



**Predicción del comportamiento de compra en tiendas de barrio utilizando modelos de machine learning**

Jhon Esteban Arredondo Parra

Gerson Gabriel Erira Sánchez

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

Universidad de Antioquia  
Facultad de Ingeniería  
Especialización en Analítica y Ciencia de Datos  
Medellín, Antioquia, Colombia  
2025

|                           |                                                                                                                                                     |
|---------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Cita</b>               | (Arredondo Parra & Erira Sánchez, 2025)                                                                                                             |
| <b>Referencia</b>         | Arredondo Parra, J. & Erira Sánchez, G. (2025). Predicción del comportamiento de compra en tiendas de barrio utilizando modelos de machine learning |
| <b>Estilo IEEE (2020)</b> | Universidad de Antioquia, Medellín, Colombia.                                                                                                       |



Especialización en Analítica y Ciencia de Datos, Cohorte VIII.

Centro de Investigación Ambientales y de Ingeniería (CIA).



## TABLA DE CONTENIDO

|                                                                         |    |
|-------------------------------------------------------------------------|----|
| RESUMEN.....                                                            | 8  |
| ABSTRACT .....                                                          | 9  |
| 1. INTRODUCCIÓN .....                                                   | 10 |
| 2. FUENTES DE DATOS .....                                               | 14 |
| 3. PROCESAMIENTO DE DATOS (METODOLOGÍA) .....                           | 17 |
| 3.1. Generación base de datos Compras Categorizadas .....               | 18 |
| 3.2. Generación de Dataset principal .....                              | 18 |
| 3.3. Selección de características y generación de dataset final .....   | 20 |
| 3.4. Predicción de ventas .....                                         | 21 |
| 3.5. Métricas de desempeño .....                                        | 22 |
| 4. RESULTADOS.....                                                      | 22 |
| 4.1. Resultados del modelo de clasificación y generación BD final ..... | 22 |
| 4.2. Analítica descriptiva.....                                         | 23 |
| 4.3. Limpieza y organización de datos para series de tiempo .....       | 25 |
| 4.4. Resultados del modelo de regresión.....                            | 27 |
| 4.4.1. Primer acercamiento modelos.....                                 | 27 |
| 4.4.2. Variación de Modelos .....                                       | 28 |
| 5. CONCLUSIONES .....                                                   | 31 |
| REFERENCIAS .....                                                       | 34 |
| ANEXOS.....                                                             | 37 |
| Anexo 1. Resumen revisión de la literatura modelos analíticos.....      | 37 |
| Anexo 2. Comparación y Selección modelo de clasificación .....          | 38 |
| Anexo 3. Detalle de columnas, tipología y cantidad de registros .....   | 40 |

---

|                                                                              |    |
|------------------------------------------------------------------------------|----|
| Anexo 4. Comparación métricas primer acercamiento modelos de regresión ..... | 43 |
| Anexo 5. Comparación métricas desempeño modelos de regresión .....           | 43 |

---

**LISTA DE TABLAS**

|                                                       |    |
|-------------------------------------------------------|----|
| <b>Tabla 1.</b> Descripción de tablas originales..... | 15 |
|-------------------------------------------------------|----|

## LISTA DE FIGURAS

|                                                                                         |    |
|-----------------------------------------------------------------------------------------|----|
| <b>Figura 1.</b> Modelo de Clasificación .....                                          | 17 |
| <b>Figura 2.</b> Modelo entidad relación para el dataset conformado .....               | 19 |
| <b>Figura 3.</b> Métricas modelo LinearSVC de clasificación de productos.....           | 23 |
| <b>Figura 4.</b> Distribución de cantidades vendidas por categoría de producto .....    | 24 |
| <b>Figura 5.</b> Ejemplo distribución de cantidades vendidas por subregión.....         | 24 |
| <b>Figura 6.</b> Distribución de cantidades vendidas por subcategoría sin atípicos..... | 26 |
| <b>Figura 7.</b> Distribución de variables socioeconómicas.....                         | 26 |
| <b>Figura 8.</b> Predicción primeros modelos regresión (red neuronal).....              | 28 |
| <b>Figura 9.</b> Predicciones modelos de regresión con relleno en cero .....            | 29 |
| <b>Figura 10.</b> Predicciones modelos de regresión con interpolación lineal .....      | 30 |

---

### **Siglas, acrónimos y abreviaturas**

|        |                                          |
|--------|------------------------------------------|
| ARIMA  | AutoRegressive Integrated Moving Average |
| BD     | Base de Datos                            |
| DL     | Deep Learning                            |
| IQR    | Interquartile Range                      |
| LSTM   | Long Short-Term Memory                   |
| MAE    | Mean Absolute Error                      |
| MAPE   | Mean Absolute Percentage Error           |
| ML     | Machine Learning                         |
| OHE    | One-Hot Encoding                         |
| R2     | Coeficiente de Determinación             |
| RANSAC | Random Sample Consensus                  |
| RMSE   | Root Mean Squared Error                  |
| SVMR   | Support Vector Machine Regression        |
| UdeA   | Universidad de Antioquia                 |

---

## RESUMEN

El proyecto busca evaluar el comportamiento de compra a nivel local, identificando los productos de la canasta familiar con mayor demanda, la base de datos principal se obtiene de la plataforma Kaggle, la cual contiene registros de compras de productos en 2022 para 217 municipios en 5 departamentos de Colombia; fue necesario complementar la base de datos con múltiples fuentes externas (plataforma Kaggle, Dane y Fuente propia) que se describen a lo largo de la monografía.

Se pretende implementar modelos de Machine Learning, con el fin de ayudar en la identificación de preferencias y demanda frente a distintas categorías de productos para poder proyectar un comportamiento de consumo. Esto proporcionará información relevante a los establecimientos comerciales para que puedan crear estrategias que les permita entrar en mercados potenciales y mejorar sus ventas.

Para la implementación de modelos de regresión, previamente se entrenaron distintos modelos de clasificación de productos en categorías, para alimentar el dataset final de predicción. El mejor modelo de clasificación fue el SVC que alcanza una precisión del 96% en la asignación de categorías. Para la predicción de demanda, se ajustan modelos de regresión tanto de machine learning como de Deep learning, sobre un dataset alimentado con variables socioeconómicas, categorías de productos y dos métodos de imputación distintos. Con el primer método de imputación con ceros, los modelos presentan un RMSE de 6 unidades. Para el segundo caso, se usa interpolación lineal obteniendo métricas de un MAPE cercano al 40% y un RMSE de aproximadamente 16 unidades.

<https://github.com/Gabriel93UdeA/AprendizajeAutomatico/tree/main>

*Palabras clave:* mercado minorista, preferencias consumidor, aprendizaje supervisado, predicción, clasificación, ventas.

---

## ABSTRACT

The main database is obtained from the Kaggle platform, which contains records of product purchases in 2022 for 217 municipalities in 5 departments of Colombia; it was necessary to complement the database with multiple external sources (Kaggle platform, Dane and own source) that are described throughout the monograph.

It is intended to implement Machine Learning models, in order to help in the identification of preferences and demand against different product categories to be able to project consumption behavior. This will provide relevant information to commercial establishments so that they can create strategies that allow them to enter potential markets and improve their sales.

For the implementation of regression models, different product classification models were previously trained to classify products into categories, to feed the final prediction dataset. The best classification model was the SVC, which achieved an accuracy of 96% in the assignment of categories. For demand prediction, both machine learning and deep learning regression models are fitted on a dataset fed with socioeconomic variables, product categories and two different imputation methods. With the first imputation method with zeros, the models present an RMSE of 6 units. For the second case, linear interpolation is used, obtaining metrics with a MAPE close to 40% and an RMSE of approximately 16 units.

*Keywords:* retail market, consumer preferences, supervised learning, prediction, classification, sales.

## 1. INTRODUCCIÓN

Una sociedad compleja con consumidores cada vez más exigentes frente a una demanda de productos y servicios conlleva a que las compañías realicen grandes inversiones de dinero y movilización de recursos humanos [1], buscando mejorar aspectos como el control de inventario [2], producción [3], logística [4], etc. Por ejemplo, una previsión eficaz de la demanda puede ahorrar más del 7 % en los gastos operativos anuales de una empresa e impacta positivamente en la experiencia de compra del consumidor final [4]. Conocer las tendencias de consumo a nivel local es un importante punto de partida para la creación de estrategias que los establecimientos comerciales puedan utilizar para entrar en mercados potenciales donde aún no tienen cobertura, o impulsar aquellos que evidencian un crecimiento significativo [5].

Aunque cada vez sean más las grandes cadenas de supermercados, algunos consumidores no descartan la idea de realizar sus compras en supermercados pequeños y tiendas de barrio. Según los resultados de una encuesta realizada en la ciudad de Bogotá, a personas de estrato 2 y 3, se encontró que los principales atributos que los usuarios prefieren en las tiendas de barrio son la oferta de unidades en presentaciones pequeñas, menores gastos de transporte y la posibilidad de negociar, convirtiéndose en un motor de desarrollo en muchas zonas del país [6]. Junto con lo anterior y como se menciona en la publicación de Espinel [7] los 2 lugares más frecuentes de adquisición de bienes de consumo en Colombia son: las tiendas de barrio (46%) y las grandes superficies (22%), evidenciando la gran importancia que tienen estas pequeñas superficies para los consumidores.

Las pequeñas superficies tienen puntos a mejorar como conseguir una amplia variedad de productos según la necesidad de los consumidores [8]; adicional a lo anterior y como lo menciona Ramírez [9] un punto clave es que los tenderos podrían fortalecer la relación con sus consumidores si implementan investigaciones enfocadas en comprender mejor las preferencias de sus clientes, con esta información podrían adaptarse mejor a las necesidades, ofreciendo un servicio personalizado que compita de manera más efectiva con los grandes supermercados.

El continuo cambio de las conductas de consumo [10] es una brecha que los empresarios deben afrontar, fomentando la implementación de tecnología, definiendo estrategias de fidelización de clientes, abriendo nuevos negocios y optimizando canales de distribución que permita la llegada de productos en un tiempo y cantidades exactas para satisfacer las necesidades de sus consumidores; además la carrera constante de la industria en cubrir estas necesidades, ha llevado a que las empresas generen distintas estrategias enfocadas en adaptarse [11]. Las compañías líderes en el sector optaron por presentar una amplia variedad de productos, otorgando a los consumidores diversidad en colores, calidad, precio, etc. [12]. Sin embargo, el sector retail no solo lo componen grandes organizaciones, sino que también están esos pequeños establecimientos o tiendas de barrio repartidas por todo el país que tienen presupuestos más ajustados donde la decisión sobre cuántos y cuáles productos tener se debe ajustar a su espacio destinado para almacenamiento y confiar en su conocimiento empírico adquirido sobre las preferencias de sus clientes [13].

