MODELO DE APRENDIZAJE AUTOMÁTICO PARA PROYECCION DE VENTAS DE LOS SERVICIOS PUBLICITARIOS EN EL METRO DE MEDELLÍN

Nombre de los estudiantes
Julio Tabares Álvarez
Código: 8993479
Sergio Villarreal Trujillo
Código: 8938622
Jhonny A. Cárdenas Rojas
Código: 8923048

Proyecto Aplicado para optar al título de
Magister en Ciencia de Datos

Director(a)
Isabel Cristina García Arboleda

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS

SANTIAGO DE CALI, JUNIO 9 DE 2025

TITULO: “MODELO DE APRENDIZAJE AUTOMÁTICO PARA PROYECCION DE VENTAS DE LOS SERVICIOS PUBLICITARIOS EN EL METRO DE MEDELLÍN”

1. ÉNFASIS: Sistemas y Computación
2. TIPO DE PROYECTO: Aplicado
3. ÁREA DE TRABAJO: Ciencias e ingeniería
4. ESTUDIANTE (S):
   - Julio Tabares Álvarez
   - Sergio Villarreal Trujillo
   - Jhonny A. Cárdenas Rojas
5. CORREO ELECTRÓNICO:
   - juliotabares@javerianacali.edu.co
   - Sergiovt@javeriancali.edu.co
   - Jhonnycr@javerianacali.edu.co
6. DIRECCIÓN Y TELÉFONO:
   - Av. 26 No 52 - 200 Apto 1147, Bello – Antioquia Tel: +57 312 766 3640
   - Calle 40# 54-30 Ciudad 2000 – Cali (Valle) Tel: +57 321 802 4358
   - Av. 2F norte # 45-83 – Cali (Valle) Tel: +57 316 394 4180
7. DIRECTOR:
   - Isabel Cristina García Arboleda
8. VINCULACIÓN DEL DIRECTOR (en la universidad): Planta
9. CORREO ELECTRÓNICO DEL DIRECTOR:
   - isabel.garcia@javerianacali.edu.co
10. GRUPO O EMPRESA QUE LO AVALA:
    - Metro de Medellín
11. PALABRAS CLAVE
    - *Pronóstico*, Aprendizaje automático, Optimización, Transporte público, Sector publicitario.
12. ODS QUE APLICA EL PROYECTO (Agenda 2030):
    - 8 trabajo decente y crecimiento económico
    - 9 industria, Innovación e infraestructura
13. FECHA DE INICIO (Desarrollo del proyecto): 18/05/2024
14. RESUMEN (máximo 400 palabras).

El proyecto "Modelo de Aprendizaje Automático para Proyección de Ventas de los Servicios Publicitarios en el Metro de Medellín" busca desarrollar un modelo predictivo basado en técnicas de aprendizaje automático para optimizar las estrategias comerciales del Metro de Medellín. El Metro, con aproximadamente 1.3 millones de usuarios diarios, busca aumentar su participación de ingresos no tarifarios al 15% para 2025, donde la publicidad juega un papel clave. Actualmente, la empresa enfrenta desafíos en la comercialización de

sus servicios publicitarios, por lo que este estudio propone una solución basada en la analítica de datos.

En el proyecto se utilizó la metodología CRISP-DM, la cual abordó cinco fases clave: comprensión del negocio, preparación de datos, modelado, proyección y validación. Se obtuvieron y se realizó la limpieza de las bases de datos de afluencia de pasajeros y ventas de publicidad desde 2020 hasta 2024, permitiendo identificar patrones de consumo y demanda.

El análisis inicial evidenció que la publicidad en estaciones, representan el 38.62% de las ventas, por lo que el proyecto decidió enfocarse en este segmento. Se evaluaron modelos de aprendizaje estadístico para seleccionar el que mejor optimice la predicción de las ventas con base en el desempeño y ajuste de cada modelo, con el fin de mejorar la toma de decisiones en la comercialización de este segmento. Además, se diseñó un tablero dinámico para el análisis de datos y la visualización de los resultados de los modelos.

Para la validación y selección del mejor modelo, se utilizaron criterios como la métrica CPM, el ajuste visual, tiempo de ejecución de los modelos, facilidad de implementación y cantidad de estaciones con mejor desempeño. Como resultado se obtuvo que Holt Winters 2 fue el mejor modelo teniendo en cuenta estos criterios.

Tabla de contenido

Tabla de ilustraciones

Listado de tablas

Listado de Ecuaciones

INTRODUCCIÓN

En el Metro de Medellín, para 2023 se movilizan aproximadamente 1.3 millones de usuarios diarios en un día laboral [2]. Actualmente, cuenta con dos tipos de ingresos operativos, los que corresponden a ingresos producto del transporte de personas (Ingresos tarifarios) y los que corresponden a negocios asociados (No tarifarios). Para el año 2025 tiene proyectado mejorar el porcentaje de participación de los ingresos no tarifarios, alcanzando un 15% del total de ingresos operativos.

Los ingresos por negocios asociados se componen de aprovechamiento del conocimiento y experiencia, y la explotación de la infraestructura del sistema. Dentro de estos últimos, se cuenta con servicios publicitarios los cuales vienen enfrentado una serie de desafíos, que no son exclusivos del Metro de Medellín, sino de la dinámica del mercado.

Para abordar este problema, el presente proyecto se apoyó en la ciencia de datos y, en particular, en técnicas de aprendizaje automático orientadas al pronóstico. Se propuso desarrollar modelos de series de tiempo que permitieran proyectar las ventas de servicios publicitarios en función de la afluencia de usuarios, usando herramientas como Holt-Winters, Prophet, ARIMA y redes neuronales tipo Elman.

El proceso de solución siguió la metodología CRISP-DM, incluyendo fases de comprensión del negocio, preparación de datos, modelado, evaluación y proyección. Se estructuraron bases de datos históricas de ventas y afluencia (2020-2024), se seleccionaron las estaciones con mayor relación entre ambas variables, y se entrenaron los modelos utilizando técnicas de validación y optimización de hiperparámetros. Además, se construyó un tablero en Power BI para la visualización y uso operativo de los resultados por parte de la empresa.

Como resultado, se identificó que el modelo Holt-Winters ofrecía el mejor balance entre precisión, facilidad de implementación y tiempos de ejecución. Este modelo fue validado en estaciones estratégicas como Poblado y San Antonio A, logrando proyecciones robustas que podrán apoyar la toma de decisiones comerciales del Metro de Medellín, incrementando así el potencial de ingresos no tarifarios.

1. CONTEXTUALIZACIÓN DEL PROYECTO

1.1. DEFINICIÓN DEL PROBLEMA

1.1.1. PLANTEAMIENTO DEL PROBLEMA

El Metro de Medellín es la Empresa de Transporte Masivo que sirve al Valle de Aburrá, zona en la que para el 2020 vivían aproximadamente 4 millones de colombianos distribuidos en los 10 municipios que contiene dicha zona [3]. En este sistema de transporte se movilizan aproximadamente 1.3 millones de usuarios en un día típico laboral. Dentro del Direccionamiento Estratégico de la Empresa, se tiene definida la Meta Grande y Ambiciosa (MEGA) como: "Ser a 2025 una Empresa innovadora con un crecimiento eficiente, articuladora de la movilidad como servicio, para conectar 1,3 millones de viajeros al día y con una participación de ingresos por negocios asociados del 15%".

Los negocios asociados que actualmente tiene la Empresa son: Aprovechamiento del conocimiento y gestión urbana por medio de la captura de valor en la explotación de los desarrollos urbanísticos. Dentro de estos negocios asociados se incluyen los servicios publicitarios, los cuales enfrentan una serie de desafíos sociales y de mercado, dadas las dinámicas de la sociedad y la tecnología que se emplea.

La ciencia de datos ofrece metodologías robustas para abordar la necesidad de proyectar ventas, especialmente a través del uso de modelos de pronóstico. En este caso, se busca optimizar la predicción de las ventas de servicios publicitarios ofrecidos por la Empresa, mediante el desarrollo de modelos de series de tiempo que integran tanto datos históricos como la afluencia de usuarios en las estaciones. Esta combinación permite generar proyecciones más precisas, facilitando la toma de decisiones informadas que contribuyan a mejorar la rentabilidad del negocio.

Lo anterior se alinea con la necesidad de la Gerencia de Desarrollo de Negocios del Metro de Medellín, quienes proyectan cumplir con el objetivo estratégico planteado de incrementar los ingresos por negocios asociados en un 15% para 2025. Es importante resaltar que para el año 2022 dichos ingresos ascendieron al 7% y para el año 2023 al 9% del total de ingresos operacionales percibidos por la Empresa [2].

En este contexto, el uso de modelos de series de tiempo para realizar proyecciones se presenta como una herramienta clave para apoyar el cumplimiento de dicha meta. Herramientas como redes neuronales, Holt Winters, Facebook Prophet y modelos autorregresivos integrados de media móvil (ARIMA) permiten generar pronósticos de ventas más precisos al analizar grandes volúmenes de datos históricos y adaptarse dinámicamente a los cambios del mercado, lo cual reduce la incertidumbre y evita depender de supuestos arbitrarios. Estos modelos también

facilitan la planificación operativa, presupuestal y comercial, al proyectar escenarios futuros que permiten anticipar variaciones en la demanda y asignar recursos de manera más eficiente. Además, su capacidad de automatización contribuye a disminuir costos operativos, optimizando la toma de decisiones basada en datos y reduciendo la carga manual y el riesgo de errores humanos.

1.1.2. FORMULACIÓN DEL PROBLEMA

Como interesados de este proyecto, es interesante conocer ¿de qué manera se puede aumentar los ingresos por servicios publicitarios según el comportamiento de los datos de las ventas y afluencia, en cada una de las estaciones y líneas? Para dar respuesta a esta pregunta se hace necesario determinar:

- ¿Cómo funciona el negocio asociado por servicios publicitarios en el Metro de Medellín?
- ¿Se cuenta con la base de datos necesaria para el desarrollo del proyecto?
- ¿Se encuentran adecuadas y preparadas las bases de datos que tiene la Empresa para la modelación por medio de los precios de los servicios publicitarios?
- Luego de definir algunos modelos con base en la metodología trabajada ¿cuál o cuáles se ajustan y son válidos para las necesidades y realidad del negocio?
- ¿Qué medio se empleará para que la Empresa pueda hacer uso del modelo, visualizando los resultados obtenidos?
- Luego de realizar pruebas de validación ¿El modelo requiere de ajustes según la realidad y circunstancias del negocio?
- ¿Qué estaciones o líneas son un potencial de mejoramiento de acuerdo con el comportamiento de las variables de afluencia y ventas en el Metro de Medellín?

1.2. OBJETIVOS DEL PROYECTO

1.2.1. OBJETIVO GENERAL

Desarrollar un modelo de proyección de ventas para los servicios publicitarios del Metro de Medellín, utilizando un algoritmo de aprendizaje automático que analice datos históricos de ventas y permita optimizar las estrategias comerciales a través de proyecciones precisas.

1.2.2. OBJETIVOS ESPECÍFICOS

1) Comprender el funcionamiento del negocio del Metro de Medellín mediante un estudio detallado de sus servicios publicitarios para identificar y definir las variables relevantes que influyen en las ventas de los servicios publicitarios.
2) Preparar los datos necesarios sobre las variables definidas utilizando técnicas de limpieza y organización de datos para asegurar su calidad y pertinencia en el análisis posterior.

3) Analizar diferentes modelos de aprendizaje automático mediante la evaluación de diferentes enfoques teóricos y metodológicos y elegir el o los modelos que mejor se ajustan a las necesidades del negocio.
4) Desarrollar un tablero dinámico mediante herramientas de visualización de datos que permita realizar análisis de la información que se genere en función del modelo de aprendizaje automático y los resultados de este.
5) Validar el o los modelos que más se ajusten a las necesidades y realidad del negocio.

1.3. MARCO DE REFERENCIA

El Metro de Medellín es uno de los principales sistemas de transporte público en Colombia, y representa un lugar estratégico para la publicidad debido a su alta afluencia de pasajeros. El Metro enfrenta el desafío de mejorar los ingresos publicitarios en todo su sistema, con el propósito de aportar al cumplimiento de la MEGA, tomando en cuenta factores como la ubicación, la afluencia de usuarios y las ventas históricas. En este contexto, el aprendizaje automático surge como una herramienta poderosa para analizar grandes volúmenes de datos y realizar predicciones precisas sobre los precios de los servicios publicitarios.

A continuación, se dará una breve explicación de los temas que se relacionan con el desarrollo del proyecto, teniendo en cuenta como base fundamental las proyecciones de venta por medio de modelos de series de tiempo y aprendizaje automático.

1.3.1. MARCO TEÓRICO

1.3.1.1. MACHINE LEARNING

El *Machine Learning* o aprendizaje automático es la ciencia de programar computadores para que puedan aprender y tomar decisiones basadas en datos. Arthur Samuel definió el *Machine Learning* en 1959 como “el campo de estudio que da a las computadoras la capacidad de aprender sin ser explícitamente programadas" [4]. Lo cual quiere decir que el *Machine Learning* utiliza algoritmos para identificar patrones en los datos y realizar predicciones o decisiones basadas en esos patrones, siendo capaces de realizar tareas como clasificación y regresión.

El *Machine Learning* se ha vuelto fundamental debido a su capacidad para analizar grandes volúmenes de datos y extraer información valiosa. Algunas de las razones clave por las que se utiliza esta ciencia incluyen:

- Automatización de tareas: Permite automatizar tareas repetitivas y basadas en datos, mejorando la eficiencia y reduciendo el error humano.

- Predicciones precisas: Los modelos de *Machine Learning* pueden hacer predicciones precisas sobre futuros eventos o comportamientos basados en datos históricos.
- Adaptabilidad: Los sistemas pueden adaptarse y mejorar con el tiempo a medida que se introducen más datos.

1.3.1.2. FORECASTING

El pronóstico mediante Aprendizaje Automático (También conocido como *Machine Learning Forecasting*) es un proceso que es ampliamente utilizado en múltiples industrias para proyectar el comportamiento de una o varias variables, según la necesidad del negocio. Esta emplea modelos computacionales que aprenden patrones a partir de datos históricos, con el fin de realizar predicciones futuras de las variables de interés. Este se basa en el uso de algoritmos y modelos diseñados para el manejo de grandes volúmenes de datos, con el fin de detectar tendencias complejas y estacionalidades, las cuales no podrían ser identificadas por modelos estadísticos tradicionales o con la visualización de los datos sin ser procesados.

Lo anterior ha sido un logro importante en la ciencia de datos, por medio de la implementación de diferentes técnicas de Aprendizaje Automático, las cuales son ampliamente utilizadas por diferentes sectores de la sociedad, como lo son: Análisis de series de tiempo, Árbol de Decisiones (Decisition Tree), Redes Neuronales Recurrentes, Modelos Atencionales entre otros [5].

1.3.1.3. SERVICIOS PUBLICITARIOS

Los servicios publicitarios, son una forma de comunicación comercial entre una empresa y el consumidor, con el propósito de promover a su vez la venta de servicios, productos o ideas. Este tipo de servicio es esencial para impulsar el crecimiento de sectores de la sociedad, generando un impacto que puede llegar a ser determinante entre el éxito o fracaso de un producto o servicio [6].

Los servicios publicitarios han evolucionado a lo largo de la historia de la humanidad, conforme han evolucionado las tecnologías y los medios de comunicación, adaptándose a las dinámicas del mercado, las preferencias y necesidades de los clientes.

Para las empresas que ofrecen este tipo de servicios, es fundamental tener claridad sobre cómo se está comportando las variables ligadas a su mercado, con el fin de definir estrategias administrativas y operativas, que aporten al crecimiento en la utilización de los servicios publicitarios. Es ahí cuando se puede hacer uso del Machine Learning Forecasting.

1.3.1.4. METODOLOGIA CRISP-DM

La metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) es un modelo de proceso ampliamente utilizado para guiar proyectos de ciencia de datos. Esta metodología proporciona un marco estructurado que ayuda a las organizaciones a extraer conocimiento y valor de sus datos. Se compone de seis fases principales, cada una de las cuales desempeña un papel crucial en el éxito del proyecto [7].

El proceso de un proyecto de ciencia de datos bajo la metodología CRISP se inicia con la comprensión del negocio, cuyo objetivo es identificar los requisitos y metas clave para asegurar que el análisis genere valor real; esto implica definir los objetivos del proyecto, entender el contexto organizacional, establecer criterios de éxito y desarrollar un plan. A continuación, en la etapa de comprensión de los datos, se busca familiarizarse con la información disponible mediante la recolección, descripción, exploración y verificación de su calidad. Luego, en la preparación de los datos, se transforman los datos para que sean adecuados para el modelado, lo cual incluye su selección, limpieza, construcción, integración y formateo. En la fase de modelado, se aplican técnicas analíticas para descubrir patrones relevantes, definiendo estrategias de prueba, construyendo modelos y evaluando su desempeño. Posteriormente, la etapa de evaluación permite verificar si los modelos cumplen con los objetivos del negocio y son viables para su implementación, analizando sus resultados y determinando los siguientes pasos. Finalmente, en la fase de despliegue, se implementan los modelos en el entorno operativo, asegurando su monitoreo, mantenimiento y una adecuada documentación del proceso y los resultados obtenidos.

