Estudio comparativo de técnicas de Data Mining para aportar valor en la toma de decisiones

Tutor técnico: Francisco Morteo

Profesora: Dra. Marcela Samela

Alumna: Leda Mariana Jazmín Varela

Trabajo Final de Carrera presentado para obtener el título de Lic. en Gestión de Tecnología informática

(Diciembre, 2022)

Resumen

En el presente trabajo final, se realizó una investigación descriptiva acerca de los procesos, métodos y técnicas correspondientes a la minería de datos que resultan más asiduamente utilizadas a nivel empresarial para apoyar la toma de decisiones de negocios.

Asimismo, en el marco de la investigación se planteó un problema a resolver mediante business intelligence en una entidad bancaria, el objetivo fue descubrir si existen agrupaciones diferentes a las establecidas en la actualidad para medir el desempeño de sus sucursales y, de resultar procedente, reformular el sistema de clasificación al que se someten periódicamente las casas, este planteo se realizó con el propósito de comparar técnicas de agrupamiento sobre un paquete de datos obtenido mediante un proceso KDD, sobre los que se ejecutaron los algoritmos correspondientes a cuatro técnicas representativas de clústering: k-means, AGNES, DBSCAN y fuzzy c-means, correspondientes a agrupamientos de los tipos basados en distancia, en jerarquía, en densidad y difuso, respectivamente.

Luego, mediante el desarrollo de una aplicación de minería de datos en lenguaje R, se entrenó el modelo con las técnicas mencionadas para obtener los diferentes agrupamientos sobre el dataset de sucursales.

Finalmente, haciendo uso de medidas de validación interna de clústeres como el método del codo y el coeficiente de silueta, y de tendencia de agrupación como el índice de Hopkins, se analizaron los resultados sobre la aplicación de los diferentes algoritmos pudiéndose establecer que la técnica que refirió mejores resultados, para el paquete de datos estudiado y el problema planteado, resultó ser k-means, seguido de clústering jerárquico.

Palabras clave: algoritmos para la gestión de datos y conocimientos, minería de datos, métodos de minería y algoritmos, métodos de agrupamiento

Abstract

In the present final work, a descriptive investigation was carried out about the processes, methods and techniques corresponding to data mining that are most frequently used at the business level to support business decision making.

Likewise, within the framework of the investigation, a problem was posed to be solved through business intelligence in a banking entity, the objective was to discover if there are groups other than those currently established to measure the performance of its branches and, if appropriate, reformulate the classification system to which houses are periodically submitted, this proposal was made with the purpose of comparing clustering techniques on a data package obtained through a KDD process, on which the algorithms corresponding to four representative clustering techniques were executed: k-means, AGNES, DBSCAN and fuzzy c-means, corresponding to groupings of the types based on distance, hierarchy, density and fuzzy, respectively.

Then, through the development of a data mining application in R language, the model was trained with the aforementioned techniques to obtain the different groupings on the branch dataset.

Finally, using internal cluster validation measures such as the elbow method and the silhouette coefficient, and clustering trends such as the Hopkins index, the results on the application of the different algorithms were analyzed, establishing that the technique used reported better results, for the data package studied and the problem posed, it turned out to be k-means, followed by hierarchical clustering.

Keywords: algorithms for data and knowledge management, data mining, mining methods and algorithms, clustering method

Dedicatoria

Desde mi más profundo sentir, quiero dedicar la realización de este trabajo final de carrera a mi familia, que me apoyó incondicionalmente para que logre mis objetivos. A mis hijos, que debieron conformarse con la ausencia de su madre en las largas horas de estudio y labor, a mi marido que me esperó pacientemente y a mis padres, que sembraron la semilla de la perseverancia en mí.

Reconocimientos

Gracias infinitas a mis profesores, los que compartieron su saber de manera dedicada y gentil, muy especialmente a la Dra. Marcela Samela, quien me supo dar confianza en los momentos de vacilación y a mi tutor Francisco Morteo, que me apoyó y guió sabiamente por el intrincado proceso del conocimiento.

Además, el respeto y cariño a mis compañeros de Licenciatura, con quienes estrechamos lazos de confianza y amistad, para lograr entre todos alcanzar nuestras metas.

Índice General

Índice de Gráficos

Índice de Tablas

Capítulo 1

Introducción

"In god we trust all others must bring data"
(Confiamos en Dios, el resto debe traer datos)
William Edwards Deming

Identificación del problema

Teléfonos inteligentes, cámaras de seguridad, redes sociales, aplicaciones móviles, tarjetas de crédito o balanzas smart son apenas una pequeña muestra de los dispositivos que integran el universo de productores de datos. La profusión de datos que generan los sistemas actuales se acumula en repositorios que difícilmente son recuperados para su estudio, ora por la falta de integración de las bases de datos, ora por la diversidad de los tipos de datos que se recaban. (Hernández Orallo y otros, 2004, págs. 3-4)

Cada vez más, términos como Big data, smart data, business intelligence o data mining ingresan al vocabulario cotidiano a la hora de lidiar con la masividad de los datos, aunque en ocasiones se asocien con un mismo concepto, cada uno de ellos representa una concepción disímil que vale la pena aclarar y explicar.

Recientemente, diferentes disciplinas relacionadas a la inteligencia de negocios han puesto de manifiesto la necesidad de contar con herramientas para descubrir el conocimiento oculto en los datos y aprovechar este activo en pos de generar valor apoyando a la toma de decisiones estratégicas y operativas de las empresas.

Planteamiento del tema de investigación

Las empresas tradicionales basan sus decisiones en diversos factores relacionados con la aplicación de la heurística, la ciencia y el arte de sus administradores. Muchas veces se delega esta facultad a las habilidades y destrezas que un individuo detente. En ocasiones su éxito o fracaso depende mayoritariamente de estas decisiones elaboradas desde el saber hacer personal.

En los últimos años se ha recurrido a la tecnología para automatizar en cierta medida estos procesos gracias a la posibilidad que ofrecen las TIC's de recabar grandes cantidades de datos, organizarlos de manera eficiente, reconocer variables o dimensiones, extraer conocimiento a partir de ellos y por último interpretarlos de modo que brinden una solución puntual a un problema real que se plantee.

Para ello se han desarrollado diversas técnicas de minería de datos aplicables a la toma de decisiones empresariales.

En el presente trabajo de investigación se pretende determinar las técnicas de segmentación que logran obtener los patrones de conocimiento de mejor calidad frente a un problema específico de negocio.

Objetivo General

El objetivo general de esta investigación es determinar qué técnica de la minería de datos basada en agrupamiento resulta más conveniente para la comprensión del comportamiento aplicable a la toma de decisiones empresariales.

Objetivos particulares

Adicionalmente se plantean como objetivos particulares los siguientes:

- Determinar las variables más significativas del modelo de datos a construir.
- Implementar las técnicas de minería de datos seleccionadas previamente.
- Comparar y evaluar los resultados de las técnicas de minería de datos puestas a prueba.

Justificación del Tema

Según una encuesta publicada recientemente por el IDC, proveedor de inteligencia de mercado, las empresas Data Driven cuentan con un 25% más de satisfacción del cliente y un 4% más de crecimiento en relación con las tradicionales (International Data Corporation [IDC], 2021).

Asimismo, resulta de vital importancia que la generación de valor a partir de los datos se produzca de manera ágil a fin de incorporarla a tiempo en los procesos operativos y comerciales de las empresas.

En la presente investigación se intentará evaluar las técnicas de minería de datos más eficientes para lograr una mejora en los procesos de toma de decisión basada en los datos.

Por otro lado, existen numerosos artículos que versan en análisis sobre segmentación de clientes con aplicación de técnicas de data mining, siendo oportuno en este caso indagar sobre los beneficios de la segmentación de sucursales en empresas de alta dispersión geográfica y clientela heterogénea, como es el caso sobre el cual se elaborará la solución de business intelligence.

Hipótesis

Entre las técnicas de data mining utilizadas para realizar agrupamiento, se encuentran las basadas en partición, las de clústering jerárquico, las de densidad o las difusas. Al aplicarlas sobre un conjunto de datos, cada una de ellas ofrece resultados disímiles aunque comparables, pudiéndose establecer la siguiente hipótesis:

La técnica de minería de datos que mejor se desempeña en una solución de business intelligence es Clústering Jerárquico.

Estructura General de la Tesis

A continuación, se detalla el contenido de los distintos capítulos que comprenden el presente trabajo final.

1.1.1 Capítulos

- En el capítulo 2, se aborda el marco teórico sobre el que se basa la investigación. Por una parte, la adopción de la perspectiva teórica y por otra la revisión de la literatura (Hernández Sampieri y otros, 1991, pág. 24).

Los temas a tratar versan sobre Business Intelligence, Sistemas de Toma de Decisiones (DSS por sus siglas en inglés), Big Data, Data Mining, Métodos y técnicas de DM y Métodos y algoritmos de agrupamiento.

- En el capítulo 3 se detalla la metodología a emplear que consta de:
    - El proceso de recolección de los datos que servirán de base sobre la que se implementarán las técnicas de minería de datos a evaluar
    - La determinación de las variables más significativas que se utilizarán para crear el dataset.
    - Las medidas que servirán para la evaluación de las agrupaciones obtenidas de los procesos previos.
- En el capítulo 4, se expone la implementación de la solución en lenguaje R que permitirá recabar los resultados de las pruebas representadas mediante tablas y gráficos, para posteriormente, realizar la evaluación de los clústeres obtenidos con las técnicas empleadas.
- Finalmente, se realiza la interpretación de los resultados, las conclusiones obtenidas fruto de la investigación y las restricciones observadas.

1.1.2 Anexos

- En el anexo I, se muestra el código en lenguaje R empleado para implementar las soluciones de clustering.

Capítulo 2

Marco Teórico

"Los datos son lo que necesitas para hacer análisis. La información es lo que necesitas para hacer negocios".
***John Owen**, un antiguo pionero de BI*

En este capítulo se exponen los aspectos teóricos que apoyan la compresión de los conceptos de Inteligencia de negocios y analítica de datos, así como también los de Big Data y Data Mining, para posteriormente realizar una revisión de los últimos avances y conocimientos extraídos sobre los temas que conciernen el presente trabajo.

Business Intelligence and Analytics

Seguidamente se realizará una introducción sobre la definición e historia de la Inteligencia de negocios y Analítica de datos, su evolución conforme el paso del tiempo, las técnicas más extendidas, además de profundizar acerca del concepto de sistemas de apoyo a la toma de decisiones y también sobre las metodologías disponibles de BI&A.

2.1.1. Historia y definición de BI & A

En la publicación que realizó el profesor Thomas Davenport en la revista de negocios de Harvard expuso que el término de inteligencia de negocios, también conocido por sus siglas en inglés BI, se popularizó en los ambientes empresariales y de tecnología a finales de la década del 80 y abarcaba una amplia gama de procesos y aplicaciones que eran utilizados para recopilar, analizar y difundir datos que apoyaban a la toma de decisiones en las empresas (Davenport, 2006, pág. 105).

Durante el WITS (Workshop de Tecnologías de la Información y Sistemas) del año 2001 se definió que la inteligencia de negocios “es el uso de inteligencia de software de alto nivel para aplicaciones de negocios. Más específicamente la inteligencia de negocios puede ser definida como la colección de tecnologías de punta que ayudan a los sistemas a ser más inteligentes” (Provost y otros, 2001, pág. 2).

Para las organizaciones comprometidas con el análisis cuantitativo basado en hechos resulta de vital importancia la analítica en sus estrategias empresariales, debiéndose adoptar cambios en las prácticas para aprovechar el poder de los datos. La inversión en tecnología, almacenamiento de datos, mejoramiento de los procesos y, sobre todo contar con las personas adecuadas, son algunas de las características que diferencian a los mayores competidores del mercado (Davenport, 2006, pág. 99)

Los sistemas de BI se componen de procesos, arquitecturas, aplicaciones, tecnologías y productos que convierten los datos crudos en información útil, ya sea a través de informes, procesamiento analítico en línea, análisis, minería de datos o minería de procesos para apoyar las decisiones comerciales (Heavy.AI, 2021).

2.1.2 Evolución

Según Chen et Al, se reconocen tres etapas evolutivas del BI&A que se diferencian básicamente por el tipo de datos a manipular, su ubicación y el modo de consumo.

La primera de ellas, *BI&A 1.0* se destaca por utilizar datos estructurados recopilados por diversos sistemas heredados y almacenados en sistemas de gestión de bases de datos (BDMS), sistemas de gestión de bases de datos relacionales (RDBMS) o data warehouses (DW), utilizando herramientas de Extracción, Transformación y Carga (ETL) para la manipulación de los datos y contando con el procesamiento analítico en línea (OLAP) además de técnicas analíticas basadas, generalmente, en métodos estadísticos tales como segmentación, asociación, agrupamiento o detección de outliers. A partir de estos datos se generan medidores de desempeño (KPI’s) que se vuelcan en los cuadros de mando de los Gestores de Procesos de Negocio (BPM).

La segunda etapa, *BI&A 2.0* se diferencia de la primera por sus características centradas en la WEB, que implican el manejo de datos semiestructurados almacenados en bases de datos capaces de contenerlos y administrarlos (NoSQL). El contenido generado por los usuarios a través de las redes sociales brinda una oportunidad de conversar con el cliente de una forma más efectiva conociendo sus intereses de manera más personalizada.

La tercera etapa, *BI&A 3.0* apunta a las aplicaciones móviles y las capacidades que brindan en los aspectos del negocio contar con la información en cualquier ámbito haciendo uso de las tecnologías propias de los smartphones como ser la geolocalización, la lectura de códigos de barras

y QR, los tags de Radiofrecuencia, los pagos con NFC, siempre apuntando a una atención centralizada en el cliente. (Chen y otros, 2012)

2.1.3 Tecnologías de BI

Según indican Diaz, Osorio y Amadeo las tecnologías se han ido diversificando para ofrecer soluciones a las distintas necesidades empresariales, entre ellas se encuentran contar con el universo datos contenidos en los registros transaccionales expresados de un mismo modo, es decir que los datos referidos a un mismo hecho o concepto se representen de manera análoga, lo antedicho tiene su fundamentación en la necesidad de mantener la integración de esos datos para ser explotados debidamente por las aplicaciones.

Otra tecnología altamente utilizada por los sistemas de BI son los Datawarehouses, capaces de albergar toda la información relevante del negocio de forma sencilla e inteligible y se encuentra orientado a responder velozmente consultas desde varios puntos de vista del usuario.

Adicionalmente a las tecnologías expresadas previamente, la disciplina OLAP permite a los usuarios finales contar con herramientas de análisis de datos, entre ellas el análisis de tendencias, la comparación de períodos o la navegación de los datos, etc.

En este sentido, una de las tecnologías que se impuso gracias al incremento de la capacidad de procesamiento de grandes y diversos volúmenes de datos es Big Data, proporcionando modelos de comportamiento que otrora pasaban desapercibidos al entendimiento.

Asimismo, entre las tecnologías a destacar se encuentran las herramientas de visualización de datos que aportan al decisor una mejor comprensión de los problemas y consecuentemente sus posibles soluciones.

Una de las tecnologías que aporta valor a BI es el análisis predictivo, que se vale de técnicas estadísticas, aprendizaje automático, modelización o minería de datos, mediante los cuales se logran anticipar comportamientos en base a indicios y señales.

Finalmente, tanto las tecnologías de tablero de control, que integran varios indicadores relacionados, como las de reporting, que permiten elaborar los informes de manera periódica, componen el amplio espectro de los sistemas de business intelligence (Diaz y otros, 2019, págs. 15-16).

