Predicción de categorías de productos en el sector minorista utilizando técnicas de aprendizaje supervisado

Maria del Mar Ipia Guzmán

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Asesora
María Bernarda Salazar Sánchez, Doctor (PhD)

Universidad de Antioquia
Facultad de Ingeniería
Especialización en Analítica y Ciencia de Datos
Medellín, Antioquia, Colombia
2023

Cita	(Ipia Guzmán, 2023)
Referencia **Estilo APA 7 (2020)**	Ipia Guzmán, M. D. M. (2023). *Predicción de categorías de productos en el sector minorista utilizando técnicas de aprendizaje supervisado*. Trabajo de grado especialización]. Universidad de Antioquia, Medellín, Colombia.

Especialización en Analítica y Ciencia de Datos, Cohorte V.

Centro de Investigación Ambientales y de Ingeniería (CIA).

UNIVERSIDAD DE ANTIOQUIA
Vicerrectoría de Docencia

Sistema de Bibliotecas

Centro de Documentación Ingeniería (CENDOI)

Repositorio Institucional: http://bibliotecadigital.udea.edu.co

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes.
Decano: Julio Cesar Saldarriaga Molina
Jefe departamento: Diego José Luis Botia Valderrama

Dedicatoria

A mi madre, por su ejemplo y por su incansable apoyo, amor y sacrificio a lo largo de mi formación, ya que sin ella no hubiera sido posible este proceso. Su constante aliento ha sido la luz que impulsó y guio este camino.

A mi profesora María Bernarda Salazar Sánchez por su orientación, sabiduría y paciencia a lo largo de este proceso académico, la cual fue fundamental en la elaboración de esta monografía.

A mis hermanos que son el motivo para lograr cada objetivo trazado y a todos aquellos que, directa o indirectamente, contribuyeron a este logro.

Agradecimientos

Agradezco sinceramente a la Universidad de Antioquia y a los profesores por su invaluable guía, apoyo y conocimientos impartidos a lo largo de esta especialización ya que fueron cruciales para la realización de esta monografía. También extiendo mis agradecimientos a mis compañeros y equipos de proyecto que colaboraron conmigo en la exploración y análisis de datos, su cooperación fue esencial para alcanzar los objetivos de esta investigación. Además, no puedo dejar de expresar mi gratitud a mi familia y amigos por su apoyo incondicional, paciencia y motivación durante el desarrollo de esta monografía, su aliento fue un pilar fundamental en cada etapa de este proyecto.

Tabla de contenido

Lista de tablas

Lista de figuras

Siglas, acrónimos y abreviaturas

ML	Machine learning
PhD	Philosophiae Doctor
UdeA	Universidad de Antioquia
MIM	Matriz de información mutua

Resumen

En esta monografía de trabajo de grado, se lleva a cabo un análisis detallado de un extenso conjunto de datos de ventas minoristas que se obtiene de la plataforma kaggle. Este conjunto de datos abarca un período de tiempo desde el año 2021 hasta el 2023 e incluye información crucial sobre las transacciones de compra realizadas en un total de 10 tiendas ubicadas en Estambul. Los datos contienen una amplia gama de variables, como identificaciones de clientes, edades, géneros, métodos de pago, categorías de productos, cantidades, precios, fechas de pedidos y nombres de tiendas.

Se pretende evaluar y comparar diversos algoritmos de aprendizaje automático, con el fin de clasificar las categorías de los productos y obtener el mejor modelo de clasificación. Para ello, se utilizan algoritmo tales como Naive Bayes, Arboles de Decisión, Random Forest, Ada Boost, Gradient Boosting y redes neuronales. El desempeño de los seis modelos se evaluará con las métricas acuracy, precisión, recall y F1 score.

Palabras clave: ventas minoristas, análisis de datos, aprendizaje supervisado, predicción, comportamiento del cliente, mercado minorista, Estambul.

Abstract

In this undergraduate thesis monograph, a detailed analysis of an extensive retail sales dataset is conducted, which is obtained from the Kaggle platform. This dataset spans a period from 2021 to 2023 and includes crucial information about purchase transactions made in a total of 10 stores located in Istanbul. The data contains a wide range of variables, such as customer IDs, ages, genders, payment methods, product categories, quantities, prices, order dates, and store names.

The goal is to assess and compare various machine learning algorithms to classify product categories and obtain the best classification model. For this purpose, algorithms such as Naive Bayes, Decision Trees, Random Forest, Ada Boost, Gradient Boosting, and neural networks are employed. The performance of the six models will be evaluated using metrics such as accuracy, precision, recall, and F1 score.

Keywords: retail sales, data analysis, supervised learning, prediction, customer behavior, retail market, Istanbul.

1. Descripción del problema

El avance de la tecnología y la proliferación de grandes cantidades de datos, han dado lugar a un creciente interés en el campo del aprendizaje automático y en las aplicaciones que se puede tener en diferentes sectores. Uno de ellos es el mercado minorista actual, en cual el uso del machine learning (ML) se ha convertido en un recurso invaluable para comprender y aprovechar la información generada a través de las transacciones de compra. En este trabajo, se exploran algunas técnicas de aprendizaje supervisado para la clasificación de categorías de compras en el mercado minorista.

1.1. Problema de negocio

La clasificación precisa de las compras en diferentes categorías es fundamental para las empresas minoristas, ya que proporciona una visión profunda de los patrones de compra, el comportamiento del consumidor y las preferencias del mercado. Sin embargo, a medida que crece la cantidad de productos disponibles y se diversifican las categorías, se vuelve cada vez más difícil realizar esta clasificación de forma manual y eficiente.

Implementar modelos de aprendizaje supervisado, podría permitir a las tiendas minoristas en Estambul mejorar su toma de decisiones estratégicas y optimizar sus operaciones, ya que pueden comprender mejor las preferencias de los clientes y las categorías de compras. Lo anterior conlleva a que las tiendas adapten su inventario, la estrategia de marketing y las campañas de promociones para satisfacer las necesidades de los consumidores de manera más efectiva y directa, generando mayor rentabilidad y competitividad en el mercado. Además de esto, la clasificación de categorías de productos en ventas minoristas mediante modelos de ML puede mejorar significativamente la eficiencia operativa, la satisfacción del cliente y la rentabilidad de la empresa, lo que la convierte en una práctica esencial en el entorno minorista moderno. En este sentido, la correcta clasificación de las categorías se puede ver reflejado en la mejora la experiencia del cliente y con ello, poder fidelizarlos con promociones personalizadas y con poder aumentar las posibilidades de que realicen compras.

1.2. Aproximación desde la analítica de datos

El aprendizaje automático se puede clasificar en dos tipos principales: aprendizaje supervisado y aprendizaje no supervisado y la implementación de estos depende de la tarea a resolver. El aprendizaje supervisado es un método en el campo del aprendizaje automático en el cual el algoritmo se entrena con datos etiquetados, lo que significa que tiene acceso tanto a las características de entrada como a las etiquetas de salida correctas. Esto se subdivide aún más en tareas de clasificación y tareas de regresión: en la clasificación, las etiquetas son categorías discretas, mientras que, en la regresión, las etiquetas son cantidades continuas (VanderPlas, Python Data Science Handbook. Essential Tools for Working with Data, 2016).

El aprendizaje no supervisado implica modelar las características de un conjunto de datos sin hacer referencia a ninguna etiqueta. Estos modelos incluyen tareas como el agrupamiento y la reducción de dimensionalidad. Los algoritmos de agrupamiento identifican grupos distintos de datos, mientras que los algoritmos de reducción de dimensionalidad buscan representaciones más concisas de los datos (Géron, 2017)

Por otra parte, el aprendizaje no supervisado se emplea cuando no se dispone de datos etiquetados para el entrenamiento. En estos métodos, la única información disponible son los datos de entrada, y el propósito es descubrir una estructura o patrón que simplifique el análisis de esos datos. Los algoritmos comunes incluyen técnicas como el agrupamiento, el análisis de componentes principales y la identificación de anomalías (Agrawal, 2021)

En este caso, se plantea un problema de clasificación, en el que las características de entrada incluirán diversos atributos como información del cliente, detalles de transacciones y datos relacionados con los productos, mientras que las etiquetas de salida representarán las categorías de productos. Aprovechando estos datos etiquetados, los algoritmos aprenderán patrones y relaciones entre las características de entrada y las categorías de productos correspondientes.

El objetivo final es construir un sólido modelo de clasificación que pueda predecir y asignar con precisión los productos a sus categorías adecuadas, brindando información valiosa a minoristas y empresas. Para ello, se utilizarán modelos como Árboles de Decisión y Bosques Aleatorios los cuales dividen el conjunto de datos en subconjuntos más pequeños basados en ciertas reglas. Los bosques aleatorios combinan múltiples árboles de decisión para mejorar la precisión. Redes Neuronales ya que especialmente las redes neuronales profundas (Deep Learning) son potentes

para clasificación al aprender representaciones complejas de los datos. Naive Bayes: Basado en el teorema de Bayes, asume independencia entre características y se usa comúnmente en clasificación de texto y minería de datos. Gradient Boosting Models (XGBoost, LightGBM, CatBoost): Estos modelos combinan múltiples modelos más débiles para mejorar la precisión general.

1.3. Origen de los datos

Para la realización de este informe, se utilizó la plataforma Kaggle para la obtención de los datos (Kaggle, 2023). Este conjunto de datos de ventas minoristas contiene información sobre las transacciones de compras realizadas en 10 diferentes tiendas en Estambul, entre los años 2021 y 2023, además de identificación de los clientes, edad, género, métodos de pago, categorías de productos, cantidad, precio, fechas de pedidos y nombres de las tiendas.

1.4. Métricas de desempeño

las métricas de desempeño son esenciales para comprender, comparar y mejorar modelos machine learning, proporcionando una evaluación objetiva y cuantitativa de su rendimiento. Con el fin de evaluar qué tan bien el modelo es capaz de predecir las clases o categorías correctas, se utilizan la matriz de confusión, exactitud, precisión, sensibilidad y F1-Score como métricas de desempeño de evaluación de un modelo de clasificación, tal que permiten evaluar el ajuste entre la salida del modelo y los datos métricas (Brownlee, 2020). De acuerdo con el objetivo principal que se planteó para la realización de este trabajo, se puede utilizar la precisión para comparar diferentes modelos de clasificación. El modelo con la mayor precisión es el que clasifica correctamente la categoría, o se podría utilizar el recall (true positive rate) para evaluar el rendimiento del modelo para cada categoría. Un recall alto indica que el modelo está clasificando correctamente la mayoría de las compras de una categoría (Agrawal, 2021).

En términos de evaluación de métricas de negocio, obtener modelos que logren la personalización y generación de recomendaciones sobre las categorías de productos a los clientes aumentará probablemente las ventas cruzadas para la empresa. Sin contar con qué esto puede generar beneficios tales cómo la gestión de inventario gracias a la clasificación precisa de

productos. Esto puede llevar a predecir la demanda de productos en cada categoría y mejorar el manejo de stock disponible para los productos en cada categoría. Así mismo, terminar generando una segmentación del mercado, ya que, al conocer las preferencias de compra de los clientes en diferentes categorías de productos, las tiendas minoristas pueden segmentar su mercado y dirigir sus esfuerzos de marketing de manera más efectiva, lo que puede influir en las decisiones de compra de los clientes y el ahorro de tiempo y recursos y niveles altos de competitividad.

En general, las métricas de desempeño del negocio deben elegirse teniendo en cuenta los objetivos del negocio. Por ejemplo, si el objetivo del negocio es maximizar las ventas, entonces la precisión podría ser la métrica más importante. Sin embargo, si el objetivo del negocio es minimizar los costos, entonces el costo de los falsos positivos podría ser la métrica más importante.