CRISP-DM es valorado por su flexibilidad y su enfoque iterativo, lo que permite a los equipos de datos refinar y mejorar sus modelos a medida que avanzan en el proyecto. Esta metodología es utilizada en diversas industrias debido a su eficacia para abordar problemas complejos de ciencia de datos y asegurar la alineación con los objetivos empresariales.

1.3.1.5. MODELO FACEBOOK PROPHET

En el contexto de la predicción de series temporales, Facebook Prophet se ha consolidado como una herramienta robusta y flexible para la elaboración de pronósticos a gran escala. Prophet es un modelo aditivo que descompone la serie temporal en tres componentes principales: tendencia, estacionalidad y eventos especiales o festivos. Esta estructura permite capturar dinámicas complejas y patrones recurrentes, como ciclos semanales o anuales, además de eventos puntuales que afectan el comportamiento de la serie[8].

Una de las principales ventajas de Prophet es su facilidad de ajuste e interpretación, lo que permite a analistas sin formación avanzada en estadística intervenir en la configuración del

modelo. La componente de tendencia puede modelarse mediante crecimiento logístico saturado o mediante una función lineal segmentada con puntos de cambio (changepoints), los cuales pueden ser definidos manualmente o seleccionados automáticamente. Las estacionalidades se representan a través de series de Fourier, lo que proporciona flexibilidad para capturar patrones periódicos complejos. Adicionalmente, Prophet permite incluir efectos específicos de días festivos y eventos especiales, aspecto crucial para series que presentan variaciones significativas asociadas a calendarios[8].

El modelo está diseñado para ser escalable y eficiente, permitiendo su aplicación en escenarios donde se requieren múltiples pronósticos simultáneamente. Esta escalabilidad, junto con la posibilidad de incorporar conocimiento experto mediante la intervención humana ("analyst-in-the-loop"), facilita la generación de pronósticos confiables y adaptables a diferentes contextos empresariales y de investigación.

1.3.1.6. MODELO ELMAN

La red neuronal Elman es una arquitectura recurrente diseñada para modelar datos secuenciales y capturar dependencias temporales entre observaciones. Su estructura se compone de una capa oculta con retroalimentación interna (capa de contexto), que permite almacenar la información de estados anteriores y utilizarla en pasos futuros. Esta capacidad de memoria hace que sea especialmente útil para el análisis de series temporales, donde el comportamiento futuro depende mucho de patrones y variaciones pasadas [9]. Gracias a esto, es una herramienta ampliamente utilizada en tareas de predicción donde el comportamiento futuro depende de datos históricos.

1.3.1.7. MODELO ARIMA

El modelo ARIMA (Autoregressive Integrated Moving Average) es un método clásico para el pronóstico de series temporales introducido por Box y Jenkins, basado en la combinación de tres componentes que capturan la dinámica de los datos de order [10]:

- p: autorregresivo (AR)
- d: integración (I)
- q: media móvil (MA)

Cuando las series presentan estacionalidad, se extiende a SARIMA(p,d,q)(P,D,Q)s con periodo s, identificado como *seasonal_order*:

- P: autorregresivo (AR)
- D: integración (I)
- Q: media móvil (MA)
- S: periodo de la estacionalidad

1.3.1.8. MODELO HOLTWINTERS

El modelo HoltWinters es una técnica de suavizamiento exponencial extendido que se utiliza para pronosticar series temporales que presentan tendencia y estacionalidad. Es una extensión del modelo de Holt (que solo considera tendencia) y del suavizamiento exponencial simple (que solo sirve para series sin tendencia ni estacionalidad)[10].

Se basa en la idea de asignar más peso a las observaciones recientes y decreciente peso a las más antiguas, por medio de factores de suavizamiento. Existen dos variantes principales: la primera variante es el aditivo el cual se emplea para analizar cuando la estacionalidad tiene un efecto constante a lo largo del tiempo (se suma), el segundo es el multiplicativo y se emplea cuando la estacionalidad tiene un efecto proporcional a la magnitud de la serie (se multiplica).

1.3.1.9. METRICAS DE DESEMPEÑO

Para garantizar una evaluación rigurosa de los modelos predictivos desarrollados, se seleccionó un conjunto de métricas que permiten analizar el desempeño de los diferentes modelos utilizados. En principio, se utilizaron cuatro métricas ampliamente reconocidas en el ámbito de la predicción de series de tiempo: el Error Absoluto Medio (MAE), la Raíz del Error Cuadrático Medio (RMSE), el Error Porcentual Absoluto Medio (MAPE) y el Indicador Compuesto de Desempeño (CPM). Sin embargo, durante la evaluación de los modelos se identificó que el MAPE para varios modelos arrojaba valores que tendían a infinito, por lo que dejo de ser una métrica confiable para el análisis.

Adicional, se hizo una prueba de una estación en el modelo Facebook Prophet para evaluar porque el MAPE arrojaba resultados con tendencia a infinito y se encontró que estos resultados obedecen a la fórmula de la métrica y la naturaleza de los datos utilizados.

- **Error Porcentual Absoluto Medio**

Ecuación 1.Error Porcentual Absoluto Medio

$$MAPE = \frac{\frac{\sum_{t=1}^{n} |y_t - F_t|}{y_t} * 100}{n}$$

Por consiguiente, se procedió a remplazar la métrica MAPE por el Error Porcentual Absoluto Medio simétrico (SMAPE), la cual funciona mejor para cuando los datos de prueba son 0, pues no hace que la métrica se indetermine.

Finalmente, las métricas seleccionadas permiten evaluar tanto la precisión general como la sensibilidad del modelo ante errores significativos, asegurando una valoración robusta en diversos contextos y escalas.

- **Error Absoluto Medio (MAE)**

Ecuación 2. Error Absoluto Medio

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - F_t|$$

El MAE mide el promedio de las diferencias absolutas entre los valores reales y los valores predichos, esta métrica indica el tamaño medio del error cometido por el modelo, y para el caso del proyecto se expresa en millones, por lo que permite una mejor interpretación en el contexto financiero de las ventas. Una de sus principales ventajas es su robustez frente a valores atípicos moderados, ya que todos los errores se ponderan por igual, por lo que lo convierten en un buen indicador para el proyecto en la precisión cuando se quiere comprender la magnitud típica del error sin incurrir en valores desproporcionados debido a errores extremos.

- **Raíz del Error Cuadrático Medio (RMSE)**

Ecuación 3. Raíz del Error Cuadrático Medio

$$RMSE = \sqrt{\sum \frac{(F_t - Y_t)^2}{n}}$$

El RMSE representa la raíz cuadrada del promedio de los errores cuadrados. A diferencia del MAE, esta métrica penaliza severamente los errores grandes, ya que los errores se elevan al cuadrado antes de promediarse. Al igual que el MAE se expresa en millones para mantener la coherencia de escala utilizadas en las métricas y en el proyecto. Esta métrica fue seleccionada debido a su capacidad para destacar desviaciones más pronunciadas, lo cual es útil cuando se busca detectar si el modelo incurre en errores significativos en momentos específicos.

- **Error Porcentual Absoluto Medio Simétrico (SMAPE)**

Ecuación 4. Error Porcentual Absoluto Medio Simétrico

$$SMAPE = \frac{1}{n} * \sum_{t=1}^{n} \frac{|Y_t - F_t|}{(|Y_t| + |F_t|)/2}$$

El SMAPE se utiliza como una métrica porcentual y simétrica que mide el error relativo entre las predicciones y los valores reales para evitar inflar los errores cuando los valores reales son cercanos o iguales a cero, por lo que se adapta para la evaluación de los resultados de este

proyecto con la información suministrada. Este indicador es útil cuando se quiere comparar el desempeño del modelo entre estaciones con diferentes escalas de ventas, ya que proporciona una visión relativa del error expresada en porcentajes, además, su simetría evita sesgos a favor de sobreestimaciones o subestimaciones.

- **Indicador Compuesto de Desempeño (CPM)**

Ecuación 5. Indicador de desempeño compuesto (CPM)

$$CPM_j = Z(MAE_j) + Z(RMSE_j) + Z(SMAPE_j)$$

El CPM es un índice que consolida múltiples métricas de desempeño en un único valor, para el proyecto se calcula el promedio de los valores normalizados de las métricas anteriormente mencionadas (MAE, RMSE y SMAPE), asegurando que cada componente contribuye equitativamente a la evaluación general. La inclusión de esta métrica permite comparar de forma homogénea el desempeño entre diferentes modelos y estaciones, el CPM facilita la toma de decisiones al ofrecer un criterio compuesto unificado.

1.3.2. ANTECEDENTES

1.3.2.1. IMPACTO DE LA PUBLICIDAD

Existen una gran cantidad de casos que demuestran la importancia de los servicios publicitarios para el crecimiento de las industrias y las empresas, tomaremos un caso en particular que ha sido referente en la actualidad. En 1984 la marca Apple difundió un anuncio llamado “1984” en medio de la final de la Liga de Futbol Americano, el cual cambió la suerte de la Compañía debido a una serie de factores innovadores, inteligentes y de gran impacto. Actualmente, Apple es la marca más valorada del mundo, de acuerdo con la publicación anual realizada por la revista Forbes [11], lo que demuestra la importancia y el impacto a largo plazo que puede llegar a generar los servicios publicitarios.

1.3.2.2. IMPORTANCIA DEL FORECASTING

Dado que el Forecasting es un proceso de predicción del comportamiento de una variable, basándose en los datos históricos y las estacionalidades de dicha variable, este es ampliamente usado en diferentes industrias como las finanzas, manufacturera, meteorología, logística, entre otros. Tal como lo señala la consultora DREW: La previsión de la demanda juega un papel de suma importancia en todas las empresas, sin importar cuál sea la industria a la que pertenece [12].

El pronóstico de ventas es una herramienta esencial en la toma de decisiones empresariales, permitiendo anticipar la demanda futura y optimizar estrategias de producción, inventario y

comercialización. Con el avance de la ciencia de datos y la inteligencia artificial, los métodos tradicionales han evolucionado hacia enfoques más sofisticados que combinan el análisis de series temporales con modelos de aprendizaje automático y el aprovechamiento de datos no estructurados. En este contexto, los pronósticos pueden beneficiarse del uso de múltiples fuentes de información, desde datos históricos hasta tendencias en redes sociales y factores externos como la economía y las políticas gubernamentales. A continuación, se presentan tres estudios que ejemplifican diferentes aplicaciones de técnicas de pronóstico en diversos sectores, destacando el uso de modelos estadísticos y de Machine Learning para mejorar la precisión y adaptabilidad de las predicciones en entornos de datos limitados o altamente dinámicos.

1.3.2.3. A MACHINE LEARNING-BASED FRAMEWORK FOR FORECASTING SALES OF NEW PRODUCTS WITH SHORT LIFE CYCLES

Este estudio propone un marco cuantitativo para predecir la demanda de nuevos productos con ciclos de vida cortos, utilizando técnicas avanzadas de aprendizaje automático y métodos estadísticos. Se enfoca en la clasificación de productos similares mediante Clustering y asignación de nuevos productos a estos grupos utilizando métodos como la integración y el "Dynamic Time Warping" (DTW). Además, emplea técnicas de aumento de datos para mejorar la capacidad de predicción con datos históricos limitados. La comparación de diferentes métodos mostró que el modelo ARIMAX tiene un mejor desempeño que las redes neuronales profundas en la mayoría de los casos, aunque los modelos basados en aprendizaje profundo demostraron mayor robustez ante datos ruidosos [13].

1.3.2.4. ON PREDICTIVE MODELING OF TWITTER-BASED SALES DATA USING A NEW PROBABILISTIC MODEL AND MACHINE LEARNING METHODS

Este trabajo introduce un nuevo modelo probabilístico basado en la distribución inversa de Weibull con una transformación coseno para modelar datos de ventas generados a partir de publicidad en Twitter. El estudio utiliza técnicas de regresión de soporte vectorial (SVR) y perceptrón multicapa (MLP) para realizar pronósticos de corto y mediano plazo, demostrando que el MLP es más eficaz en predicciones a corto plazo, mientras que SVR es más preciso en pronósticos a mayor plazo. La metodología empleada permite capturar el impacto de las campañas publicitarias digitales en las ventas, lo que proporciona una perspectiva innovadora en la intersección del análisis de datos y la publicidad en redes sociales [14].

1.3.2.5. NEW ENERGY VEHICLES SALES FORECASTING USING MACHINE LEARNING

Este estudio analiza la predicción de ventas de vehículos eléctricos mediante modelos de aprendizaje automático y técnicas de descomposición. Se evalúan múltiples enfoques, incluyendo modelos basados en redes neuronales recurrentes (RNN), ARIMAX y métodos híbridos que integran descomposición de series temporales. Los resultados muestran que la combinación de múltiples fuentes de datos, como volumen de búsqueda en línea, reseñas de clientes y datos

macroeconómicos, mejora significativamente la precisión del pronóstico. Se destaca que los modelos de descomposición y ensamblaje ofrecen un rendimiento superior en comparación con los modelos individuales [15].

1.4. METODOLOGIA

Para la definición de una metodología apropiada es importante definir el tipo de proyecto que se llevara a cabo. Para ello, se precisó que el proyecto es de carácter teórico-practico, pues el objetivo general pretende desarrollar un modelo de aprendizaje automático que permita la proyección de ventas de los servicios de publicidad en el metro de Medellín; mediante datos históricos del flujo de pasajeros en las estaciones del metro y las ventas de los servicios publicitarios a través del tiempo.

Realizando la revisión de los objetivos específicos se puede observar que estos se encuentran en su mayoría alineados con las etapas de la metodología CRISP-DM *(Cross-Industry Standard Process for Data Mining),* la cual es la metodología seleccionada para el desarrollo del proyecto.

A continuación, se presenta mediante la Tabla 1 las actividades asociadas a cada uno de los objetivos y etapas de la metodología seleccionada.

Tabla 1. Metodología y actividades

Etapas (CRISP-DM)	Objetivo	Actividad
Entendimiento del negocio	Comprender el funcionamiento del negocio del Metro de Medellín mediante un estudio detallado de sus servicios publicitarios para identificar y definir las variables relevantes que influyen en la proyección de ventas.	Reunión con Coordinador de inteligencia de Negocios en el metro de Medellín.
		Obtener el portafolio de servicios publicitario del metro de Medellín
Entendimiento de los datos		Realizar un estudio de mercado del sector publicitario en la ciudad de Medellín
		Definir las variables a utilizar en el proyecto
Preparación de los datos	Preparar los datos necesarios sobre las variables definidas utilizando técnicas de limpieza y organización de datos para asegurar su	Obtener los datos de las variables previamente definidas
		Realizar un análisis exploratorio de los datos de

Etapas (CRISP-DM)	Objetivo	Actividad
	calidad y pertinencia en el análisis posterior.	cada una de las variables definidas
		Realizar limpieza de los datos obtenidos de cada una variable
	Desarrollar un tablero dinámico mediante herramientas de visualización de datos que permita realizar análisis de la información a utilizar del modelo de aprendizaje automático y los resultados de este.	Validar la cantidad y calidad de la información
		Desarrollar el tablero para la visualización y el análisis de la información
Modelado	Analizar diferentes modelos de aprendizaje automático de series de tiempo tales como: ARIMA, Holt Winters, Facebook Prophet y ELMAN, para determinar cuál se ajusta mejor a las necesidades del negocio y los datos.	Investigar qué tipo de modelos de aprendizajes automático se ajustan mejor a temas de proyección-pronóstico de ventas
		Escoger los modelos a utilizar para el análisis
		Definir los datos de entrenamiento y prueba
		Realizar el entrenamiento de los modelos seleccionados
Evaluación	Validar el o los modelos que más se ajusten a las necesidades y realidad del negocio.	Evaluar los resultados de los modelos
		Escoger el mejor modelo
		Realizar recomendaciones y posibles pasos a seguir
		(Realizar simulación sobre los resultados de prestar los servicios publicitarios utilizando los datos arrojados por el modelo)

2. COMPRENSIÓN DEL NEGOCIO

2.1. SOBRE LA EMPRESA

De acuerdo con lo documentado en su sitio web oficial: “El Metro de Medellín es una sociedad pública sujeta al régimen de las empresas industriales y comerciales del estado, compite de forma directa con el sector privado, por lo que goza de autonomía administrativa y financiera. Sus socios son, en igual proporción, el Departamento de Antioquia y el Municipio de Medellín. Asimismo, es una entidad descentralizada adscrita al Municipio de Medellín, haciendo parte de su Modelo de Gestión de Conglomerado Público” [16].

2.1.1. RED METRO

La principal fuente de ingreso de la Empresa el ofrecer el servicio de transporte de personas, el cual se desarrolla por medio de los siguientes modos de transporte y sus líneas asociadas:

Ilustración 1. Mapa esquematico del Metro de Medellín tomado de [16]

2.1.2. OTRAS FUENTES DE INGRESO

La Empresa cuenta con la marca Negocios Metro, la cual es una unidad de negocios responsable de los negocios no tarifarios de la Empresa. Se encarga de diversificar y fortalecer las fuentes de ingreso distintas a las tarifas del servicio de transporte. Las principales líneas de negocio, se acuerdo al sitio web oficial son [17]:

- **Ecosistemas de pago:** soluciones de pago rápidas, prácticas y seguras, que brindan acceso instantáneo a múltiples empresas de transporte a través de tarjetas y aplicaciones móviles
- **Consultorías y formaciones:** compartimos nuestro conocimiento y experiencia para contribuir a la transformación de organizaciones y territorios, aportando mejores prácticas que se traducen en mayor calidad de vida.
- **Captura de Valor:** enfocados en el aprovechamiento de la Infraestructura existente y en los corredores de movilidad proyectados formulación de operaciones urbanas y desarrollos inmobiliarios.