En la Figura 1 se grafican los conceptos antes mencionados:

Figura 1

Tecnologías Aplicadas a BI

Nota. Tecnologías para el Análisis de datos basadas en software libre, 2019.

2.1.4 Modelo de datos dimensional

Los datos resultantes de la operatoria diaria se recopilan y acumulan en bases de datos, planillas de cálculo, información externa a la empresa, etc. siendo necesario integrarlos previamente en un modelo de datos denominado datawarehouse, que permita manipularlos de manera sencilla y veloz (Ferrari & Russo, 2008, pág. 9).

Los DW deben cumplir una serie de principios a fin de facilitar la información de negocios requerida en el menor tiempo factible, para ello se valen de dos conceptos básicos: el de hechos y dimensiones. Las tablas de hechos están integradas por los valores mensurables de un proceso, por ejemplo, para las ventas de un comercio sus posibles campos podrían ser: la fecha de la transacción, el producto vendido, la sucursal, si estaba sujeto a alguna promoción, el cliente, el vendedor, la transacción, el monto y las unidades vendidas. Por otro lado, se ubican las dimensiones, que se corresponden con la información textual que describen los hechos, para el caso anterior una dimensión podría ser la tabla de productos (Diaz y otros, 2019, págs. 25-27).

Las tablas de hechos se vinculan con las de dimensiones a través de las claves foráneas, mientras que las tablas de dimensiones contienen una única clave primaria para mantener la integridad referencial.

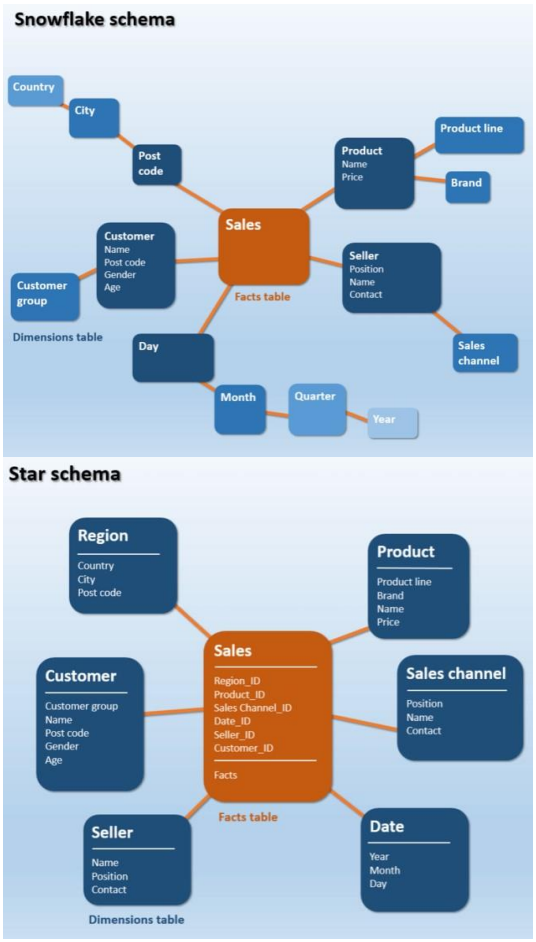
Con el propósito de mantener la simplicidad de la estructura, que redunda en velocidad, el esquema más conveniente es el de estrella, en el cual la tabla de hechos se vincula a las distintas tablas de dimensiones. Adicionalmente para casos específicos, existe la posibilidad de vincular

una dimensión con otra dimensión, denominándose esta topología copo de nieve (Ferrari & Russo, 2008, págs. 21-22).

Seguidamente en la Figura 2 se muestran el esquema en Estrella y Copo de Nieve:

Figura 2

Ejemplo de esquema en estrella y Copo de Nieve



Nota. Adaptado de Digital Guide Ionos, 2022 (www.ionos.es/digitalguide/online-marketing/analisis-web/los-data-warehouses-en-la-business-intelligence).

Es importante destacar además que el esquema dimensional acepta algún grado de desnormalización, es decir los registros pueden tener datos duplicados, esto permite mayor velocidad de acceso a las consultas.

2.1.5 Metodologías de BI

Existen variadas metodologías para crear una solución de Business Intelligence, aunque solamente dos resultan las más conocidas y probadas.

La primera de las metodologías fue formulada por William Inmon en su libro Creación de un Datawarehouse. En él describe los pasos para desarrollar un sistema de almacén de datos que sea la "recopilación de datos variables en el tiempo, integrada, no volátil y orientada al tema en apoyo de las decisiones de gestión" (Inmon, 2005, pág. 29). El proceso se inicia identificando las áreas temáticas con las que trabaja la empresa, como ser clientes, productos, ventas, etc. y en función de ello se crea el modelo lógico con los atributos asociados a cada entidad. Este modelo cumple con la normalización de los sistemas de bases de datos relacionales. Se trata de una visión desde lo general hacia lo particular que ayuda a identificar precozmente los requisitos comerciales y las posibles inconsistencias entre las relaciones, volviéndolo más robusto desde el comienzo, aunque también más costoso, porque se debe contemplar todo el escenario antes de diseñar el DW (Ferrari & Russo, 2008, pág. 30). Posteriormente se pasa a la etapa de construcción del modelo físico manteniendo la estructura normalizada, esto puede conllevar a la creación de múltiples tablas y relaciones. Se propone la construcción de data marts separados para cada entidad a fin de que el usuario consuma de ellos la información que requiera. Es de destacar que todos los datos que ingiere el DW están integrados y éste funciona como una única fuente de datos para varios data marts, de esta manera se logra la integridad y la coherencia en toda la empresa.

Por otra parte, con la intención de lograr sistemas de business intelligence menos complejos, surge una nueva metodología llamada de "arquitectura bus" que fue desarrollada e impulsada por Ralph Kimball con un enfoque opuesto a la visión de Inmon, debido a que el análisis comienza de abajo hacia arriba, es decir los requisitos comerciales son los que originan los diferentes data marts que conformarán a su vez el Data warehouse. Contrariamente a la política de Inmon de alimentar el DW desde el modelo lógico, en este caso los datos se colectan desde los diferentes sistemas transaccionales que conviven en una empresa (OLTP)[1], se tratan generalmente de sistemas heredados, como ser un CRM[2] u otro medio capaz de contener los registros de la administración del negocio. Como consecuencia de la diversidad sobre la fuente de los datos, se

[1] OLTP On-line Transaction Processing
[2] CRM Customer Relationship Manager

requiere seguidamente realizar un proceso de limpieza conocido como ETL que se encarga de la extracción, la transformación o normalización y la carga de éstos en un área de ensayo del DW que sirve para el almacenamiento temporario de las tablas durante el mencionado proceso, adicionalmente se considera una buena práctica la creación de una base de datos de configuración que interactúe durante la producción de los data marts (Kimball & Ross, 2013, págs. 7-27).

Los data marts podrían definirse como un subconjunto autónomo del data warehouse completo que se encuentra funcionalmente definido por hechos y dimensiones. Los hechos son la representación en términos mensurables de las distintas interacciones de los clientes con la empresa, por ejemplo, la venta de un producto a un cliente, o el monto total de las ventas de un artículo durante un tiempo determinado. Las dimensiones son objetos analíticos, como ser una lista de proveedores o de productos con sus atributos específicos: el color, año de fabricación, fabricante, etc. (Ferrari & Russo, 2008, pág. 20)

Se puede entender al data warehouse como la base de datos que contiene todas las tablas, vistas, procedimientos y códigos que los usuarios utilizan para la generación de información acerca del negocio.

2.1.6 Support Business Decisions DSS

También conocido como Sistema de apoyo a la Toma de Decisiones, es un sistema de información de la administración estratégica de una empresa diseñado específicamente para apoyar las habilidades gerenciales en todas las etapas del proceso de toma de decisiones. (Shim y otros, 2002, págs. 111-126)

Entre las características principales que deben cumplir los sistemas DSS para lograr un alto desempeño se destacan las siguientes: ser capaces de generar informes dinámicos, flexibles e interactivos, no ser requisito previo para su uso la adquisición de conocimientos técnicos, brindar la información en tiempos de respuesta ágiles, ser capaces de integrarse naturalmente con los sistemas existentes y departamentos de la compañía, proveer a cada usuario la disponibilidad de información adecuada para su perfil y contar la información histórica requerida a fin de proveerla en caso de requerirse.

Según Gartner, en su informe del 2021 sobre el Cuadrante mágico para Plataformas de Análisis e Inteligencia de Negocios, en los últimos años el uso de este tipo de herramientas se vio incrementada gracias a la incorporación de funcionalidades tales como la seguridad de usuarios y

posibilidad de realizar auditorías, las capacidades de administración para mantener controlada la divulgación de la información, los servicios en la nube, la actualización de los sistemas ETL para volverlos más intuitivos, el uso de machine learning para generar automáticamente hallazgos y mostrarlos a los usuarios finales, capacidad de visualización y manipulación de imágenes y gráficos, la posibilidad de realizar las consultas a las bases en lenguaje natural, la capacidad de crear el story telling de forma sencilla y la habilidad de crear y distribuir informes de alta calidad visual de forma programada, entre otros (Gartner, 2021).

En la Figura 3 se aprecia el resultado de la investigación mencionada precedentemente acerca del posicionamiento competitivo de los proveedores de tecnología de mercados de rápido crecimiento medidos en cuatro áreas: líderes, visionarios, jugadores de nicho y desafiantes.

Figura 3

Cuadrante Mágico para Plataformas de Análisis e Inteligencia de Negocios

Nota. Gartner, 2021.

Big Data

En esta sección se analizan los aspectos más destacados del Big Data, el proceso KDD para descubrimiento del conocimiento a partir de bases de datos incluyendo la fase de extracción, transformación y carga de los datos que debe realizarse previamente a la minería de datos.

2.2.1 Conceptos generales de Big Data

Como se ha indicado en el capítulo 1, la explosión de datos que generan los sensores y dispositivos en la actualidad permite contar con una cantidad y variedad antes inusitada. Si bien resulta beneficioso disponer de estos datos para obtener predicciones más certeras y veloces, también resulta desafiante encontrar los métodos más apropiados para lograr su correcto tratamiento (Gutierrez Puebla, 2017, pág. 2).

Los entornos de Big Data brindan la capacidad de manipular datos no estructurados como ser correos electrónicos, registros de flujo de clics de Internet, transcripciones de call centers, respuestas de encuestas, archivos de registro, imágenes, publicaciones en redes sociales, datos de sensores, etc. incapaces de ser procesados, gestionados ni analizados con las tecnologías previas (Stedman, 2022).

Acorde a la definición brindada por Doug Laney en 2001 para MetaGroup, Big Data se caracteriza por tres capacidades que inician con V (Stedman, 2022):

- Volumen, porque generalmente constituyen una gran cantidad de datos medibles en terabytes o más.
- Variedad, hace referencia a los tipos de datos que puede manejar (Gutierrez Puebla, 2017, pág. 6):
    - Estructurados, generalmente almacenados en bases de datos relacionales.
    - Semiestructurados, como ser ficheros HTML y JSON
    - No estructurados, mencionados en el segundo párrafo del punto 2.2.1 son administrados generalmente por bases de datos NoSQL
- Velocidad, es común en este tipo de sistemas la manipulación de datos en tiempo real que exige su tratamiento en cortos períodos de tiempo.

Adicionalmente se incorporaron al listado más características iniciadas con V:

- Veracidad. es requisito contar con conjuntos de datos en los que la precisión y fiabilidad se analicen previamente.
- Valor, se trata del valor comercial que los datos puedan proporcionar.

Entre los beneficios que la adopción de Big Data aporta a las empresas se pueden mencionar que, un mejor conocimiento del cliente, sus preferencias y comportamiento de compra resulta valioso para desarrollar una mayor inteligencia sobre las tendencias del mercado, productos y competidores. Además, la detección precoz de los problemas sobre la cadena de valor permite resolverlos a tiempo y lograr fidelizar a los clientes. Asimismo, la publicidad dirigida a los distintos segmentos de clientes evita la difusión masiva de pautas a aquellos poco interesados o con baja capacidad de conversión. Adicionalmente las técnicas de Big Data ayudan al mejoramiento de los procesos internos (Stedman, 2022).

2.2.2 Knowledge Discovery in Databases KDD

La definición que brinda Fayyad et al. para el término KDD "es el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos" (Fayyad y otros, 1996, pág. 40). Entre las propiedades mencionadas se entiende la palabra válido en el sentido que los patrones deben ajustarse a los nuevos conjuntos de datos manteniendo su precisión, con respecto al término novedoso, hace referencia al aporte de algo desconocido tanto para el sistema como para los usuarios, en cuanto a potencialmente útiles, se explica en la idea de beneficio para el usuario y por último comprensible, es un aspecto invalidante porque si la información extraída resulta incomprensible no proporciona conocimiento desde el punto de vista de su utilidad (Hernández Orallo y otros, 2004, pág. 13)

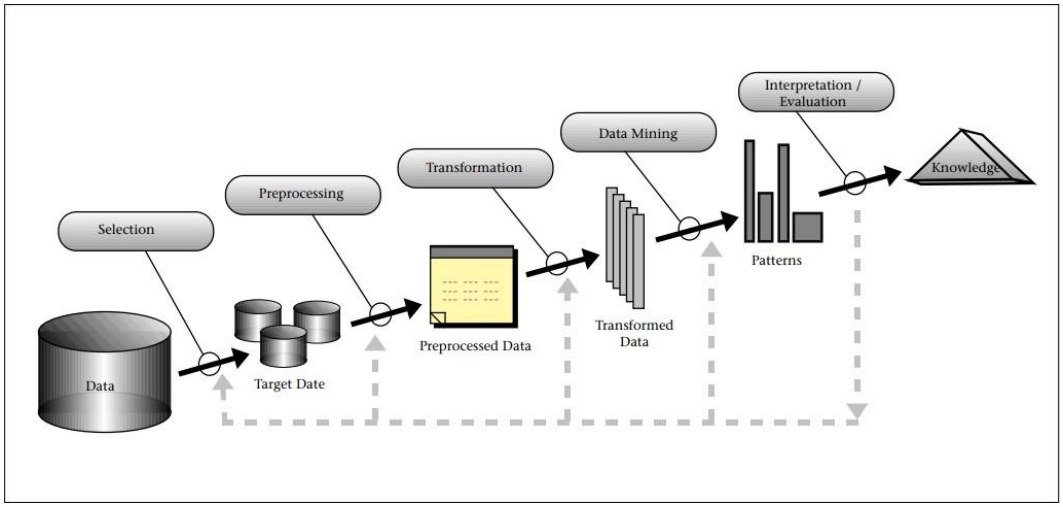
Según explican Han et al. KDD es el acrónimo de descubrimiento de conocimiento a partir de bases de datos, es un proceso compuesto por la siguiente secuencia iterativa de pasos: Primero: Se efectúa la limpieza de los datos, en esta fase se elimina el ruido y los datos inconsistentes. Segundo: Se lleva a cabo la integración de datos, aquí se combinan las diversas fuentes de datos. Tercero: se seleccionan los datos más relevantes en la tarea de análisis. Cuarto: sucede la etapa de transformación de datos, para permitir el adecuado análisis éstos se transforman y consolidan mediante la realización de operaciones de resumen o agregación. Quinto: En esta etapa, se realiza un proceso esencial llamado minería de datos, en el cual se aplican métodos inteligentes para

extraer patrones de comportamiento. Sexto: Se evalúan los patrones descubiertos y se identifican los más relevantes para la representación del conocimiento basado en medidas de interés y Séptimo: Se llega así a la presentación del conocimiento, el cual se vale de técnicas de visualización y representación para informar el conocimiento extraído a los usuarios (Han y otros, 2012, págs. 6-8) (Hernández Orallo y otros, 2004, págs. 19-39).