2. Objetivos

2.1. Objetivo general

Desarrollar un modelo para predecir las categorías de productos en el sector minorista utilizando técnicas de aprendizaje supervisado como herramienta útil para la gestión de inventario y la personalización de recomendaciones en tiendas minoristas.

2.2. Objetivos específicos

- Identificar el conjunto de características que más aporten a la clasificación de las categorías de productos en el sector minorista.
- Evaluar varios algoritmos de aprendizaje supervisado como estrategia de clasificación de categorías de productos en el sector minorista.
- Evaluar la precisión y el rendimiento del modelo utilizando las métricas precisión, recall y F1-score.

3. Fuente de datos y análisis descriptivo

Inicialmente, se realizó el procesamiento del conjunto de datos, con el fin de obtener unos datos limpios para entrenar los modelos. El conjunto de datos tiene 15 variables entre las cuales se encuentran la categoría que es la variable que se quiere predecir. Esta variable se elimina y se dejan las demás características. Lo anterior se realiza porque en el entrenamiento de modelos de aprendizaje automático (ML), es crucial eliminar la variable objetiva (o etiqueta) del conjunto de datos de entrenamiento, ya que esta variable es precisamente lo que el modelo intenta predecir. Si se incluye la variable objetiva en los datos de entrada, el modelo simplemente aprendería a asociar directamente esta variable consigo misma, lo que resulta en un sobre ajuste extremo y una incapacidad para generalizar o hacer predicciones significativas sobre datos no vistos. Esto es equivalente a darle al modelo las respuestas durante su entrenamiento, lo cual invalida el propósito de aprender patrones subyacentes en los datos para realizar predicciones precisas.

Se utilizó la función de sckit-learn, train_test_split para obtener los conjuntos de datos de entrenamiento y de testeo. Esta división se hace en un 80 % para entrenar los modelos, es decir 75. 568 registros y un 20 % para evaluarlos, que equivale a 18.893.

Se implementan los modelos anteriormente mencionados y cómo metodología primero se entrenan los modelos, se realizan las predicciones y se validan las métricas de desempeño. Con el fin de buscar los mejores hiperparámetros se utiliza GridSearchCV que es una técnica de validación cruzada incluida en el paquete de scikit learn. Lo que hace es ejecutarse a través de los diferentes parámetros que se introducen en la cuadrícula de parámetros y extraer los mejores valores y combinaciones de parámetros y finalmente se realiza una validación por medio de la función de sckit-learn, cross_val_score con el fin de que el modelo no haga sobreajuste.

El conjunto de datos utilizados para la fase de entrenamiento tiene un tamaño de 94.461 filas y 14 columnas.

3.1. Datos originales y análisis descriptivo

El conjunto de datos "Customer Shopping Dataset" disponible en Kaggle contiene datos de ventas minoristas que proporciona información sobre las transacciones de compras realizadas en 10 tiendas minoristas en Turquía desde el 2021 hasta el 2023. El conjunto de datos se presenta en un archivo CSV que contiene 99,457 registros y 10 columnas (ver **¡Error! No se encuentra el o rigen de la referencia.**).

Variable	Descripción	Tipo de dato
invoice_no	Número de factura	Cadena de texto
customer_id	Número de cliente	Cadena de texto
gender	Género del cliente	Cadena de texto
age	Edad del cliente	Entero
category	Categoría del producto comprado	Cadena de texto
quantity	Las cantidades de cada producto (artículo) por transacción	Entero
price	Precio del producto por unidad en liras turcas (TL).	Flotante
payment_method	Método de pago (efectivo, tarjeta de crédito o tarjeta de débito) utilizado para la transacción.	Cadena de texto
invoice_date	El día en que se generó una transacción	Cadena de texto
shopping_mall	Nombre del centro comercial donde se realizó la transacción	Cadena de texto

Tabla 1. Variables del conjunto de datos

3.2. Datasets

En el contexto del aprendizaje automático, la separación de un conjunto de datos en conjuntos de entrenamiento y prueba es esencial para desarrollar y evaluar modelos predictivos. Este proceso implica dividir el conjunto de datos original en dos partes: el conjunto de entrenamiento y el conjunto de prueba. El conjunto de entrenamiento se utiliza para alimentar al modelo durante su fase de entrenamiento, permitiéndole aprender patrones y relaciones en los datos. Por otro lado, el conjunto de prueba se reserva exclusivamente para evaluar la capacidad del modelo de generalizar lo que ha aprendido. Esta división es fundamental para medir la calidad y el

rendimiento del modelo, ya que permite comprobar cómo se comporta con datos no vistos previamente.

Para este problema, se utiliza la función de sckit-learn, train_test_split para obtener los conjuntos de datos de entrenamiento y de testeo. Esta división se hace en un 80 % para entrenar los modelos y un 20 % para evaluarlos.

3.3. Analítica descriptiva

El conjunto de datos "Customer Shopping Dataset" contiene información sobre las transacciones de compras realizadas en 10 tiendas en Turquía desde el año 2021 hasta el 2023 (Kaggle, 2023). Contiene 99.457 registros y 10 columnas, 3 numéricas y 7 categóricas. Inicialmente se realiza una división del conjunto de datos para hacer un análisis de las variables que tiene el dataset.

Variables Categóricas

Al analizar la información contenida en la data, se encuentra que las mujeres tienen una mayor participación en el total de las compras realizadas, con una diferencia de 19.507 (ver Figura 1). En el conjunto de datos se registran tres métodos de pago: efectivo, tarjeta débito y tarjeta crédito (ver Figura 2**¡Error! No se encuentra el origen de la referencia.¡Error! No se encuentra el origen de la referencia.¡Error! No se encuentra el origen de la referencia.¡Error! No se encuentra el origen de la referencia.**). El 44.7% de las transacciones se realizan con efectivo, es decir 44.447, mientras que 34.931 se realizan con tarjeta de crédito y 20.079 con tarjeta débito. Se evidencia además que, un poco más de la mitad del total de las transacciones realizadas en el periodo de tiempo estudiado, es decir el 55.07% se centra en 3 centros comerciales: Mall of Istanbul, Kanyon y Metrocity (ver Figura 3).

Figura 1. Distribución por género en el total de compras realizadas

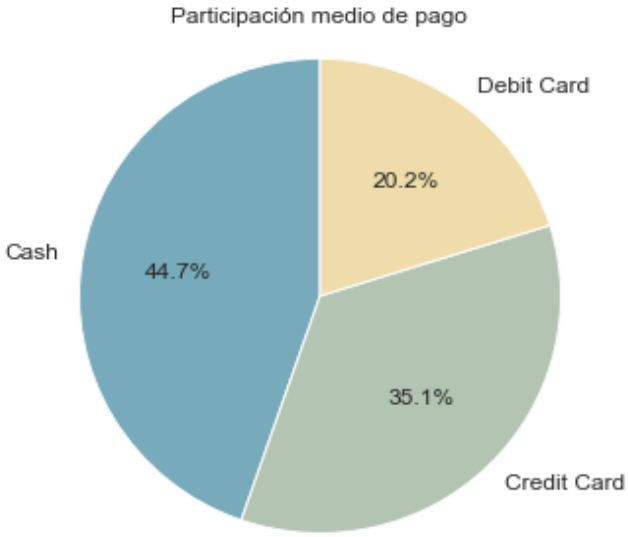

Figura 2. Distribución por medio de pago en el total de compras realizadas

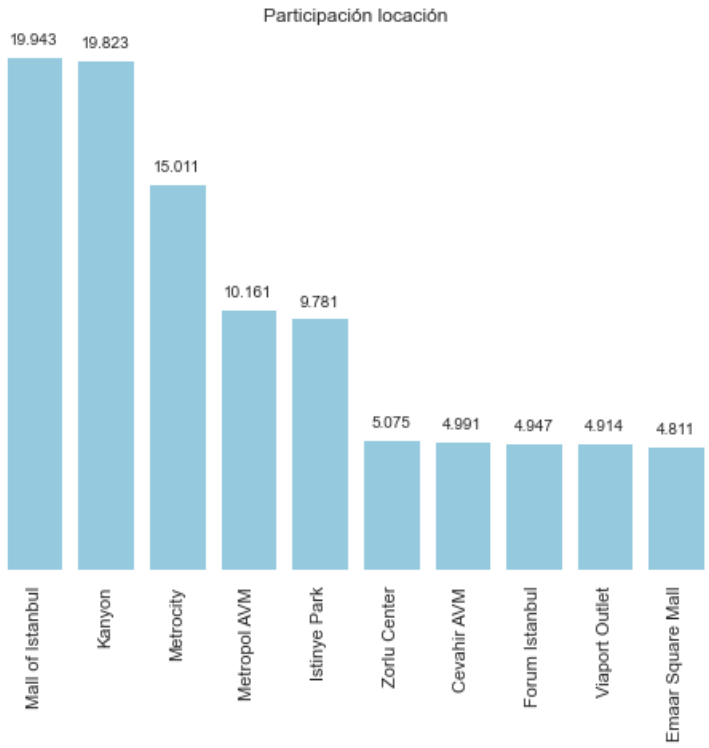

Figura 3. Distribución por locación en el total de compras realizadas

El total de las transacciones de compra se encuentran distribuidas en ocho categorías de compra (ver Figura 4). Los productos de la categoría de ropa ocupan el porcentaje más alto con el 34.7% que representa 34.487 transacciones, seguido de los cosméticos con 15.2% equivalente a 15.907 transacciones y comidas y bebidas con 14.9% equivalentes a 14.776 transacciones. Debido a que el objetivo de esta monografía es tratar de predecir las categorías de compra, es importante realizar un análisis de la distribución de los datos de las 8 categorías, ya que cerca del 35% de las transacciones es de ropa y cinco de las categorías, no alcanzan una participación mayor al 10% de los datos, lo que evidencia que hay un desequilibrio en los datos a predecir.

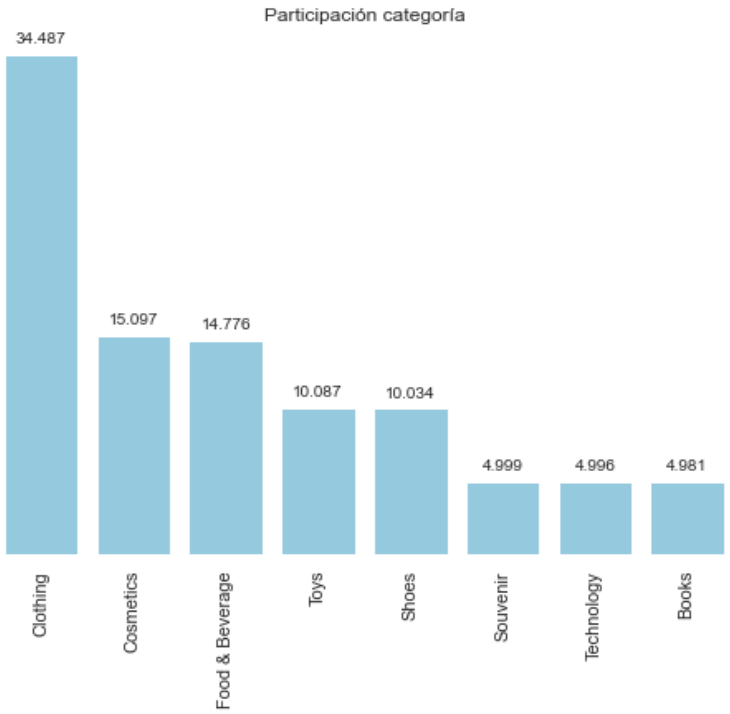

Figura 4. Distribución por categoría del producto en el total de compras realizadas.