2.1.3. PORTAFOLIO DE SERVICIOS PUBLICITARIOS

La Empresa cuenta con una serie de servicios publicitarios los cuales se describen a continuación:

Tabla 22. Resumen portafolio servicios publicitarios Metro de Medellín

Elemento publicitario	Descripción
BTL	Mensajes de audio que se transmiten en trenes y tranvía.
Buses	Vallas publicitarias en estaciones de buses
Cabinas	calcomanías en las cabinas del Metrocable
Estación	Publicidad en estaciones
Megamuppis	Variedad especifica de pantalla en estaciones
Pantalla	Pantallas instaladas en estaciones
Programática	Publicidad personalizada y programada en pantallas digitales
Tranviario	Publicidad adentro de los tranvías
Tren CAF	calcomanías o empapelado de trenes nuevos
Tren MAF	calcomanías o empapelado de trenes viejos

2.2. OBTENCIÓN DE LOS DATOS Y DEFINICIÓN DE VARIABLES

Dado que el proyecto se desarrolla con información de bases de datos, el equipo de trabajo tuvo una primera reunión con la Coordinación de Inteligencia de Negocios y el equipo que lidera los servicios publicitarios en el Metro de Medellín, con el ánimo de obtener toda la información necesaria para el desarrollo del proyecto. Como resultado se obtuvieron las siguientes bases de datos:

- Afluencia de Usuarios: Base de datos que contiene el flujo de usuarios en toda la Red Metro entre el año 2018 al 2024. Para efectos del trabajo se acotó la información entre el año 2020 y el año 2024, dado que la base de datos ventas cuenta con registros para dicho periodo.
- Ventas: Esta base de datos contiene las facturas de las ventas realizadas en los servicios publicitarios a lo largo de los años 2020 a 2024.

2.2.1. AFLUENCIA DE USUARIOS

La Empresa cuenta con un Dashboard el cual se emplea para proyectar unas visualizaciones en función del comportamiento de la afluencia de usuarios a lo largo de la red Metro. Luego de analizar dicho tablero, ser determinó que la información contenida en este y proyectada de forma mensual puede ser útil para el alcance del proyecto, dado que este permitía obtener la información de afluencia para cada estación de la Red Metro, con periodicidad mensual.

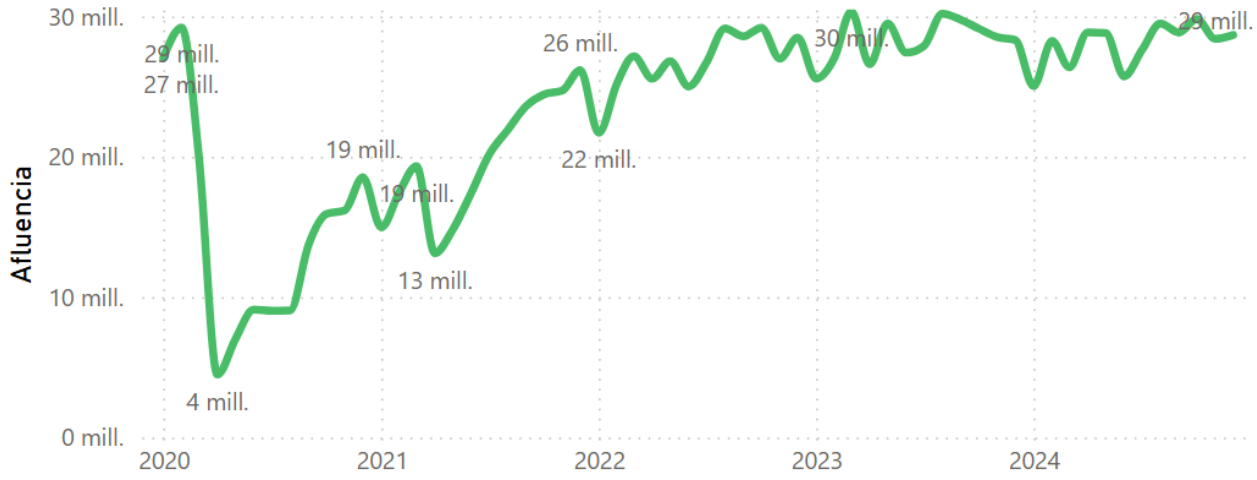

Ilustración 2. Comportamiento de la afluencia en el Metro de Medellín, entre el 2020 y 2024.

2.2.2. VENTAS POR SERVICIOS PUBLICITARIOS

Paralelamente se realizó una segunda reunión con el personal del Metro, específicamente el departamento de Gestión urbana, para obtener acceso a la información correspondiente al portafolio de servicios publicitarios.

Como se mencionó en el apartado 2.2, Se logró obtener acceso a una base de datos (tipo base de datos de hechos) en la cual se discriminaba la facturación relacionada con los servicios publicitarios. Esta base de datos es administrada por el aliado estratégico del Metro de Medellín, el cual tiene bajo su responsabilidad el administrar y ofertar los elementos publicitarios.

3. PREPARACIÓN DE LOS DATOS

La preparación de los datos es fundamental para el desarrollo del proyecto, por lo que el equipo de trabajo se enfocó en las siguientes actividades, con el propósito de poder contar con la información necesaria y suficiente para el desarrollo del proyecto, teniendo en cuenta las técnicas y teoría aprendida a lo largo de la Maestría.

3.1. ANÁLISIS EXPLORATORIO DE LOS DATOS

Inicialmente, con el conocimiento obtenido de la empresa, el equipo desarrolló un diccionario de datos para cada una de las fuentes de información, el cual sirvió para entender que representa cada una de las variables obtenida. A continuación, se presenta el diccionario de datos:

3.1.1. BASE DE DATOS DE AFLUENCIA

Tabla 33. Diccionario de datos BD Afluencia

Diccionario de Datos		
Nombre de la variable	**Tipo de Variable**	**Descripción**
Fecha	Fecha	Fecha en la que se registra la afluencia. Esta variable tiene una frecuencia mensual.
Modo	Texto/Categórica nominal	Hace referencia al modo de transporte del Metro de Medellín (Cables, Buses, Trenes y Tranvía)
Línea	Texto/Categórica nominal	Hace referencia a la ruta del modo de transporte
Estación	Texto/Categórica nominal	Hace referencia a la estación donde se presenta la afluencia
Usos (laboral)	Entero/Numérica	Cantidad de validaciones realizadas por los usuarios en cada una de las líneas. Estos son contabilizados para días laborales.

Viajes (laboral)	Entero/Numérica	Cantidad de viajes que realiza los usuarios desde que ingresan a una de las estaciones hasta que salen del sistema Metro. Estos son contabilizados para días laborales.
Usos (sábado)	Entero/Numérica	Cantidad de validaciones realizadas por los usuarios en cada una de las líneas. Estos son contabilizados para días laborales y sábados.
Viajes (sábado)	Entero/Numérica	Cantidad de viajes que realiza los usuarios desde que ingresan a una de las estaciones hasta que salen del sistema Metro. Estos son contabilizados para sábados
Usos (Domingo y Festivo)	Entero/Numérica	Cantidad de validaciones realizadas por los usuarios en cada una de las líneas. Estos son contabilizados para domingos y festivos.
Viajes (Domingo y festivo)	Entero/Numérica	Cantidad de viajes que realiza los usuarios desde que ingresan a una de las estaciones hasta que salen del sistema Metro. Estos son contabilizados para domingos y festivos.
Usos (Total)	Entero/Numérica	Total de usos realizados en el sistema incluyendo las trasferencias con otros modos de transporte en un mismo viaje. Incluye laboral, sábado, domingo y festivo.
Viajes (Total)	Entero/Numérica	Total de viajes realizados en el sistema. Incluye laboral, sábado, domingo y festivo.

3.1.2. BASE DE DATOS DE VENTAS

Tabla 44. Diccionario de datos BD Ventas

Diccionario de Datos		
Variable	**Tipo de Variable**	**Descripción**
Fecha	Date	Corresponde a la fecha de facturación contable
Ejecutivo	Texto/Categórica nominal	Persona que realiza el registro contable
Detalle de la factura	Texto/Categórica nominal	Descripción del movimiento contable
Cantidad	Numérica	Cantidad de facturas realizadas por servicio
Valor Unitario	Numérica	Valor del servicio por unidad
Valor Total	Numérica	Valor del servicio total
Descripción del ítem	Texto/Categórica nominal	Descripción general de la factura
Cliente al que se le factura	Texto/Categórica nominal	Es la persona natural o jurídica que paga el servicio publicitario
Tema	Texto/Categórica nominal	Tema del servicio publicitario
Cliente campaña	Texto/Categórica nominal	Es la persona natural o jurídica que responsable del servicio publicitario
Año	Entero/Numérica	Año de la facturación
Mes	Entero/Numérica	Mes de la facturación

Diccionario de Datos		
Variable	**Tipo de Variable**	**Descripción**
Ciudad	Texto/Categórica nominal	Ciudad donde se presta el servicio publicitario
Expr_04, Expr_05 y Prefijo	Texto/Categórica nominal	Conceptos contables del metro de Medellín
Estación	Texto/Categórica nominal	Hace referencia al lugar o estación donde se prestó el servicio publicitario
Elemento	Texto/Categórica nominal	Es el elemento publicitario utilizado
Concepto	Texto/Categórica nominal	Tipo de servicio por producción o alquiler*
Grupo	Texto/Categórica nominal	Existen campañas que se manejan por medio de centrales y otras directamente con el cliente, aquí se detalla esta información

*Según los expertos, producción Son las producciones que se le cobran a cliente por sus publicidades. Por otra parte, alquiler es el valor que se cobra por el arrendamiento del espacio, en este caso a metro se le paga un determinado monto.

3.2. LIMPIEZA Y TRANSFORMACIÓN DE LAS BASES DE DATOS

En este apartado se describe todo el proceso de limpieza y transformación de datos que se realizó sobre las dos bases de datos que hacen parte del proyecto.

3.2.1. BASE DE DATOS DE AFLUENCIA

3.2.1.1. TRANSFORMACIÓN DE LA BASE DE DATOS

Inicialmente se tuvo que realizar un proceso de transformación de los datos para la base de datos de afluencia, dado que la información adquirida era una tabla tipo pivote y se requiere como una tabla estructurada. Por tal razón, por medio de Python se generó un código el cual permitió realizar la transformación de la base de datos.

Luego de analizar la información de las tablas tipo pivote, se determinó que la información a emplear sería usos(total) por estación y mes, dado que este indica la afluencia de los usuarios en cada estación. El resto de las métricas sobre cada tabla, son datos que emplea la Empresa para el manejo del negocio, pero que no representan mayor utilidad para el alcance del proyecto.

Durante la transformación de los datos se identificaron una serie de dificultades las cuales fueron sorteadas positivamente, tal como:

- Entre 2020 y 2023 se realizó una incorporación de líneas y estaciones a la Red Metro, con motivo de la ampliación de ésta. Por tal razón, algunas estaciones tienen datos faltantes en la afluencia para todo el periodo de tiempo que se analizará en el proyecto. Un ejemplo de esto es las estaciones de la Línea P, que comenzó a operar en junio de 2021, por lo que para dicha línea y sus estaciones, no se tendrá registro de afluencia de usuarios entre 2020.01.01 y 2021.05.31.

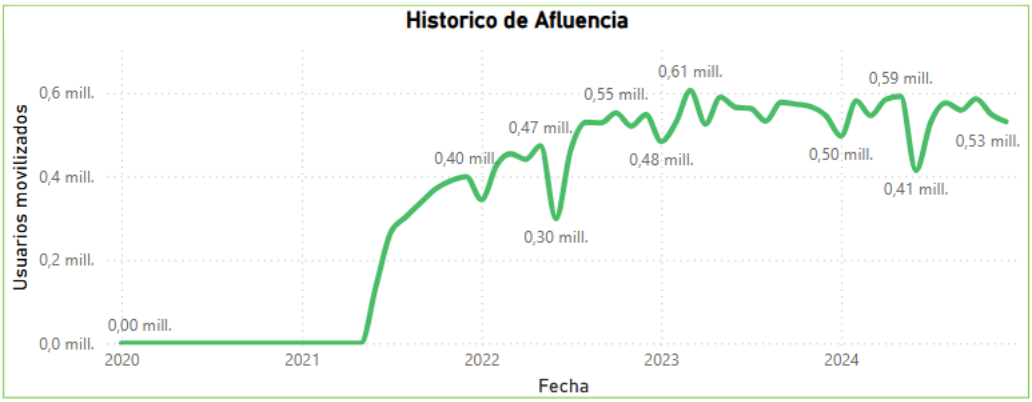

Ilustración 3. Histórico de afluencia para la línea P.

- Con relación a las estaciones, se identificaron unos valores que no corresponden a ningún tipo de estación y otras que están catalogadas como estaciones, pero en realidad son servicios conexos a la Red Metro que no hacen parte del alcance del proyecto. Con lo anterior se realizó un filtro de la variable estación y se procedió a eliminar 73 “Estaciones” las cuales corresponden a otros negocios de recaudo que tiene el Metro de Medellín y que también se ven reflejado en dicha información. Por lo anterior quedaron sólo las siguientes:
- Estaciones = [“ACE", “ACEK”, "ALP", "AYU", "BAN", "BEO", "BER", "CAR", "CAT", "ENV", "HOS", "IND", "ITA", "MAD", "NIQ", "POB", "PRA", "SAA", "STA", "TRI", "UNI", "XPO", "CIS", "EST", "FLO", "JAV", "LUC", "SAM", "ALE", "BIT", "BUA", "JOS", "LOY", "MIR", "ORI", "PAG", "SAT", "TOR", "VIS", "AUR", "JUA", "VAL", "AND", "DOM", "POP", "MIM", "NOV", "PIN", "ACEP", "OCT", "PRO", "SEN", "ARV", "DOL", "ARA", "BEL", "CHG", "CIB", "ESM", "FAT", "GAR", "HOB", "INB", "LIN", "LPE", "MAR", "MIN", "NUT", "PAL", "PLM", "PVEM", "ROS", "UDA", "UDM", "CAD", "COL", "LN2", "PES", "PLA", "SJO", "PT80", “L1” , “L2” , “LB” , “LJ”, “LK”, “LL”, “LM”, “LO”]. Es importante resaltar que cada dato corresponde a las siglas de cada estación de la Red Metro, codificación interna que está estandarizada a lo largo de la operación comercial de la Empresa.

Luego de toda la transformación la base de datos afluencia quedó con la siguiente información:

- Fecha: Año/mes correspondiente a cada registro

- Estación: Sigla correspondiente a cada estación.
- Afluencia: Cantidad de usuarios movilizados por estación.

3.2.1.2. DATOS ATÍPICOS

Adicional a la validación relacionada en el anterior numeral, con base a las líneas que entraron a operar durante el periodo de análisis, se validó la existencia de datos faltantes en líneas y estaciones que vienen operando desde antes del periodo, concluyéndose que no tienen datos faltantes.

También se realizó un análisis con relación al comportamiento de la afluencia en época de la enfermedad Covid-19.

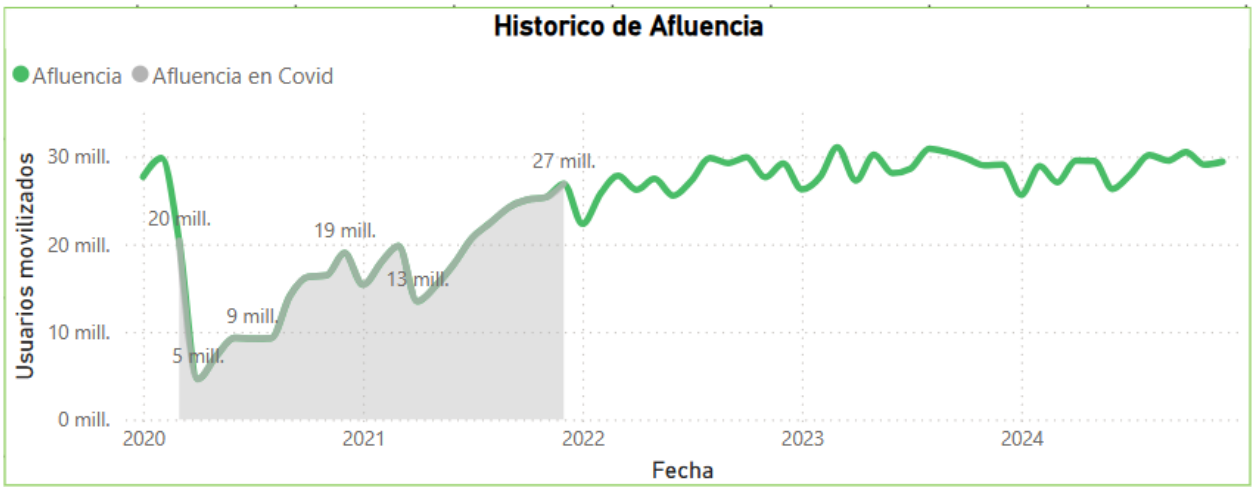

Ilustración 4. Afluencia atípica en época de Covid 19.

