Grado en Estadística
Trabajo de Fin de Grado

Retos en Machine Ciencia de Datos: Resolución de un caso de estudio mediante el uso de técnicas estadísticas y de Machine Learning

Patricia Contreras Parra

UNIVERSIDAD DE GRANADA

Departamento de Ciencias de la Computación e I.A.

Universidad de Granada

Tutor

Alberto Luis Fernández Hilario

Granada, julio 2024

Grado en Estadística
Trabajo de Fin de Grado

Retos en Machine Ciencia de Datos: Resolución de un caso de estudio mediante el uso de técnicas estadísticas y de Machine Learning

UNIVERSIDAD DE GRANADA

Declaro explícitamente que el trabajo presentado es original, entendido en el sentido de que no he utilizado fuentes sin citarlas debidamente.

Patricia Contreras Parra

Autoevaluación

Esta autoevaluación está dedicada a la exposición de los esfuerzos, logros y dificultades que han supuesto elaborar este Trabajo Final de Grado.

Antes de comenzar quiero poner en contexto al lector. Este Trabajo Final de Grado lo elegí personalmente por sus características únicas con respecto al resto de propuestas. Principalmente habían dos motivos, enfrentarme a un problema con datos reales de los cuales desconocía y utilizar el famoso lenguaje de programación Python para resolverlo.

La primera etapa del trabajo se basó en una comunicación continua y efectiva con el tutor para consensuar el tipo de problema que se pretendía plantear para este trabajo. Mi propósito era resolver un problema relacionado con los problemas de aprendizaje no supervisado, más en concreto un problema relacionado con el uso de técnicas clustering, ambos temas contaban con mi total atención al no poseer las mismas propiedades que poseen los problemas de aprendizaje supervisado y además de ser un tema del que desconocía al no ser tan sonado en el grado.

La segunda etapa de este trabajo está relacionada con la búsqueda de datos en la plataforma Kaggle. Este proceso no resultó fácil ya que esta búsqueda requería encontrar una base de datos de un número adecuado de individuos para estudiar el fenómeno de interés. Además de buscar un conjunto de datos con variables relevantes que caracterizaran a estos individuos. La calidad de los datos fue otra cuestión importante que se tuvo en cuenta en la búsqueda. Finalmente, se encontró una base de datos que se moldeaba al problema planteado.

Continuamos a la tercera etapa de este trabajo. Esta etapa se caracterizó por sentar las bases del mismo, es decir, se procedió a la documentación tanto de qué es la ciencia de datos, los algoritmos de aprendizaje automático (más en concreto, el aprendizaje no supervisado), las técnicas de clustering y, por último pero no menos importante, el uso de Python como herramienta para el análisis y presentación de los resultados. Considero esta etapa como la que supuso un reto de mayor dificultad con respecto a las demás, por las siguientes razones: *(1)* La búsqueda de informes que fueran eficientes para documentarme sobre las técnicas de agrupamiento jerárquico, las cuales parecían escasas en un principio, lo que me llevó a recurrir a la literatura del grado y, a partir de ahí, pude realizar una búsqueda más lógica de los temas en cuestión. *(2)* Aprender el lenguaje de programación Python, del que tenía un conocimiento básico. Esta ha resultado la tarea más laboriosa de todas, ya que al principio no era una tarea intuitiva para mí, pero gracias a la documentación y diversos foros en internet, pude seguir adelante y conocer mejor este lenguaje de programación. Tampoco puedo olvidar las recomendaciones bibliográficas y la eficaz comunicación por parte del tutor, que han sido de gran ayuda en los momentos de posibles atascos e inconvenientes, consiguiendo que todo continúe según lo indicado.

Una vez que conocía y disponía de la documentación necesaria para afrontar este trabajo, se procedió primero, a implementar los códigos en Python para la obtención de resultados, y segundo, a la redacción del presente Trabajo Final de Grado. La estructura de este trabajo no tendría sentido sin la colaboración del tutor quién ha sido el director y guionista en cada punto de este trabajo.

En general, estoy muy satisfecha con el resultado final. Este Trabajo Final de Grado ha permitido que investigue un campo que me llamaba la atención cuando lo conocí en el grado. El hecho de poder poner en práctica todo lo aprendido y obtener un informe profesional que se adecúa a una estructura clara y concisa es reconfortante. La labor de documentarme, aspirar a conocer Python, una herramienta de la cual desconocía, y el querer afrontar un problema de datos desconocidos han supuesto un reto que ha generado

una mayor confianza y seguridad. Dado los objetivos iniciales y observando el resultado final, considero que el trabajo presentado se adecúa bastante a los objetivos propuestos, además de tener una calidad profesional tras describir cada una de las técnicas utilizadas con la máxima precaución para que en todo momento el lenguaje sea preciso, claro y objetivo. En conclusión, ha sido un Trabajo Final de Grado en el que se ha intentado abordar cada uno de los temas mencionados con la mayor apreciación posible.

Resumen

La inteligencia artificial (IA) ha emergido como una tecnología disruptiva que está transformando todas las áreas de la ciencia y la industria. La capacidad de las máquinas para analizar grandes volúmenes de datos, aprender patrones y tomar decisiones basadas en algoritmos sofisticados ha optimizado los procesos en múltiples sectores. Este avance se debe no solo al incremento masivo de los datos digitales y a la mayor capacidad de computación, sino también a las innovaciones en los algoritmos de inteligencia artificial y el aprendizaje automático.

La ciencia de datos, el trabajo estadístico y el aprendizaje automático se destacan como áreas críticas en nuestra era de automatización y gestión de grandes volúmenes de datos. La ciencia de datos integra estadística, matemáticas y ciencia de la computación para diseñar, recopilar, analizar e interpretar datos, mientras que, el trabajo estadístico extrae patrones y tendencias significativas mediante análisis descriptivo e inferencia estadística y el aprendizaje automático diseña algoritmos que aprenden de los datos. Estas disciplinas tienen una aplicabilidad directa en sectores como las matemáticas, estadística, informática, ingeniería, ciencias y finanzas.

En la ciencia de datos, el principal desafío radica en el análisis matemático de los datos. Cuando el objetivo es interpretar el modelo y cuantificar la incertidumbre en los datos, este análisis suele denominarse aprendizaje estadístico. En cambio, cuando el énfasis está en hacer predicciones utilizando datos a gran escala, se habla de aprendizaje automático. Existen dos objetivos principales en el modelado de datos: predecir con precisión canti-

dades futuras de interés a partir de datos observados y descubrir patrones inusuales o interesantes. Estos objetivos se conocen respectivamente como aprendizaje supervisado y no supervisado. En el aprendizaje supervisado, podemos comprobar nuestro trabajo viendo qué tan bien nuestro modelo predice la respuesta, lo que lo convierte en un área bien comprendida. Por el contrario, el aprendizaje no supervisado es más desafiante, ya que no existe un objetivo simple como la predicción de una respuesta y el ejercicio tiende a ser más subjetivo.

Nuestro objetivo es aplicar el ciclo de vida de la ciencia de datos en un caso de segmentación, utilizando métodos de aprendizaje no supervisado y el lenguaje de programación Python, para analizar un caso de estudio concreto: la segmentación de clientes de iFood, la principal aplicación de entrega de comida a domicilio en Brasil. Este estudio tiene como finalidad no solo aplicar la teoría de técnicas de segmentación, sino también desarrollar soluciones efectivas y prácticas para abordar los desafíos reales del sector.

La aplicación de técnicas de aprendizaje no supervisado en nuestro estudio ha ofrecido múltiples beneficios. Estos métodos no solo han permitido una segmentación eficaz de los clientes, revelando patrones y correlaciones previamente indetectables, sino que también han mejorado nuestra capacidad de tomar decisiones basadas en datos. A través de este enfoque, hemos podido extraer valor de grandes volúmenes de información de manera no intrusiva, maximizando así la eficiencia y efectividad de nuestras estrategias de análisis. La utilización de estas técnicas nos ha proporcionado una ventaja significativa en el entendimiento de comportamientos y preferencias del consumidor, lo cual es esencial para la adaptación y evolución en un mercado competitivo.

Summary

Artificial intelligence (AI) has emerged as a disruptive technology that is transforming all areas of science and industry. The ability of machines to analyse large volumes of data, learn patterns, and make decisions based on sophisticated algorithms has optimised processes in many sectors. This progress is due not only to the massive increase in digital data and greater computing power but also to innovations in artificial intelligence algorithms and machine learning.

Data science, statistical work, and machine learning stand out as critical areas in our age of automation and big data management. Data science combines statistics, mathematics, and computer science to design, collect, analyse, and interpret data, while statistical work extracts significant patterns and trends through descriptive analysis and statistical inference. Machine learning designs algorithms that learn from the data. These disciplines have direct applicability in fields such as mathematics, statistics, computer science, engineering, science, and finance.

In data science, the main challenge is the mathematical analysis of data. When the goal is to interpret the model and quantify uncertainty in the data, this analysis is often called statistical learning. On the other hand, when the focus is on making predictions using large-scale data, it is referred to as machine learning. There are two main objectives in data modelling: accurately predicting future quantities of interest from observed data and discovering unusual or interesting patterns. These objectives are known respectively as supervised and unsupervised learning. In supervised learning, we can check our work

by seeing how well our model predicts the response, making it a well-understood area. In contrast, unsupervised learning is more challenging as there is no simple goal like predicting a response, and the exercise tends to be more subjective.

Our goal is to apply the data science lifecycle in a segmentation case, using unsupervised learning methods and the Python programming language, to analyse a specific case study: the segmentation of iFood customers, the leading food delivery app in Brazil. This study aims not only to apply the theory of segmentation techniques but also to develop effective and practical solutions to address real sector challenges.

The application of unsupervised learning techniques in our study has offered multiple benefits. These methods have not only enabled effective customer segmentation, revealing previously undetectable patterns and correlations but also improved our ability to make data-driven decisions. Through this approach, we have been able to extract value from large volumes of information non-intrusively, thereby maximising the efficiency and effectiveness of our analysis strategies. Using these techniques has given us a significant advantage in understanding consumer behaviours and preferences, which is essential for adapting and evolving in a competitive market.

Índice general

1. Introducción **14**

2. Marco teórico **18**

2.1. Ciencia de datos 18

2.1.1. ¿Quién es el científico de datos? 19

2.1.2. Etapas de un proyecto de Ciencia de Datos 19

2.2. Aprendizaje automático 30

2.2.1. Aprendizaje supervisado 31

2.2.2. Aprendizaje no supervisado 32

2.2.3. Aprendizaje por refuerzo 32

3. Agrupación **33**

3.1. Definición 33

3.2. Etapas del análisis clúster 34

3.3. Precauciones en el uso de las técnicas clustering 35

3.4. Medición de proximidad 36

3.4.1. Medidas de similaridad para datos categóricos 37

3.4.2. Medidas de distancia para datos continuos 39

3.4.3. Estandarización 41

3.5. Clasificación de las técnicas clustering 42

3.5.1. Los métodos jerárquicos 42
3.5.2. Los métodos no jerárquicos 48

4. Propuesta metodológica **50**
4.1. Una pregunta y unos datos que la responden 51
4.2. Análisis exploratorio de los datos 51
4.3. Preprocesamiento de los datos 52
4.4. Aplicación de la técnica clúster 53
4.4.1. Paso 1. Elección de las variables 53
4.4.2. Paso 2. Elección de la medida de proximidad 53
4.4.3. Paso 3. El método de Ward 54
4.4.4. Paso 4. Validación de los resultados e interpretación de los mismos. 58
4.5. Informe y publicación de los códigos 58
4.6. Herramientas para el proyecto de ciencia de datos: lenguaje de programación con Python 59
4.6.1. Instalando Python 59
4.6.2. Bibliotecas esenciales y herramientas 60
4.6.3. Versiones utilizadas en este trabajo 67

5. Marco experimental: caso de estudio sobre segmentación de clientes **69**
5.1. Caso de estudio 70
5.2. Análisis exploratorio de los datos 71
5.3. Análisis preliminar de las variables 76
5.4. Aplicación del algoritmo de Ward 77
5.4.1. Paso 1. Elección de las variables 77
5.4.2. Paso 2. Elección de la medida de proximidad 78
5.4.3. Paso 3. El método de Ward 79
5.4.4. Paso 4. Validación de los resultados 81

5.5. Estrategias de marketing sobre la base de clientes 87
5.5.1. Estrategias de marketing según los clústeres obtenidos utilizando la distancia Euclídea 88
5.5.2. Estrategias de marketing según los clústeres obtenidos utilizando la distancia de Manhattan 89

6. Conclusiones finales **90**

7. Anexo **97**

Índice de tablas

3.1. *Recuentos de resultados binarios para dos objetos.* Fuente: Elaboración propia. 38

3.2. *Medidas de similaridad de variables binarias.* Fuente: Everitt et al. (2011). 39

5.1. *Conjunto de variables del problema de iFood.* Fuente: Elaboración propia. . 72

5.2. *Variables que finalmente se consideraron para el estudio.* Fuente: Elaboración propia. 77

5.3. *Recuento de las variables finales para el estudio según su naturaleza.* Fuente: Elaboración propia. 78

5.4. *Etiquetas de los cinco grupos a partir de la aplicación de algoritmo de Ward en una matriz de distancias Euclídea.* Fuente: Elaboración propia. 85

5.5. *Etiquetas de los cuatros grupos a partir de la aplicación de algoritmo de Ward en una matriz de distancias de Manhattan.* Fuente: Elaboración propia. 87

Índice de figuras

2.1. *Ciclo de vida de la Ciencia de Datos.* Fuente: Elaboración propia. 20

2.2. *Diagrama de cajas y bigotes.* Fuente: Elaboración propia. 26

3.1. *Diferentes formas de encontrar grupos en el mismo conjunto de datos.* Fuente: Adaptado de Kumar et al. (2006). 43

3.2. *Algunas terminologías utilizadas en la descripción de un dendrograma.* Fuente: Elaboración propia. 44

3.3. *Diferentes formas de encontrar tres grupos en el mismo conjunto de datos utilizando el algoritmo K-Medias.* Fuente: Adaptado de Kumar et al. (2006). 48

4.1. *Ejemplo del algoritmo de Ward.* Fuente: Adaptado de Pedret (2003). 57

4.2. *Gráfico simple de la función seno utilizando* `matplotlib`. Fuente: Elaboración propia. 63

4.3. *Gráfico simple utilizando* `seaborn`. Fuente: Elaboración propia. 64

5.1. Izquierda: *Distribución de la edad de los clientes.* Derecha: *Ingreso anual del hogar de los clientes.* Fuente: Elaboración propia. 73

5.2. Izquierda: *Nivel de educación según el cliente.* Derecha: *Estado marital de los clientes.* Fuente: Elaboración propia. 73

5.3. *Gasto en Vino (a), Gasto en Fruta (b), Gasto en Carne (c), Gasto en Pescado (d), Gasto en Dulces (e) y Gasto en Oro (f).* Fuente: Elaboración propia. 74
5.4. *Distribución del número de compras y visitas en la página web.* Fuente: Elaboración propia. 75
5.5. *Dendrograma construido a partir de la matriz de distancias Euclídea.* Fuente: Elaboración propia. 80
5.6. *Dendrograma construido a partir de la matriz de distancias de Manhattan.* Fuente: Elaboración propia. 81

Capítulo 1

Introducción

La inteligencia artificial (IA) ha emergido como una tecnología disruptiva que está transformando todas las áreas de la ciencia y la industria. Es una herramienta de amplio alcance que permite a las personas repensar cómo integramos la información, analizamos los datos y utilizamos los conocimientos resultantes para mejorar la toma de decisiones (Fabián, 2019) . A su vez, la capacidad de las máquinas para analizar grandes volúmenes de datos, aprender patrones y tomar decisiones basadas en algoritmos sofisticados ha optimizado los procesos en múltiples sectores. Este avance se debe no solo al incremento masivo de los datos digitales y a la mayor capacidad de computación, sino también a las innovaciones en los algoritmos de inteligencia artificial y del aprendizaje automático (Bryson, s. f.).

La ciencia de datos, el trabajo estadístico y el aprendizaje automático se destacan como áreas críticas en nuestra era de automatización y gestión de grandes volúmenes de datos. Es importante mencionar la diferencia entre los conceptos mencionados ya que muchas veces se tiende a una idea errónea de cada uno de ellos. En primer lugar, la ciencia de datos es un campo multidisciplinar que incluye el lenguaje y las técnicas necesarias para comprender y tratar los datos. Implica el diseño, recopilación, análisis e interpretación de datos numéricos, con el objetivo de extraer patrones y otra información útil (Kroese et al., 2023). En ciencia de datos, el tratamiento de datos requiere el dominio de una variedad de habilidades y conceptos, incluidos muchos tradicionalmente asociados con los campos de Estadística, Matemáticas y Ciencia de la Computación. La ciencia de

datos no es la intersección de estas áreas, sino una integración efectiva de ellas (Gutierrez, 2024). En segundo lugar, el trabajo estadístico en el tratamiento de datos es darle sentido a la información para extraer patrones y tendencias importantes, y comprender "lo que los datos dicen". El aporte estadístico en la ciencia de datos comprende desde el análisis descriptivo de los mismos, así como el análisis e interpretación de cuadros y gráficos estadísticos. Además, la validación de las conclusiones son pruebas de inferencia estadística. Otro aporte de la Estadística son los modelos probabilísticos y bayesianos y que ayuda a comprender situaciones inciertas. También se estudian modelos para encontrar relaciones entre acciones y consecuencias como modo de comprender una situación y adaptarse a la misma, que son bien interpretadas usando modelos de causalidad (Gutierrez, 2024). Y por último, el aprendizaje automático se ocupa del diseño de algoritmos y recursos informáticos para aprender de los datos (Kroese et al., 2023). La creciente popularidad de estos conceptos proviene tanto de su aplicabilidad directa a problemas reales como de su capacidad para integrar diversas disciplinas, como el marketing, finanzas, transporte, deportes, entre otras. En marketing, optimizan estrategias y campañas publicitarias; en finanzas, identifican patrones de mercado y detectan fraudes; en transporte, optimizan rutas y horarios, y predicen flujos de tráfico; en deportes, analizan el rendimiento de deportistas y estrategias de juego.

En la ciencia de datos, el principal desafío radica en el análisis matemático de los datos. Cuando el objetivo es interpretar el modelo y cuantificar la incertidumbre en los datos, este análisis suele denominarse aprendizaje estadístico. En cambio, cuando el énfasis está en hacer predicciones utilizando datos a gran escala, se habla de aprendizaje automático. En ciencia de datos existen dos objetivos principales en el modelado de datos: *(1)* predecir con precisión cantidades futuras de interés a partir de datos observados y *(2)* descubrir patrones inusuales o interesantes (Kroese et al., 2023). Estos objetivos se conocen respectivamente como aprendizaje supervisado y no supervisado. En el aprendizaje supervisado, podemos comprobar nuestro trabajo viendo qué tan bien nuestro modelo predice la res-

puesta, lo que lo convierte en un área bien comprendida. Por el contrario, el aprendizaje no supervisado es más desafiante, ya que no existe un objetivo simple como la predicción de una respuesta y el ejercicio tiende a ser más subjetivo lo que conlleva a que no hay una forma de comprobar nuestro trabajo porque no sabemos la verdadera respuesta ya que el problema no está supervisado.

Nuestro objetivo en este Trabajo Final de Grado es aplicar el ciclo de vida de la ciencia de datos en un caso de segmentación, utilizando métodos de aprendizaje no supervisado y el lenguaje de programación Python, para analizar un caso de estudio concreto: la segmentación de clientes de iFood, la principal aplicación de entrega de comida a domicilio en Brasil.

Este estudio tiene como finalidad no solo aplicar la teoría de técnicas de segmentación, sino también desarrollar soluciones efectivas y prácticas para abordar los desafíos reales del sector. Al agrupar a los clientes en función de características de compra específicas, la empresa puede dirigirse con más precisión a estos grupos y personalizar sus campañas de marketing. Esto permite crear contenido altamente relevante y personalizado para su público, lo que a su vez puede contribuir a mejorar las relaciones con los clientes y a hacer campañas de marketing más eficaces (*Machine Learning En Marketing — Mailchimp*, s. f.)

Concretamente a lo largo de este trabajo describiremos el ciclo de vida de la ciencia de datos necesario para abordar el problema de segmentación del cliente. Para ello, comenzaremos el Capítulo 2 abordando conceptos clave para el estudio, comenzando con la ciencia de datos, el rol y habilidades del científico de datos, y las herramientas necesarias para su trabajo. Incluye el concepto de aprendizaje automático, sus ramas principales y fundamentos teóricos.

El Capítulo 3 será crucial para conocer las técnicas de agrupamiento del aprendizaje no supervisado. Se explicarán tanto su concepto, el procedimiento y teoría necesaria para su debida aplicación.

En el Capítulo 4 se describirá la metodología utilizada en el caso de estudio de segmentación del cliente con el uso de técnicas de agrupamiento, incluyendo la recopilación de datos, análisis exploratorio, preprocesamiento y finalmente utilizando la técnica de agrupamiento jerárquica aglomerativa de Ward. Se detallará cómo se determinaron los grupos y la interpretación de los resultados. Este capítulo incluye, además, una sección donde se detalla la implementación de Python en este trabajo así como las bibliotecas fundamentales para cumplir con nuestro objetivo.

El Capítulo 5 presenta los resultados del caso de estudio sobre la segmentación del cliente aplicando la metodología detallada en el Capítulo 4. Se presentará la visualización de los grupos de clientes obtenidos y se analizarán las características de cada grupo, destacando así la relevancia para la estrategia de marketing de iFood y la optimización de futuras campañas según sus clientes.

El Capítulo 6 está dedicado a la exposición de las conclusiones obtenidas de este trabajo, los logros y dificultades de haber aplicado las técnicas clustering en un problema de segmentación además de aportar información relevante para la empresa.

Finalmente este Trabajo Final de Grado concluye, cumpliendo así todos los objetivos planteados, aportando algunas propuestas o ideas para las futuras campañas de marketing de la empresa, buscando la finalidad de optimizar los procesos y llegar a grupos de clientes de la forma más precisa posible para la empresa y facilitando a la empresa datos de valor de su organización para que puedan establecer estrategias futuras con mayor tasa de éxito.

Capítulo 2

Marco teórico

En este segundo capítulo, se abordará los conceptos que serán necesarios para comprender el estudio. Comenzaremos por el concepto "Ciencia de datos", quién es el científico de datos, las habilidades que requiere un científico de datos y las herramientas necesarias para desarrollar su proyecto siendo estas las que seguiremos en este trabajo. Se incluye el concepto "aprendizaje automático", una disciplina que se sitúa en la intersección de la estadística, la IA y la ciencia de la computación. También se conocerán las ramas principales de aprendizaje automático y en qué se basan.

2.1. Ciencia de datos

La ciencia de datos es la práctica de utilizar datos para tratar de comprender y resolver problemas del mundo real. Este concepto no es exactamente nuevo; La gente ha estado analizando cifras de ventas y tendencias desde la invención del cero. Sin embargo, en la última década hemos obtenido acceso a una cantidad exponencial de más datos de los que existían antes. La llegada de las computadoras ha ayudado a generar todos esos datos, pero la informática es también nuestra única manera de procesar los grandes volúmenes de información. Mediante código informático, un *científico de datos* puede transformar o agregar datos, ejecutar análisis estadísticos o entrenar modelos de *aprendizaje automáti-*

co.[1]

2.1.1. ¿Quién es el científico de datos?

Hasta hace poco tiempo, el científico de datos no existía como tal, y por lo tanto su definición también es reciente y sigue sufriendo cambios. Una definición práctica es la que propone Vázquez (2018), la cual dice que "*Un Científico de Datos es una persona encargada de analizar problemas de negocios/organizaciones ofreciendo una solución estructurada que comienza convirtiendo este problema en una pregunta válida y completa, luego usa programación y herramientas computacionales para desarrollar códigos que preparan, limpian y analizan los datos para crear modelos y responder la pregunta inicial*". Ferreira (2022) nombra una serie de habilidades en las que un científico de datos debe destacar, y estas son: habilidades en matemáticas y estadística, habilidades en el área de la computación (programación, redes, seguridad, bases de datos, entre otras) y tener conocimiento de negocio (clientes, ventas, artículos, marketing, entre otras).

2.1.2. Etapas de un proyecto de Ciencia de Datos

El ciclo de vida de la ciencia de datos se refiere a las distintas etapas por las que suele pasar un proyecto de ciencia de datos, desde la concepción inicial y la recopilación de datos hasta la comunicación de resultados y perspectivas. A pesar de que cada proyecto de ciencia de datos es único, la mayoría de los proyectos siguen un ciclo de vida similar. Este ciclo de vida proporciona un enfoque estructurado para manejar datos complejos, sacar conclusiones precisas y tomar decisiones basadas en datos. La Figura 2.1 muestra el ciclo de vida de la Ciencia de Datos.

Vamos a detallar las estapas de un proyecto de ciencia de datos en las siguientes subsecciones.

[1]Robinson, E., & Nolis, J. (2020). *Build a career in data science*. Manning Publications Co., 5.

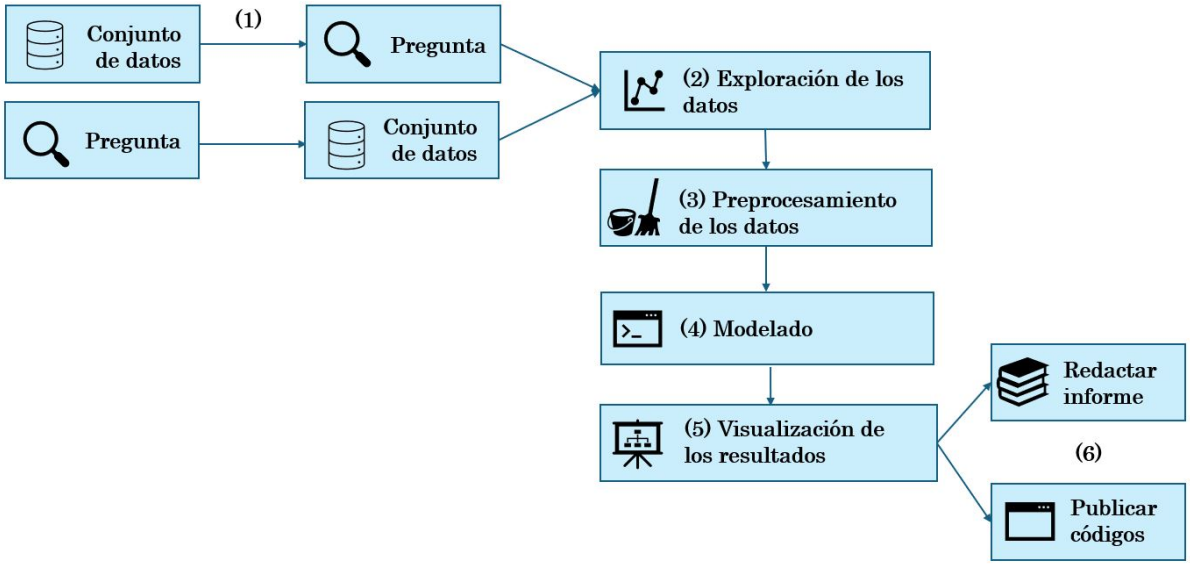


Figura 2.1: *Ciclo de vida de la Ciencia de Datos.* Fuente: Elaboración propia.