Variables numéricas

En el data set se presentan tres variables numéricas, la edad de los participantes que oscila entre 18 y 69 años, con una media de 43 años y una distribución uniforme (ver Figura 5). La cantidad de productos comprados, que oscilan entre uno y cinco productos y tiene una distribución uniforme (ver Figura 6) y el precio de venta en el que la mayoría de las transacciones involucran productos de menor precio, aunque hay casos de productos significativamente más caros (ver Figura 7).

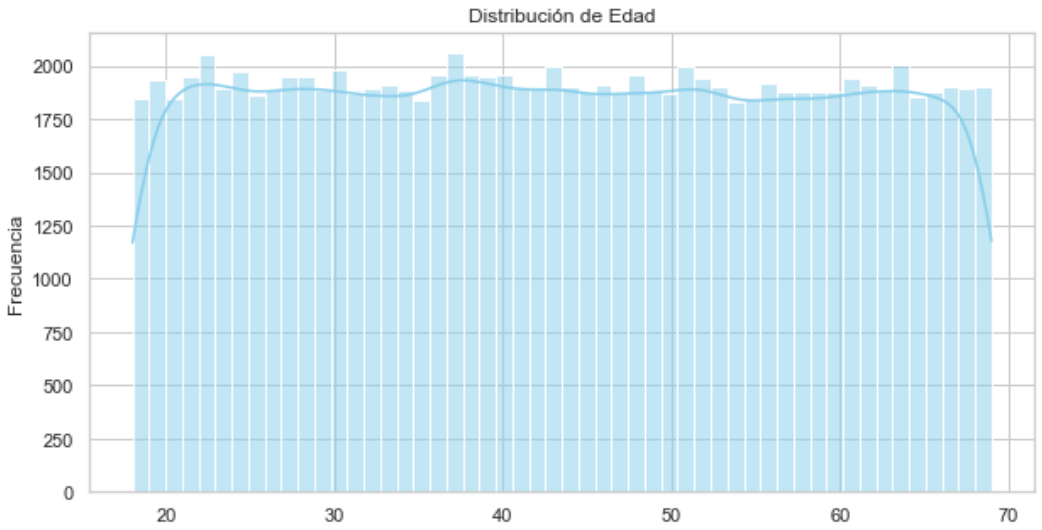

Figura 5. Distribución de edades de los participantes en el conjunto de datos.

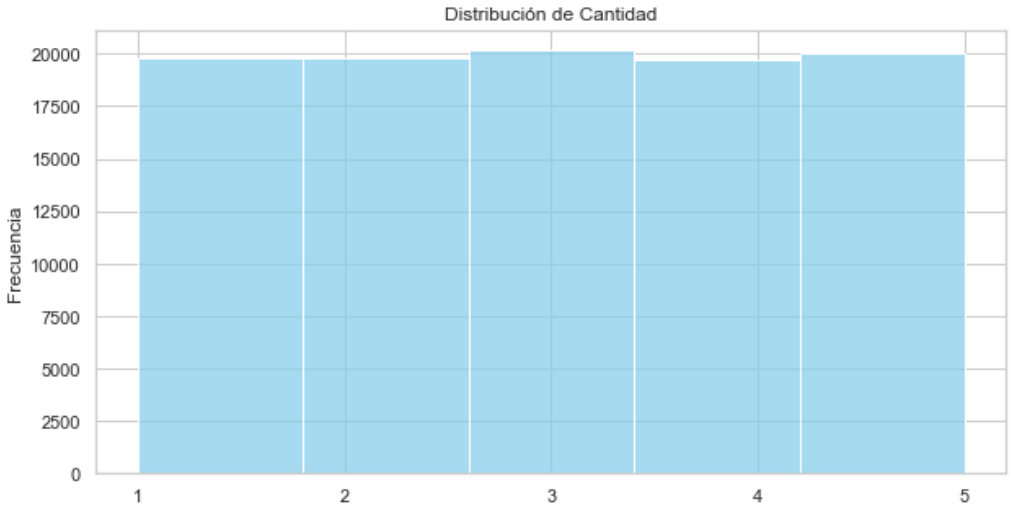

Figura 6. Distribución de cantidades de productos comprados en el conjunto de datos.

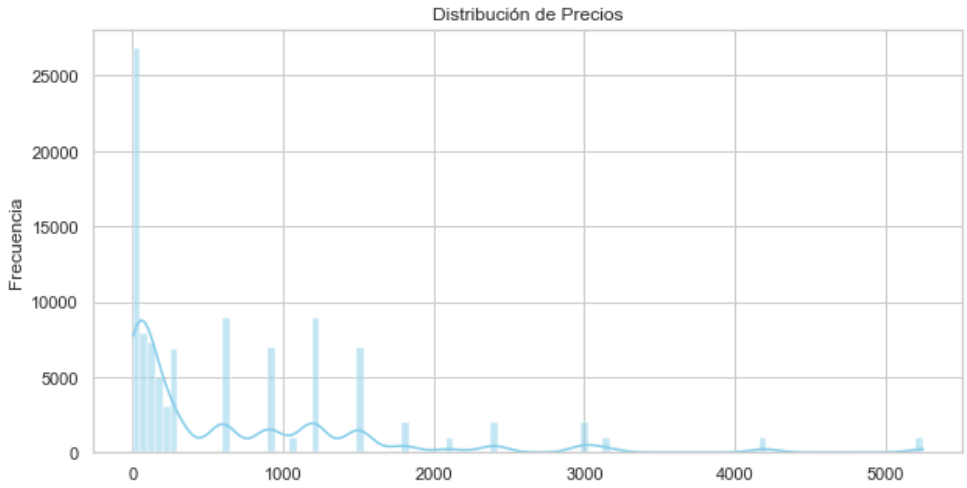

Figura 7. Distribución de precios de los productos en el conjunto de datos.

4. Metodología

El pipeline, o flujo de trabajo, en un proyecto de análisis de datos y modelado abarca un proceso secuencial de varias etapas cruciales (ver **¡Error! No se encuentra el origen de la r eferencia.**).

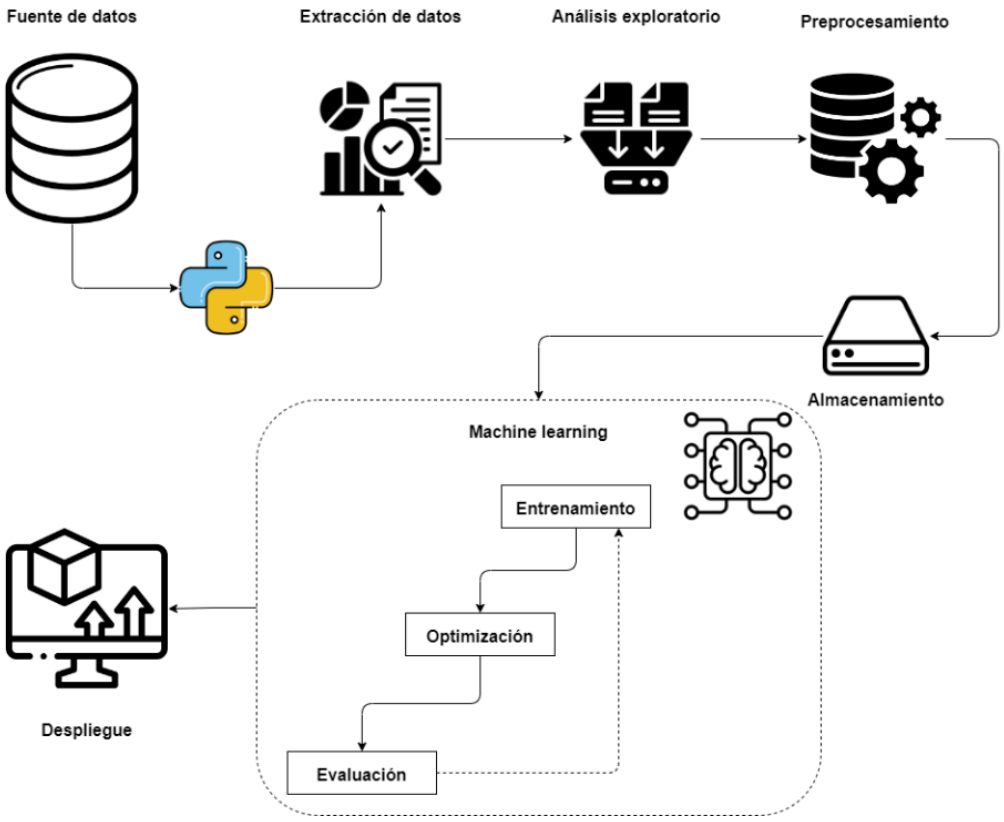

Figura 8. Flujo de trabajo para el desarrollo y validación del modelo

Se inicia con la etapa de extracción de datos, donde se recopilan fuentes de información relevantes para el análisis. Posteriormente, el análisis exploratorio de datos se lleva a cabo, permitiendo descubrir patrones, tendencias y posibles relaciones, lo que proporciona una comprensión inicial de los datos. La fase de preprocesamiento implica la limpieza, transformación y normalización de los datos para su uso en modelos predictivos. Acto seguido, el almacenamiento de datos asegura un acceso seguro y organizado para futuros análisis. La ejecución de modelos abarca la implementación de algoritmos y técnicas de machine learning para extraer patrones y

generar predicciones. Finalmente, el despliegue de modelos permite la implementación de los resultados para su aplicación práctica, facilitando la toma de decisiones informadas. Este pipeline destaca la importancia de cada etapa, subrayando la necesidad de una integración sin problemas entre ellas para garantizar la obtención de resultados precisos y aplicables

4.1. Análisis estadístico

Para realizar un análisis de las variables categóricas, se utilizan tres herramientas estadísticas: tabla de contingencia, Test chi-cuadrado y matriz de información mutua.

4.1.1. Tablas de contingencia.

Esta es una herramienta estadística que se utiliza para resumir y analizar la relación entre dos o más variables categóricas. En una tabla de contingencia, se registran las frecuencias conjuntas de las categorías de las variables categóricas, lo que permite visualizar y analizar patrones de asociación o independencia entre ellas. En una tabla de contingencia, se registran las frecuencias conjuntas de las categorías de las variables categóricas, lo que permite visualizar y analizar patrones de asociación o independencia entre ellas. (Quintero, 2017)

Del total de las transacciones del conjunto de datos, las mujeres realizan el 59.8 % de, mientras que los hombres el 40.2 %. El medio de pago preferido por los consumidores es el efectivo ya que, de las 59.482 transacciones realizadas por mujeres, el 44.6 % se hacen con este medio de pago y de las 39.975 transacciones hechas por hombres, el 44.9 % se hacen con efectivo (Ver Tabla 2). Aunque las mujeres presentan más registros de ventas en el conjunto de datos, el comportamiento de las compras y el medio de pago utilizado es similar para los dos géneros.

payment_method	Cash	Credit Card	Debit Card	Total
gender				
Female	26509	21011	11962	59482
Male	17938	13920	8117	39975
Total	44447	34931	20079	99457

Tabla 2. Transacciones realizadas por medio de pago y género.