Seguidamente en la Figura 4 se grafican los pasos del proceso KDD mencionados precedentemente:

Figura 4

Una Descripción General de los Pasos que Componen el KDD



Nota. Adaptado de From Data Mining to Knowledge Discovery in Databases, 1996

Data Mining

Una de la etapas del proceso KDD descripto en el punto precedente es la minería de datos, según Witten y Frank la definen como “el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos” (Witten & Frank, 2000), si bien esta concepción resulta ser bastante similar a la de KDD explica su importancia y también el motivo por el cual uno y otro término son empleados indistintamente.

2.3.1 Introducción a Data Mining

Según Hernández Orallo la minería de datos busca extraer conocimiento a partir de los datos, que pueden presentarse en forma de relaciones, patrones, reglas o también en forma de resumen. También existen diversas maneras de representar esos modelos que dependerán del tipo de técnica que se emplee para inferirlos (Hernández Orallo y otros, 2004, pág. 12).

La minería de datos se diferencia en dos tipos de modelos. El primero de ellos recibe el nombre de predictivos ya que intenta anticipar o estimar valores desconocidos de variables de interés denominadas objetivo o dependientes usando otras variables almacenadas que se denominan variables independientes o predictivas. El segundo se trata de los modelos descriptivos cuya función es brindar una descripción más concisa o resumida de los datos y pretenden señalar situaciones que representan los datos identificando patrones que expliquen sus propiedades. Las tareas de minería de datos que producen los modelos predictivos son la clasificación y la regresión, mientras que las que dan lugar a los modelos descriptivos son el agrupamiento, las reglas de asociación y el análisis correlacional (Hernández Orallo y otros, 2004, pág. 12).

2.3.2 Tareas y métodos

Las tareas mencionadas precedentemente pueden ser realizadas usando distintos métodos, por ejemplo, una clasificación de elementos podría llevarse a cabo mediante árboles de decisión (DT) o redes neuronales (RNA).

Tareas

A continuación, se presentan las distintas tareas para el modelo descriptivo:

- **Agrupamiento** (clústering): La idea es obtener conjuntos de elementos que presentan cierta similitud entre sí y de detectar los focos de agrupación más sobresalientes, al inicio del proceso se desconoce la cantidad de grupos y sus características, las que se irán descubriendo durante el aprendizaje.
- **Correlaciones y factorizaciones**: Se centran en los atributos numéricos e intentan establecer una relación lineal o de otro tipo entre dos valores, el resultado es una función no orientada.
- **Reglas de asociación**: Es similar a las tareas correlacionales, pero con atributos nominales. Pueden darse del tipo orientadas o no.

- **Detección de valores anómalos**: El objetivo es encontrar aquellas instancias que no son similares a ninguna otra, con el objetivo de detectar circunstancias que impliquen algún riesgo.

Con respecto al modelo prescriptivo se pueden mencionar las siguientes:

- **Clasificación y clasificación suave**: se intenta aprender una función que represente la correspondencia entre los ejemplos, para cada valor del conjunto de entrada existe únicamente un valor en el conjunto de salida S, teniendo en cuenta que S, al ser nominal, puede adoptar un conjunto de valores (clases). En el caso de la clasificación suave, además se aprende otra función que significa el grado de certeza de la predicción.
- **Estimación de probabilidad de clasificación**: Es una generalización de la anterior, para elementos de dos conjuntos, se trata de aprender la probabilidad de que una instancia pertenezca a alguna clase previamente definida.
- **Categorización**: se pueden asignar varias categorías a una misma instancia.
- **Priorización**: como su nombre lo indica, se trata de determinar a partir de 2 o mas ejemplos el orden de preferencia.
- **Regresión**: el objetivo es aprender una función que represente la correspondencia existente entre los ejemplos que son exclusivamente numéricos.

Métodos

- **Técnicas algebraicas**: Expresan modelos mediante fórmulas algebraicas, funciones lineales, no lineales, distribuciones o valores estadísticos.
- **Técnicas bayesianas:** Utilizan el teorema de Bayes para estimar la probabilidad de pertenencia mediante las probabilidades condicionales inversa y a priori.
- **Técnicas basadas en conteos de frecuencias:** cuando dos o mas sucesos se presentan conjuntamente.
- **Técnicas basadas en árboles de decisión y sistemas de aprendizajes de reglas**: representan los modelos en forma de reglas, los algoritmos más conocidos son ID3, C4.5, CART y CN2.
- **Técnicas relacionales, declarativas y estructurales**: utilizan lenguajes declarativos, como los lógicos, funcionales o lógicos-funcionales. Las técnicas IPL (Programación lógica inductiva) se utilizan en la minería de datos relacional.

- **Técnicas basadas en redes neuronales**: Aprenden un modelo a través del entrenamiento de los pesos y activan nodos cercanos, el modelo se forma a partir de las conexiones y pesos. Existen diversos tipos: perceptrón simple, redes multicapa, redes de Kohonen, etc.
- **Técnicas basadas en núcleo y máquinas de soporte vectorial**: Intentan maximizar el margen entre los grupos o clases formadas.
- **Técnicas estocásticas y difusas**: Se pueden contar entre ellas simulated annealing, métodos evolutivos y genéticos o las funciones de pertenencia difusas.
- **Técnicas basadas en casos, densidad o distancia**: los algoritmos más conocidos son K-Medias, Two-step y COBWEB. (Hernández Orallo y otros, 2004, págs. 146-147)

Los modelos predictivos requieren ser entrenados con un conjunto de datos de aprendizaje en los que la variable objetivo es conocida, en la medida que se realiza este proceso los modelos se van ajustando a la realidad obteniendo cada vez mejores resultados, a este tipo de aprendizaje se lo denomina supervisado.

Por otra parte, en los modelos descriptivos, donde no se cuenta con los valores resultantes de antemano, el tipo de aprendizaje es no supervisado, dentro de esta categorización se encuentran el agrupamiento y las reglas de asociación.

2.3.3 Análisis de clústeres

La intención del análisis de clústeres es capturar la estructura natural de los datos dividiendo el conjunto de datos en grupos que sean significativos en forma tal que los objetos dentro de un mismo clúster tengan entre sí gran similitud y sean muy diferentes con respecto a los que integran los otros clústeres. En ocasiones se utiliza para acotar la cantidad de observaciones y reducir la complejidad del análisis proporcionando una abstracción de los datos individuales que integran esos clústeres, encontrando los prototipos de conglomerados más representativos (Tan y otros, 2006, pág. 487) (Han y otros, 2012, pág. 443).

Lo expresado en el párrafo anterior tiene sustento en que muchas técnicas de análisis de datos, tales como la regresión, insumen tiempo y espacio por su complejidad y resulta más eficiente aplicar el algoritmo a los prototipos de cada clúster que a todo el conjunto de datos, máxime si se tratare de un gran volumen de datos. En otras oportunidades, los tipos de datos son extensos y

requieren ser comprimidos previamente a su tratamiento, esta situación se da con datos de imagen, sonido y video, para ello, a cada prototipo se le asigna un índice de posición asociado a cada clúster evitando trabajar con el dato completo que no aportaría mayor relevancia. Adicionalmente, la aplicación de la técnica de búsqueda de vecinos más cercanos puede requerir la comparación de a pares de las distancias de todos los puntos del conjunto, demandando ello una gran cantidad de cálculos, una manera de optimizar este proceso es calcular exclusivamente la distancia de los objetos en grupos cercanos o pertenecientes a un mismo clúster (Tan y otros, 2006, pág. 489).

Las agrupaciones se pueden diferenciar en jerárquica o particional, en la primera se da el anidamiento de subconjuntos en otros mayores que los contienen representando un esquema de árbol, en el segundo las particiones no son inclusivas entre sí. También, según el tipo de asignación de los objetos a cada grupo se denominan exclusivos, superpuestos o difusos. Por último, las agrupaciones pueden ser completas o parciales en la medida que asignen la totalidad o no de sus objetos a un grupo (Han y otros, 2012, págs. 447-448)

2.3.4 Métodos de clústering

Existen varios algoritmos de agrupamiento que pueden clasificarse según su método, los de partición, por ejemplo, se identifican por dividir el espacio de datos en grupos de una manera que cada uno contenga al menos un objeto, generalmente se basan en distancia euclidiana. Crean una partición inicial y luego utilizan una técnica de reubicación iterativa que intenta mejorar la partición reposicionando los objetos de un grupo a otro. Entre las aplicaciones se distinguen los algoritmos k-means y k-medoids.

Otro tipo lo componen los métodos jerárquicos, que pueden clasificarse en aglomerativos o divisivos, según el enfoque que se adopte para iniciar el proceso: de lo particular a lo general o viceversa. En el primero cada objeto se adopta como un grupo separado que se fusiona iterativamente con los otros grupos conforme su cercanía hasta que todos los grupos se consolidan en uno. Del modo inverso, en el enfoque divisivo el conjunto de datos se toma como un único grupo que se separa en cada iteración de acuerdo con sus características, hasta que cada elemento se transforma en un grupo distinto. Este tipo de modelos, tanto los aglomerativos como los divisivos, adolecen de ser poco flexibles, ya que una vez que se fusiona o divide un grupo el

proceso no se puede deshacer, esta característica también puede tomarse como virtud debido a que redunda en menores costos de cálculo al reducir las combinaciones posibles.

El tercer método es el basado en la densidad, la idea general es incorporar elementos en un grupo determinado mientras la densidad en el vecindario exceda cierto umbral. Resulta útil para los casos de filtrado de ruido o para valores anómalos y para descubrir grupos de forma arbitraria y de estructuras no esféricas.

El método basado en cuadrículas reparte el espacio de datos en celdas, este tipo de enfoque cuenta con la ventaja de independizarse de la cantidad de objetos que se deban procesar tornándolo muy rápido (Han y otros, 2012, págs. 448-450).

Finalmente, el clústering difuso, considerado del tipo suave, por asignar los objetos a más de un clúster mediante el coeficiente de afiliación dependiente de su grado de pertenencia. Esta concepción difiere de los métodos llamados duros que relacionan cada objeto a un único clúster según su similitud. El algoritmo Fuzzy c-means (FCM) por sus siglas, es uno de los más utilizados para este tipo de clústering, el centroide de cada grupo se calcula como la distancia promedio de todos los puntos ponderada por su coeficiente de pertenencia (Kassambara, 2017, pág. 167).

A continuación, se detallan las diferentes técnicas de agrupamiento para los métodos mencionados precedentemente:

K-means: "Es una técnica particional basada en prototipos que intenta encontrar un número de agrupamientos (K) especificado por el usuario, que están representados por sus centroides" *(Tan y otros, 2006, pág. 495)*

El procedimiento es el siguiente: Primero se especifica la cantidad de clústeres buscados que será el parámetro K. Luego cada punto se asigna al centroide más cercano conforme su distancia euclídeana y cada conjunto de puntos asignados a un centroide será un grupo. Seguidamente el centroide de cada grupo se actualiza en función de la media. Este proceso de asignación y actualización se repite hasta que ningún punto cambie de clúster o, lo que es lo mismo, hasta que los centroides sigan siendo los mismos (Witten & Frank, 2000, pág. 139) (Tan y otros, 2006, pág. 497)

Este algoritmo puede arrojar diferencias según la elección de los centroides iniciales, por esa razón elegirlos correctamente resulta un paso clave del procedimiento. Una técnica válida se

basa en realizar varias ejecuciones con diferentes centroides iniciales elegidos al azar y luego seleccionar el conjunto de clústeres con menor SSE (sumatoria del error cuadrático) o dispersión. Un técnica válida para mejorar el agrupamiento es seleccionar el primer punto al azar y luego, para cada centroide sucesivo, elegir el punto que esté más alejado de los anteriores centroides, de esta forma se logra un conjunto de valores iniciales separados. Es de destacar que esta técnica puede seleccionar valores atípicos, para mitigar este riesgo se suele seleccionar una muestra aleatoria del conjunto y aplicar el algoritmo con el objeto de reducir la probabilidad de elección de estos puntos como centroides.

K-medoids: Se trata de un algoritmo similar al anterior que intenta disminuir la sensibilidad a los valores anómalos, para ello recurre el uso del criterio del error absoluto.

En el algoritmo Partitioning Around Medoids (PAM) los objetos representativos iniciales llamados semillas son elegidos arbitrariamente y se prueba iterativamente la calidad del agrupamiento según el costo de disimilitud entre un objeto y el objeto representativo de su grupo o medoide, al cambiar el objeto representativo por otro. Si bien este algoritmo es más robusto con respecto a k-means también resulta más complejo y por ende menos escalable.

Para conjuntos de datos grandes se puede usar un método basado en muestreo llamado CLARA (Clustering LARge Applications) que toma una muestra aleatoria del conjunto de datos y luego aplica el algoritmo PAM. La efectividad de CLARA dependerá de la calidad de la muestra, si los mejores k-medoids no fueron elegidos, jamás se logrará el agrupamiento óptimo. Lo anterior se puede mitigar con el algoritmo CLARANS (Clustering LARge Applications based on RANdomized) que compensa el costo y la efectividad del uso de muestras para obtener el agrupamiento.

Clústering Jerárquico: Un agrupamiento jerárquico generalmente se diagrama en forma de árbol llamado dendograma, que muestra las relaciones y el orden en el que los grupos se fusionaron o dividieron.

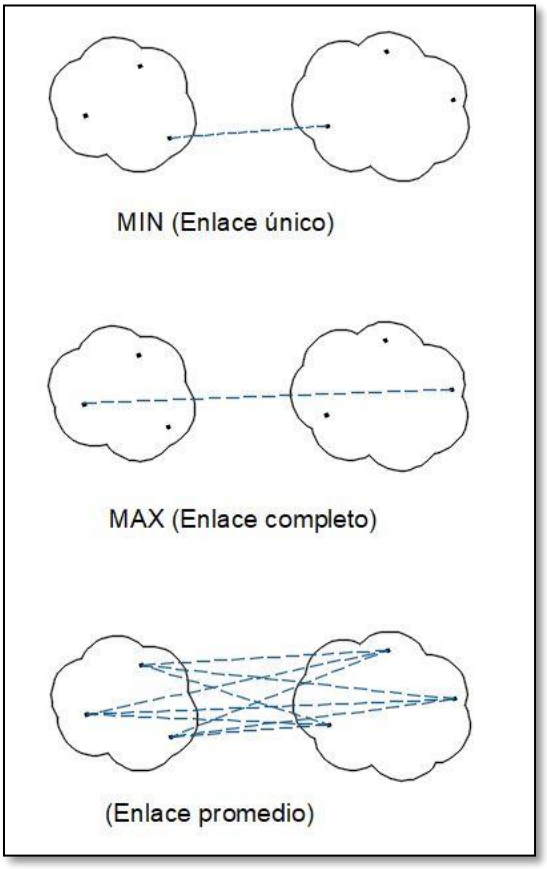
En el clústering jerárquico aglomerativo el procedimiento es el siguiente: se calcula la matriz de proximidad, se fusionan los dos grupos más cercanos, luego se actualiza la matriz de proximidad y se repite el proceso hasta que solo quede un grupo.