2.1.2.1. Etapa 1. Encontrar los datos y realizar una pregunta

Comenzamos con dos puntos importantes: un conjunto de datos que sea interesante y una pregunta que hacer al respecto. La combinación de esta pregunta y los datos son el núcleo del proyecto de ciencia de datos.

Cuando se plantea los datos que se desean utilizar, lo más importante es encontrar datos que le interesen a el científico de datos en cuestión. Su elección de datos es una forma de mostrar su personalidad o el conocimiento del dominio que tiene de su carrera o estudios anteriores. Esto no tiene por qué ocurrir siempre así, se puede comenzar explorando conjuntos de datos y ver si ocurren preguntas interesantes. Aquí hay algunas sugerencias sobre dónde podría comenzar a buscar estos datos:

- **Kaggle:** comenzó como un sitio web para concursos de ciencia de datos. Las empresas publican un conjunto de datos y una pregunta y, por lo general, ofrecen un premio a la mejor respuesta. Kaggle también tiene foros de discusión y "núcleos" en los que las personas comparten su código para poder aprender cómo otros aborda-

ron el conjunto de datos. Como resultado, Kaggle tiene miles de conjuntos de datos acompañados de preguntas y ejemplos de cómo los analizaron otras personas.

El mayor beneficio de Kaggle es también su mayor inconveniente: al entregar un conjunto de datos y un problema (generalmente limpios), ha hecho gran parte del trabajo. Una forma de utilizar Kaggle es tomar un conjunto de datos pero plantear una pregunta diferente o realizar un análisis exploratorio. Pero, en general, Kaggle es una herramienta útil para aprender a abordar un proyecto y luego viendo cómo se desempeñó en comparación con otros, aprendiendo así de lo que hicieron sus modelos.[2]

- **Datos públicos españoles:** muchos datos gubernamentales están disponibles en línea. Varios ejemplos son,

  - *datos.gob.es*, su fin es promocionar la apertura de la información pública y desarrollo de servicios avanzados basados en datos. Es promovida por el Ministerio para la Transformación Digital y de la Función Pública y la Entidad Pública Empresarial Red.es[3].

  - El *Instituto Nacional de Estadística (INE)*, es un organismo autónomo adscrito al Ministerio de Economía y Hacienda. Su tarea principal y de mayor tradición es la de elaborar estadísticas públicas, que son estudios oficiales sobre la situación y evolución de la población, la economía y la sociedad de España[4].

  - *El Centro de Investigaciones Sociológicas (CIS)*, es un organismo autónomo adscrito al Ministerio de la Presidencia, Relaciones con las Cortes y Memoria Democrática que tiene por finalidad el estudio científico de la sociedad

[2]Robinson, E., & Nolis, J. *Build a career in data science*. cit., 57.
[3]datos.gob.es. (2024). Acerca de la iniciativa Aporta. Última consulta 28 de junio 2024.
[4]INE. (s. f.). *El INE y sus principales funciones.* Última consulta 28 de junio 2024.

española[5].

- Cada Ministerio español tiene un departamento de estadística para realizar aquellos estudios directamente relacionados con su actividad[6].

- **Datos propios:** hay muchas aplicaciones donde el usuario propio puede descargar datos sobre su cuenta registrada en la aplicación. Los sitios web de redes sociales y los servicios de correo electrónico son dos de los más importantes.

2.1.2.2. Etapa 2. Análisis exploratorio de datos

Una vez que tenemos la pregunta y hemos seleccionado el conjunto de datos a través de las diferentes plataformas que ofrecen conjuntos de datos se procede a realizar un análisis exploratorio de los datos. Este paso conlleva un análisis de los datos para comprender sus patrones, características y posibles anomalías.

En esta sección describiremos una serie de técnicas relativamente simples que nos ayudarán a entender los datos. La mayoría de los métodos que mencionaremos se basan en un examen descriptivo y gráfico de las variables marginales univariados o bivariados directos de los datos multivariados.

Pero, al manipular, resumir y mostrar datos, es importante especificar correctamente el tipo de variables con los que estamos trabajando. Para definir la estructura de los datos vamos a seguir la literatura de Kroese et al. (2023). Generalmente podemos clasificar las variables como *cuantitativas* o *cualitativas*. Las variables cuantitativas poseen una "cantidad numérica", como la altura, la edad, el número de nacimientos, etc., y pueden ser *continuas* o *discretas*. Las variables cuantitativas continuas toman valores en un rango continuo de valores posibles, como altura, voltaje o rendimiento del cultivo; estas variables capturan la idea de que las mediciones siempre se pueden realizar con mayor precisión.

[5] *Funciones*. (s. f.). CIS. Última consulta 28 de junio de 2024.
[6] INE. (s. f.). *El INE y sus principales funciones.*. Última consulta 28 de junio 2024.

Las variables cuantitativas discretas tienen un número contable de posibilidades, como un recuento.

Por el contrario, las variables *cualitativas* no tienen un significado numérico, pero sus posibles valores se pueden dividir en un número fijo de categorías, como (M, F) para el género o (azul, negro, marrón, verde) para el color de ojos. Por esta razón, estas variables también se denominan *categóricas*. Una regla general simple es: si no tiene sentido promediar los datos, es categórico. Por ejemplo, no tiene sentido promediar los colores de ojos. Por supuesto, todavía es posible representar datos categóricos con números, como 1 = azul, 2 = negro, 3 = marrón, pero esos números no tienen ningún significado cuantitativo: las variables categóricas a menudo se denominan *factores*. Hay rasgos cualitativos que tienen un orden, estos son denominados como *ordinales*, un ejemplo sería el nivel de educación (universitario, graduado, máster, posgrado). En contraste con los rasgos cualitativos sin orden, estos se denominan *nominales* un ejemplo de estos sería, por ejemplo, el número de identidad de una persona.

Una vez conocido el tipo de datos que nos podemos encontrar en una base de datos, procederemos a describir las técnicas descriptivas y gráficas. Los detalles de estas técnicas están basadas principalmente en la literatura de Kroese et al. (2023).

- **Tablas resumen.**

  A menudo resulta útil resumir una base de datos en una forma más condensada. Una tabla de recuentos o una tabla de frecuencias facilita la comprensión de la distribución subyacente de una variable, especialmente si los datos son cualitativos.

- **Resumen estadístico.**

  Los cuadros de resumen estadístico suelen incluir una columna para las variables a describir y una fila para los diferentes elementos que resumen características cualitativas, mediante el recuento más frecuente y el número de elementos únicos.

Sea $\mathbf{x} = [x_1, ..., x_n]'$ un vector columna de n elementos. Los cuadros de resumen estadístico incluyen las siguientes medidas descriptivas de una variable.

- La *media muestral* de $\mathbf{x}$, denotada por $\bar{x}$, es simplemente el promedio de los valores de los datos, entonces, sea $i = 1, \ldots, n$, tenemos,

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{2.1}$$

- La *varianza muestral*, denotada por s^2, tiene por fórmula, entonces, sea $i = 1, \ldots, n$, tenemos,

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2, \tag{2.2}$$

  donde la *desviación estándar* es $s = \sqrt{s^2}$.

- El *cuartil muestral*, p, donde $p \in (0, 1)$, de $\mathbf{x}$ es un valor x tal que al menos una fracción p de los datos es menor o igual que x y al menos una fracción $1 - p$ de los datos es mayor o igual a x. La *mediana muestral* es el cuartil 0,5 de la muestra. El cuartil muestral también se denomina *percentil*, $100 \times p$. Los percentiles muestrales 25, 50 y 75 se denominan primer, segundo y tercer *cuartiles* de los datos.

- La media y la mediana muestrales brindan información sobre la ubicación de los datos, mientras que la distancia entre los cuartiles muestrales dan alguna indicación de la *dispersión* de los datos. Otras medidas de dispersión son el *rango muestral*, $max_i x_i$ - $min_i x_i$ siendo $i = 1, \ldots, n$.

A continuación, vamos a decribir varios métodos para visualizar datos. El principal punto al que se pretende llegar es que la forma de visualizar las variables debe adaptarse siempre a la naturaleza de los mismas. Por ejemplo, los datos cualitativos deben representarse de forma diferente a los cuantitativos.

- **Trazado de variables cualitativas.**

  Representar gráficamente datos cuantitativos discretos o cualitativos se puede realizar mediante un *diagrama de barras*. Son representativos los rectángulos verticales u horizontales, conocidos como barras. El tamaño de cada barra se ajusta proporcionalmente al valor que representa. Este tipo de gráficos proporcionan una comparación visual de cantidades o frecuencias, lo que facilita la interpretación de los datos.

- **Trazado de variables cuantitativas.**

  - *Histograma:* es una representación gráfica común de la distribución de una característica cuantitativa. Comenzamos dividiendo el rango de valores en varios contenedores o pistas. Contamos los recuentos de los valores que caen en cada pista y luego hacemos el gráfico dibujando rectángulos cuyas bases son los intervalos de la pista y cuyas alturas son los recuentos.

  - *Diagrama de cajas y bigotes:* puede verse como una representación gráfica de cinco herramientas descriptivas de variables, estas son, el mínimo, el máximo y el primer, segundo y tercer cuartil. Además, el diagrama permite intuir la morfología y simetría de una variable e identificar valores atípicos y comparar distribuciones con respecto a otras variables.

    Observe la Figura 2.2. La caja se dibuja desde el primer cuartil (Q_1) hasta el tercer cuartil (Q_3). La línea horizontal dentro de la caja indica la ubicación de la mediana. Los llamados "bigotes" se extienden en la parte superior e inferior de la caja. El tamaño de la caja se llama rango intercuartil: $IQR = Q_3 - Q_1$. El bigote inferior se extiende hasta el mayor de *(a)* el mínimo de los datos y *(b)* $Q_1 - 1,5 \times IQR$. De manera similar, el bigote superior se extiende hasta el menor de *(a)* el máximo de los datos y a *(b)* $Q_3 + 1,5 \times IQR$. Cualquier punto de datos fuera de los bigotes se indica mediante un pequeño punto entendido

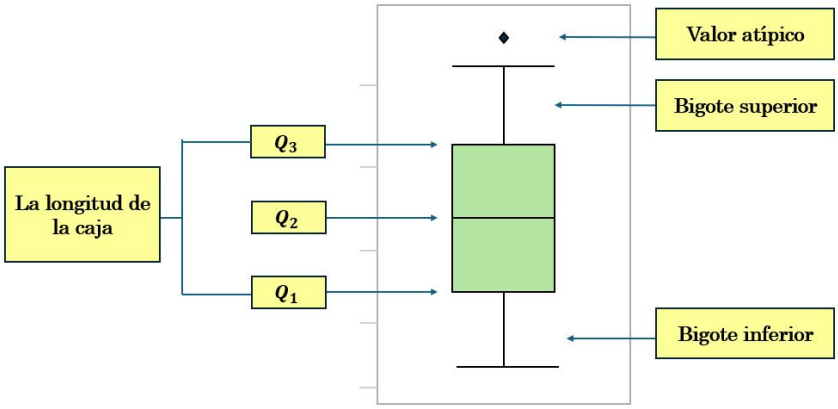


Figura 2.2: *Diagrama de cajas y bigotes.* Fuente: Elaboración propia.

como un punto sospechoso o desviado (valor atípico). Un diagrama de cajas y bigotes se suele utilizar más para representar variables cuantitativas continuas.

2.1.2.3. Paso 3. Preprocesamiento de los datos

Dado que los datos reales pueden ser incompletos, ruidosos e inconsistentes, el preprocesamiento de datos es una parte esencial dentro de las etapas del ciclo de vida de la ciencia de datos.

Vamos a mencionar algunas de las principales tareas involucradas en el preprocesamiento de datos que propone Janssen et al. (2012). Estas son:

- **Limpieza de datos:** completar valores faltantes, suavizar datos ruidosos, identificar o eliminar valores atípicos, corregir inconsistencias y resolver redundancias causadas por la integración o fusión de datos de varias fuentes. Se pueden utilizar varias técnicas para limpiar los datos, de las que sólo trataremos brevemente la forma en que se pueden tratar los datos faltantes y los valores atípicos.

  (i) Manejo de datos faltantes. Es posible que falten valores debido a que no se recopila información o las variables no son aplicables en todos los casos (por

ejemplo, ingresos anuales para aquellos clientes que tienen 18 y están sólo estudiando). Una forma obvia de manejar los datos faltantes es simplemente eliminar los objetos de datos correspondientes y analizar solo la parte del conjunto de datos que está completa. Esta estrategia no conduce al uso más eficiente de los datos y se recomienda sólo en situaciones donde el número de valores faltantes es muy pequeño. Otra opción, llamada imputación, para lidiar con los datos faltantes es reemplazar los valores faltantes por una constante global o por una estimación, por ejemplo, la media, mediana, un valor más probable.

(ii) Manejo de valores atípicos. Los valores atípicos son valores de datos que son extremadamente grandes o pequeños en relación con el resto de los datos. Por lo tanto, se sospecha que tergiversan la población de la que fueron recolectados.

El primer paso en el manejo de valores atípicos consiste en su detección. El segundo paso implica el tratamiento previo de los valores atípicos antes de realizar la modelización de los datos. Se pueden aplicar tres estrategias generales: (a) utilizar los datos periféricos en el análisis, teniendo en cuenta sus efectos en los resultados; (b) *trimming*: eliminar los datos atípicos del conjunto de datos; (c) *winsorizing*: reemplazar los valores atípicos por una variante truncada, por ejemplo un percentil específico del conjunto de datos, o un valor de corte asociado del diagrama de caja sesgado.

- **Reducción de datos:** se refiere a obtener una representación reducida del volumen de los datos que producen resultados analíticos iguales o similares.

  En situaciones en las que el conjunto de datos es muy grande, la reducción de datos tiene como objetivo reducir el tiempo de ejecución y los problemas de almacenamiento. El desafío es obtener una representación reducida del conjunto de datos que tenga un volumen mucho menor pero que produzca los mismos (o casi los mismos) resultados analíticos. Para lograrlo existen diversas estrategias de reducción:

(i) Agregación. Consiste en combinar dos o más variables (u objetos) en una sola variable (u objeto), dando como resultado un número reducido de variables u objetos. Esto también puede implicar un cambio de escala y puede generar datos más "estables" (menos variabilidad), aunque el precio de perder información más detallada escala.

(iii) *Muestreo.* En lugar de procesar el conjunto de datos completo, se puede decidir procesar parte del conjunto de datos que se obtiene seleccionando una muestra restringida (aleatoria).

(iii) Selección de características. La selección de características consiste en identificar y eliminar características (o, equivalentemente, variables) que son redundantes o irrelevantes.

- **Transformación de datos:** poner los datos en formularios que sean apropiados para un análisis más extenso. Esto incluye la normalización y la realización de operaciones de resumen o agregación de los datos, por ejemplo.

Puede encontrar una lista más extensa en Janssen et al. (2012).

2.1.2.4. Paso 4. Modelado

En esta etapa, los científicos de datos utilizan algoritmos de aprendizaje automático y modelos estadísticos para identificar patrones, hacer predicciones o descubrir información. El objetivo aquí, es obtener algo significativo de los datos que se ajuste a los objetivos del proyecto, ya sea predecir resultados futuros, clasificar datos o descubrir patrones ocultos.

2.1.2.5. Paso 5. Visualización de los resultados

Una vez modelado los datos es importante interpretar y comunicar los resultados. No basta con obtener información, sino que se debe comunicar con eficacia, mediante un lenguaje claro y conciso y dando uso de elementos visuales convincentes.

2.1.2.6. Paso 6. Informe y publicación de los códigos

Divulgar la metodología y resultados de un proyecto de ciencia de datos es excelente para celebrar los progresos como científico de datos. Existen varias plataformas populares donde se pueden publicar estos proyectos. A parte de la ya mencionada plataforma Kaggle, otras también reconocidas son,

- **GitHub:** es una plataforma donde se puede almacenar, compartir y trabajar junto con otros usuarios para escribir código. Almacenar un código en un "repositorio" en GitHub permite: presentar o compartir el trabajo, seguir y administrar los cambios en el código a lo largo del tiempo, dejar que otros usuarios revisen el código y realicen sugerencias para mejorarlo y colaborar en un proyecto compartido, sin preocuparse de que los cambios afectarán al trabajo de los colaboradores antes de que esté listo para integrarlos.[7]

- **GitLab:** es una plataforma de gestión de repositorios de código que utiliza el sistema de control de versiones Git. Es un servicio en la nube que permite a los desarrolladores y equipos de trabajo almacenar y gestionar el código fuente de sus proyectos.[8]

- **Stack Overflow:** es un sitio de preguntas y respuestas sobre casi cualquier rama de la informática.[9]

Comprender y aplicar este ciclo de vida permite un enfoque más sistemático y satisfactorio de los proyectos de ciencia de datos.

[7] *Acerca de GitHub y Git - documentación de GitHub.* (s. f.). GitHub Docs. Última consulta 13 de junio de 2024.

[8] *The most-comprehensive AI-powered DevSecOps platform.* (s. f.). GitLab. Última consulta 13 de junio de 2024.

[9] *Stack Overflow en español.* (s. f.). Última consulta 13 de junio de 2024.

2.2. Aprendizaje automático

Aunque esta sección puede parecer un tema más de IA, lo cierto es que actualmente varios algoritmos y técnicas usadas en ciencia de datos utilizan la IA, especialmente de aprendizaje automático, el cual ha dado un gran avance al área de ciencia de datos. Otro punto importante es que es difícil hablar de ciencia de datos sin IA y viceversa, ya que la gran cantidad de datos que se generan actualmente ha propiciado un gran impulso a la IA que bajo ciertos esquemas específicos, funcionan mejor cuando se tiene una gran cantidad de datos.[10]

El aprendizaje automático o más conocido por su término en inglés *machine learning* consiste en extraer conocimiento de los datos. Es un campo de investigación en la intersección de la estadística, la IA y la informática. En los últimos años, la aplicación de métodos de aprendizaje automático se ha vuelto omnipresente en la vida cotidiana. Desde recomendaciones automáticas sobre qué películas ver, qué comida pedir o qué productos comprar, hasta personalizar la radio en línea y reconocer a tus amigos en tus fotos, muchos sitios web y dispositivos modernos tienen algoritmos de aprendizaje automático en su núcleo. Cuando observa un sitio web complejo como Facebook, Amazon o Netflix, es muy probable que cada parte del sitio contenga múltiples modelos de aprendizaje automático. Fuera de las aplicaciones comerciales, el aprendizaje automático ha tenido una enorme influencia en la forma en que se realiza hoy la investigación basada en datos. Sin embargo, no es necesario que su aplicación sea de gran escala o que cambie el mundo como estos ejemplos para beneficiarse de aprendizaje automático.[11]

El aprendizaje automático no es una excepción y se le requiere un buen flujo de datos variados y organizados para una solución de aprendizaje automático sólida. Los algorit-

[10]Ferreira, R. (2022). *Ciencia de Datos Teoría y Aplicaciones*. Tecnológico Nacional de México, 71.
[11]Müller, A. C., & Guido, S. (2016). *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly Media Inc., 1.

mos de aprendizaje automático tienen la capacidad de mejorarse a sí mismos mediante el entrenamiento[12]. Hoy en día, los algoritmos de aprendizaje automático se entrenan utilizando tres métodos destacados. Estos tres tipos de aprendizaje automático son el *aprendizaje supervisado*, el *aprendizaje no supervisado* y el *aprendizaje por refuerzo*.

2.2.1. Aprendizaje supervisado

El aprendizaje supervisado es uno de los tipos más básicos de aprendizaje automático. En este caso, el algoritmo de aprendizaje automático se entrena con *datos etiquetados*. En el aprendizaje supervisado, el algoritmo aprendizaje automático recibe un pequeño conjunto de *datos de entrenamiento* con el que trabajar. Este conjunto de datos de entrenamiento es una parte más pequeña del conjunto de datos más grande y sirve para darle al algoritmo una idea básica del problema, la solución y los puntos de datos a tratar. El conjunto de datos de entrenamiento también es muy similar al conjunto de datos final en sus características y proporciona al algoritmo los parámetros etiquetados necesarios para el problema. Luego, el algoritmo encuentra relaciones entre los parámetros dados, esencialmente estableciendo una relación de causa y efecto entre las variables del conjunto de datos. Al final del entrenamiento, el algoritmo tiene una idea de cómo funcionan los datos y la relación entre la entrada y la salida. Esta solución se implementa para usarla con el conjunto de datos final, que aprende de la misma manera que el conjunto de datos de entrenamiento. Esto significa que los algoritmos de aprendizaje automático supervisados seguirán mejorando incluso después de su implementación, descubriendo nuevos patrones y relaciones a medida que se entrenan con nuevos datos.[13]

[12]Sakarkar, G., Patil G., & Dutta, P. (2021). *Internet of Things and Machine Learning: Machine Learning Algorithms Using Python Programming*, Nova Science Publishers Inc., 39.
[13]Sakarkar, G., Patil G., & Dutta, P. *Internet of Things and Machine Learning.* cit., 39.

2.2.2. Aprendizaje no supervisado

El aprendizaje no supervisado tiene la ventaja de poder trabajar con *datos sin etiqueta*. Las etiquetas permiten que el algoritmo encuentre la naturaleza exacta de la relación entre dos puntos de datos. Sin embargo, el aprendizaje no supervisado no tiene etiquetas con las que trabajar, lo que da como resultado la creación de estructuras ocultas. Las relaciones entre puntos de datos son percibidos por el algoritmo de una manera abstracta. La creación de estas estructuras ocultas es lo que hace que los algoritmos de aprendizaje no supervisados sean versátiles. En lugar de un planteamiento del problema definido y establecido, los algoritmos de aprendizaje no supervisados pueden adaptarse a los datos cambiando dinámicamente estructuras ocultas[14]. En el Capítulo 3 se profundizará más sobre el aprendizaje no supervisado.

2.2.3. Aprendizaje por refuerzo

El aprendizaje por refuerzo se inspira directamente en cómo los seres humanos aprenden de los datos en sus vidas. Cuenta con un algoritmo que se mejora a sí mismo y aprende de nuevas situaciones mediante un método de prueba y error. Los resultados favorables se alientan o "refuerzan", y los resultados no favorables se desalientan o "castigan". Basado en el concepto psicológico de condicionamiento, el aprendizaje por refuerzo funciona colocando el algoritmo en un entorno de trabajo con un intérprete y un sistema de recompensa. En cada iteración del algoritmo, el resultado se entrega al intérprete, quien decide si el resultado es favorable o no. En caso de que el programa encuentre la solución correcta, el intérprete refuerza la solución proporcionando una recompensa al algoritmo. Si el resultado no es favorable, el algoritmo se ve obligado a repetir hasta encontrar un resultado mejor.[15]

[14]Sakarkar, G., Patil G., & Dutta, P. *Internet of Things and Machine Learning.* cit., 43-44.
[15]Sakarkar, G., Patil G., & Dutta, P. *Internet of Things and Machine Learning.* cit., 46.

Capítulo 3

Agrupación

En este tercer capítulo nos centraremos en las técnicas de segmentación o llamadas similarmente técnicas de agrupamiento, del aprendizaje no supervisado. No solo nos quedaremos en el concepto sino que se explicará el procedimiento que se debe seguir para utilizar estas técnicas así como los conceptos teóricos que se necesitan conocer para su correcta aplicación.

3.1. Definición

El entorno de aprendizaje no supervisado cuenta con herramientas destinadas a entornos en los que sólo tenemos un conjunto de variables $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_p}$ medidas en n observaciones u objetos. El objetivo sería descubrir cosas interesantes sobre las mediciones en $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_p}$. ¿Podemos descubrir subgrupos entre las variables o entre las observaciones? El aprendizaje no supervisado nos ayudará a responder preguntas como esta.[1]

La *agrupación* o más conocido por su nombre en inglés *clustering* se refiere a un conjunto muy amplio de técnicas para encontrar subgrupos o grupos en un conjunto de datos. Cuando agrupamos las observaciones de un conjunto de datos, buscamos dividirlas en grupos distintos de modo que las observaciones dentro de cada grupo sean bastante similares

[1] James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An Introduction to Statistical Learning: With Applications in Python.* (1.ªed.). Springer, 503.

entre sí, mientras que las observaciones en diferentes grupos sean bastante diferentes entre sí. Una observación puede describirse mediante un conjunto de medidas o por su relación con otros objetos. Por supuesto, para concretar esto, debemos definir qué significa que dos o más observaciones sean *similares* o *diferentes*.

Entonces, el centro de los objetivos del análisis clúster[2] es la noción del grado de similitud (o disimilitud) entre las observaciones individuales que se agrupan. Un método de agrupación, intenta agrupar las observaciones según la definición de similitud que se le proporciona.[3]

Los métodos de clustering se clasifican en dos grandes grupos: los métodos de *agrupamiento jerárquico* y los métodos de *agrupamiento no jerárquico*. El primer grupo se divide en dos ramas: los métodos de agrupación *aglomerativa* y los *divisivos*. Cada uno de estos enfoques de agrupación tiene sus ventajas y desventajas.

3.2. Etapas del análisis clúster

Gallardo (s. f.) resume las etapas a seguir en el empleo de una técnica clúster en los siguientes puntos:

1. **Elección de las variables.**

   La elección inicial del conjunto concreto de variables usadas para describir a cada individuo constituye un marco de referencia para establecer las agrupaciones o clústeres.

   En esta etapa Gallardo (s. f.) menciona dos cuestiones que se deben considerar, la primera cuestión sería decidir cuáles son las variables que se consideran relevantes y

[2]Real Academia Española. (2005). Clúster. En *Diccionario panhispánico de dudas (DPD)*. (2.ª ed.). "grupo de elementos similares o cercanos, o agrupados en función de alguna característica o variable". Última consulta 5 de junio 2024.

[3]Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. (2.ªed.). Springer Series in Statistics, Springer, 503-520.

que puedan dar lugar a la agrupación de grupos o clústeres. La segunda cuestión sería determinar el número de variables a emplear. En muchas aplicaciones es probable que tomar demasiadas medidas dé origen a diversos problemas, bien sea a nivel computacional o bien porque dichas variables adicionales oscurezcan la estructura de los grupos.

2. **Elección de una medida de proximidad.**

   El objetivo del análisis de grupos es dividir las observaciones en grupos de modo que las diferencias por pares entre las asignadas al mismo clúster tiendan a ser menores que las de otras observaciones asignadas a otros clústeres. Esto se refleja en la elección de la medida de proximidad de distancia entre los objetos.

3. **Elección de la técnica de clustering a emplear en el estudio.**

   Los métodos de clustering que se han propuesto y desarrollado en los últimos años son bastante numerosos y muy diversos en cuanto a su concepción, clasificándose, en un primer estado, en jerárquicos y no jerárquicos (que más tarde detallaremos).

   En algunos problemas prácticos, la elección del método a emplear será relativamente natural, dependiendo, sobre todo, de la naturaleza de los datos usados y de los objetivos finales perseguidos, si bien en otros, la elección no será tan clara.

4. **Validación de los resultados e interpretación de los mismos.**

   Finalmente, una vez obtenidos los clústeres, es necesario interpretar los clústeres resultantes de las técnicas.

3.3. Precauciones en el uso de las técnicas clustering

Gallardo (s. f.) menciona algunas precauciones que hay tener sobre los métodos de agrupación,

- Algunos de los métodos de análisis clúster son procedimientos que, en la mayor parte de los casos, no están soportados por un cuerpo de doctrina estadística teórica. En otras palabras, la mayor parte de los métodos son heurísticos.

