usuarios y por lo tanto realizar una oferta más personalizada a las necesidades y gustos de sus clientes. Plataformas como Spotify o Netflix son un claro ejemplo. Cada vez que los consumidores hacen clic estas páginas aprenden sobre las preferencias de sus consumidores ofreciendo una oferta más personalizada basándose en dichas preferencias. (Espino Timón, 2017 & Ladrero, 2018).

3.3.3. Sector Marketing

Algunas de las principales aplicaciones del análisis predictivo en el sector marketing son:

- Marketing directo: El análisis predictivo permite a la empresa identificar que individuos son más sensibles a las técnicas de marketing, es decir, que consumidores son más influenciables. Igualmente, un modelo predictivo puede indicar a que personas no se debe contactar de manera directa, ya que ello podría terminar siendo contraproducente. Con esta información, la empresa es capaz de hacer una distinción entre potenciales clientes y rentabilizar y centrar sus esfuerzos de marketing en comunicarse con aquellos que tienen una mayor probabilidad de responder si son contactados de forma directa. Este aspecto es también interesante a la hora de captar votantes durante las campañas electorales. De este modo se pueden centrar los esfuerzos durante la campaña para acceder a aquellos votantes que pueden cambiar de voto. (Espino Timón, 2017).

- Publicidad predictiva: Al igual que con en el lanzamiento de nuevos productos, un modelo predictivo es capaz de predecir el grado de aceptación que un anunció tendrá basándose en el producto que se quiere vender, anuncios o promociones anteriores, hora a la que se retransmitirá el anuncio, plataforma por la que se retransmitirá, segmento de la población para el que va dirigido el anuncio, etc. (Espino Timón, 2017).

- Anuncios más personalizados: Es normal que nada más recibir un anuncio sobre el coche del que hemos hablado de comprar hace tan solo diez minutos, pensemos que las empresas nos leen la mente o que nos escuchan. Sin embargo, esto es simplemente un modelo predictivo recogiendo y procesando los datos vocales de tu conversación para desarrollar una oferta más personalizada en función de tus gustos. Esto igualmente sucede cuando realizas búsquedas en internet o cuando frecuentas distintos sitios con la función gps del teléfono móvil activada. El modelo predictivo utilizará esos datos que generas para personalizar anuncios basándose en tu actividad.

4. MACHINE LEARNING

A lo largo de los puntos anteriores se han hecho numerosas referencias al aprendizaje computacional como aspecto clave para realizar un análisis predictivo. Hoy en día se ha vuelto prácticamente imposible hablar de análisis predictivo sin hacer alguna referencia al aprendizaje computacional, sin embargo, ¿qué es el aprendizaje computacional?

4.1. ¿QUÉ ES?

El aprendizaje computacional o machine learning, se trata de una rama de la inteligencia artificial que se centra en aprender un modelo determinado en base a unos datos. Para ello, se aplican distintos algoritmos o técnicas de análisis de datos que detecten de forma automática patrones significativos en los datos.

Vamos a demostrar un ejemplo típico de una tarea de machine learning. El ejemplo está obtenido del artículo *Understanding Machine Learning: From Theory to Algorithms* de Shalev-Shwartz y Ben-David: Supongamos que queremos programar una máquina para que sea capaz de aprender a filtrar correo basura o considerado como *spam*. Una solución podría ser hacer que la máquina simplemente memorizase todos aquellos correos que en el pasado han sido considerados como *spam* por el usuario. Así, cuando llegue un nuevo correo electrónico, la máquina lo buscará en el conjunto de los anteriores correos electrónicos considerados como *spam* y si coincide con uno de ellos, será destruido. De lo contrario, se moverá a la bandeja de entrada del usuario. No obstante, el aprendizaje por memorización, aunque útil, se encuentra incompleto. Con este método la máquina carece de la capacidad de etiquetar los correos electrónicos no vistos por el usuario. Para que el aprendizaje sea un éxito, la máquina debe ser capaz de aprender por razonamiento intuitivo, es decir, la máquina debe ser capaz de pasar de los ejemplos individuales a una generalización más amplia. (Shalev-Shwartz, S. & Ben-David, S., 2014). Para alcanzar la generalización en el ejemplo del filtrado de correo basura, Shalev-Shwartz y Ben Davis explican que el usuario puede escanear los correos electrónicos pasados, y extraer un conjunto de palabras cuya aparición en un mensaje de correo electrónico es indicativo de *spam.* Ahora, al llegar un nuevo correo electrónico, la máquina puede comprobar si una de las palabras pertenecientes al conjunto aparece en él y predecir su calificación en

consecuencia. Con ello, somos capaces de superar el problema que los correos electrónicos no vistos planteaban. (Shalev-Shwartz, S. & Ben-David, S., 2014).

El artículo prosigue describiendo que, sin embargo, el razonamiento inductivo es capaz de llevarnos a conclusiones precipitadas y falsas. Esto es más fácil de entender con el experimento llevado a cabo por el psicólogo B. F. Skinner: en este experimento el psicólogo colocó a un grupo de palomas dentro de una jaula que contenía un mecanismo automático que alimentaba a las palomas en intervalos regulares. La primera vez que la comida fue suministrada, las palomas se encontraban en movimiento (picando, aleteando, etc.). Al cabo de un tiempo las palomas asimilaron una conexión entre su movimiento y la entrega de comida. Esto resultó en un mayor movimiento y actividad por parte de las palomas que asociaban la entrega de alimento con sus acciones. Así, en el aprendizaje, el razonamiento intuitivo es a veces capaz de generar una cadena de eventos que refuerza la asociación de una acción con una causa, de manera errónea. (Skinner, 1947). En el caso de los humanos a veces es posible superar este problema aplicando el sentido común para desechar conclusiones erróneas, mientras que en el caso de las máquinas que carecen de sentido común, es crucial proporcionarles principios bien definidos y nítidos que protejan al programa de llegar a conclusiones falsas. (Shalev-Shwartz, S. & Ben-David, S., 2014). La máquina no debe adoptar cualquier explicación para la ocurrencia de un suceso, debe adoptar la explicación correcta. Para ello, será clave la introducción de conocimiento previo en los algoritmos de aprendizaje. En términos generales, cuanto más fuerte sea el conocimiento previo (o los supuestos previos) con los que se inicia el proceso de aprendizaje, más fácil será aprender de otros ejemplos. (Shalev-Shwartz, S. & Ben-David, S., 2014).

El nivel de conocimiento previo del algoritmo de machine learning estará determinado por la cantidad de datos de la que dispongamos, es decir, cuantos más datos, más adecuado será el algoritmo usado. Los algoritmos son instrucciones a ordenadores o máquinas para que sepan cómo interactuar, manipular y transformar los datos. A diferencia de otros algoritmos donde primero se crea el algoritmo y más tarde se añaden los datos, los algoritmos de machine learning son creados a partir de los datos recopilados. Por último, Shalev-Shwartz y Ben-David explican que existen tres situaciones donde, sobre todo, es necesario el machine learning sobre un modelo de programación estándar:

- Tareas demasiado complejas para programar, como aquellas llevadas a cabo por personas o animales. Ejemplos como conducir o el reconocimiento facial y de voz son tareas que los humanos realizamos de manera rutinaria y sin embargo son increíblemente complejas de programar; parece que nuestra introspección sobre cómo las hacemos no es lo suficientemente elaborada como para lograr un programa bien definido. Los programas de machine learning, por otro lado, parecen obtener resultados bastante satisfactorios replicando este tipo de tareas una vez son expuestos a suficientes modelos o ejemplos de entrenamiento gracias a su capacidad de aprender de la experiencia. (Shalev-Shwartz, S. & Ben-David, S., 2014).

- Tareas que se encuentran más allá de la capacidad humana, como aquellas relacionadas con el análisis de conjuntos de datos complejos y de gran volumen: la conversión de archivos médicos en conocimiento médico, predicción del tiempo, análisis de datos de genomas o de datos astronómicos, motores de búsqueda en la web y el comercio electrónico son otra amplia gama de tareas que se benefician de las técnicas de machine learning. La enorme cantidad de datos disponibles y registrados digitalmente hace que se vuelva evidente la existencia de grandes tesoros de información significativa enterrados en archivos de datos que son demasiado grandes y complejos para ser comprendidos por los humanos. Por lo tanto, la capacidad de aprender a detectar patrones significativos dentro de conjuntos de datos complejos de los programas de machine learning abre nuevos horizontes y posibilidades de aprendizaje y desarrollo a los seres humanos. (Shalev-Shwartz, S. & Ben-David, S., 2014).

- Tareas que requieran de un alto grado de adaptabilidad. Como se ha mencionado antes, los algoritmos de métodos de programación estándar son rígidos, se crean y luego se añaden los distintos datos, se mantienen sin cambios. Los algoritmos de machine learning en cambio, al ser creados a partir de sus datos de entrada se adaptan por naturaleza a cualquier cambio en el entorno con el que interactúen. Algunos ejemplos de tareas que requieren de adaptación y pueden beneficiarse de los algoritmos de machine learning son: decodificación de textos escritos a mano, donde el programa es capaz de adaptarse a las variaciones de caligrafía ente usuarios, la detección de correos basura (como se ha ejemplificado antes), donde

el programa se adapta automáticamente a los cambios en los correos considerados como spam, o el reconocimiento de voz, donde el programa se adapta a los distintos tonos y timbres de voz de los usuarios. (Shalev-Shwartz, S. & Ben-David, S., 2014).

4.2. TIPOS DE APRENDIZAJE SUPERVISADO VS NO SUPERVISADO

El ámbito del aprendizaje es, por supuesto, muy amplio. En consecuencia, el campo del machine learning se ha ramificado en varios subcampos que se ocupan de diferentes tipos de labores de aprendizaje (Shalev-Shwartz, S. & Ben-David, S., 2014). De todas estas modalidades, este trabajo destacará dos: el aprendizaje supervisado y no supervisado.

Dado que el aprendizaje implica una interacción entre el alumno y el entorno, se pueden dividir las tareas de aprendizaje según la naturaleza de esa interacción. (Shalev-Shwartz, S. & Ben-David, S., 2014). Con esta división, surgen el aprendizaje supervisado y no supervisado. Como ejemplo ilustrativo vamos a considerar, de nuevo, la tarea de detección de anomalías frente a la tarea anterior sobre detección de correo basura o *spam* expuesta por Shalev-Shwartz y Ben-David. Recordemos que, para esta última tarea, el programa recibe a modo de entrenamiento mensajes previamente categorizados como *spam* y no *spam*. En base a este entrenamiento, el algoritmo debe establecer una regla para etiquetar un mensaje de correo electrónico recién llegado. Por el contrario, en la tarea de detección de anomalías lo único que el algoritmo recibe a modo de entrenamiento es un gran número de datos sin categorizar y sobre los que debe encontrar alguna anomalía. (Shalev-Shwartz, S. & Ben-David, S., 2014).

Si entendemos el aprendizaje como un proceso que utiliza la experiencia para adquirir conocimiento; el aprendizaje supervisado describe un escenario donde la experiencia (en este caso el modelo que se ha empleado para entrenar al programa) contiene información significativa (etiquetas que categorizan la información) que no se encuentra en los ejemplos de prueba no vistos por el programa. En este contexto, “la experiencia adquirida tiene por objeto predecir la información que falta en los ejemplos de prueba, pudiendo considerar al entorno como una especie de profesor que supervisa al programa ofreciéndole información extra (las etiquetas)” (Shalev-Shwartz, S. & Ben-David, S., 2014). Por otro lado, en el aprendizaje no supervisado no existe distinción entre un

ejemplo de entrenamiento y otro de prueba, es decir, los modelos de entrenamiento no ofrecen ninguna información extra que pueda ser empleada por el algoritmo para más tarde ser aplicado en los modelos de prueba. El programa procesa los datos con el objetivo de obtener una versión resumida o comprimida de esos datos. (Shalev-Shwartz, S. & Ben-David, S., 2014).