Tradicionalmente, se ha intentado resolver el problema mediante múltiples modelos estadísticos basados en series de tiempo. Silva *et al.* [14] afirman que estos modelos tradicionales no permiten alcanzar métricas de precisión adecuadas para el entorno actual, que es de gran incertidumbre y presenta cambios constantes. En línea con la crítica a los métodos tradicionales, Güven y Şimşir [15] proponen el uso de redes neuronales para la predicción de la demanda de distintos productos, dado que estos no necesitan la verificación de supuestos y son más flexibles en la inclusión de variables exógenas. Por ello, es necesario adoptar nuevos paradigmas basados en la modelación de sistemas no lineales y proponen que las redes neuronales y las máquinas de vectores de soporte son ideales para este tipo de problemas.

Con este objetivo las técnicas de Machine Learning (ML) y Deep Learning (DL) han sido bastante usadas en la predicción de situaciones basadas en grandes cantidades de datos, aportando información valiosa en distintos contextos [16]. Dada la importancia que han tomado estas herramientas, el presente trabajo busca construir modelos de ML que aporten a la identificación de posibles preferencias de compra, tendencias de consumo y con esto proyectar un comportamiento por subcategoría de producto.

Una de las implementaciones de ML en el sector alimenticio lo hicieron Chae *et al.* [17] y Punia y Shankar [18], los cuales usaron variables como tiempo, socioeconómicas, datos de ventas, promociones, medios de comunicación, entre otras; teniendo como resultado mejoras en la planificación de la demanda, ajustarse a un entorno con alta incertidumbre y disminución de costos en inventario y transporte. De forma similar, Rodrigues *et al.* [19] demuestran que el uso de modelos predictivos permite reducir entre un 14% y un 52% los productos desperdiciados. Además, estiman que basar las decisiones en estas predicciones los servicios de restaurante podrían reducir la demanda no satisfecha entre un 3% y un 16%.

Dentro de la revisión de proyectos similares se encontró casos de aplicación de modelos analíticos para la estimación de la demanda (ver Anexo 1. Resumen revisión de la literatura modelos analíticos), por ejemplo, Correa y Mejia [20] utilizaron el histórico de pedidos entre enero de 2019 a marzo de 2023 de una distribuidora de artículos de ferretería, para predecir la demanda mensual de los 3 artículos más vendidos utilizando modelos ARIMA, Regresión Lineal, Random Forest y redes neuronales. De forma similar Jaramillo y Gutiérrez [21] utilizaron datos abiertos en Kaggle implementando redes neuronales para realizar la predicción de ventas diarias en tiendas de la Corporación Favorita de Ecuador.

A pesar de que algunos de los modelos anteriores permiten el uso de múltiples variables explicativas, los artículos utilizaron solo el tiempo como variable predictora obteniendo métricas de error MAE entre 10 y 1000, mostrando que la variable tiempo no es el único factor que influye en la demanda de estos productos. En ese sentido, no fue explorada la posibilidad de discriminar y predecir ventas por tienda y producto, y tampoco el uso de variables exógenas como el entorno socioeconómico. En nuestro caso, esperamos poder hacer uso de un dataset más amplio que permita evaluar la influencia de este tipo de factores.

Un punto clave en la definición de modelos es el manejo de outliers. Por ejemplo, Hincapie [22] en su trabajo para la implementación de modelos de machine learning para predecir la demanda sobre productos vendidos, se utilizaron los modelos Random Forest Regressor, Plataforma H2O AutoML y H2O AutoML con análisis de tipicidad, los cuales fueron comparados

basándose en su MAPE y su capacidad para predecir la demanda de 17 tipos de productos. El uso de H2O AutoML permite la evaluación automática de múltiples modelos de ML sobre un conjunto de datos, combinando distintos hiperpárametros y facilitando la identificación de los modelos con menor margen de error [23]. Así mismo Lizcano y Romero [24] buscaron predecir las ventas para una cadena de tiendas haciendo uso de algoritmos de inteligencia artificial como lo son Random Forest, Redes Neuronales recurrentes y el modelo RANSAC (RANdom SAmple Consensus), el cual es una técnica de regresión robusta para abordar los outliers [25]. Este último modelo mencionado fue relevante para dar un buen manejo a los outliers presentados en la base de datos, evitando impactar negativamente la construcción de modelos.

Otro ítem importante es la selección de las características a utilizar para el entrenamiento de estos modelos predictivos ya que el dataset final es crítico en el éxito de los mismos. Por ejemplo, Ipia [26] hace uso de técnicas de aprendizaje supervisado para obtener el mejor modelo de clasificación según las ventas de 10 tiendas ubicadas en Estambul, para esto se utiliza y evalúan 5 modelos (Naive Bayes, Árbol de Decisión, Random Forest, Adaptive Boost y Redes Neuronales). Aunque este estudio no está centrado en la predicción de la demanda, su enfoque en clasificación podría ser útil en caso de dificultades en el entrenamiento debido a la naturaleza de los datos y resulte necesario caracterizar el comportamiento de consumo según la ubicación geográfica de las tiendas.

En ese sentido, el presente trabajo analizará el comportamiento de compra de distintos productos de supermercado en 217 municipios de los departamentos de Antioquia, Caldas, Quindío, Risaralda y Valle del Cauca. Con la implementación de modelos de Machine Learning se busca herramientas que aporten en la identificación de preferencias de compra, tendencias de consumo y con esto proyectar un comportamiento por subcategoría de producto; lo anterior considerando el entorno socioeconómico y el histórico de ventas del año 2022. Además, se pretende dar una visión integral del comportamiento de compra de la población en estos departamentos, detectando oportunidades de crecimiento en sectores industriales y comerciales clave.

En la sección 2 se presentan las fuentes de datos utilizadas, provenientes principalmente de conjuntos de datos abiertos y tablas de elaboración propia, dentro de las cuales resalta la recolección de datos para enriquecer la información descargada de Kaggle (técnica WebScraping para adicionar nuevas categorías de productos) y la categorización de productos que fue el resultado de un Modelo de Clasificación, la sección 3 describe el procesamiento de datos, que incluye categorización de productos (sección 3.1. Generación base de datos Compras Categorizadas), generación Dataset (sección 3.2. Generación de Dataset principal), Extracción de características (sección 3.3. Selección de características y generación de dataset final) y que modelos se usaron para predecir las ventas (sección 3.4. Predicción de ventas), en la sección 4 se encuentran los diferentes resultados iniciando con el modelo de clasificación de productos, hallazgos más relevantes obtenidos durante el análisis exploratorio (EDA), proceso de limpieza y resultados de los modelos de regresión, sección en la cual se abordó el proceso de regresión bajo dos enfoques de interpolación y finalmente en la sección 5 se presentan las conclusiones más relevantes de los modelos implementados, tanto de modelo de clasificación como regresión, comparando nuestros resultados con los más relevantes de la literatura revisada.

## 2. FUENTES DE DATOS

Para el desarrollo del trabajo, se obtuvieron datos de ventas de productos masivos en 217 municipios de 5 departamentos de Colombia para el año 2022. Los datos están disponibles en Kaggle, y son de uso abierto. Adicionalmente, se usaron tablas auxiliares que complementan el dataset, provenientes del DANE, Kaggle y de un modelo de clasificación de propia autoría (La generación del dataset se explica en la sección 3.1. Generación base de datos Compras Categorizadas), toda la información anterior está en formato de Excel (xlsx). Las distintas fuentes permiten nutrir el dataset final y aplicar los diferentes modelos de ML y DL que permitan predecir los valores demandados de una lista de productos de consumo masivo en distintas zonas geográficas.

La tabla principal de datos (ComprasDda) contiene información de 16.383 registros de compras de productos en 217 municipios de 5 departamentos de Colombia para el año 2022, la

información está desagregada en 6 variables las cuales describen la compra según la fecha, nombre producto, ID Ubicacion, cantidad, unidad de medida y precio de venta (en COP). El resumen del origen de los datos se encuentra en la **Tabla 1**. Descripción de tablas originales:

**Tabla 1.** Descripción de tablas originales

| Tabla                                | Descripción                                                                               | Modelo        | Filas | Variables Categóricas | Variables Numéricas | Fuente                                                                                                                                                                                | Tamaño    |
|--------------------------------------|-------------------------------------------------------------------------------------------|---------------|-------|-----------------------|---------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------|
| ComprasDda                           | Registro de la cantidad vendida de un producto en una zona geográfica y fecha determinada | Regresión     | 16383 | 4                     | 2                   | Kaggle[ <a href="https://www.kaggle.com/datasets/davidms8/shopping-map">https://www.kaggle.com/datasets/davidms8/shopping-map</a> ]                                                   | 978 KB    |
| Ubicaciones                          | Zona geográfica, municipio, región y departamento                                         | Regresión     | 1725  | 9                     | 2                   | Kaggle[ <a href="https://www.kaggle.com/datasets/davidms8/shopping-map">https://www.kaggle.com/datasets/davidms8/shopping-map</a> ]                                                   | 3.574 KB  |
| Fechas                               | Indica si una fecha determinada es laboral, quincena, fin de semana                       | Regresión     | 365   | 3                     | 0                   | Propia                                                                                                                                                                                | 16 KB     |
| Variables socioeconómicas            | Variables de seguridad, nivel económico de cada municipio                                 | Regresión     | 1133  | 2                     | 49                  | DANE [ <a href="https://territoria.dnp.gov.co/index-app.html#/descargas">https://territoria.dnp.gov.co/index-app.html#/descargas</a> ]                                                | 17.437 KB |
| Lista de productos de consumo masivo | Lista de productos de supermercado con su respectiva subcategoría para entrenar modelo    | Clasificación | 25638 | 10                    | 3                   | Kaggle[ <a href="https://www.kaggle.com/datasets/camiloe martinez/productos-consumo-masivo/data">https://www.kaggle.com/datasets/camiloe martinez/productos-consumo-masivo/data</a> ] | 1.950 KB  |
| Nuevas Categorías Clasificación      | Lista de productos de categorías "Útiles Escolares" y "Frutas y Verduras"                 | Clasificación | 1830  | 2                     | 0                   | Fuente Propia                                                                                                                                                                         | 52 KB     |
| ComprasCategorizadas_                | Resultado tras aplicar un modelo de clasificación                                         | Regresión     | 16383 | 6                     | 2                   | Modelo de clasificación                                                                                                                                                               | 677 KB    |

|     |                                                                                    |  |  |  |  |        |  |
|-----|------------------------------------------------------------------------------------|--|--|--|--|--------|--|
| SVC | a la lista de productos de consumo masivo, agrupándolos en 16 categorías distintas |  |  |  |  | Propio |  |
|-----|------------------------------------------------------------------------------------|--|--|--|--|--------|--|

Las bases de datos usadas para nutrir nuestro dataset principal incluyeron información como la localización de las zonas que reportaron los registros de demanda, municipio, departamento, región y la ubicación geográfica dada en latitud y longitud (Base de datos “Ubicaciones”), categorización de las fechas de 2022 entre días “Laborales”, “Festivos”, “Fines de semana” y “Laboral-Quincena” (Base de datos “Fechas”), agregación de información demográfica, económica y de seguridad asociada a cada municipio (Base de datos “Variables Socioeconómicas”). Finalmente, la tabla de “Lista de productos de consumo masivo” y “Nuevas Categorías Clasificación” fueron usadas para entrenar un modelo de clasificación de textos, cuya salida es “Categorías de productos” y esta última se usó en el entrenamiento de modelos de predicción de demanda.