Par calcular la proximidad entre los clústeres existen varios métodos, el de enlace único o MIN: define la proximidad de un conglomerado como la proximidad entre los puntos más cercanos que se encuentren en distintos clústeres, el de enlace completo o MAX: toma la proximidad entre los dos puntos más alejados de dos clústeres y el de Enlace Promedio del grupo: calcula la proximidad de dos clústeres como la media de todos los puntos.

En la Figura 5 se grafican los 3 tipos de enlaces mencionados previamente.

Figura 5

Tipos de enlaces para medir proximidad de clústeres



BIRCH: Por sus siglas en inglés Balanced Iterative Reducing and Clustering using Hierarchies es un algoritmo que está pensado para tratar grandes cantidades de datos numéricos integrando dos métodos de agrupamiento, como ser el agrupamiento jerárquico y la partición iterativa, este enfoque brinda mayor escalabilidad y capacidad para deshacer pasos anteriores aunque, al utilizar la noción de radio o diámetro para controlar el límite de un cúmulo, solamente es efectivo para agrupamientos con forma esférica.

El procedimiento es el siguiente: En la primera fase BIRCH escanea el dataset creando un árbol de características de agrupamiento (CF), este tipo de construcción es incremental a medida que se insertan los objetos, asignándolo al objeto hoja más cercano, luego se realiza la comparación del diámetro del subclúster al que pertenece esa hoja con respecto al objeto y, si resultase mayor al umbral establecido se procede a dividirlo en otro subclúster. En la segunda fase se aplica un procedimiento para agrupar los nodos hoja del árbol CF con la intención de eliminar los valores atípicos y también reagrupar los clústeres más densos entre sí (Han y otros, 2012, págs. 462-465).

Chameleon: Este algoritmo del tipo jerárquico utiliza modelos dinámicos para evaluar la similitud entre pares de agrupamientos a través de las conexiones y la proximidad de ellos. Se construye un gráfico de k-nearest neighbours en el que cada nodo representa un objeto enlazado por aristas entre los de mayor similitud. Luego se realiza una partición de gráfico minimizando el corte de borde (peso de la arista) para lograr una gran cantidad de subgrupos, que finalmente utiliza un algoritmo de agrupamiento jerárquico aglomerativo para fusionar iterativamente los grupos en función de su similitud. Chameleon demostró mayor capacidad para descubrir clústeres de forma arbitraria de alta calidad que BIRCH o DBSCAN, sin embargo, el costo del procesamiento en paquetes de datos de alta dimensionalidad puede requerir hasta $O(n^2)$ tiempo para n objetos *(Han y otros, 2012, págs. 466-467).*

DBSCAN: Es un algoritmo basado en densidad que produce un agrupamiento particional en el que el algoritmo determina automáticamente la cantidad de clústeres, es del tipo incompleto debido a que las regiones de baja densidad se clasifican como ruido y se omiten.

Este algoritmo encuentra objetos centrales o con vecindarios densos, es decir con una cantidad determinada de objetos cercanos, y los conecta para formar regiones densas en grupos.

La determinación del parámetro para delimitar una vecindad es especificada por el usuario con antelación en el inicio del proceso.

El procedimiento es el siguiente: Se marcan todos los objetos del conjunto de datos como “no visitados”, seguidamente DBSCAN elije aleatoriamente uno de ellos, lo marca como visitado y verifica si el vecindario contiene al menos tantos objetos como el parámetro preestablecido para crear un nuevo grupo e incluir a los objetos del vecindario como candidatos, en caso de no satisfacer la ecuación, el punto se marca como ruido. El proceso se repite con los objetos no visitados hasta que no queden objetos por visitar.

OPTICS: El nombre de este algoritmo se debe al acrónimo de Ordering Points to Identify the Clustering Structure ya que genera un orden de grupos según su densidad y representa la estructura de agrupamiento. OPTICS no requiere que el usuario proporcione de antemano el límite de densidad sobre el cual se definirá una vecindad. Los objetos son procesados siguiendo el orden hallado lo que permite adelantar el proceso de búsqueda de clústeres que posean mayor densidad. Esto puede ser útil para extraer información básica de grupos, como ser centros de conglomerados o encontrar conglomerados de forma arbitraria.

STING: Es una técnica basada en cuadrícula que particiona el espacio de datos en celdas independientemente de la distribución de los objetos. Además, el espacio se puede dividir jerárquicamente y contener nuevas celdas en los niveles inferiores. STING explora la información de manera estadística, extrayendo para cada celda la cantidad de objetos, media, desviación estándar, mínimo, máximo y tipo de distribución, y en base a estos parámetros se realizan las agrupaciones. Si bien el método es muy eficiente para grandes volúmenes de datos multidimensionales, la calidad del agrupamiento está estrechamente relacionada al nivel de granularidad que se adopte. Además, como los límites de los conglomerados resultantes pueden ser exclusivamente horizontales y verticales, no resulta factible detectar límites diagonales.

CLIQUE (CLustering In QUEst): En ocasiones se requiere buscar grupos en diferentes subespacios de los datos, este método intenta encontrar clústeres acordes la densidad de las cuadrículas según un umbral preestablecido para identificar celdas densas o dispersas. Luego usa las celdas densas para ensamblar grupos que pueden tener una forma arbitraria. Esta técnica es insensible al orden de los objetos de entrada y tiene una buena escalabilidad frente al aumento de dimensiones en los datos, sin embargo, lograr una agrupación significativa depende del tamaño de las cuadriculas y del umbral de densidad fijados.

2.3.5 Evaluación de Clústering

Existen tres tareas principales para determinar la calidad de la agrupación, la primera es la evaluación de la tendencia de agrupación, la segunda es la determinación del número de conglomerados y la tercera, la medición de la calidad del agrupamiento (Han y otros, 2012, págs. 483-490).

Al inicio del análisis, resulta necesario verificar el tipo de distribución que presentan los datos, si la distribución es uniforme no existe un patrón de agrupamiento y aplicar algoritmos sobre ellos no devolverá información significativa, ya que al volver a ejecutarlos los conglomerados que se logren pueden adoptar cualquier otra forma.

Una manera de verificar si existe una estructura no aleatoria de los datos es aplicar la Estadística de Hopkins, se trata de una prueba que analiza la aleatoriedad espacial de una variable distribuida en el espacio y se representa de la siguiente forma:

Para cada punto p_i en un espacio de datos D x_i es la distancia entre p_i y su vecino más cercano en D. Asimismo para cada punto q_i y_i es la distancia entre q_i y su vecino más cercano en $D - \{ q_i \}$

$$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i}$$

La ecuación precedente indica que para distribuciones homogéneas en un espacio de datos D, o sea que las distancias de los vecinos más cercanos sean similares para todos los puntos de D, el valor de H rondaría 0.5. Sin embargo, si D estuviera muy sesgado H tendería a 0. De esta manera

se puede realizar la prueba iterativamente empleando 0.5 como umbral. Si $H = 0.5$ entonces es poco probable que *D* tenga conglomerados estadísticamente significativos (Kassambara, 2017, págs. 123-124).

Otro factor para tener en cuenta es la cantidad correcta de conglomerados en el conjunto de datos, ya que este valor determinará la granularidad adecuada. Si todo el conjunto de datos se agrupara en un mismo clúster maximizaría la compresión, aunque no aportaría información válida de los datos contenidos, de manera inversa, si cada elemento del conjunto de datos se agrupara en un clúster cada uno, el resultado sería altamente preciso, aunque tampoco aportaría mayor valor. Es por esta razón que se requiere de un equilibrio entre la compresibilidad y la precisión en el análisis de conglomerados.

Un procedimiento para determinar la cantidad de clústeres es el denominado método del codo, que surge de verificar que, a medida que se incrementa el número de conglomerados la suma de la varianza dentro del conglomerado se reduce ya que permite capturar grupos más finos de objetos similares entre sí, esta relación cambia si se definen demasiados conglomerados. El punto de inflexión en la curva generada por la función conforme varía la cantidad de clústeres, resulta el valor óptimo de agrupaciones.

Existen varios métodos para medir la calidad del agrupamiento diferenciados según se disponga o no de información para medir el grado de acierto del patrón generado. En clústering generalmente se desconoce de antemano la cantidad de clústeres y la forma que adoptarán, razón por la cual se utilizan los métodos intrínsecos, evaluando las medidas de separación y cohesión entre clústeres para determinar si representan una buena agrupación. Una medida es el coeficiente de silueta, que determina la compactabilidad de un conglomerado y el grado de separación entre ellos. Cuando el valor del coeficiente de silueta se aproxima a 1, el grupo que contiene *o* es compacto y está alejado de otros grupos, en cambio si *s(o)* es negativo, *o* está más cercano a objetos de otro grupo que al de pertenencia. Su fórmula es la siguiente:

$$s(o) = \frac{b(o) - a(o)}{\max\{ao, bo\}}$$

Donde *a(o)* es la distancia promedio entre *o* y todos los demás objetos al que pertenece *o* y *b(o)* es la distancia mínima promedio entre *o* y todos los grupos a los que *o* no pertenece (Han y otros, 2012, págs. 489-490).

Capítulo 3

Metodología

"Los datos son la nueva ciencia.
El Big Data son las respuestas"
Pat Gelsinger

Introducción

En este capítulo se explica la metodología utilizada en la presente investigación. La misma consiste en la aplicación del método KDD para la obtención del paquete de datos sobre el que se ejecutarán las diversas técnicas de agrupamiento de minería de datos a fin de descubrir los clústeres más significativos del modelo mediante la evaluación de desempeño de los algoritmos empleados.

Metodología KDD

En complemento a lo expresado en el punto 2.2.2, en el que se explican los conceptos teóricos de la metodología KDD, se detallan las siguientes fases:

- **Selección**

Dentro de esta etapa se prioriza el entendimiento del negocio desde el punto de vista del cliente, se trazan los objetivos del proceso KDD y se destina el análisis al conocimiento previo y a la determinación del dominio del problema a resolver. En función de ello se crea el conjunto de datos seleccionando las variables más significativas que permitan el descubrimiento de los patrones de comportamiento insertos en los datos.

- **Preprocesamiento**

Aquí se aplica mediante operaciones básicas la limpieza de los datos, tales como la eliminación del ruido, la verificación de campos vacíos, la estandarización en la representación, además de la reducción de datos redundantes o de dimensiones poco significativas. Se intenta encontrar características útiles para representar los datos en función del objetivo de la tarea.

- **Transformación**

En este paso los atributos de los datos sufren modificaciones para obtener una adecuada calidad que mejore los resultados de las técnicas de minería que se aplicarán seguidamente. Entre

ellos se pueden mencionar la reducción de dimensionalidad por transformación mediante Análisis de Componentes Principales (PCA) o por análisis factorial, aumento de la dimensionalidad, la discretización o la numerización (conversión de un dato numérico en valor nominal y viceversa) o el escalado.

- **Minería de datos**

El propósito de esta etapa es elegir el algoritmo de la minería de datos que mejor satisfaga el objetivo según las diferentes tareas y métodos disponibles para lograr la búsqueda de patrones de interés que aporten conocimiento acerca de los datos analizados.

- **Evaluación e implantación**

Por último, se procede a la interpretación de los patrones extraídos mediante la minería, puede implicar también la visualización de los modelos encontrados para aportar una mayor comprensión. En esta etapa se realiza el uso del conocimiento minado en otro sistema o su documentación para presentación en informes que apoyen la toma de decisiones.

Metodología utilizada en este trabajo

La metodología de esta investigación se basa en la aplicación de KDD para la obtención de patrones de agrupamiento en un conjunto de datos proporcionados por el Datawarehouse de una empresa del ramo bancario para descubrir la similitud natural de sus sucursales y aplicar eventualmente una recategorización en función de sus resultados.

Asimismo, se evalúa el desempeño de las diferentes técnicas de clústering, implementadas mediante el software RStudio, utilizando diversas medidas para calificar los agrupamientos encontrados.

El primer paso en la metodología KDD es realizar el análisis del dominio del problema de negocios acorde a los objetivos organizacionales con el propósito de descubrir, mediante la aplicación de técnicas de minería de datos, patrones de comportamiento entre las sucursales de la empresa, verificando si existen agrupaciones nuevas a las ya existentes que aporten valor en su administración.

Las actividades de negocio principales son las relacionadas con préstamos, depósitos, cobro de comisiones y tarjetas de crédito; en función de ellas se extrajeron los datos transaccionales

almacenados en el datawarehouse propio, incorporándose al cubo la cantidad de operaciones de cada filial. Como resultado de la consulta, se generó un dataset de 21 columnas y 640 filas.

Seguidamente se procede a la etapa de limpieza de los datos crudos, detectando valores omitidos, outliers, posibles errores en la recolección, etc. Como resultado de este análisis, se detectaron 5 casas con actividades especializadas, como ser operativa judicial, aduanera y de comercio exterior, por lo que se estimó conveniente retirarlas del análisis a fin evitar que los datos se vean sesgados por registros que no representan la actividad normal.

En el paso siguiente, denominado de transformación, se ejecutó un procedimiento de relativización por máximos que tiene la finalidad de homogeneizar el peso de las variables y permitir que éstas sean más comparables entre sí, luego se realizó el análisis de componentes principales (PCA) para evaluar en qué medida las variables influyen en el modelo de datos.

En la instancia de minería de datos, se entrenó el modelo con 4 técnicas diferentes de clústering haciendo uso de la herramienta de código abierto RStudio versión 2022.07.1 que ejecuta lenguaje R v.4.2.2 para Windows.

Las técnicas a probar son del tipo basadas en distancia, en jerarquía, en densidad y difusa, implementadas mediante los algoritmos de k-medias, hclust, dbscan y Fanny.

Para la etapa de evaluación e implementación de KDD, se lleva a cabo el análisis de los clústeres descubiertos por cada técnica mencionada, realizando la medición del coeficiente de silueta descrito en el punto 2.3.5, mediante la función fviz_nbclust. Obteniendo una tabla de comparación de métodos como resultado del proceso y las gráficas correspondientes que mejoran su comprensión.

Finalmente, se realiza una breve interpretación de los resultados.

Capítulo 4

Implementación de la propuesta y evaluación de resultados

"Cuantos más datos tengamos,
más posibilidades tenemos de ahogarnos en ellos"
Nassim Taleb

Introducción

En este capítulo se presenta la implementación de la solución en la que se prueban las distintas técnicas sobre el paquete de datos obtenido mediante la metodología detallada precedentemente.

Datos

Los datos operados en el presente trabajo final pertenecen a una entidad bancaria argentina en actividad, y debido a su procedencia transaccional no resultan del dominio público, razón por la cual se mantiene el resguardo de privacidad de los datos. Para ello se realizó la modificación del dataset, evitando contener información que pudiera causar algún perjuicio comercial.