Un comportamiento similar se observa si se analizan los medios de pago utilizados en las compras realizadas en los centros comerciales registrados. El 55.07 % de las transacciones se concentran en tres centros comerciales, Kanyon con 19.823 transacciones, Mall de Estambul con 19.943 y Metrocity con 15.011 y en ellos, la participación del uso del efectivo como medio de pago en las compras es similar con un 44.7 %, 44.6 % y 44.1 % respectivamente (Ver Tabla 3). La distribución de compras realizadas en los tres centros comerciales mencionados se distribuye en un 60 % aproximadamente para las mujeres y 40 % para los hombres (Ver Tabla 4).

shopping_mall	Cevahir AVM	Emaar Square Mall	Forum Istanbul	Istinye Park	Kanyon	Mall of Istanbul	Metrocity	Metropol AVM	Viaport Outlet	Zorlu Center	Total
payment_method											
Cash	2228	2114	2183	4436	8853	8894	6625	4559	2231	2324	44447
Credit Card	1779	1696	1750	3422	6916	7019	5347	3521	1721	1760	34931
Debit Card	984	1001	1014	1923	4054	4030	3039	2081	962	991	20079
Total	4991	4811	4947	9781	19823	19943	15011	10161	4914	5075	99457

*Tabla 3***.** Transacciones realizadas por medio de pago y locación.

shopping_mall	Cevahir AVM	Emaar Square Mall	Forum Istanbul	Istinye Park	Kanyon	Mall of Istanbul	Metrocity	Metropol AVM	Viaport Outlet	Zorlu Center	Total
gender											
Female	2940	2842	3016	5874	11906	11902	8941	6144	2949	2968	59482
Male	2051	1969	1931	3907	7917	8041	6070	4017	1965	2107	39975
Total	4991	4811	4947	9781	19823	19943	15011	10161	4914	5075	99457

*Tabla 4***.** Transacciones realizadas por género y locación.

4.1.2. Test chi-cuadrado

La prueba de chi-cuadrado, es una prueba estadística que se utiliza para determinar si existe una asociación significativa entre dos variables categóricas en una tabla de contingencia. El objetivo principal de la prueba de chi-cuadrado es evaluar si las dos variables son independientes o si existe una relación significativa entre ellas. La distribución chi-cuadrado es una componente importante de la prueba y estimación de hipótesis estadísticas. (Ronald E. Walpole, Raymond H. Myers, Sharon L. Myers, & Keying Ye, 2007). El resultado de aplicar esta prueba al problema en estudio se relaciona en la Tabla 5.

Dado que el valor de p (~0.85) supera un nivel de significancia preestablecido (comúnmente 0.05), no se dispone de pruebas suficientes para afirmar que existe una relación significativa entre el género y el método de pago. En otras palabras, no es posible afirmar con confianza que las variables categóricas están vinculadas de manera significativa.

Tabla	Chi-cuadrado	Valor-p
Género – Medio de pago	2.665	0.8496
Género – Locación	12.4466	0.8998
Medio de pago – Locación	14.0861	0.9939

*Tabla 5***.** Resultados test chi-cuadrado

4.1.3. Matriz de Información Mutua (MIM)

La matriz de información mutua es una herramienta utilizada en el campo de la estadística y el análisis de datos para medir la relación o la dependencia entre dos o más variables aleatorias, especialmente en el contexto de variables discretas o categóricas. Es una medida que cuantifica la dependencia entre dos variables aleatorias al medir cuánta información proporciona una variable sobre la otra, es decir que mide la cantidad de información que una variable aleatoria contiene sobre la otra. El objetivo es obtener la mayor cantidad de información posible de dos variables aleatorias, y esto se conseguirá al maximizar la información mutua. (Pacheco Melo & Rojas Vilches, 2014). Los valores de información mutua son muy pequeños por lo que se puede concluir que las variables categóricas no están relacionadas (ver Figura 9).

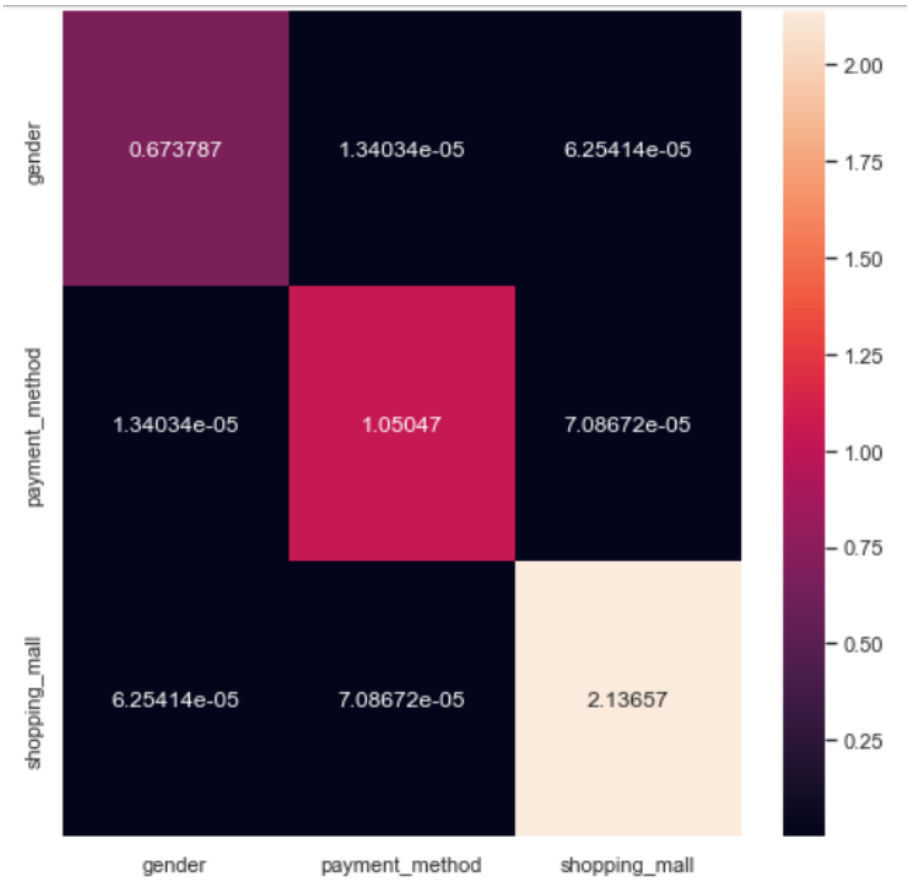

Figura 9. Matriz de información mutua de las variables categóricas.

4.1.4. Correlación entre variables numéricas

Para realizar un análisis de las variables numéricas, se utiliza el coeficiente de correlación de Pearson, que es una medida estadística que evalúa la relación lineal entre dos variables continuas. El rango de valores del coeficiente de correlación de Pearson va de -1 a 1, donde un valor de 1 indica una correlación lineal positiva perfecta, es decir que a medida que una variable aumenta, la otra también lo hace en la misma proporción. (Restrepo & González, 2007) Un valor de -1 indica una correlación lineal negativa perfecta, lo que significa que a medida que una variable aumenta, la otra disminuye en la misma proporción. Un valor cercano a 0 indica una correlación débil o inexistente entre las dos variables. Además, con el coeficiente de correlación de Pearson, se obtiene el valor-p, el cual indica la significancia de la correlación observada entre dos variables (ver **¡Error! No se encuentra el origen de la referencia.**).

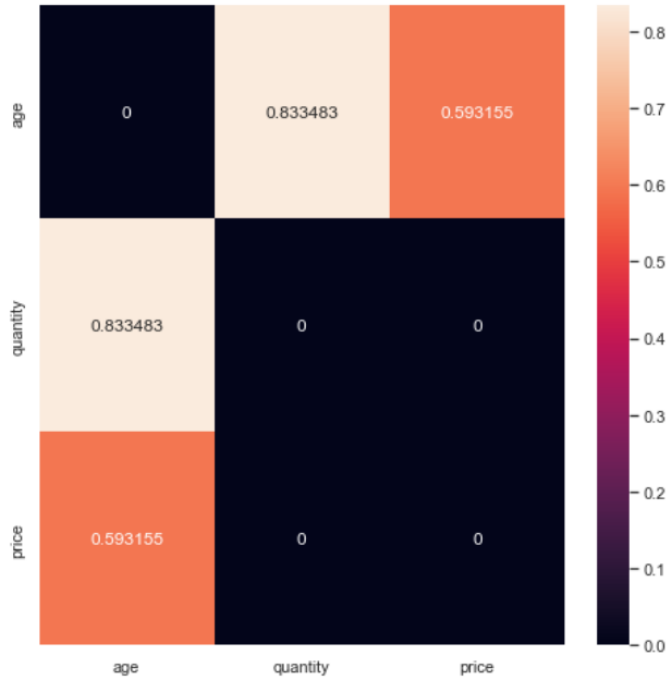

Figura 10. Mapa de calor de las variables numéricas.

Si el valor p es bajo (por lo general, menor a 0.05), se considera que la correlación observada es significativa. Esto sugiere que es poco probable que la correlación entre las variables sea el resultado del azar y que realmente existe una relación significativa entre ellas en la población.

Si el valor p es alto (generalmente mayor a 0.05), esto indica que la correlación observada es más probable que sea el resultado de la aleatoriedad en lugar de una verdadera relación entre las variables en la población. Por lo tanto, Precio y Cantidad, están correlacionadas linealmente.

4.2. Preprocesamiento

4.2.1. Imputación de datos

Tras realizar un análisis para identificar posibles valores faltantes, se ha constatado que el conjunto de datos está completo, sin ausencia de información en ninguna de sus columnas, por lo tanto, no se hace imputación de datos. La ausencia de datos faltantes en el conjunto de datos es indicativa de una buena recopilación y mantenimiento de los registros, lo que a su vez facilita el proceso de análisis y asegura una mayor fiabilidad en los resultados obtenidos a partir de estos datos.

4.2.2. Datos atípicos

Inicialmente, se analizan las tres variables numéricas en conjunto por medio de un diagrama de cajas con el fin de poder visualizar la presencia de datos atípicos. Posterior a ello, se realiza un análisis de manera individual para ver el comportamiento de cada variable (ver Figura 11).

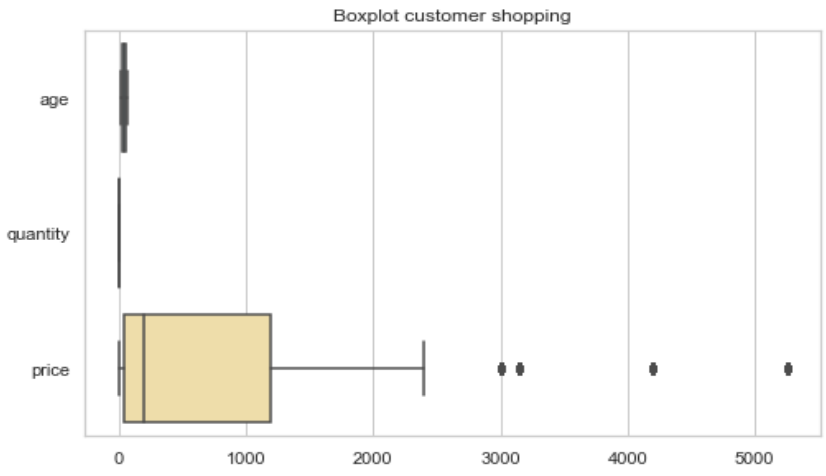

Figura 11. Diagrama de cajas de las variables numéricas del conjunto de datos.

4.2.3. Normalización de los datos

Se normalizan los datos de las tres variables numéricas por medio del método MinMaxScaler, con el objetivo de transformar los atributos para que estén en una escala similar, lo cual podría mejorar el rendimiento y la estabilidad del entrenamiento del modelo. Se puede inferir de una manera visual, que se encuentran valores atípicos en la variable precio (Ver **¡Error! No se encuentra el origen de la referencia.** y **¡Error! No se encuentra el origen de la referencia.**).