4.3. TÉCNICAS DE APRENDIZAJE SUPERVISADO

"El aprendizaje supervisado se emplea en aplicaciones financieras para la calificación crediticia, el trading algorítmico y la calificación de bonos; en aplicaciones biológicas para la detección de tumores y el descubrimiento de fármacos; en aplicaciones energéticas para la predicción de carga y precio; y en aplicaciones de reconocimiento de patrones para el habla y las imágenes." (MWb n.d.) El aprendizaje supervisado incluye dos categorías de algoritmos: regresión y clasificación. (MWb, n.d.)

4.3.1. Modelos de Regresión

Los modelos de regresión permiten estudiar y cuantificar la relación entre una variable dependiente o de respuesta continua y una o más variables independientes o predictores con el fin de averiguar en qué medida la variable dependiente puede estar explicada por la variable o variables independientes, y de predecir nuevas observaciones en la variable dependiente a partir de la variable o variables independientes. (Análisis de Regresión lineal, n.d.). Algunos de los modelos de regresión más comunes son:

<u>Regresión lineal:</u>

Un modelo de regresión lineal permite cuantificar las relaciones entre una variable dependiente (**Y**) y una o varias variables independientes (**X**) a través de la creación de una ecuación lineal. Una función de regresión lineal debe ser lineal en los parámetros, lo cual restringe la ecuación a una sola forma básica (Support.minitab, n.d.), una recta:

$$Y = \beta_0 + \sum \beta_\iota X_\iota$$

Donde **β** representa las estimaciones de parámetros lineales que se ajustan para que la medida sea óptima.

Regresión no lineal:

Un modelo de regresión no lineal permite generar una función no lineal que cuantifica las relaciones no lineales entre una variable dependiente (**Y**) y una o varias variables independientes (**X**) creando una ecuación no lineal. Una función de regresión no lineal no requiere de parámetros lineales, por lo que una ecuación no lineal puede adoptar muchas formas diferentes (Support.minitab, n.d.):

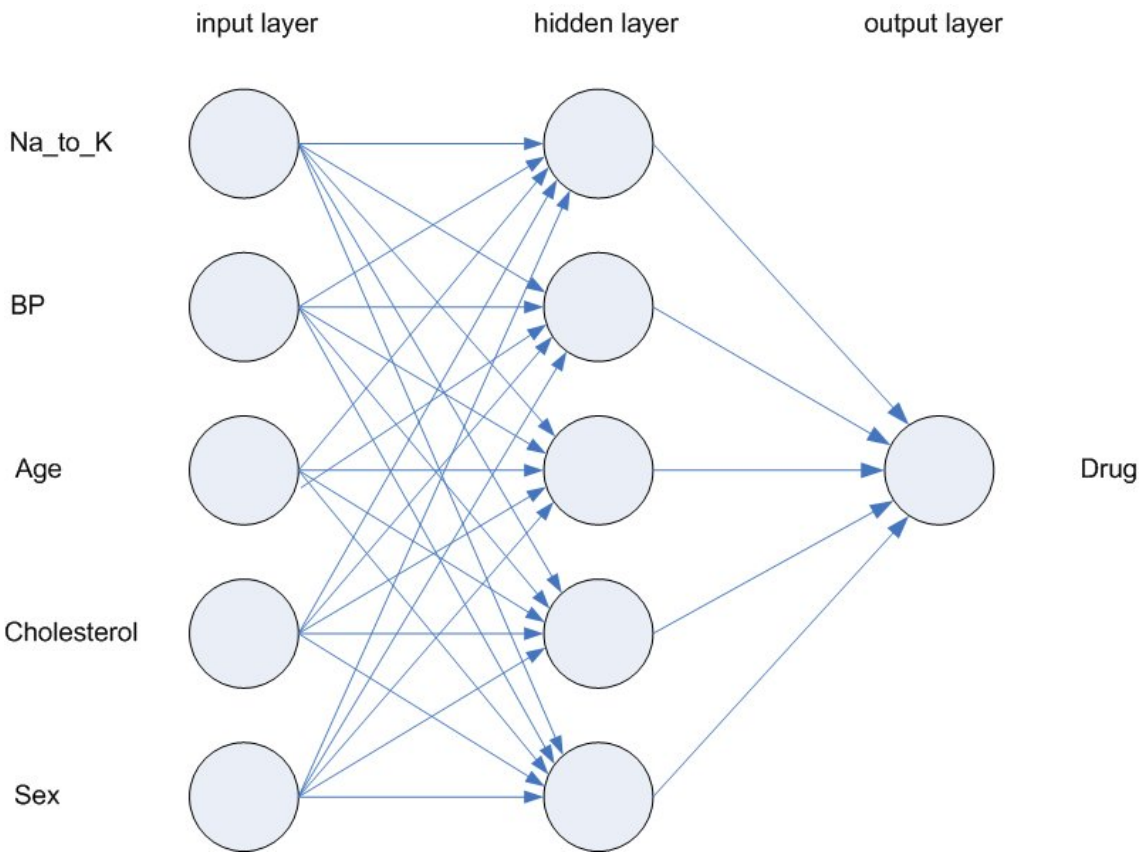
$$\mathbf{Y} = f(\mathbf{X}, \beta)$$

Donde f es alguna función no lineal respecto unos parámetros β no lineales. Las formas no lineales posibles incluyen: convexa, cóncava, curva sigmoidal, curvas asintóticas, etc. (Support.minitab, n.d.)

Redes neuronales:

De acuerdo con la definición dada por la empresa IBM: "El modelo de red neuronal es un modelo simplificado que emula el modo en que el cerebro humano procesa la información." Consiste en un conjunto de nodos o unidades de procesamiento, llamadas neuronas artificiales, conectadas entre sí para transmitirse señales. Las unidades de procesamiento o neuronas se agrupan en capas. Según IBM, hay tres capas normalmente en una red neuronal: una capa de entrada (input layer), con unidades que representan campos de entrada por donde acceden los datos; una capa oculta (hidden layer), donde se procesan y modifican los datos; y una capa de salida (output layer), con una unidad o unidades que representa el campo o los campos de destino por donde se extrae la información media que se ha ido transfiriendo de neurona en neurona. Los datos se propagan desde cada neurona hasta cada neurona de la capa siguiente y cada neurona posee a su vez un peso. "La red aprende examinando los registros individuales, generando una predicción y realizando ajustes a las ponderaciones cuando realiza una predicción incorrecta. Este proceso se repite muchas veces y la red sigue mejorando sus predicciones hasta haber alcanzado uno o varios criterios de parada." (IBM, n.d.).

Por lo general, las redes neuronales se utilizan para describir una relación entre los datos de entrada y de salida y con ello poder realizar predicciones. El funcionamiento puede observarse mejor con el siguiente esquema:



**Fuente –El modelo de redes neuronales (IBM)*

Este modelo se puede aplicar igualmente para modelos de clasificación.

4.3.2. Modelos de Clasificación

"Un método de clasificación es la partición del conjunto de datos en dos conjuntos de datos más pequeños que serán utilizados con los siguientes fines: entrenamiento y test. El subconjunto de datos de entrenamiento es utilizado para estimar los parámetros del modelo y el subconjunto de datos de test se emplea para comprobar el comportamiento del modelo estimado." (Parra, 2019) Un modelo de clasificación ideal es capaz de entrenar al modelo empleando un conjunto de datos independiente de los de los datos con

los que probamos el modelo. (Parra, 2019). Algunos de los modelos de clasificación más comunes son:

Clasificador Bayesiano:

Atendiendo a la explicación de Francisco Parra (2019): "Naïve Bayes es uno de los clasificadores más utilizados por su simplicidad y rapidez. Se trata de una técnica de clasificación y predicción supervisada que construye modelos que predicen la probabilidad de posibles resultados, en base al Teorema de Bayes o de la probabilidad condicionada":

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)}$$

**Fuente – Estadística y Machine Learning con R: Clasificador Bayesiano (Parra, 2019)*

Análisis Discriminante:

Se trata de una técnica que se utiliza para "clasificar la pertenencia de uno o más individuos en un grupo o población alternativos a partir de un conjunto de predictores." (de la Fuente Crespo, na). El objetivo del análisis discriminante "es entender las diferencias de los grupos y predecir la verosimilitud de que una persona o un objeto pertenezca a una clase o grupo basándose en los valores que toman los predictores." (Parra, 2019). Uno de los ejemplos más comunes de análisis discriminante es identificar el riesgo de impago de un préstamo:

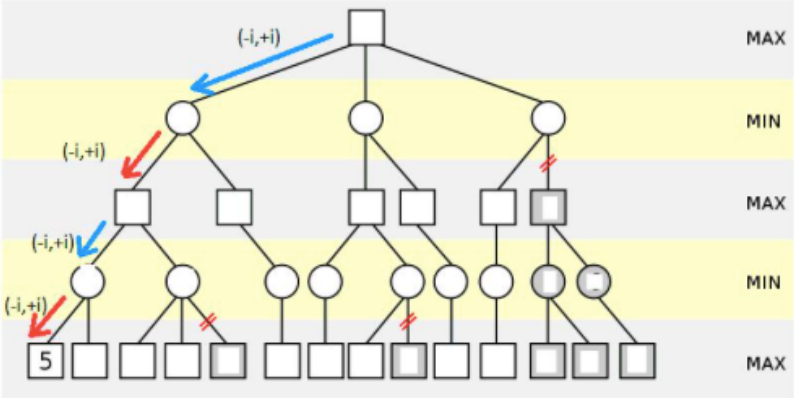
El ejemplo está obtenido del artículo Análisis Discriminante por Laura de la fuente Crespo: Cuando un banco concede un préstamo se enfrenta a la posibilidad de que no sea reintegrado. En caso de no ser reintegrado será clasificado como fallido. En esta línea, se pueden considerar dos grupos de clientes: cumplidores o fallidos. Así, el banco puede utilizar la información existente que posee sobre préstamos concedidos en el pasado en la concesión de préstamos futuros de forma que se evite o, por lo menos se reduzca la concesión de préstamos a clientes que entren en la categoría de fallidos. Para ello, lo primero que tendría que hacer el banco es analizar la información y características de los clientes a los que haya concedido un préstamo. Como es probable que los clientes cumplidores tengan unas características distintas de los clientes fallidos, el siguiente paso será utilizar estas características establecer unas funciones que clasifiquen a los clientes

a los que se les ha concedido un préstamo de la manera más correcta y precisa posible en cumplidores y fallidos. (de la Fuente Crespo, na)

Árboles de decisión:

Los árboles de decisión o de clasificación son un modelo de Machine Learning que, partiendo de una base de datos, crea diagramas de construcciones lógicas que nos ayudan a resolver problemas. (Parra, 2019). Si atendemos al manual de Francisco Parra (2019) se nos explica que: "los árboles de decisión se componen de: nodos, ramas y hojas. Los nodos son las variables de entrada, las ramas representan los posibles valores de las variables de entrada y las hojas son los posibles valores de la variable de salida." Cabe destacar que la variable con mayor relevancia del proceso de clasificación se encontrará en el primer elemento del árbol conocido como nodo raíz. (Parra, 2019).

En cuanto a su funcionamiento, Francisco Parra lo describe de la siguiente manera: "Se comienza con el nodo inicial o raíz, dividiendo la variable dependiente a partir de una partición de una variable independiente que se escoge de modo tal que de lugar a dos conjuntos homogéneos de datos. Por ejemplo, se elige la variable χ y se determina un punto de corte *c*, de modo que se puedan separar los datos en dos conjuntos: aquellos con $\chi \leq c$ y los que tienen $\chi > c$. De este nodo inicial saldrán dos: uno al que llegan las observaciones con $\chi \leq c$ y otro al que llegan las observaciones con $\chi > c$. En cada uno de estos nodos se repite el proceso de seleccionar una variable independiente o punto de corte para dividir la muestra. El proceso termina una vez se hayan clasificado todas las observaciones correctamente en su grupo." (Parra, 2019) Todos los algoritmos de los árboles de decisión obtienen modelos más o menos complejos y consistentes respecto a la evidencia. Para evitar incoherencias en los datos o posibles problemas en las predicciones, Parra sugiere limitar el crecimiento del árbol consiguiendo modelos más generales. (Parra, 2019).