Es importante resaltar que la tabla de categorías es el resultado de un modelo de clasificación, el cual tuvo como objetivo agrupar la lista de 3.830 productos (Base de datos “Compras Dda”) en 16 categorías, facilitando el análisis y posterior implementación de modelos; en la metodología, se dará una breve descripción de la generación del modelo de clasificación para obtener la base de datos final utilizada en la predicción de las compras. El proceso de generación de las tablas se resume en la **Figura 1. Modelo de Clasificación.**



**Figura 1.** Modelo de Clasificación

La descripción y pasos seguidos para este proceso de clasificación se encuentran disponibles en el siguiente notebook [AprendizajeAutomatico/ModeloClasificacion.ipynb](#) at [main · Gabriel93UdeA/AprendizajeAutomatico](#)

A partir de los anteriores conjuntos de datos se generó el dataset final usado en la predicción de ventas realizadas para el presente artículo.

### 3. PROCESAMIENTO DE DATOS (METODOLOGÍA)

Este capítulo describe la estructura y el contenido de los datos de demanda de producto, así como el proceso de unión, limpieza y construcción del conjunto de datos utilizado en el presente trabajo. También se realiza un análisis descriptivo de las variables del dataset principal, proporcionando una comprensión inicial de sus características y patrones. Finalmente, se estandarizan las variables según lo requiera su naturaleza y se realiza la división en conjuntos para entrenamiento y validación de los modelos de machine learning.

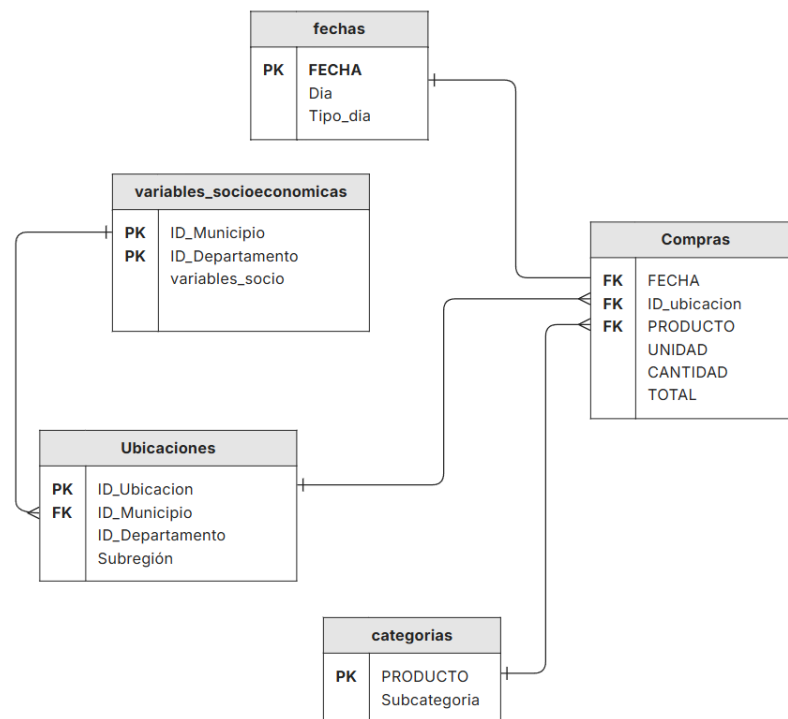
### 3.1. Generación base de datos Compras Categorizadas

En primera instancia para el modelo se comenzó usando la base de datos “Productos consumo masivo” (Información de Kaggle), la cual contiene más de 25 mil registros clasificados en 14 categorías, sin embargo se identificó que no contenía las categorías “Útiles Escolares” y “Frutas y Verduras”; teniendo en cuenta lo anterior y con el uso de la técnica “Web Scraping” se construyó una base de datos “Nuevas\_Categorías\_Clasificación” que contiene 1.830 registros, los cuales enriquecieron la base original de Kaggle. Posterior a la unificación de los datasets se realizó una depuración de las variables y un balanceo por categoría de productos con la técnica “RandomUnderSampler”, dando como resultado una base de datos balanceada con 629 registros por categoría.

Con la base de datos resultante se procedió a entrenar 6 modelos de clasificación (LinearSVC, LogisticRegression, RandomForestClassifier, MultinomialNB, GradientBoostingClassifier y KNeighborsClassifier) (ver Anexo 2. Comparación y Selección modelo de clasificación) teniendo en cuenta una separación del dataset (entrenamiento (80%) y testeo (20%)) y configurados con un Cross-Validation de 10. Se seleccionó el modelo con la mayor precisión para luego clasificar los productos de la base de datos de “ComprasDda” en 16 categorías.

### 3.2. Generación de Dataset principal

El proceso de generación de las tablas se resume en la **Figura 2. Modelo entidad relación** para el dataset conformado.



**Figura 2.** Modelo entidad relación para el dataset conformado

El dataset resultante contiene 16.383 registros con 67 columnas, que consideran variables de tipo de día, 51 indicadores socioeconómicos, departamento, subregiones, productos, subcategoría de producto y la cantidad vendida. El detalle de las columnas, su tipología (numérica o categórica) y su cantidad de registros no nulos se encuentran en el Anexo 3. Detalle de columnas, tipología y cantidad de registros, en el cual se evidencia que existen columnas con demasiados valores faltantes.

Por lo anterior, se decidió eliminar las columnas con más del 10% de valores nulos. Adicionalmente, para efectos del trabajo de interés, solo se usarán los productos que sean medibles por unidad (omitiendo aquellas basados en su peso) y agregando los valores numéricos obtenidos a un detalle por fecha, departamento y categoría de producto. Con lo anterior, el dataset se reduce a 5.364 registros con 60 columnas.

Finalmente, las columnas categóricas se convirtieron en variables binarias mediante One Hot Encoding (OHE) y las variables numéricas se estandarizaron utilizando Min Max Scaler, con

el objetivo de tener la misma escala de valores (0 a 1) en todas las variables consideradas para el entrenamiento de los modelos ML. Para la variable respuesta de regresión también se construyó un Min Max Scaler que estandarizará sus valores.

### **3.3. Selección de características y generación de dataset final**

Dado que el objetivo del análisis era predecir cantidades demandadas a lo largo del tiempo, fue necesario incorporar características adicionales derivadas de la serie temporal. Para ello, se calcularon las características Rolling Mean, Rolling Standard Desviation, Rolling Min, Rolling Max, Rolling Skew y Slope, utilizando ventanas móviles de 1, 3 y 7 días, lo que permitió capturar tendencias, volatilidad y variaciones en la distribución de los datos para cada categoría y departamento.

Una vez generadas estas variables, se llevó a cabo un análisis exploratorio para evaluar la distribución de las variables cuantitativas y su correlación a la variable respuesta. En este proceso, se identificó una alta colinealidad entre diversas variables socioeconómicas, lo que sugiere redundancia en la información. Además, se observó que varias de estas variables presentaban una baja relación con la cantidad demandada, medida a través de los coeficientes de correlación de Pearson y Spearman.

Para abordar estos problemas, se realizó una depuración de variables, eliminando aquellas con alta colinealidad y baja relevancia estadística. Posteriormente, se seleccionaron las características más significativas, combinando las variables generadas a partir de la serie temporal con un subconjunto optimizado de variables socioeconómicas. Este proceso permitió reducir la dimensionalidad del dataset a 34 columnas y mejorar la calidad de la información utilizada en los modelos predictivos.

### 3.4. Predicción de ventas

#### 3.4.1. División de los datos en entrenamiento, validación y prueba

El dataset resultante en la selección de características se reservó un 10% como conjunto de prueba, mientras que el 90% restante se dividió en conjuntos de entrenamiento y validación para el ajuste de los modelos de regresión. Para esta última división, se utilizó la función `TimeSeriesSplit` de `scikit-learn`.

#### 3.4.2. Modelos de Predicción de demanda

Para la predicción de demanda de productos de consumo, se implementaron cinco modelos de regresión utilizando Python con las librerías `Scikit-learn` y `TensorFlow`, donde los hiperpárametros fueron optimizados mediante `GridSearchCV`. En cada modelo, se evaluaron entre 3 y 4 valores por hiperpárametro para encontrar la mejor configuración. Los hiperpárametros ajustados para cada modelo fueron los siguientes:

- **Support Vector Machine (SVM):** Se utilizó el modelo SVM de `Scikit-learn`, probando distintos kernel e hiperpárametros  $C$ ,  $\text{degree}$  y  $\epsilon$ , buscando capturar relaciones no lineales en los datos.
- **Elastic Net:** Se implementó con `ElasticNet` de `Scikit-learn`, combinando las penalizaciones  $L1$  (Lasso) y  $L2$  (Ridge) para regularización mediante el hiperpárametro  $\text{l1\_ratio}$ . Se ajustaron los parámetros  $\alpha$  y la proporción de mezcla mediante validación cruzada.
- **Random Forest:** El modelo `RandomForestRegressor` de `Scikit-learn` es útil para capturar relaciones no lineales y manejar características de alta dimensión. Se ajustó el número de árboles ( $\text{n\_estimators}$ ), el máximo de features ( $\text{max\_features}$ ) y la profundidad ( $\text{max\_depth}$ ) para maximizar el desempeño.
- **Gradient Boosting:** Se aplicó `GradientBoostingRegressor`, optimizando la tasa de aprendizaje ( $\text{learning\_rate}$ ), el número de estimadores ( $\text{n\_estimators}$ ) y la profundidad ( $\text{max\_depth}$ ) para mejorar la predicción iterativa y minimizar el error residual.
- **Redes neuronales:** Se diseñaron distintas redes neuronales con `TensorFlow`, utilizando capas densas (`Dense`) con activaciones `ReLU` y función de pérdida `MSE`.