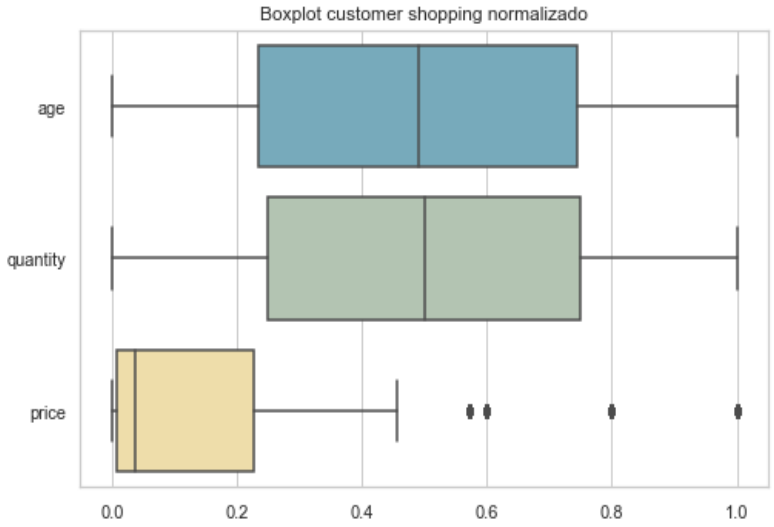

Figura 12. Diagrama de cajas normalizado de las variables numéricas.

Sin embargo, es importante analizar que la variable precio en el conjunto de datos presenta un rango intercuartil (IQR) de 1154.87, lo que indica una amplia dispersión entre el primer cuartil (Q1) y el tercer cuartil (Q3). Con un primer cuartil a 45.45 y un tercer cuartil a 1200.32, se observa que hay valores de precio tanto muy bajos como muy altos. Esta significativa dispersión sugiere que hay productos con una gran variedad de precios dentro del conjunto de datos. Los valores extremadamente bajos pueden representar artículos de menor calidad, promociones, descuentos significativos o productos básicos, mientras que los valores extremadamente altos pueden reflejar artículos de lujo, de alta demanda o con características premium.

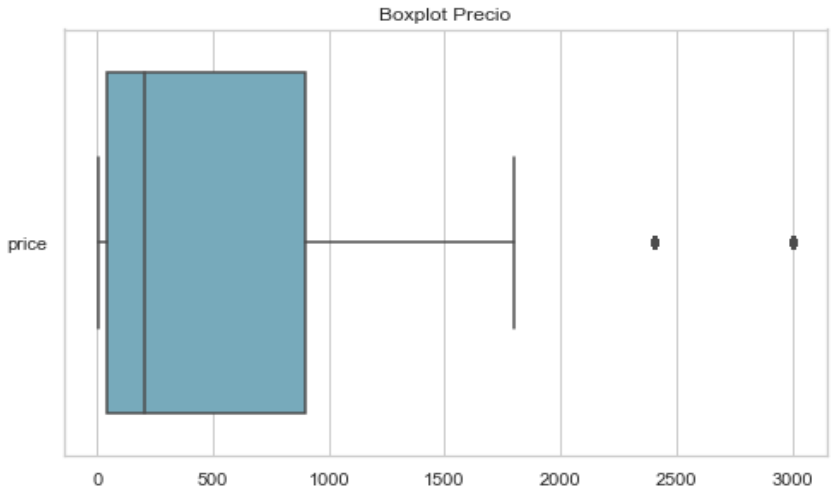

Figura 13. Diagrama de cajas normalizado de la variable precio.

Debido a esto, cada entrada de datos, tras una cuidadosa revisión, parece caer dentro de rangos esperados y razonables, sin indicaciones de errores de entrada o mediciones que distorsionen significativamente los patrones subyacentes. Por lo anterior, se mantienen la totalidad de los datos, permitiendo que el análisis posterior y los modelos de aprendizaje automático que se desarrollen reflejen fielmente la variabilidad natural y las tendencias inherentes al conjunto de datos. Por otra parte, la variable edad no presenta valores atípicos (Ver Figura 14) ya que las edades de los participantes se encuentran en un rango de edades que no es considerado anómalo (18 y 63 años). La cantidad de productos comprados que se encuentra en un rango de uno a cinco productos (Ver Figura 15), tampoco presenta datos atípicos.

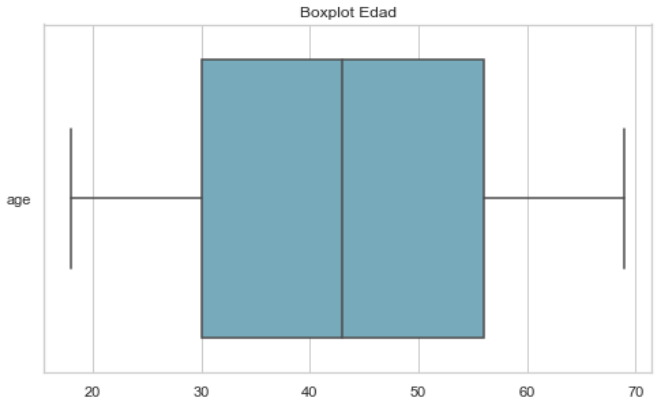

Figura 14. Diagrama de cajas de la edad de los participantes en el conjunto de datos.

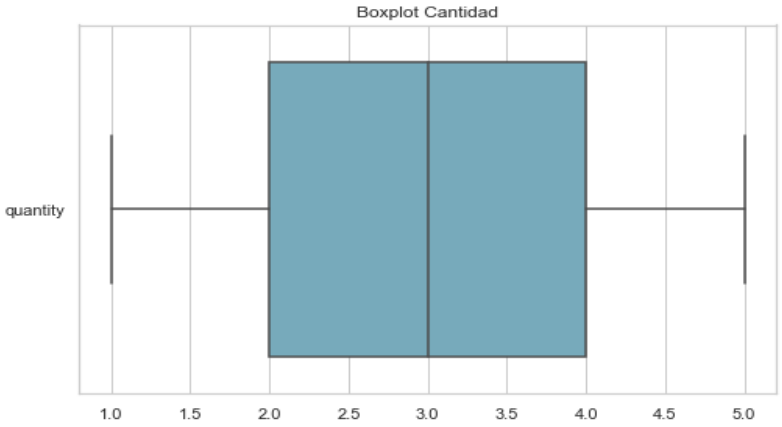

Figura 15. Diagrama de cantidad de productos comprados en el conjunto de datos.

4.2.4. Transformaciones

Agrupación de columnas de bajo porcentaje. El conjunto de datos cuenta con 8 categorías, las cuales no están distribuidas de forma uniforme (Ver **¡Error! No se encuentra el o rigen de la referencia.**). Debido a esto, se procede a realizar un análisis de las variables que no tengan suficientes datos para aportar al modelo y así poder saber si se deben eliminar o agrupar para balancear el conjunto de datos.

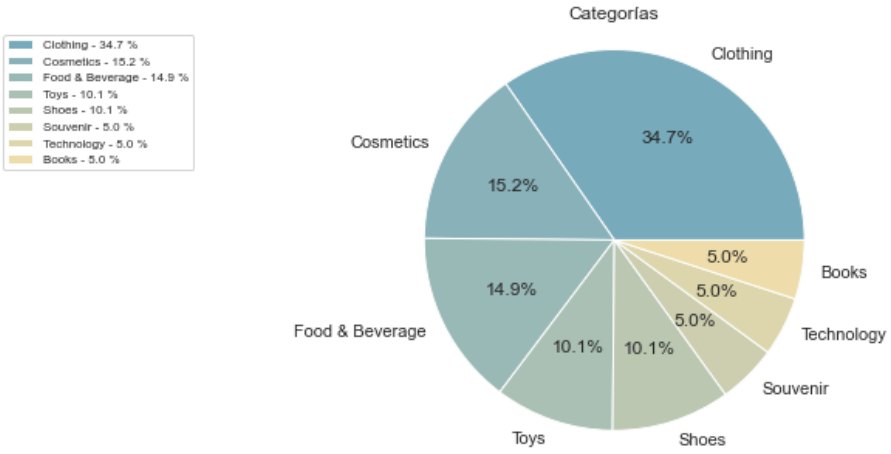

Figura 16. Distribución de las transacciones por categoría de los productos.

Libros, tecnología y souvenir sólo ocupan un 5% de los datos, por lo que se decide eliminar la categoría de tecnología y agrupar las dos restantes. Lo anterior debido a que estas pueden

contener características similares, mientras que la característica que se elimina no comparte alguna similitud con las demás variables. Después de hacer este proceso, se tiene un conjunto de datos con 6 características las cuales están distribuidas de manera más uniforme (Ver Figura 17).

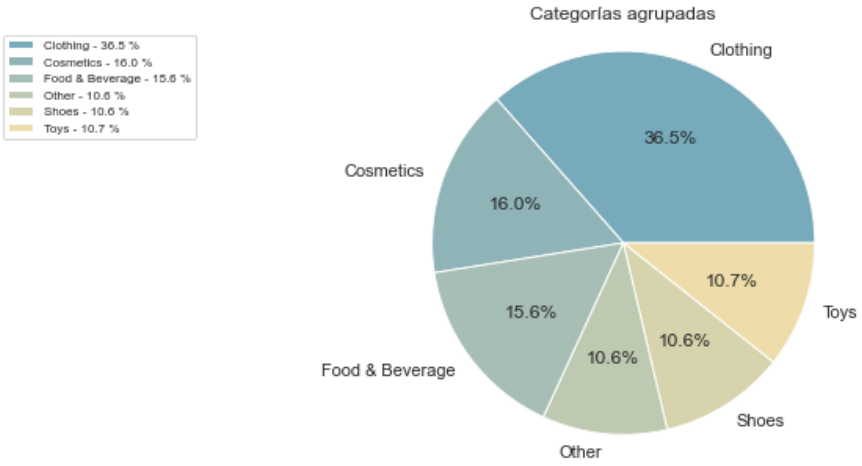

Figura 17. Distribución de las transacciones por categorías agrupadas.

Creación de nuevas columnas. La necesidad de utilizar get_dummies en el preprocesamiento de datos surge principalmente por la presencia de variables categóricas en conjuntos de datos. Los modelos de machine learning y los algoritmos estadísticos requieren que estas variables sean representadas de manera numérica para su procesamiento. La función get_dummies se convierte en una herramienta fundamental para transformar estas variables, permitiendo la conversión de variables categóricas en una forma numérica que los modelos puedan utilizar eficazmente. Además, esta transformación garantiza que las categorías no introduzcan un orden o jerarquía innecesarios, evitando sesgos en el análisis y proporcionando una representación fiel de las relaciones entre las variables categóricas y la variable objetivo. De esta, se realiza el proceso sobre las características categóricas, además, se elimina la "cantidad" del conjunto de datos al estar correlacionada linealmente con "precio". Así, el conjunto de datos para el entrenamiento de los modelos resulta con 14 variables.

4.3. Modelos implementados

4.3.1. Naive Bayes

Este clasificador se basa en el teorema de Bayes, que describe la relación entre probabilidades condicionales. En la clasificación bayesiana, se busca la probabilidad de una etiqueta dadas ciertas características $P(E \mid c)$, donde E es la etiqueta y c son las características. El teorema de Bayes muestra cómo expresarlo en términos de cantidades que podemos calcular más directamente (Géron, 2017)

El algoritmo Naive Bayes funciona de la siguiente manera:

1. Se calcula la probabilidad de que cada etiqueta sea verdadera.
2. Para cada instancia de datos, se calcula la probabilidad de que cada característica sea verdadera para cada clase.
3. La clase con la probabilidad más alta se asigna a la instancia de datos.
4. La probabilidad de que cada clase sea verdadera se calcula utilizando la fórmula de Bayes:

$$P(E \mid c) = \frac{P(c \mid E)\, P(E)}{P(c)} \quad (1)$$

Dónde,

- $P(E)$ es la probabilidad de que la etiqueta sea verdadera.
- $P(c|E)$ es la probabilidad de que la instancia de datos sea verdadera dado que la etiqueta es verdadera.
- $P(c)$ es la probabilidad de que la instancia de datos sea verdadera.