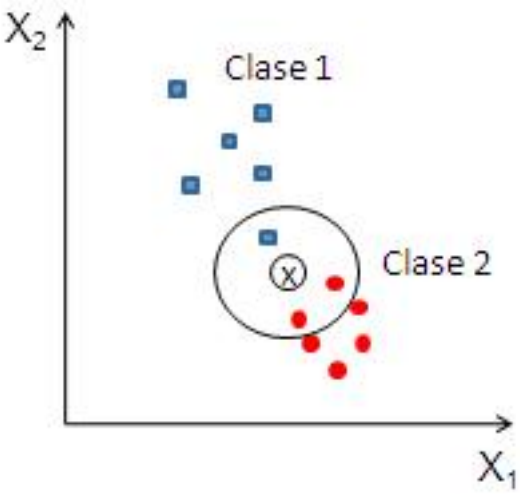


**Fuente – Estadística y Machine Learning con R: Arboles de clasificación (Parra, 2019)*

Al igual que con las redes neuronales, este modelo se puede aplicar también en métodos de regresión.

K-NN:

El método K-vecinos más cercanos es un método de clasificación supervisada que sirve para clasificar cada dato nuevo en el grupo que corresponda, según tenga K vecinos más cerca de un grupo o de otro. (Ruiz, 2017). Es decir, "calcula la distancia del elemento nuevo a cada uno de los existentes, y ordena dichas distancias de menor a mayor para ir seleccionado el grupo al que pertenece." (Ruiz, 2017). El método K-vecinos supone que los vecinos más cercanos nos dan la mejor clasificación. El rendimiento del algoritmo está influenciado por tres factores principales: la variable K o número de vecinos para clasificar la muestra, "con distintos valores de K podemos obtener resultados muy distintos" (Ruíz, 2017); la medida de distancia utilizada para localizar los K-vecinos más cercanos; la regla de decisión usada para derivar una clasificación de los K-vecinos más cercanos.



**Fuente – Estadística y Machine Learning con R: Algorítmo K-vecinos más cercanos (Parra, 2019)*

En la figura, observamos por el círculo que se han seleccionado tres vecinos (K=3). De los tres vecinos más cercanos a χ en la figura, uno de ellos pertenece a la clase uno (el cuadrado azul) y los otros dos a la clase dos (los círculos rojos). (Parra, 2019). Por lo tanto, si los vecinos más cercanos dan la mejor clasificación la regla 3-vecinos, asignará χ a la clase dos. En cambio, si redujéramos K a 1 (K=1), el modelo agruparía a χ con su vecino más cercano que en este caso sería de la clase uno, asignando a χ como parte de la clase uno. (Parra, 2019). Todo ello, considerando que la métrica de distancia empleado sea la distancia euclidea. De no ser así los resultados podrían variar.

4.4. TÉCNICAS DE APRENDIZAJE NO SUPERVISADO

El aprendizaje no supervisado se emplea en aplicaciones financieras para la detección de fraudes en forma de anomalías; en aplicaciones de marketing para la segmentación de los clientes, marketing dirigido y sistemas de recomendación; en aplicaciones biológicas para la clasificación de plantas y animales (González, 2018); en aplicaciones científicas para la visualización de grandes volúmenes de datos; en aplicaciones sismológicas para determinar aquellas zonas más propensas a sufrir un terremoto (González, 2018); en aplicaciones de análisis de redes sociales para la identificación de hubs y autoridades. El aprendizaje no supervisado incluye tres categorías de algoritmos: clúster o agrupación, reducción de la dimensionalidad y reglas de asociación.

4.4.1. Modelo Clúster o de Agrupación

“El análisis clúster es un conjunto de técnicas multivariantes cuyo principal propósito es agrupar objetos basándose en las características que poseen. El modelo clúster clasifica los objetos en clases o conglomerados de tal forma que cada objeto sea parecido a los que hay en el conjunto de su conglomerado. Los conglomerados resultantes deberán tener un alto grado de homogeneidad interna (dentro del conglomerado) y de heterogeneidad externa (entre conglomerados).” (Parra, 2019). Algunos de los métodos de agrupación más comunes son:

Modelos de Partición:

El modelo de clúster de partición o *partitioning clustering*, se compone de una serie técnicas o algoritmos que requieren que se especifique el número de clústeres que se quiere crear. (Amat Rodrigo, 2017). El algoritmo más común de este modelo es:

- K-Medias: El algoritmo K-Medias tiene como objetivo encontrar y agrupar en clases los puntos de datos que tienen una alta similitud entre ellos. Esta similitud se entiende como lo opuesto de la distancia entre datos, es decir, cuanto más cerca estén los datos entre sí, más similares serán y la probabilidad de pertenecer a un mismo clúster será más alta (Roman, 2019). La métrica o medida de distancia que este algoritmo aplica para encontrar similitud entre las observaciones es la distancia euclídea. La distancia euclídea entre dos puntos p y q se define como la longitud del segmento que une ambos puntos. Así, en un espacio de dos dimensiones en el que cada punto está definido por las coordenadas (x, y), la distancia euclídea entre p y q será:

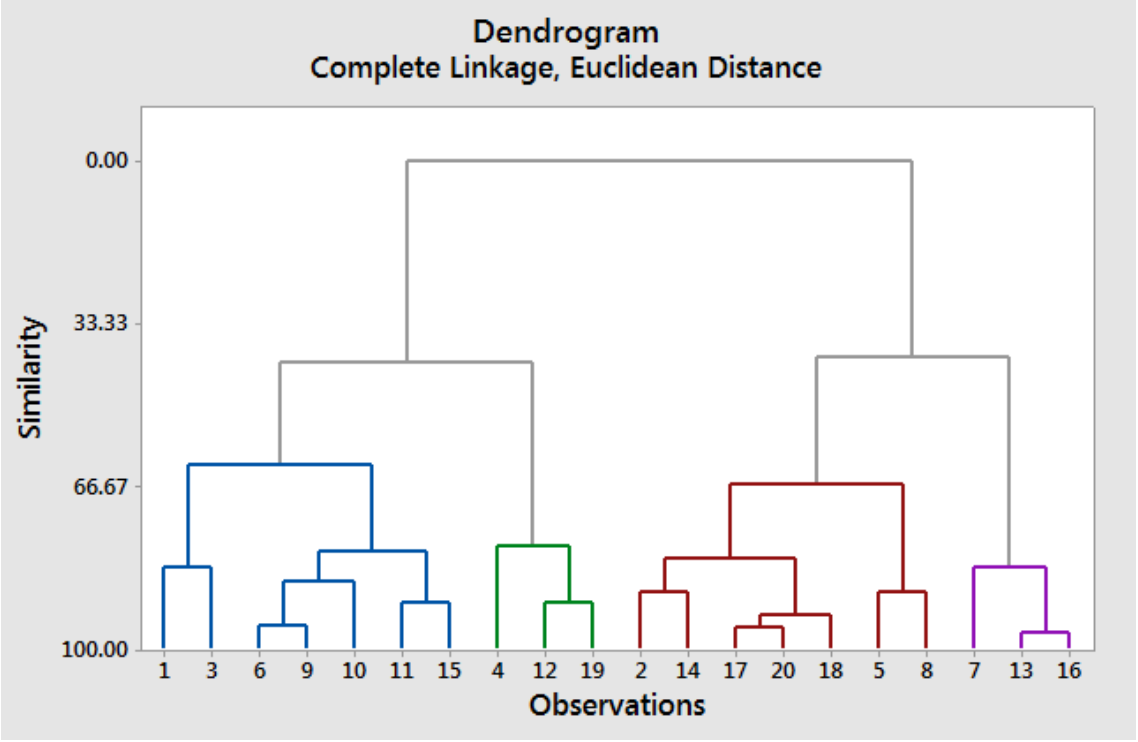
$$d_{euc}(p,q) = \sqrt{(x_p - x_q)^2 + (y_p - y_q)^2}$$

Métodos Jerárquicos:

El modelo de clúster jerárquico no requiere que se pre-especifique el número de clústeres. (Amat Rodrigo, 2017). Los métodos que engloban el agrupamiento jerárquico se subdividen en dos tipos dependiendo de la estrategia seguida para crear los grupos:

- *Agglomerative clustering (bottom-up)*: El agrupamiento se inicia en la base del árbol, donde cada observación forma un clúster individual. Los clústeres se van combinando a medida que la estructura crece hasta converger en una única rama central. (Amat Rodrigo, 2017).
- *Divisive clustering (top-down)*: Es la estrategia opuesta, se inicia con todas las observaciones contenidas en un mismo clúster y se suceden divisiones hasta que cada observación forma un clúster individual. (Amat Rodrigo, 2017).

En ambos casos, los resultados se pueden representar de forma muy intuitiva en una estructura de árbol llamada dendograma:



**Fuente – Aprendizaje No supervisado en Machine Learning: Agrupación (Roman, 2019.).*

Atendiendo a la explciación de Joaquín Amat: "La base del dendograma está compuesta por terminaciones individuales conocidas como hojas del árbol. A medida que el dendograma asciende, las hojas se van fusionando y forman ramas llamadas nodos. Estos nodos a su vez pueden conectarse a hojas individuales u otros nodos. Los nodos corresponden a los pares de observaciones más similares." Así, cuanto más próxima a la base ocurra la fusión de hojas, mayor es la similitud. Por ejemplo, la observación 9 es la más similar a la 6 y viceversa, ya que es la primera conexión que recibe. Los

dendogramas, por lo tanto, se deben interpretar únicamente en base al eje vertical y no al horizontal. (Amat Rodrigo, 2017).

4.4.2. Reducción de la Dimensionalidad

Los métodos de reducción de la dimensionalidad son procedimientos que mapean un conjunto de datos a espacios de menor dimensión, derivados del espacio original. (Arroyo-Hernandez, 2016). En otras palabras, se busca reducir el número de variables de un conjunto de datos. Los modelos de reducción de la dimensionalidad favorecen la compresión, eliminación de redundancia del conjunto de datos y permite mejorar procesos de clasificación y visualización de los datos. Algunos de los métodos de reducción de dimensionalidad más comunes son:

Técnicas de Selección de Variables:

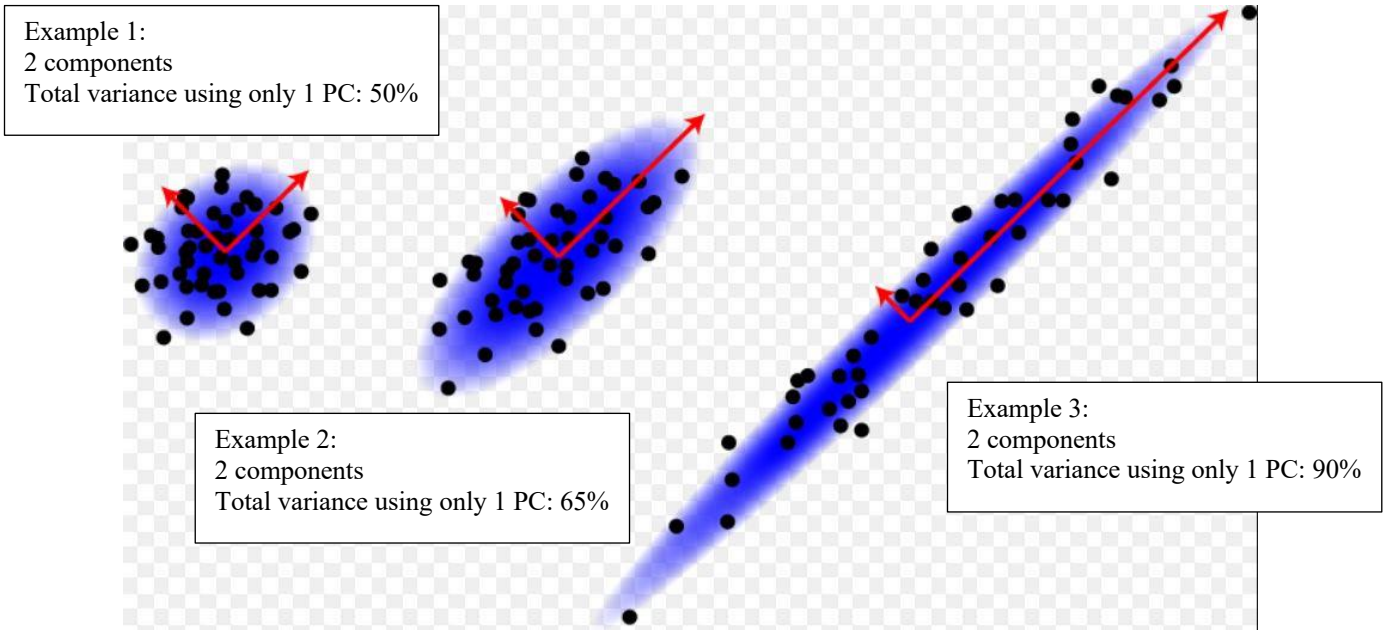
"Un conjunto de variables óptimo para un conjunto de datos será el que contiene las variables más significativas del conjunto de datos original." (Lafuente, 2018). Las técnicas de selección de variables permiten obtener este conjunto óptimo. Uno de los criterios que se utiliza es la correlación. Por ejemplo, "para un modelo de predicción de cáncer en humanos, el conjunto de datos con el que se trabaje probablemente contenga variables como la edad, antecedentes de cáncer en su familia, si fuma o no etc." (Lafuente, 2018). Mientras que variables como el color de ojos, en principio no se deberían de encontrar dentro del conjunto de datos. No obstante, no siempre podemos estar seguros de que una variable no influya en la probabilidad de que surja otra, en este caso no podemos estar seguros de que el color de los ojos no influya para nada en padecer cáncer. (Lafuente, 2018).

Para evitar este problema, podemos emplear diferentes métodos matemáticos como estudiar la correlación o la consistencia. El criterio de correlación permite medir el nivel de asociación entre variables permitiéndonos deshacernos de aquellas que tengan correlación negativa o también igual o próxima a 0. Otro criterio que podemos emplear es el de la consistencia. (Lafuente, 2018). Este criterio nos permite deshacernos de aquellas variables que puedan resultar redundantes. Por ejemplo, en el modelo de predicción de cáncer en humanos teniendo variables como la capacidad pulmonar o cardiovascular del paciente, quizás no sea necesario saber si es fumador o no, ya que

puede que esta última esté, de una forma y otra, ya recogida en las dos primeras. (Lafuente, 2018)

Análisis de Componentes Principales:

"El análisis de componentes principales se utiliza para identificar tendencias en un conjunto de datos amplio, reduciéndolo a sus componentes principales." (numerenutr.org, n.d.). Los componentes principales o ejes son aquellas variables dentro del conjunto de datos original, que son capaces de explicar al resto, es decir, son las variables que más información contienen o que más varianza explican. Estos componentes principales se ordenan en función de la varianza del conjunto de datos que expliquen. De esta forma, los primeros ejes o componentes principales describen la mayor parte de la varianza de los datos, es decir, contienen aspectos más significativos de la información que otros componentes. Si los primeros componentes explican una gran cantidad de la varianza, podremos ignorar aquellos componentes principales más bajos. En caso contrario, tendremos que continuar generando componentes principales o ejes hasta que la mayoría de la varianza quede explicada. Esto se entiende mejor con el siguiente ejemplo:



**Fuente - Fundamentos en Business Analytics: Analítica Predictica (Monfort Vinueasa, n.d.).*

En la imagen observamos tres casos diferentes. En todos los casos el conjunto de datos se separa en dos componentes principales. Empezando por la izquierda, observamos que en el primer caso el primer componente explica el 50% de la varianza, lo que significa

que el segundo eje no se puede ignorar, ya que el primer componente por sí mismo no explica lo suficiente sobre el modelo. En el segundo caso, el primer componente explica el 65% de la varianza; aquí se podría considerar si ignorar o no el segundo eje, pues el primer eje explica la mayoría del modelo. En el tercer ejemplo, el primer componente explica el 90% de la varianza, por lo que el primer eje resume prácticamente la totalidad del modelo, siendo innecesario acudir al segundo componente principal. Cabe señalar, que a mayor correlación entre los datos los primeros componentes serán capaces de describir más varianza. De esta forma, podemos asumir que en el ejemplo 3 los datos se encuentran más correlacionados que en el ejemplo 2 y estos a su vez se encuentran más correlacionados que los del ejemplo 1.

Un ejemplo que muestra muy bien la utilidad del análisis de componentes principales es el siguiente: Si quisiéramos realizar un análisis socio económico de varios países, se tendrían que analizar varios indicadores como el PIB total del país, el PIB per cápita, tasa de desempleo, índice de ruralidad, etc. Aplicando un análisis de componentes principales, se podría encontrar aquellos indicadores que son capaces de explicar la mayoría del modelo, ahorrando una gran cantidad de tiempo. Así, por ejemplo, si se demuestra que el PIB total del país y el índice de ruralidad describen la mayoría de la varianza del modelo, significa que las variables restantes se encuentran muy correlacionadas con estas. Por lo tanto, si el PIB total o el índice de ruralidad varían, las demás variables lo harán de igual manera, lo que nos permite desecharlas.

4.4.3. Reglas de Asociación

El propósito de las reglas de asociación es encontrar relaciones de asociación en grandes conjuntos de datos. Cuando buscamos las reglas de asociación, hay que definir algunos parámetros:

- Soporte: Porcentaje del conjunto de datos que contiene un conjunto de elementos específico, es decir, el número de veces que aparece un *ítem*. Un soporte alto indica que un producto aparezca un porcentaje alto de veces en una tienda, lo que significa que a más variedad de producto menor soporte. Establecimientos como grandes almacenes o supermercados suelen tener un soporte bajo, mientras que tiendas con productos artesanales tienden a tener un soporte alto.

- Confianza: Porcentaje de éxito de la regla (A $\Rightarrow$ B), es decir, significa que la regla es cierta al menos para ese valor de confianza. La regla tiene que seguir la ecuación de probabilidad condicionada: la probabilidad de que ocurra el suceso B si ha ocurrido el suceso A.

- *Lift* de la regla: Indica el nivel de dependencia de los sucesos. Un valor de $Lift > 1$ indica dependencia positva entre los eventos, es decir será más probable que suceda B habiendo sucedido A primero. Un valor de $Lift = 1$ indica que los eventos son independientes, es decir será igual de probable que suceda B habiendo sucedido A primero. Un valor de $Lift < 1$ indica una dependencia negativa entre sucesos, es decir será menos probable que suceda B habiendo sucedido A primero.

Por ejemplo, en un supermercado encontramos la siguiente regla con un soporte del 40%, una confianza del 80% y un valor de *lift* de 2.

{paquete de frutos secos, Coca-Cola} => {patatas fritas}

Esto implica, que un 40% de los casos (ventas) incluían frutos secos y Coca-Cola, y en los casos en que se vendieron frutos secos y Coca-Cola, también se vendieron patatas fritas al menos en un 80% de las veces. Por último, al ser el valor de lift > 1 sabemos que el antecedente de la regla, comprar frutos secos y Coca-Cola, aumenta la probabilidad de ocurrencia del consecuente de la regla, comprar patatas fritas. Por ello, podemos concluir que la probabilidad de comprar patatas fritas habiendo comprado primero frutos secos y Coca-Cola es mayor que la probabilidad a priori de comprar patatas fritas.

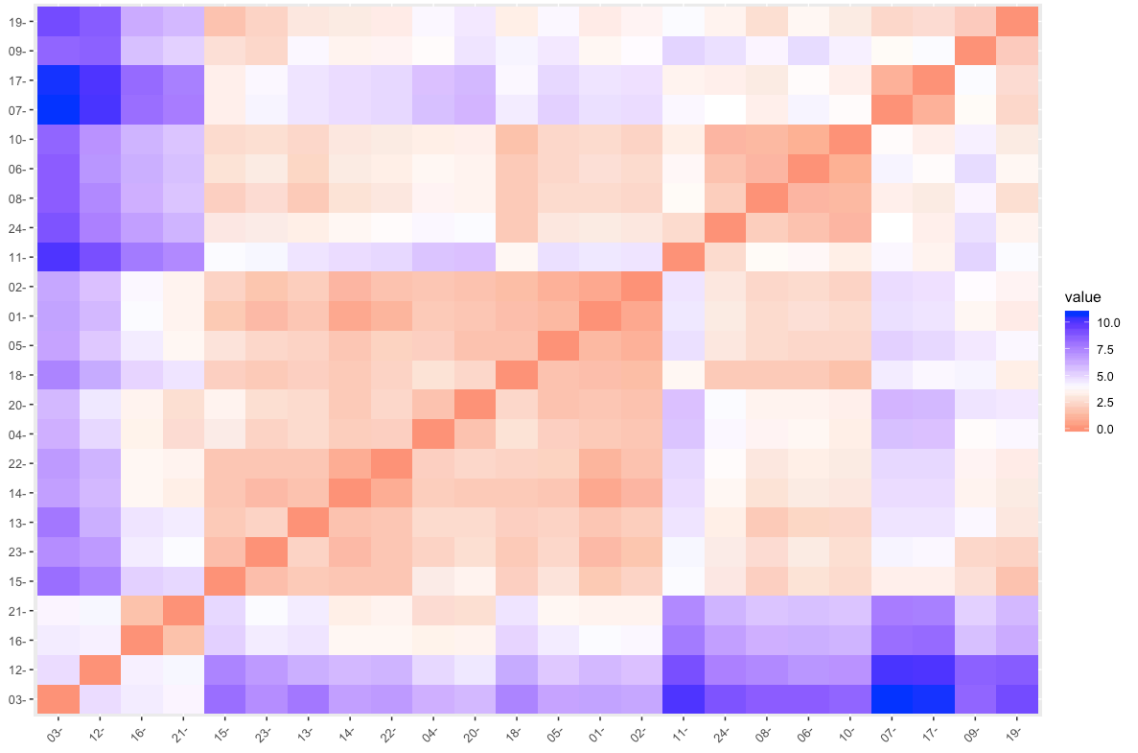
5. EJEMPLOS DE CLÚSTER EN LA HERRAMIENTA R

El lenguaje R es ampliamente utilizado entre los estadísticos e ingenieros de software para el desarrollo de softwares estadísticos y análisis de datos. R ofrece una amplia variedad de funcionalidades estadísticas como modelado lineal y no lineal, clasificación o agrupamiento. (Espino Timón, 2017). No obstante, antes de comenzar con cada ejemplo hace falta resaltar tres aspectos:

Primero, para llevar a cabo nuestro ejemplo de clusterización, primero con K-Medios y luego con un Clustering Jerárquico, utilizaremos las funciones "library" e "install.packages" del programa RStudio para instalar los siguientes paquetes: [Dplyr, factoextra, FactoMineR, arules, arulesViz, cluster.datasets, DataExplorer, ggplot2, cluster].

Segundo, El conjunto de datos empleado para ambos métodos contiene información sobre la rentabilidad como porcentaje del capital social de 24 sectores de la economía estadounidenses entre 1959 y 1968. La base de datos contiene 12 variables como puede apreciarse en el anexo: dos categóricas; código y nombre de cada sector y diez numéricas; la rentabilidad como porcentaje del capital social de cada sector en cada año.

Tercero, como explica Joaquín Amat antes de aplicar algún método de agrupación o clúster primero es "conveniente evaluar si dentro del conjunto de datos a emplear hay indicios de que realmente existe algún tipo de agrupación." A este proceso se le conoce como *assesing clustering tendency* y puede llevarse a cabo mediante test estadísticos o de forma visual (*visual assesment of cluster tendency*). (Amat Rodrigo, 2017). Esta última será la forma en la que evaluaremos si los datos muestran indicios de agrupación:



**Fuente – Elaboración propia en la herramienta RStudio.*

El método de *VAT* confirma que en el conjunto de datos sí hay una estructura de grupos. Con este método, igualmente, podemos empezar a intuir los clústeres que se van a ir formando. El color rojo indica una distancia corta entre componentes, mientras que el color azul indica una distancia más larga. La diagonal tiene una distancia de 0, ya que es cada componente u observación consigo misma. Así, podemos intuir que el sector de productos químicos y similares (21) y el sector de madera y productos de madera (11) no formarán parte del mismo clúster, pero el sector de productos químicos y similares (21) y el sector de la fabricación de tabaco (16) sí.