- Distintos procedimientos clúster pueden generar soluciones diferentes sobre el mismo conjunto de datos. Una razón para ello radica en el hecho de que los métodos clúster se han desarrollado a partir de diversas disciplinas que han dado origen a reglas diferentes de formación de grupos. De esta manera, lógicamente, es necesaria la existencia de técnicas que puedan ser usadas para determinar qué método produce los grupos naturalmente más homogéneos en los datos.

3.4. Medición de proximidad

Una cuestión central a la hora de agrupar objetos es el conocimiento de qué tan "cerca" están estos objetos entre sí, o qué tan lejos están. Esto se refleja en la elección de las medidas de proximidad.

Muchas investigaciones de agrupamiento tienen como punto de partida una matriz unimodal $n \times n$, cuyos elementos reflejan, en cierto sentido, una medida cuantitativa de cercanía, más comúnmente denominada en este contexto disimilitud, distancia o semejanza, siendo el término general *proximidad*.

Generalmente, las similaridades están acotadas en el rango de cero a uno; un aumento de la similaridad implica un aumento de la semejanza entre individuos, y toda similaridad de un individuo consigo mismo debería ser igual al máximo valor posible, es decir, uno. Las distancias, en cambio, disminuyen con un aumento del parecido, no son negativas y la distancia de un elemento consigo mismo es cero.[4]

[4]Demey, J. R., Pla, L., Vicente-Villardón, J. L., Di Rienzo, J. A., & Casanoves, F. (2011). *Valoración y análisis de la diversidad funcional y su relación con los servicios ecosistémicos*, CATIE, 47.

Las proximidades se pueden determinar directa o indirectamente, aunque esto último es más común en la mayoría de las aplicaciones. Por un lado, las proximidades determinadas directamente ocurren con mayor frecuencia en áreas como la psicología y la investigación de mercados. Por otro lado, las proximidades indirectas generalmente se derivan de la matriz multivariada (bimodal) $\mathbf{X}$ de dimensión $n \times p$, donde n es el número de objetos y p el número de variables. Existe una amplia gama de posibles medidas de proximidad, muchas de las cuales veremos en este capítulo.

Para analizar las medidas de proximidad, primero consideraremos medidas adecuadas para variables categóricas y luego aquellas útiles para variables continuas. Gran parte de la literatura de las siguientes subsecciones está basada en Everitt et al. (2011), Demey et al. (2011) y Gallardo (s. f.).

3.4.1. Medidas de similaridad para datos categóricos

Con datos en los que todas las variables son categóricas, las medidas de similitud son las más comúnmente utilizadas. Las medidas generalmente se escalan para estar en el intervalo [0. 1], aunque ocasionalmente se expresan como porcentajes en el rango 0-100 %. Dos individuos i y j tienen un coeficiente de similitud, s_{ij}, de unidad si ambos tienen valores idénticos para todas las variables. Un valor de similitud de cero indica que los dos individuos difieren al máximo en todas las variables. La definición de similaridad es la siguiente.

Definición 3.4.1. *Sea U un conjunto finito o infinito de elementos. Una función $s : U \times U \longrightarrow \mathbb{R}$ se llama similaridad si cumple las siguientes propiedades: $\forall x, y \in U$:*

1. $s(x, y) \leq s_o$
2. $s(x, x) = s_0$
3. $s(x, y) = s(y, x)$

donde s_0 es un número real finito arbitrario.

Definición 3.4.2. *Una función s, verificando las condiciones de la definición 2.4.1, se llama similaridad métrica si, además, verifica:*

1. $s(x, y) = s_o \Rightarrow x = y$

2. $|s(x, y) + s(y, z)|s(x, z) \geq s(x, y)s(y, x), \forall z \in U$

donde s_0 es un número real finito arbitrario.

Notemos que la segunda definición corresponde al hecho de que la máxima similaridad sólo la poseen dos elementos idénticos.

3.4.1.1. Medidas de similaridad para datos binarios

El tipo más común de datos categóricos multivariados es aquel en el que todas las variables son binarias y se ha propuesto una gran cantidad de medidas de similitud para dichos datos. Todas las medidas se definen en términos de las entradas en una clasificación cruzada de los recuentos de coincidencias y discrepancias en las p variables para dos individuos; la versión general de esta clasificación cruzada se muestra en la Tabla 3.1.

En la Tabla 3.2 se muestra una lista de algunas de las medidas de similitud sugeridas para datos binarios; se puede encontrar una lista más extensa en Gower y Legendre (1986).

Tabla 3.1: *Recuentos de resultados binarios para dos objetos.* Fuente: Elaboración propia.

	Objeto j			
	Resultado	1	0	*Total*
Objeto i	1	a	b	$\mathbf{a+b}$
	0	c	d	$\mathbf{c+d}$
	Total	$\mathbf{a+c}$	$\mathbf{b+d}$	$\mathbf{p}$

Tabla 3.2: *Medidas de similaridad de variables binarias.* Fuente: Everitt et al. (2011).

Medida	**Similaridad**
S1: Coeficiente de coincidencia simple	$s_{ij} = (a+d) \ / \ (a+b+c+d)$
S2: Coeficiente de Jaccard	$s_{ij} = a \ / \ (a+b+c)$
S3: Rogers y Tanimoto	$s_{ij} = (a+d) \ / \ [a+2(b+c)+d]$
S4: Sneath y Sokal	$s_{ij} = a \ / \ [a+2(b+c)]$
S5: Gower y Legendre	$s_{ij} = (a+d)/[a+0.5(b+c)+d]$
S6: Gower y Legendre	$s_{ij} = a \ /[a+0.5(b+c)]$

3.4.1.2. Medidas de similaridad para datos categóricos de dos o más niveles

Los datos categóricos en los que las variables tienen más de dos niveles podrían tratarse de manera similar a los datos binarios, considerando cada nivel de una variable como una única variable binaria. Sin embargo, este no es un enfoque atractivo ya que inevitablemente habrá muchas coincidencias negativas. Un enfoque mucho mejor es asignar una puntuación s_{ijk} de cero o uno a cada variable k, dependiendo de si los dos objetos i y j son iguales en esa variable. Luego, estas puntuaciones se promedian sobre todos las p variables para obtener el coeficiente de similitud requerido como:

$$s_{ij} = \frac{1}{p} \sum_{k=1}^{p} s_{ijk}. \tag{3.1}$$

Observe que este coeficiente de similitud es una generalización del coeficiente de coincidencia **S1** para datos binarios.

3.4.2. Medidas de distancia para datos continuos

Cuando todas las variables registradas son continuas, las proximidades entre individuos generalmente se cuantifican mediante medidas de disimilitud o medidas de distancia, donde una medida de disimilitud, se denomina *distancia métrica* si cumple la siguiente definición.

Definición 3.4.3. *Sea U un conjunto finito o infinito de elementos. Una función dist : $U \times U \longrightarrow \mathbb{R}$ se llama una distancia métrica si $\forall x, y \in U$ se tiene:*

1. $dist(x, y) \geq 0$
2. $dist(x, y) = 0 \Leftrightarrow x = y$
3. $dist(x, y) = dist(y, x)$
4. $dist(x, z) \leq dist(x, y) + dist(y, z)$, $\forall x, y, z \in U$

La distancia Euclídea es la más conocida, la de mayor uso y es la herramienta fundamental de cálculo de la mayoría de los métodos multivariados basados en distancias. Su expresión es,

$$dist(i, j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}, \tag{3.2}$$

donde x_{ik} y x_{jk} son, respectivamente, el k-ésimo valor de la variable de las observaciones p-dimensionales para los individuos i y j. Esta medida de distancia tiene la atractiva propiedad de que $dist(i, j)$, puede interpretarse como distancias físicas entre dos puntos p-dimensionales $x_i^{'} = (x_{i1}, ..., x_{ip})$ y $x_j^{'} = (x_{j1}, ..., x_{jp})$ en el espacio euclidiano. Demey et al. (2011) menciona varios inconvenientes de esta distancia: no está acotada, es sensible a cambios de escalas y considera las p variables estocásticamente independientes. El problema de la independencia aleatoria se resuelve introduciendo la *distancia de Manhattan* o también conocida como métrica *city-block*, que tiene en cuenta las correlaciones entre variables y por lo tanto la redundancia que existe entre las mismas. Su expresión es:

$$dist(i, j) = \sum_{k=1}^{p} |x_{ik} - x_{jk}| . \tag{3.3}$$

La distancia de Manhattan describe distancias en una configuración rectilínea, porque mide distancias viajado en configuración de calle. Tanto la distancia euclidiana (q = 2)

como la distancia de Manhattan (q = 1) son casos especiales de la distancia general de Minkowski. La distancia de Minkowski ($q \geq 1$), sería,

$$dist(i,j) = \left(\sum_{k=1}^{p} |x_{ik} - x_{jk}|^q \right)^{\frac{1}{q}}. \tag{3.4}$$

Para $q \longrightarrow \infty$ se convierte en la distancia de Chebyshev. En general, el parámetro q permite ajustar la sensibilidad de la métrica a las diferencias entre las coordenadas de los puntos.

3.4.3. Estandarización

En muchas aplicaciones de agrupación, las variables que describen los objetos que se van a agrupar no se medirán en las mismas unidades. Está claro que no sería sensato tratar, digamos, el peso medido en kilogramos, la altura medida en metros y la ansiedad clasificada en una escala de cuatro puntos como equivalentes en ningún sentido para determinar una medida de similitud o distancia. Cuando todas las variables se han medido en una escala continua, la solución que se sugiere con mayor frecuencia para abordar el problema de las diferentes unidades de medida es simplemente estandarizar cada variable a la varianza unitaria antes de cualquier análisis. Para este propósito se proponen utilizar una serie de medidas de variabilidad. Cuando se utilizan las desviaciones estándar calculadas a partir del conjunto completo de objetos que se van a agrupar, la técnica a menudo se denomina escala automática, puntuación estándar o puntuación z. Las alternativas son la división por las desviaciones absolutas medianas, o por los rangos, y se ha demostrado que este último supera al autoescalado en muchas aplicaciones de agrupación.

3.5. Clasificación de las técnicas clustering

Como destaqué anteriormente las técnicas de clustering se dividen en dos grandes grupos, los métodos de agrupamiento jerárquico y los métodos de agrupamiento no jerárquico.

3.5.1. Los métodos jerárquicos

Los llamados métodos jerárquicos tienen por objetivo agrupar clústeres para formar uno nuevo o bien separar alguno ya existente para dar origen a otros dos, de tal forma que, si sucesivamente se va efectuando este proceso de aglomeración o división, se minimice alguna distancia o bien se maximice alguna medida de similitud.[5] Cuando se desconoce o no se tiene idea del número de clústeres que se puedan formar, es recomendable aplicar un método jerárquico para encontrar el número de clústeres apropiados a partir de estos datos. Observe la Figura 3.1, a partir de un mismo conjunto de datos (a) se muestra cómo se forman dos (b), cuatro (c) y seis (d) clústeres según las similaridades o distancias.

Los métodos jerárquicos, además, permiten la construcción de un árbol de clasificación, que recibe el nombre de *dendrograma*, en el cual se puede seguir de forma gráfica el procedimiento de unión seguido, mostrando que clústeres se van uniendo, en qué nivel concreto lo hacen, así como el valor de la medida de asociación entre los clústeres cuando estos se agrupan. Es más, el término jerarquía hace referencia al hecho de que los grupos obtenidos cortando el dendrograma a una altura dada están necesariamente anidados dentro de los clústeres obtenidos.[6]

[5] Gallardo, J. (s. f.). *Introducción al Análisis Clúster*. Universidad de Granada.

[6] James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. *An Introduction to Statistical Learning*, cit., 521-528.

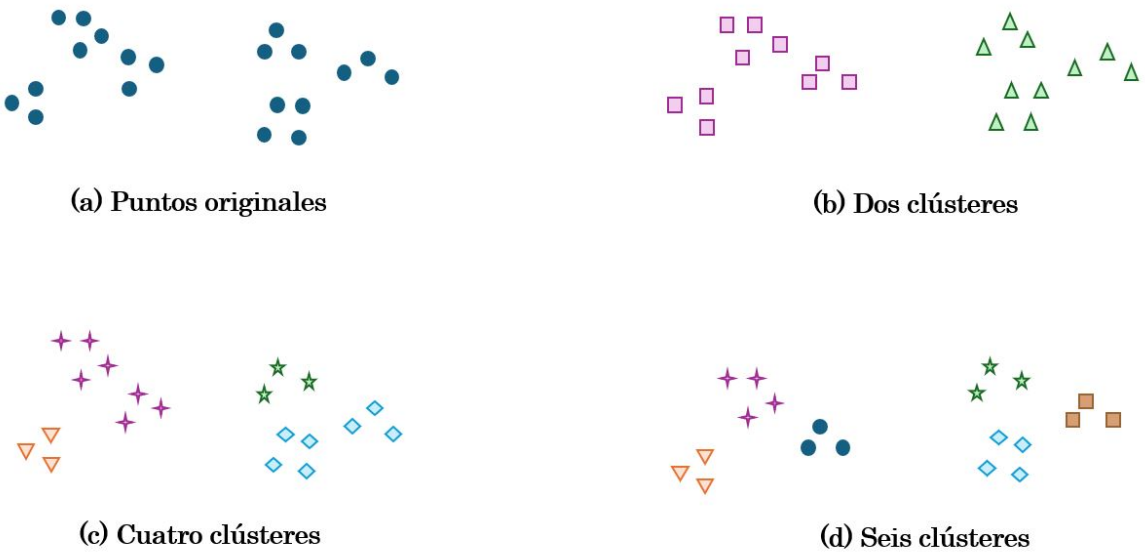


Figura 3.1: *Diferentes formas de encontrar grupos en el mismo conjunto de datos.* Fuente: Adaptado de Kumar et al. (2006).

3.5.1.1. Dendrograma

Como ya hemos mencionado, el dendrograma, o diagrama de árbol, es una representación matemática y pictórica del procedimiento jerárquico. Aquí se proporciona alguna terminología (ver Figura 3.2). Los *nodos* del dendrograma representan grupos, y las longitudes de los tallos (*alturas*) representan las distancias a las que se unen los grupos (más adelante se detallará esta vinculación). Entonces, cuanto antes se produzcan las fusiones (en la parte derecha del árbol), más similares serán los grupos de observaciones entre sí. Por otro lado, las observaciones que se fusionan más tarde (cerca de la copa del árbol) pueden ser bastante diferentes. La mayoría de los dendrogramas tienen dos aristas que emanan de cada nodo (*árboles binarios*). La disposición de nodos y tallos es la *topología* del árbol. Los nombres de los objetos adjuntos a los nodos terminales se conocen como *etiquetas*.

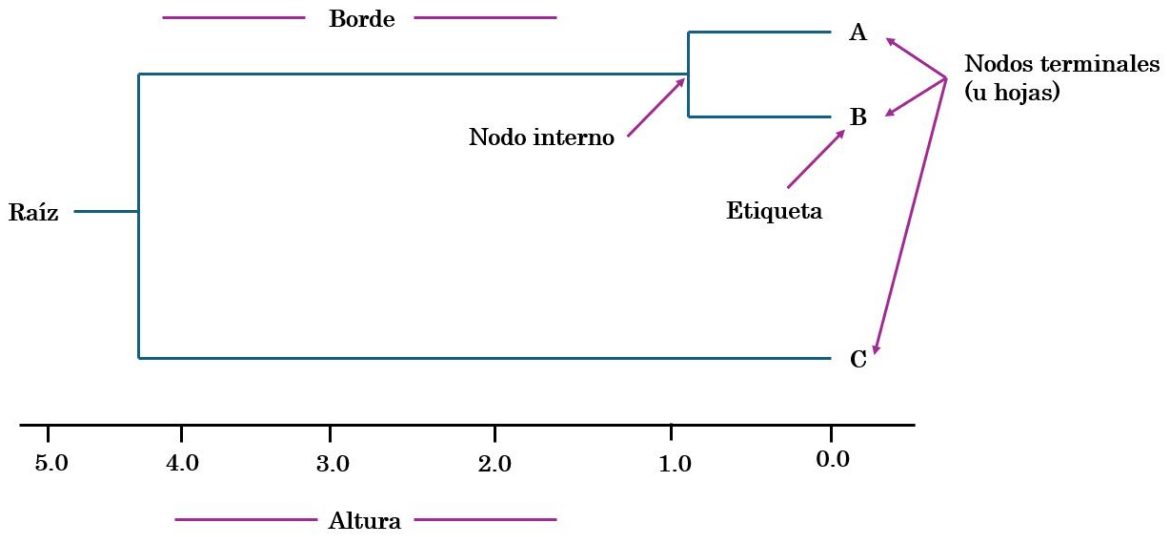


Figura 3.2: *Algunas terminologías utilizadas en la descripción de un dendrograma.* Fuente: Elaboración propia.

3.5.1.2. Elección del número de clústeres

La técnica más sencilla para determinar el número de clústeres que se pueda citar se basa en cortar el dendrograma de forma subjetiva tras visualizarlo. Siguiendo el ejemplo de la Figura 3.2, si realizamos un corte a la altura $1,5$ obtendríamos los distintos conjuntos de observaciones debajo del corte que pueden interpretarse como clústeres. Como resultado hay dos clústeres distintos, uno formado por las etiquetas {A, B} y otro por la etiqueta {C}. Se pueden hacer más cortes a medida que se desciende por el dendrograma para obtener cualquier número de clústeres.

3.5.1.3. Tipos de métodos jerárquicos

Everitt et al. (2011), James et. al (2023), Kroese et al. (2023) y Gallardo (s. f.), establecen dos formas en la que los métodos jerárquicos se subdividen y son,

- Los métodos aglomerativos.

- Los métodos disociativos o divisivos.

Métodos aglomerativos

También conocidos como ascendentes, comienzan el análisis con tantos grupos como individuos haya. A partir de estas unidades iniciales se van formando grupos, de forma ascendente, hasta que al final del proceso todos los casos tratados están englobados en un mismo grupo. Este es el tipo más común de agrupamiento jerárquico.

Algoritmo de los métodos aglomerativos

El dendrograma de agrupación aglomerativa se obtiene mediante un algoritmo extremadamente simple. Comenzamos definiendo algún tipo de medida de disimilitud entre cada par de observaciones. El algoritmo procede de forma iterativa. Comenzando en la parte inferior del dendrograma, cada una de las n observaciones se trata como su propio grupo. Los dos grupos que son más similares entre sí se fusionan para que ahora haya $n - 1$ grupos. A continuación, los dos grupos que son más similares entre sí se fusionan nuevamente, de modo que ahora hay $n - 2$ grupos. El algoritmo continúa de esta manera hasta que todas las observaciones pertenecen a un solo grupo y el dendrograma está completo. Este algoritmo parece bastante simple, pero hay un problema que no se ha abordado. ¿Cómo determinamos que un grupo debería fusionarse con otro grupo? Tenemos un concepto de disimilitud entre pares de observaciones, pero ¿cómo definimos la disimilitud entre dos grupos si uno o ambos grupos contienen múltiples observaciones? El concepto de disimilitud entre un par de observaciones debe extenderse a un par de grupos de observaciones. Esta extensión se logra desarrollando la noción de *vínculo* o más conocido por su término en inglés *linkage*, que define la disimilitud entre dos grupos de observaciones[7].

[7] James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. *An Introduction to Statistical Learning*, cit., 529.

Existen muchos tipos diferentes de agrupamiento jerárquico, dependiendo de cómo se defina la distancia entre dos puntos de datos y el vínculo entre dos grupos. A continuación vamos a detallar las más importantes siguiendo la literatura de Kroese et al. (2023). Si denotamos el conjunto de datos $\mathcal{X} = \{x_i, i = 1, ..., n\}$. Sea $dist(x_i, x_j)$ la distancia entre dos puntos x_i y x_j del conjunto de datos. Por defecto, suele ser la distancia euclídea, $dist(x_i, x_j) = ||x_i - x_j||$, siendo $|| \cdot ||$ la norma euclídea. Sean $\mathcal{I}$ y $\mathcal{J}$ dos subconjuntos disjuntos (es decir, grupos) de $\{1, ..., n\}$ donde $\mathcal{X} : (x_i, i = \mathcal{I})$ y $(x_j, j = \mathcal{J})$. Denotamos la distancia entre estos dos clústeres por $d(\mathcal{I}, \mathcal{J})$. Al establecer la función d, indicamos el criterio de vinculación. Tendríamos estos ejemplos:

- **Enlace único.** La distancia más cercana entre los grupos.

$$d_{min}(\mathcal{I}, \mathcal{J}) = \min_{i \in I, j \in J} dist(x_i, x_j). \tag{3.5}$$

- **Enlace completo.** La mayor distancia entre los grupos,

$$d_{max}(\mathcal{I}, \mathcal{J}) = \max_{i \in \mathcal{I}, j \in \mathcal{J}} dist(x_i, x_j). \tag{3.6}$$

- **Promedio del grupo.** La distancia media entre los grupos. Tenga en cuenta que esto depende de los tamaños de los grupos.

$$d_{avg}(\mathcal{I}, \mathcal{J}) = \frac{1}{|\mathcal{I}||\mathcal{J}|} \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} dist(x_i, x_j). \tag{3.7}$$

- **Varianza mínima de Ward.** Aquí, la distancia entre grupos se expresa como la cantidad adicional de "varianza" (expresada en términos de la suma de cuadrados) que se introduciría si los dos grupos se fusionaran. Más precisamente, para cualquier conjunto $\mathcal{K}$ de índices (etiquetas), sea $\bar{x}_{\mathcal{K}} = \sum_{k \in \mathcal{K}} \frac{x_k}{|K|}$ la media del grupo

correspondiente. Entonces,

$$d_{Ward}(\mathcal{I},\mathcal{J}) = \sum_{k\in\mathcal{I}\cup\mathcal{J}} ||x_k - \bar{x}_{\mathcal{I}\cup\mathcal{J}}||^2 - \left(\sum_{i\in\mathcal{I}} ||x_i - \bar{x}_{\mathcal{I}}||^2 + \sum_{j\in\mathcal{J}} ||x_j - \bar{x}_{\mathcal{J}}||^2\right). \quad (3.8)$$

Puede demostrarse el que el vínculo de Ward depende solo de la media de los clústeres y del tamaño de los clústeres $\mathcal{I}$ y $\mathcal{J}$:

$$d_{Ward}(\mathcal{I},\mathcal{J}) = \frac{|\mathcal{I}||\mathcal{J}|}{|\mathcal{I}|+|\mathcal{J}|}||\bar{x}_{\mathcal{I}} - \bar{x}_{\mathcal{J}}||^2. \quad (3.9)$$

Métodos disociativos o divisivos

En contraste con el enfoque ascendente (aglomerativo) de la agrupación jerárquica, según Kroese et al. (2023), el enfoque divisivo comienza con un grupo, que se divide en dos grupos que son lo más "disimilares" posibles, luego pueden dividirse aún más, y así sucesivamente. Podemos utilizar los mismos criterios de vinculación de la agrupación aglomerativa para dividir un grupo principal en dos grupos secundarios maximizando la distancia entre los grupos secundarios. Aunque es natural intentar agrupar datos separando los diferentes en la medida de lo posible, la implementación de esta idea tiende a escalar mal con n datos. El problema está relacionado con el conocido problema del *corte máximo*: dada una matriz $n \times n$ de costos positivos c_{ij}, $i, j \in \{1, ..., n\}$, dividir el conjunto de índices $\mathcal{I} = \{1, ..., n\}$ en dos subconjuntos $\mathcal{J}$ y $\mathcal{K}$ tales que el costo total en los conjuntos, esto es,

$$\sum_{j\in\mathcal{J}}\sum_{k\in\mathcal{K}} d_{jk}, \quad (3.10)$$

sea máximo. Recuerde que d indica el criterio de vinculación. Si en cambio maximizamos según la distancia promedio, obtenemos el criterio de vinculación promedio del grupo.

3.5.2. Los métodos no jerárquicos

En cuanto a los métodos no jerárquicos, también conocidos como partitivos o de optimización, tienen por objetivo realizar una sola partición de los individuos en **K** grupos. Ello implica que el investigador debe especificar a priori los grupos que deben ser formados, siendo ésta, posiblemente, la principal diferencia respecto de los métodos jerárquicos. La asignación de individuos a los grupos se hace mediante algún proceso que optimice el criterio de selección. Vea la Figura 3.3, mediante el algoritmo K-Medias (método no jerárquico) se puede observar como existen diferentes formas óptimas de encontrar tres grupos en el mismo conjunto de datos. Otra diferencia de estos métodos respecto a los jerárquicos reside en que trabajan con la matriz de datos original y no precisan su conversión en una matriz de distancias o similitudes.

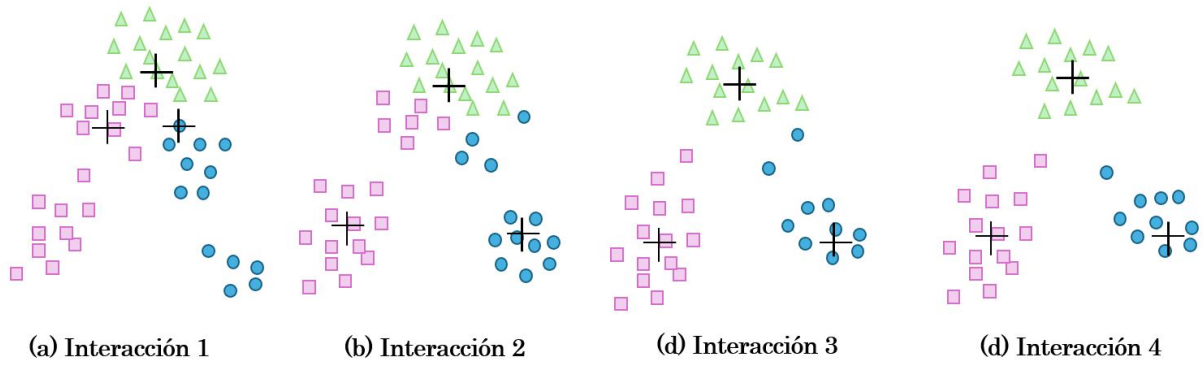


Figura 3.3: *Diferentes formas de encontrar tres grupos en el mismo conjunto de datos utilizando el algoritmo K-Medias.* Fuente: Adaptado de Kumar et al. (2006).

Según Gallardo (s. f.) los métodos no jerárquicos se pueden agrupar en cuatro familias,

- Los métodos de reasignación.
- Búsqueda de la densidad.
- Métodos directos.
- Reducción de dimensiones.

Métodos de reasignación

Permiten que un individuo asignado a un grupo en un determinado paso del proceso sea reasignado a otro grupo en un paso posterior, si ello optimiza el criterio de selección. El proceso acaba cuando no quedan individuos cuya reasignación permita optimizar el resultado que se ha conseguido. Nos encontramos con una variedad de métodos, por ejemplo, K-Medias, Quick-Clúster análisis, el método de Forgy y el método de las nubes dinámicas.