3.2.2. BASE DE DATOS DE VENTAS

Como se mencionó en el apartado 2.2.2, Se recibió un libro de Excel el cual contenía toda la facturación entre el año 2020 y 2024, con un total de 13189 facturas a lo largo de dicho periodo de tiempo. Estas facturas corresponden a los servicios de publicidad en la Red Metro.

3.2.2.1. ELIMINACIÓN DE ATÍPICOS

Luego de validar toda la información, se identificó que la base de datos cuenta con unas facturas tipo Nota a Crédito. De acuerdo con lo indagado con los expertos de la Empresa de Publicidad, esto corresponde a facturas que tuvieron que ser anuladas debido a diferentes factores, por lo cual no podrían ser tenidas en cuenta durante el proyecto. Por tal razón, dicha información debía

ser eliminada, hecho que tuvo que ser realizado manualmente por los encargados del proyecto dado que no fue posible recibir la base de datos limpia por parte de los responsables de esta.

El proceso de limpieza de dichas facturas fue dispendioso dado que no solo era identificar la factura (Nota a crédito) que debía ser anulada, sino la factura a la que se le aplicaba dicho proceso, por lo que se debía cotejar una a una cuál era su factura par y eliminar ambas. El total de notas a crédito identificadas correspondía al 4,22% (557 facturas) por lo que en total tuvieron que ser eliminadas 1114 facturas.

3.2.2.2. LIMPIEZA DE LAS VARIABLES

Se observó que los títulos de cada base de datos de cada año estaban escritos de diferente forma y una de las tablas no contenía la misma cantidad de variables que las otras. Por lo tanto, se realizó la estandarización de los títulos de cada tabla y se complementaron las columnas faltantes de la tabla mencionada anteriormente.

Luego de validar toda la información, la proyección sobre los modelos a emplear y el tipo de variable que podría emplearse, el equipo tomó la decisión de emplear las siguientes variables sobre esta base de datos:

- Fecha
- Valor Total
- Elemento
- Estación

3.2.3. BASE DE DATOS AFLUENCIA – VENTAS

Dado que el proyecto requiere de una base de datos estructurada y única, se unificó la base de datos ventas y afluencia, empleando como llave entre ambas bases de datos una variable creada llamada “Ubicación_AñoMes”, mediante la unión de las variables “Estación” y “AñoMes”. Del resultado de dicha compilación se deriva la base de datos Venta-Afluencia la cual contiene las siguientes variables:

- Fecha
- Elemento
- Estación
- Afluencia
- Valor Total

Para facilidades en la generación de visualizaciones, dado las características de la Red Metro, se agregaron las siguientes tablas:

- dimFecha: La cual contiene una lista de todas las fechas entre 2020 y 2024, esta decisión se tomó dada la experiencia del equipo de trabajo en las visualizaciones elaboradas.
- dimEstaciones: tabla que contiene las siguientes variables:
    - Línea: Línea de la red metro a la que pertenece cada estación. Es una variable categórica tipo texto.
    - Modo: Modo de transporte correspondiente a cada línea, los cuales podrían ser Férreo (Tren), Tranvia, Bus y Cable Aéreo.

Con las anteriores tablas, se creó un modelo relacional tipo Estrella, de acuerdo con lo observado en la siguiente ilustración, donde la tabla de hechos es la base de datos VentasAfluencia y las dimensiones serían dimFecha y dimEstaciones.

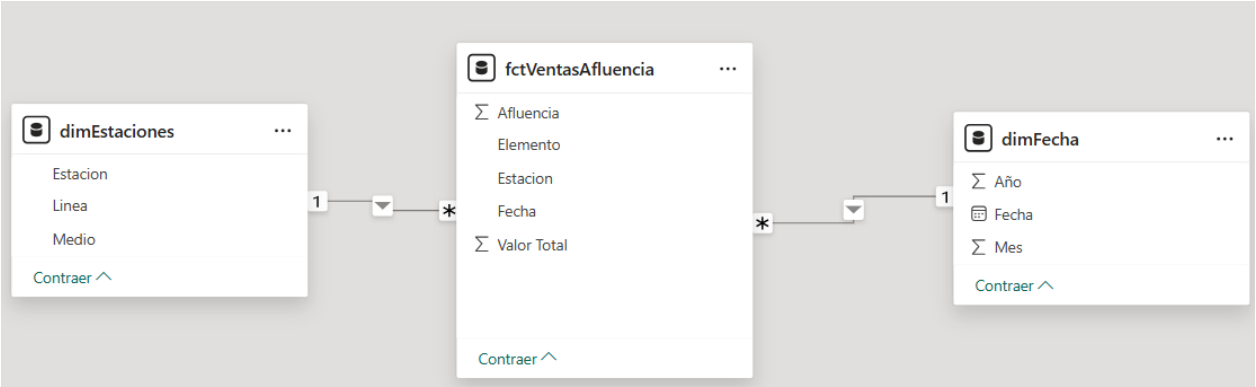

Ilustración 5. Modelo relacional de base de datos VentasAfluencia

3.2.3.1. ESTACIONES SIN REGISTROS DE VENTAS

Se identificó que era necesario incluir registros de no ventas en las diferentes estaciones y servicios publicitarios. Para lo anterior, por medio de una función cíclica, se analizó cada servicio publicitario, en las diferentes estaciones y fechas, si no se registraba ventas, se adicionó una venta con valor de cero y afluencia correspondiente a la fecha y estación de cada registro. El total de registros adicionados fue de 1433.

Luego de avanzar en las diferentes etapas del proyecto, se identificó esta necesidad dado que era preciso ayudar a los diferentes modelos a identificar las fechas, en las que los diferentes servicios publicitarios no registraron venta.

3.3. ANÁLISIS DESCRIPTIVO DE LOS DATOS

3.3.1. AFLUENCIA

3.3.1.1. AFLUENCIA POR MODO

El comportamiento del flujo de usuarios (Afluencia) es fundamental, dado que dicha variable indica el comportamiento de los usuarios a lo largo de la Red Metro. En la siguiente ilustración se puede identificar el Modo más empleado por los usuarios a lo largo de los años de análisis.

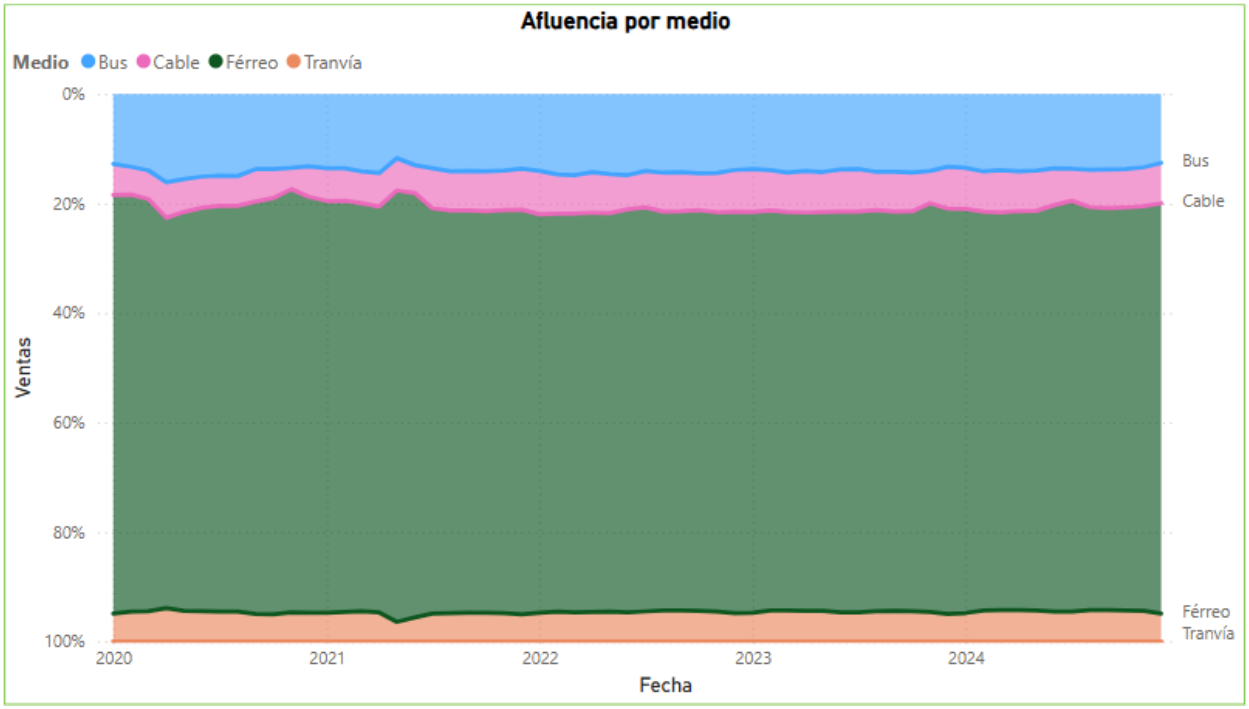

Ilustración 6. Histórico de afluencia de usuarios por modo de transporte.

Se puede observar que el modo más empleado es el Férreo, con una media de 73,81% de los usuarios movilizados, le sigue Bus con 13.93%, posteriormente Cables Aéreos con el 6,86% y tranvía con 5.4% del total.

3.3.1.2. AFLUENCIA POR ESTACIÓN

Dicha relevancia para el transporte vía Férreo puede identificarse en la siguiente gráfica, donde se observa que en las primeras 19 estaciones con mayor afluencia, son

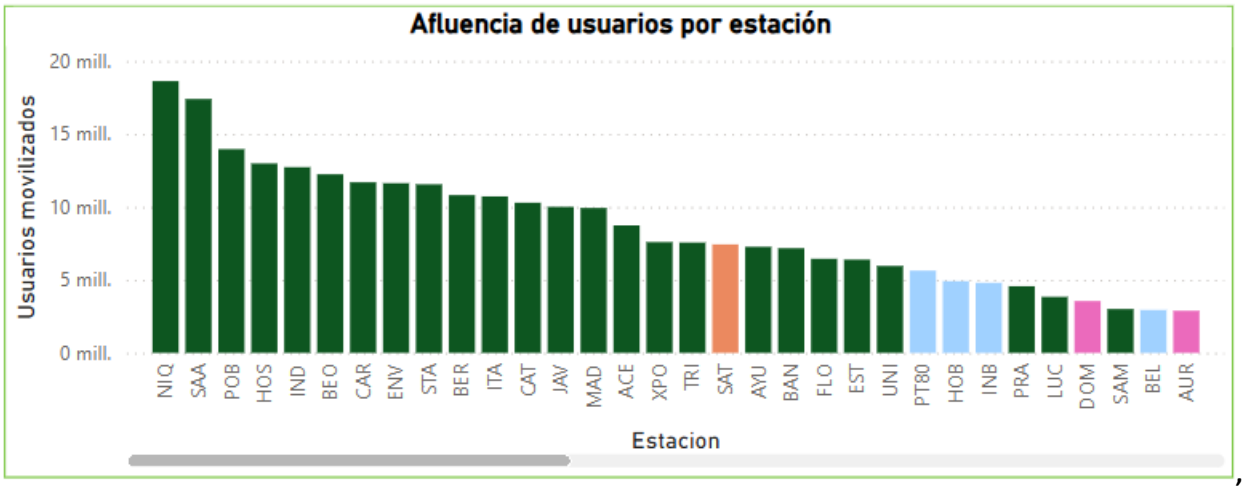

Ilustración 7. Usuarios movilizados por estación en 2024, para estaciones del modo Férreo (verde) Tranvía (Naranja), Buses (Azul) y Cables Aéreos (Fucsia)

La anterior ilustración contrasta con lo señalado en la ilustración 6, con relación a la dominación que tiene el modo Férreo (verde) para el transporte de personas, dado que las primeras 17 estaciones con mayor cantidad de personas movilizadas son de dicho modo de transporte. Posteriormente, comienza a identificarse la principal estación de Tranvía (Naranja), algunas de buses (azul) y de Cables Aéreos (Fucsia).

3.3.2. COMPORTAMIENTO DE LAS VENTAS

Inicialmente, se procedió a realizar un análisis de Pareto para lograr identificar qué tipo de servicios publicitarios son los que más afectan las ventas del Metro de Medellín, tal como lo señala la siguiente visualización:

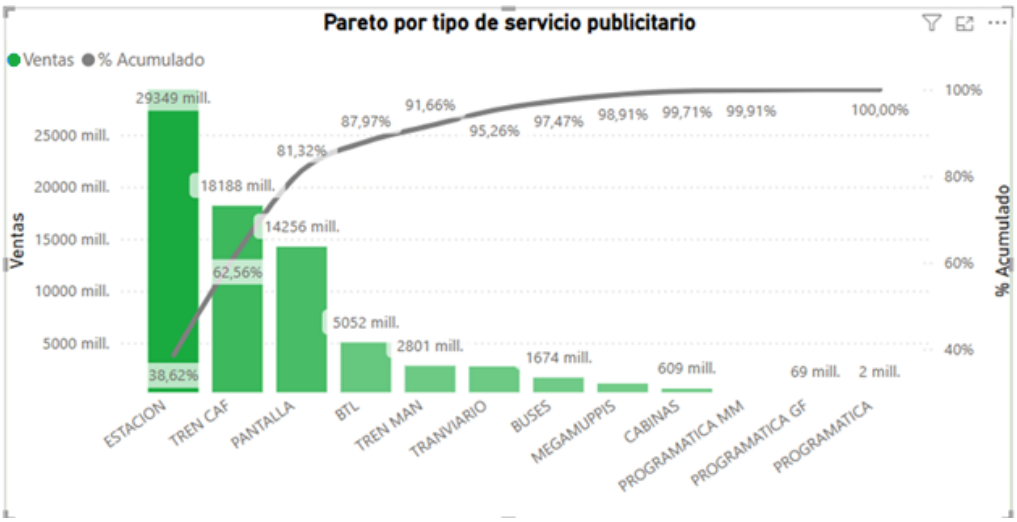

Ilustración 8. Análisis de Pareto para las ventas de los tipos de servicios publicitario en el metro de Medellín

De la anterior grafica se puede observar que los servicios publicitarios más significativos en ventas para el metro de Medellín, en los últimos 4 años fueron las publicidades de las estaciones, la publicidad de la línea de trenes CAF y las pantallas que se ubican en diferentes estaciones, los cuales representan un 81,32% de las ventas totales.

Por consiguiente, el equipo decidió enfocar los esfuerzos en realizar modelos predictivos para el tipo de servicios publicitarios en las estaciones, con la finalidad de mejorar la calidad de las proyecciones para tomar decisiones informadas con más certeza que permitan el incremento de utilidades del metro.

3.3.2.1. VENTAS POR ESTACIÓN

En la siguiente visualización se puede identificar las estaciones que más han realizado ventas en el periodo de análisis. En ésta se puede observar que sigue predominando las estaciones de trenes, fenómeno que también se repite en la variable afluencia, tal como ya se expuso.

Es importante resaltar que se presenta un cambio en la forma de medir esta variable para las estaciones de Tranvía, buses y cables, ya que la facturación es relacionada para líneas completas de cada modo. También es relevante comprender que las líneas con la unión de varias estaciones, en las cuales las unidades se desplazan entre estas haciendo recorridos cíclicos.

Las estaciones que más resaltan dentro de esta medición, a diferencia de Férreo son “Tranv” y L1, la cual corresponden a la única línea de Tranvía y la principal línea de Buses respectivamente.