5.1. EJEMPLO DE K-MEDIOS

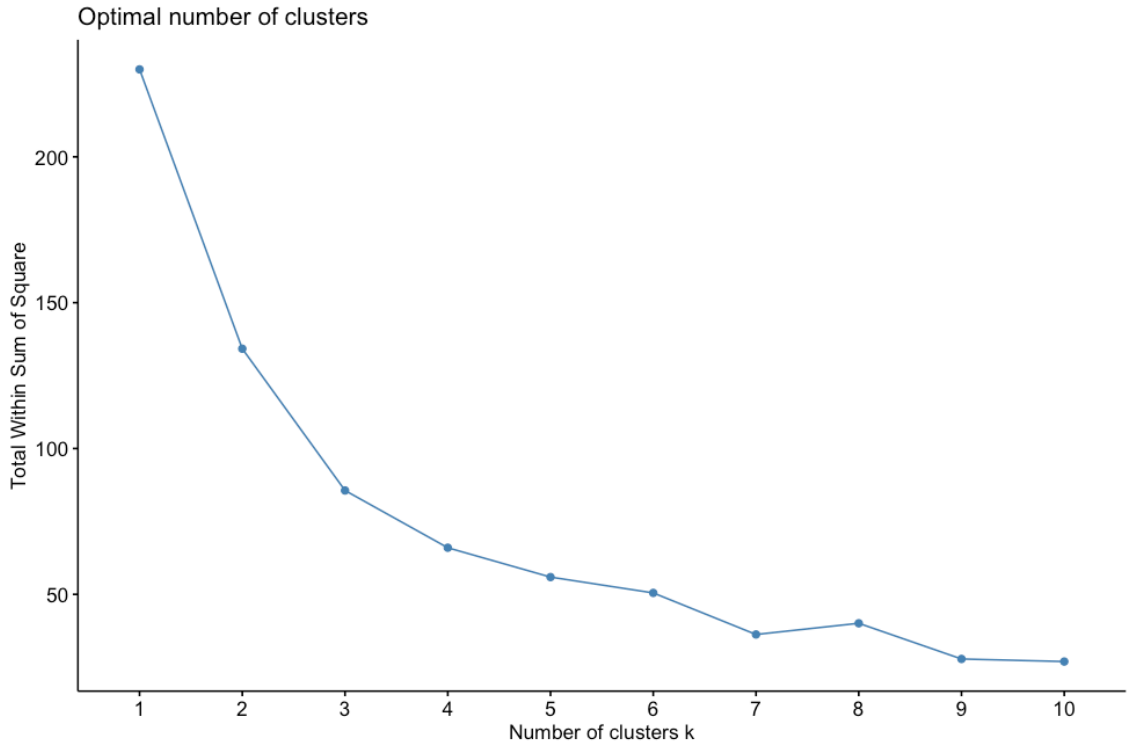
1. Determinamos el valor óptimo de K:

"Determinar el número óptimo de clústeres es uno de los pasos más complicados a la hora de aplicar métodos de agrupación, ya que no existe una única forma de averiguar el número adecuado de clústeres." (Amata Rodrigo, 2017). Se trata de un proceso bastante

subjetivo que depende en gran medida del método empleado. A pesar de ello, se han desarrollado varias estrategias que ayudan en el proceso. (Amat Rodrigo, 2017).

Así, para determinar el valor óptimo de clústeres podemos, o bien utilizar el método *Elbow* o el método *Silhouette*:

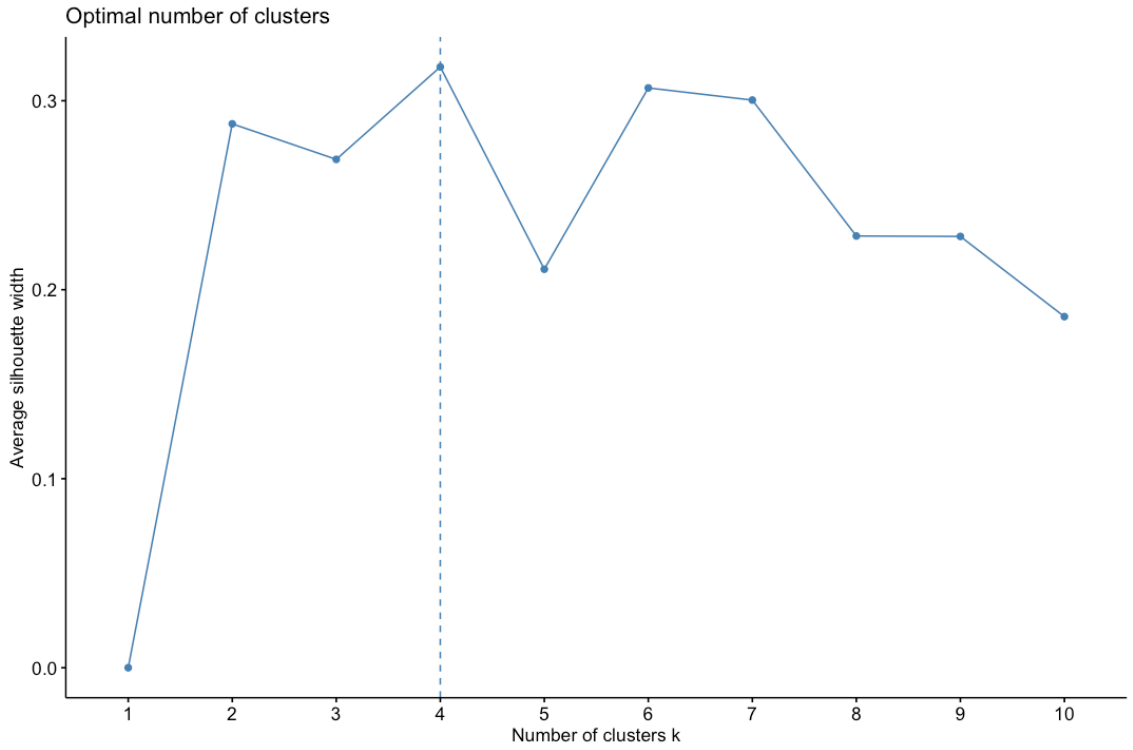
- El método *Elbow* considera como número óptimo de agrupaciones aquel que minimiza la varianza total intra-clústeres o la suma total de cuadrados entre clústeres (*total inter-cluster sum of squares*). "El método *Elbow* calcula la varianza total intra-clústeres en función del número de agrupaciones y escoge como óptimo aquel valor a partir del cual añadir más clústeres apenas consigue mejoría." (Amat Rodrigo, 2017):



**Fuente – Elaboración propia en la herramienta RStudio.*

La curva indica que a partir del tercer clúster la mejora es mínima. Es decir, el método *Elbow* nos indica que el número óptimo de clústeres son 3.

- El método *Shillouette*, considerar el número óptimo de agrupaciones aquel que maximiza la media de los *silhouette coeficient* o índices silueta, en lugar de aquel que minimiza la varianza total intra-clústeres. (Amat Rodrigo, 2017):



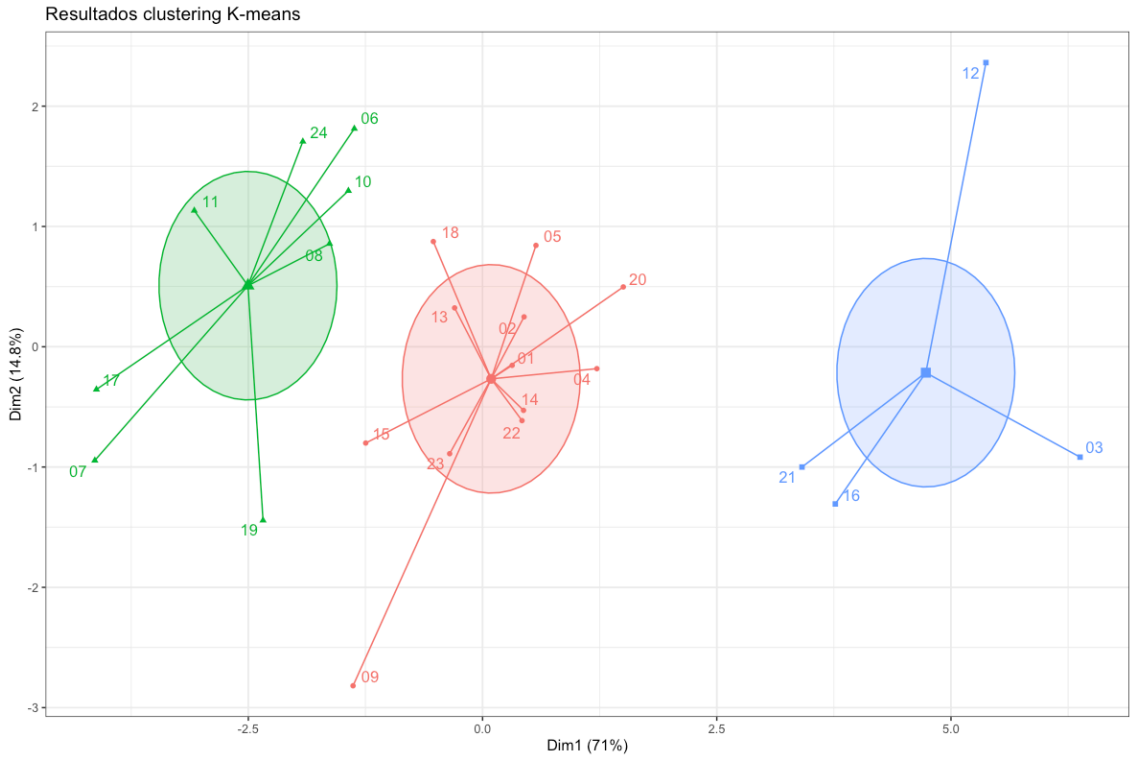
**Fuente – Elaboración propia en la herramienta RStudio.*

El método *Silhoutte* nos indica que el valor que maximiza la media de coeficientes silueta es el 4. Por lo tanto, a diferencia del método *Elbow*, el método *Silhoutte* considera que el número óptimo de clústeres es 4.

2. Se crean los clústeres:

El algoritmo seleccionará aleatoriamente los centroides de cada grupo. Entendiendo por centroide "la posición definida por la media de cada una de las dimensiones (variables) de las observaciones que forman el clúster, esto puede entenderse como el centro de gravedad de cada uno de los clústeres." (Amat Rodrigo, 2017). Cada punto se asignará al centroide más cercano utilizándose la distancia euclídea como medida por defecto.

Siguiendo el criterio del método *Elbow*, obtenemos tres clústeres bien diferenciados:



**Fuente – Elaboración propia en la herramienta RStudio.*

De izquierda a derecha:

- **Grupo 1**: Productos de metal fabricados (06), Industria primaria de hierro y acero (07), Industria primaria de metales no ferrosos (08), Muebles y accesorios (10), Madera y productos de madera, excepto los muebles (11), Productos de fábrica textil (17), Productos de papel y similares (19), Cuero y productos de cuero (24).

- **Grupo 2**: Todas las corporaciones manufactureras, excepto los periódicos (01), Durabilidad total (02), Maquinaria, equipo y suministros eléctricos (04), Maquinaria excepto la eléctrica (05), Piedra, arcilla y productos de vidrio (09), Manufactura diversa, incluida la artillería (13), Total no duradero (14), Alimentos y productos afines (15), Ropa y productos relacionados (18), Imprenta y publicación, excepto los periódicos (20), Refinamiento de petróleo (22), Caucho y productos plásticos diversos (23).