Se recopilaron datos de las actividades principales de todas las filiales durante el año 2021 volcando el valor promedio en 20 variables numéricas que se detallan seguidamente:

Tabla 1

Variables Ponderadas para el Análisis

Nombre de variable	Descripción
TR	Cantidad de transacciones
CCOMN	Cantidad de préstamos de cartera comercial normal (sin deuda)
SCOMN	Saldo de préstamos de cartera comercial normal (sin deuda)
CCOMI	Cantidad de préstamos de cartera comercial irregular (con deuda)
SCOMI	Saldo de préstamos de cartera comercial irregular (con deuda)

CCONSN	Cantidad de préstamos de cartera de consumo normal
SCONSN	Saldo de préstamos de cartera de consumo normal
SCONSI	Saldo de préstamos de cartera de consumo irregular
CCONSI	Cantidad de préstamos de cartera de consumo irregular
CCC	Cantidad de Cuentas Corrientes
SCC	Saldo de Cuentas Corrientes
CCA	Cantidad de Cajas de Ahorros
SCA	Saldo de Cajas de Ahorros
CCCE	Cantidad de Cuentas Especiales
SCCE	Saldo de Cuentas Especiales
CPFMI	Cantidad de Plazos Fijos minoristas
SPFMI	Saldo de Plazos Fijos minoristas
CPFMA	Cantidad de Plazos Fijos mayoristas
SPFMA	Saldo de Plazos Fijos mayoristas
COM	Comisiones e Ingresos por Servicios

Preparación de los datos

Según se indicó en el capítulo anterior se procedió a eliminar 5 registros correspondientes a sucursales con operativa específica que no representan la normalidad de las casas, separándolos para un análisis pormenorizado posterior que no resulta resorte de esta investigación.

Asimismo, se ejecutaron procedimientos para encontrar valores nulos y tomar acciones en consecuencia, como resultado se detectaron 3 casos que fueron verificados respecto a la base de datos, determinándose que esos campos no contenían datos, por lo tanto se procedió a reemplazarlos por el valor medio de la columna a la que pertenecían.

Por último, se normalizaron los nombres de las filiales a fin de evitar los caracteres que dificulten su operación.

Transformación de los datos

En primera instancia se importó el archivo del tipo .CSV, que contiene los datos generados en los pasos previos, al aplicativo RStudio para iniciar el proceso de análisis de tendencia de agrupación mediante dos métodos, el primero es el coeficiente que arroja la estadística de Hopkins

desarrollado en el punto 2.3.5 y el segundo un método visual VAT, en el que se puede verificar si existen agrupaciones entre los datos mediante la formación de figuras cuadradas de tono oscuro linderos a la diagonal de la imagen, ambos implementados mediante la función get_clust_tendency() del paquete factoextra,. En las Figuras 6 y 7 se exponen ambos resultados.

Figura 6

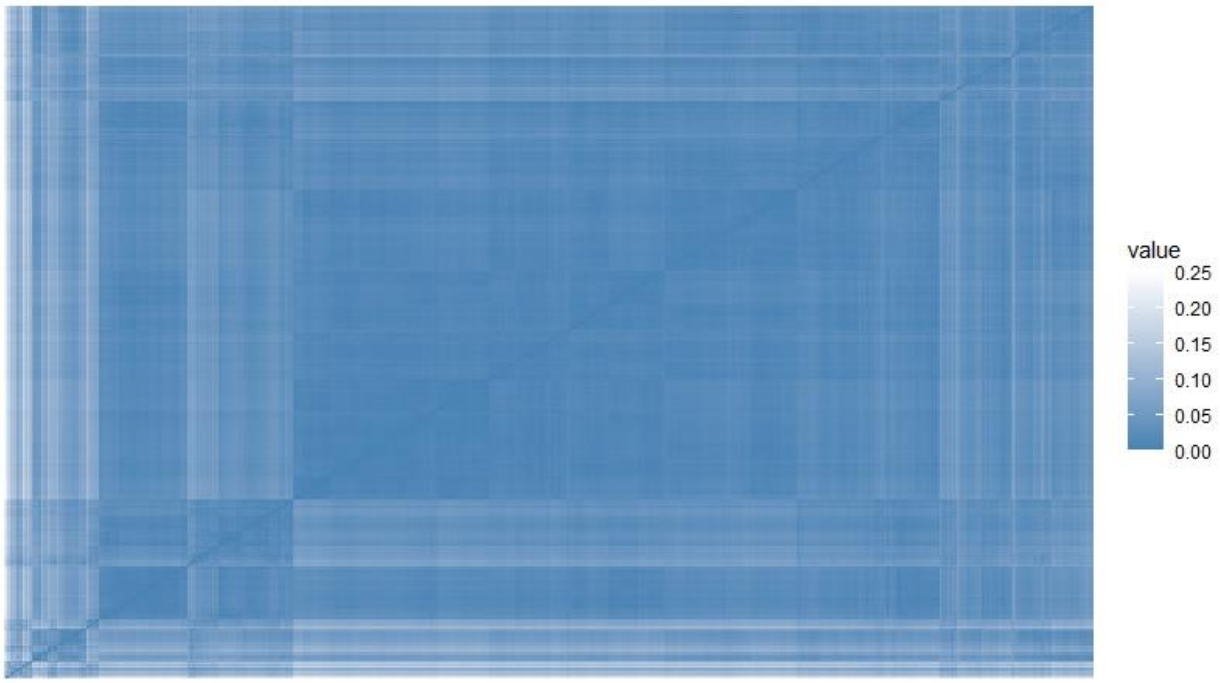
Coeficiente de Hopkins

```
> get_clust_tendency(dtr1, n = 50,
+                    gradient = list(low = "steelblue",  high = "white"))
$hopkins_stat
[1] 0.9261484
```

Nota. El valor que arroja el coeficiente de Hopkins es mayor a 0.5 por lo tanto se puede inferir que los datos contenidos en el dataset dtr1 presentan una distribución no uniforme y por lo tanto existen agrupaciones significativas.

Figura 7

VAT Evaluación Visual de la Tendencia de Agrupamiento



Nota. Las formas cuadradas oscuras denotan la presencia de agrupaciones.

Minería de los datos

En esta etapa de extracción de conocimiento se prueban las cuatro técnicas de agrupamiento seleccionadas con el propósito de extraer los valores que permitan su posterior evaluación: K-means, jeráquico, Fanny y DBSCAN.

4.5.1 Clústering basado en distancia: K-means

Una vez verificada la tendencia de agrupamiento se procede a ejecutar el algoritmo k-means sobre el dataset para distintos valores de k, que representan los centroides de cada clúster. La implementación se realiza mediante la función kmeans() del paquete stats, con k=2, k=3, k=4, k=5, k=6 y 100 iteraciones para cada corrida, inicializando previamente el vector que genera aleatoriamente los valores de los centroides, esta acción se realiza para reproducir los mismos resultados en el futuro. En la Figura 8 y en la Tabla 2 se representan el código y los resultados obtenidos.

Figura 8

Ejecución de Algoritmo K-means() para K entre 2 y 6

```
> set.seed(123)
> k2 <- kmeans(x = dtr1 , centers = 2, nstart = 100)
> k3 <- kmeans(x = dtr1 , centers = 3, nstart = 100)
> k4 <- kmeans(x = dtr1 , centers = 4, nstart = 100)
> k5 <- kmeans(x = dtr1 , centers = 5, nstart = 100)
> k6 <- kmeans(x = dtr1 , centers = 6, nstart = 100)
```

Nota. Se aplica la función sobre el mismo paquete de datos variando la cantidad de clústeres predefinidos.

Tabla 2

Resultados de K-means para K entre 2 y 6

		k2	k3	k4	k5	k6
Tamaño	**c1**	594	450	409	146	292
de	**c2**	41	21	20	90	15
clústeres	**c3**		164	80	348	33
	c4			126	17	128

	c5				34	126
	c6					41
Indice de silueta de clústeres	**c1**	0.7131463	0.5740307	0.5458543	0.4328884	0.46056348
	c2	0.1528854	0.17141867	0.173659	0.162005	0.14351021
	c3		0.03453065	0.1530301	0.1616086	0.15223334
	c4			0.163352	0.2175396	0.27475544
	c5				0.1405838	0.12675321
	c6					0.07507839
Indice de silueta promedio		0.6769719	0.4213805	0.4087439	0.3133823	0.3084704

4.5.2 Clústering basado en jerarquía

En este punto se implementa la técnica jerárquica aglomerativa o AGNES, que tiene una visión desde lo particular a lo general o de abajo hacia arriba, en principio cada objeto se define como un clúster y se combina iterativamente con otros según su similitud, al final del proceso el conjunto de datos completo compone un único clúster. Estas particiones se representan en un gráfico llamado dendograma o diagrama de árbol que puede ser cortado transversalmente a diversas alturas para determinar la cantidad de clústeres y su respectivo nivel de abstracción.

Para iniciar el proceso, se calcula la matriz de distancias usando, para este caso la del tipo euclideana. En la Figura 9 se muestra el código en R. Luego se realiza el proceso de clústering empleando la función hclust().

Figura 9

Cálculo de Matriz de Distancia y Aplicación de Algoritmo Hclust

```
> d <- dist(dtr1, method = "euclidean")
> hc1 <- hclust(d , method="ward.D2")
```

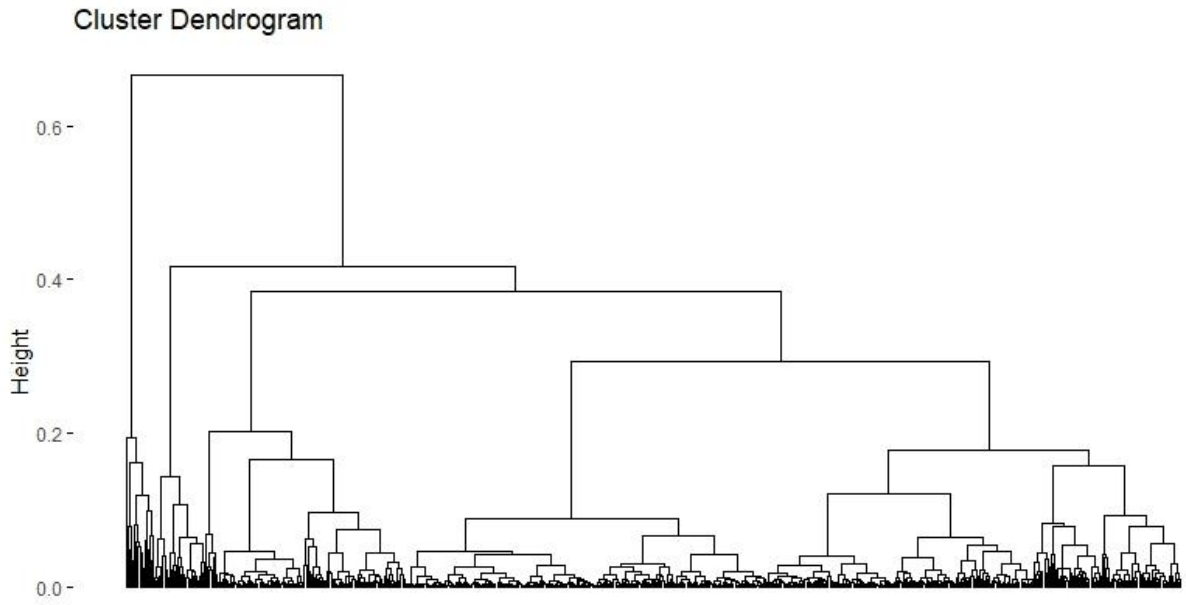
Nota. Se utilizó la liga de Ward, que minimiza el total de la varianza intra-clústeres.

La manera de interpretar la conformación de agrupamientos que generó el algoritmo es mediante el dendograma, al que se accede a través de la función fviz_dend() del paquete factoextra que se representa en la Figura 10. Cada hoja del árbol representa un objeto y a medida que se

asciende en la altura se expone el modo en que se fusionaron. La altura se denomina distancia cofenética entre dos objetos (Kassambara, 2017, pág. 72).

Figura 10

Dendograma de Hc1



Nota. Para mejor visualización se retiraron las etiquetas de los objetos.

Como se indicó en el punto 2.3.4 existen diferentes tipos de enlaces para medir la proximidad de los clústeres, estas son completa o máxima, única o mínima, promedio, centroide y de Ward. Una forma de averiguar cuál de las mencionadas mantiene mejor correlación con las distancias entre objetos es ejecutar el algoritmo con cada una de las medidas mencionadas, luego calcular su distancia cofenética y finalmente comparar la correlación con la matriz de distancias. A medida que el coeficiente se aproxime a 1 indica que la solución de clústering empleada es más precisa. Para implementar este procedimiento se utiliza la función cophenetic() y cor() obteniendo los resultados que se detallan en la Figura 11 y en la Tabla 3 respectivamente.

Figura 11

Cálculo de Distancia Cofenética y Correlación con Matriz de Distancia

```
> cof <- cophenetic(hc1)
> cor(d,cof)
```

Tabla 3

Resultados de Clústering Jerárquico con Diversos Enlaces

Tipo de enlace	Coeficiente de correlación
Completo	0.8300547
Único	0.910691
Promedio	0.9170063
Centroide	0.9116647
Ward	0.6901838

Nota. La liga promedio refiere mejor resultado

Una vez determinado el tipo de enlace a utilizar se puede cortar el árbol para separar los clústeres, mediante la altura o la cantidad de agrupaciones. Este proceso se implementa con la función cutree(). Obteniendo luego la representación del árbol con la función fviz_dend() que se detallan en las Figuras 12 y 13.

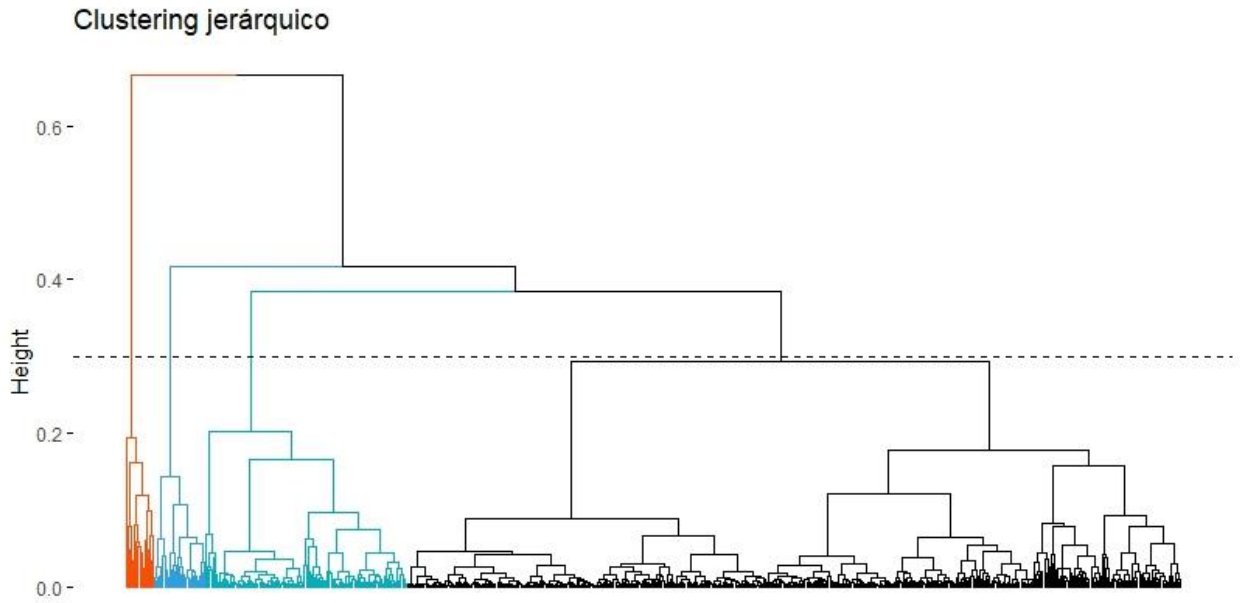
Figura 12

Código Generación de Dendograma

```
> clust<-cutree(hc1,k=4)
> clusterh <- cbind(dtr1, cluster=clust)
> fviz_dend(hc1, cex = 0.2, k = 4,
+           k_colors= c("#FC4E07","#2E9FDF","#00AFBB", "black" ),
+           show_labels = FALSE )+
+           labs(title = "Clustering jerárquico")+
+           geom_hline(yintercept = 0.3, linetype = "dashed")
```

Figura 13

Dendograma de Hc1 con 4 clústeres



A continuación, en la Tabla 4 se presentan los resultados de efectuar clústering jerárquico para 2 a 6 agrupaciones, indicado para cada una de ellas la cantidad de elementos asignados a los grupos y sus respectivos índices de silueta, además del factor de silueta promedio para cada corrida.