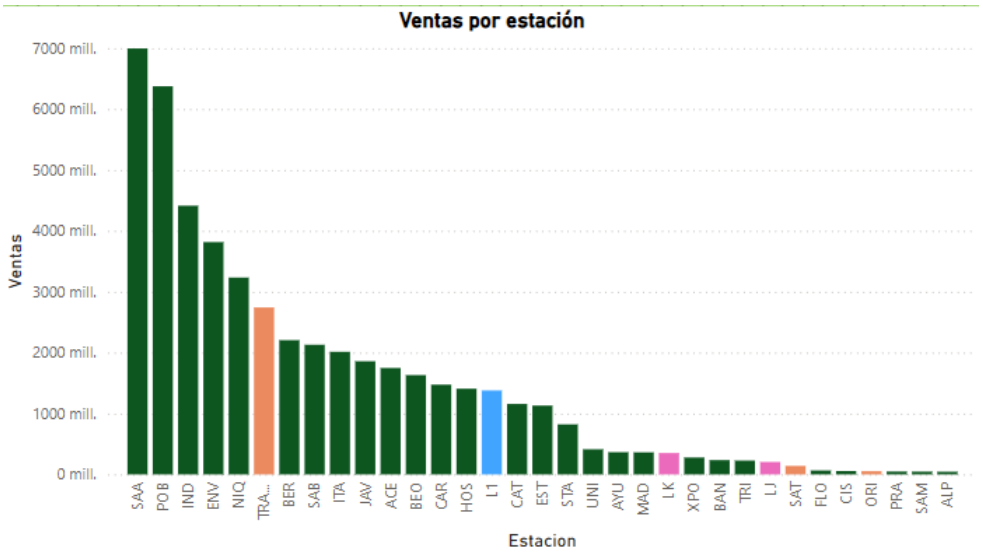

Ilustración 9. Ventas realizadas por estación en 2024, para estaciones del modo Férreo (verde) Tranvía (Naranja), Buses (Azul) y Cables Aéreos (Fucsia)

3.3.2.2. VENTAS POR MODO

El comportamiento de las ventas para cada modo de transporte se puede observar en la siguiente ilustración. Se puede identificar que tiene una diferencia importante con la ilustración

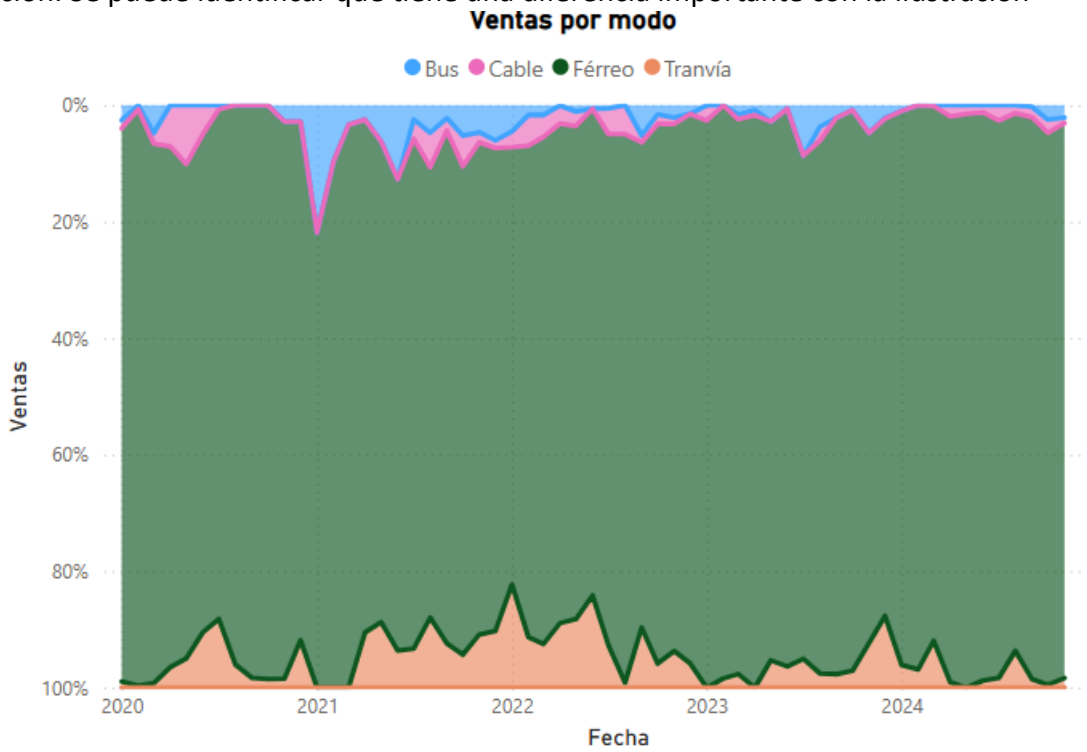

Ilustración 10. Comportamiento de las ventas por modo de transporte

3.3.2.3. SERVICIOS PUBLICITARIOS ELEGIDOS

Se analizó el comportamiento de las ventas para todos los servicios publicitarios, por medio de la siguiente ilustración.

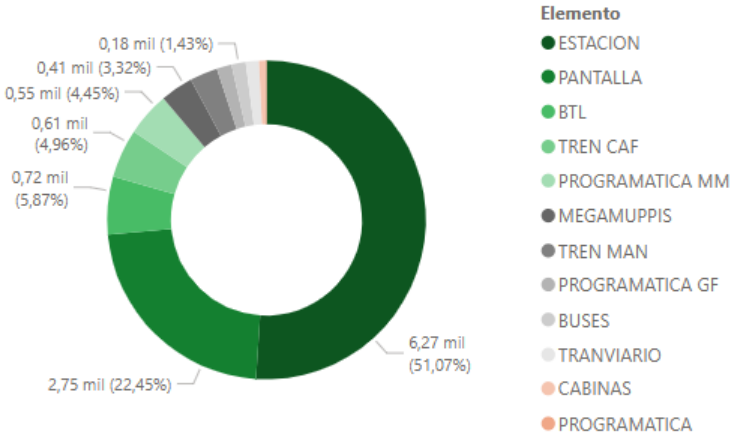

Ilustración 11. Distribución de ventas para todos los tipos de servicios publicitarios

Avanzando en las diferentes etapas del proyecto, incluyendo el modelado, el equipo tomó la decisión de concentrar el proyecto en los servicios publicitarios tipo “Estación” con base en los siguientes hechos:

- Estos servicios son el 51.7% del total de las ventas a lo largo de todo el periodo de análisis.
- Cuentan con la mayor cantidad de registros de ventas, en comparación a otros servicios, lo cual favorece la cantidad de datos para el modelado y la calidad de este.
- El análisis de relación entre las variables afluencia y ventas indica que para la categoría Estación, en las diferentes estaciones, dichas variables tienen mayor colinealidad y cointegración, respecto al resto de servicios publicitarios.

3.3.3. VENTAS-AFLUENCIA

En las diferentes conversaciones de contexto que se realizaron con los expertos de la Empresa, se ha identificado por medio de la experiencia y conocimiento adquirido a lo largo de los años que las variables afluencia y ventas guardan una estrecha relación. Tal es el caso de las estaciones SAA y POB las cuales presentan los mayores niveles de ventas y a su vez hacen parte de las estaciones con mayor afluencia (Ver Ilustración 7 y 9)

3.3.3.1. DISPERSIÓN POR SERVICIOS PUBLICITARIOS

La siguiente ilustración permite observar cómo se comporta la afluencia versus las ventas. Se puede identificar una agrupación importante en algunos registros los cuales se encuentran cerca al límite de cero ventas o pocas ventas a pesar de que cuentan con valores de afluencia, mientras que otros se encuentran dispersos sobre las ventas en la medida que aumenta la afluencia.

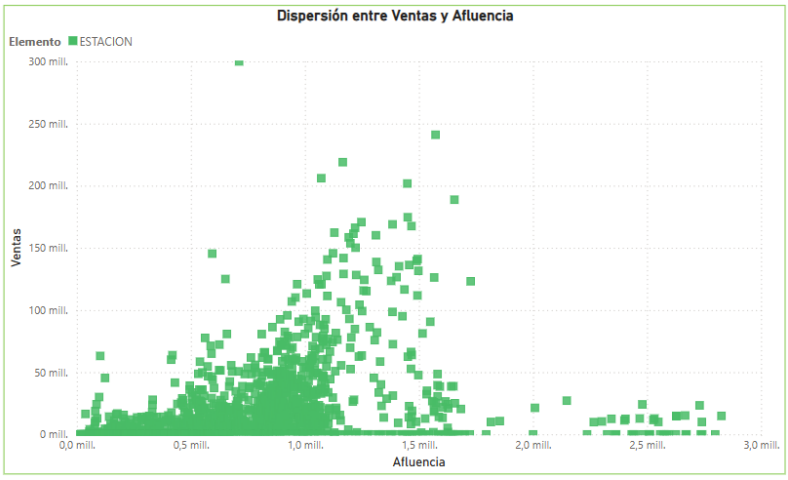

Ilustración 12. Grafico de dispersión entre las variables Afluencia y Ventas para los servicios publicitarios para cada servicio publicitario.

3.3.4. PRUEBA DE COLINEALIDAD Y COINTEGRACIÓN

Se inició el análisis realizando una prueba de colinealidad con el objetivo de evaluar estadísticamente si existe alguna relación entre las variables Afluencia y Ventas. Para llevar a cabo el análisis de esta relación se empleó las siguientes métricas:

- Factor de Inflación de la Varianza (VIF), para identificar si las variables son colineales o lo que significa, que tienen una correlación en el tiempo.
- Cointegración, la cual permite identificar si las variables de series temporales tienen una relación de equilibrio a largo plazo.

Al realizar las pruebas de las variables para cada estación, se identificó que existen 3 grupos principales, los cuales se dividieron así:

Tabla 55. Métricas y rangos empleados para identificar la fuerza de la relación entre las variables

Relación	VIF	Cointegración	Significado
Alta	2.5 a 10	Si	Buena relación entre las variables y registros a lo largo de todo el periodo de análisis
Media	1.6 a 2.5	Si	Relación con una intensidad media entre las variables, pero cointegradas.
Baja	1 a 1.2	SI/NO	Relaciones con una baja, nula intensidad, o pocos datos en ventas para hacer el análisis (Menos de 10 datos)

Con base en los criterios definidos previamente para identificar variables con alta colinealidad y, posteriormente, aplicando la prueba de cointegración, se identificaron diez estaciones que cumplían con ambos requisitos: BEO, CAR, CAT, IND, ITA, JAV, NIQ, POB, SAA y XPO. A partir de esta selección, se procedió a escoger las estaciones POB (El Poblado) y SAA (San Antonio A) para el desarrollo de los modelos predictivos, dado que son las que presentan el mayor número de observaciones y volumen de ventas.

Los análisis realizados muestran que en ambas estaciones existe una alta colinealidad entre ventas y afluencia, lo cual sugiere que la demanda de servicios publicitarios está estrechamente relacionada con el flujo de personas. Además, el hecho de que estas variables estén cointegradas respalda la existencia de una relación de equilibrio a largo plazo, lo cual fortalece su inclusión como base en los modelos de predicción.

4. MODELADO

4.1. PERPARACIÓN DE DATASET

En un proyecto de ciencia de datos es importante siempre definir cómo se van a particionar los datos para el entrenamiento y prueba de los modelos, y en los modelos de series de tiempo no es la excepción. En este apartado, se describe brevemente como fue la selección de la partición de los datos de acuerdo con la disponibilidad, cantidad y los tipos de modelos a utilizar.

A partir de las pruebas de colinealidad y cointegración mencionadas en las secciones anteriores, se definieron 3 grupos los cuales clasificaron las diferentes series de tiempo de acuerdo con los resultados de las pruebas (Ver tabla 5).

Se utilizó esta metodología principalmente por la naturaleza de los datos (ventas y afluencia), que al ser series de tiempo, estas deben agruparse de tal forma que ambas series correspondieran a una misma estación y servicio publicitario en la misma ventana de tiempo, es decir, en el grupo de alta colinealidad y cointegración sólo deben existir series de tiempo completas para una estación y para un tipo de servicio publicitario en particular y no una parte de la serie, razón por la cual un modelo de clustering no era suficiente para la selección de los datos.

Anteriormente, se mencionó que se tomó un muestreo de 2 estaciones por cada grupo con la finalidad de ser utilizadas en el modelado, las estaciones seleccionadas fueron aquellas que más datos disponibles tenían, teniendo un total de 6 estaciones con 59 datos cada una. A continuación, se mencionan las estaciones de cada grupo:

Tabla 66. Descripción de las estaciones utilizadas en el modelado

Servicio publicitario	**Estación**	
Grupo	**Ubicación**	**Datos**
Relación alta	Poblado (POB) San Antonio A (SAA)	Periodicidad: Mensual **Cantidad:** Desde 01-enero de 2020 hasta 30 noviembre de 2024 (59 meses)
Relación media	Acevedo (ACE) Hospital (HOS)	
Relación baja	Linea H (LH) Linea J (LJ)	

Otro factor importante a la hora de seleccionar la partición fueron los modelos a utilizar, se seleccionaron los modelos ARIMA, Holt-Winters, Prophet y Red Neuronal Elman, debido a su capacidad para capturar distintos patrones en series temporales mensuales.

ARIMA modela relaciones lineales y tendencias simples; Holt-Winters aborda estacionalidades explícitas; Prophet permite incorporar cambios de tendencia y gestionar datos irregulares; y Elman captura dinámicas no lineales y memorias temporales complejas. Esta combinación asegura un análisis robusto y flexible, adaptado a la naturaleza diversa de los datos evaluados.

Algo que tienen en común los modelos a evaluar es la cantidad de datos necesarios para la realización de un buen pronóstico, lo cual para este proyecto es una de las principales restricciones al solo disponer 59 periodos de tiempo para todos los modelos

Por lo tanto, todos los modelos están sujetos a ser de memoria larga, lo que significa que entre más datos se utilicen para entrenar las predicciones serán más confiables a largo plazo. Por lo tanto, inicialmente se plantearon diferentes formas de partir los datos en entrenamiento y prueba las cuales se muestran a continuación:

Tabla 77. Particiones definidas

Opción	Entrenamiento	Prueba
1	48 meses	11 meses
2	53 meses	6 meses
3	55 meses	4 meses
4	56 meses	3 meses

Se seleccionó la opción 2, debido a que proporciona un equilibrio adecuado entre un conjunto de entrenamiento suficientemente amplio para capturar de manera robusta los patrones de tendencia, estacionalidad y dinámica subyacente de la serie temporal, y un conjunto de prueba de tamaño razonable (6 meses) que permite evaluar de forma confiable la capacidad predictiva del modelo. Un entrenamiento con 53 meses asegura que los modelos, especialmente aquellos que requieren capturar comportamientos a largo plazo como Prophet y Elman, dispongan de suficiente información histórica para ajustar sus parámetros de forma precisa, mientras que una prueba de 6 meses minimiza el riesgo de sobreajuste (overfitting) y proporciona una ventana temporal representativa para validar la estabilidad y generalización del modelo en el corto y mediano plazo.

4.2. DESCOMPOSICIÓN TEMPORAL

La descomposición de series temporales es una técnica fundamental en el análisis y pronóstico de datos secuenciales. Permite descomponer una serie en componentes como tendencia, estacionalidad y residuo, facilitando una comprensión más profunda de los patrones subyacentes y mejorando la precisión de las predicciones.

Según Hyndman y Athanasopoulos en su obra Forecasting: Principles and Practice [18], la descomposición ayuda a mejorar la comprensión de la serie temporal y puede utilizarse para mejorar la precisión del pronóstico. Al separar los componentes sistemáticos (tendencia y estacionalidad) del ruido aleatorio, se facilita la selección y ajuste de modelos predictivos adecuados, como ARIMA, Holt-Winters o Prophet, y se mejora la interpretación de los resultados.

Además, la descomposición permite detectar anomalías y evaluar la estabilidad de los patrones estacionales a lo largo del tiempo, lo cual es esencial para una modelación efectiva y para la toma de decisiones informadas en diversos contextos, como el análisis de ventas, tráfico o comportamiento del consumidor.

Por lo tanto, se procedió a realizar la descomposición para cada una de las series mencionadas en la tabla 5.

4.2.1. POBLADO (POB)

La descomposición de la serie temporal correspondiente a la estación POB permitió analizar de manera detallada la estructura interna de los datos. El componente de tendencia evidenció un crecimiento sostenido desde el año 2020 hasta mediados del 2023, indicando una dinámica positiva en los valores observados durante la mayor parte del período analizado. No obstante, hacia finales del horizonte temporal se observa una ligera estabilización e incluso un incipiente descenso, lo que podría asociarse a cambios en el comportamiento del mercado o factores externos que afectaron el desempeño.

Por otra parte, el componente estacional reveló patrones anuales recurrentes, caracterizados por picos y valles que se repiten de manera consistente, reflejando la existencia de ciclos de alta y baja en la serie, posiblemente ligados a fenómenos propios del calendario como temporadas de alta demanda.

Finalmente, el análisis de los residuos mostró que las variaciones no explicadas son moderadas y en su mayoría aleatorias, aunque con algunos eventos atípicos que generan desviaciones específicas, sugiriendo que la mayor parte de la variabilidad de la serie puede ser explicada adecuadamente por la tendencia y la estacionalidad identificadas.