4.3.2. Árbol de decisión (Decision Tree)

Esta técnica construye de forma recursiva un árbol, comenzando con un conjunto de datos y una raíz. En cada paso, el árbol se divide en dos subconjuntos en función de la característica que mejor separa las clases. Este proceso se repite hasta que cada subconjunto contiene solo una clase (Agrawal, 2021)

1. Inicializar el árbol con una raíz. La raíz representa el conjunto de datos completo.

2. Dividir el conjunto de datos en dos subconjuntos. Para dividir el conjunto de datos, se utiliza una función de división. La función de división debe elegir la característica que mejor separa las clases.
3. Repetir los pasos 2 y 3 para cada subconjunto. El proceso se repite hasta que cada subconjunto contiene solo una clase.
4. Asociar una clase a cada hoja. La clase de una hoja es la clase más frecuente en el subconjunto de datos que la representa.

4.3.3. Random Forest

Un bosque aleatorio es un conjunto de árboles de decisión, cada uno de los cuales se construye a partir de un subconjunto aleatorio de los datos de entrenamiento. Para predecir la clase de una instancia de datos, un bosque aleatorio calcula la predicción de cada árbol de decisión y luego vota por la clase más popular. Este es un método de ensamble, también conocidos como métodos combinados. El modelo está basado en dos conceptos importantes:

- **Ensemble Learning.** El aprendizaje conjunto implica combinar múltiples modelos de aprendizaje automático para mejorar el rendimiento y la capacidad de generalización del modelo final. En lugar de depender de un solo modelo, se utilizan varios modelos y se combinan sus predicciones para obtener resultados más robustos y precisos.
- **Bagging.** Es una técnica de aprendizaje conjunto que se utiliza para reducir el sobreajuste de los modelos individuales. Consiste en crear múltiples estimadores (modelos) que entrenan en diferentes conjuntos de datos, donde cada conjunto se obtiene mediante el muestreo aleatorio con reemplazo de los datos de entrenamiento originales. Luego, se promedian las predicciones de estos modelos individuales para obtener una predicción más robusta y generalizada.
- **Boosting.** Es un término que se refiere a cualquier método de conjunto (ensemble) que combina varios "aprendices débiles" (weak learners) para crear un "aprendiz fuerte" (strong learner). Los aprendices débiles son modelos de aprendizaje automático simples, pero aún no son muy precisos por sí mismos. El objetivo del boosting es convertir estos modelos débiles en un modelo fuerte y preciso combinándolos de manera inteligente.

La idea general detrás de la mayoría de los métodos de boosting es entrenar a los predictores de manera secuencial, donde cada uno intenta corregir los errores de su predecesor.

4.3.4. Adaptive Boost (AdaBoost)

Para construir un adaboost se require:

1. Entrenamiento del Primer Modelo (Aprendiz Débil): Comienza con un primer modelo base, como un árbol de decisión simple, que se entrena en el conjunto de datos de entrenamiento. Este modelo hace predicciones iniciales sobre el conjunto de datos.
2. Actualización de los Pesos de las Instancias: AdaBoost evalúa el rendimiento del primer modelo y aumenta el peso de las instancias que fueron clasificadas incorrectamente. Esto significa que las instancias que el primer modelo "subajustó" o no pudo clasificar correctamente obtendrán un peso más alto. La idea detrás de esto es que el próximo modelo se enfocará en corregir las instancias difíciles.
3. Entrenamiento del Segundo Modelo: Con los pesos actualizados, se entrena un segundo modelo. Este segundo modelo intenta corregir las instancias que el primer modelo no pudo clasificar correctamente. De nuevo, este modelo es un aprendiz débil y hace predicciones.
4. Actualización de Pesos nuevamente: Los pesos se actualizan nuevamente en función de cómo se desempeñó el segundo modelo. Las instancias que todavía se clasifican incorrectamente obtienen pesos más altos, y el proceso se repite.
5. Repeticiones: Los pasos 3 y 4 se repiten para un número predeterminado de veces o hasta que se cumpla un cierto criterio de parada. Cada nuevo modelo se entrena para enfocarse en las instancias difíciles que los modelos anteriores no pudieron manejar.
6. Combinación de Modelos: Finalmente, todas las predicciones de los modelos se combinan para formar el modelo final. Las instancias que fueron más difíciles de clasificar reciben un peso mayor en la predicción final.

4.3.5. Redes Neuronales

Un perceptrón es uno de los tipos más simples de una red neuronal, ideal para clasificar datos que son linealmente separables. Consiste en una sola neurona con pesos ajustables y un sesgo (bias), que aprende de los datos de entrenamiento. Cada entrada en el perceptrón está asociada con un

peso que representa su importancia. Cuando una entrada llega al perceptrón, se multiplica por su peso correspondiente y se suman todas estas entradas ponderadas. A esta suma se le puede añadir un término de sesgo para ajustar el umbral de activación. La función de activación es una operación matemática aplicada a la suma ponderada de las entradas y el sesgo. Transforma la entrada del perceptrón a una salida deseada, por ejemplo, 0 o 1 para problemas de clasificación binaria. Una función de activación comúnmente utilizada es la función sigmoide, que comprime los valores de entrada a un rango entre 0 y 1, lo que es útil para probabilidades o clasificaciones binarias. El entrenamiento del perceptrón ajusta los pesos y el sesgo en respuesta a los errores en las predicciones, utilizando reglas como la regla de aprendizaje del perceptrón o algoritmos más avanzados como el descenso del gradiente. Este proceso se repite con múltiples ejemplos de entrenamiento hasta que el modelo minimiza el error y mejora la precisión de sus predicciones. (Fausett, 1994).

4.4. Métricas

4.4.1. Matriz de confusión.

Esta es una medida de rendimiento para los problemas de clasificación del aprendizaje automático donde el resultado puede ser dos o más clases. Se define como la tabla que se utiliza a menudo para describir el rendimiento de un modelo de clasificación en un conjunto de datos de prueba cuyos valores verdaderos se conocen [4]. A partir de esta matriz (ver Figura 18) se pueden construir otras métricas como: Exactitud, Exhaustividad, Accuracy y Valor F1 score.

Figura 18. Matriz de confusión, donde VP corresponde a los verdaderos positivos, FP a los falsos positivos, FN son los falsos negativos y VN los verdaderos negativos.

4.4.2. Exactitud (Accuracy)

La exactitud mide la proporción de predicciones correctas en general (verdaderos positivos más verdaderos negativos) con respecto a todas las predicciones. Es una de las métricas más utilizadas para evaluar un modelo. la métrica de precisión tiene algunas limitaciones: no funciona bien con clases desequilibradas que pueden tener muchos elementos de la misma clase y pocas otras clases (Brownlee, 2020).

$$accuracy = \frac{VP + VN}{VP + FP + VN + FN}$$

4.4.3. Sensibilidad (Recall)

La recuperación mide la proporción de verdaderos positivos con respecto a todos los casos verdaderamente positivos en los datos. Es una medida de la capacidad del modelo para identificar

todos los casos positivos (VanderPlas, Python Data Science Handbook: Essential Tools for Working with Data., 2022)

$$recall = \frac{VP}{VP + FN}$$

4.4.4. Precisión (Precision)

La precisión mide la proporción de predicciones positivas correctas (verdaderos positivos) con respecto a todas las predicciones positivas (verdaderos positivos + falsos positivos). Es una medida de la exactitud del modelo en las predicciones positivas.

$$precision = \frac{VP}{VP + FP}$$

4.4.5. F1 score

Esta métrica es una combinación de métricas de precisión y recuperación que sirve como componente. La mejor puntuación de F1 es igual a 1, mientras que la peor es 0.

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

5. Resultados y discusión

El conjunto de datos *Customer Shopping Dataset - Retail Sales Data* contiene variables numéricas y categóricas las cuales en su mayoría están distribuidas de forma bastante equilibradas. Lo anterior facilitó el desarrollo del trabajo y el análisis de estas y la implementación de los diferentes modelos ya que, por ejemplo, no se encontraron datos faltantes ni tampoco encontraron datos atípicos.

Se analizaron las compras realizadas por hombres y mujeres, en seis categorías de productos comprados en diez centros comerciales, con tres medios de pago y en un periodo de tiempo estudiado de tres años. Se observo que las mujeres son las que realizan la mayoría de las compras y que los productos más comprados pertenecen a las categorías de ropa, cosméticos, comida y bebidas (ver Figura 19).

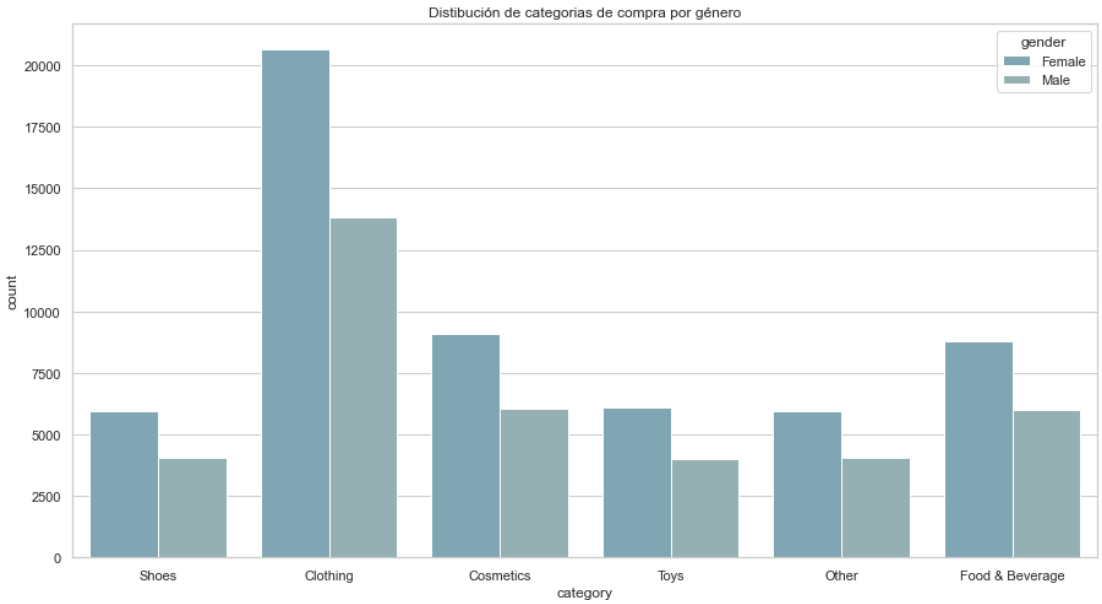

Figura 19. Distribución por categorías y género en el total de compras realizadas.

Otras categorías como la tecnología, los souvenirs y los libros aparecen con el mismo porcentaje de participación en las compras con un 5%. Lo anterior es relevante para el análisis ya que actualmente, todos los productos de tecnología hacen parte de las principales compras de los usuarios a nivel mundial, independiente de la cultura y sus costumbres, fenómeno que tomo

relevancia durante y después de la pandemia, debido al cambio de hábitos de las personas como el trabajo remoto, el uso de las redes sociales que para el 2022 el 80.8% de la población tenía al menos una red social y el incremento en el volumen de ventas a través de comercio electrónico que en Turquía para el 2021 registró un incremento del 69 %, hasta alcanzar los 382.000 millones de liras turcas (más de 36.300 millones de euros según el cambio medio de ese año) (Lázaro, 2022). A pesar de ello, se eliminó la categoría de productos de tecnología porque la cantidad de información que había en el conjunto de datos no era lo suficientemente grande para aportar a los modelos que se iban a desarrollar.

Se analizan, además, las compras realizadas por mujeres y hombres, de acuerdo con la edad y a las categorías de compras (ver Figura 20). Para la realización de este gráfico, se crea una nueva columna que indica el total de compras realizadas, es decir, la cantidad de productos por su precio de venta. En el gráfico, se puede ver de manera comparativa, el total de ventas por categorías y edades que gastan en sus compras tanto hombres como mujeres. Las mujeres entre los 26 y 35 años son las que más compran ropa, mientras que los hombres lo hacen en un rango de edad de 46 a 55 años.