Se empleó el optimizador Adam y el ajuste de hiperpárametros incluyó la selección de neuronas por capa y la tasa de regularización (dropout).

### **3.5. Métricas de desempeño**

Con la revisión de la literatura realizada, se detectó que para el presente problema es común utilizar el RMSE y el MAPE como métricas de desempeño a los modelos de regresión. En algunos casos también es útil el uso del R<sup>2</sup>. En los trabajos analizados, el MAPE varía entre 0.2 y 0.4 [17] [19] [22] [24].

Dado el contexto de negocio, los pequeños comerciantes podrían tomar como base la información obtenida del modelo para reducir costos de inventario, mejorar la satisfacción del cliente por la disponibilidad de productos, mejorar la rotación de productos, entre otros.

## **4. RESULTADOS**

Este capítulo presenta los principales resultados a los procedimientos realizados para la clasificación de productos y su predicción de demanda. Se presenta los resultados de entrenamiento de modelos de clasificación de productos en subcategorías, definiendo el modelo con menor margen de error. Posteriormente, se realiza el análisis descriptivo del dataset de demanda generado y su respectiva limpieza. Finalmente, se presentan los resultados a los modelos de regresión para la predicción de demanda de productos.

### **4.1. Resultados del modelo de clasificación y generación BD final**

Como se observa en las secciones anteriores, para el modelo de clasificación se entrenaron 6 modelos, de los cuales el LinearSVC y RandomForestClassifier obtuvieron los mejores valores de Accuracy con 0,96 y 0,94 respectivamente, cabe resaltar que se implementó Grid-Search para identificar los mejores hiperpárametros, sin embargo, no se evidenciaron diferencias significativas (0,96 y 0,94 respectivamente) a comparación con los resultados iniciales.

En la **Figura 3** se muestran las métricas obtenidas por el modelo seleccionado (LinearSVC), con el cual se procedió a clasificar la base de datos “ComprasDda”.

|                                | Predicted |        |          |         |
|--------------------------------|-----------|--------|----------|---------|
|                                | precision | recall | f1-score | support |
| Aseo de hogar                  | 1.00      | 0.98   | 0.99     | 132     |
| Bebidas                        | 0.95      | 0.98   | 0.97     | 107     |
| Belleza                        | 1.00      | 0.97   | 0.98     | 128     |
| Charcutería                    | 0.97      | 0.98   | 0.98     | 120     |
| Cuidado Personal               | 0.98      | 0.98   | 0.98     | 126     |
| Cuidado de ropa y calzado      | 0.98      | 0.99   | 0.99     | 117     |
| Despensa                       | 0.92      | 0.90   | 0.91     | 136     |
| Dulces y postres               | 0.93      | 0.95   | 0.94     | 128     |
| Frutas y Verduras              | 0.88      | 0.92   | 0.90     | 113     |
| Limpieza de cocina             | 0.95      | 0.99   | 0.97     | 134     |
| Lácteos, huevos y refrigerados | 0.95      | 0.95   | 0.95     | 130     |
| Panadería y Pastelería         | 0.96      | 0.93   | 0.95     | 122     |
| Pasabocas                      | 0.92      | 0.91   | 0.91     | 121     |
| Productos Congelados           | 0.98      | 0.99   | 0.99     | 137     |
| Utiles Escolares               | 0.99      | 0.95   | 0.97     | 126     |
| Vinos y Licores                | 1.00      | 0.99   | 1.00     | 136     |
| accuracy                       |           |        | 0.96     | 2013    |
| macro avg                      | 0.96      | 0.96   | 0.96     | 2013    |
| weighted avg                   | 0.96      | 0.96   | 0.96     | 2013    |

**Figura 3.** Métricas modelo LinearSVC de clasificación de productos

A partir de los subconjuntos de datos definidos en la **Tabla 1**, se construyó el dataset principal, el cual permitió entrenar modelos de regresión orientados a la predicción de múltiples categorías de productos. La carga, unión, limpieza y exploración de datos se encuentra disponible en el repositorio de GitHub asociado ([Gabriel93UdeA/AprendizajeAutomatico](https://github.com/Gabriel93UdeA/AprendizajeAutomatico)).

## 4.2. Analítica descriptiva

En la **Figura 4**, se ilustra la distribución de las cantidades de venta por subcategoría de producto. En la misma, se evidencian valores atípicos en cada una de las subcategorías. Estos valores atípicos en su mayoría se consideran excepciones dado el contexto de negocio, y tendrán que ser tratados más adelante para evitar problemas de ajuste en los modelos.

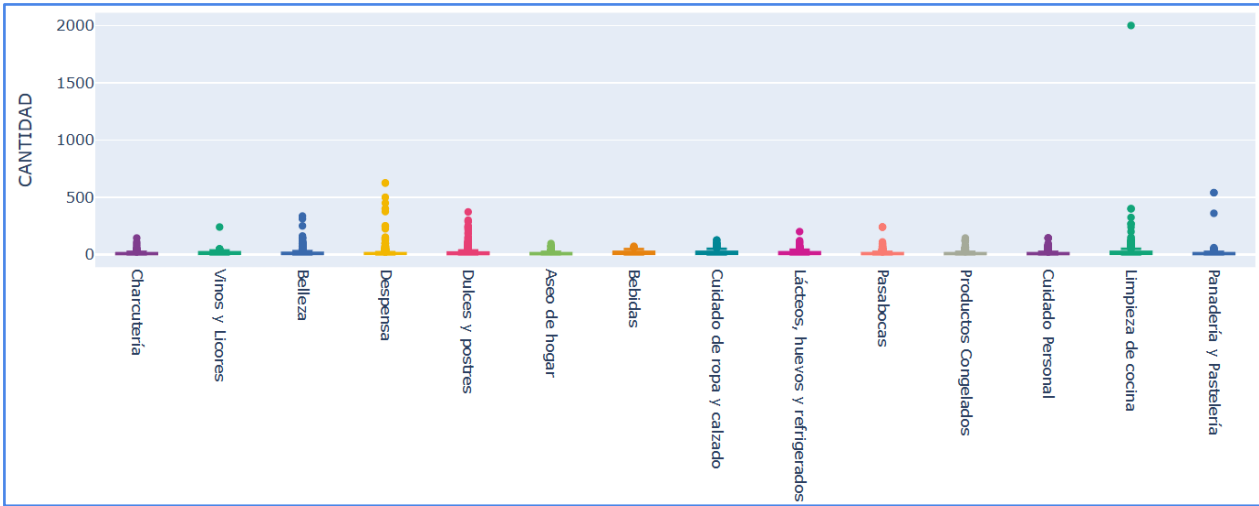


Figura 4. Distribución de cantidades vendidas por categoría de producto

Adicionalmente, se calculan los valores vendidos en cada subregión mes a mes, como se ejemplifica en la **Figura 5**. De este análisis, es importante resaltar dos puntos clave:

- 1. Existen subregiones con pocos registros, e incluso con 0 registros en algunos meses.
- 2. En el último trimestre, se observa una reducción significativa en los valores de venta registrados. Esta situación es inusual dentro del contexto del negocio en cuestión, lo que lleva a sospechar sobre la confiabilidad de los datos en los últimos tres meses.

| mes                | 1      | 2      | 3      | 4      | 5      | 6      | 7      | 8      | 9      | 10    | 11    | 12    |
|--------------------|--------|--------|--------|--------|--------|--------|--------|--------|--------|-------|-------|-------|
| Subregión (SGR)    |        |        |        |        |        |        |        |        |        |       |       |       |
| Alto Occidente     | 348.0  | 459.0  | 555.0  | 430.0  | 380.0  | 255.0  | 209.0  | 380.0  | 615.0  | 89.0  | 274.0 | 212.0 |
| Alto Oriente       | 336.0  | 185.0  | 227.0  | 76.0   | 208.0  | 167.0  | 85.0   | 294.0  | 136.0  | 104.0 | 35.0  | 82.0  |
| Bajo Cauca         | 744.0  | 577.0  | 983.0  | 613.0  | 535.0  | 552.0  | 348.0  | 749.0  | 438.0  | 113.0 | 117.0 | 317.0 |
| Bajo Occidente     | 94.0   | 267.0  | 153.0  | 132.0  | 115.0  | 64.0   | 71.0   | 310.0  | 90.0   | 20.0  | 36.0  | 338.0 |
| Capital            | 729.0  | 334.0  | 396.0  | 286.0  | 163.0  | 288.0  | 361.0  | 405.0  | 317.0  | 125.0 | 90.0  | 176.0 |
| Capital Cordillera | NaN    | 49.0   | 82.0   | 31.0   | 1.0    | 9.0    | 8.0    | NaN    | 34.0   | NaN   | NaN   | 48.0  |
| Centro             | 1753.0 | 2608.0 | 1619.0 | 1549.0 | 1837.0 | 1510.0 | 1236.0 | 2220.0 | 1272.0 | 708.0 | 410.0 | 727.0 |
| Cordilleranos      | 25.0   | 12.0   | 52.0   | NaN    | 16.0   | 13.0   | 9.0    | 9.0    | 8.0    | 10.0  | 17.0  | 66.0  |