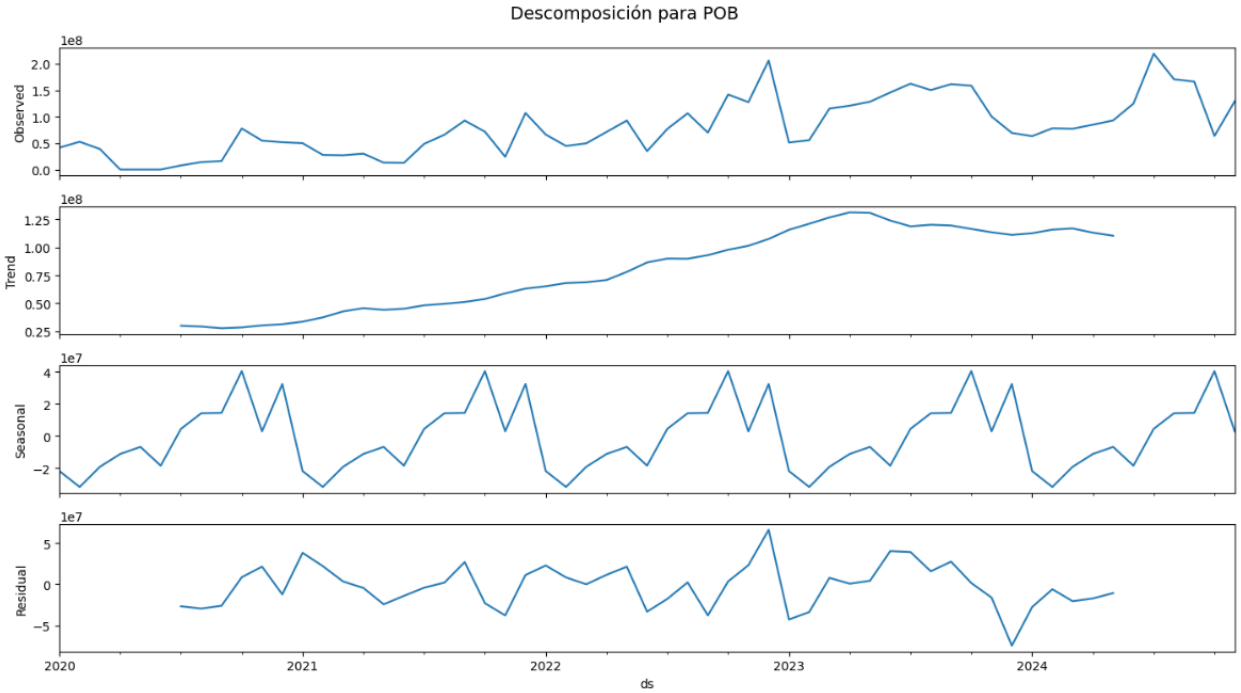

Ilustración 13. Descomposición serie de tiempo estación poblado (POB)

4.2.2. SAN ANTONIO A (SAA)

La descomposición de la serie temporal correspondiente a la estación SAA permitió evidenciar una dinámica caracterizada por alta variabilidad y crecimiento generalizado en el período de análisis. El componente de tendencia mostró un comportamiento ascendente sostenido desde el año 2020 hasta mediados de 2023, seguido de una etapa de ligera estabilización. Esta evolución sugiere un entorno inicialmente favorable con un posible cambio en las condiciones del mercado o en los patrones de comportamiento hacia el final del período.

Por otro lado, el componente estacional reflejó la existencia de patrones recurrentes anuales con una amplitud considerable, indicando que los ciclos de alta y baja demanda son intensos y forman parte inherente del comportamiento de la serie.

Finalmente, el análisis de los residuos evidenció una mayor dispersión en comparación con otras estaciones, lo que sugiere la ocurrencia de eventos extraordinarios que afectan de manera significativa los valores observados, confirmando la necesidad de considerar factores externos o episodios particulares al momento de interpretar la serie.

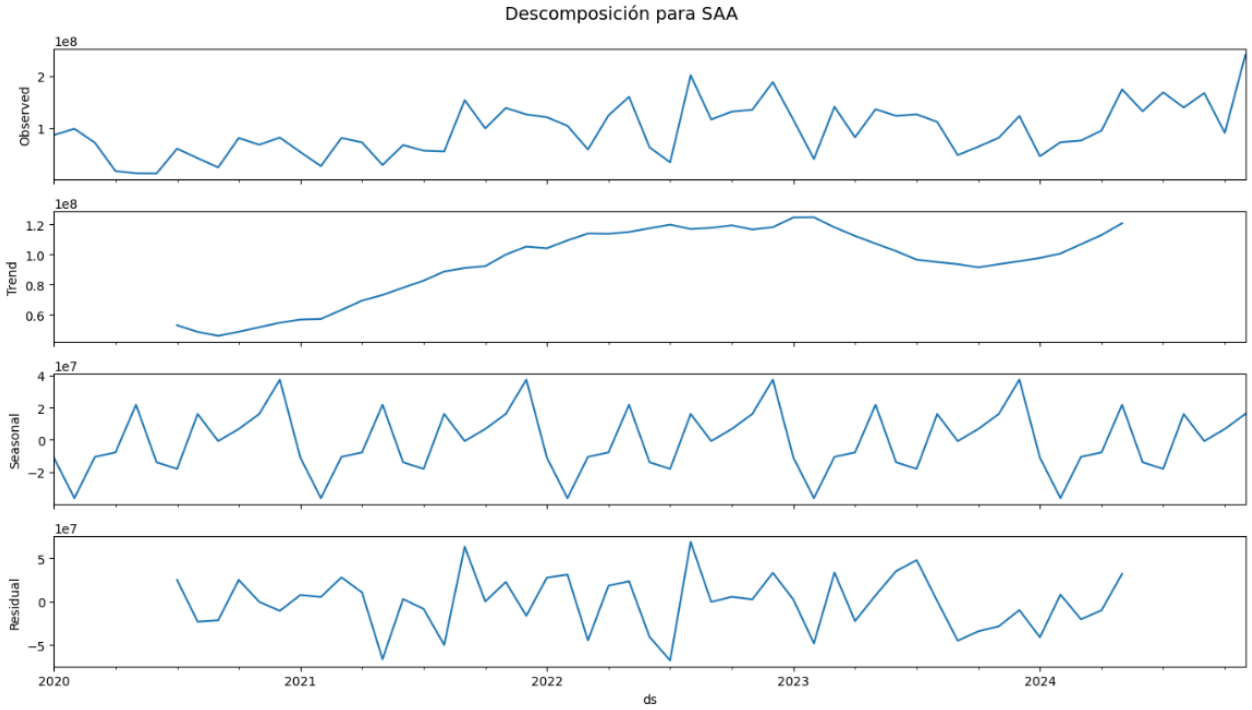

Ilustración 14. Descomposición serie de tiempo estación San Antonio A (SAA)

4.2.3. ACEVEDO (ACE)

La descomposición de la serie temporal correspondiente a esta estación permitió identificar un comportamiento caracterizado inicialmente por un crecimiento sostenido, que se mantuvo hasta aproximadamente mediados del año 2023.

Posteriormente, se evidenció una disminución progresiva en la tendencia, lo cual podría estar asociado a cambios en las condiciones de mercado, a alteraciones en la demanda o a factores estructurales no contemplados inicialmente. El componente estacional reveló patrones cíclicos claros, donde se observan aumentos y disminuciones periódicas en los valores, sugiriendo la existencia de eventos recurrentes que influyen en la dinámica de la serie.

Sin embargo, el análisis de los residuos mostró una alta dispersión, indicando que, además de la tendencia y la estacionalidad, existen factores aleatorios o eventos extraordinarios que afectan significativamente el comportamiento de los datos, lo que resalta la necesidad de considerar fuentes adicionales de variabilidad en el análisis predictivo.

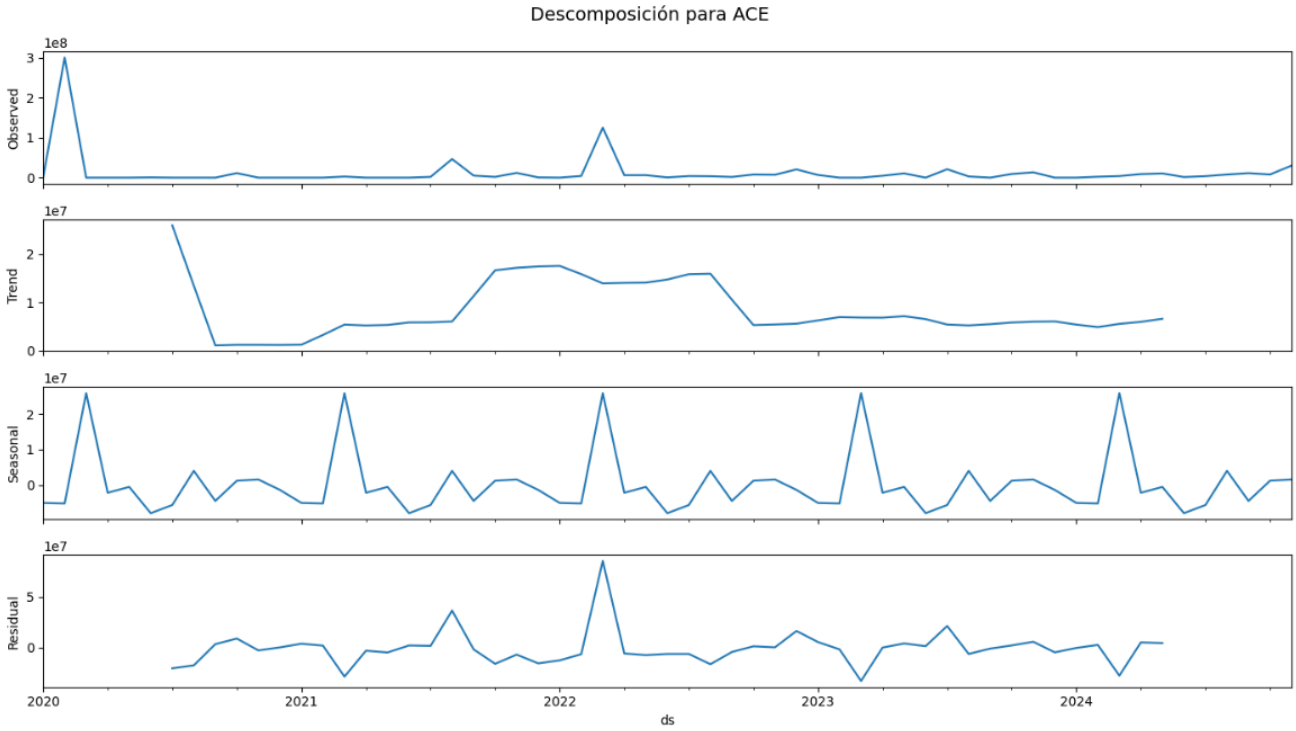

Ilustración 15. Descomposición serie de tiempo estación Acevedp (ACE)

4.2.4. HOSPITAL (HOS)

La descomposición de la serie temporal para la estación HOS revela un comportamiento caracterizado por baja actividad general, interrumpida ocasionalmente por picos de gran magnitud.

El análisis del componente de tendencia mostró un crecimiento progresivo desde el año 2020 hasta principios de 2022, seguido de un descenso sostenido durante los períodos posteriores, lo cual podría reflejar una disminución en la demanda o cambios estructurales en el comportamiento de los usuarios o consumidores asociados a esta estación. En cuanto a la estacionalidad, se identificaron patrones repetitivos anuales de alta intensidad, con oscilaciones marcadas entre meses de alta y baja actividad.

Finalmente, los residuos presentan una variabilidad considerable, evidenciando que factores aleatorios o eventos extraordinarios tienen un impacto significativo en la serie, lo que implica la necesidad de considerar elementos adicionales en futuros análisis o modelos predictivos.

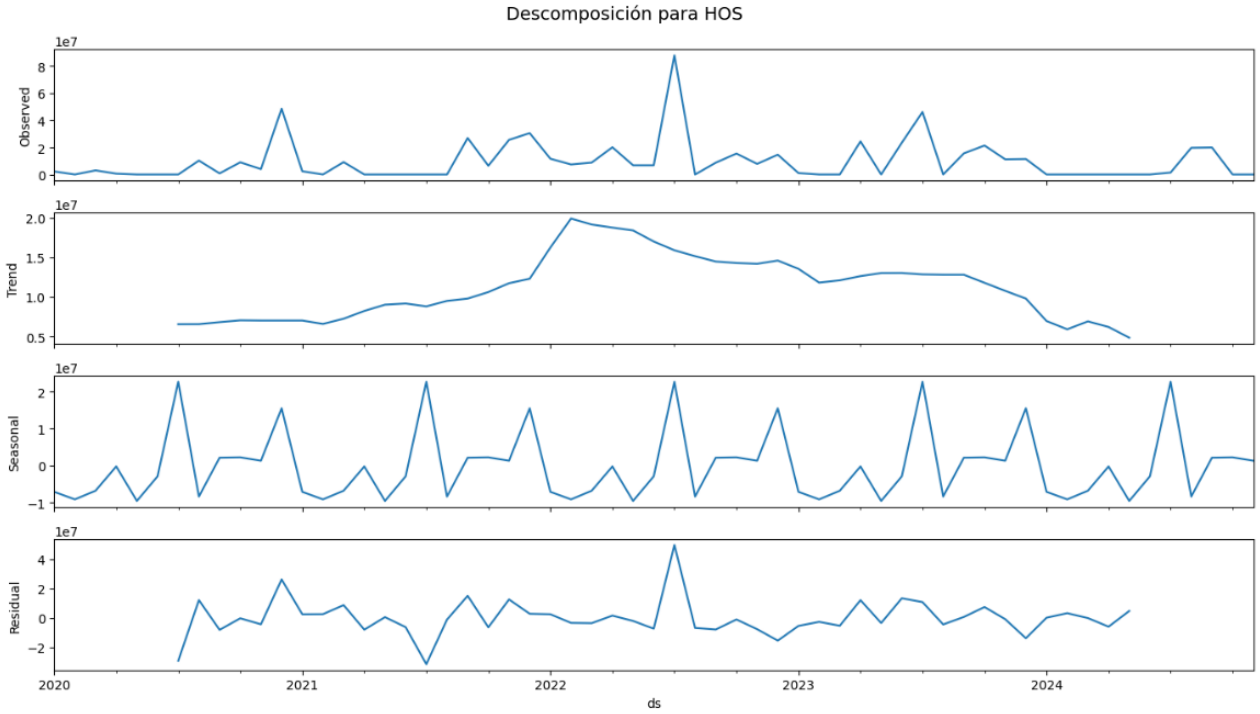

Ilustración 16. Descomposición serie de tiempo estación Hospital (HOS)

4.2.5. LÍNEA H (LH)

La descomposición de la serie temporal para la estación LH evidencia un comportamiento particular, caracterizado por una inactividad casi total desde el año 2020 hasta mediados de 2023, periodo durante el cual los valores observados permanecieron cercanos a cero.

A partir de mediados de 2023, se detecta un cambio sustancial, reflejado en un aumento progresivo de la tendencia y en la aparición de valores significativamente mayores, lo cual sugiere el inicio reciente de operaciones o de un nuevo proceso de generación de datos en la estación. La componente estacional, aunque presenta patrones recurrentes a lo largo de los años, muestra una mayor relevancia a partir del surgimiento de la actividad.

Finalmente, los residuos son reducidos durante los años de inactividad, pero presentan mayor dispersión en el período de actividad reciente, indicando la aparición de factores adicionales que contribuyen a la variabilidad de la serie en esta nueva fase.

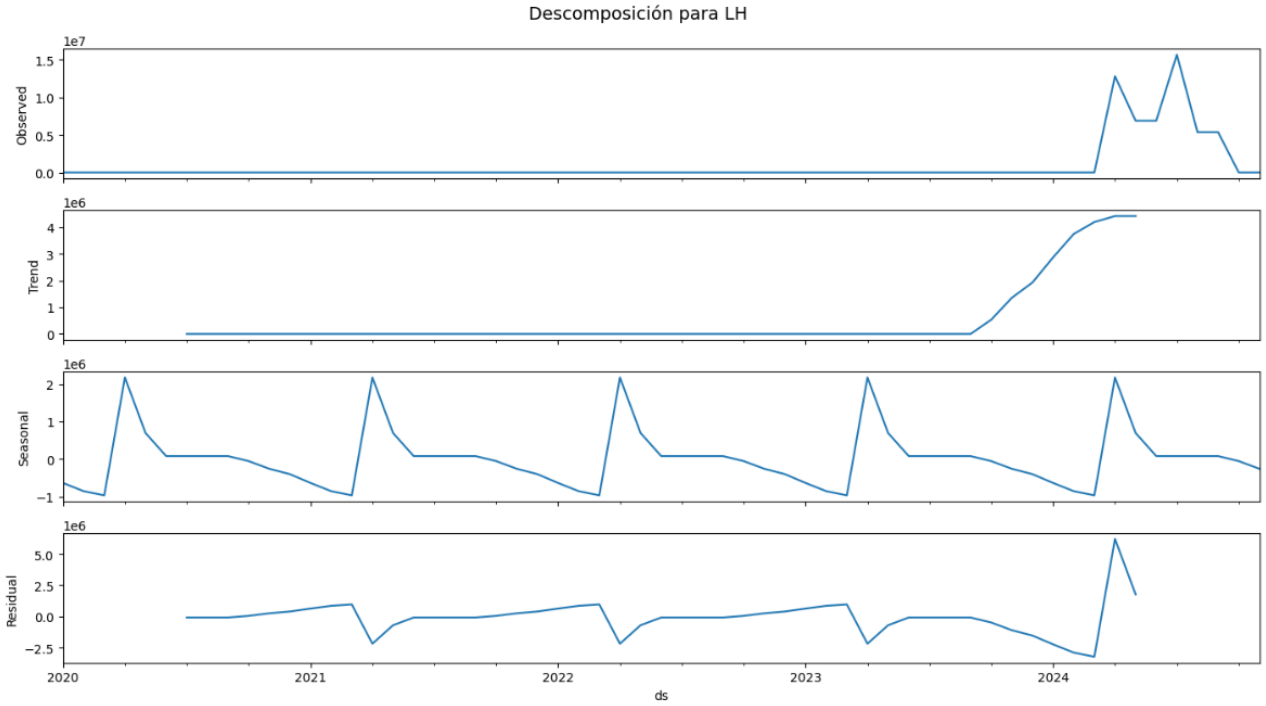

Ilustración 17. Descomposición serie de tiempo estación Línea H (LH)

4.2.6. LÍNEA J (LJ)

La descomposición de la serie temporal correspondiente a la estación LJ evidencia un comportamiento de inactividad prolongada, con valores prácticamente nulos entre los años 2020 y 2023.