- **Grupo 3**: Vehículos de motor y equipo (03), Instrumentos y productos conexos (12), Fabricación de tabaco (16), Productos químicos y similares (21).

5.2. EJEMPLO DE CLUSTERING JERÁRQUICO

Como se ha mencionado anteriormente, existen dos estrategias de agrupación jerárquica: aglomerativa y divisiva. Sin embargo, para decidir qué grupos deben combinarse, en el caso de los clústeres aglomerativos, o dónde debe dividirse un grupo, en el caso de los clústeres divisivos, primero es necesario definir la métrica y el criterio de enlace (*linkage*) o método de agrupación de clústeres que queremos emplear.

En cuanto a la métrica, para no complicar el ejemplo e introducir conceptos matemáticos avanzados sobre otro tipo de distancias entre observaciones, emplearemos la distancia euclídea como métrica para la similitud entre observaciones; al igual que en el ejemplo anterior con el algoritmo K-Medios. Así, la métrica de la distancia euclídea estipula que cuanto más próximas se encuentren dos observaciones entre sí, más similares serán.

Por otro lado, el concepto de criterio de enlace sirve para definir con que método se va a cuantificar la similitud entre dos clústeres. Es decir, el concepto de distancia entre pares de observaciones se tiene que extender para que sea aplicable a pares de grupos, cada uno formado por varias observaciones. A este proceso se le conoce como *linkage*. (Amat Rodrigo, 2017). En resumen, la métrica o medida de distancia sirve para cuantificar la similitud entre observaciones y el criterio de enlace (*linkage*) sirve para determinar la similitud entre grupos de observaciones o clústeres. Es importante subrayara, que el criterio de enlace solo es necesario definirlo para clústeres aglomerativos no divisivos. (Amat Rodrigo, 2017). Los métodos de agrupación o criterios de enlace más comunes son:

- Enlace completo (*complete linkage*): También conocido como método del vecino más lejano. Se calcula la distancia entre todos los posibles pares formados por una observación del clúster A y una del clúster B. “La mayor de todas ellas se selecciona como la distancia entre los dos clústeres.” (Amat Rodrigo, 2017). (Anexo 2.1).

- Enlace simple (*single linkage*): También conocido como método del vecino más próximo. “Se calcula la distancia entre todos los posibles pares formados por una observación del clúster A y una del clúster B. La menor de todas ellas se

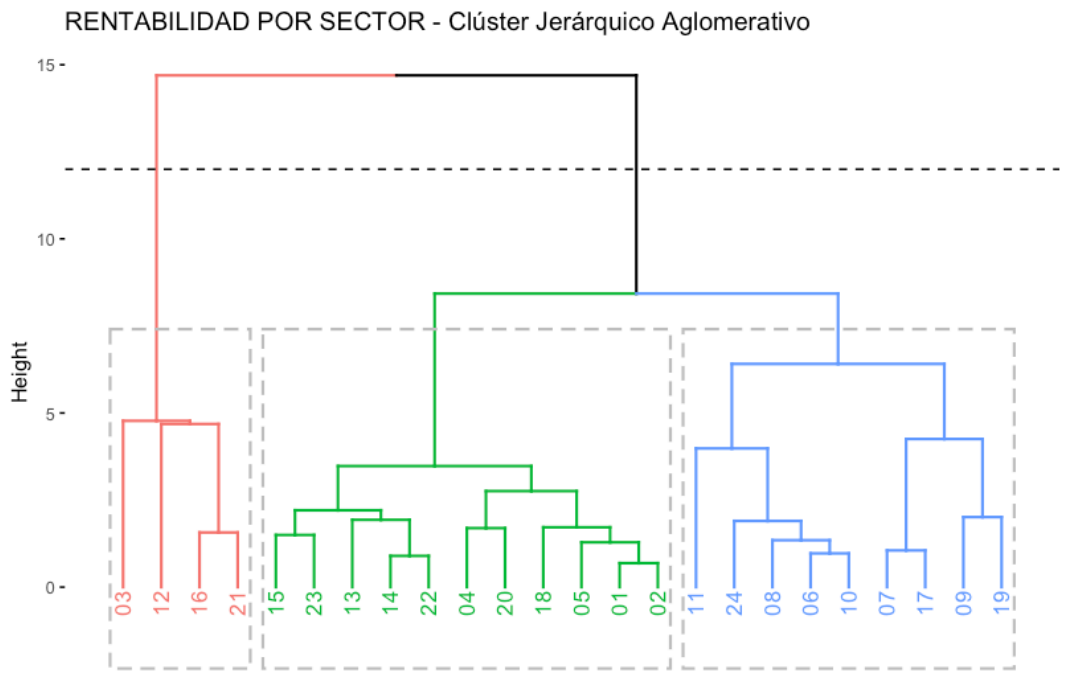
selecciona como la distancia entre los dos clústeres." (Amat Rodrigo, 2017). (Anexo 2.2).

- Enlace promedio (*average linkage*): "Se calcula la distancia entre todos los posibles pares formados por una observación del clúster A y una del clúster B. El valor promedio de todas ellas se selecciona como la distancia entre los dos clústeres." (Amat Rodrigo, 2017). (Anexo 2.3).

- Enlace centroide (*centroid linkage*): "Se calcula el centroide de cada uno de los clústeres y se selecciona como la distancia entre los dos clústeres." (Amat Rodrigo, 2017). (Anexo 2.4).

- Enlace Ward (*Ward linkage*): Se busca minimizar la varianza que intra-clúster. Se identifican aquellos clústeres cuya fusión conlleva menor incremento de la varianza total intra-clúster. (Amat Rodrigo, 2017).

A continuación, aplicaremos ambas estrategias a nuestro conjunto de datos para poder visualizar distintos tipos de clústeres. De esta forma, al hacer un clúster aglomerativo con el método Ward de mínima varianza, el dendograma que obtenemos es:



**Fuente – Elaboración propia en la herramienta RStudio.*

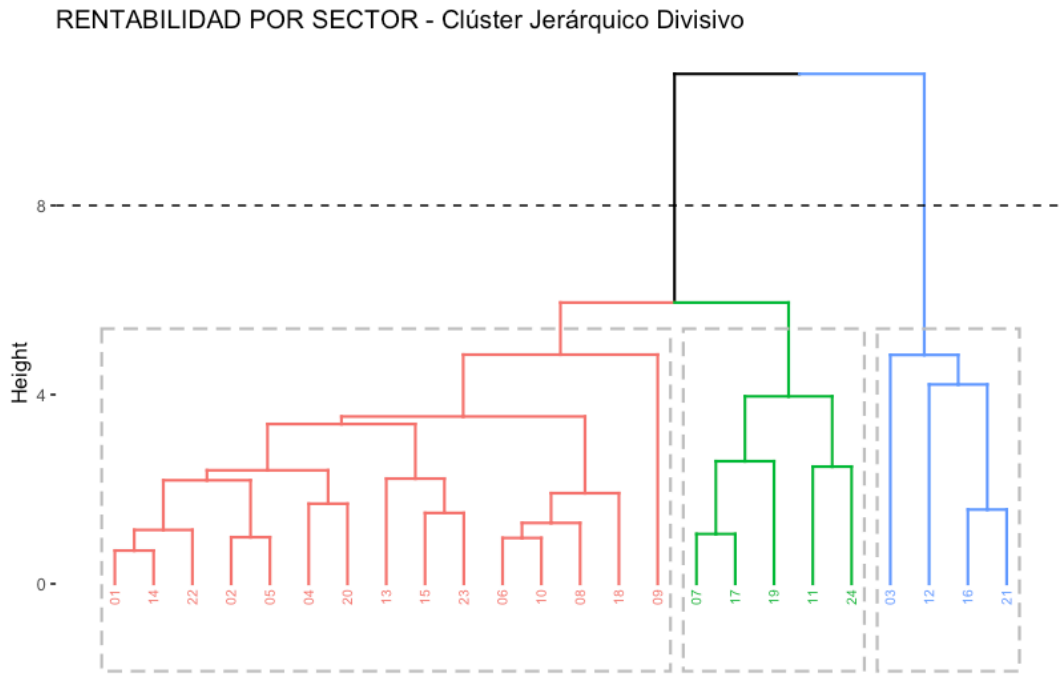
El dendograma nos indica que las empresas se pueden agrupar hasta en 2 clústeres como mínimo. Esto se puede visualizar con ayuda de la línea discontinua de arriba. Sin embargo, hemos decidido crear 5 clústeres que aparecen dentro de 5 cajas formadas por líneas discontinuas, para poder comparar con el método K-Medios.

- **Grupo 1**: Vehículos de motor y equipo (03), Instrumentos y productos conexos (12), Fabricación de tabaco (16), Productos químicos y similares (21).

- **Grupo 2**: Todas las corporaciones manufactureras, excepto los periódicos (01), Durabilidad total (02), Maquinaria, equipo y suministros eléctricos (04), Maquinaria excepto la eléctrica (05), Manufactura diversa, incluida la artillería (13), Total no duradero (14), Alimentos y productos afines (15), Ropa y productos relacionados (18), Imprenta y publicación, excepto los periódicos (20), Refinamiento de petróleo (22), Caucho y productos plásticos diversos (23).

- **Grupo 3**: Productos de metal fabricados (06), Industria primaria de hierro y acero (07), Industria primaria de metales no ferrosos (08), Piedra, arcilla y productos de vidrio (09), Muebles y accesorios (10), Madera y productos de madera, excepto los muebles (11), Productos de fábrica textil (17), Productos de papel y similares (19), Cuero y productos de cuero (24).

Pasando a la estrategia de clúster divisivo, no es necesario definir un criterio de enlace como en el clúster aglomerativo. Al hacer un clúster jerárquico divisivo, el dendograma que obtenemos es el siguiente:



**Fuente – Elaboración propia en la herramienta RStudio.*

Al igual que en el dendograma anterior, este nos indica que las empresas se pueden agrupar hasta en 2 clústeres como mínimo. Igualmente, hemos decidido crear 3 clústeres para poder comparar. En este caso:

- **Grupo 1**: Todas las corporaciones manufactureras, excepto los periódicos (01), Durabilidad total (02), Maquinaria, equipo y suministros eléctricos (04), Maquinaria excepto la eléctrica (05), Productos de metal fabricados (06), Industria primaria de metales no ferrosos (08), Piedra, arcilla y productos de vidrio (09), Muebles y accesorios (10), Manufactura diversa, incluida la artillería (13), Total no duradero (14), Alimentos y productos afines (15), Ropa y productos relacionados (18), Imprenta y publicación, excepto los periódicos (20), Refinamiento de petróleo (22), Caucho y productos plásticos diversos (23).

- **Grupo 2**: Industria primaria de hierro y acero (07), Madera y productos de madera, excepto los muebles (11), Productos de fábrica textil (17), Productos de papel y similares (19), Cuero y productos de cuero (24).

- **Grupo 3**: Vehículos de motor y equipo (03), Instrumentos y productos conexos (12), Fabricación de tabaco (16), Productos químicos y similares (21).

Extraemos las siguientes observaciones de los resultados obtenidos:

- Observamos que los resultados obtenidos por cada método son muy similares entre sí. Las diferencias entre clústeres son mínimas. Por ejemplo, Los métodos K-Medios y de Clustering Jerárquico Aglomerativo modelo Ward, ofrecen resultados prácticamente idénticos. La única diferencia es el sector Piedra, arcilla y productos de vidrio (09), el cual pasa a formar parte de otro clúster. El método de Clustering Jerárquico Divisivo es el que ofrece datos más dispares, con respecto a los otros dos modelos. Con este método, los componentes 06, 08, y 10 forman parte de otro clúster.