Tabla 4

Resultados de Clústering Jerárquico para K entre 2 y 6

		k2	k3	k4	k5	k6
	c1	618	587	466	216	216
Tamaño	**c2**	17	17	17	250	250
de	**c3**		31	31	17	17
clústeres	**c4**			121	31	31
	c5				121	113
	c6					8
	c1	0.7521773	0.5642331	0.379613	0.536515758	0.534851628

Indice de silueta de clústeres	**c2**	0.3335457	0.2102379	0.1987717	0.006629676	-0.007089077
	c3		0.3420638	0.3203493	0.198771653	0.059747991
	c4			0.2030784	0.242767481	0.240201938
	c5				0.126026077	0.187151696
	c6					0.413136192
Indice de silueta promedio		0.7409699	0.54391	0.3382395	0.2262975	0.2309778

4.5.3 Fuzzy Clústering

La técnica Fuzzy o difusa se puede implementar con la función eclust() del paquete factoextra, esta brinda la posibilidad de calcular el coeficiente de silueta y permitir ser graficado mediante ggplot2. En la Figura 14 se muestra el código de la función para 4 particiones, es de destacar que la asignación de los objetos al clúster expresa el coeficiente de su grado de pertenencia, y en la Figura 15 el porcentaje de asignación de los 10 primeros elementos a cada clúster.

Figura 14

Implementación de Fuzzy Clústering

```
> fn4 <- eclust(dtr1,"fanny", k=4,
+     hc_metric =  "euclidean",graph = FALSE, seed = 123)
```

Nota. Se aplica la función sobre el paquete de datos. K = 4

Figura 15

Asignación de Coeficientes con Fuzzy Clústering

```
> head(fn4$membership,10)
           [,1]      [,2]      [,3]      [,4]
 [1,] 0.3027529 0.2663332 0.2154564 0.2154575
 [2,] 0.3387911 0.2695171 0.1958452 0.1958465
 [3,] 0.3015194 0.2672981 0.2155907 0.2155918
 [4,] 0.2348449 0.2568047 0.2541749 0.2541755
 [5,] 0.3471582 0.2705208 0.1911598 0.1911612
 [6,] 0.3265655 0.2755857 0.1989237 0.1989252
 [7,] 0.2316689 0.2447279 0.2618018 0.2618014
 [8,] 0.2829074 0.2652080 0.2259418 0.2259428
 [9,] 0.2530942 0.2590665 0.2439193 0.2439199
[10,] 0.2937151 0.2692924 0.2184957 0.2184968
```

Nota. Primeras 10 observaciones del dataset fn4 generado con eclust()

Seguidamente, en la Tabla 5 se presentan los resultados obtenidos sobre la técnica difusa para valores de k desde 2 a 6, sus índices de silueta y el promedio.

Tabla 5

Resultados de Fuzzy Clústering para K entre 2 y 6

		k2	**k3**	**k4**	**k5**	**k6**
Tamaño de clústeres	**c1**	357	340	294	276	268
	c2	278	57	74	62	61
	c3		238	223	227	236
	c4			44	70	67
	c5					3
	c6					
Indice de silueta de clústeres	**c1**	0.6638932	0.2885444	0.29163785	0.28374753	0.28895498
	c2	-0.1189353	0.1290874	0.09745353	-0.05108967	-0.02100228
	c3		-0.2324099	-0.22435075	-0.22116139	-0.26169623
	c4			0.2030784	0.09511793	0.03764161
	c5					-0.16241953
	c6					
Indice de silueta promedio		0.3211746	0.07897563	0.0635978	0.04976595	0.02587908

4.5.4 Clústering basado en densidad: DBSCAN

La siguiente técnica, llamada DBSCAN Density-Based Spatial Clustering of Applications with Noise, requiere que se determine el radio o umbral y la cantidad mínima de puntos por clúster. Para lograr una aproximación al valor del radio, se puede utilizar la función KNNdisplot() integrada en el paquete dbscan según se muestra en la Figura 16, que proporciona una curva con el promedio de las distancias de cada punto a sus k vecinos más cercanos, en este caso la distancia resulta próxima a 0.06, según se grafica en la Figura 17.

Figura 16

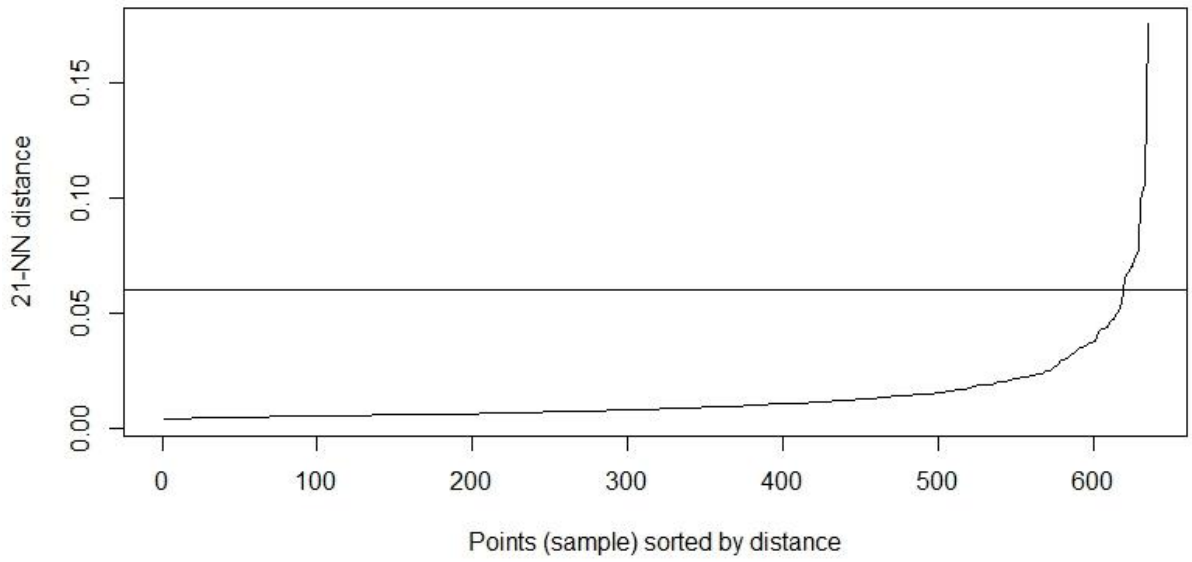
Determinación del Radio de Clústeres

```
> dbscan::kNNdistplot(dtr1, k=21)
> abline(h=0.06,lty=1)
```

Nota. Se aplica la función sobre el paquete de datos. K = dimensiones + 1.

Figura 17

Curva de Distancia del Dataset



Seguidamente se ejecuta la función dbscan() del paquete fpc para realizar el proceso de clústering, asignando el valor de épsilon aproximado en el paso anterior y la cantidad mínima de objetos para considerar un agrupamiento, posteriormente se puede graficar la asignación con la función hullplot() del paquete dbscan arrojando los resultados que se detallan en la Figuras 18 y 19.

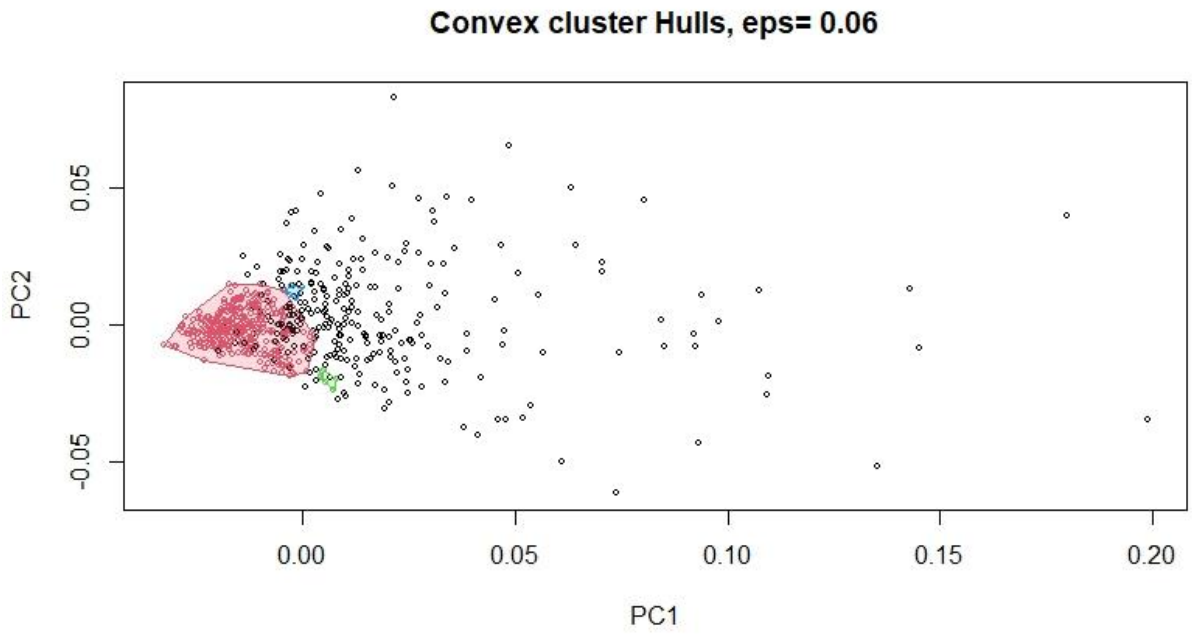
Figura 18

Implementación de Clústering DBSCAN

```
> set.seed(123)
> f<- fpc::dbscan(dtr1 , eps = .006 , MinPts = 5)
> f
dbscan Pts=635 MinPts=5 eps=0.006
         0   1 2 3
border 269  51 6 4
seed     0 303 1 1
total  269 354 7 5
>
> hullplot(dtr1,f$cluster, main = "Convex cluster Hulls, eps= 0.06")
```

Figura 19

Gráfico de Clústering con DBSCAN



Nota. Los puntos rojos, verdes y azules representan los objetos agrupados en clústeres, los puntos de color negro se consideran ruido.

Evaluación de clústeres

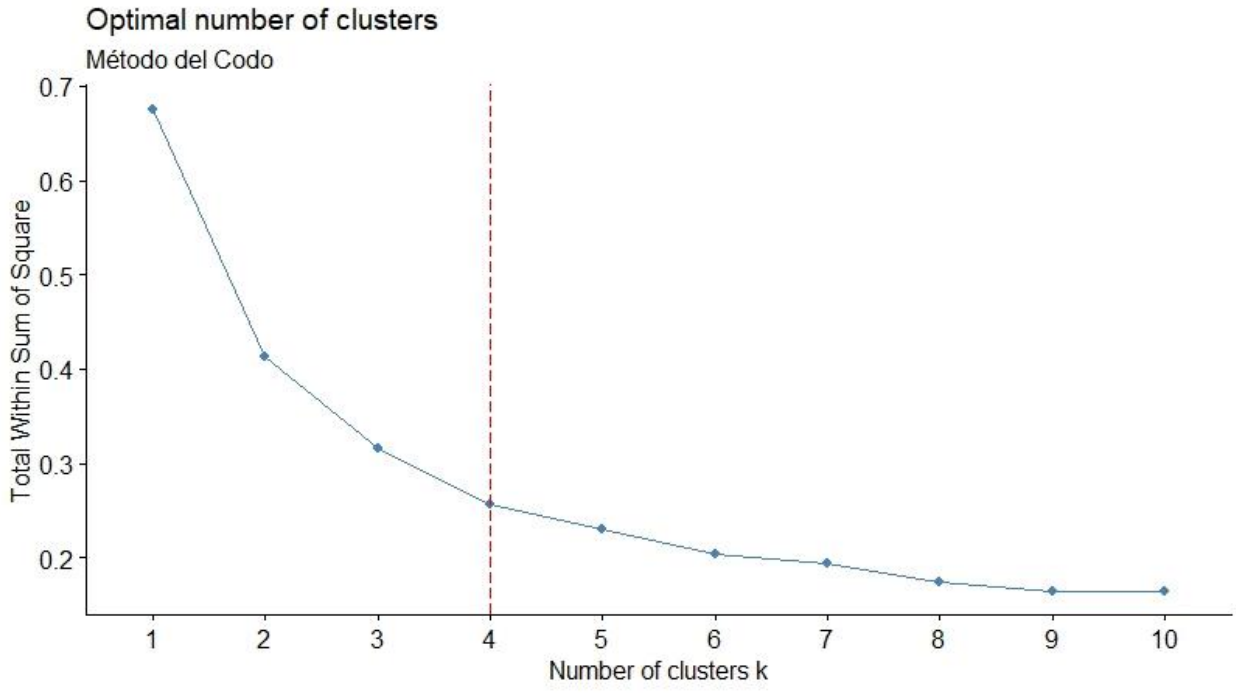
Según se expone en el punto 2.3.5, la evaluación de los aglomerados que generó la minería de datos tiene varias aristas por analizar, la primera es la tendencia del agrupamiento, que se llevó

a cabo al inicio del proceso para determinar si la distribución de los datos resulta o no homogénea y consecuentemente si resulta válido aplicar las técnicas de clústering. En este caso el coeficiente de Hopkins arrojó un valor alejado de 0.5, lo que permite interpretar que la distribución de los datos no resulta homogénea y por lo tanto es susceptible de ser separada en grupos con algún nivel de coherencia.

Otra medida de evaluación es la determinación del número óptimo de clústeres en los que se particiona el dataset. Para ello se utiliza una medida denominada WSS que es la suma total de cuadrados en un clúster y representa la compactación de los datos en un grupo. Para ello se traza una curva en la que el eje de las abscisas corresponde a la cantidad de clústeres que se van a evaluar y el eje de las ordenadas al valor WSS para cada k. El gráfico se interpreta en el siguiente sentido, a medida que aumentan los clústeres disminuye el coeficiente WSS, la curva adquiere una forma de codo cuando la pendiente no resulta significativa con respecto al siguiente valor de k. En la Figura 20 se representa el gráfico resultante del método del codo con el algoritmo k-means.