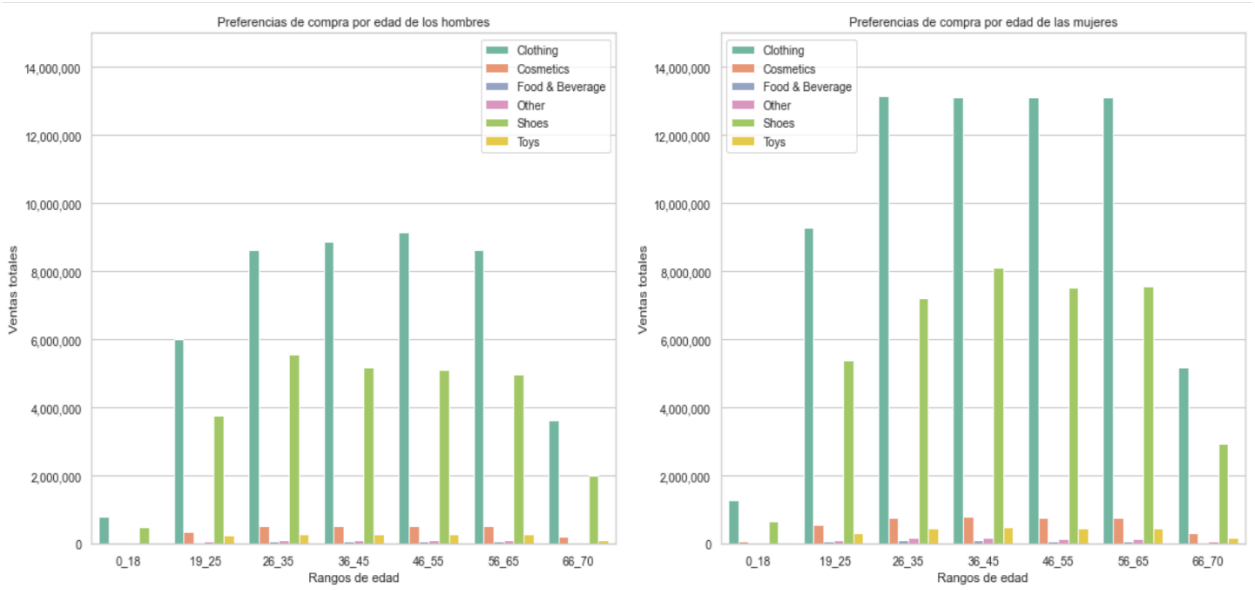

Figura 20. Distribución de compras por categorías y rango de edades.

El conjunto de datos registra tres medios de pago, efectivo, tarjetas debito y tarjetas crédito, siendo el efectivo el medio de pago preferido por los consumidores. Las mujeres en un rango de 36 a 45 años son las que más utilizan este medio de pago, mientras que los hombres que lo utilizan son de 26 a 35 años (ver Figura 21). Lo anterior puede obedecer a las preferencias de pago que existen en Turquía, Las cifras del Banco Central de la República de Türkiye (CBRT) muestran que el efectivo emitido ha aumentado desde 2019. Las emisiones de moneda crecieron más rápidamente durante la pandemia de Covid-19. El efectivo representó el 89% de las transacciones minoristas y el 75% en términos de valor en 2020 (Çevik & Teber, 2021). Por otra parte, las categorías de compras son de productos de bajo monto, lo que podría explicar por que el uso de las tarjetas en las compras es menor al del dinero en efectivo.

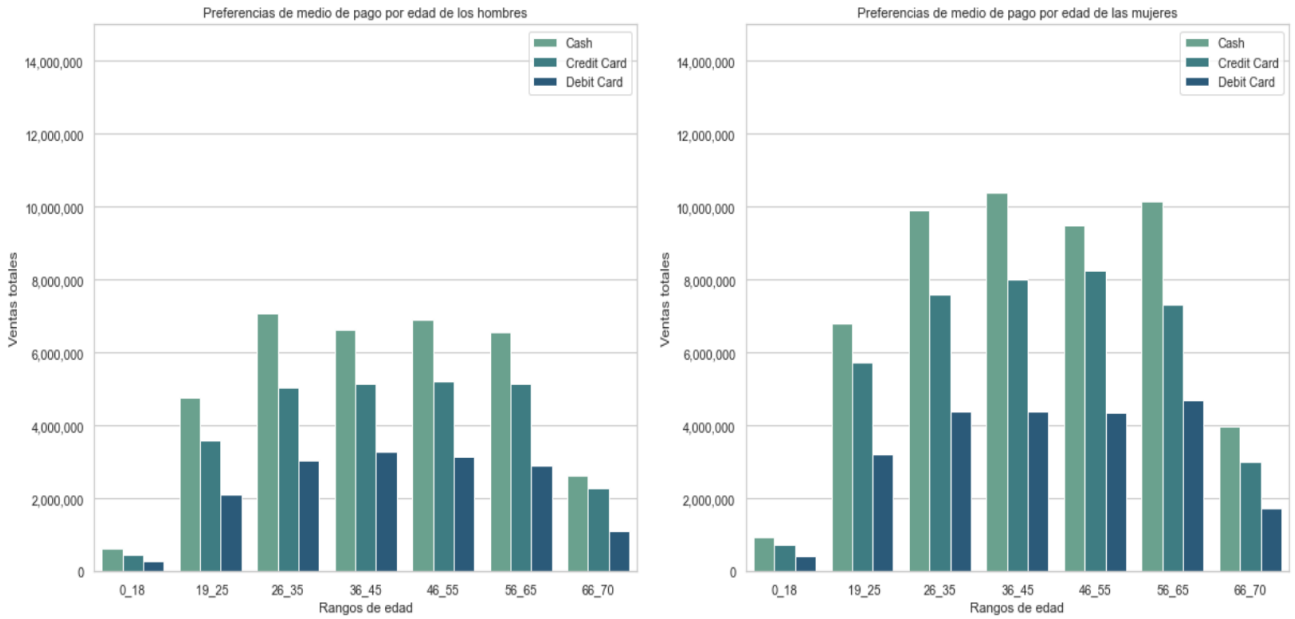

Figura 21. Distribución de compras por medios de pago y rango de edades.

Con respecto a la distribución de la información de los centros comerciales, se evidencia que los centros comerciales llamados Kanyon, Mall Of Istanbul y Metrocity están en la cima del ranking. Zorlu Center y Cevahir están en las categorías inferiores. Se podría suponer que tal vez esta diferencia se deba al tamaño de los centros comerciales, o la cantidad de tiendas disponibles para que las personas hagan sus compras, sin embargo, los centros comerciales que se encuentran en el conjunto de datos, cuentan con un promedio de 150 tiendas, zonas recreativas y de comida con características similares entre ellos. (YILDIZ, 2023)

La aplicación de técnicas de Machine Learning (ML) en la clasificación múltiple para categorías de ventas es fundamental en la optimización y la toma de decisiones estratégicas en un negocio. Los modelos aplicados a datos de ventas pueden prever qué productos o servicios tendrán una mayor demanda. Esto conduce a una mejor planificación del inventario y evitar el exceso de stock o la escasez. Además, la clasificación precisa facilita la personalización del marketing, dirigido a incrementar la eficacia de las campañas publicitarias y promociones basadas en predicciones de tendencias de compra. Con esto como objetivo, se implementaron diferentes modelos para el desarrollo de esta monografía, los cuales arrojaron diferentes puntajes (ver Tabla 6).

	Modelo	Scores
4	RandomForestClassifier	1.000000
5	NeuralNetwork	0.779019
0	GaussianNB	0.775367
6	AdaBoostClassifier	0.706452
3	DecisionTreeClassifier	0.617477
2	BernoulliNB	0.427301
1	MultinomialNB	0.364209

Tabla 6. Resultados de los modelos implementados.

Inicialmente, se implementa el algoritmo clasificador Naïve-Bayes (NBC), que es un clasificador probabilístico simple con fuerte suposición de independencia. Aunque la suposición de la independencia de los atributos es generalmente una suposición pobre y se viola a menudo para los conjuntos de datos verdaderos. (Mosquera, Castrillón, & Parra, 2018). Los resultados de precisión (accuracy) para los tres modelos de clasificación bayesiana indican que el Gaussian Naive Bayes (modelGNB) con un 77.54% de precisión es el más efectivo de los tres para el conjunto de datos específico con el que se trabajó. Esto sugiere que los datos se ajustan bien a una distribución gaussiana, la cual es asumida por este modelo. Por otro lado, el Bernoulli Naive Bayes (modelBNB) y el Multinomial Naive Bayes (modelMNB) muestran una precisión considerablemente más baja,

con 42.73% y 36.42% respectivamente, lo que podría indicar que la suposición de una distribución de Bernoulli o multinomial no es adecuada para estos datos, o que el modelo podría no estar bien calibrado o carece de suficientes datos para entrenarse efectivamente.

El modelo Random Forest Classifier, entendido como un conjunto de Árboles de Decisión, generalmente entrenados a través del método de bagging (o a veces pasting), típicamente con max_samples configurado al tamaño del conjunto de entrenamiento (Géron, 2017), fue el que arrojo el resultado de precisión más alto, 100% con hiperparametros de: 'max_depth': 9, 'max_features': 12, 'n_estimators': 20. Debido a estos resultados, se realiza un análisis de sobreajuste (Overfitting) ya que este rendimiento podría sugerir la posibilidad de sobreajuste, donde el modelo se ajusta demasiado a los datos de entrenamiento y no generaliza bien a nuevos datos. El modelo podría estar memorizando el conjunto de datos de entrenamiento en lugar de aprender patrones generales, lo que podría afectar su capacidad para predecir con precisión datos completamente nuevos. Se realiza, entre otros análisis, validación cruzada que puede ayudar a evaluar la capacidad del modelo para generalizar a datos no vistos. Si el modelo muestra un rendimiento del 100% en múltiples divisiones de datos, es más probable que sea generalizable.

Los resultados de la validación cruzada muestran que para cada pliegue (fold) o división del conjunto de datos, el modelo Random Forest ha logrado una precisión (accuracy) perfecta de 1.0, lo que indica que ha predicho correctamente todas las muestras en cada uno de los pliegues. La medida de la precisión promedio (mean acc) es también del 100%, lo que sugiere que el modelo tiene un rendimiento excelente y ha sido capaz de generalizar bien a datos no vistos en cada división del conjunto de datos durante la validación cruzada. Un resultado de validación cruzada perfecto (1.0 o 100%) en cada pliegue implica que el modelo es muy preciso y se desempeña de manera consistente en diferentes particiones del conjunto de datos. (Ver Tabla 7) Esto es un indicio bastante fuerte de que el modelo es altamente efectivo en predecir los datos de forma general, aunque siempre es importante tener en cuenta la posibilidad de sobreajuste si no se ha realizado una validación adecuada en un conjunto de datos de prueba independiente.

```
cross_val_score --> fold 1: 1.0
cross_val_score --> fold 2: 1.0
cross_val_score --> fold 3: 1.0
cross_val_score --> fold 4: 1.0
cross_val_score --> fold 5: 1.0
mean acc: 1.0
```

Tabla 7. Validación cruzada del modelo Random Forest

Se implementó el modelo Decision Tree Classifier, y se obtuvo un valor de precisión del 56% el cual indica que el modelo es capaz de predecir correctamente el 56% de las muestras en el conjunto de datos de prueba. Si bien esta precisión no es tan alta como la que se obtuvo con el modelo Random Forest (100%), aún es significativa. Este resultado se obtuvo bajo los hiperparámetros: ccp_alpha': 0.001, 'criterion': 'entropy', 'max_depth': 9, 'max_features': 'auto', 'min_samples_leaf': 2, 'min_samples_split': 5. Se ha utilizado 'entropy' como criterio para medir la calidad de las divisiones del árbol de decisión, que se basa en la teoría de la información de Shannon. Max_depth (profundidad máxima del árbol) para limitar la profundidad del árbol a 9 niveles y ayudar a prevenir el sobreajuste y simplificar el modelo y se fija un número mínimo de muestras requeridas para dividir un nodo interno: 5, lo que ayuda a evitar divisiones en nodos con muy pocas muestras. Posterior a ellos, se evalúa nuevamente el modelo basado en los datos de prueba con el modelo y se obtiene un resultado de 61.75 %. Esto significa que, al utilizar el modelo ajustado con los mejores parámetros en un conjunto de datos de prueba separado, se logró predecir correctamente el 61.75 % de las muestras. Esta precisión del 61.75 % muestra un ligero aumento respecto al 56 % obtenido anteriormente. Esto sugiere que el modelo ajustado con los hiperparámetros optimizados está funcionando un poco mejor al predecir las categorías de productos en datos que no fueron utilizados para el entrenamiento del modelo.