- En los tres métodos se termina formando un clúster compuesto por las observaciones 03, 12, 16 y 21. Cabe destacar que el proceso de formación de este clúster en ambos métodos jerárquicos, aglomerativo modelo Ward y divisivo, es idéntico. El primer nodo que se forma es entre el sector de productos químicos y similares (16) con el sector de fabricación (16) de tabaco. El siguiente sector en anexionarse es el de instrumentos y productos conexos (12), siendo el sector de vehículos de motor y equipo (03) el último. Esto nos indica que la similitud entre los sectores de productos químicos y similares y de fabricación de tabaco guardan una mayor similitud entre sí que con los sectores de instrumentos y productos conexos y vehículos de motor y equipos. Esto se refuerza con el gráfico que obtenemos con el modelo K-Medios, donde se puede visualizar la mayor proximidad entre los componentes 16 y 21.

Aunque mínimas, existen diferencias entre métodos. Sin embargo, no se puede determinar que uno sea mejor que otro, ya que depende del caso de estudio en cuestión.

6. CONCLUSIONES

La utilidad del análisis predictivo es algo que no ha pasado desapercibido por gobiernos, grandes corporaciones e incluso universidades. Su popularidad ha crecido tanto que se ha convertido en parte integral del "Business Inteligence" o Inteligencia de negocio, encontrándose a disposición de todo tipo de empresas, personas y organizaciones.

Mi intención con este trabajo ha sido la de explicar y destacar los aspectos que hacen de los modelos predictivos una ventaja competitiva para las empresas y organizaciones, además de describir las herramientas y técnicas que el Machine Learning nos proporciona para crearlos. Para abordar un tema tan amplio, decidí empezar desde los más general, describir que es el Big Data, para luego ir concretando cada vez más en los temas de Análisis Predictivo, sus distintos modelos, el Machine Learning y sus algoritmos, terminando con un ejemplo del algoritmo clúster.

Si algo debe de quedar claro es que el requerimiento principal para poder crear un modelo predictivo es la existencia de un conjunto lo suficientemente amplio de datos como para permitir detectar en ellos patrones que permitan formular reglas capaces de anticipar previsiones. Igualmente, la capacidad de almacenar y gestionar esos datos se convierte en un requerimiento fundamental para la correcta elaboración de modelos predictivos. En otras palabras, no podemos hablar de Análisis Predictivo sin mencionar el Big Data, pues este representa un pilar esencial del proceso.

Un aspecto con el que no contaba pero que me ha llamado mucho la atención y encuentro que merece la pena comentar, es el concepto de la anonimidad de los datos. Uno de los aspectos más negativos, podríamos decir, sobre la revolución digital y la aparición constante de información, es la perdida de privacidad a todos los niveles. Muchas veces no somos conscientes de lo expuestos que estamos al acceder a ciertas páginas web, comprar online o comentar en redes sociales. Lo que conseguimos a través de estas acciones es facilitar nuestros datos y generar una huella rastreable en internet. Puede que algunos aspectos como que la gente conozca nuestros gustos u opiniones sobre ciertos temas no nos parezcan tan alarmantes. Sin embargo, a medida que crece y se expande esta revolución digital y nuestra huella se hace más grande, se vuelve más fácil el acceso

a otro tipo de información más confidencial como nuestra dirección, cuenta bancaria o número de la seguridad social.

Tampoco nos debe de sorprender que la parte de profundización del Análisis Predictivo y Machine Learning haya terminado siendo la más amplia. Ambos aspectos encierran una enorme complejidad sobre la que hay mucho que profundizar. Como he dicho antes, en este trabajo se ha intentado destacar solamente los aspectos más importantes y de mayor utilidad para crear modelos predictivos, esto representa una pequeña parte del universo de posibilidades que ofrecen el Análisis Predictivo y el Machine Learning.

Finalmente, en cuanto a la parte del ejemplo del método clúster el resultado obtenido con los dos algoritmos empleados, propios del método, muestran que la herramienta RStudio cumple con las prestaciones exigibles para la creación de modelos predictivos. He de añadir también, que he disfrutado mucho buscando una base de datos apropiada y creando los modelos con la herramienta RStudio.

7. BIBLIOGRAFÍA

Agencia Española de Protección de Datos. (n.d.). La K-anonimidad como medida de la privacidad. Disponible en: https://www.aepd.es/sites/default/files/2019-09/nota-tecnica-kanonimidad.pdf

Amat Rodrigo, J. (septiembre, 2017): Clustering y heatmaps: aprendizaje no supervisado. Disponible en: https://rpubs.com/Joaquin_AR/310338

análisiscientífico.com. (22 de junio, 2017). Análisis predictivo, descriptivo y prescriptivo. Disponible en: https://www.analisiscientifico.com/single-post/Análisis-predictivo-descriptivo-y-prescriptivo

Arroyo-Hernández, J. (enero 2016). Métodos de reducción de dimensionalidad: Análisis comparativo de los métodos APC, ACPP y ACPK.

Barrueta Meza, R. A. & Castillo Villarreal, E. J. P. (6 de diciembre, 2018). Modelos de análisis predictivo para determinar clientes con tendencias a la deserción en bancos peruanos. Disponible en: https://repositorioacademico.upc.edu.pe/bitstream/handle/10757/626023/Barrueta%20MR.pdf?sequence=1&isAllowed=y

de la Fuente Crespo, L. Análisis Discriminante. Disponible en: http://www.estadistica.net/Master-Econometria/Analisis_Discriminante.pdf

es.mathworks.com. (MWa). (n.d.). Análisis Predictivo, tres cosas que es necesario saber. Disponible en: https://la.mathworks.com/discovery/predictive-analytics.html

es.mathworks.com. (MWb). (n.d.). Técnica de Machine Learning para crear modelos predictivos a partir de datos de entrada y respuesta conocidos. Disponible en: https://la.mathworks.com/discovery/supervised-learning.html

Espino Timón, C. (16 de enero, 2017). Análisis predictivo: técnicas y modelos utilizados y aplicaciones del mismo – herramienta *Open Source* que permite su uso. Disponible en: http://openaccess.uoc.edu/webapps/o2/bitstream/10609/59565/6/caresptimTFG0117memm%C3%B2ria.pdf

González, L. (10 agosto 2018). Todo sobre aprendizaje no supervisado en Machine Learning. Disponibles en: https://ligdigonzalez.com/todo-sobre-aprendizaje-no-supervisado-en-machine-learning/

halweb.uc3m.es (n.d.). Capítulo 18, análisis de regresión lineal, el procedimiento Regresión lineal. Disponible en: http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/GuiaSPSS/18reglin.pdf

Hartigan, J. A. (1975). Clustering Algorithms, John Wiley, New York.

IBM.com. (n.d.). El modelo de redes neuronales. Disponible en: https://www.ibm.com/support/knowledgecenter/es/SS3RA7_sub/modeler_mainhelp_client_ddita/components/neuralnet/neuralnet_model.html

IBM.com. (2012). ¿Qué es Big Data? Disponible en: https://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/index.html

iic.uam.es. (2016). Las 7 V del Big Data: Características más importantes. Disponible en: https://www.iic.uam.es/innovacion/big-data-caracteristicas-mas-importantes-7-v/

iep.edu.es. (n.d.). 5 tipos de datos en el Big Data. Disponible en: https://www.iep.edu.es/5-tipos-de-datos-en-el-big-data/

Juan F. Cía. (2015). El ranking de las mejores soluciones de análisis predictivo para empresas. Disponible en: https://bbvaopen4u.com/es/actualidad/el-ranking-de-las-mejores-soluciones-de-analisis-predictivo-para-empresas

Ladrero, I. (2018). Big Data en el sector financiero: 10 casos de uso. Disponible en: https://www.baoss.es/big-data-sector-financiero-10-casos-uso/

Lafuente, A. (22 de mayo 2018) Reducción de la dimensionalidad (o por qué más datos no siempre es mejor). Disponible en: https://aukera.es/blog/reduccion-dimensionalidad/

masterbigdataucm.com (n.d.) ¿Qué es Big Data? Facultad de Estudios Estadísticos, Universidad Complutense de Madrid. Disponible en: https://www.masterbigdataucm.com/que-es-big-data/

numerentur.org. (n.d.) PCA-KPCA. Disponible en: http://numerentur.org/pca-kpca/

Parra, F. (25 de enero, 2019). Estadística y Machine Learning con R. Disponible en: https://bookdown.org/content/2274/metodos-de-clasificacion.html

Powerdata.es. (n.d.). Big Data: ¿En qué consiste? Su importancia, desafíos y gobernabilidad. Disponible en: https://www.powerdata.es/big-data

Román, V. (12 de junio, 2019). Aprendizaje no supervisado Cluster. Disponible en: https://medium.com/datos-y-ciencia/aprendizaje-no-supervisado-en-machine-learning-agrupación-bb8f25813edc

Rouse, M. (n.d.) Analítica predictiva o análisis predictivo. Disponible en: https://searchdatacenter.techtarget.com/es/definicion/Analitica-predictiva-o-analisis-predictivo

Ruíz, S. (20 de julio, 2017). El algoritmo K-NN y su importancia en el modelado de datos. Disponible en: https://www.analiticaweb.es/algoritmo-knn-modelado-datos/

Shalev-Shwartz, S. & Ben-David, S. (2014). Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press. Disponible en: https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/understanding-machine-learning-theory-algorithms.pdf

Shimada, T. & López, F. Analítica Predictiva: cómo convertir la información en ventaja competitiva. Disponible en: http://www.il- latam.com/images/articulos/articulo-revista-109-como-convertir-la-informacion-en-ventaja-competitiva.pdf

Siegel, E. (2013). Predictive Analytics. The power to predict who will click, buy, lie, or die.

Skinner, B.F. (1947). Superstition in the Pigeon. Indiana University. Journal of Experimental Psychology, 38, 168-172.

support.minitab.com. (n.d.). Explicación Regresión no lineal. Disponible en: https://support.minitab.com/es-mx/minitab/18/help-and-how-to/modeling-statistics/regression/supporting-topics/nonlinear-regression/understanding-nonlinear-regression/

8. ANEXOS

Anexo 1: El cuadro contiene las ganancias como porcentaje del capital social de diversos sectores económicos de la economía estadounidense para los años 1959 a 1968.