Métodos de búsqueda de la densidad

Dentro de estos métodos están los que proporcionan una aproximación tipológica y una aproximación probabilística. En el primer tipo, los grupos se forman buscando las zonas en las cuales se da una mayor concentración de individuos. Entre ellos destacan el análisis modal de Wishart, el método Taxmap, el método de Fortin. El segundo tipo, se parte del postulado de que las variables siguen una ley de probabilidad según la cual los parámetros varían de un grupo a otro. Se trata de encontrar los individuos que pertenecen a la misma distribución. Entre los métodos de este tipo destaca el método de las combinaciones de Wolf.

Métodos directos

Busca clasificar en simultáneo a individuos y variables. Un método conocido es el Block Clustering.

Métodos de reducción de dimensiones

Busca factores en el espacio de los individuos, cada factor corresponde a un grupo. Por ejemplo, el análisis Factorial tipo Q.

Capítulo 4

Propuesta metodológica

Este cuarto capítulo está dedicado a la metodología empleada en un caso de estudio con datos reales mediante la aplicación de técnicas clustering.

La propuesta metodológica sigue el ciclo de vida de un proyecto de ciencia de datos, comenzando con el planteamiento de una pregunta: cómo segmentar los clientes objetivo de una empresa de servicio a domicilio de alimentos usando técnicas clustering. Basándonos en esta pregunta, se buscó y recopiló un conjunto de datos adecuado. Posteriormente, se realizó un análisis exploratorio y preprocesamiento de datos para identificar tendencias y patrones, preparándonos para la aplicación del modelo de clustering. Se optó por la técnica de clustering jerárquica aglomerativa de Ward por sus propiedades estadísticas. Finalmente, se determinó el número de clústeres a través del dedrograma y se interpretaron los resultados obtenidos. Los códigos empleados en la metodología se publicarán en Kaggle para asegurar su accesibilidad y permitir a otros usarios revisar y utilizar el trabajo realizado.

Este capítulo metodológico incluye una última sección en la que se detalla la herramienta utilizada para el trabajo de ciencia de datos. Esta herramienta es Python. Se incluye una sección que detalla la instalación de Python mediante Anaconda. Además, se introducen las bibliotecas más importantes y usadas en este trabajo, como NumPy, Pandas y Scipy. También se proporcionan ejemplos de códigos relevantes desarrollados durante el trabajo para facilitar la replicación de los métodos y resultados así como las versiones utilizadas.

4.1. Una pregunta y unos datos que la responden

Se propuso la pregunta de cómo se pueden clasificar individuos utilizando técnicas de aprendizaje no supervisado. Es por esta pregunta que se procedió a buscar conjuntos de datos que estuvieran estrechamente relacionados en la aplicación del algoritmo de clustering.

Buscar un conjunto de datos supuso la búsqueda de un número adecuado de objetos o individuos para estudiar el fenómeno de interés. Además, se trató de buscar conjuntos de datos con variables relevantes que caracterizaran a estos objetos o individuos basados en que, posteriormente, los grupos deberán subdividirse en subgrupos homogéneos. La calidad de los datos fue otra cuestión importante que se tuvo en cuenta en la búsqueda, se tuvieron en cuenta varios aspectos como, por ejemplo, la exactitud, integridad, representatividad, coherencia, puntualidad, credibilidad, interpretabilidad y accesibilidad de los datos, presencia de ruido y valores atípicos, valores faltantes, datos duplicados, etc.

El conjunto de datos con el que trabajaremos finalmente está relacionado en la investigación de mercados, más concretamente, se propuso el objetivo de segmentar los clientes de una compañía de alimentación según sus características. Dividir a los clientes en grupos homogéneos es una de las estrategias básicas del marketing. La plataforma que se propuso para la recogida de datos en este trabajo fue Kaggle.

4.2. Análisis exploratorio de los datos

Una vez recogido el conjunto de datos sobre los clientes de la empresa, se procedió a realizar un análisis exploratorio de los datos, es decir, se utilizaron herramientas estadísticas para observar cada uno de las variables del conjunto y así tratar de comprender la estructura subyacente de los datos.

En esta etapa se utilizaron histogramas, diagrama de barras, diagrama de cajas y

bigotes, tablas resumen y cuadros de resumen estadístico para cada variable separada. Pero, según Everitt et al. (2012) en general esto no será de gran ayuda para descubrir la estructura de grupos en los datos, porque la distribución marginal de cada variable puede no reflejar con precisión la distribución multivariada del conjunto completo de variables.

4.3. Preprocesamiento de los datos

Es necesario que los datos recopilados deben ser característicos, pertinentes y de buena calidad para permitir un análisis útil. Dado que los datos reales pueden ser incompletos, ruidosos e inconsistentes, el preprocesamiento de datos es una parte esencial.

En la sección 2.1.2.3 se mencionaron algunas de las principales tareas involucradas en el preprocesamiento de datos. En nuestro caso, se decidió aplicar dos de las cuatro tareas mencionadas y son: la limpieza de datos y la reducción de los datos.

- **Limpieza de datos.**

  (i) Manejo de datos faltantes: se decidió eliminar los valores faltantes y analizar solo la parte del conjunto de datos que está completa.

  (ii) Manejo de valores atípicos. Se decidió aplicar la técnica trimming que consistía en eliminar los datos atípicos del conjunto de datos.

- **Reducción de datos.**

  (i) Agregación. Se combinó dos o más variables en una sola variable, reduciendo el número de variables.

  (ii) Selección de características. Se realizó una selección de variables o características dejando de lado variables que son redundantes o irrelevantes para realizar clústeres.

Tras realizar estos pasos obtuvimos una base de datos reestructurada.

4.4. Aplicación de la técnica clúster

Esta etapa está enfocada en la aplicación de una técnica clúster en la base de datos del cliente. Recordemos, de la sección 3.2, que para poder aplicar una técnica clúster se debe seguir una serie de etapas.

4.4.1. Paso 1. Elección de las variables

Respondiendo a las dos cuestiones que proponía Gallardo (s. f.).

- De la base de datos reestructurada se seleccionó las variables que podían resumir mejor el comportamiento del cliente.
- Se creó una tabla resumen y un cuadro de resumen estadístico que contabilizaba y daba a comprender la naturaleza de las variables.

4.4.2. Paso 2. Elección de la medida de proximidad

Como bien se ha recalcado a lo largo de este trabajo, la elección de la medida de proximidad es muy importante, ya que tiene un fuerte efecto en el dendrograma resultante y, en general, se debe prestar cuidadosa atención al tipo de datos que se agrupan.

Dado que la naturaleza de las variables era continua la idea principal fue aplicar, en primer lugar, la distancia Euclídea y así, por ejemplo, los compradores que hayan comprado muy pocos artículos en total se agruparan. Esto puede no ser deseable ya que no se tienen en cuenta otros aspectos. En segundo lugar, se aplicó la distancia de Manhattan para evitar los problemas de la distancia Euclídea.

Consideré estandarizar las variables para tener una desviación estándar uno antes de calcular la disimilitud entre las observaciones. Así a cada variable se le dará la misma importancia en el agrupamiento jerárquico utilizado. Otro motivo por el que decidí utilizar la técnica de estandarización fue porque las variables están medidas en escalas diferentes.

En resumen, se aplicó en primer lugar una estandarización de las variables y seguidamente se aplicaron las medidas de disimilitud más famosas, es decir, la distancia Euclídea y la de Manhattan. Para detalles como la descripción y formulación de las medidas mencionadas puede acceder a la sección 3.4.

4.4.3. Paso 3. El método de Ward

Una vez que se seleccionó las variables de interés, estandarizarlas y posteriormente aplicar dos medidas de distancia entre los individuos, se obtuvo finalmente dos matrices de entrada para el algoritmo jerárquico.

Esta tercera etapa corresponde a la aplicación de la técnica de clustering. Recordemos que se dividían en dos ramas, la rama jerárquica y la rama no jerárquica. Los métodos jerárquicos se subdividen en dos, las técnicas aglomerativas y divisivas, mientras que los métodos no jerárquicos son cuatro tipos, los métodos de reasignación, los métodos de búsqueda de la densidad, los métodos directos y los métodos de reducción de dimensiones.

La que pareció más interesante para utilizar en este estudio fue la técnica jerárquica aglomerativa de Ward. La elección de esta técnica fue por las propiedades que ofrecía con respecto a las demás técnicas.

Cuando se unen dos grupos, con independencia del método utilizado, la varianza aumenta. El método de Ward une los casos basándose en un criterio clásico de suma de cuadrados, produciendo grupos que minimizan la dispersión dentro del grupo en cada fusión binaria. Además, el método de Ward es interesante porque busca clústeres en el espacio euclidiano multivariado.[1]

Gallardo (s. f.) demuestra el algoritmo del método de Ward del siguiente modo. Notemos por,

[1]Murtagh, F., & Legendre, P. (2014). *Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?*. Montfort University & Université de Montréal, 274-275.

- x_{ij}^k al valor de la $j-$ésima variable sobre el $i-$ésimo individuo del $k-$ésimo clúster, suponiendo que dicho clúster posee n_k individuos.

- m_k al centroide del clúster k, con componentes m_j^k (el valor de la j-ésima componente del centroide del clúster k-ésimo).

- E_k a la suma de cuadrados de los errores del clúster k, o sea, la distancia Euclídea al cuadrado entre cada individuo del clúster k a su centroide,

$$E_k = \sum_{i=1}^{n_k} \sum_{j=1}^{n} \left(x_{ij}^k - m_j^k\right)^2 = \sum_{i=1}^{n_k} \sum_{j=1}^{n} \left(x_{ij}^k\right)^2 - n_k \sum_{j=1}^{n} \left(m_j^k\right)^2. \tag{4.1}$$

- E a la suma de cuadrados de los errores para todos los clústeres, o sea, si suponemos que hay h clústeres,

$$E = \sum_{k=1}^{h} E_k. \tag{4.2}$$

El proceso comienza con m clústeres, cada uno de los cuales está compuesto por un solo individuo, por lo que cada individuo coincide con el centro del clúster y por lo tanto en este primer paso se tendrá E_k es cero para cada clúster y con ello, E es cero. El objetivo del método de Ward es encontrar en cada etapa aquellos dos clústeres cuya unión proporcione el menor incremento en la suma total de errores, E.

Supongamos ahora que los clústeres C_p y C_q se unen resultando un nuevo clúster C_t donde $n_t = n_p + n_q$ individuos. Entonces el incremento de E será,

$$\Delta E_{pq} = E_t - E_p - E_q = \tag{4.3}$$

$$= \left[\sum_{i=1}^{n_t} \sum_{j=1}^{n} \left(x_{ij}^t\right)^2 - n_t \sum_{j=1}^{n} \left(m_j^t\right)^2\right] - \left[\sum_{i=1}^{n_p} \sum_{j=1}^{n} \left(x_{ij}^p\right)^2 - n_p \sum_{j=1}^{n} \left(m_j^p\right)^2\right] -$$

$$- \left[\sum_{i=1}^{n_q} \sum_{j=1}^{n} \left(x_{ij}^q\right)^2 - n_q \sum_{j=1}^{n} \left(m_j^q\right)^2\right].$$

Al expandir esta expresión, los términos $\sum_{i=1}^{n_p}\sum_{j=1}^{n}\left(x_{ij}^{p}\right)^2$ y $\sum_{i=1}^{n_q}\sum_{j=1}^{n}\left(x_{ij}^{q}\right)^2$ se cancelan, ya que suman exactamente los mismos valores antes y después de la unión de los clústeres. Así, el incremento de ΔE_{pq} se reduce a,

$$\Delta E_{pq} = n_p \sum_{j=1}^{n}\left(m_j^p\right)^2 + n_q \sum_{j=1}^{n}\left(m_j^q\right)^2 - n_t \sum_{j=1}^{n}\left(m_j^t\right)^2. \tag{4.4}$$

Se observa que la media ponderada del nuevo clúster C_t es,

$$m_j^t = \frac{1}{n_t}\left(\sum_{i=1}^{n_p}\sum_{j=1}^{n} x_{ij}^p + \sum_{i=1}^{n_q}\sum_{j=1}^{n} x_{ij}^q\right) = \frac{1}{n_t}\left(n_p m_j^p + n_q m_j^q\right) = \frac{n_p m_j^p + n_q m_j^q}{n_t}.$$

Ahora bien,

$$n_t m_j^t = n_p m_j^p + n_q m_j^q,$$

de donde,

$$n_t^2 (m_j^t)^2 = n_p^2 (m_j^p)^2 + n_q^2 (m_j^q)^2 + 2 n_p n_q m_j^p m_j^q,$$

y como,

$$2 m_j^p m_j^q = (m_j^p)^2 + (m_j^q)^2 - (m_j^p - m_j^q)^2,$$

sustituyendo y simplificando,

$$n_t^2 (m_j^t)^2 = n_p(n_p + n_q)(m_j^p)^2 + n_q(n_q + n_p)(m_j^q)^2 - n_p n_q (m_j^p - m_j^q)^2.$$

Dado que $n_t = n_p + n_q$, dividiendo por n_t^2 se obtiene,

$$(m_j^t)^2 = \frac{n_p}{n_t}(m_j^p)^2 + \frac{n_q}{n_t}(m_j^q)^2 - \frac{n_p n_q}{n_t^2}(m_j^p - m_j^q)^2, \tag{4.5}$$

con lo cual, sustituyendo en la expresión (3.5), la expresión (3.4), se obtiene ΔE_{pq}. Sim-

plificamos en primer lugar,

$$n_p \sum_{j=1}^{n} \left(m_j^p\right)^2 + n_q \sum_{j=1}^{n} \left(m_j^q\right)^2 - n_t \sum_{j=1}^{n} \left[\frac{n_p}{n_t}(m_j^p)^2 + \frac{n_q}{n_t}(m_j^q)^2 - \frac{n_p n_q}{n_t^2}(m_j^p - m_j^q)^2\right] =$$

$$= n_p \sum_{j=1}^{n} \left(m_j^p\right)^2 + n_q \sum_{j=1}^{n} \left(m_j^q\right)^2 - n_p \sum_{j=1}^{n}(m_j^p)^2 - n_q \sum_{j=1}^{n}(m_j^q)^2 - \frac{n_p n_q}{n_t^2} \sum_{j=1}^{n}(m_j^p - m_j^q)^2.$$

Y obtenemos ΔE_{pq},

$$\Delta E_{pq} = \frac{n_p n_q}{n_t^2} \sum_{j=1}^{n}(m_j^p - m_j^q)^2. \tag{4.6}$$

Así el menor incremento de los errores cuadráticos es proporcional a la distancia Euclídea al cuadrado de los centroides de los clústeres unidos. Es decir, la medida de proximidad es la distancia cuadrada entre los centroides pesada por un factor.

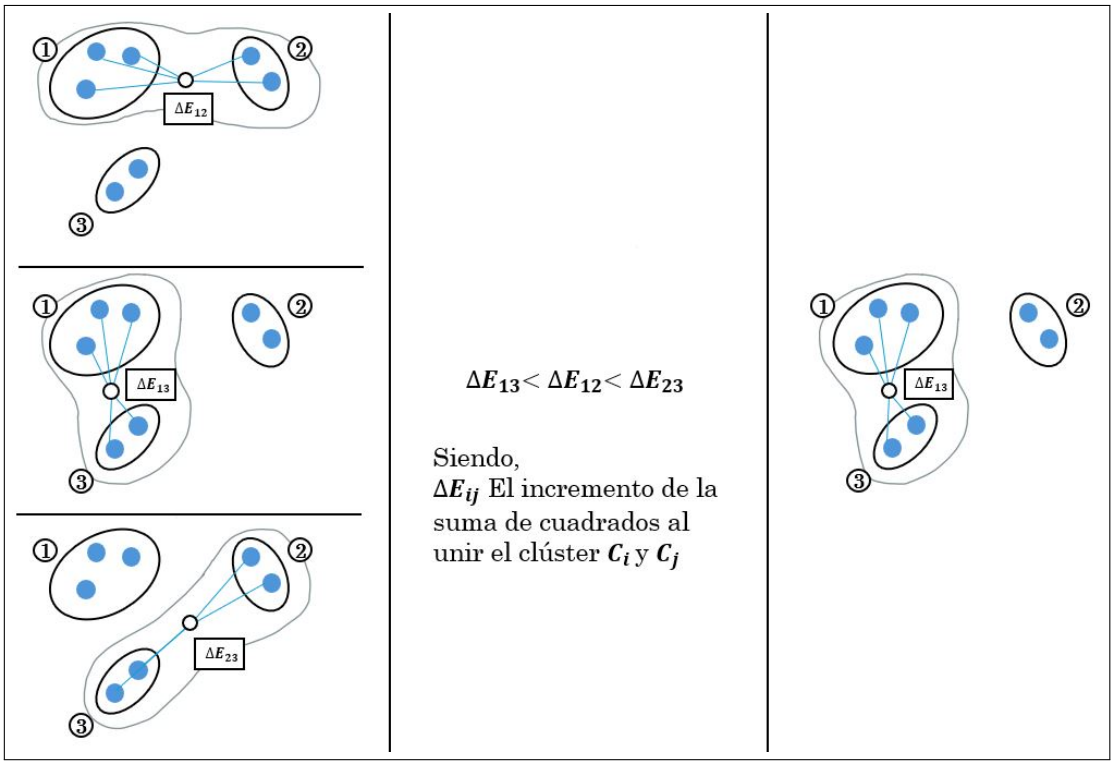


Figura 4.1: *Ejemplo del algoritmo de Ward.* Fuente: Adaptado de Pedret (2003).

La Figura 4.1 ilustra un ejemplo de cómo trabaja el algoritmo del método de Ward: el índice ΔE_{ij} hace referencia al incremento de la suma de cuadrados al unir los clústeres C_i y C_j. El algoritmo estudia el menor error de los errores en la unión del clúster C_1 y clúster C_2, luego el clúster C_1 y el clúster C_3 y por último el clúster C_3 y el clúster C_2. La unión de C_1 y C_3 tiene menor error y el algoritmo elige este para la unión.

Una vez aplicado el algoritmo de Ward se procedió graficar los dendrogramas correspondientes a cada uno de los resultados obtenidos mediante las diferentes medidas de distancia y se determinó el número de clústeres contando el dendrograma. Cada cliente fue etiquetado según el grupo al que pertenecía tras realizar el corte en el dendrograma.

4.4.4. Paso 4. Validación de los resultados e interpretación de los mismos.

Una vez aplicado el algoritmo y determinado el número de clústeres fue necesario interpretar la clasificación obtenida. Mediante cuadros de resumen estadístico, diagramas de barras y diagramas de caja y bigotes dependiendo de la naturaleza de la variable se obtuvo una visión más clara de los clústeres y nos ayudó a obtener las conclusiones.

4.5. Informe y publicación de los códigos

Los códigos desarrollados para este trabajo serán subidos a Kaggle para facilitar su acceso y uso. Esto permitirá a otros usuarios revisar, ejecutar y modificar los scripts utilizados en el análisis y modelado. Además, se proporcionarán instrucciones detalladas para la ejecución de los códigos.

4.6. Herramientas para el proyecto de ciencia de datos: lenguaje de programación con Python

Python, R, Julia, Scala y C++ son los lenguajes más utilizados en ciencia de datos. De los anteriores mencionados, la más utilizada es Python, el cual, no solamente es un lenguaje de propósito general, si no que es el más utilizado para hacer análisis de datos de alto nivel debido a la extensa biblioteca que ofrece y su amplia gama de funciones.

Esta sección comenzará introduciendo como puede instalar Python en su ordenador a través de la distribuidora Anaconda. Seguiremos con la mención de las bibliotecas que necesitaremos en el desarrollo del trabajo y terminaremos mencionando las versiones utilizadas de cada biblioteca.

4.6.1. Instalando Python

El sitio web principal de Python[2] ofrece para principiantes documentación, un tutorial, guías, ejemplos de software, etc. Es importante señalar que hay dos versiones de Python, llamadas Python 3 y Python 2. Dado que Python 2 no recibe actualizaciones de seguridad ni soporte por parte de su comunidad de desarrolladores desde el 1 de enero de 2020 este trabajo se desarrolló en Python 3.

Como hay muchos paquetes interdependientes que se utilizan frecuentemente con una instalación de Python, es conveniente instalar una distribución. En nuestro caso, se usó la distribución *Anaconda*[3] Python. El instalador de Anaconda instala automáticamente los paquetes más importantes y también proporciona un conveniente entorno de desarrollo interactivo, llamado *Visual Studio Code* (VS Code).

[2] *Welcome to Python.org.* (2024). Python.org. Última consulta 29 de junio 2024.
[3] Anaconda. (2024). *Anaconda | The Operating System for AI.* Última consulta 29 de junio 2024.

4.6.2. Bibliotecas esenciales y herramientas

A continuación, mencionamos las bibliotecas y funciones más importantes que ofrecen así como las que se utilizaron para el estudio.

4.6.2.1. NumPy

Numpy ha sido durante mucho tiempo la piedra angular de la computación numérica en Python. Ofrece las estructuras de datos, los algoritmos y el pegamento de biblioteca necesarios para la mayoría de las aplicaciones científicas que involucran datos numéricos.[4]

El paquete NumPy (`numpy`) contiene funciones matemáticas estándar, como `sin`, `cos`, `tan`, etc., así como funciones eficientes para la generación de números aleatorios, álgebra lineal y cálculo estadístico.[5] La funcionalidad principal de NumPy es la clase `ndarray`, una matriz multidimensional (n-dimensional). Todos los elementos de la matriz deben ser del mismo tipo. Una matriz NumPy se ve así:

In[1]

```
1 import numpy as np
2
3 x = np.array([[1, 2, 3], [4, 5, 6]])
4 print("x:\n{}".format(x))
```

Out[1]:

```
x:
[[1 2 3]
 [4 5 6]]
```

[4] McKinney, W. (2017). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython.* (2.ª ed.). O'Reilly Media Inc., 4.

[5] *NumPy documentation - NumPy v2.0 Manual.* (s. f.). Última consulta 13 de junio de 2024.

Usaremos mucho NumPy en este trabajo y nos referiremos a los objetos de la clase `ndarray` NumPy como "matrices NumPy" o simplemente "matrices".

4.6.2.2. Pandas

La biblioteca de Pandas (`pandas`) proporciona varias herramientas y estructuras de datos para el análisis de datos, incluida la clase fundamental `DataFrame`.

Un `DataFrame` de Pandas es una tabla, similar a una hoja de cálculo de Excel. Pandas proporciona una gran variedad de métodos para modificar y operar en esta tabla; en particular, permite consultas similares a SQL y uniones de tablas. A diferencia de NumPy, que requiere que todas las entradas de una matriz sean del mismo tipo, Pandas permite que cada columna tenga un tipo independiente (por ejemplo, números enteros, fechas, números de punto flotante y cadenas). Otra herramienta valiosa proporcionada por Pandas es su capacidad para ingerir datos desde una gran variedad de formatos de archivos y bases de datos, como archivos SQL, Excel y archivos de valores separados por comas (CSV).[6] Aquí le mostramos un ejemplo de como se crea un `DataFrame` usando un diccionario:

In[2]

```
import pandas as pd
# Vamos a crear un simple conjunto de datos sobre personas
data = {'Nombre': ["John", "Anna", "Peter", "Linda"],
        'Localidad' : ["New York", "Paris", "Berlin", "London"],
        'Edad' : [24, 13, 53, 33]
        }
data_pandas = pd.DataFrame(data)
print(data_pandas)
```

Out[2]:

[6] *pandas - Python Data Analysis Library.* (s. f.). Última consulta 13 de junio de 2024.

```
  Nombre Localidad  Edad
0   John  New York    24
1   Anna     Paris    13
2  Peter    Berlin    53
3  Linda    London    33
```

Hay varias formas posibles de consultar esta tabla. Por ejemplo, se puede seleccionar todas las filas que tengan una columna de edad superior a 30 de la siguiente forma:

In 3

```
# Seleccionar todas las filas que tengan una columna de edad
# superior a 30
print(data_pandas[data_pandas.Edad > 30])
```

Out[3]:

```
  Nombre Localidad  Edad
2  Peter    Berlin    53
3  Linda    London    33
```

4.6.2.3. Matplotlib

La biblioteca de Matplotlib es la principal biblioteca de trazado científico en Python. Proporciona funciones para realizar visualizaciones con calidad de publicación, como gráficos de líneas, histogramas, diagramas de dispersión, etc. Visualizar los datos y diferentes aspectos de un análisis puede brindar información importante; es por ello que usaremos Matplotlib para todas nuestras visualizaciones.[7] Vamos a ejemplificar esta librería.

[7] *Matplotlib - Visualization with Python.* (s. f.). Última consulta 13 de junio de 2024.

In[4]

```
import matplotlib.pyplot as plt
# Generamos una secuencia de numeros de -10 a 10 con 100
# pasos intermedios
x = np.linspace(-10, 10, 100)
# Creamos una segunda matriz usando el seno
y = np.sin(x)
# La funcion plot hace un grafico lineal de un array contra otro
plt.plot(x, y, marker="x")
```

Out[4]:

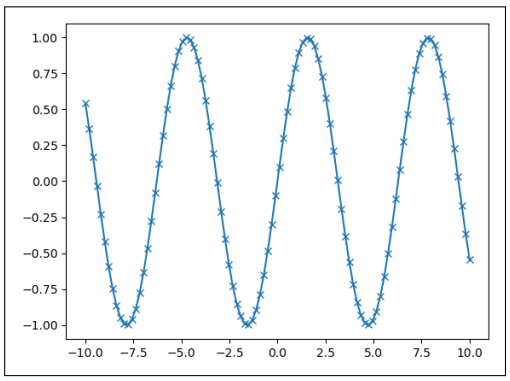


Figura 4.2: *Gráfico simple de la función seno utilizando* ***matplotlib****.* Fuente: Elaboración propia.

4.6.2.4. Seaborn

Seaborn es una biblioteca para crear gráficos estadísticos en Python. Se basa en Matplotlib y se integra estrechamente con las estructuras de datos de Pandas. Seaborn ayuda a explorar y comprender los datos. Sus funciones de trazado operan en marcos de datos y matrices que contienen conjuntos de datos completos y realizan internamente el mapeo semántico y la agregación estadística necesarios para producir gráficos informativos. Su API declarativa orientada a conjuntos de datos le permite centrarse en lo que significan

los diferentes elementos de sus gráficos, en lugar de en los detalles de cómo dibujarlos.[8] Un ejemplo sería la Figura 4.3.

In[5]

```
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
# Generamos 1000 numeros aleatorios de una distribucion normal:
data = np.random.randn(1000)
# Creamos una grafica con seaborn:
sns.histplot(data, kde=True)

plt.title("Histograma de Numeros Aleatorios")
plt.xlabel("Valor")
plt.ylabel("Frecuencia")
plt.show()
```

Out[5]:

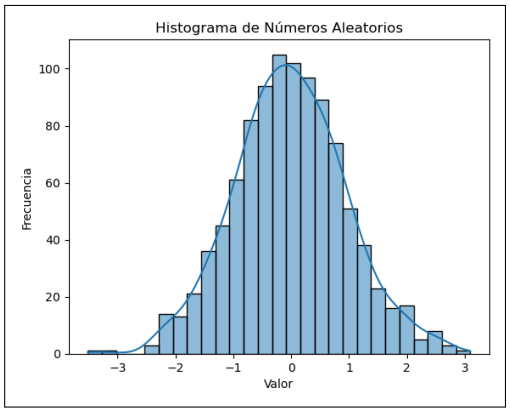


Figura 4.3: *Gráfico simple utilizando* ***seaborn****.* Fuente: Elaboración propia.