Figura 5. Ejemplo distribución de cantidades vendidas por subregión

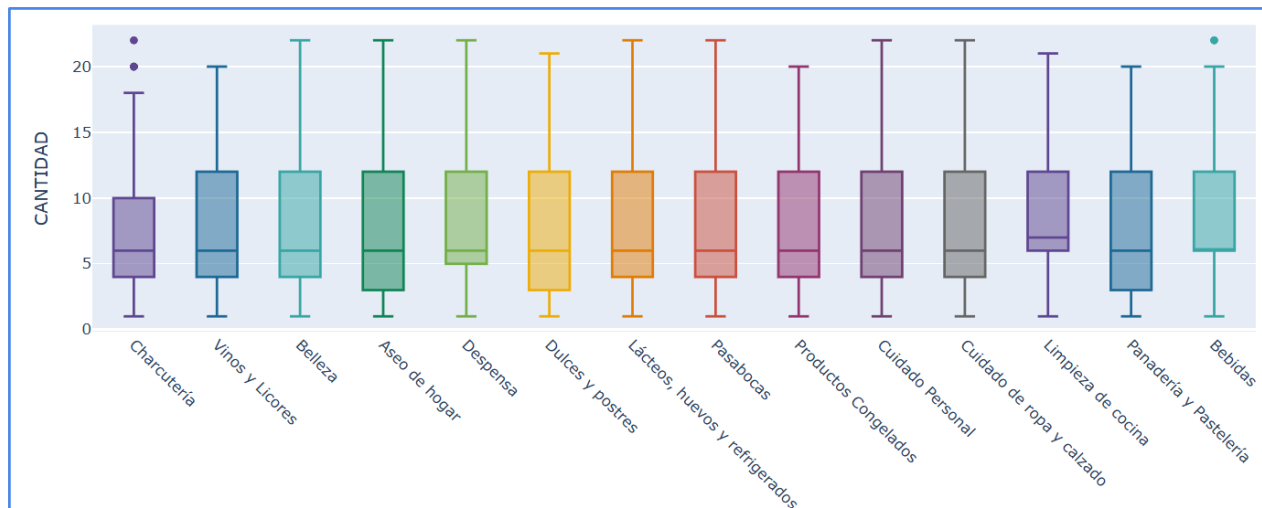
Finalmente, se calcularon los coeficientes de correlación de Pearson y Spearman de las variables numéricas respecto a la cantidad vendida. Para ninguno de los casos la correlación calculada indica una relación fuerte con la demanda de productos, obteniendo valores por debajo del 5%.

#### **4.3. Limpieza y organización de datos para series de tiempo**

El dataset presenta diversas inconsistencias, incluyendo variaciones significativas en las ventas del último trimestre del año, subregiones con pocos o nulos registros, sesgos en algunos indicadores socioeconómicos, columnas con valores nulos persistentes y valores atípicos en la demanda de productos. Para evitar problemas de ajuste futuros, se le dio tratamiento a cada una de las falencias mencionadas con el fin de trabajar con un dataset consistente.

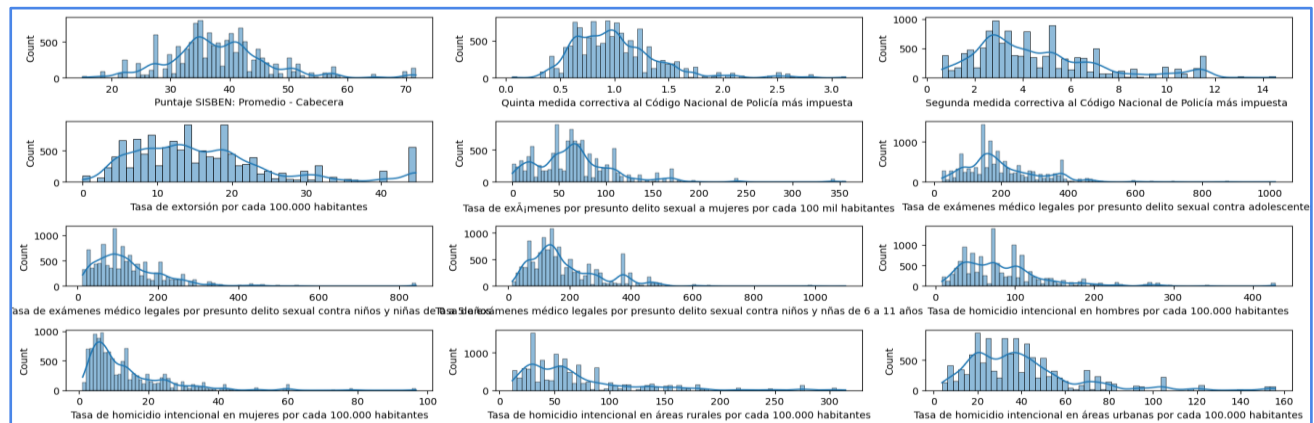
De esta manera, se omite el último trimestre del año 2022, al igual que se eliminan 6 subregiones (Montes de María, Cordilleranos, Capital Cordillera, Sabanas, Fría, Capital y Risaralda Sur) que demostraban tener una cantidad mínima de datos. En cuanto a los valores faltantes en variables socioeconómicas, se realiza la imputación mediante un promedio nacional para los campos que lo requirieron.

Por otro lado, para el caso de los datos atípicos, se eliminan los valores de venta por encima del tercer cuartil (Q3) más 1.5 veces el rango intercuartil (IQR), dado que estos representan los valores extremos en la distribución de las cantidades de productos. Esto nos permite limpiar el dataset de valores anómalos que podrían distorsionar los resultados del análisis. En la **Figura 6** se evidencia la distribución tras la eliminación de datos atípicos.



**Figura 6.** Distribución de cantidades vendidas por subcategoría sin atípicos

En la **Figura 7** se muestra la distribución de algunas variables socioeconómicas incluidas en el dataset. En esta visualización también se pueden observar valores atípicos dentro de las distribuciones de cada variable. Adicionalmente, se identifican columnas con sesgos de distribución significativos, debido a la presencia de un número excesivo de valores en 0, entre otros.



**Figura 7.** Distribución de variables socioeconómicas

Finalmente, como se evidenciará en la siguiente sección de resultados, durante el análisis de los datos se identificaron falencias en los registros de ventas, principalmente relacionadas con la falta de información en ciertos días del año para diversas categorías y subregiones. Ante esta situación, se llevaron a cabo experimentos con tres tipos de bases de datos: la original, una con los

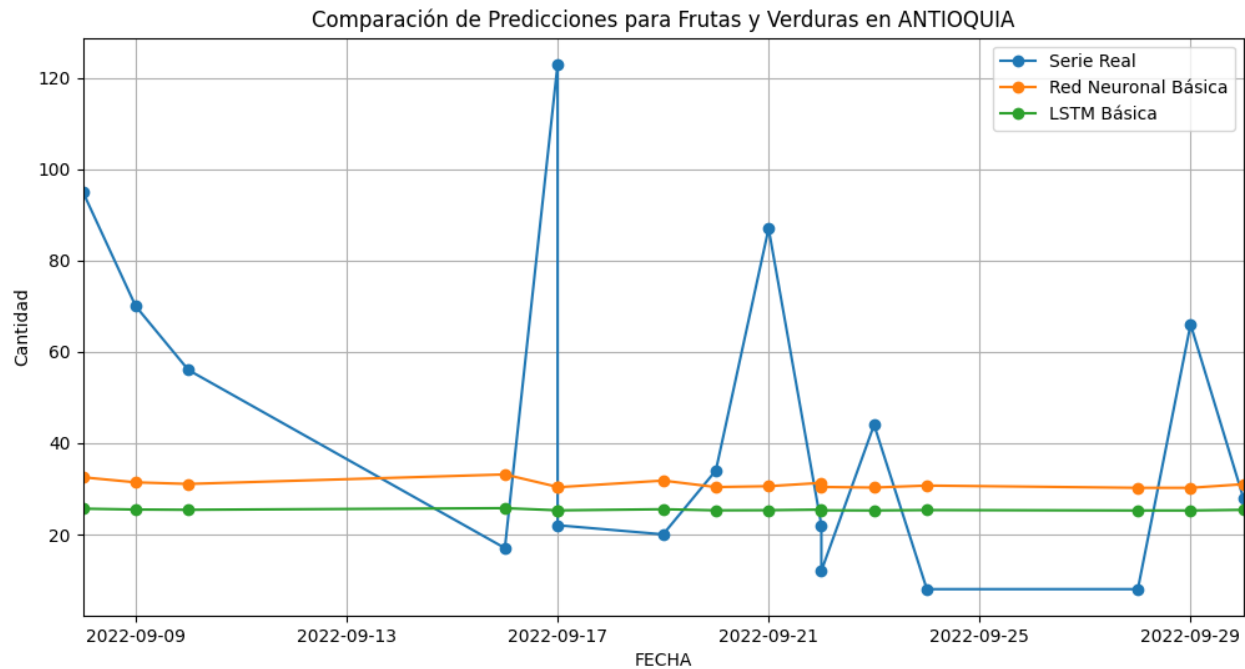
días faltantes reemplazados por ceros y otra con imputaciones mediante interpolación lineal. En la sección de resultados se analizará el impacto de cada enfoque, mientras que en la sección de conclusiones se discutirá la pertinencia y el uso de estos métodos.

#### **4.4. Resultados del modelo de regresión**

Como se describe en la sección 3.4.2, se usaron 5 modelos para predecir el comportamiento de ventas en las diferentes subregiones de los 5 departamentos, sin embargo, en la construcción de modelos se identificaron algunas particularidades que se describirán a continuación.

##### **4.4.1. Primer acercamiento modelos**

En la implementación de los primeros modelos de regresión se identificó que no se obtuvieron métricas que muestren un buen comportamiento, lo anterior se evidencia en el Anexo 4. Comparación métricas primer acercamiento modelos de regresión donde se recolectaron los valores de RMSE y MAPE en múltiples pruebas, evidenciando que a nivel general los modelos obtuvieron un MAPE superior al 100% y RMSE entre 16 y 92, mostrando que la predicción tiene un error promedio que es mayor al valor real y por tanto estos no son adecuados para predicciones, junto con lo anterior en la **Figura 8** podemos observar que los modelos de redes neuronales no capturando las variaciones en ventas y están prediciendo valores sin que se adapten a la variación de la demanda.



**Figura 8.** Predicción primeros modelos regresión (red neuronal)

#### 4.4.2. Variación de Modelos

En el proceso de análisis de los resultados anteriores, identificamos falencias en el dataset final, principalmente la ausencia de registros de cantidades de distintas categorías en ciertas fechas. Esta falta de información representaba un desafío, ya que impedía una evaluación continua del comportamiento de compras a lo largo del tiempo y causaba errores de generalización en los modelos entrenados.