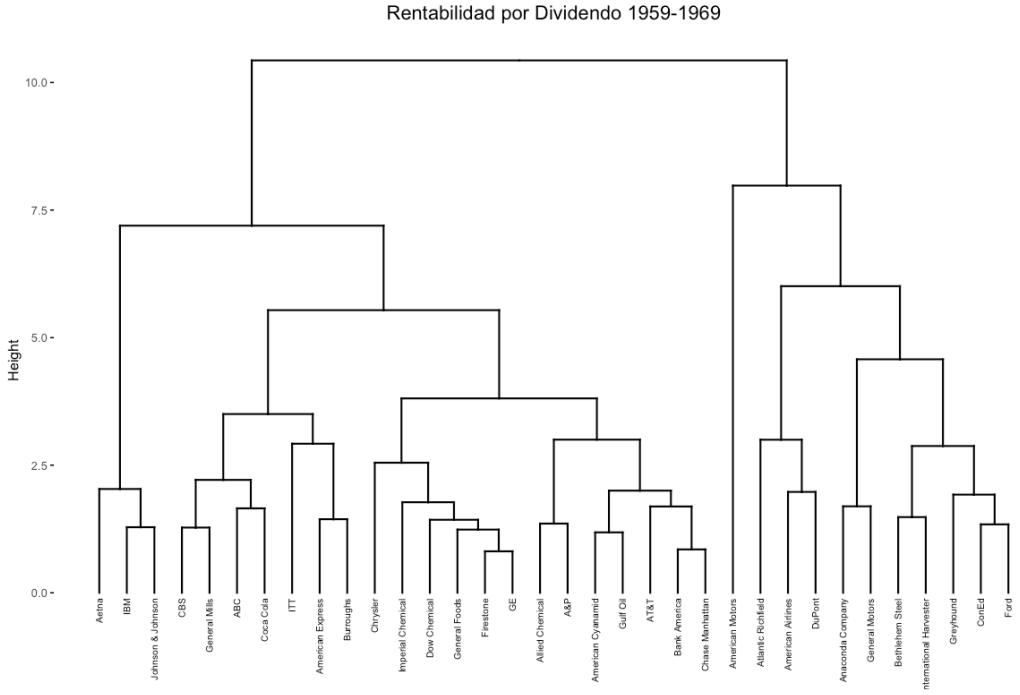
	code	sector	y.1959	y.1960	y.1961	y.1962	y.1963	y.1964	y.1965	y.1966	y.1967	y.1968
1	1	All manufacturing corporations except newspapers	10	9	9	10	10	12	13	13	12	12
2	2	Total durable	10	9	8	10	10	12	14	14	12	12
3	3	Motor vehicles and equipment	14	14	11	16	17	17	20	16	12	15
4	4	Electrical machinery, equipment and supplies	13	10	9	10	10	11	14	15	13	12
5	5	Machinery except for electrical	10	8	8	9	10	13	14	15	13	12
6	6	Fabricated metal products	8	6	6	8	8	10	13	15	13	12
7	7	Primary iron and steel industry	8	7	6	5	7	9	10	10	8	8
8	8	Primary non-ferrous metal industry	8	7	7	8	8	10	12	15	11	11
9	9	Stone, clay and glass products	13	10	9	9	9	10	10	10	8	9
10	10	Furniture and fixtures	9	7	5	8	8	10	13	14	12	12
11	11	Lumber and wood products except furniture	9	4	4	6	8	10	10	10	9	15
12	12	Instruments and related products	13	12	11	12	12	14	18	21	18	17
13	13	Miscellaneous manufacture, including ordnance	9	9	10	9	9	10	11	15	13	12
14	14	Total nondurable	10	10	10	10	10	12	12	13	12	12
15	15	Food and kindred products	9	9	9	9	9	10	11	11	11	11
16	16	Tobacco manufacture	13	13	14	13	13	13	14	14	14	14
17	17	Textile mill products	8	6	5	6	6	9	11	10	8	9
18	18	Apparel and related products	9	8	7	9	8	12	13	13	12	13
19	19	Paper and allied products	10	9	8	8	8	9	9	11	9	10
20	20	Printing and publishing, except newspapers	11	11	9	10	9	13	14	16	13	13
21	21	Chemical and allied products	14	12	12	12	12	14	15	15	13	13
22	22	Petroleum refining	10	10	10	10	11	11	12	12	13	12
23	23	Rubber and miscellaneous plastic products	11	9	9	10	9	11	12	12	10	12
24	24	Leather and leather products	9	6	4	7	7	11	12	13	12	13

**Fuente – Clustering Algorithms (Hartigan, 1975)*

.

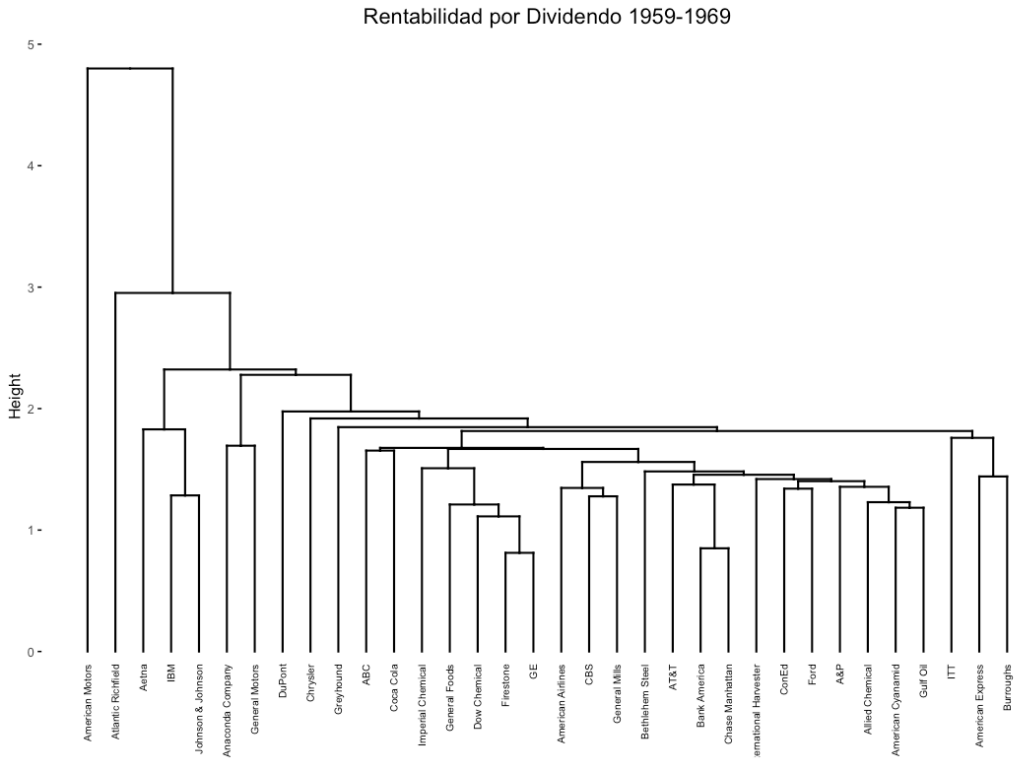
Anexo 2: Ejemplo de criterios de enlace o métodos de clustering jerárquicos.

Anexo 2.1: Enlace Completo o *Complete Linkage*



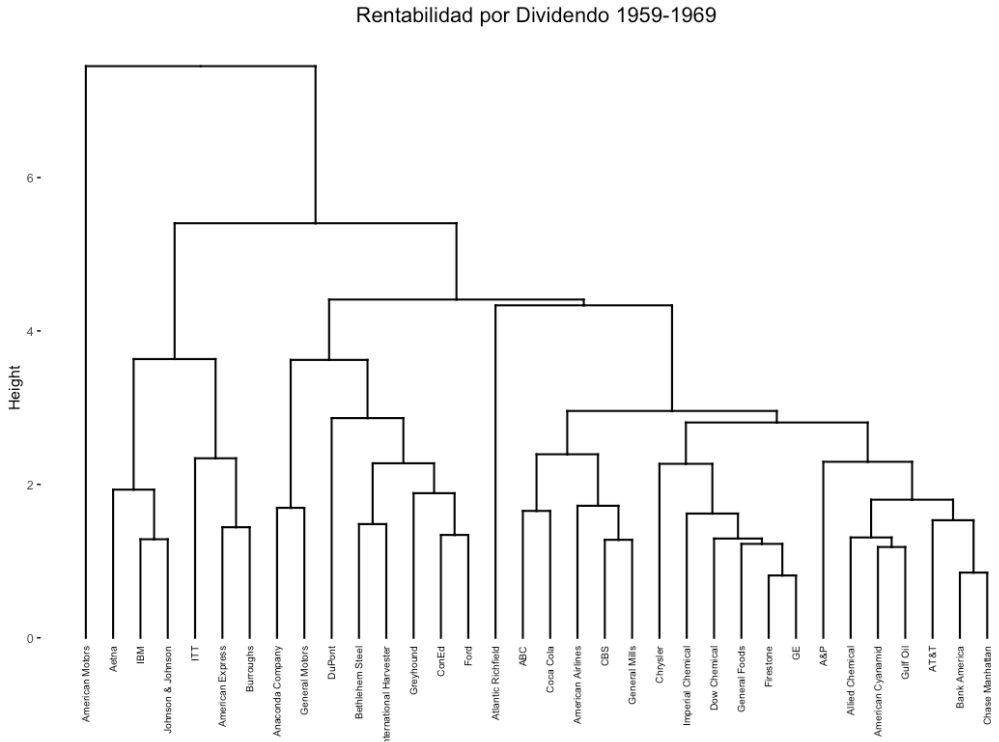
**Fuente – Elaboración propia en la herramienta RStudio.*

Anexo 2.2: Enlace Simple o *Single Linkage*



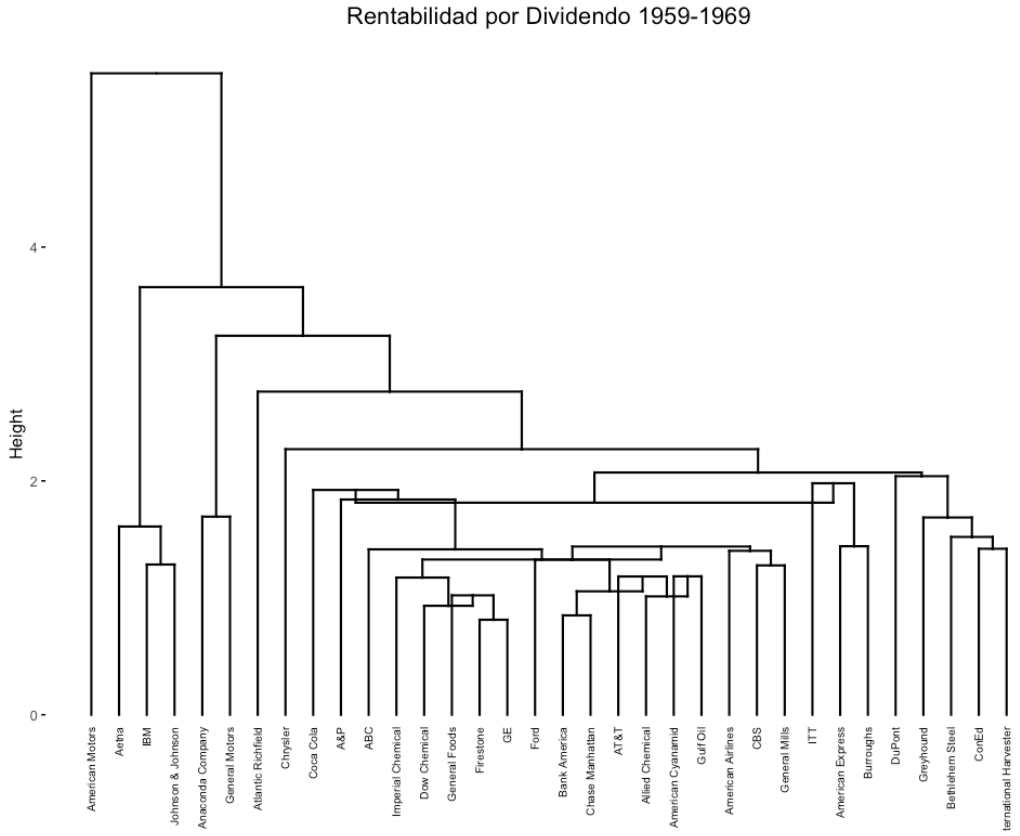
**Fuente – Elaboración propia en la herramienta RStudio.*

Anexo 2.3: Enlace Ponderado o *Average Linkage*



**Fuente – Elaboración propia en la herramienta RStudio.*

Anexo 2.4: Enlace Centroide o *Centroid Linkage*



**Fuente – Elaboración propia en la herramienta RStudio.*

Anexo 3: Lista de Acrónimos.

API: Interfaz de Programación de Aplicaciones. Es un conjunto de definiciones y protocolos que se utiliza para desarrollar e integrar el software de las aplicaciones.

GDPR: Reglamento General de Protección de Datos.

HTML: HyperText Markup Language. Es el lenguaje de marcado para la elaboración de páginas web.

IBM: Empresa conocida como International Business Machines.

IoT: Internet of the Things.

K-NN: K Nearest Neighbours.

NoSQL: "no solo SQL". Es una amplia clase de sistemas de gestión de bases de datos que difieren del modelo clásico de SGBDR (Sistema de Gestión de Bases de Datos Relacionales) en aspectos importantes, siendo el más destacado que no usan SQL como lenguaje principal de consultas.

PC: Principal Component.

PIB: Producto Interior Bruto.

SAS: Empresa conocida como Software y Soluciones Analíticas.

SDC: Statistical Disclousure Control. Es una técnica utilizada en la investigación basada en datos para garantizar que ninguna persona u organización sea identificable a partir de los resultados de un análisis de datos de encuestas o administrativos, o en la divulgación de micro datos.

SNA: Social Network Analysis. Es un área de investigación enfocada en el estudio de las redes sociales.

TI: Tecnología e Información.

UE: Unión Europea.