[8] *seaborn: statistical data visualization - seaborn 0.13.2 documentation.* (s. f.). Última consulta 13 de junio de 2024.

4.6.2.5. Scikit-Learn

La biblioteca Scikit-Learn es un proyecto de código abierto, lo que significa que su uso y distribución son gratuitos, y cualquiera puede obtener fácilmente el código fuente para ver qué sucede detrás de escenas. El proyecto Scikit-Learn se desarrolla y mejora constantemente y cuenta con una comunidad de usuarios muy activa. Contiene una serie de algoritmos de aprendizaje automático de última generación, así como documentación completa sobre cada algoritmo.[9]

La biblioteca incluye una variedad de algoritmos que son muy utilizado por su sencillez y su amplitud. El nombre del módulo es `sklearn`. La función que utilizaremos será `StandardScaler` que es aquella que estandariza los datos.

4.6.2.6. SciPy

La biblioteca SciPy (nombre del módulo `scipy`) es una colección de funciones para informática científica en Python. Proporciona, entre otras funciones, rutinas avanzadas de álgebra lineal, optimización de funciones matemáticas, procesamiento de señales, funciones matemáticas especiales y distribuciones estadísticas.[10] Scikit-Learn se basa en la colección de funciones de SciPy para implementar sus algoritmos. La parte más importante de Scipy para nosotros son los submódulos `sklearn.preprocessing` y `scipy.cluster.hierarchy`. El primer submódulo es porque incluye la función `pdist`, que es la que nos permite calcular las distancias utilizando el parámetro `metric`. El segundo submódulo es el que proporciona las funciones para realizar análisis de clúster mediante métodos jerárquicos. Este submódulo incluye varias funciones, siendo `linkage` la función de vínculo entre clústeres, `dendrogram` la función que nos permite visualizar el dendrograma resultante y `cut_tree`,

[9] *scikit-learn: machine learning in Python - scikit-learn 1.5.0 documentation.* (s. f.). Última consulta 13 de junio de 2024.

[10] *SciPy* -. (s. f.). Última consulta 13 de junio de 2024.

función que nos permite cortar el árbol jerárquico y obtener las etiquetas de los individuos.

Vamos a ejemplificar estos códigos. En primer lugar vamos a generar una matriz de datos a través de `numpy`, usaremos la distancia euclídea para calcular las distancias entre las observaciones, posteriormente aplicaremos el método jerárquico mediante el vínculo único (esto se especifica a través de `method="single"`). Para mostrar el dendrograma necesitamos de una biblioteca que realice gráficas, esta sería `matplotlib`. Y por último, utilizaremos `cut_tree` para cortar el árbol y obtener las etiquetas de los individuos.

In[6]

```
import numpy as np
from scipy.cluster.hierarchy import linkage, dendrogram, cut_tree
import matplotlib.pyplot as plt
from scipy.spatial.distance import pdist

X = np.array([[1.5, 2.9], [2.7, 3.5], [3.7, 4.9],
              [5.4, 6.9], [8.5, 8.5]])

# Calculamos las distancias euclideas entre las observaciones
distances = pdist(X, metric='euclidean')

# Aplicamos el metodo jerarquico mediante el vinculo unico
Z = linkage(distances, method='single')

# Dendrograma
dendrogram(Z)
plt.title("Dendrograma de Cluster Jerarquico (Vinculo unico)")
plt.xlabel("Indice de Muestras")
plt.ylabel("Distancia")

# Cortamos el arbol y obtenemos las etiquetas de los individuos
labels = cut_tree(Z, n_clusters=3)
```

Puede consultar el Anexo I recogido en este documento donde se indican las numerosas herramientas mencionadas y utilizadas en este trabajo.

4.6.3. Versiones utilizadas en este trabajo

Las siguientes versiones de las bibliotecas mencionadas anteriormente con las que se trabajó fueron:

In[7]

```
import sys
print("Python version: {}".format(sys.version))
import numpy as np
print("NumPy version: {}".format(np.__version__))
import pandas as pd
print("pandas version: {}".format(pd.__version__))
import matplotlib
print("matplotlib version: {}".format(matplotlib.__version__))
import seaborn
print("matplotlib version: {}".format(seaborn.__version__))
import sklearn
print("scikit-learn version: {}".format(sklearn.__version__))
import scipy as sp
print("SciPy version: {}".format(sp.__version__))
```

Out[6]:

```
Python version: 3.10.9 | packaged by Anaconda, Inc.
                       | (main, Mar  1 2023, 18:18:15)
                         [MSC v.1916 64 bit (AMD64)]
NumPy version: 1.23.5
pandas version: 1.5.3
matplotlib version: 3.7.0
matplotlib version: 0.12.2
```

```
scikit-learn version: 1.2.1
SciPy version: 1.10.0
```

Ahora que tenemos todo configurado, podemos sumergirnos en la aplicación de aprendizaje automático y las técnicas clústering en un caso con datos reales.

Capítulo 5

Marco experimental: caso de estudio sobre segmentación de clientes

En este quinto capítulo se presenta un caso de estudio basado en los datos de clientes de iFood, la principal aplicación de entrega de comida a domicilio en Brasil. Utilizando un conjunto de datos de 2,240 clientes obtenidos tras una campaña piloto, se aplicará una metodología de análisis y segmentación de clientes.

A continuación, se mostrarán los resultados obtenidos mediante la metodología descrita del anterior capítulo, que incluye el análisis exploratorio de datos, la limpieza de datos, el tratamiento de valores atípicos y la creación de nuevas variables. Utilizaremos el método jerárquico aglomerativo de Ward para agrupar a los clientes en función de sus similitudes. Este proceso incluirá la normalización de los datos, la selección de variables relevantes y la generación de dendrogramas para visualizar la formación de clústeres.

Finalmente, se presentarán y analizarán los resultados obtenidos. Se detallarán las características de cada clúster y su relevancia para la estrategia de marketing de iFood, proporcionando una comprensión profunda del comportamiento de los clientes y permitiendo la optimización de futuras campañas de marketing.

5.1. Caso de estudio

iFood es la principal aplicación de entrega de comida a domicilio de Brasil, con presencia en más de mil ciudades. Según su página web: [1] "*Somos la mayor empresa foodtech de América Latina y, con gran responsabilidad, también queremos promover un cambio social positivo en nuestra sociedad. Conocemos el tamaño de este desafío, pero también sabemos que podemos llegar mucho más lejos.*" El término *foodtech* -un vocablo inglés que fusiona *food* (comida) y *technology* (tecnología)- hace referencia a la unión de la industria tecnológica y la innovación con la industria alimentaria. Para ser más precisos, iFood es una empresa que tiene por proyectos aprovechar las tecnologías como el big data y la Inteligencia Artificial (IA), entre otras, para transformar la industria agroalimentaria en un sector más moderno, sostenible y eficiente en todas sus etapas, que abarcan desde la elaboración de los alimentos hasta la distribución y el consumo.

En 2012, la empresa iFood, en base a las pérdidas futuras, decidió invertir en la creación de un modelo predictivo para obtener mayores ganancias en la próxima campaña de marketing programada para el siguiente mes. Así fue que, el equipo de marketing realizó una campaña piloto en la que participaron 2,240 clientes. Los clientes fueron seleccionados al azar y contactados por teléfono para la adquisición de un nuevo dispositivo. Durante los meses siguientes, los clientes que compraron la oferta fueron debidamente etiquetados. Debido a que la tasa de éxito de la campaña fue del 15 % el objetivo actual del equipo era entonces desarrollar un modelo que estudie el comportamiento del cliente y aplicarlo al resto de la base de clientes.[2]

El conjunto de datos que se recogió dos años después de estos clientes en base al estudio de comportamiento del cliente para su última campaña son los que utilizaremos

[1] iFood. (2021). iFood: Relato de Impacto. https://institucional.ifood.com.br/wp-content/uploads/2024/01/Relato-de-Impacto-iFood-2021-4.pdf

[2] Nailson. (2020). GitHub. ifood-data-business-analyst-test/iFood Data Analyst Case.pdf.

para nuestro caso de estudio. Recuperados de la plataforma Kaggle[3].

Analizar la personalidad del cliente y descubrir información del cliente ayuda a la empresa a comprender mejor a sus clientes y les facilita la modificación de los productos en función de las necesidades, comportamientos y preocupaciones específicas de los distintos tipos de clientes. Por ejemplo, en lugar de gastar dinero en comercializar un nuevo producto a todos los clientes de la base de datos de la empresa, ésta puede analizar qué segmento de clientes tiene más probabilidades de comprar el producto y comercializarlo sólo en ese segmento concreto.

5.2. Análisis exploratorio de los datos

El conjunto de datos contiene las características sociodemográficas y firmográficas de 27 variables de estudio sobre los 2,240 clientes. La información tomada de los clientes incluye también el gasto en la venta de las cinco categorías de productos principales de la empresa: vinos, productos cárnicos raros, frutas exóticas, pescados especialmente preparados y productos dulces. Estos se pueden dividir además en oro y productos regulares. También se tomó en cuenta el número de compras en los tres canales de venta de iFood: tiendas físicas, catálogos y página web de la empresa. Estas y todas las características de estudio se encuentran recogidos en la Tabla 5.1 así como una breve descripción de las mismas. Para más información puede consultar el cuadro resumen en el Anexo II recogido en este documento.

El mayor número de clientes está concentrado entre los 30 y 60 años. Siendo en su mayoría mayores de 35 y 45 años, con un pico máximo alrededor de los 40 años. Este grupo representa la mayor parte de los clientes, con casi 400 personas en ese rango de edad. Además, los ingresos anuales del hogar de los clientes están mayoritariamente por

[3] *Customer Personality analysis.* (2021). Kaggle. https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis/data

Tabla 5.1: *Conjunto de variables del problema de iFood.* Fuente: Elaboración propia.

Variable	**Descripción**
ID	Identificador único del cliente
Edad	Edad del cliente
Educación	Nivel de educación del cliente
Estado_Marital	Estado civil del cliente
Ingresos	Ingresos anuales del hogar del cliente
Niños_Hogar	Número de niños en el hogar del cliente
Adolescentes_Hogar	Número de adolescentes en el hogar del cliente
Dt_Customer	Fecha de alta del cliente en la empresa
Recency	Número de días desde la última compra del cliente
Queja	Si el cliente se quejó en los últimos 2 años
Vino	Cantidad gastada en vino en los últimos 2 años
Frutas	Monto gastado en frutas en los últimos 2 años
Carne	Cantidad gastada en carne en los últimos 2 años
Pescado	Cantidad gastada en pescado en los últimos 2 años
Dulces	Cantidad gastada en dulces en los últimos 2 años
Oro	Cantidad gastada en oro en los últimos 2 años
NumDesCompras	Número de compras realizadas con descuento
AcceptedCmp1	Si el cliente aceptó la oferta en la primera campaña
AcceptedCmp2	Si el cliente aceptó la oferta en la segunda campaña
AcceptedCmp3	Si el cliente aceptó la oferta en la tercera campaña
AcceptedCmp4	Si el cliente aceptó la oferta en la cuarta campaña
AcceptedCmp5	Si el cliente aceptó la oferta en la quinta campaña
AcceptedLastCmp	Si el cliente aceptó la oferta en la última campaña
NumWebCompras	Número de compras realizadas a través del sitio web de la empresa
NumCatalogCompras	Número de compras realizadas mediante un catálogo
NumTiendaCompras	Número de compras realizadas directamente en tiendas
NumWebVisitasMes	Número de visitas al sitio web en el último mes

debajo de las 100,000 unidades monetarias con un alcance mayor alrededor de 50,000 unidades monetarias, lo que indica que tienen ingresos bajos a moderados. Observe el panel izquierdo y derecho de la Figura 5.1, respectivamente.

Los clientes encuestados son en su mayoría graduados que no decidieron continuar con sus estudios, exactamente el 50,36 % de los clientes. Le sigue el 21.71 % de clientes que

estudiaron estudiado un posgrado, luego los clientes que estudiaron un máster (el 16.47 %) y por último los que están estudiando un grado y no han terminado (el 11.46 %). Además, el 64.53 %, son solteros cuando un 36.47 % están casados. Observe los paneles izquierdo y derecho de la Figura 5.2, respectivamente.

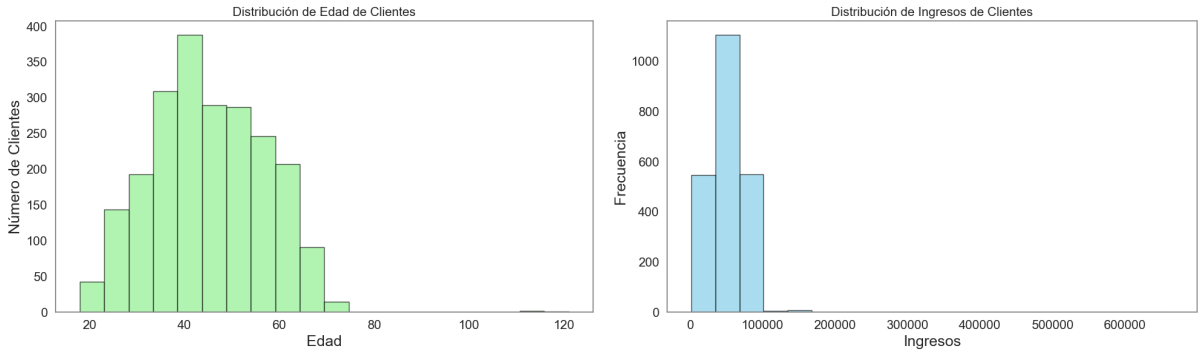


Figura 5.1: Izquierda: *Distribución de la edad de los clientes.* Derecha: *Ingreso anual del hogar de los clientes.* Fuente: Elaboración propia.

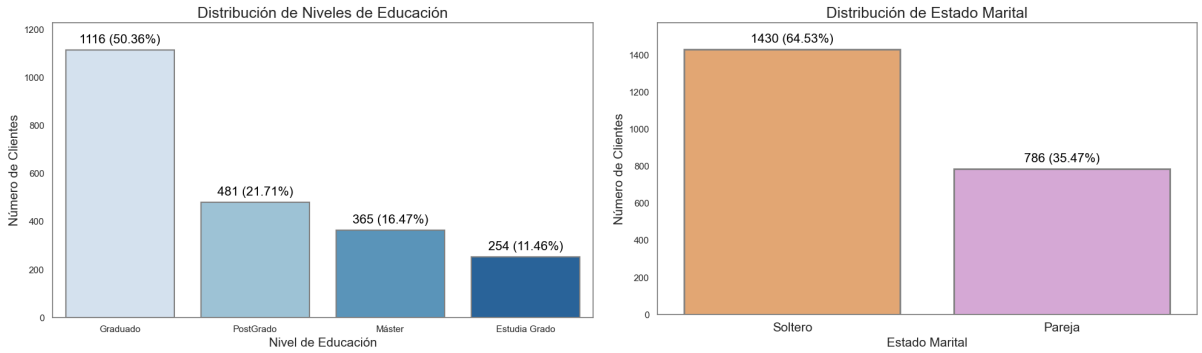


Figura 5.2: Izquierda: *Nivel de educación según el cliente.* Derecha: *Estado marital de los clientes.* Fuente: Elaboración propia.

Si observamos la Figura 5.3 que hace referencia a los gastos de los clientes en los principales productos de la empresa: los clientes tienden a gastar más en vino (a) y con un gasto entre 100 y 500 unidades monetarias. Esto puede sugerir que el vino es un artículo popular y que algunos clientes hacen compras considerablemente grandes. En

contraste, los gastos en frutas (b) son generalmente bajos, con la mayoría de gastos en 25 unidades monetarias. Sin embargo, hay muchos clientes que gastan más de 75 unidades monetarias, mostrando una gran cantidad de valores atípicos. Esto puede indicar que, aunque la mayoría de las compras son pequeñas, hay clientes que compran frutas en grandes cantidades ocasionalmente. Los clientes también gastan bastante en carne (c), con un gasto entre 100 y 300 unidades monetarias. Ocurre que la carne también es otro artículo en el que los clientes invierten considerablemente. Similar a las frutas, los gastos en pescado (d) son bajos, gastando alrededor de 25 unidades monetarias. Los gastos en dulces (e) son moderados, con la mayoría de gastos entre 25 y 75 unidades monetarias. Hay algunos valores atípicos, pero no tan extremos como en el caso del vino o la carne. Esto sugiere que los dulces son un artículo en el que los clientes gastan de forma moderada y más constante. En cuanto al gasto en artículos de mayor calidad (f), los clientes tienden a gastar una cantidad considerable, similar a los dulces. Esto indica que aunque la mayoría de las compras son moderadas, hay compras ocasionales de oro en grandes cantidades.

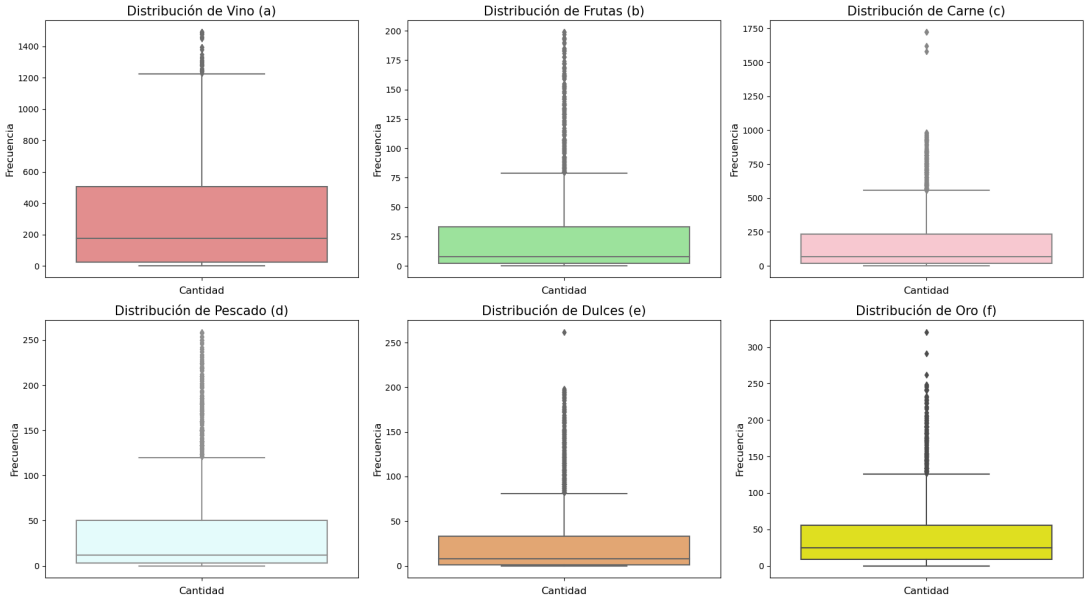


Figura 5.3: *Gasto en Vino (a), Gasto en Fruta (b), Gasto en Carne (c), Gasto en Pescado (d), Gasto en Dulces (e) y Gasto en Oro (f).* Fuente: Elaboración propia.

La Figura 5.4 muestra la distribución del número de compras realizadas con descuento, realizadas a través de la página web, por catálogo y por tienda. Además también se muestra el número de visitas de los clientes de la página web en el último mes. El mayor número de compras con descuento se encuentra entre 1 y 3 compras. Mientras que las compras que se realizan en la web, mayoritariamente son entre 1 y 4 compras. En el caso de las compras por catálogo, muchos clientes realizan entre 0 y 3 compras, donde se observa que hay un pico en 0 compras por catálogo lo que indica que muchos de los clientes prefieren otras opciones de compra. Las compras en tienda están más distribuidas, con muchos clientes realizando entre 1 y 5 compras y un número significativo de clientes que realizan hasta 10 compras en tienda. Las visitas web al mes muestran una distribución diferente, con muchos clientes visitando la web entre 0 y 8 veces al mes y la frecuencia disminuye gradualmente después de 8 visitas.

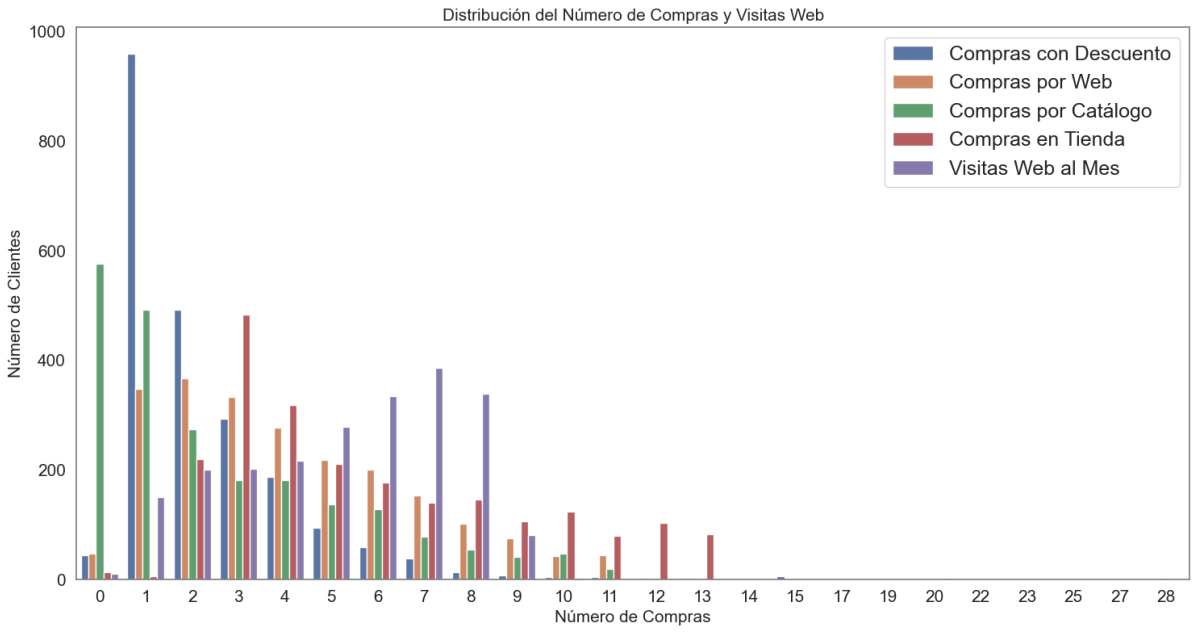


Figura 5.4: *Distribución del número de compras y visitas en la página web.* Fuente: Elaboración propia.

5.3. Análisis preliminar de las variables

Se realizó un conteo de los valores faltantes en la base de datos a través de un cuadro resumen y se observó que la procedencia de los valores faltantes procedían de una única variable, los *ingresos*. Siendo 24 los registros encontrados. Se procedió a eliminarlos dado que el número de datos faltantes suponía el 1,07 %. Consulte el Anexo III recogido en este trabajo para el cuadro resumen del conteo de valores faltantes.

Para comprender además el fenómeno en estudio de estas variables, se realizó un cuadro de resumen estadístico por variable de la base de datos que puede consultar en el Anexo IV recogido en este documento. Nos encontramos ante la presencia de datos atípicos en las variables *edad* e *ingresos*. La *edad* tiene un pico en 121 años mientras que, los *ingresos*, existen clientes que se encuentran significativamente por encima del rango intercuartil superior. El hecho de que los casos atípicos se localicen en la parte superior de la distribución indica que se trata de distribuciones asimétricas positivas, lo cual se deberá corregir antes de seguir con los siguientes pasos. Se decidió entonces aplicar la técnica trimming en la cual se eliminó aquellos registros de clientes con una edad superior a los 80 años y también aquellos donde los ingresos anuales eran mayor de 200,000 unidades monetarias. Son 4 registros los eliminados suponiendo un 0,18 % menos de la base de datos.

Para el problema planteado, se decidió agregar una variable al estudio combinando tres variables del conjunto de datos además de realizar una selección de variables que pueden resultar interesantes en lugar de otras. Por un lado, la variable que se decidió incluir fue *Unidad_familiar* (el número total personas que viven en el mismo hogar). La variable *Unidad_familiar* fusiona las variables estado marital, el número de niños y el número de adolescentes proporcionando el total de personas que viven en un mismo hogar: si el cliente tiene pareja y/o está casado constituyen dos personas las que viven en un mismo hogar, en el caso contrario, una persona. A esto se le añade el número de niños y adolescentes.

Por otro lado, la selección de variables que se han considerado para el estudio y que pueden resultar de interés fueron la edad, los ingresos, el nivel de educación, el estado marital, la unidad familiar, las cantidad gastada por el cliente en cada una de las cinco categorías de productos principales que vende la empresa, el número de compra realizadas con descuento, mediante la página web, el catálogo y la tienda y por último el número de visitas a la página web de iFood.

Finalmente, contamos contamos con un conjunto de datos reestructurado de 16 variables de estudio de 2,212 clientes. Siendo el 1,25 % los registros eliminados.

5.4. Aplicación del algoritmo de Ward

5.4.1. Paso 1. Elección de las variables

Respondiendo a la primera cuestión que propone Gallardo (s.f.) en esta primera etapa, las variables que se consideraron relevantes del conjunto de datos reestructurado son las que se muestran en la Tabla 5.2. Puede consultar el cuadro resumen de estas mismas recogido en el Anexo V de este documento.

Tabla 5.2: *Variables que finalmente se consideraron para el estudio.* Fuente: Elaboración propia.

Tipo		**Variable**	
Cuantitativa	*Continua*	Edad	Carne
		Ingresos	Pescado
		Vino	Dulces
		Frutas	Oro
	Discreta	NumDesCompras	NumTiendaCompras
		NumWebCompras	NumWebVisitasMes
		NumCatalogCompras	Unidad_familiar
Cualitativa	*Ordinal*	Educación	
	Nominal	Estado_Marital	

Tabla 5.3: *Recuento de las variables finales para el estudio según su naturaleza.* Fuente: Elaboración propia.

Tipo de variable		**Total**
Cuantitativa	*Continua*	8
	Discreta	6
Cualitativa	*Ordinal*	1
	Nominal	1

Respondiendo a la segunda cuestión que propone Gallardo (s.f.), contamos con 16 variables de estudio de los 2,212 clientes. Puede consultar el recuento de variables según su naturaleza en la Tabla 5.3.

Las variables discretas las consideramos como variables continuas teniendo en cuenta los futuros inconvenientes y manteniendo precaución en las interpretaciones. En el caso de las variables ordinales, pueden considerarse como variables cuantitativas si la asignación del ranking no es caprichosa sino que refleja en cierta forma una diferencia entre los estados de la variable.[4] Por ejemplo, la variable que indica los niveles de educación junto a sus categorías: estudia grado, graduado, máster y posgrado; puede ser razonable asignarle valores: 0, 1, 2 y 3, respectivamente ya que las categorías consecutivas pueden considerarse como equidistantes. En cuanto al estado marital sobre el cliente, podemos interpretar que aquellos que tienen pareja y/o están casados constituyen una pareja de dos personas que viven en un mismo hogar y si es soltero sería una. De esta manera, la nuevas variables numéricas podría ser tratada como variables cuantitativas.

5.4.2. Paso 2. Elección de la medida de proximidad

Se escaló el conjunto de datos y posteriormente se aplicó la distancia Euclídea y la de Manhattan obteniendo dos matrices de distancias que determinarán la unión de las

[4]Demey, J. R., Pla, L., Vicente-Villardón, J. L., Di Rienzo, J. A., & Casanoves, F. *Valoración y análisis de la diversidad funcional*, cit., 53.

observaciones para la formación de grupos final.