Figura 20

Gráfico de Elbow Method con K-means

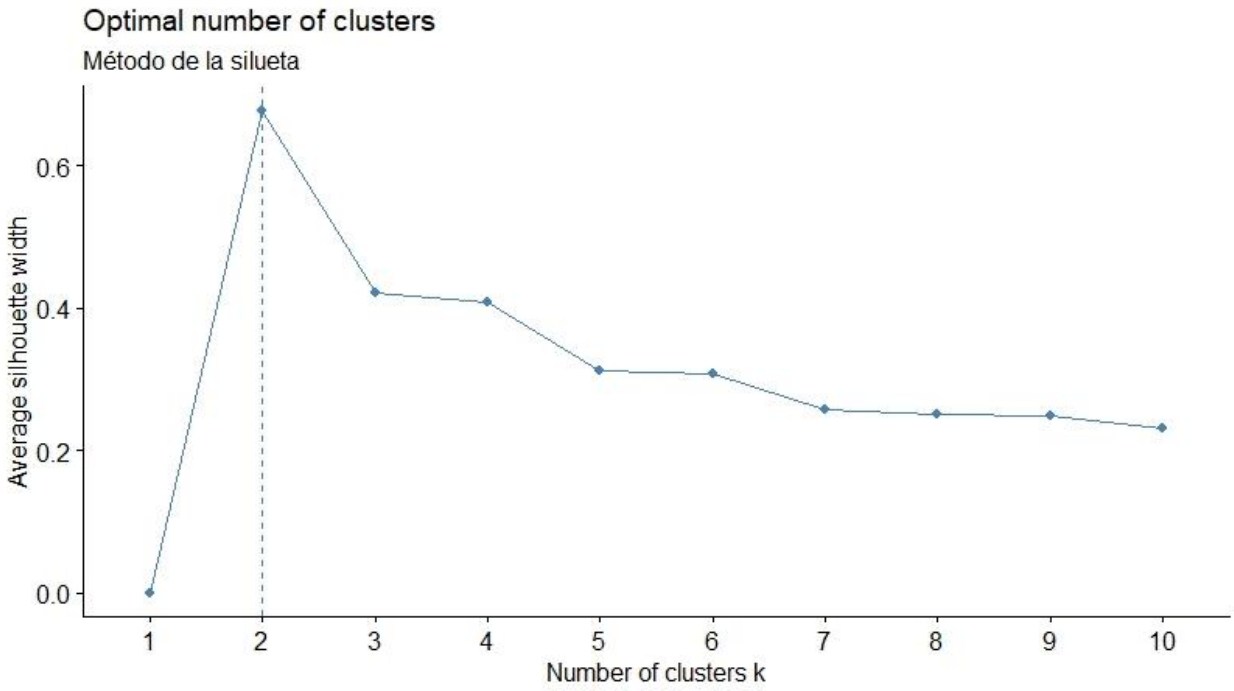


Nota. El punto de inflexión está próximo a k=4

El siguiente método para determinar el número más apropiado de clústeres es el que utiliza el coeficiente de silueta que se expresa en el gráfico de la Figura 21.

Figura 21

Gráfico de la Silueta con K-means

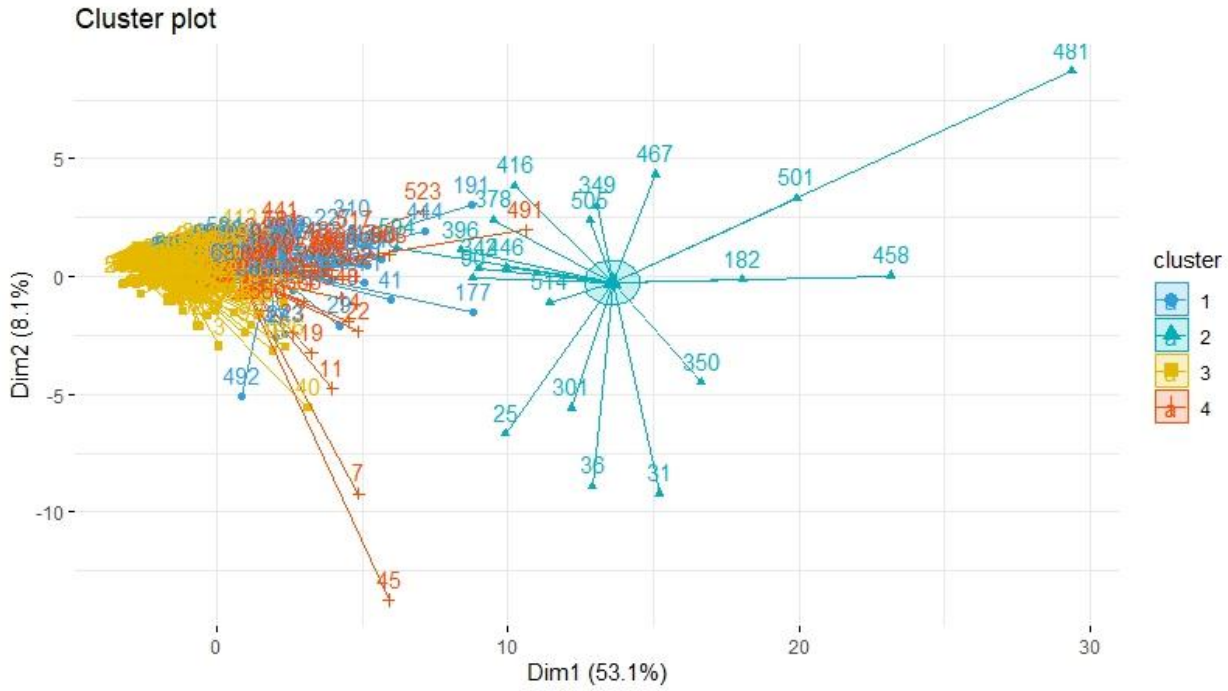


Nota. El algoritmo arroja 2 clústeres como medida óptima.

Si bien el método del codo aproxima a 4 la cantidad correcta de clústeres para minimizar el WSS intra clústeres y el método de la silueta expone que el valor óptimo es 2, ambas medidas son aproximativas y debe tenerse en cuenta el problema que se pretende resolver. En el caso de negocio que se estudia, dividir 635 sucursales en dos grupos no resulta suficientemente explicativo sobre las particularidades de las diversas casas, ya que separaría en dos clústeres de 41 y 594 sucursales respectivamente. La opción de 4 clústeres brinda una respuesta más acorde a las necesidades del negocio con agrupamientos de 20, 80, 126 y 409 casas. Según se grafica en la Figura 22.

Figura 22

K-means de Sucursales en 4 Grupos



Nota. Se grafican las 2 dimensiones más significativas

Con respecto al desempeño del algoritmo DBSCAN, no resulta conveniente su aplicación en este problema específico de negocio, debido a que la asignación de los objetos no es completa, es decir contempla como ruido observaciones correspondientes a sucursales en los que la vecindad o radio no se puede establecer suficientemente acotada para generar una cantidad de clústeres manejables o lógicos. Para un radio de 0.06 y una cantidad mínima de puntos de 5, se establecieron 3 agrupaciones con un tamaño de 354, 7 y 5 puntos, detectando 269 objetos como ruido. Este inconveniente se debe a la variación de la densidad en la población estudiada (Tan y otros, 2006, pág. 532) .

Los gráficos de silueta promedio implementados mediante la función fviz_silhouette() del paquete factoextra, representan de manera visual los valores que se detallaron en las Tablas 2, 4 y 5, demostrando que, para agrupaciones de 4 clústeres el método más apropiado resulta ser k-means. Seguidamente se presentan las figuras 23, 24 y 25 con lo expresado precedentemente.

Figura 23

Silueta Promedio K-means 4 Clústeres

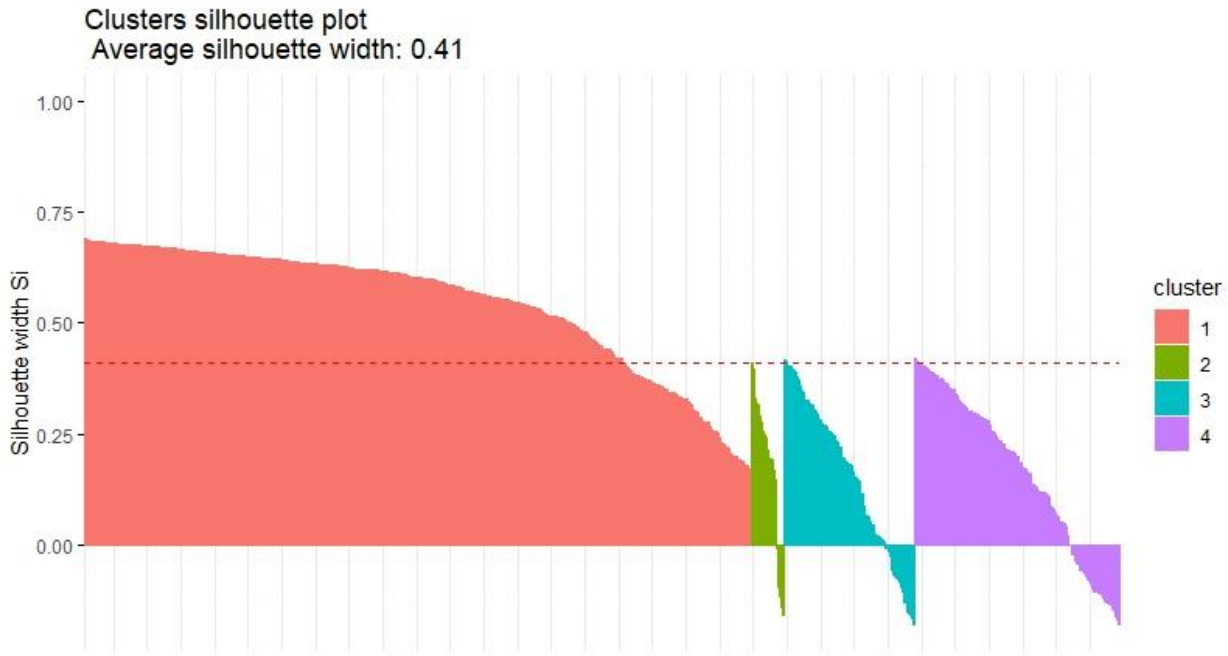


Figura 24

Silueta Promedio Método Jerárquico 4 Clústeres

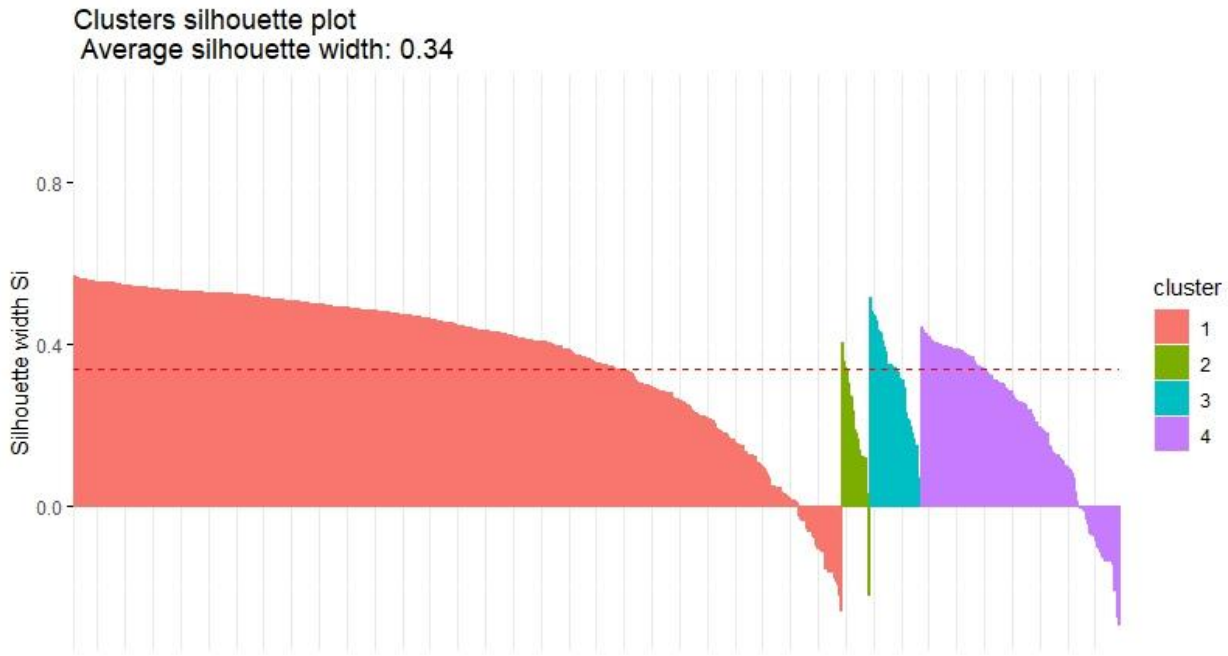
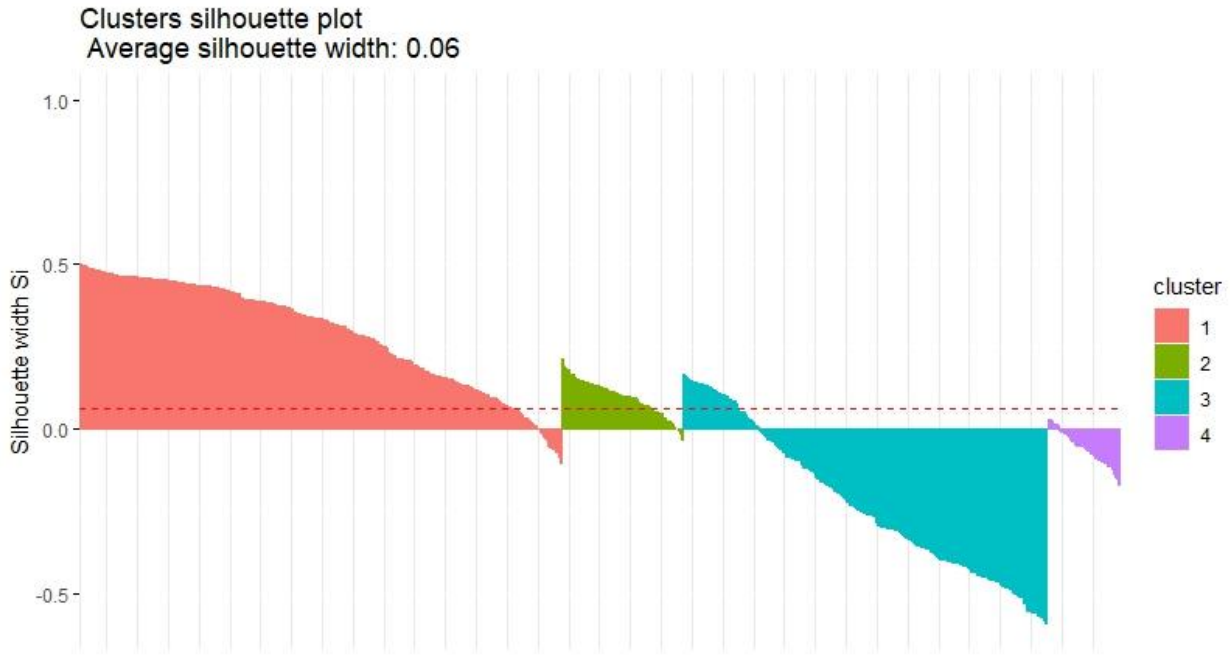


Figura 25

Silueta Promedio Fanny 4 Clústeres



Es preciso destacar que un mayor ancho promedio de silueta significa que la asignación de los elementos en cada clúster es más apropiada, para las técnicas estudiadas el resultado arroja los siguientes valores: K-means obtuvo 0.41, método jerárquico arrojó 0.34 y Fanny logró 0.06. Adicionalmente, en las gráficas se pueden observar valores menores a 0, lo que indica que esos puntos podrían pertenecer a algún otro clúster por presentar similitudes con ellos, en consecuencia, los picos hacia abajo denotan una mala agrupación. Uniendo ambos conceptos la evidencia indica que la aplicación de K-means es más adecuada para segmentar este paquete de datos en 4 clústeres.