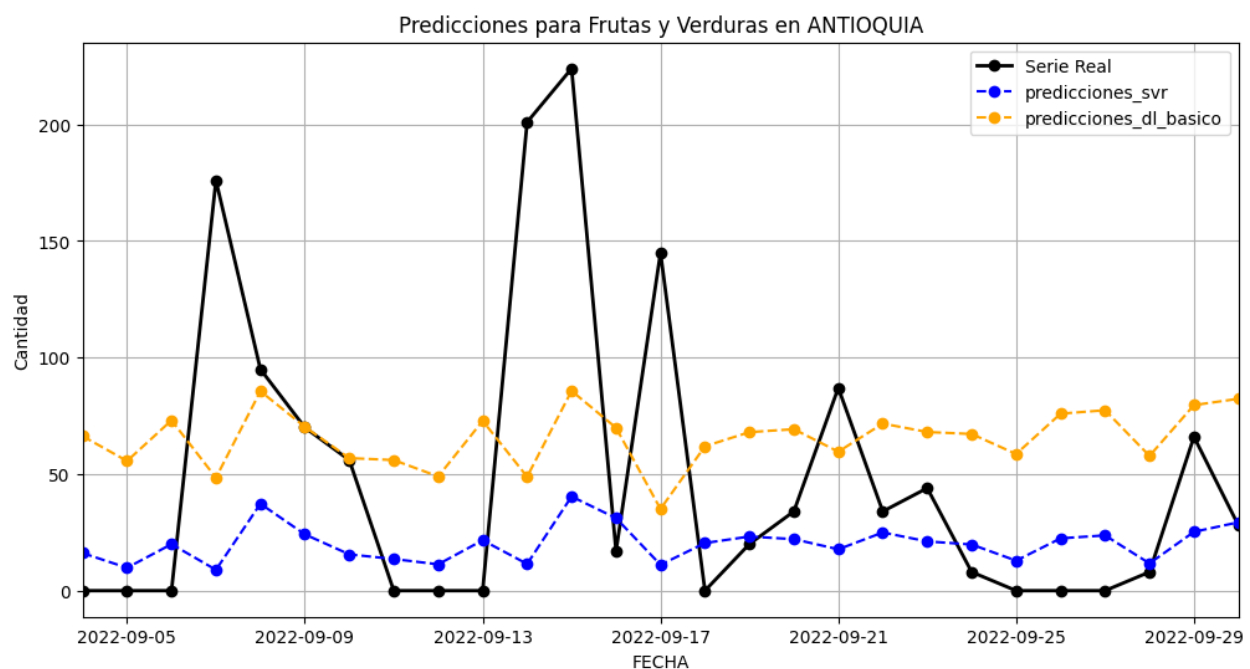
Para abordar este problema, implementamos un proceso de imputación con dos enfoques distintos:

- **Relleno con Cero:** En este enfoque, los valores faltantes se asignan como cero, asumiendo que la ausencia de registros corresponde efectivamente a la inexistencia de ventas en esas fechas. Esta estrategia podría ser útil cuando se quiere reflejar fielmente los períodos sin actividad comercial, evitando posibles sesgos en la proyección de tendencias. Sin embargo, para nuestro caso significaría que productos de consumo masivo (como despensa) no

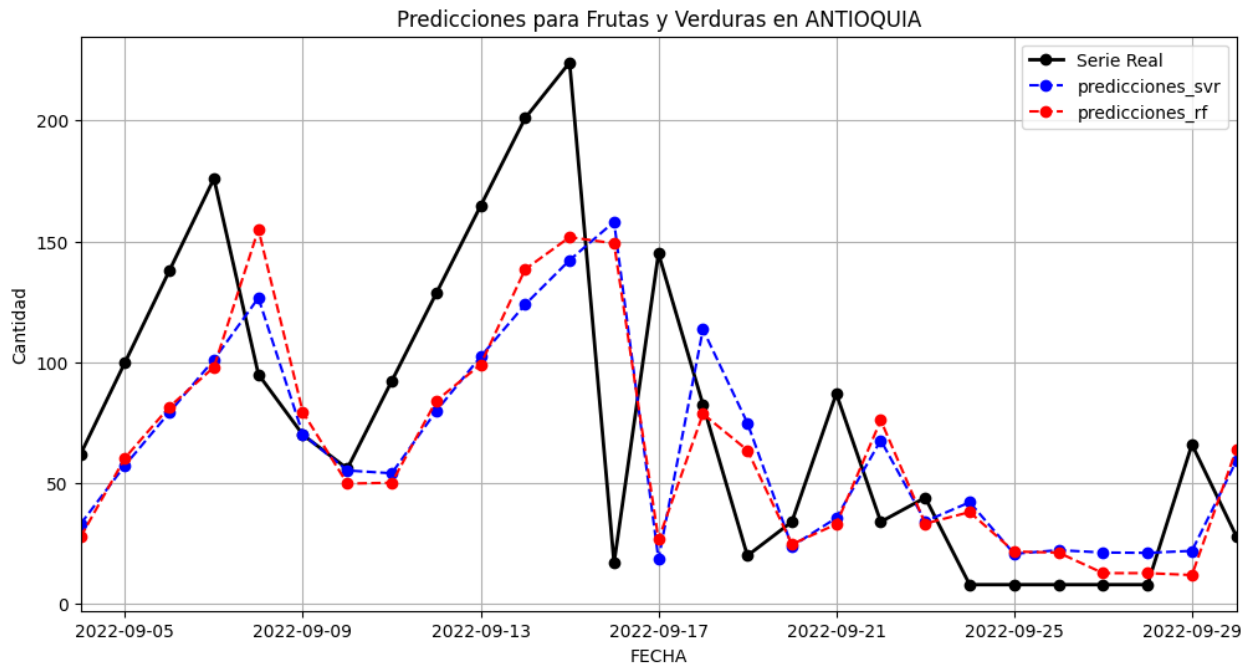
tuvieron ventas por varios días consecutivos en las tiendas mapeadas, lo que parece poco probable dada la alta densidad poblacional de departamentos como Antioquia.

- **Interpolación lineal:** Este método calcula los valores faltantes tomando como referencia los datos de otros días para las categorías y departamentos que presenten vacíos de información. Se considera una opción efectiva cuando el patrón de venta de productos es relativamente estable, permitiendo generar valores más representativos y alineados con la tendencia general. Sin embargo, en casos donde existe una alta variabilidad en la demanda se puede incurrir en errores al generar un registro que está afectado por valores extremos.

Las **Figura 9** y **Figura 10** ilustran los resultados de los mejores modelos obtenidos en cada enfoque, permitiendo comparar el impacto de ambas técnicas en la estructuración de los datos.



**Figura 9.** Predicciones modelos de regresión con relleno en cero



**Figura 10.** Predicciones modelos de regresión con interpolación lineal

Para ambos casos, se evidencia un mejor ajuste de las predicciones de los modelos de regresión, en comparación de la **Figura 8**. Adicionalmente, se obtuvieron valores de RMSE más acotados, con mínimos de hasta 6 unidades en algunos modelos. Estas métricas se pueden visualizar en el Anexo 5. Comparación métricas desempeño modelos de regresión. Para ambos enfoques de imputación, resalta el ajuste de algoritmos como SVMR, Gradient Boosting y LSTM.

## 5. CONCLUSIONES

A partir de la revisión de la literatura descrito en el Anexo 1. Resumen revisión de la literatura modelos analíticos , se conocieron las principales metodologías usadas en la predicción de la demanda, permitiendo concluir que el enfoque de nuestro proyecto, que combina el análisis de variables socioeconómicas y temporales, tiene el potencial de ofrecer un avance significativo en la predicción de ventas en el sector retail. A diferencia de los estudios previos, nuestros modelos buscaron no solo prever la demanda agregada, sino también comprender las variaciones en la demanda por tipo de producto en un entorno geográfico más amplio, analizando el comportamiento de las ventas en 5 departamentos de Colombia, lo que genera un valor agregado dado que no solo se implementan modelos de machine learning, sino que generaría información de interés sobre el comportamiento de consumo en sectores del territorio, útil para la toma de decisiones estratégicas.

La clasificación de productos mediante técnicas de procesamiento de datos y aprendizaje automático fue un procedimiento clave para el desarrollo del trabajo. A partir de una base existente y complementada con Web Scraping, se logró enriquecer la información disponible, incorporando nuevas categorías que no estaban presentes originalmente. La experimentación con seis modelos de clasificación permitió evaluar diferentes enfoques y seleccionar el más preciso, garantizando una asignación óptima de productos dentro de las 16 subcategorías. En este procedimiento resalta la importancia de combinar fuentes de datos externas, estrategias de preprocesamiento y técnicas avanzadas de clasificación para obtener un sistema confiable y optimizado.

Aunque el mejor modelo de clasificación logró obtener un accuracy de 0,96, este resultado está sujeto a mejora, dado que al realizar una prueba aleatoria entre los registros clasificados de la base de datos “ComprasDda” identificamos que se cometían errores puntuales en la clasificación de algunos productos. Indicando que, a pesar del buen desempeño del modelo, existen limitaciones frente algunos productos, lo que se podría abordar obteniendo o construyendo una base de datos que abarque una mayor cantidad de productos con lo que el modelo de clasificación se pueda entrenar.

El proceso de optimización de hiperpárametros es crucial para que los modelos logren alcanzar su máximo potencial con los datos que fueron entrenados. Sin embargo, particularmente en la optimización para los modelos de clasificación, identificamos que el uso de GridSearch no aportó una mejora significativa en el resultado final, dado que el modelo LinearSVC obtuvo el mismo valor de Accuracy (0,966912) entre el modelo base y con hiperpárametros optimizados, mostrando que los parámetros generados por defecto eran los adecuados para nuestro caso particular.

Para el entrenamiento de modelos de regresión, fue fundamental acompañar el análisis de resultados con visualizaciones que permitieron identificar falencias en las predicciones generadas por los distintos algoritmos. Uno de los principales problemas detectados fue la dificultad de los modelos para generalizar adecuadamente las variaciones de las series debido a la presencia de huecos de información en varias fechas dentro del dataset. Para abordar esta problemática, se exploraron dos enfoques distintos: el primero asumió que no existieron ventas en los días faltantes, completando los valores con ceros, mientras que el segundo aplicó imputación mediante interpolación lineal, considerando que las ventas deberían existir de manera continua a lo largo del año 2022. Los resultados mostraron que el enfoque de interpolación permitió una mejor generalización de los modelos, logrando seguir con mayor precisión las variaciones de la serie gracias a la mayor continuidad en los datos. Sin embargo, en ambos casos, las métricas de desempeño siguen dependiendo del comportamiento real de las ventas en el año 2022, lo que puede generar sesgos en la evaluación de los modelos.