No obstante, a finales de 2024, se observa un cambio abrupto caracterizado por un incremento significativo en los valores observados, acompañado de un ascenso marcado en la tendencia, lo cual sugiere el inicio reciente de un proceso de actividad relevante en esta estación. La componente estacional, aunque se mantiene presente a lo largo del periodo analizado, adquiere mayor representatividad a partir del surgimiento de los valores positivos.

Por su parte, los residuos mostraron baja dispersión durante la fase de inactividad, pero experimentaron un aumento notable tras la aparición de los eventos extraordinarios, indicando la existencia de factores adicionales que afectaron la dinámica reciente de la serie.

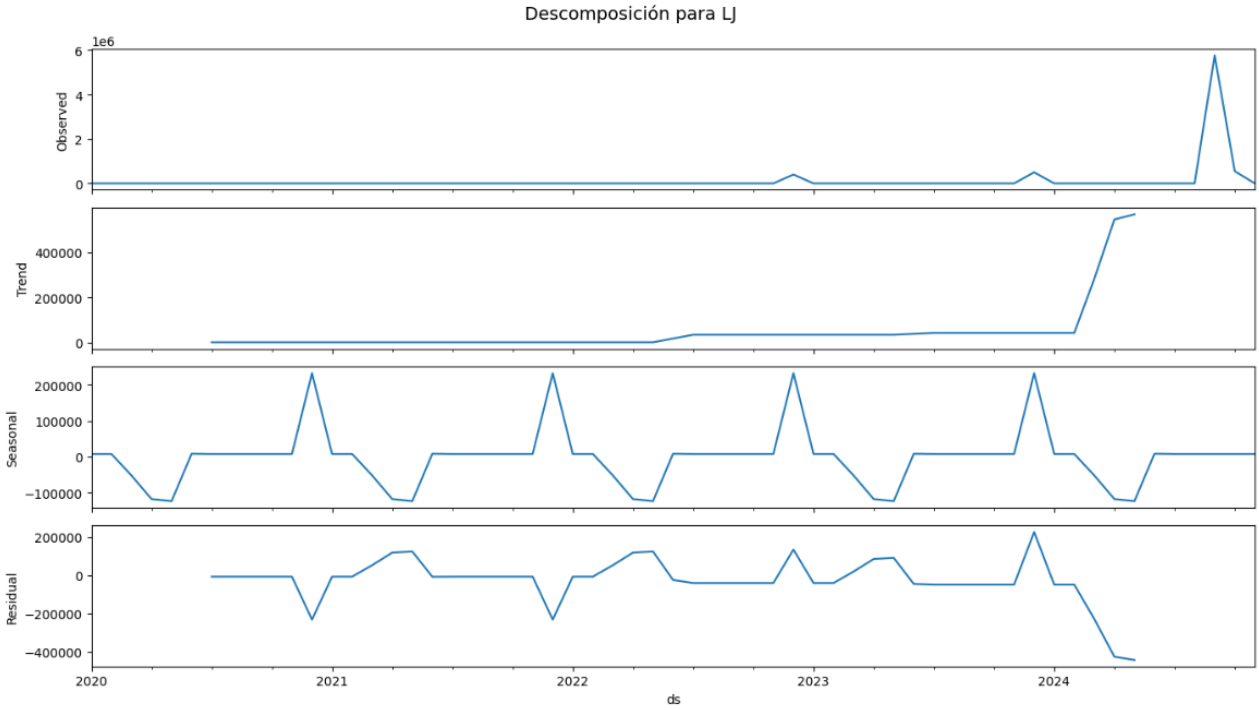

Ilustración 18. Descomposición serie de tiempo estación Línea J (LJ)

4.3. MODELO FACEBOOK PROPHET

Para la implementación de este modelo se utilizó Google Coolab como notebook para la escritura del código en Python, se utilizaron librerías como Numpy, Pandas, statsmodels, Matplotlib, Prophet y Sklearn.

4.3.1. HIPERPARAMETROS EMPLEADOS

En un principio, se definieron los diferentes hiperparámetros del modelo que se utilizaron para optimizar los resultados de las predicciones mediante una búsqueda por grilla, para diferentes combinaciones de entrenamiento y prueba. A continuación, se describe cada uno de los hiperparámetros utilizados, la búsqueda de grilla y su función real dentro del modelado.

Tabla 88. Hiperparametros utilizados en el modelo

Hiperparámetros	Función
Change point prior scale (**CPS**)	Controla la flexibilidad del modelo para ajustarse a cambios en la tendencia. Valores más altos permiten que el modelo detecte cambios más abruptos, pero también incrementan el riesgo de sobreajuste. Valores típicos van de 0.001 a 0.5.

N_changepoints (**NC**)	Indica el número de puntos de cambio potenciales en la tendencia durante el periodo de entrenamiento. El modelo elegirá estos puntos automáticamente de forma equidistante si no se especifican manualmente. A mayor número, más capacidad para adaptarse a cambios estructurales. El rango típico puede ir de 0 a 50 o incluso más dependiendo del tamaño del conjunto de datos
Seasonality_mode (**SM**)	Define cómo se modela la estacionalidad "additive": la estacionalidad se suma a la tendencia. "multiplicative": la estacionalidad se multiplica por la tendencia, útil cuando la magnitud de los efectos estacionales cambia con el nivel de la serie.

Posteriormente, se definió la grilla de búsqueda, la cual se muestra a continuación

Tabla 99. Grilla definida por experimentos

CPS	NC	SM	Combinaciones evaluadas en la grilla
[0,01;0,1;0,5]	[10,25;50]	[additive,multiplicative]	18
np.linspace(0.01,0.5,10)	np.arange(0,56,7)	[additive,multiplicative]	160

4.3.2. DISEÑO DE PRUEBAS

De la anterior tabla, es importante mencionar que para la segunda grilla se utilizaron las funciones linspace y arange de la librería numpy. Para el caso de CPS np.linspace(0.01,0.5,10) devuelve un vector de 10 valores equidistantes entre 0,01 y 0,5, por otro lado, para el NC se utilizó np.arange (0,56,7) el cual devuelve un vector entre 0 y 56 con incremento de 7 unidades, que para este caso el vector tendría 8 datos.

Luego de la definición de la grilla se estableció el modelo con la variable afluencia como un regresor y se realizó la respectiva ejecución para cada partición previamente definida (Ver tabla 6). A continuación, se muestra un resumen de los experimentos ejecutados

Tabla 1010. Combinaciones por experimento

Particiones	Estaciones	Grilla	Modelos ejecutados por partición	Modelos totales
4	6	18	108	432
		160	960	3840

Finalmente, se obtuvieron resultados para cada experimento, para cada estación de la partición seleccionada.

Tabla 1111. Metricas modelo FB prophet

Escenario	Estación	CPS	SM	CP	SMAPE	MAE	RMSE
1	POB	0,1	multiplicative	10	31,96	43.755.453	58.229.607
	SAA	0,01	multiplicative	25	49,47	66.330.959	76.173.419
	ACE	0,5	multiplicative	25	103,65	8.524.938	12.691.529
	HOS	0,5	additive	25	147,6	6.037.356	7.104.404
	LH	0,5	additive	10	94,37	4.395.837	5.819.320
	LJ	0,1	additive	10	190,55	1.056.798	2.335.751
2	POB	0,01	multiplicative	0	31,99	43.790.840	58.218.310
	SAA	0,01	multiplicative	28	49,57	66.431.130	76.297.080
	ACE	0,39	multiplicative	28	62,81	7.711.403	10.899.090
	HOS	0,12	additive	49	134,19	6.801.034	8.293.559
	LH	0,50	additive	7	94,36	4.395.733	5.819.152
	LJ	0,28	additive	35	189,23	1.060.835	2.330.371

4.4. MODELO ARIMA

Para este modelo se empleó como variable exógena la afluencia, dados los resultados obtenidos en el análisis de colinealidad y cointegración entre las variables de ventas y afluencia.

Para la implementación de este modelo se utilizó Google Colab como entorno de notebook en Python, empleando librerías fundamentales:

- NumPy y Pandas para carga y preprocesamiento de datos.
- Statsmodels para estimar ARIMA/SARIMA.
- Scikit-learn para la búsqueda por grilla (GridSearchCV).
- Matplotlib para la visualización de resultados y diagnóstico de residuos.

4.4.1. HIPERPARAMETROS EMPLEADOS

Luego de varias pruebas, el conocimiento sobre los datos, la cantidad de datos y lo recomendado por Box & Jenkins [10] se ha optado por la configuración básica de SARIMAX, centrada en el order y seasonal_order.

Es por esto por lo que los hiperparámetros empleados fueron los siguientes:

Tabla 1212. Hiperparámetros utilizados en el modelo Arima/Sarimax

Hiperparámetros	Función
order	En el proyecto, se exploraron valores de p, d, q, en rangos preestablecidos para ajustar la dinámica no estacional.

Hiperparámetros	Función
Seasonal_order	Se usa para capturar ciclos periódicos en la serie. En el proyecto, se exploraron valores de P, D, Q, en rangos preestablecidos. Para el proyecto el s = 12, dado que los datos son estacionales a 12 meses.
enforce_stationarity	Transforma los parámetros autorregresivos (AR) para que el polinomio AR tenga todas sus raíces dentro del círculo unitario, garantizando que el proceso sea estacionario (es decir, que medios y varianzas no diverjan con el tiempo). Para los modelos analizados este hiperparametro es = False
enforce_invertibility	Transforma los parámetros de media móvil (MA) para que el polinomio MA tenga sus raíces fuera del círculo unitario, asegurando que el proceso sea invertible (es decir, que los choques pasados puedan expresarse adecuadamente como combinaciones finitas de errores). Para los modelos analizados este hiperparametro es = False.

4.4.2. DISEÑO DE PRUEBAS

El equipo analizó las opciones disponibles para el diseño de las diferentes pruebas a implementar. Se concluyó que las más optimas a trabajar son las siguientes:

4.4.2.1. AUTOARIMA

Dado que simplifica y acelera la búsqueda de los valores de order y seasonal_order, se empleó la función autoARIMA. El resto de hiperparámetros empleados en este modelo están especificados en la Tabla 11. Para el rango de los valores order y seasonal_order, se incluyó que todos los valores deben variar entre 0 y 2.

4.4.2.2. BÚSQUEDA POR GRILLA

Se diseño una búsqueda por grilla, con las siguientes características:

- Order y seasonal_order, enforce_stationarity y enforce_invertibility, según lo definido en la Tabla 11.
- La elección de la mejor combinación de hiperparámetros se realiza por medio de la métrica: CPM (Ver numeral 5. EVALUACIÓN), la cual se calcula con base a todas las pruebas que se realizan durante la búsqueda por grilla. Esta elección se dio producto del uso que tiene dicha métrica para la evaluación de los modelos.

Luego de la definición de la grilla se estableció el modelo con la variable afluencia como un regresor y se realizó la respectiva ejecución para cada partición previamente definida (Ver tabla 12). A continuación, se muestra un resumen de los experimentos ejecutados

Tabla 1313. Combinaciones por experimento

Modelo	Particiones*	Estaciones	Order	Seasonal_Order	Total para 6 estaciones
autoARIMA	4	6	3^3	3^3	17496
Arima Grilla	4	6	3^3	3^3	17496

** Particiones, hace referencia a lo documentado en el numeral 4.5*

Finalmente se obtuvieron las mejores predicciones, con los siguientes hiperparámetros:

Tabla 1414. Hiperparametros para los mejores modelos en autoARIMA y Grid Search

Estación	autoARIMA	Arima búsqueda por Grilla
POB	order (0, 2, 2), Seasonal order (0, 2, 2, 12)	order (0, 0, 0) Seasonal order (0, 0, 0, 12)
SAA	order (0, 2, 2) Seasonal order (0, 2, 2, 12)	order (0, 0, 0) Seasonal order (0, 0, 0, 12)
HOS	order (0, 2, 2) Seasonal order (0, 2, 2, 12)	order (0, 0, 0) Seasonal order (0, 0, 0, 12)
ACE	order (0, 2, 2) Seasonal order (0, 2, 2, 12)	order (0, 0, 0) Seasonal order (0, 0, 0, 12)
LH	order (0, 2, 2) Seasonal order (0, 2, 2, 12)	order (0, 0, 0) Seasonal order (0, 0, 0, 12)
LJ	order (0, 0, 0) Seasonal Order (2, 2, 0, 12)	order (0, 0, 0) Seasonal order (0, 0, 0, 12)

4.5. MODELO HOLT WINTERS

Dadas las bondades del modelo Holtwinters y las características de las bases de datos, se empleó el modelo de suavizado exponencial, según el siguiente diseño de pruebas e hiperparámetros empleados.

4.5.1. DISEÑO DE PRUEBA 1

Se implementó una búsqueda por grilla para Holtwintes con los siguientes hiperparámetros:

Tabla 1515. Hiperparametros utilizados en el modelo Holtwinters 1

Hiperparametro	Función	Grilla
Trend	Componente de crecimiento o decrecimiento para analizar el comportamiento de tendencia de las ventas	Add, mul, none
Seasonal	Componente de estacionalidad, ya sea aditiva o multiplicativa	Add, mul, none
Seasonal_periods	Longitud del ciclo de la estacionalidad. Para este modelo también sería 12 (meses)	12
α (smoothing level)	Controla el peso asignado al nivel de la serie Nivel (ℓ_t)	[0.1, 0.3, 0.5, 0.7, 0.9]
β (smoothing slope)	Ajusta la tasa de cambio de la Tendencia (b_t)	[0.1, 0.3, 0.5, 0.7, 0.9]
γ (smoothing seasonal)	Modula la reacción ante variaciones estacionales (s_t):	[0.1, 0.3, 0.5, 0.7, 0.9]

El total de combinaciones a evaluar es:

$$3\ (Trend) \times\ 3\ (Seasonal) \times\ 5\ (\alpha) \times\ 5\ (\beta) \times\ 5\ (\gamma)\ \times Estaciones\ =\ 6750\ pruebas$$

Finalmente se obtuvo el mejor resultado para cada una de las estaciones:

Tabla 1616. Metricas modelo Holtwinters 1

Estación	Holtwinters	MAE (M)	RSME (M)	SMAPE (%)
POB	α=0.2, β=0.1, γ= 0.3 damped=True, trend='add', seasonal='mul'	67,07	75,04	52,98
SAA	α=0.2, β=0.2, γ= 0.2 damped=True, trend='add', seasonal='mul'	55,46	64,05	40,38
ACE	α=0.2, β=0.1, γ= 0.3 damped=True, trend='add', seasonal='mul'	14,02	15,84	156,84
HOS	α=0.2, β=0.1, γ= 0.3 damped=True, trend='add', seasonal='add'	9,71	13,67	177,4
LH	α=0.2, β=0.1, γ= 0.3 damped=True, trend='add', seasonal='add'	5,57	6,57	94,01
LJ	α=0.2, β=0.1, γ= 0.3 damped=True, trend='add', seasonal='add'	67,07	75,04	52,98

4.6. MODELO ELMAN

Se implementó una Red Neuronal Recurrente de tipo Elman con el propósito de desarrollar un modelo predictivo que permita estimar las ventas totales mensuales en cada una de las estaciones analizadas. La red Elman se distingue por su capacidad para capturar dependencias temporales, ya que incorpora conexiones de retroalimentación que le permiten almacenar información de

estados anteriores. Esta característica la hace particularmente adecuada para el tratamiento de series temporales, como es el caso de los datos utilizados en este proyecto.

El comportamiento de las ventas en el contexto evaluado evidencia una dependencia significativa respecto a los valores históricos de afluencia de pasajeros. Específicamente, los registros de afluencia de meses anteriores influyen directamente en los valores futuros de ventas, configurando patrones estacionales y dinámicos que la red Elman puede modelar de manera eficiente para lograr predicciones más precisas.

El proceso de implementación comenzó con la carga de los datos desde un archivo ".csv", el cual contenía las variables clave necesarias para el modelado: afluencia de pasajeros, ventas totales y las fechas correspondientes a cada registro. Posteriormente, se realizó la organización y limpieza de la base de datos, asegurando la correcta estructuración de los datos de fecha y la eliminación de inconsistencias. Cada estación fue filtrada individualmente, permitiendo realizar un análisis de descomposición temporal para entender mejor sus patrones de tendencia, estacionalidad y residuales.

Una vez realizado el análisis, el filtrado y la separación de los datos para el entrenamiento y prueba, se procede a realizar el preprocesamiento de los datos, esto con el objetivo de mejorar la eficiencia del entrenamiento y evitar problemas de escala, por lo que se normalizan entre un rango de [0,1] las variables de entrada (afluencia y venta total) mediante el método MinMaxSacler, lo cual facilita la convergencia del modelo y mejora la estabilidad numérica del proceso de aprendizaje.