5.4.3. Paso 3. El método de Ward

A continuación se presentarán los diferentes grupos obtenidos bajo la implementación del método jerárquico de Ward sobre la matriz de distancias Euclídea y la matriz de distancias de Manhattan obtenidas de la base reestructurada de los datos. Se utilizará el dendrograma para la representación visual de los clústeres así como el corte que determina el número de grupos.

5.4.3.1. Clasificación y agrupación de clientes a través del método Ward en la matriz de distancias Euclídea

El análisis de la segmentación del cliente mediante el dendrograma obtenido de aplicar el método Ward en la matriz de distancias Euclídea clasificó y agrupó los 2,212 clientes en estudio de acuerdo al nivel de disimilitud entre los clientes.

En la Figura 5.5 puede observar el dendrograma obtenido y la línea horizontal que corta el dendrograma y determina el número de clústeres. El eje de abscisas muestra el orden en el que los 2,212 clientes se agrupan en base a la distancia marcada por el eje de ordenadas. El umbral de corte establecido se encuentra a la altura 40. El número de clústeres considerados a partir del umbral son cinco y puede observarlo mediante los colores que aparecen por debajo de la línea de corte y que indica que clústeres ha considerado el algoritmo.

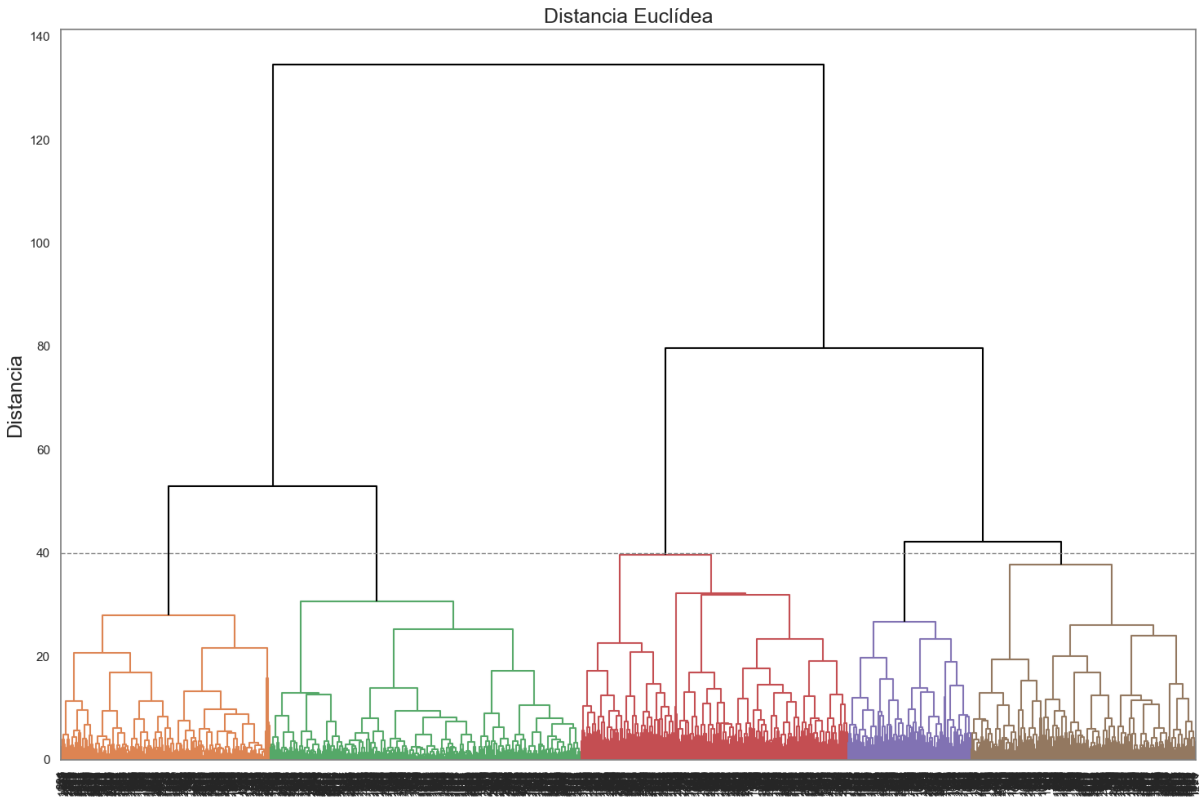


Figura 5.5: *Dendrograma construido a partir de la matriz de distancias Euclídea.* Fuente: Elaboración propia.

5.4.3.2. Clasificación y agrupación de clientes a través del método Ward en la matriz de distancias de Manhattan

El análisis de la segmentación del cliente mediante el dendrograma obtenido de aplicar el método Ward en la matriz de distancias de Manhattan clasificó y agrupó los 2,212 clientes en estudio de acuerdo al nivel de disimilitud entre los clientes.

En la Figura 5.6 puede observar el dendrograma obtenido y la línea horizontal que corta el dendrograma y determina el número de clústeres. El eje de abscisas muestra el orden en el que los 2,212 clientes se agrupan en base a la distancia marcada por el eje de ordenadas. El umbral de corte establecido se encuentra a la altura 150. El número de clústeres considerados a partir del umbral son cuatro y puede observarlo mediante

los colores que aparecen por debajo de la línea de corte y que indica que clústeres ha considerado el algoritmo.

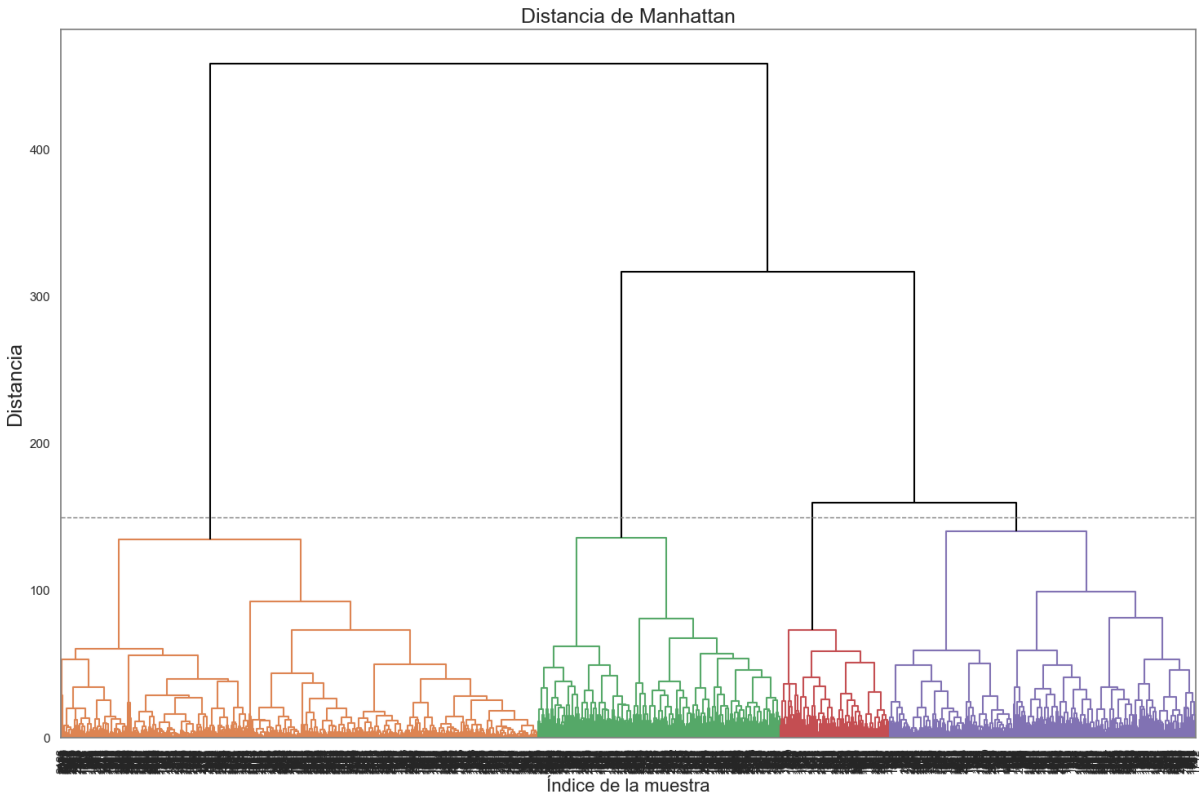


Figura 5.6: *Dendrograma construido a partir de la matriz de distancias de Manhattan.* Fuente: Elaboración propia.

5.4.4. Paso 4. Validación de los resultados

En base a los resultados anteriores, vamos a interpretar los resultados obtenidos. Para ello, se utilizará tablas que resumen la información de los grupos formados a partir del algoritmo, descripción de cada uno de ellos y anexos donde están las gráficas correspondientes a cada análisis. Además, se etiquetará cada uno de los grupos.

Vamos a identificar los distintos grupos obtenidos a través de las dos entradas de datos, la matriz de distancias euclídeas y la matriz de distancias de Manhattan.

5.4.4.1. grupos a través de la distancia Euclídea y el método de Ward

A partir de la entrada de la matriz de distancias euclídeas y el uso del algoritmo de Ward, obtuvimos el número de cinco clústeres, los cuales vamos a detallar aquí:

- **Clúster 1:** formado por 438 clientes que se encuentran mayormente en un rango de edad entre los 40 y 60 años. Predominan los clientes graduados que decidieron no continuar sus estudios (39.7 %), siguiéndole clientes que estudiaron un posgrado (32.6 %), un máster (22.6 %) y, por último, los que están estudiando un grado, que son un 5 % de personas restantes. Estos clientes destacan por tener pareja y/o estar casados, exactamente un 80.1 %, además de que esta mayoría de clientes conviven en el hogar con un hijo (49.5 %). Destaca también que la unidad familiar llega hasta el número de 5 personas (el 2.7 %). Los ingresos son medio-altos con respecto a los demás clústeres y destacan por tener unos gastos significativos en vino, carne, pescado y productos exclusivos considerados como "oro". Estos clientes aprovechan los descuentos, rondando entre 1 y 7 productos que han comprado en los dos últimos años. Donde más suelen comprar es a través de la página web y se observa que es muy frecuente para ellos visitar la página web de la empresa.

- **Clúster 2:** formado por 520 clientes que se encuentran mayormente en un rango de edad de 35 a 65 años, aproximadamente. Son en su mayoría clientes que terminaron un grado y decidieron no continuar sus estudios (un 54.2 %), les siguen aquellos que estudian un posgrado (22.7 %), el máster (el 15 %) y estudiantes de grado (el 8,1 %). Son en su mayoría clientes que tienen pareja y/o están casados (un 63.8 %). Aproximadamente, un poco más de la mitad de los clientes se caracteriza por convivir con un hijo u adolescente (el 50.6 %), siendo también considerable el número de personas que viven solas (el 31.3 %) y pocos clientes los que alcanzan convivir con dos hijos más en el hogar (17.9 %). Estos clientes se caracterizan por tener unos ingresos más altos con respecto a los demás clústeres, un valor que ronda

entre las 70,000 unidades monetarias y 85,000 unidades monetarias. Gastan dinero de manera considerable en todas las categorías de productos de la marca además de los productos de "oro". El mayor número de compras es en tienda y por catálogo, y las visitas a la página web son escasas.

- **Clúster 3:** formado por 606 clientes, siendo el clúster que concentra un número mayor de jóvenes de edades entre los 35 y aproximadamente 50 años. La mayoría son clientes que están graduados y decidieron no continuar con sus estudios (el 48.7 %) y el resto de clientes se reparten de forma equivalente en porcentajes de los distintos niveles de educación, por ejemplo, aquellos que estudian un posgrado son el 18.2 % y los que estudian un máster un 16.2 %. Debemos destacar que en este clúster es el que contiene al mayor número de clientes que están estudiando un grado y no han terminado. En cuanto al estado marital, no hay ninguno que esté soltero. Los clientes alcanzan un pico en la convivencia con un hijo en el hogar (el 58.3 %), seguido de dos hijos más (el 28.5 %) y alcanzando a cinco personas que conviven en un mismo hogar (el 3.1 %). Con respecto a los demás clústeres, tienen un nivel medio-bajo de ingresos, esto puede deberse a que la mayoría de los clientes en el clúster son jóvenes y están estudiando. En cuanto al gasto en los productos de la empresa, son el clúster con menos gastos. Las compras suelen hacerlas mayoritariamente en tienda y en catálogo. Además, son muchos los clientes que visitan la página web en el último mes.

- **Clúster 4:** formado por 240 clientes, con un rango de edad entre los 35 a 55 años. La mayoría son graduados (68.3 %), le siguen los que están estudiando un grado (un 13.8 %), un máster (9.6 %) y unos pocos el posgrado (el 8.3 %). En cuanto al estado marital, un poco más de la mitad tiene pareja y/o está casado. El número de personas en el hogar que conviven con el cliente no alcanza a las cinco personas, habiendo un pico en un hijo que convive con el cliente del 45 %. Tienen unos ingresos

altos que rondan entre las 60,000 y 80,000 unidades monetarias. Estos clientes suelen gastar bastante en las cinco principales categorías de productos de la empresa, así como en los productos considerados como "oro". Suelen comprar pocos productos en descuento y destacan por comprar a través de las tres vías de compra de la empresa. Además, suelen visitar moderadamente la página web de la empresa.

- **Clúster 5:** formado por 408 clientes con una edad que ronda entre los 35 y 55 años aproximadamente. Son clientes que en su mayoría están graduados (un 49 %), les siguen aquellos que estudiaron un posgrado (21.8 %), el máster y los que aún estudian un grado. El 99 % de estos clientes son solteros que tienen descendencia de al menos un hijo (el 56.6 %), dos hijos (28.5 %) y tres hijos (4.4 %). Los ingresos son medios con respecto a los demás clústeres. Se observa que donde suelen gastar más es en vino, pescado y productos considerados como "oro". La principal vía de compra es por tienda y aprovechan los productos que tienen descuento. También se caracterizan por frecuentar mucho la página web de la empresa.

Puede consultar los anexos recogidos en este documento para un detalle exhaustivo de los distintos clústeres, estos son el Anexo VI que incluye cuadros de resumen estadístico por clúster; el Anexo VII que incluye la visualización en diagramas de cajas y bigotes de las variables ingresos, vino, frutas, carne, pescado, dulces y oro por clúster; el Anexo VIII que incluye los diagramas de barras de las variables educación, unidad familiar y estado marital por clúster; el Anexo IX que incluye el diagrama de barras de compras con descuento, compras web, catálogo, tienda y número de visitas a la web de empresa en el último mes por clúster.

Finalmente, etiquetamos los clústeres de clientes en la manera que está descrita en la Tabla 5.4).

Tabla 5.4: *Etiquetas de los cinco grupos a partir de la aplicación de algoritmo de Ward en una matriz de distancias Euclídea.* Fuente: Elaboración propia.

Clústeres a través del método de Ward y distancia Euclídea
Clúster 1: "Familias acomodadas"
Clúster 2: "Profesionales de alto ingreso"
Clúster 3: "Jóvenes con familias"
Clúster 4: "Graduados con altos ingresos"
Clúster 5: "Solteros con hijos y consumo moderado"

5.4.4.2. grupos a través de la distancia de Manhattan y el método de Ward

A partir de la entrada de la matriz de distancias de Manhattan y el uso del algoritmo de Ward, obtuvimos el número de cuatro clústeres, los cuales vamos a detallar aquí:

- **Clúster 1:** formado por 598 clientes, donde la edad ronda entre los 40 años y 55 aproximadamente. La mayoría son graduados (el 48.7 %), le siguen aquellos que estudiaron un posgrado (un 25.1 %), un máster (el 19.1 %) y el restante estudia aún un grado. El 98.2 % tienen pareja y/o están casados y son el 66.7 % de los clientes los que tienen un hijo, le siguen aquellos que tienen dos hijos (el 20.6 %), viven sin hijos (el 10.4 %) y conforman una familia de 5 personas (el 2.3 %). El número de ingresos es medio-alto, entre las 50,000 unidades monetarias y 70,000 unidades monetarias aproximadamente. Este grupo se caracteriza por un gasto considerable en vino, frutas, pescado y en productos considerados como "oro". Son clientes que compran frecuentemente en la empresa, sobre todo cuando los productos tienen descuento. Destacan por comprar a través de la tienda y la página web. Además, suelen frecuentar la página web.

- **Clúster 2:** formado por 472 clientes, donde la edad ronda entre los 45 y 57 años. Un poco más de la mitad de los clientes son graduados (un 55.5 %), le siguen aquellos que estudian un posgrado (un 20.3 %), un máster (15.3 %) y el resto están todavía en un grado. El 57.8 % de los clientes tienen pareja y/o están casados y son pocos los que

tienen un hijo o más. Los ingresos de estos clientes son más altos que con respecto al resto de los clústeres (entre las 71,000 unidades monetarias y 85,000 unidades monetarias aproximadamente). Se caracterizan por gastar de forma considerable en los distintos productos de la empresa, siendo el vino y la carne los productos estrella. Además, también gastan en productos que se consideran como "oro". Suelen comprar por la página web y catálogo. El número de visitas por estos clientes en la página web no suele ser frecuente.

- **Clúster 3:** es el clúster con el mayor número de clientes, exactamente 930 clientes. La edad de estos clientes ronda entre los 35 y 50 años. Aproximadamente el 50 % son graduados, les siguen aquellos que estudiaron un posgrado (el 18.6 %) y es el mismo porcentaje de clientes que estudiaron un máster y aún estudian un grado. La mayoría son clientes que tienen pareja y/o están casados (el 61 %) y son clientes que mayormente tienen un hijo (el 46.3 %), viven con sus parejas (el 28.1 %) y tienen dos hijos (el 18.5 %). Los ingresos de estos clientes son los más bajos con respecto a los demás clústeres (entre las 25,000 unidades monetarias y 42,000 unidades monetarias). Son un grupo de clientes donde suelen gastar muy poco en los productos de la empresa, destacando el poco gasto en vino y carne, aunque no se puede ignorar el hecho de que lo poco que gastan es para productos que se consideran como "oro". Suelen comprar en la tienda y a través de la página web. Además, frecuentan la página web de la empresa.

- **Clúster 4:** formado por 212 clientes, donde la edad de estos ronda mayormente entre los 40 y 57 años de edad. El 99.5 % de los clientes son solteros y tienen un hijo (el 75 %). Son clientes que tienen un ingreso medio con respecto a los demás clústeres, además de que suelen gastar de manera considerable en vino, frutas, carne y pescado. También gastan en productos de calidad. Suelen comprar productos con descuento y el mayor número de productos se compran a través de la página web.

Además, suelen frecuentar la página web de la empresa.

Puede consultar los anexos recogidos en este documento para un detalle exhaustivo de los distintos clústeres, estos son el Anexo X que incluye cuadros de resumen estadístico por clúster; el Anexo XI que incluye la visualización en diagramas de cajas y bigotes de las variables ingresos, vino, frutas, carne, pescado, dulces y oro por clúster; el Anexo XII que incluye los diagramas de barras de las variables educación, unidad familiar y estado marital por clúster; el Anexo XIII que incluye el diagrama de barras de compras con descuento, compras web, catálogo, tienda y número de visitas a la web de empresa en el último mes por clúster.

Finalmente, etiquetamos los clústeres de clientes en la manera que está descrita en la Tabla 5.5).

Tabla 5.5: *Etiquetas de los cuatros grupos a partir de la aplicación de algoritmo de Ward en una matriz de distancias de Manhattan.* Fuente: Elaboración propia.

Clústeres a través del método de Ward y distancia de Manhattan
Clúster 1: "Familias con compras frecuentes"
Clúster 2: "Profesionales de alto poder adquisitivo"
Clúster 3: "Jóvenes con ingresos bajos"
Clúster 4: "Solteros con hijos y consumo moderado"

5.5. Estrategias de marketing sobre la base de clientes

A continuación presentamos algunas estrategias específicas para manejar a los clientes de iFood, segmentados en clústeres distintos obtenidos de haber utilizado dos medidas de proximidad, es decir, la distancia Euclídea y la distancia de Mnahattan. El objetivo es la búsqueda de mejorar la interacción y satisfacción del cliente, así como optimizar las estrategias de marketing y ventas. A continuación, se detallan las recomendaciones para cada

grupo de clientes según su perfil y preferencias siguiendo la información proporcionada por Patel y Snigdha (2024).

5.5.1. Estrategias de marketing según los clústeres obtenidos utilizando la distancia Euclídea

Recordemos que se obtuvieron cinco clústeres mediante el uso de la distancia Euclídea y método de Ward. Las etiquetas de cada clúster están en la Tabla 5.4.

- **Familias acomodadas:** estos clientes suelen aprovechar descuentos y compran frecuentemente en la página web. Para manejarlos, se recomienda a la empresa enfocarse en brindar ofertas especiales y descuentos personalizados además de asegurar un sitio web fácil de navegar y una experiencia de compra en línea fluida.

- **Profesionales de alto ingreso:** gustan de comprar en tiendas físicas y por catálogo, y no visitan frecuentemente la página web. Se recomienda ofrecer atención personalizada y exclusiva en tiendas y enviar catálogos detallados con productos premium y exclusivos.

- **Jóvenes con familias:** compran principalmente en tiendas y catálogos y visitan la página web regularmente. Proporciona una experiencia de compra ágil tanto en línea como en tiendas físicas. Se recomienda ofrecer promociones familiares y paquetes de productos.

- **Graduados con altos ingresos:** prefieren comprar a través de múltiples canales y visitan la página web moderadamente. Se recomienda brindar una experiencia de compra coherente en todos los puntos de contacto y asegurar que las promociones sean accesibles en todas las plataformas.

- **Solteros con hijos y consumo moderado:** frecuentan la página web y aprovechan descuentos. Se recomienda utilizar marketing digital para enviar ofertas y descuentos

personalizados a través de correo electrónico y redes sociales además de asegurar una plataforma web optimizada para facilitar las compras.

5.5.2. Estrategias de marketing según los clústeres obtenidos utilizando la distancia de Manhattan

Recordemos que se obtuvieron cuatro clústeres mediante el uso de la distancia de Manhattan y método de Ward. Las etiquetas de cada clúster están en la Tabla 5.5.

- **Familias con compras frecuentes:** estos clientes compran frecuentemente en la tienda y la página web, aprovechando descuentos. Para manejarlos, se recomienda enfocarse en ofrecer promociones especiales y descuentos personalizados y asegurarse de mantener una experiencia de compra en línea y en tienda fluida y satisfactoria.

- **Profesionales de alto poder adquisitivo:** prefieren comprar por la página web y catálogo, pero no frecuentan la página web. Se recomienda ofrecer atención personalizada y exclusiva a través de catálogos detallados con productos premium además de proporcionar una experiencia de compra de alta calidad tanto en línea como en tienda.

- **Jóvenes con ingresos bajos:** estos clientes suelen gastar poco y prefieren compras en tienda y a través de la página web. Se recomienda ofrecer promociones y descuentos accesibles para productos populares además de asegurarse de que la experiencia de compra en línea y en tienda sea ágil y conveniente.

- **Solteros con hijos y consumo moderado:** frecuentan la página web y aprovechan descuentos. Se recomienda utilizar marketing digital para enviar ofertas y promociones personalizadas por correo electrónico y redes sociales. También, se requiere asegurarse de que la plataforma web esté optimizada para una experiencia de compra fácil y rápida.

Capítulo 6

Conclusiones finales

En este Trabajo Final de Grado, hemos analizado las técnicas de segmentación del aprendizaje no supervisado en un caso concreto de segmentación del cliente basándonos en el ciclo de vida de un proyecto de ciencia de datos y utilizando la herramienta de Python para la adecuada implementación de la metodología del estudio. Las conclusiones de este trabajo son las siguientes.

En el contexto actual, la IA ha emergido como una tecnología disruptiva que está transformando múltiples sectores de la ciencia y la industria. Su capacidad para analizar grandes volúmenes de datos, aprender patrones y tomar decisiones basadas en algoritmos sofisticados ha optimizado procesos en una variedad de campos. Además, tanto la ciencia de datos, el aprendizaje estadístico y el aprendizaje automático se han convertido en una disciplina crítica en este mismo contexto.

Vimos como, en el campo de aprendizaje automático, el clustering representa un tipo de aprendizaje no supervisado, lo cual implica que no se basa en clases predefinidas ni en datos de entrenamiento etiquetados. A diferencia del aprendizaje supervisado, que aprende a partir de ejemplos específicos, el clustering aprende a partir de la observación directa de los datos. Este punto venía relacionado con las medidas de proximidad ya que diversas métricas de similitud o disimilitud derivadas del mismo conjunto de individuos pueden llevar, y frecuentemente lo hacen, a diferentes resultados cuando se aplican a problemas de clustering.

Centrándonos en el caso de estudio de segmentación del cliente, la elección de las

variables y la metodología utilizada fueron fundamentales para el éxito del análisis. La recolección de datos, el análisis exploratorio y el preprocesamiento de los mismos permitieron una clasificación precisa y útil de los clientes para la empresa. La elección del algoritmo de Ward permitió identificar los distintos grupos de clientes. Destacar que el uso de Python y bibliotecas específicas para el análisis de los datos resultó ser una herramienta importante para la implementación del trabajo.

Los resultados obtenidos proporcionan a iFood información valiosa sobre sus clientes, permitiendo una mejor comprensión de sus preferencias y comportamientos de compra. Esta información puede ser utilizada para optimizar las campañas de marketing y mejorar la experiencia del cliente, lo que podría traducirse en un incremento en las ventas y la lealtad del cliente. Se recomienda la implementación de las estrategias de marketing propuestas y la continua evaluación de su efectividad. Además, es aconsejable realizar estudios periódicos de segmentación de clientes para adaptarse a los cambios en el comportamiento del consumidor y las tendencias del mercado.

El principal desafío de este trabajo fue la elección de las técnicas clustering ya que, cuando se trata de estas técnicas, entonces no se puede recomendar un método de agrupación específico, ya que aquellos con propiedades matemáticas favorables a menudo no producen resultados empíricamente interpretables. Además, usar los resultados implica elegir una partición adecuada, lo cual no siempre es claro cómo hacerlo mejor. Otro inconveniente era la decisión de las medidas de proximidad. Sería de gran ayuda conocer cuáles de las métricas son las "óptimas" en el sentido de crear grupos dentro de un conjunto de datos. Sin embargo, no existe una respuesta definitiva y la elección de la métrica adecuada depende en gran medida del tipo de variables utilizadas y del juicio propio.

Bibliografía

[1] *Acerca de GitHub y Git - documentación de GitHub.* (s. f.). GitHub Docs. Última consulta 13 de junio de 2024.

[2] Anaconda. (2024). *Anaconda | The Operating System for AI.* Última consulta 29 de junio 2024.

[3] konstan. (s. f.). *Anaconda - Python Logo.* CLEANPNG. https://www.cleanpng.com/png-anaconda-pip-installation-python-6740669/

[4] Bryson, J. J. (s.f.). *La última década y el futuro del impacto de la IA en la sociedad — OpenMind.* OpenMind. última consulta 3 de julio 2024.

[5] *Customer Personality Analysis.* (2021). Kaggle. https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis/data

[6] datos.gob.es. (2024). Acerca de la iniciativa Aporta. Última consulta 28 de junio 2024.