La red neuronal construida con el framework de tensorflow consta de las siguientes etapas secuenciales:

1. Sequential (): Inicializa una red neuronal secuencial, que es un conjunto lineal de capas.

2. Primera capa Dense: Añade una capa densa (completamente conectada) con 256 neuronas. La función de activación 'relu' (unidad lineal rectificada) se usa para añadir no linealidad al modelo. El parámetro input_shape se establece para coincidir con el número de características de x_train.

3. BatchNormalization (): Normaliza las activaciones de la capa anterior, lo que puede llevar a una convergencia más rápida durante el entrenamiento y una mayor estabilidad general.

4. Dropout (0.5): Aplica el Dropout, que es una técnica de regularización donde aleatoriamente "apaga" un porcentaje de neuronas (en este caso, el 50 %) durante el entrenamiento para prevenir el sobreajuste.

Las siguientes capas repiten este patrón de una capa Dense seguida por Batch Normalization y Dropout, pero con un número decreciente de neuronas (128, 64, 32). Esto forma una arquitectura que se estrecha, una práctica común en el diseño de redes neuronales (ver Figura 22). Finalmente, la última capa Dense tiene 6 neuronas con una activación 'softmax', lo que indica que el modelo está diseñado para clasificar las entradas en una de seis categorías. La función softmax asegura que la salida de la red neuronal son probabilidades que suman uno, lo que hace que sea adecuado para problemas de clasificación multiclase.

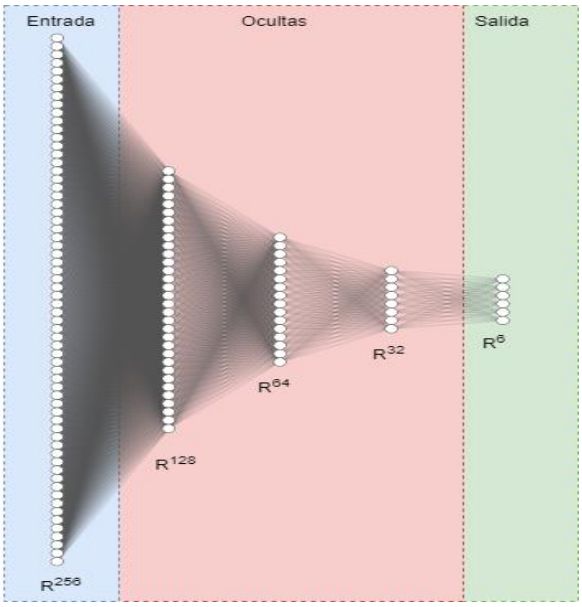

Figura 22. Arquitectura de red neuronal.

Para el despliegue de la información (ver Figura 23), se desarrolla una aplicación que funciona como una herramienta de clasificación de categorías de ventas diseñada para predecir la categoría de una venta basada en características específicas del cliente y la transacción. La

funcionalidad permitiría a los usuarios ingresar datos como la edad, el precio del artículo comprado, el género del cliente, el método de pago utilizado y el centro comercial donde se realizó la compra. Luego, al enviar esta información, la aplicación proporcionaría una predicción de la categoría de venta correspondiente.

Esta herramienta sería útil en entornos de retail o e-commerce para analizar y entender mejor las tendencias de compra y personalizar la experiencia de compra. Por ejemplo, podría ayudar a los minoristas a identificar qué productos son populares entre diferentes grupos demográficos o qué métodos de pago prefieren los clientes, lo que puede influir en las decisiones de marketing y stock de inventario. Además, el botón "Flag" podría usarse para marcar resultados inusuales o incorrectos, posiblemente para mejorar el modelo de machine learning subyacente mediante un aprendizaje activo o una revisión humana.

Figura 23. Aplicación para la clasificación de categorías de ventas.

6. Conclusiones

El análisis exploratorio de datos de las transacciones de clientes muestra una distribución equitativa por género y una preferencia desigual entre las categorías de productos,

lo que sugiere oportunidades de marketing dirigido y ajustes en el inventario. Los métodos de pago y la afluencia a diferentes centros comerciales reflejan hábitos de consumo y confianza en las opciones de pago, ofreciendo datos valiosos para estrategias de venta basadas en la ubicación.

En cuanto a la demografía de los clientes, la edad promedio indica un mercado principal mientras que la cantidad de artículos por compra y la distribución de precios sugieren que las compras tienden a ser de bajo volumen y costo. Los datos atípicos en precio pueden requerir una revisión adicional para garantizar la integridad del análisis y podrían señalar oportunidades de mercado o errores de entrada de datos.

Los resultados de precisión (accuracy) de los modelos aplicados al problema de clasificación de compras de clientes indican que el RandomForestClassifier ha obtenido una precisión perfecta de 1.00. Esto podría sugerir un sobreajuste, ya que es inusual que los modelos de aprendizaje automático alcancen una precisión del 100% en datos no vistos a menos que el problema sea extremadamente simple o los datos sean muy consistentes. Es crucial validar este resultado con un conjunto de datos de prueba no utilizado durante el entrenamiento o la validación cruzada para confirmar que el modelo generaliza bien.

El NeuralNetwork y GaussianNB (Naive Bayes gaussiano) muestran resultados similares con precisiones de aproximadamente 0.78 y 0.77, respectivamente, lo que los hace modelos razonablemente buenos, pero con margen de mejora. AdaBoostClassifier reduce la precisión a 0.71, lo que aún es respetable, pero destaca la variación en la eficacia de diferentes algoritmos de aprendizaje automático.

DecisionTreeClassifier muestra una precisión significativamente menor de 0.62, lo que podría indicar que el modelo es demasiado simple o no se ha ajustado correctamente. Por último, BernoulliNB y MultinomialNB (Naive Bayes de Bernoulli y multinomial) tienen un rendimiento mucho menor, con precisiones de 0.43 y 0.36, respectivamente, lo que podría ser resultado de la suposición de independencia de características que hacen estos modelos y que raramente se cumple en datos reales.

El uso de GridSearch y validación cruzada asegura que la selección de hiperparámetros y la evaluación de los modelos sean robustas y menos propensas al sobreajuste. Sin embargo, la gran disparidad en las precisiones sugiere que ciertos modelos son mucho más adecuados para este conjunto de datos específico. Sería importante revisar la distribución y características de los datos para entender por qué algunos modelos tienen un rendimiento muy superior a otros y asegurarse de que los resultados del RandomForestClassifier sean verificables y no un artefacto de sobreajuste.

El despliegue de la aplicación de clasificación de categorías de ventas es un ejemplo destacado de cómo los pipelines de MLOps pueden facilitar la interacción entre modelos de machine learning y usuarios finales. Utilizando Gradio, una herramienta que permite crear interfaces amigables para modelos de ML, se ha configurado un demo interactivo. Cada modelo, que ha sido entrenado y optimizado con técnicas de GridSearch y validación cruzada, puede recibir datos completamente nuevos a través de la interfaz y proporcionar una clasificación en tiempo real. Esto no solo mejora la experiencia del usuario, sino que también permite tener feedback instantáneo sobre el rendimiento de los modelos en situaciones del mundo real.

Referencias

Agrawal, S. K. (2021). *Analytics Vidhya*. Obtenido de https://www.analyticsvidhya.com/blog/2021/07/metrics-to-evaluate-your-classification-model-to-take-the-right-decisions/#h-classification-metrics-in-machine-learning

Brownlee, J. (2020). *Machine learning mastery*. Obtenido de https://machinelearningmastery.com/metrics-evaluate-machine-learning-algorithms-python/

Çevik, S., & Teber, D. (2021). *The Determinants of Consumer Cash Usage in Turkey.* Obtenido de Central Bank of the Republic of Turkey: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.tcmb.gov.tr/wps/wcm/connect/992bb543-8cac-4079-9590-3eb145411ebf/wp2135.pdf?MOD=AJPERES&CACHEID=ROOTWORKSPACE-992bb543-8cac-4079-9590-3eb145411ebf-nTIUJUF

Fausett, L. V. (1994). *Fundamentals of neural networks.* Prentice-Hall. Obtenido de chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://dl.matlabyar.com/siavash/Neural%20Network/Book/Fausett%20L.-Fundamentals%20of%20Neural%20Networks_%20Architectures,%20Algorithms,%20and%20Applications%20(1994).pdf

Géron, A. (2017). *Hands-on Machine Learning with Scikit-Learn, Keras, and Tensorflow-Concepts, Tools, and Techniques to Build Intelligent Systems.* O'Reilly Media, Inc.,.

Kaggle. (17 de 10 de 2023). *Exploring Market Basket Analysis in Istanbul Retail Data*. Obtenido de Kaggle: https://www.kaggle.com/datasets/mehmettahiraslan/customer-shopping-dataset

Lázaro, P. N. (2022). *ICEX*. Obtenido de Informe e-País: El comercio electrónico en Turquía : chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://www.icex.es/content/dam/es/icex/oficinas/033/documentos/2022/11/documentos-anexos/DOC2022917675.pdf

Mosquera, R., Castrillón, O., & Parra, L. (2018). *Scielo*. Obtenido de Support Vector Machines, Naïve Bayes Classifier and Genetic Algorithms for the Prediction of Psychosocial Risks in

Teachers of Colombian Public Schools.: https://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-07642018000600153

Pacheco Melo, F., & Rojas Vilches, J. (2014). *Selección de variables de entrada para procesos autorregresivos, mediante el cálculo de la información mutua usando los k-vecinos más cercanos.*

Quintero, J. P. (2017). *TABLAS DE CONTINGENCIA PARA LA ENSEÑANZA DEL CONCEPTO DE ASOCIACIÓN ENTRE VARIABLES ALEATORIAS CUALITATIVAS.* Obtenido de https://repositorio.unal.edu.co/handle/unal/62305

Restrepo, L., & González, J. (2007). De Pearson a Spearman. *Revista Colombiana de Ciencias Pecuarias*. Obtenido de https://www.redalyc.org/pdf/2950/295023034010.pdf

Ronald E. Walpole, Raymond H. Myers, Sharon L. Myers, & Keying Ye. (2007). *Probability & Statistics for Engineers & Scientists.* 9th Edition, Pearson Education, Inc.

VanderPlas, J. (2016). *Python Data Science Handbook. Essential Tools for Working with Data.* O'Reilly Media.

VanderPlas, J. (2022). *Python Data Science Handbook: Essential Tools for Working with Data.* O'Reilly Media.

YILDIZ, P. (Noviembre de 2023). *Istanbul Tourist information*. Obtenido de https://istanbul-tourist-information.com/es/guia-de-compras-de-estambul/#t-1632978577123