En el caso de los modelos entrenados con el dataset en el que los días faltantes fueron completados con ceros, se observó que métricas como  $R^2$  y MAPE fueron significativamente afectadas por la presencia de estos valores, dificultando la comparación con otros enfoques. En cuanto al RMSE, los modelos mostraron una desviación promedio de entre 6 y 19 unidades de venta, siendo SVR y LSTM los que presentaron el mejor desempeño en este aspecto. Sin embargo, también es posible que la inclusión de ceros haya generado un efecto indirecto en el ajuste de los modelos. Por otro lado, para el dataset con interpolación, las métricas de RMSE rondaron las 16 unidades, con un MAPE entre 0.4 y 0.6 en la mayoría de los modelos, resultados que son

comparables con los reportados en la literatura. Estos resultados resaltan la importancia de una adecuada selección de estrategias de imputación e indican la necesidad de analizar el comportamiento temporal de los datos para evitar distorsiones en las predicciones.

Adicionalmente, como se evidencia en la **Figura 10**, algunos algoritmos presentan un retraso en el seguimiento de los patrones de movimiento de la serie real. Esto puede deberse a la forma en que los modelos procesan la información temporal, especialmente aquellos que dependen de relaciones secuenciales en los datos.

---

## REFERENCIAS

- [1] S. Vivanco, *La investigación de mercados*, Santiago, 2010.
- [2] K. Hodzic, H. Hasic, E. Cogo y Z. Juric, «Warehouse Demand Forecasting based on Long Short-Term Memory neural networks,» 2019.
- [3] J. Luyo y e. al., «A Predictive Sales System Based on Deep Learning,» *International Journal of Advanced Computer Science and Applications*, 2024.
- [4] A. Mitra, J. A., A. Kishore y e. al., «A Comparative Study of Demand Forecasting Models for a Multi-Channel Retail Company: A Novel Hybrid Machine Learning Approach,» *Oper. Res. Forum* 3, 2022.
- [5] H. Rincón y et al., «Retail en Colombia 2010 - 2015: Un estudio a partir del análisis financiero integral como elemento de soporte para la toma de decisiones,» *Revista Espacios*, 2017.
- [6] D. Gaitán, «Caracterización de las tiendas de barrio de Cartagena,» *PANORAMA*, pp. 59-71, 2010.
- [7] B. Espinel y .. et al, «Factores que influyen en el comportamiento del consumidor de los negocios al detal y supermercados en el Caribe colombiano,» *Revista Lasallista de Investigación*, 2019.
- [8] A. Ortiz y et al, «La economía de las tiendas de barrio en Colombia,» *Alianza EFI*, 2021.
- [9] E. Ramírez, «¿Por qué las tiendas de barrio en colombia no han fracasado frente a la llegada de las grandes cadenas de supermercados?,» *Revista Entornos*, pp. 37--50, 2008.
- [10] A. López, *Estrategias de diferenciacion. Desafios para un retail eficiente*, Madrid: ESIC, 2017.
- [11] H. Silva, «Comportamiento de las superficies de retail en Colombia,» *Pensamiento & Gestión*, 2011.
- [12] L. Emperatriz y N. María, «TIENDAS ARA IMPACTO EN EL MERCADO DEL RETAIL: CASO CARTAGENA DE INDIAS, COLOMBIA,» *Global de Negocios*, pp. 1-13, 2018.

- 
- [13] I. De la Hoz y C. Pérez, «Método Cluster-RNA para clasificar, caracterizar y pronosticar Perfiles competitivos del sector Tiendas Minoristas en la ciudad de Barranquilla,» *INGE CUC*, pp. 234-240, 2020.
- [14] J. Silva, J. C. M. R. H. Herazo, O. B. Lezama, G. Morgado y N. Varela, «Early warning method for the commodity prices based on artificial neural networks,» *Procedia Computer Science*, pp. 1243-1248, 2019.
- [15] İ. Güven y F. Şimşir, «Demand forecasting with color parameter in retail apparel industry using artificial neural networks (ANN) and support vector machines (SVM) methods,» *Computers and Industrial Engineering*, 2020.
- [16] W. Forero y F. Negre, «Técnicas y aplicaciones del Machine Learning e Inteligencia Artificial en educación: una revisión sistemática,» *Revista Iberoamericana de Educación a Distancia*, vol. 27, n° 1, pp. 209-253, 2024.
- [17] B. Chae, C. Sheu y E. Park, «The value of data, machine learning, and deep learning in restaurant demand forecasting: Insights and lessons learned from a large restaurant chain,» *Decision Support Systems*, 2024.
- [18] S. Punia y S. Shankar, «Predictive analytics for demand forecasting: A deep learning-based decision support system,» *Knowledge-Based Systems*, 2022.
- [19] M. Rodrigues, V. Miguéis, S. Freitas y T. Machado, «Machine learning models for short-term demand forecasting in food catering services: A solution to reduce food waste,» *Journal of Cleaner Production*, 2024.
- [20] A. Correa Loaiza y T. Mejía Lemus, «Análisis de modelos basados en Machine Learning para la predicción de la demanda de productos en la empresa Dyna & Cía. S.A.,» *Repositorio Institucional Universidad de Antioquia*, 2023.
- [21] A. Jaramillo Mira y J. P. Gutiérrez Muriel, «Pronóstico de ventas de comestibles de Corporación Favorita,» *Repositorio Institucional Universidad de Antioquia*, 2021.
- [22] E. Hincapie, «Prediccion de la demanda usando modelos de machine learning,» *Repositorio Institucional Universidad de Antioquia*, 2021.

- [23] E. LeDell y S. Poirie, «H2O AutoML: Scalable Automatic Machine Learning,» *SEMANTIC SCHOLAR*, 2020.
- [24] V. Lizcano y J. Romero, «Predicción de ventas futuras,» *Repositorio Institucional Universidad de Antioquia*, 2021.
- [25] S. Raschka y V. Mirjalili, «Python Machine Learning: Machine Learning and Deep Learning with Python, Scikit-learn, and TensorFlow,» 2017.
- [26] M. Ipia, «Predicción de categorías de productos en el sector minorista utilizando técnicas de aprendizaje supervisado,» *Repositorio Institucional Universidad de Antioquia*, 2023.

## ANEXOS

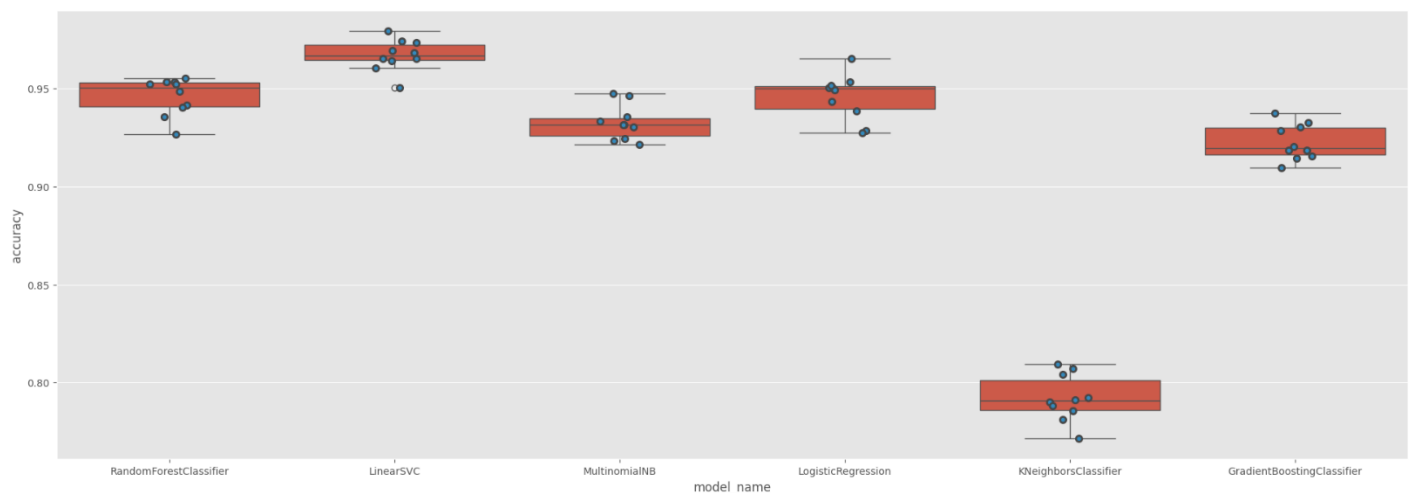
### Anexo 1. Resumen revisión de la literatura modelos analíticos

| Autores                             | Variables Entrada                                                                                        | Modelos                                                                      | Métricas Evaluación | Resultados                                                                      |
|-------------------------------------|----------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------|---------------------|---------------------------------------------------------------------------------|
| (Chae et al., 2024)                 | Tiempo, Promociones, Socioeconómicas                                                                     | Regresión Regularizada<br>Bosque aleatorio<br>SVM<br>Redes neuronales        | MAPE                | Valores MAPE entre 0.1 y 0.8                                                    |
| (Rodrigues et al., 2024)            | Tiempo, Clima, Cantidad, estudiantes percibidos                                                          | Bosque aleatorio<br>Light GBM<br>LSTM<br>Transformer<br>Media móvil<br>Näive | RMSE<br>MAPE<br>R2  | MAPE entre 0.2 y 0.4 en el producto más representativo                          |
| (Punia & Shankar, 2022)             | Características productos, Puntos de venta, Promociones, Búsquedas en Google<br>Clima<br>Socioeconómicas | ARIMA<br>Bosque Aleatorio<br>LSTM                                            | ME<br>MAE<br>RMSE   | RMSE alrededor de 2.3 para la mayoría de productos y modelos                    |
| (Güven & Şimşir, 2020)              | Características productos, Tiempo, Puntos de venta, Clima, Socioeconómicas                               | SVM<br>Redes neuronales                                                      | RMSE                | RMSE entre 110 y 170, siendo en general mejor la predicción de redes neuronales |
| (Silva et al., 2019)                | Tiempo, Socioeconómicas                                                                                  | SVM<br>Redes neuronales                                                      | MAPE                | MAPE de 0.10 para las redes neuronales, siendo el mejor modelo                  |
| (Correa Loaiza & Mejía Lemus, 2023) | Tiempo                                                                                                   | ARIMA<br>Regresión Lineal<br>Random Forest<br>Redes neuronales               | MAE                 | MAE entre 100 y 10000 unidades                                                  |
| (Jaramillo Mira & Gutiérrez)        | Tiempo                                                                                                   | Redes neuronales                                                             | RMSE                | RMSE 0.56                                                                       |

|                          |                                                                                                                                     |                                                                                         |                                                  |                    |
|--------------------------|-------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------|--------------------------------------------------|--------------------|
| Muriel, 2021)            |                                                                                                                                     |                                                                                         |                                                  |                    |
| (Hincapié, 2021)         | Tiempo, precio, ubicación, descuentos, categoría de producto, canal de venta, tipo de pago, contenido de producto, público objetivo | Random Forest Regressor, Plataforma H2O AutoML                                          | MAPE                                             | MAPE 0.35          |
| (Lizcano & Romero, 2021) | Tiempo, tipo producto, Ubicación, precio                                                                                            | Random Forest<br>Redes Neuronales<br>RANSAC                                             | RMSE                                             | RMSE entre 10 y 20 |
| (Ipia, 2023)             | Tiempo, categoría de producto, precio, cliente, método de pago, ubicación                                                           | Naive Bayes<br>Árbol de Decisión<br>Random Forest<br>Adaptive Boost<br>Redes Neuronales | Precisión (modelo de clasificación de productos) | Accuracy 78%       |