[7] Demey, J. R., Pla, L., Vicente-Villardón, J. L., Di Rienzo, J. A., & Casanoves, F. (2011). *Valoración y análisis de la diversidad funcional y su relación con los servicios ecosistémicos*, CATIE, 47-54.

[8] devsujay19. (s. f.). *Visual Studio Logo.* GitHub. https://github.com/topics/visual-studio-logo

[9] Everitt, B. S., Landau S., Leese M. & Stahl D. (2011). *Cluster Analysis*. (5.ªed.). John Wiley & Sons, Ltd., 16-46-47-48-49-50-67.

[10] Fabián Serrano, M. (2019). El liderazgo militar en los tiempos de la inteligencia artificial. Visión Conjunta, *11*(20), 26. Última consulta 3 de julio 2024.

[11] Ferreira, R. (2022). *Ciencia de Datos Teoría y Aplicaciones*. Tecnológico Nacional de México, 22-23-71.

[12] *Funciones.* (s. f.). CIS. Última consulta 28 de junio de 2024.

[13] Gallardo, J. (s. f.). *Introducción al Análisis Clúster*. Universidad de Granada.

[14] Gower, J. C. & Legendre, P. (1986). *Metric and Euclidean properties of dissimilarity coefficients*. Journal of Classification, 23.

[15] Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. (2.ªed.). Springer Series in Statistics, Springer, 503-520.

[16] iFood. (2021). iFood: Relato de Impacto. https://institucional.ifood.com.br/wp-content/uploads/2024/01/Relato-de-Impacto-iFood-2021-4.pdf

[17] INE. (s. f.). *El INE y sus principales funciones.* Última consulta 28 de junio 2024.

[18] James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An Introduction to Statistical Learning: With Applications in Python*. (1.ªed.). Springer, 503-521-528-529.

[19] Janssen, P., Walther, C., & Luedeke, M. (2012). *Cluster Analysis to Understand Socio-ecological Systems*. PIK Report, 9-10-11-14.

[20] Kroese, D. P., Botev, Z. I., Taimre, T., & Vaisman, R. (2023). *Data Science and Machine Learning: Mathematical and Statistical Methods*. Chapman and Hall/CRC, Boca Raton, xiii-xiv-3-7-8-9-10-11-19-147-148-149-151.

[21] *Machine learning en marketing — Mailchimp*. (s.f.). Mailchimp. Última consulta 4 de julio 2024.

[22] *Matplotlib - Visualization with Python*. (s. f.). Última consulta 13 de junio de 2024.

[23] McKinney, W. (2017). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. (2.ªed.). O'Reilly Media Inc., 4.

[24] Mewara, S. (2020). *Data Visualization - Insights with Matplotlib*. Learn By Insight. https://learnbyinsight.com/2020/09/13/data-visualization-insights-with-matplotlib/

[25] Müller, A. C., & Guido, S. (2016). *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly Media Inc., 1.

[26] Muñoz, L. A. S. (2021). *Curso intensivo de Python NumPy: Como construir arreglos n-dimensionales para aprendizaje automático*. freeCodeCamp.org. https://www.freecodecamp.org/espanol/news/curso-intensivo-de-python-numpy-como-construir-arreglos-n-dimensionales-para-aprendizaje-automatico/

[27] Murtagh, F., & Legendre, P. (2014). *Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?*. Montfort University & Université de Montréal, 274-275.

[28] Mwaskom. (s. f.). *Discussion of Seaborn Logo · Issue #2243 · mwaskom/seaborn*. GitHub. https://github.com/mwaskom/seaborn/issues/2243

[29] Sneath, P. H. A. & Sokal, R. R. (1973). N*umerical Taxonomy*. W. H. Freeman, San Francisco.

[30] Nailson. (2020). GitHub. ifood-data-business-analyst-test/iFood Data Analyst Case.pdf.

[31] *NumPy documentation - NumPy v2.0 Manual.* (s. f.). Última consulta 13 de junio de 2024.

[32] Gutierrez Pachas, D. A. (2024). *Ciencia de Datos y Estadística: Saber "Lo que dicen los datos" a partir de sus tratamiento.* Universidad Católica San Pablo. Última consulta 3 julio de 2024.

[33] *pandas - Python Data Analysis Library.* (s. f.). Última consulta 13 de junio de 2024.

[34] *File: Pandas logo.svg.* (2019). Wikimedia. https://en.wikipedia.org/wiki/File:Pandas_logo.svg

[35] Patel, S., & Snigdha. (2024). *25 Types of Customers and How to Deal With Them.* REVE Chat. https://www.revechat.com/blog/types-of-customers/

[36] Pedret R., Sagnier, L. & Camp, F. (2003). *Herramientas para segmentar mercados y posicionar productos.* Barcelona, España, Planeta.

[37] *File: Python logo and wordmark.svg* (2008). Wikimedia Commons. https://commons.wikimedia.org/wiki/File:Python_logo_and_wordmark.svg

[38] *Welcome to Python.org.* (2024). Python.org. Última consulta 29 de junio 2024.

[39] Real Academia Española. (2005). *Clúster.* En *Diccionario panhispánico de dudas (DPD)* (2.ª ed.). Última consulta 5 junio 2024.

[40] Robinson, E., & Nolis, J. (2020). *Build a career in data science.* Manning Publications Co., 5-57.

[41] Sakarkar, G., Patil G., & Dutta, P. (2021). *Internet of Things and Machine Learning: Machine Learning Algorithms Using Python Programming*, Nova Science Publishers Inc., 39-43-44-46.

[42] *SciPy* -. (s. f.). Última consulta 13 de junio de 2024.

[43] *scikit-learn: machine learning in Python - scikit-learn 1.5.0 documentation.* (s. f.). Última consulta 13 de junio de 2024.

[44] *Scikit learn Logo PNG Vector.* (s. f.). seeklogo. https://seeklogo.com/vector-logo/337685/scikit-learn

[45] *seaborn: statistical data visualization - seaborn 0.13.2 documentation.* (s. f.). Última consulta 13 de junio de 2024.

[46] *Stack Overflow en español.* (s. f.). Última consulta 13 de junio de 2024.

[47] Tan, P.-N., Steinbach, M., & Kumar, V. (2016). *Introduction to Data Mining.* Pearson, 490-502.

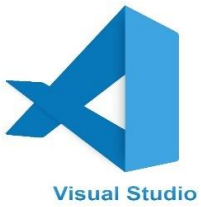
[48] *The most-comprehensive AI-powered DevSecOps platform.* (s. f.). GitLab. Última consulta 13 de junio de 2024.

[49] *Tutorial de SciPy — Interactive Chaos.* (s. f.). https://interactivechaos.com/es/manual/tutorial-de-scipy/tutorial-de-scipy

Capítulo 7

Anexo

ANEXO I: TECNOLOGÍAS UTILIZADAS [1]



[1] Fuente: Elaboración propia.

ANEXO II: VARIABLES DEL CONJUNTO DE DATOS [2]

Tipo de Variable		Variable	Descripción
Continua	Numérica	Edad[3]	Edad del cliente
		Ingresos	Ingresos anuales del hogar del cliente
		Vino	Cantidad gastada en vino en los últimos 2 años
		Frutas	Cantidad gastada en fruta en los últimos 2 años
		Carne	Cantidad gastada en carne en los últimos 2 años
		Pescado	Cantidad gastada en pescado en los últimos 2 años
		Dulces	Cantidad gastada en dulces en los últimos 2 años
		Oro	Cantidad gastada en oro en los últimos 2 años
	Discreta	Niños_Hogar	Número de niños en el hogar del cliente
		Adolescentes_Hogar	Número de adolescentes en el hogar del cliente
		Dt_Customer	Fecha de alta del cliente en la empresa
		Recency	Número de días desde la última compra del cliente
		NumDesCompras	Número de compras realizadas con descuento
		NumWebCompras	Número de compras realizadas a través del sitio web
		NumCatalogCompras	Número de compras realizadas mediante un catálogo
		NumTiendaCompras	Número de compras realizadas directamente en tiendas
		NumWebVisitasMes	Número de visitas al sitio web de la empresa en el último mes
Cualitativa	Ordinal	Educación	Nivel de educación del cliente. *Niveles:* Estudia grado, graduado, máster y posgrado
		Queja	El cliente se quejó en los últimos 2 años *Niveles:* Sí y No
		AcceptedCmp1	El cliente aceptó la oferta en la primera campaña *Niveles:* Sí y No
		AcceptedCmp2	El cliente aceptó la oferta en la segunda campaña *Niveles:* Sí y No
		AcceptedCmp3	El cliente aceptó la oferta en la tercera campaña *Niveles:* Sí y No
		AcceptedCmp4	El cliente aceptó la oferta en la cuarta campaña *Niveles:* Sí y No
		AcceptedCmp5	El cliente aceptó la oferta en la quinta campaña *Niveles:* Sí y No
		Response	El cliente aceptó la oferta en la última campaña Niveles: Sí y No
	Nominal	ID	Identificador único del cliente
		Estado_Marital	Estado civil del cliente *Niveles:* Casado, tiene pareja, divorciado, viudo y otros que no incluye pareja

[3] *Customer Personality Analysis*. (2021). Kaggle. https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis/data

ANEXO III: CONTEO DE VALORES FALTANTES POR VARIABLE [4]

	ID	Educacion	Marital_Status	Niños_Hogar	Adolescentes_Hogar	Recency	Edad	Ingresos	Vino	Frutas	Carne
contador	0	0	0	0	0	0	0	24	0	0	0

	Pescado	Dulces	NumDesCompras	NumWebCompras	NumCatalogCompras	NumTiendaCompras	NumWebVisitaMes
contador	0	0	0	0	0	0	0

	AcceptedCmp1	AcceptedCmp2	AcceptedCmp3	AcceptedCmp4	AcceptedCmp5	AcceptedLastCmp	Queja
contador	0	0	0	0	0	0	0

[4] Fuente: Elaboración propia.

ANEXO IV: ANÁLISIS DESCRITIVO DE LAS VARIABLES DEL PROBLEMA[5]

	Educacion	Marital_Status	Niños_Hogar	Adolescentes_Hogar	Recency	Edad	Ingresos	Vino	Frutas	Carne
contador	2216.0	2216.0	2216.0	2216.0	2216.0	2216.0	2216.0	2216.0	2216.0	2216.0
media	1.48	1.65	0.44	0.51	49.01	45.18	52247.25	305.09	26.36	167.0
std	0.96	0.48	0.54	0.54	28.95	11.99	25173.08	337.33	39.79	224.28
min	0.0	1.0	0.0	0.0	0.0	18.0	1730.0	0.0	0.0	0.0
25%	1.0	1.0	0.0	0.0	24.0	37.0	35303.0	24.0	2.0	16.0
50%	1.0	2.0	0.0	0.0	49.0	44.0	51381.5	174.5	8.0	68.0
75%	2.0	2.0	1.0	1.0	74.0	55.0	68522.0	505.0	33.0	232.25
max	3.0	2.0	2.0	2.0	99.0	121.0	666666.0	1493.0	199.0	1725.0

	Pescado	Dulces	NumDesCompras	NumWebCompras	NumCatalogCompras	NumTiendaCompras	NumWebVisitaMes
contador	2216.0	2216.0	2216.0	2216.0	2216.0	2216.0	2216.0
media	37.64	27.03	2.32	4.09	2.67	5.8	5.32
std	54.75	41.07	1.92	2.74	2.93	3.25	2.43
min	0.0	0.0	0.0	0.0	0.0	0.0	0.0
25%	3.0	1.0	1.0	2.0	0.0	3.0	3.0
50%	12.0	8.0	2.0	4.0	2.0	5.0	6.0
75%	50.0	33.0	3.0	6.0	4.0	8.0	7.0
max	259.0	262.0	15.0	27.0	28.0	13.0	20.0

	AcceptedCmp1	AcceptedCmp2	AcceptedCmp3	AcceptedCmp4	AcceptedCmp5	AcceptedLastCmp	Queja
contador	2216.0	2216.0	2216.0	2216.0	2216.0	2216.0	2216.0
media	0.06	0.01	0.07	0.07	0.07	0.15	0.01
std	0.24	0.12	0.26	0.26	0.26	0.36	0.1
min	0.0	0.0	0.0	0.0	0.0	0.0	0.0
25%	0.0	0.0	0.0	0.0	0.0	0.0	0.0
50%	0.0	0.0	0.0	0.0	0.0	0.0	0.0
75%	0.0	0.0	0.0	0.0	0.0	0.0	0.0
max	1.0	1.0	1.0	1.0	1.0	1.0	1.0

[5] Fuente: Elaboración propia.

ANEXO V: VARIABLES FINALES [6]

Tipo de Variable		Variable	Descripción
Continua	Numérica	Edad[7]	Edad del cliente
		Ingresos	Ingresos anuales del hogar del cliente
		Vino	Cantidad gastada en vino en los últimos 2 años
		Frutas	Cantidad gastada en fruta en los últimos 2 años
		Carne	Cantidad gastada en carne en los últimos 2 años
		Pescado	Cantidad gastada en pescado en los últimos 2 años
		Dulces	Cantidad gastada en dulces en los últimos 2 años
		Oro	Cantidad gastada en oro en los últimos 2 años
	Discreta	NumDesCompras	Número de compras realizadas con descuento
		NumWebCompras	Número de compras realizadas a través del sitio web
		NumCatalogCompras	Número de compras realizadas mediante un catálogo
		NumTiendaCompras	Número de compras realizadas directamente en tiendas
		NumWebVisitasMes	Número de visitas al sitio web de la empresa en el último mes
		Unidad_Familiar	Número total de personas que viven en el mismo hogar
Cualitativa	Ordinal	Educación	Nivel de educación del cliente. *Niveles:* Estudia grado, graduado, máster y posgrado
	Nominal	Estado_Marital	Estado civil del cliente *Niveles:* Casado, tiene pareja, divorciado, viudo y otros que no incluye pareja

[6] Fuente: Elaboración propia.

[7] *Customer Personality Analysis*. (2021). Kaggle. https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis/data

ANEXO VI

CLÚSTER 1- A TRAVÉS DE LA DISTANCIA EUCLÍDEA Y MÉTODO DE WARD [8]

	Educacion	Marital_Status	Ingresos	Vino	Frutas	Carne	Pescado	Dulces	Oro	NumDesCompras	NumWebCompras
contador	438.0	438.0	438.0	438.0	438.0	438.0	438.0	438.0	438.0	438.0	438.0
media	1.83	1.8	57499.97	456.03	16.01	130.41	22.37	14.63	48.23	3.97	5.65
std	0.95	0.4	15124.56	325.79	20.05	118.6	27.82	19.03	43.18	2.63	2.43
min	0.0	1.0	24401.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
25%	1.0	2.0	49182.75	199.25	3.0	60.0	4.0	1.0	15.0	2.0	4.0
50%	2.0	2.0	57008.5	372.0	9.0	96.5	13.0	9.0	35.0	3.0	6.0
75%	3.0	2.0	65042.25	640.5	21.0	155.75	32.0	21.0	66.0	6.0	7.0
max	3.0	2.0	162397.0	1486.0	159.0	860.0	223.0	150.0	215.0	15.0	11.0

	NumCatalogCompras	NumTiendaCompras	NumWebVisitaMes	Edad	Unidad_Familiar
contador	438.0	438.0	438.0	438.0	438.0
media	2.73	7.01	5.85	48.63	2.96
std	1.71	2.48	2.08	10.4	0.81
min	0.0	0.0	0.0	20.0	1.0
25%	1.0	5.0	5.0	41.0	2.0
50%	2.0	7.0	6.0	48.0	3.0
75%	4.0	9.0	8.0	57.75	3.0
max	10.0	13.0	10.0	71.0	5.0

[8] Fuente: Elaboración propia.

ANEXO VI

CLÚSTER 2- A TRAVÉS DE LA DISTANCIA EUCLÍDEA Y MÉTODO DE WARD [9]

	Educacion	Marital_Status	Ingresos	Vino	Frutas	Carne	Pescado	Dulces	Oro	NumDesCompras	NumWebCompras
contador	520.0	520.0	520.0	520.0	520.0	520.0	520.0	520.0	520.0	520.0	520.0
media	1.52	1.64	76103.14	626.4	67.4	464.97	99.04	64.5	75.23	1.38	5.2
std	0.93	0.48	11986.42	315.99	51.22	254.12	65.93	50.14	61.37	1.37	2.15
min	0.0	1.0	2447.0	1.0	0.0	29.0	0.0	0.0	0.0	0.0	0.0
25%	1.0	1.0	69652.5	387.0	26.0	268.25	43.75	26.0	30.0	1.0	4.0
50%	1.0	2.0	75912.5	583.5	52.0	427.0	89.0	50.0	54.0	1.0	5.0
75%	2.0	2.0	82332.25	833.5	102.0	614.75	149.0	94.25	108.0	1.0	7.0
max	3.0	2.0	160803.0	1493.0	197.0	1725.0	259.0	198.0	249.0	15.0	11.0

	NumCatalogCompras	NumTiendaCompras	NumWebVisitaMes	Edad	Unidad_Familiar
contador	520.0	520.0	520.0	520.0	520.0
media	6.34	8.48	2.8	46.35	1.87
std	2.93	2.89	1.75	13.42	0.7
min	1.0	0.0	0.0	19.0	1.0
25%	4.0	6.0	1.0	36.0	1.0
50%	6.0	8.5	2.0	46.0	2.0
75%	8.0	11.0	4.0	58.0	2.0
max	28.0	13.0	9.0	73.0	4.0

[9] Fuente: Elaboración propia.

ANEXO VI

CLÚSTER 3- A TRAVÉS DE LA DISTANCIA EUCLÍDEA Y MÉTODO DE WARD [10]

	Educacion	Marital_Status	Ingresos	Vino	Frutas	Carne	Pescado	Dulces	Oro	NumDesCompras	NumWebCompras
contador	606.0	606.0	606.0	606.0	606.0	606.0	606.0	606.0	606.0	606.0	606.0
media	1.35	2.0	33603.29	35.9	3.91	18.91	5.89	4.25	13.08	1.89	1.96
std	0.97	0.0	11477.33	47.77	5.72	18.08	7.37	5.88	13.54	1.04	1.08
min	0.0	2.0	7500.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0
25%	1.0	2.0	24945.5	8.0	0.0	7.0	0.0	0.0	4.0	1.0	1.0
50%	1.0	2.0	33566.5	19.0	2.0	13.0	3.0	2.0	9.0	2.0	2.0
75%	2.0	2.0	41407.75	41.75	5.0	23.0	8.0	6.0	18.0	2.0	3.0
max	3.0	2.0	64413.0	309.0	47.0	137.0	43.0	62.0	108.0	7.0	6.0

	NumCatalogCompras	NumTiendaCompras	NumWebVisitaMes	Edad	Unidad_Familiar
contador	606.0	606.0	606.0	606.0	606.0
media	0.45	3.16	6.44	42.27	3.25
std	0.63	0.92	1.6	10.67	0.67
min	0.0	2.0	1.0	18.0	2.0
25%	0.0	3.0	5.0	35.0	3.0
50%	0.0	3.0	7.0	41.0	3.0
75%	1.0	4.0	8.0	49.0	4.0
max	4.0	8.0	10.0	68.0	5.0

[10] Fuente: Elaboración propia

ANEXO VI

CLÚSTER 4- A TRAVÉS DE LA DISTANCIA EUCLÍDEA Y MÉTODO DE WARD [11]

	Educacion	Marital_Status	Ingresos	Vino	Frutas	Carne	Pescado	Dulces	Oro	NumDesCompras	NumWebCompras
contador	240.0	240.0	240.0	240.0	240.0	240.0	240.0	240.0	240.0	240.0	240.0
media	1.12	1.56	62706.94	413.16	47.85	195.41	64.1	62.9	89.88	2.6	6.84
std	0.74	0.5	13445.37	248.41	40.45	109.99	56.19	50.54	62.99	1.56	3.07
min	0.0	1.0	4428.0	6.0	0.0	3.0	0.0	0.0	2.0	0.0	2.0
25%	1.0	1.0	55414.75	224.0	16.0	108.5	20.0	23.75	37.75	1.75	5.0
50%	1.0	2.0	62856.0	380.0	36.0	170.5	54.0	48.0	75.5	2.0	6.0
75%	1.0	2.0	70854.5	561.5	71.0	259.0	90.75	97.0	131.5	4.0	8.0
max	3.0	2.0	113734.0	1324.0	199.0	536.0	237.0	262.0	321.0	9.0	27.0

	NumCatalogCompras	NumTiendaCompras	NumWebVisitaMes	Edad	Unidad_Familiar
contador	240.0	240.0	240.0	240.0	240.0
media	3.58	8.41	4.84	44.86	2.46
std	1.98	3.1	1.84	10.59	0.79
min	0.0	0.0	0.0	22.0	1.0
25%	2.0	6.0	4.0	36.0	2.0
50%	3.0	8.0	5.0	44.0	3.0
75%	5.0	11.0	6.0	54.0	3.0
max	10.0	13.0	9.0	69.0	4.0

[11] Fuente: Elaboración propia.

ANEXO VI

CLÚSTER 5- A TRAVÉS DE LA DISTANCIA EUCLÍDEA Y MÉTODO DE WARD [12]

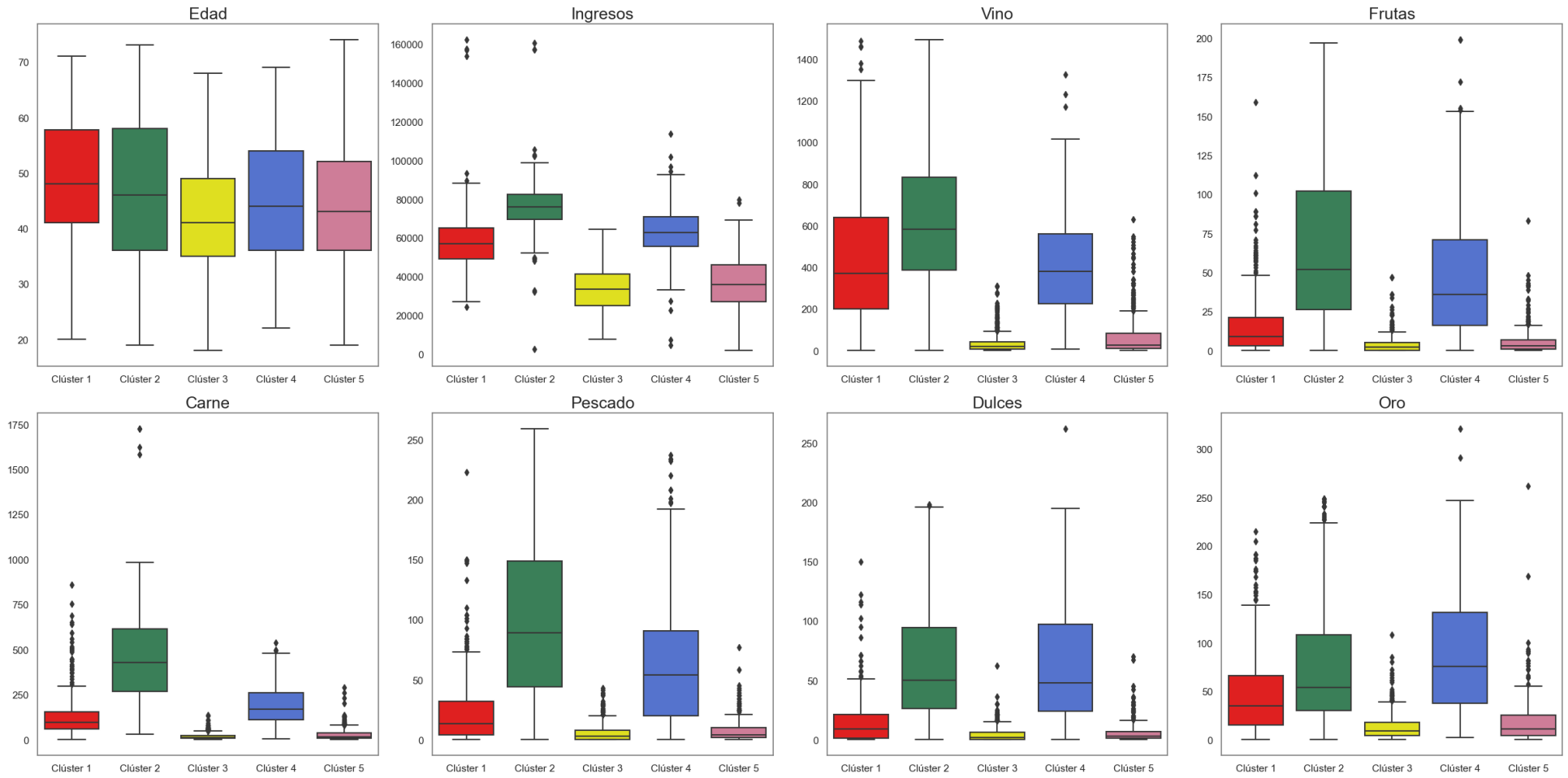
	Educacion	Marital_Status	Ingresos	Vino	Frutas	Carne	Pescado	Dulces	Oro	NumDesCompras	NumWebCompras
contador	408.0	408.0	408.0	408.0	408.0	408.0	408.0	408.0	408.0	408.0	408.0
media	1.47	1.01	36178.94	70.87	5.7	29.92	7.41	5.42	18.2	2.25	2.55
std	0.97	0.1	13765.19	105.03	8.74	36.21	9.5	8.23	22.82	1.68	1.89
min	0.0	1.0	1730.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
25%	1.0	1.0	26874.75	9.0	1.0	8.0	1.75	1.0	4.0	1.0	1.0
50%	1.0	1.0	35919.5	26.0	3.0	17.0	4.0	3.0	11.0	2.0	2.0
75%	2.0	1.0	46089.0	81.25	7.0	37.0	10.0	7.0	25.0	3.0	3.0
max	3.0	2.0	79761.0	631.0	83.0	288.0	77.0	70.0	262.0	15.0	10.0

	NumCatalogCompras	NumTiendaCompras	NumWebVisitaMes	Edad	Unidad_Familiar
contador	408.0	408.0	408.0	408.0	408.0
media	0.71	3.51	6.59	43.99	2.23
std	0.85	1.68	2.37	11.6	0.72
min	0.0	0.0	1.0	19.0	1.0
25%	0.0	3.0	5.0	36.0	2.0
50%	1.0	3.0	7.0	43.0	2.0
75%	1.0	4.0	8.0	52.0	3.0
max	5.0	11.0	20.0	74.0	4.0

[12] Fuente: Elaboración propia.