En resumen, una vez obtenido el dataset se evaluó la tendencia de agrupamiento para verificar que éste fuera agrupable de manera no aleatoria mediante el índice de Hopkins que arrojó un resultado de 0.9261, lo cual indica que la distribución de los objetos no es uniforme. Luego se procedió a la aplicación de las diferentes técnicas de agrupamiento para 2 a 6 clústeres, sus resultados fueron volcados en las tablas 2, 3 y 4. Por último se realizó la evaluación de los clústeres obtenidos con cada técnica para determinar según las métricas de coeficiente de silueta promedio y WSS, cuál de ellas resulta más adecuada al paquete de datos y al problema de negocio planteado.

Conclusiones

En el presente apartado se exponen las conclusiones arribadas luego de la investigación. En primera instancia se puede colegir que los objetivos planteados oportunamente han sido alcanzados de manera satisfactoria pudiéndose corroborar que la aplicación de técnicas de data mining aportan valor en la toma de decisiones empresariales, ofreciendo información que de otra manera no resulta intuitiva o es demasiado extensa para la comprensión intelectual directa. En este caso particular, la manipulación y extracción de conocimiento sobre el paquete de datos que contiene 635 observaciones de 20 dimensiones obtenido mediante la aplicación de KDD, es factible gracias a la capacidad de análisis que ofrece la minería de datos. En este caso de negocio se logró agrupar en 4 clústeres significativos las filiales bancarias estudiadas.

Con respecto a la hipótesis planteada, en la que se proponía que la técnica de data mining con mejor desempeño en una solución de business intelligence resultaba ser Clústering Jerárquico, de acuerdo con los datos obtenidos durante el proceso de la investigación, se evidenció que no fue apoyada. Es decir que, de la aplicación de las cuatro técnicas establecidas oportunamente, (k-means, dbscan, método jerárquico aglomerativo y técnica difusa) la que mejor resultado obtuvo, en base a las métricas de coeficiente de silueta promedio y WSS, fue k-means con un promedio de silueta de 0.41 mientras que el método jerárquico alcanzó 0.34, ambos con 4 clústeres.

No obstante lo expuesto, es necesario destacar que no existe una técnica de clústering con desempeño absoluto. Esta debe ser evaluada previa y posteriormente a su aplicación, teniendo en cuenta el tipo de problema que se desea resolver y las características de los datos con los que se cuentan: su volumen, variedad y veracidad. Destacando además que la distribución de las observaciones resulta ser un factor fundamental para determinar si el sometimiento de los datos a estas prácticas arrojará un resultado coherente, ya que la aplicación de los algoritmos sobre objetos distribuidos uniformemente particiona el espacio de manera aleatoria con un resultado distinto cada vez.

En cuanto al primer objetivo particular de determinación de las variables más significativas, se logró gracias a la aplicación del método Knowledge Discovery in Databases, que apunta a la comprensión del negocio y a la resolución del problema, obteniéndose 20 variables que representan la actividad comercial de las filiales bancarias estudiadas.

El tercer objetivo particular, implementar las técnicas de clústering indicadas previamente, se ha podido alcanzar mediante el uso de la herramienta RStudio y la ejecución de las funciones desarrolladas en las bibliotecas de software libre, las cuales arrojaron las métricas que permitieron la comparación posterior de los conglomerados descubiertos con los diferentes algoritmos, facilitando de esta manera la prosecución de la investigación con el cuarto objetivo particular tendiente a la evaluación de los resultados de las técnicas de minería de datos puestas a prueba, que también fue logrado.

Por último, se entiende que la investigación desarrollada en el presente trabajo final ofrece al lector un abordaje general de los conceptos de big data, business intelligence y data mining, profundizando en las técnicas de minería de datos basadas en agrupamiento y ofreciendo un estudio comparativo de técnicas de clústering sobre un caso específico de negocio apoyado en la metodología KDD. Se espera que la misma logre incentivar al uso de las tecnologías de la información para el descubrimiento de patrones de comportamiento en grandes y variados volúmenes de datos.

Líneas Futuras de Investigación

Se propone como línea futura de investigación, la ampliación del ámbito de estudio incorporando otros algoritmos, como ser la combinación de dos métodos, por ejemplo los particionales y jerárquicos en hierachical k-means, o la incorporación de clústering en bases de datos temporales, espaciales o con contenido multimedia.

Adicionalmente, para el caso de negocio estudiado, podrían incluirse otras variables relacionadas con la adopción de las nuevas tecnologías por parte de los clientes, teniendo en cuenta el vuelco generalizado hacia las aplicaciones fintech que está adoptando el mercado.

Acrónimos

BI	**Business Intelligence**
BPM	**Business Process Management**
CRM	**Customer Relationship Management**
DM	**Data Mining**
DSS	**Decision Support System**
DW	**Data Warehouse**
ETL	**Extracción, Transformación y Carga**
IDC	**International Data Corporation**
KPI	**Key Performance Indicator**
NoSQL	**No solo SQL**
OLAP	**Online Analytical Processing**
OLTP	**On-Line Transaction Processing**
PCA	**Análisis de Componentes Principales**
RDBMS	**Relational Data Base Management System**
TIC	**Tecnologías de la Información y la Comunicación**
WITTS	**Workshop de Tecnologías de la Información y Sistemas**
WSS	**Within Sum Square (Suma de cuadrados intra clúster)**

Anexo 1

```
pkgs <- c("factoextra" , "FactoMineR", "readr", "rgl", "clValid")
pkgs1 <-c("NbClust" , "cluster" , "fpc" , "dendextend")
install.packages(pkgs)
install.packages(pkgs1)
library(factoextra);library(FactoMineR);library(readr);library(rgl)
library(clValid);library(NbClust);library(cluster)
library(fpc);library(dendextend)

#Cargar el archivo CSV
setwd("C:/Users/ledav/Documents")
dtr <- read.csv("Sucursales01.csv", header = T)

#HOPKINS valores cercanos a 1 siginifica que el dataset es clusterizable,
la imagen tiene que tener cuadrados azules
get_clust_tendency(dtr1, n = 50,
      gradient = list(low = "steelblue",  high = "white"))

## k-means
set.seed(123)
k2 <- kmeans(x = dtr1 , centers = 2, nstart = 100)
k3 <- kmeans(x = dtr1 , centers = 3, nstart = 100)
k4 <- kmeans(x = dtr1 , centers = 4, nstart = 100)
k5 <- kmeans(x = dtr1 , centers = 5, nstart = 100)
k6 <- kmeans(x = dtr1 , centers = 6, nstart = 100)

k2
k2$withinss
k2$totss
k2$betweenss

k3
k3$withinss
k3$betweenss
k3$totss

k4
k4$withinss
k4$betweenss
k4$totss
```

```
k5
k5$withinss
k5$betweenss
k5$totss

k6
k6$withinss
k6$betweenss
k6$totss

d <- dist(dtr1, method = "euclidean")
d
hc1 <- hclust(d , method="ward.D2")
hc1

fviz_dend(hc2, cex = 0.5,show_labels = FALSE)
plot(hc1,cex=1, hang = -1)

#distancia cofenética

hc1 <- hclust(d , method="ward.D2")
cof <- cophenetic(hc1)
cor(d,cof)

hc2 <- hclust(d , method="average")
cof <- cophenetic(hc2)
cor(d,cof)

hc3 <- hclust(d , method="complete")
cof <- cophenetic(hc3)
cor(d,cof)

hc4 <- hclust(d , method="single")
cof <- cophenetic(hc4)
cor(d,cof)

hc5 <- hclust(d , method="centroid")
cof <- cophenetic(hc5)
cor(d,cof)

hc6 <- hclust(d , method="median")
```

```
cof <- cophenetic(hc6)
cor(d,cof)

clust2 <- cutree(hc1,k=2)
table(clust2)

clust3 <- cutree(hc1,k=3)
table(clust3)

clust4 <- cutree(hc1, k=4)
table(clust4)


clust5 <- cutree(hc1, k=5)
table(clust5)

clust6 <- cutree(hc1, k=6)
table(clust6)

fviz_dend(hc1, cex = 0.2, k = 4,
          k_colors= c("#FC4E07","#2E9FDF","#00AFBB", "black" ),
          show_labels = FALSE )+
  labs(title = "Clustering jerárquico")+
  geom_hline(yintercept = 0.3, linetype = "dashed")

##otro algoritmo
hc2 <- eclust(dtr1, "hclust", k = 3, graph = FALSE)
fviz_dend(hc2, rect = TRUE, show_labels = FALSE)

fviz_silhouette(hc2)

fviz_cluster(list(data=dtr1, cluster=clust))

clusterh <- cbind(dtr1, cluster = clust)
View(clusterh)
?fviz_dend
fviz_dend(hc1)
fviz_dend(hc1, cex = 0.5, k = 3, color_labels_by_k = TRUE)+ labs(title =
"Clustering jerárquico")
+ geom_hline(yintercept = 50, linetype = "dashed")

km2 <- eclust(dtr1, "kmeans", k = 2, graph = FALSE, seed = 123)
```

```
km2$size
km2$silinfo

km3 <- eclust(dtr1, "kmeans", k = 3, graph = FALSE, seed = 123)
km3$size
km3$silinfo

km4 <- eclust(dtr1, "kmeans", k = 4, graph = FALSE, seed = 123)
fviz_silhouette(km)
km4$size
km4$silinfo

km5 <- eclust(dtr1, "kmeans", k = 5, graph = FALSE, seed = 123)
km5$size
km5$silinfo

km6 <- eclust(dtr1, "kmeans", k = 6, graph = FALSE, seed = 123)
km6$size
km6$silinfo

## algoritmo eclust con fviz_silhouette

hc2 <- eclust(dtr1,"hclust", k=2, hc_method = "ward.D2")
fviz_silhouette(hc2)
hc2$silinfo
hc2$size

hc3 <- eclust(dtr1,"hclust", k=3, hc_method = "ward.D2")
fviz_silhouette(hc3)
hc3$silinfo
hc3$size

hc4 <- eclust(dtr1,"hclust", k=4, hc_method = "ward.D2")
fviz_silhouette(hc4)
hc4$silinfo
names(silinfo)
hc4$size
hc4$nbclust

head(silinfo$widths[,1:3],10)
silinfo$clus.avg.widths
silinfo$avg.width
```

```
      hc5 <- eclust(dtr1,"hclust", k=5, hc_method = "ward.D2")
      fviz_silhouette(hc5)
      hc5$silinfo
      hc5$size

      hc6 <- eclust(dtr1,"hclust", k=6, hc_method = "ward.D2")
      fviz_silhouette(hc6)
      hc6$silinfo
      hc6$size

      #Método basado en densidad

      dbscan::kNNdistplot(dtr1, k=21)
      abline(h=0.06,lty=1)

      set.seed(123)
      f<- fpc::dbscan(dtr1 , eps = .006 , MinPts = 5)
      f
      ?hullplot
      hullplot(dtr1,f$cluster, main = "Convex cluster Hulls, eps= 0.06")

      db <- dbscan::dbscan(dtr1 , eps = 0.01, minPts = 21)
      db

      ?fviz_cluster()
      fviz_cluster(f, dtr1, geom= "point")
      View(f)
      f$cluster

      ##Método Fuzzy

      fn2 <- eclust(dtr1,"fanny", k=2, hc_metric =  "euclidean",graph = FALSE,
seed = 123)
      fn2$silinfo
      datasetcluster <- cbind(dtr1,fn2$clustering)
      count(datasetcluster,21)
      fviz_silhouette(fn2)

      fn3 <- eclust(dtr1,"fanny", k=3, hc_metric =  "euclidean",graph = FALSE,
seed = 123)
      fn3$silinfo
```

```
      fviz_silhouette(fn3)

      fn4 <- eclust(dtr1,"fanny", k=4,
                    hc_metric =  "euclidean",graph = FALSE, seed = 123)
      fn4$silinfo
      fviz_silhouette(fn4)
      head(fn4$membership,10)

      fn5 <- eclust(dtr1,"fanny", k=5, hc_metric =  "euclidean",graph = FALSE,
seed = 123)
      fn5$silinfo
      fviz_silhouette(fn5)
      fn5$membership

      fn6 <- eclust(dtr1,"fanny", k=6, hc_metric =  "euclidean",graph = FALSE,
seed = 123)
      fn6$silinfo
      fviz_silhouette(fn6)
```

Referencias

Chen, H., Chiang, R., & Storey, V. (December de 2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly, 36*(4), 1165-1188. http://www.jstor.org/stable/41703503

Davenport, T. (January de 2006). Competing on Analytics. *Harvard Business review*, 98-107.

Diaz, J., Osorio, M., & Amadeo, A. (2019). *Tecnologías para el análisis de datos basadas en software libre.* La Plata: Editorial de la Universidad de La Plata. www.editorial.unlp.edu.ar

Fayyad, U., Piatetsky-Shapiro, G., & Padhraic, S. (1996). From Data Mining to Knowledge Discovery in Databases. *AI MAGAZINE*, 37-54.

Ferrari, A., & Russo, M. (2008). *Introduction to the SQLBI*. sqlbi: https://www.sqlbi.com/wp-content/uploads/Introduction-to-SQLBI-Methodology-draft-1.0.pdf

Gartner. (2021). *Cuadrante Mágico para Plataformas de Análisis e Inteligencia de Negocios.* https://analyticslatam.com/wp/wp-content/uploads/2021/03/Cuadrante-Magico-de-Gartner.pdf

Gutierrez Puebla, J. (2017). La huella digital de las actividades humanas. *Big Data y nuevas geografías* (pág. 23). Barcelona: Universidad Autónoma de Barcelona. https://doi.org/https://doi.org/10.5565/rev/dag.526

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques.* San Diego: Academic Press.

Heavy.AI. (09 de 11 de 2021). *Inteligencia de negocios*. https://www.heavy.ai/technical-glossary/business-intelligence

Hernández Orallo, J., Ramirez Quintana, M., & Ferri Ramirez, C. (2004). *Introducción a la minería de datos.* Madrid: Pearson Educación S.A.

Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucío, P. (1991). *Metodología de la Investigación.* México: Mc Graw-Hill Interamericana de México S.A de C.V.

Inmon, W. (2005). *Building the DataWarehouse.* Indianapolis: Wiley Publishing.

International Data Corporation [IDC]. (2021). *How Data Culture Fuels Business Value in Data-Driven Organizations.* https://489b825afrm2p9lw23oekybe-wpengine.netdna-ssl.com/wp-content/uploads/2021/07/Tableau_WhitePaper_IDC_English.pdf

Kassambara, A. (2017). *Practical guide to cluster analysis in R. Unsupervised mahine learning.* STHDA. http://www.sthda.com

Kimball, R., & Ross, M. (2013). *The data warehouse toolkit: The complete guide to dimensional modelling* (3era ed.). Indiana: Willey Computer Publishing.

Provost, F., Bernstein, A., & Grosof, B. (2001). Business intelligence: The next Frontier for information systems research? *Workshop on Information Technologies and Systems.* New Orlean, USA.

Shim, J., Warkentin, M., Courtney, J., Power, D., Sharda, R., & Carlsson, C. (2002). Past, present, and future of decision support technology. *Decision Support Systems, 33*, 111-126. https://doi.org/10.1016/S0167-9236(01)00139-7

Stedman, C. (2022). *The ultimate guide to big data for businesses.* TechTarget.com: https://www.techtarget.com/searchdatamanagement/The-ultimate-guide-to-big-data-for-businesses

Tan, P.-N., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining.* Pearson Addison-Wesley.

Witten, I., & Frank, E. (2000). *Data Mining. Practical Machine Learning Tools and Techniques with Jaca Implementations.* Morgan Kaufmann Publishing.