## Anexo 2. Comparación y Selección modelo de clasificación

Se compararon 6 modelos de clasificación



Teniendo en cuenta el valor de accuracy y la optimización de parámetros con GridSearch, los dos



|                                | precision | recall | f1-score | support |
|--------------------------------|-----------|--------|----------|---------|
| Aseo de hogar                  | 1.00      | 0.98   | 0.99     | 132     |
| Bebidas                        | 0.95      | 0.98   | 0.97     | 107     |
| Belleza                        | 1.00      | 0.97   | 0.98     | 128     |
| Charcutería                    | 0.97      | 0.98   | 0.98     | 120     |
| Cuidado Personal               | 0.98      | 0.98   | 0.98     | 126     |
| Cuidado de ropa y calzado      | 0.98      | 0.99   | 0.99     | 117     |
| Despensa                       | 0.92      | 0.90   | 0.91     | 136     |
| Dulces y postres               | 0.93      | 0.95   | 0.94     | 128     |
| Frutas y Verduras              | 0.88      | 0.92   | 0.90     | 113     |
| Limpieza de cocina             | 0.95      | 0.99   | 0.97     | 134     |
| Lácteos, huevos y refrigerados | 0.95      | 0.95   | 0.95     | 130     |
| Panadería y Pastelería         | 0.96      | 0.93   | 0.95     | 122     |
| Pasabocas                      | 0.92      | 0.91   | 0.91     | 121     |
| Productos Congelados           | 0.98      | 0.99   | 0.99     | 137     |
| Utiles Escolares               | 0.99      | 0.95   | 0.97     | 126     |
| Vinos y Licores                | 1.00      | 0.99   | 1.00     | 136     |
| accuracy                       |           |        | 0.96     | 2013    |
| macro avg                      | 0.96      | 0.96   | 0.96     | 2013    |
| weighted avg                   | 0.96      | 0.96   | 0.96     | 2013    |

### Anexo 3. Detalle de columnas, tipología y cantidad de registros

| #<br>Colum | Nombre de la columna | Non-Null<br>Count | Tipo de<br>dato    |
|------------|----------------------|-------------------|--------------------|
| 1          | PRODUCTO             | 16383             | object             |
| 2          | FECHA                | 16383             | datetime<br>64[ns] |
| 3          | IDUbicacion          | 16383             | int64              |
| 4          | CANTIDAD             | 16383             | float64            |
| 5          | UNIDAD               | 16383             | object             |
| 6          | TOTAL                | 16383             | int64              |
| 7          | subcategory          | 16383             | object             |
| 8          | Fecha                | 16383             | datetime<br>64[ns] |
| 9          | Dia                  | 16383             | object             |
| 10         | Tipo Dia             | 16383             | object             |
| 11         | IDDto                | 16383             | int64              |
| 12         | IDMpio               | 16383             | int64              |
| 13         | Ubicación            | 16383             | int64              |
| 14         | Dpto                 | 16383             | object             |

|    |                                                                                                                                |       |         |
|----|--------------------------------------------------------------------------------------------------------------------------------|-------|---------|
| 15 | Municipio                                                                                                                      | 16383 | object  |
| 16 | Subregión (SGR)                                                                                                                | 16383 | object  |
| 17 | Código Departamento                                                                                                            | 16383 | int64   |
| 18 | Código Entidad                                                                                                                 | 16383 | int64   |
| 19 | Cuarta medida correctiva al Código Nacional de Policía más impuesta en la entidad territorial                                  | 16348 | float64 |
| 20 | Fallecidos por siniestros viales                                                                                               | 16383 | float64 |
| 21 | Población total de hombres                                                                                                     | 16383 | float64 |
| 22 | Población total de mujeres                                                                                                     | 16383 | float64 |
| 23 | Porcentaje población de hombres                                                                                                | 16383 | float64 |
| 24 | Porcentaje población de mujeres                                                                                                | 16383 | float64 |
| 25 | Primera medida correctiva al Código Nacional de Policía más impuesta en la entidad territorial                                 | 16383 | float64 |
| 26 | Proporción en que son impuestas otras medidas correctivas al Código Nacional de Policía más impuesta en la entidad territorial | 16348 | float64 |
| 27 | Puntaje SISBEN: Promedio - Cabecera                                                                                            | 16383 | float64 |
| 28 | Quinta medida correctiva al Código Nacional de Policía más impuesta en la entidad territorial                                  | 16348 | float64 |
| 29 | Segunda medida correctiva al Código Nacional de Policía más impuesta en la entidad territorial                                 | 16348 | float64 |
| 30 | Tasa de extorsión por cada 100.000 habitantes                                                                                  | 16383 | float64 |
| 31 | Tasa de exámenes por presunto delito sexual a mujeres por cada 100 mil habitantes                                              | 14870 | float64 |
| 32 | Tasa de exámenes médico legales por presunto delito sexual contra adolescentes                                                 | 16008 | float64 |
| 33 | Tasa de exámenes médico legales por presunto delito sexual contra niños y niñas de 0 a 5 años                                  | 14971 | float64 |
| 34 | Tasa de exámenes médico legales por presunto delito sexual contra niños y niñas de 6 a 11 años                                 | 15854 | float64 |
| 35 | Tasa de homicidio intencional en hombres por cada 100.000 habitantes                                                           | 16383 | float64 |
| 36 | Tasa de homicidio intencional en mujeres por cada 100.000 habitantes                                                           | 15341 | float64 |
| 37 | Tasa de homicidio intencional en áreas rurales por cada 100.000 habitantes                                                     | 16383 | float64 |
| 38 | Tasa de homicidio intencional en áreas urbanas por cada 100.000 habitantes                                                     | 15844 | float64 |
| 39 | Tasa de homicidio intencional por cada 100.000 habitantes                                                                      | 16383 | float64 |
| 40 | Tasa de homicidios en accidentes de tránsito por cada 100.000 habitantes                                                       | 16383 | float64 |
| 41 | Tasa de homicidios en hombres por cada 100.000 habitantes                                                                      | 16383 | float64 |
| 42 | Tasa de homicidios en mujeres por cada 100.000 habitantes                                                                      | 16383 | float64 |
| 43 | Tasa de homicidios en áreas rurales por cada 100.000 habitantes                                                                | 16383 | float64 |
| 44 | Tasa de homicidios en áreas urbanas por cada 100.000 habitantes                                                                | 16383 | float64 |
| 45 | Tasa de homicidios por cada 100.000 habitantes                                                                                 | 16383 | float64 |

|    |                                                                                                |       |         |
|----|------------------------------------------------------------------------------------------------|-------|---------|
| 46 | Tasa de hurto a personas por cada 100.000 habitantes                                           | 16383 | float64 |
| 47 | Tasa de hurto común por cada 100.000 habitantes                                                | 16383 | float64 |
| 48 | Tasa de lesiones en accidentes de tránsito por cada 100.000 habitantes                         | 16383 | float64 |
| 49 | Tasa de muertes por eventos de transporte en adolescentes (12 a 17 años)                       | 11562 | float64 |
| 50 | Tasa de muertes por eventos de transporte en niños y niñas de 0 a 5 años                       | 5945  | float64 |
| 51 | Tasa de muertes por eventos de transporte en niños y niñas de 6 a 11 años                      | 6047  | float64 |
| 52 | Tasa de secuestro por cada 100.000 habitantes                                                  | 16383 | float64 |
| 53 | Tasa de suicidio de hombres por cada 100 mil habitantes                                        | 12654 | float64 |
| 54 | Tasa de suicidio de mujeres por cada 100 mil habitantes                                        | 12654 | float64 |
| 55 | Tasa de suicidios en niños y niñas de 6 a 11 años                                              | 2223  | float64 |
| 56 | Tasa de suicidios en niños y niñas de adolescentes (12 a 17 años)                              | 10215 | float64 |
| 57 | Tasa de violencia contra adolescentes 12 a 17 años                                             | 16383 | float64 |
| 58 | Tasa de violencia contra niños y niñas de 6 a 11 años                                          | 16383 | float64 |
| 59 | Tasa de violencia contra niños y niñas de primera infancia                                     | 16383 | float64 |
| 60 | Tasa de violencia de pareja a mujeres por cada 100 mil habitantes                              | 15371 | float64 |
| 61 | Tasa de violencia interpersonal por cada 100.000 habitantes                                    | 16383 | float64 |
| 62 | Tasa de violencia intrafamiliar en adolescentes (12 a 17 años)                                 | 13558 | float64 |
| 63 | Tasa de violencia intrafamiliar en niños y niñas de 0 a 5 años                                 | 12424 | float64 |
| 64 | Tasa de violencia intrafamiliar en niños y niñas de 6 a 11 años                                | 12569 | float64 |
| 65 | Tasa de violencia intrafamiliar por cada 100.000 habitantes                                    | 16383 | float64 |
| 66 | Tercera medida correctiva al Código Nacional de Policía más impuesta en la entidad territorial | 16348 | float64 |
| 67 | Índice de Interrelación de Problemáticas                                                       | 16383 | float64 |



---

|  |                        |       |      |     |       |      |     |
|--|------------------------|-------|------|-----|-------|------|-----|
|  | Red Neuronal Multicapa | 21.27 | 0.58 | 71% | 17.41 | 0.63 | 67% |
|  | LSTM                   | 22.99 | 0.51 | 69% | 17.40 | 0.63 | 55% |
|  | Keras Regresor         | 23.32 | 0.50 | 79% | 18.70 | 0.57 | 76% |