ANEXO VII: CLÚSTERES A TRAVÉS DE LA DISTANCIA EUCLÍDEA Y MÉTODO DE WARD

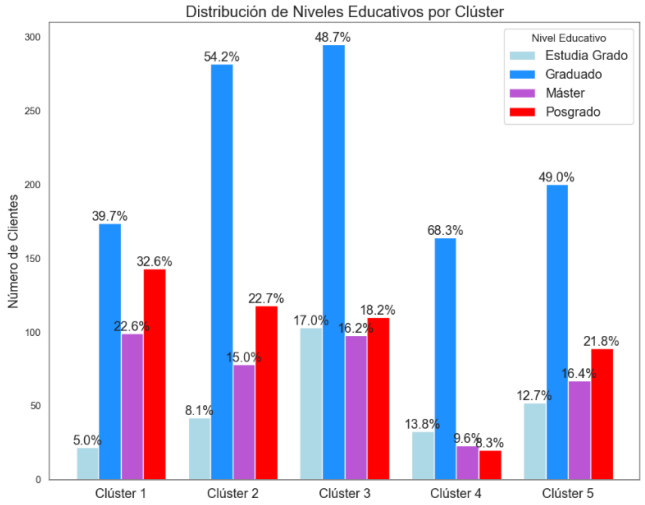
(*INGRESOS, VINO, FRUTAS, CARNE, PESCADO, DULCES* Y *ORO*) [13]

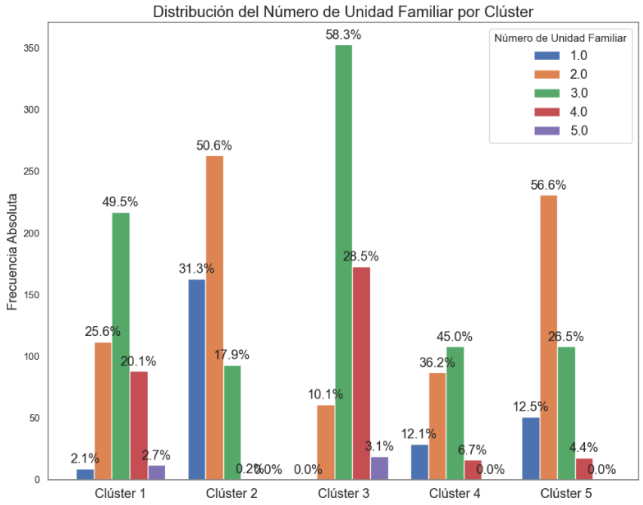


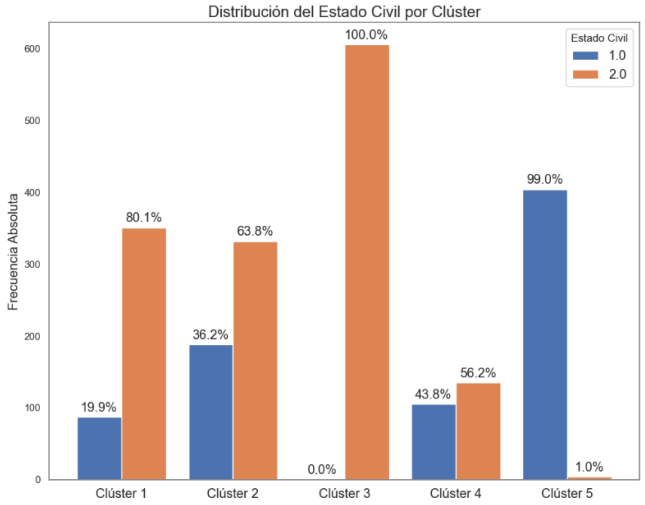
[13] Fuente: Elaboración propia.

ANEXO VIII: CLÚSTERES A TRAVÉS DE LA DISTANCIA EUCLÍDEA Y MÉTODO DE WARD

(*EDUCACIÓN, UNIDAD FAMILIAR Y ESTADO MARITAL*) [14]



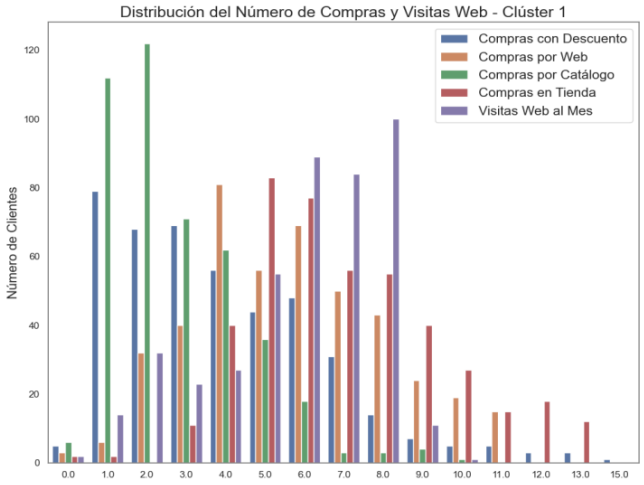


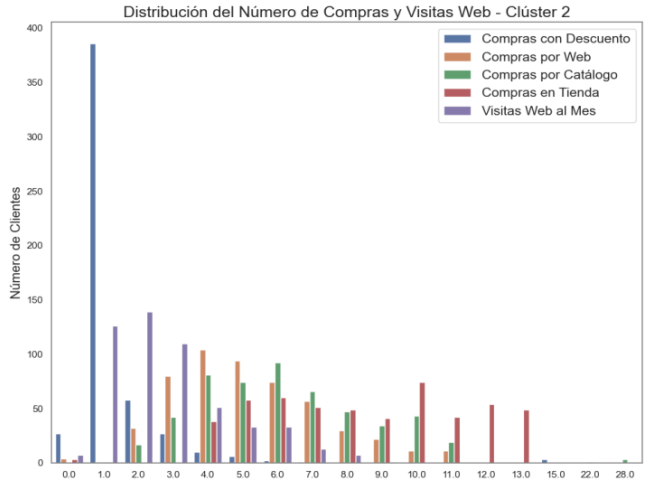


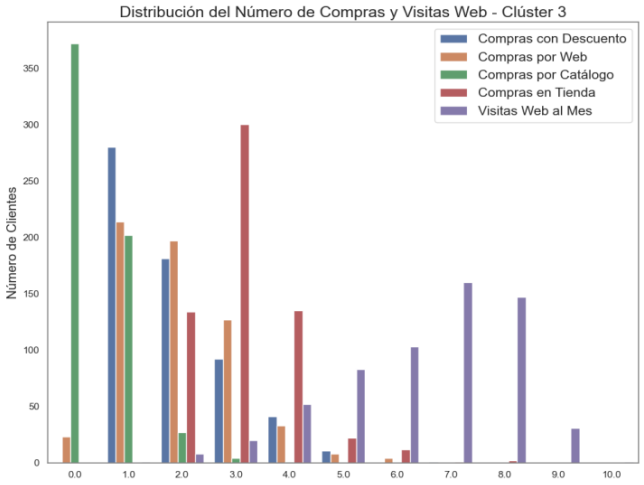
[14] Fuente: Elaboración propia.

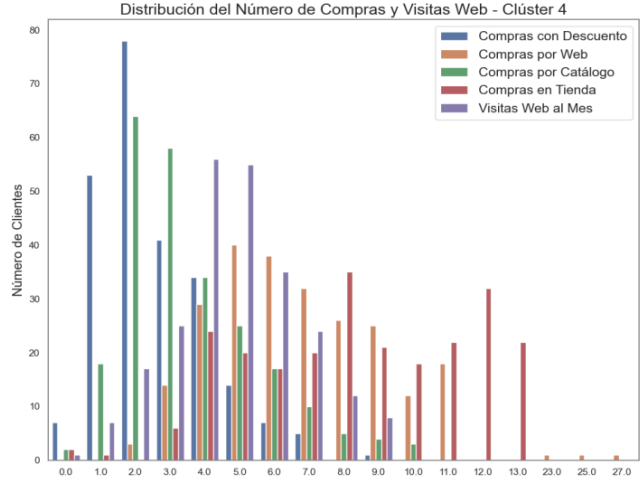
ANEXO IX: CLÚSTERES A TRAVÉS DE LA DISTANCIA EUCLÍDEA Y MÉTODO DE WARD

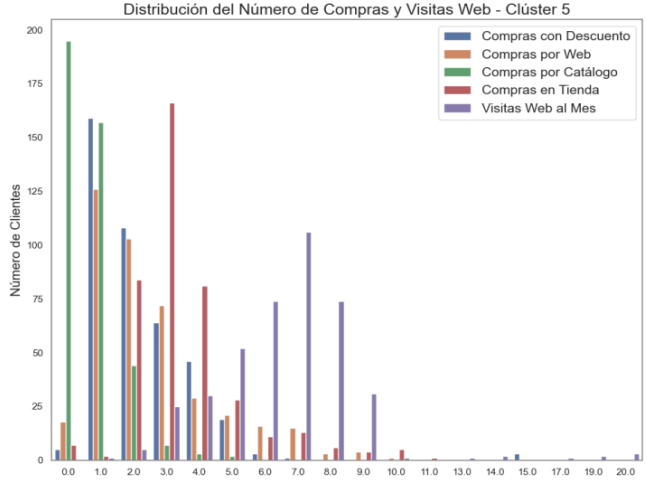
(*COMPRAS POR DECUENTO, WEB, CATÁLOGO, TIENDA* Y *NÚMERO DE VISTAS POR MES*) [15]











[15] Fuente: Elaboración propia.

ANEXO X

CLÚSTER 1- A TRAVÉS DE LA DISTANCIA DE MANHATTAN Y MÉTODO DE WARD [16]

	Educacion	Marital_Status	Ingresos	Vino	Frutas	Carne	Pescado	Dulces	Oro	NumDesCompras	NumWebCompras
contador	598.0	598.0	598.0	598.0	598.0	598.0	598.0	598.0	598.0	598.0	598.0
media	1.62	1.98	57783.6	419.78	27.02	141.39	35.5	28.88	56.64	3.5	5.87
std	0.94	0.13	12688.06	308.03	34.64	112.31	44.14	38.11	51.68	2.25	2.41
min	0.0	1.0	16927.0	7.0	0.0	6.0	0.0	0.0	0.0	0.0	1.0
25%	1.0	2.0	49671.25	176.25	4.0	61.0	6.0	4.0	17.0	2.0	4.0
50%	1.0	2.0	57991.0	349.0	13.0	107.5	19.0	13.5	39.0	3.0	6.0
75%	2.75	2.0	66310.5	575.75	35.0	182.75	50.0	35.0	80.0	5.0	8.0
max	3.0	2.0	93404.0	1492.0	193.0	650.0	237.0	195.0	248.0	15.0	11.0

	NumCatalogCompras	NumTiendaCompras	NumWebVisitaMes	Edad	Unidad_Familiar
contador	598.0	598.0	598.0	598.0	598.0
media	3.0	7.34	5.55	47.65	3.15
std	2.1	2.78	1.93	10.47	0.62
min	0.0	2.0	1.0	22.0	2.0
25%	1.0	5.0	4.0	39.0	3.0
50%	2.0	7.0	6.0	47.0	3.0
75%	4.0	9.0	7.0	57.0	3.0
max	11.0	13.0	10.0	71.0	5.0

[16] Fuente: Elaboración propia.

ANEXO X

CLÚSTER 2- A TRAVÉS DE LA DISTANCIA DE MANHATTAN Y MÉTODO DE WARD [17]

	Educacion	Marital_Status	Ingresos	Vino	Frutas	Carne	Pescado	Dulces	Oro	NumDesCompras	NumWebCompras
contador	472.0	472.0	472.0	472.0	472.0	472.0	472.0	472.0	472.0	472.0	472.0
media	1.47	1.58	77205.79	620.29	66.1	482.07	101.36	68.36	76.71	1.05	4.87
std	0.91	0.49	9139.08	306.94	50.59	228.03	66.2	51.36	61.39	0.48	1.86
min	0.0	1.0	48192.0	68.0	0.0	74.0	0.0	0.0	0.0	0.0	1.0
25%	1.0	1.0	70944.25	379.5	25.0	311.0	43.75	27.75	30.75	1.0	3.75
50%	1.0	2.0	77488.5	564.0	51.0	447.0	92.0	54.0	54.0	1.0	5.0
75%	2.0	2.0	82577.5	833.5	100.0	653.25	150.0	102.25	111.25	1.0	6.0
max	3.0	2.0	105471.0	1493.0	197.0	984.0	258.0	198.0	249.0	4.0	11.0

	NumCatalogCompras	NumTiendaCompras	NumWebVisitaMes	Edad	Unidad_Familiar
contador	472.0	472.0	472.0	472.0	472.0
media	6.1	8.31	2.48	45.65	1.68
std	2.26	2.79	1.54	13.79	0.57
min	2.0	4.0	0.0	19.0	1.0
25%	4.0	6.0	1.0	34.0	1.0
50%	6.0	8.0	2.0	45.0	2.0
75%	8.0	10.0	3.0	57.0	2.0
max	11.0	13.0	9.0	73.0	4.0

[17] Fuente: Elaboración propia.

ANEXO X

CLÚSTER 3- A TRAVÉS DE LA DISTANCIA DE MANHATTAN Y MÉTODO DE WARD [18]

	Educacion	Marital_Status	Ingresos	Vino	Frutas	Carne	Pescado	Dulces	Oro	NumDesCompras	NumWebCompras
contador	930.0	930.0	930.0	930.0	930.0	930.0	930.0	930.0	930.0	930.0	930.0
media	1.37	1.61	34118.8	33.97	4.3	25.86	5.85	4.58	14.34	1.95	2.01
std	0.96	0.49	16152.8	46.38	6.3	109.39	7.21	10.35	21.7	1.5	1.77
min	0.0	1.0	1730.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
25%	1.0	1.0	24723.75	7.0	0.0	7.0	0.0	0.0	4.0	1.0	1.0
50%	1.0	2.0	33428.5	18.0	2.0	13.0	3.0	2.0	9.0	2.0	2.0
75%	2.0	2.0	41448.25	40.0	6.0	23.0	8.0	6.0	18.0	2.0	3.0
max	3.0	2.0	162397.0	464.0	47.0	1725.0	43.0	262.0	321.0	15.0	27.0

	NumCatalogCompras	NumTiendaCompras	NumWebVisitaMes	Edad	Unidad_Familiar
contador	930.0	930.0	930.0	930.0	930.0
media	0.57	3.02	6.48	42.3	2.84
std	1.84	0.93	2.05	10.98	0.85
min	0.0	0.0	0.0	18.0	1.0
25%	0.0	2.0	5.0	35.0	2.0
50%	0.0	3.0	7.0	41.0	3.0
75%	1.0	4.0	8.0	49.0	3.0
max	28.0	8.0	20.0	74.0	5.0

[18] Fuente: Elaboración propia.

ANEXO X

CLÚSTER 4- A TRAVÉS DE LA DISTANCIA DE MANHATTAN Y MÉTODO DE WARD [19]

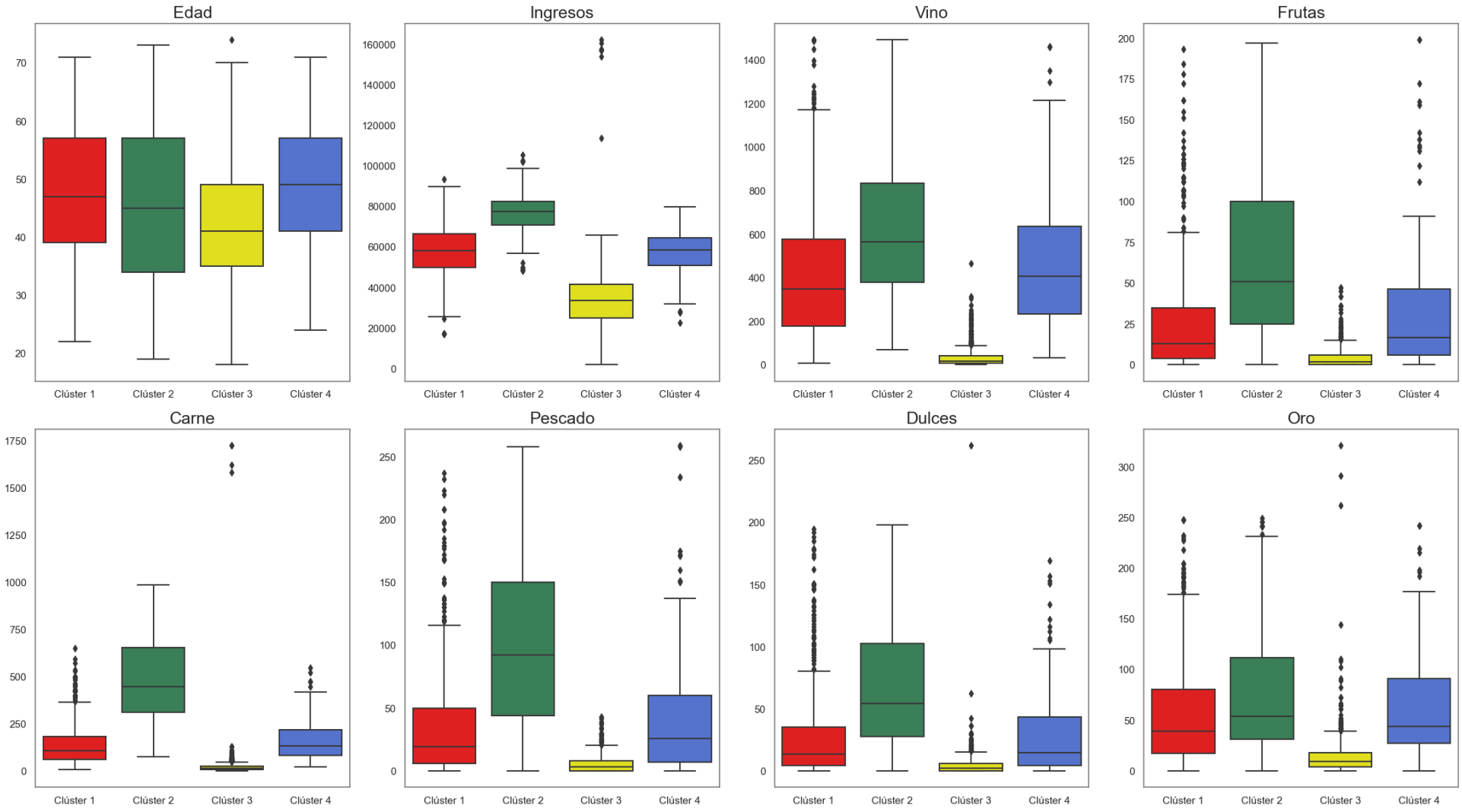
	Educacion	Marital_Status	Ingresos	Vino	Frutas	Carne	Pescado	Dulces	Oro	NumDesCompras	NumWebCompras
contador	212.0	212.0	212.0	212.0	212.0	212.0	212.0	212.0	212.0	212.0	212.0
media	1.64	1.0	57578.67	471.23	32.47	157.22	41.36	28.46	64.86	3.49	6.46
std	0.98	0.07	10432.35	304.7	39.38	103.53	48.46	34.19	53.01	2.12	2.4
min	0.0	1.0	22507.0	33.0	0.0	21.0	0.0	0.0	0.0	1.0	2.0
25%	1.0	1.0	50650.75	234.0	6.0	81.0	7.0	4.0	27.0	2.0	5.0
50%	1.0	1.0	58330.0	408.0	16.5	130.5	25.5	14.5	44.0	3.0	6.0
75%	3.0	1.0	64461.0	635.0	46.25	217.0	60.0	43.0	91.25	5.0	8.0
max	3.0	2.0	79761.0	1462.0	199.0	546.0	259.0	169.0	242.0	11.0	11.0

	NumCatalogCompras	NumTiendaCompras	NumWebVisitaMes	Edad	Unidad_Familiar
contador	212.0	212.0	212.0	212.0	212.0
media	3.32	8.12	5.89	48.8	2.0
std	2.13	2.69	1.71	10.07	0.5
min	1.0	3.0	2.0	24.0	1.0
25%	2.0	6.0	5.0	41.0	2.0
50%	3.0	8.0	6.0	49.0	2.0
75%	4.0	10.0	7.0	57.0	2.0
max	11.0	13.0	9.0	71.0	3.0

[19] Fuente: Elaboración propia.

ANEXO XI: CLÚSTERES A TRAVÉS DE LA DISTANCIA DE MANHATTAN Y MÉTODO DE WARD

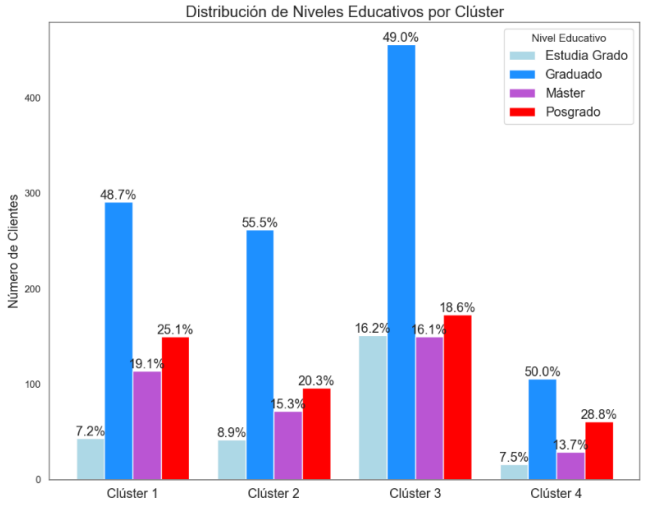
(*INGRESOS, VINO, FRUTAS, CARNE, PESCADO, DULCES* Y *ORO*) [20]

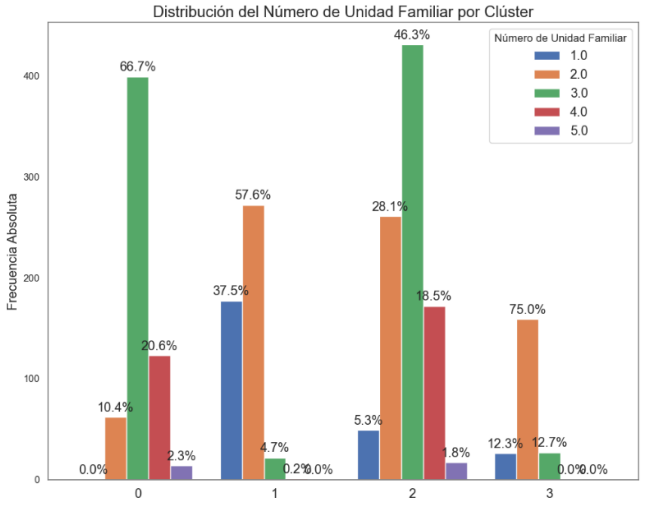


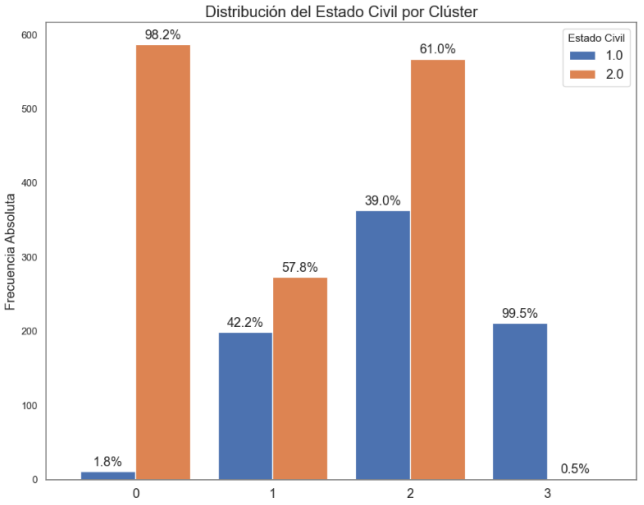
[20] Fuente: Elaboración propia.

ANEXO XII: CLÚSTERES A TRAVÉS DE LA DISTANCIA DE MANHATTAN Y MÉTODO DE WARD

(*EDUCACIÓN, UNIDAD FAMILIAR Y ESTADO MARITAL*) [21]



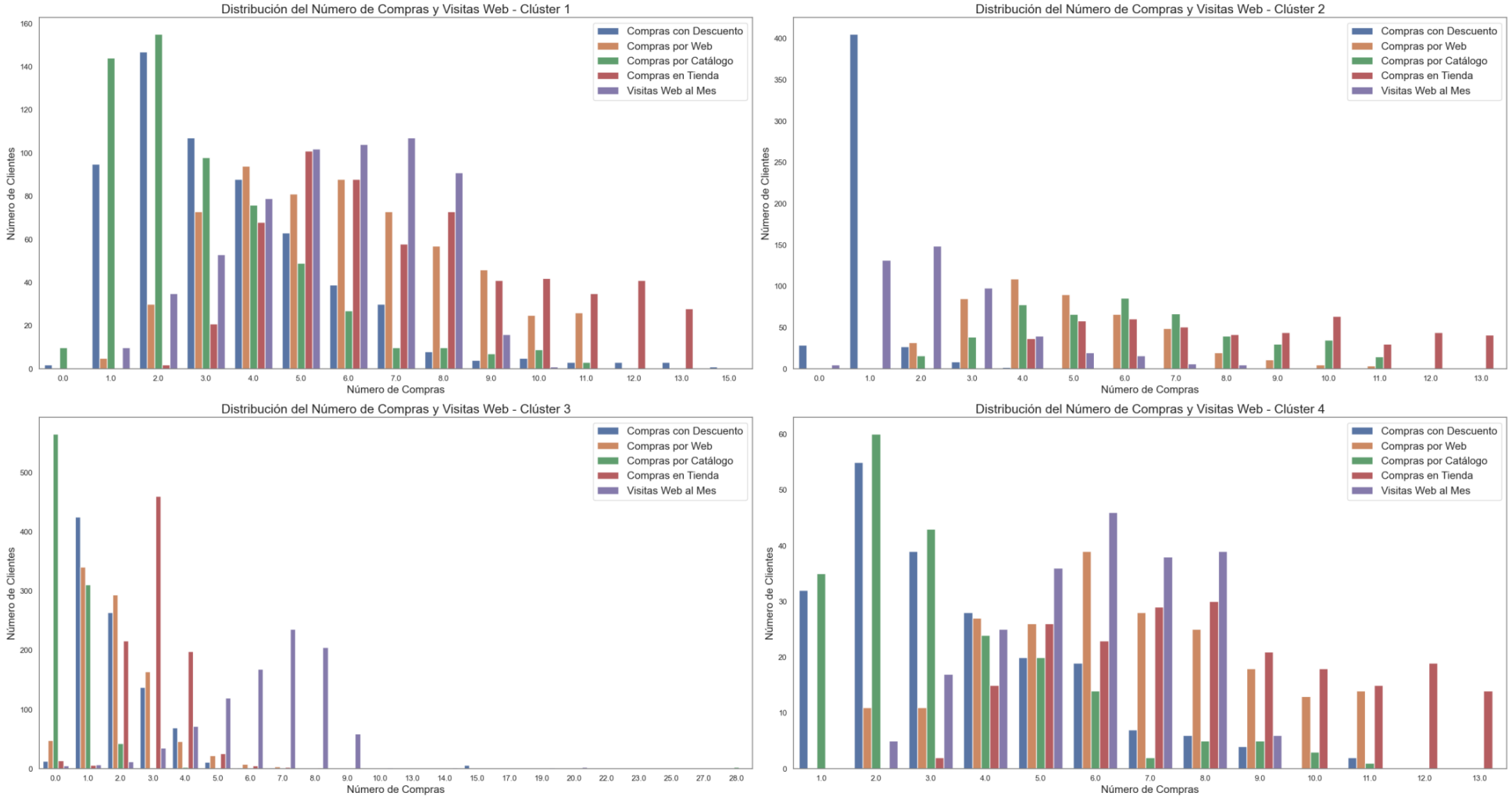




[21] Fuente: Elaboración propia.

ANEXO XIII: CLÚSTERES A TRAVÉS DE LA DISTANCIA DE MANHATTAN Y MÉTODO DE WARD

(*COMPRAS POR DECUENTO, WEB, CATÁLOGO, TIENDA* Y *NÚMERO DE VISTAS POR MES*) [22]



[22] Fuente: Elaboración propia.