El modelo este compuesto por una capa recurrente SimpleRNN, la cual está encargada de procesar la información secuencial y de capturar las dependencias temporales de los registros, seguida de una capa densa de salida con una sola neurona, cuya función es emitir la predicción mensual del valor de ventas. El modelo fue compilado por el optimizador Adam, el cual es reconocido por su eficiencia en problemas de redes neuronales.

Para seleccionar los hiperparámetros más adecuados, se implementó un procedimiento de búsqueda exhaustiva tipo Grid Search. Se evaluaron combinaciones de los siguientes hiperparámetros:

Tabla 1717. Hiperparámetros Seleccionados ELMAN

Hiperparámetro	Descripción y función	Valores
Units	Número de unidades en la capa oculta recurrente, controla la capacidad del modelo para aprender patrones complejos.	[10, 20, 30]
Activaction	Función de activación en las capas de la RNN, esto define como se transforma la salida de cada neurona.	[Tanh, relu]

Hiperparámetro	Descripción y función	Valores
Batch_Size	Numero de muestras usadas para actualizar los pesos en una iteración, controla cuantos ejemplos se procesan antes de realizar un paso de propagación.	[4, 8]
Epochs	Número de veces que el modelo verá todo el conjunto de entrenamiento.	[100, 200]

Cada combinación de hiperparámetros fue entrenada y evaluada de forma independiente sobre el conjunto de prueba. La combinación de hiperparámetros que produjo los mejores resultados (mínimos errores de predicción) fue seleccionada para el modelo final de cada estación. Finalmente, las métricas obtenidas y los hiperparámetros seleccionados fueron consolidados y exportados a un archivo ".csv", facilitando su análisis comparativo y la generación de reportes de desempeño por estación.

Finalmente, por el método de Grid Search se obtuvieron las mejores predicciones, con los siguientes hiperparámetros, y sus respectivas métricas:

Tabla 1818. Métricas modelo ELMAN

Estación	ELMAN por Grid Search	MAE (M)	RSME (M)	SMAPE (%)
POB	Units=30 Activation=tanh Batch_size=4 Epochs=100	61,40	68,27	46,66
SAA	Units=30 Activation=tanh Batc_size=4 Epochs=100	55,89	64,80	37,65
ACE	Units=30 Activation=tanh Batc_size=4 Epochs=100	13,80	14,61	86,73
HOS	Units=30 Activation=tanh Batc_size=4 Epochs=100	10,14	10,50	128,83
LH	Units=20 Activation=tanh Batc_size=4 Epochs=100	5,20	7,53	187,66
LJ	Units=10 Activation=tanh Batc_size=4 Epochs=200	1,85	2,42	122,39

5. EVALUACIÓN

Para todas las estaciones se evaluaron nueve modelos de predicción: Elman, 4 variaciones de Holt-Winters, 2 variaciones de Prophet y 2 variaciones de Arima, con base en su comportamiento gráfico y su desempeño cuantitativo medido mediante la métrica compuesta de desempeño (CPM).

5.1. RESULTADOS POR ESTACION (METRICAS DE DESEMPEÑO)

Tabla 1919. Resultados Metricas de Desempeño por Estación

Métrica	Estación	AR1	AR2	HW1	HW2	FP1	FP2	ELMAN
CPM	POB		2,7	3,0	2,3	**2,0**	**2,0**	2,7
	SAA	2,9	2,1	2,2	**1,5**	2,7	2,7	2,2
	HOS	3,0	2,0	2,6	2,3	**1,7**	**1,7**	1,9
	ACE	1,9	2,1	2,6	**1,4**	2,1	2,1	2,5
	LH	2,5	2,5	2,2	**1,9**	**1,9**	**1,9**	2,7
	LJ	1,9	**1,9**	2,5	2,6	2,5	2,5	2,6

Se evidencia que el desempeño varía significativamente entre estaciones, lo que sugiere que la dinámica de las ventas de publicidad tiene patrones locales específicos que impactan la precisión de los modelos. En términos generales, los modelos **Holt-Winters** (particularmente **HW1** y **HW2**) y **Facebook Prophet** (**FP1** y **FP2**) presentan un rendimiento consistente, ubicándose frecuentemente entre las dos mejores alternativas. Por ejemplo, en estaciones como **ACE** y **SAA**, los modelos **HW1 y HW2** logran captar con precisión los componentes estacionales y de tendencia, alcanzando CPMs de 1,4 y 1,5, respectivamente. En contraste, en **POB** y **HOS**, modelos como **FP1 y FP2** sobresalen, lo que podría atribuirse a su capacidad de incorporar estacionalidades múltiples y factores externos, lo cual es relevante en estas estaciones de mayor complejidad comercial. En el caso de **LH**, se observa un empate entre cuatro modelos (HW3, HW4, FP1 y FP2), lo que sugiere un equilibrio entre la capacidad de modelar variaciones abruptas y mantener estabilidad predictiva. Finalmente, en **LJ**, los modelos **AR1 y AR2**, ambos autorregresivos, ofrecen la mejor precisión con un CPM de 1,9, lo que indica que, en esa estación, la dinámica temporal de ventas puede explicarse de forma más parsimoniosa mediante dependencia directa de sus propios rezagos. Este análisis refuerza la importancia de una selección de modelo por estación, en lugar de una aproximación global, e indica que los modelos Holt-Winters (HW3, HW4) y Prophet (FP1, FP2) deben considerarse como candidatos robustos en escenarios donde se espera alta variabilidad o complejidad estacional.

Por otro lado, la diferencia entre FP1 y FP2 consiste principalmente en la búsqueda de grilla utilizada. Para el caso del FP2 se utilizó una grilla más amplia la cual permitió evaluar 3840 modelos, el cual hace que se eleve el consumo de recursos computacionales y por ende el tiempo

de ejecución del modelo ascendió a unas 4 horas y media aproximadamente; en contraste con el modelo FP1 el cual utilizo una grilla más acotada la cual evaluó 432 modelos con un tiempo de ejecución de 27 min aproximadamente. Finalmente, se opta por utilizar el FP1 ya que los resultados visuales y estadísticos no presentan gran diferencia y el consumo de recursos computacionales es menor, por lo cual es un mejor modelo para el ámbito operativo del metro de Medellín.

A pesar de que el modelo Elman es uno de los modelos más robustos para la predicción de variables en series de tiempo, en el proyecto sus resultados no fueron los más destacados en comparación con otros modelos como Holt-Winters o Prophet. Esto debido principalmente a la cantidad y calidad de los datos dispuestos por la empresa Metro de Medellín, especialmente luego de segmentar para realizar la proyección por estaciones. Al realizar esta segmentación el volumen de datos para cada modelo entrenado disminuye significativamente, afectando de esta manera la capacidad de aprendizaje de la red neuronal, la cual requiere gran volumen de datos para capturar adecuadamente los patrones temporales y minimizar el riesgo de sobreajuste o sub-ajuste. En contextos donde hay pocos datos y mucha variación entre estaciones, el rendimiento del modelo puede verse afectado, especialmente si se compara con otros modelos más simples que no necesitan tanta información para funcionar bien.

5.2. RESULTADOS POR ESTACION (Gráficos)

A continuación, se muestran los resultados obtenidos para los mejores modelos según las métricas de desempeño relacionadas en la sesión anterior.

5.2.1. POBLADO (POB)

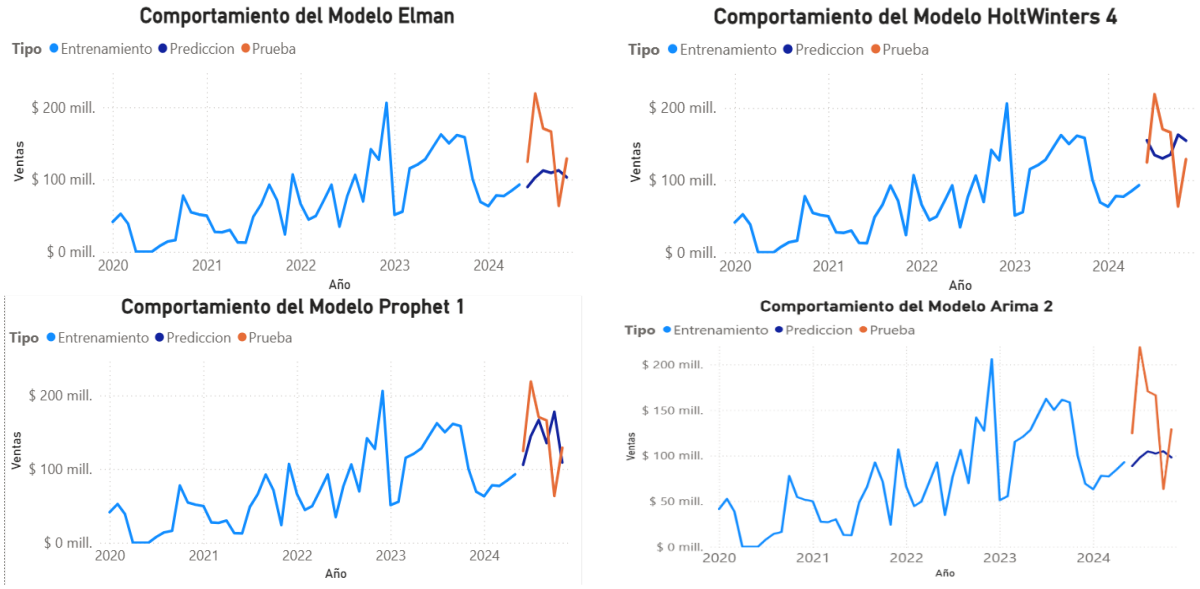

Ilustración 19. Comparación entre modelos para la estación Poblado (POB)

Visual y estadísticamente, el modelo Prophet 1 mostró el mejor ajuste a la serie real, replicando con mayor precisión las oscilaciones, picos y caídas observadas en los datos del período de prueba, lo cual indica una buena capacidad para capturar tanto la tendencia como la estacionalidad. Por su parte, el modelo Holt-Winters 4, aunque no tan preciso en la forma de las curvas, obtuvo el segundo mejor resultado cuantitativo con un valor CPM de aproximadamente 2.3, lo que lo posiciona como el segundo modelo más eficiente en términos de minimización del error compuesto.

El modelo Elman también presentó un ajuste aceptable, pero mostró desviaciones en la sincronización temporal y una tendencia a suavizar los picos, con un CPM de 2.7. Finalmente, el modelo Arima 2, si bien siguió una tendencia general coherente, fue menos eficaz en capturar la variabilidad de la serie.

En conclusión, Prophet 1 ofrece el mejor desempeño visual y estadístico, resulta más adecuado cuando se busca interpretar y seguir fielmente la dinámica temporal de la serie en contextos operativos.

5.2.2. SAN ANTONIO A (SAA)

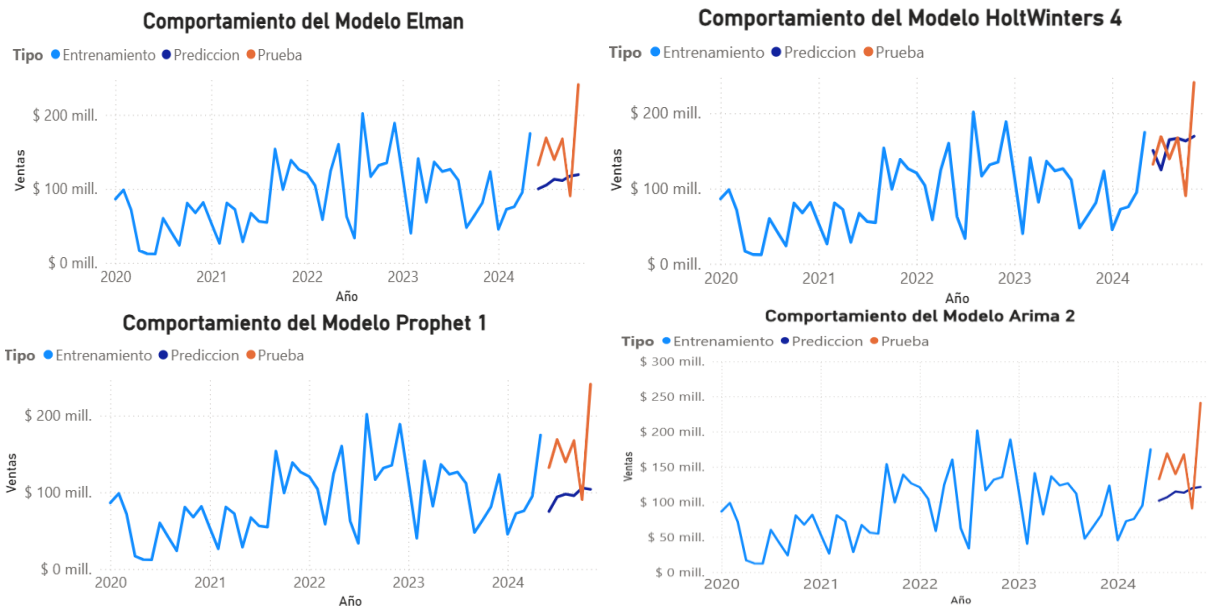

Ilustración 20. Mejores modelos para la estacion San Antonio A (SAA)

En la estación San Antonio A (SAA), el modelo Holt-Winters 4 se destaca como la mejor alternativa tanto en términos visuales como estadísticos. Este modelo logra capturar de forma efectiva la estacionalidad de la serie, aunque presenta un leve desfase temporal. A pesar de esta ligera

desincronización, obtiene un valor de CPM de 1.5, el más bajo entre todos los modelos evaluados para esta estación, lo que respalda su precisión en la predicción.

Por otro lado, los modelos Prophet 1, Elman y Arima 2 muestran comportamientos visuales similares entre sí, con una correcta identificación de la tendencia creciente de la serie, pero una marcada suavización de los picos. Esto sugiere que, si bien estos modelos logran seguir la dirección general de los datos, presentan limitaciones para capturar adecuadamente las variaciones abruptas, lo cual puede afectar la sensibilidad de sus predicciones frente a cambios relevantes en el comportamiento de la serie.

5.2.3. HOSPITAL (HOS)

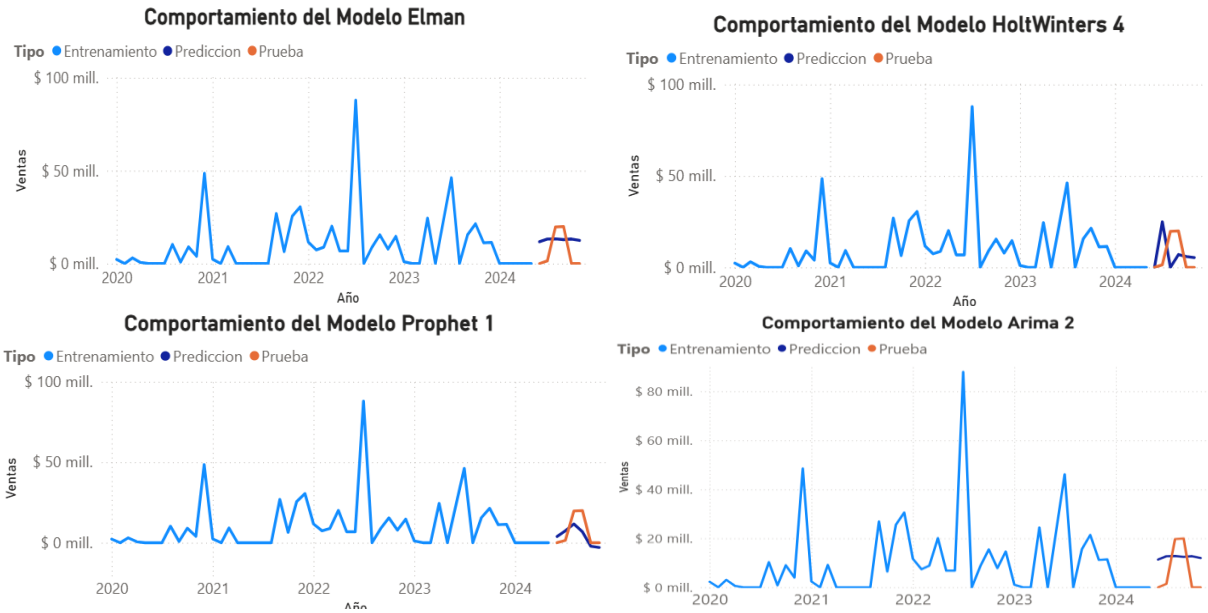

Ilustración 21. Mejores modelos para la estación Hospital (HOS)

En la estación Hospital, se observa que el modelo de Elman y Arima 2 tienen predicciones muy similares, siendo el modelo de Elman un poco más sobre ajustado que el modelo ARIMA. Sin embargo, estos dos modelos no muestran visualmente un buen ajuste, son algo conservadores y no siguen bien la distribución de los datos.

Por otra parte, Holt Winters 4 pareciera tener un mejor ajuste y estadísticamente es el segundo mejor modelo con un CPM de 2.3. A pesar de tener un buen ajuste no logra superar el ajuste del modelo Prophet 1, el cual es el más atractivo visual y estadísticamente tiene el mejor CPM para esta estación con un valor de 1,65

5.2.4. ACEVEDO (ACE)

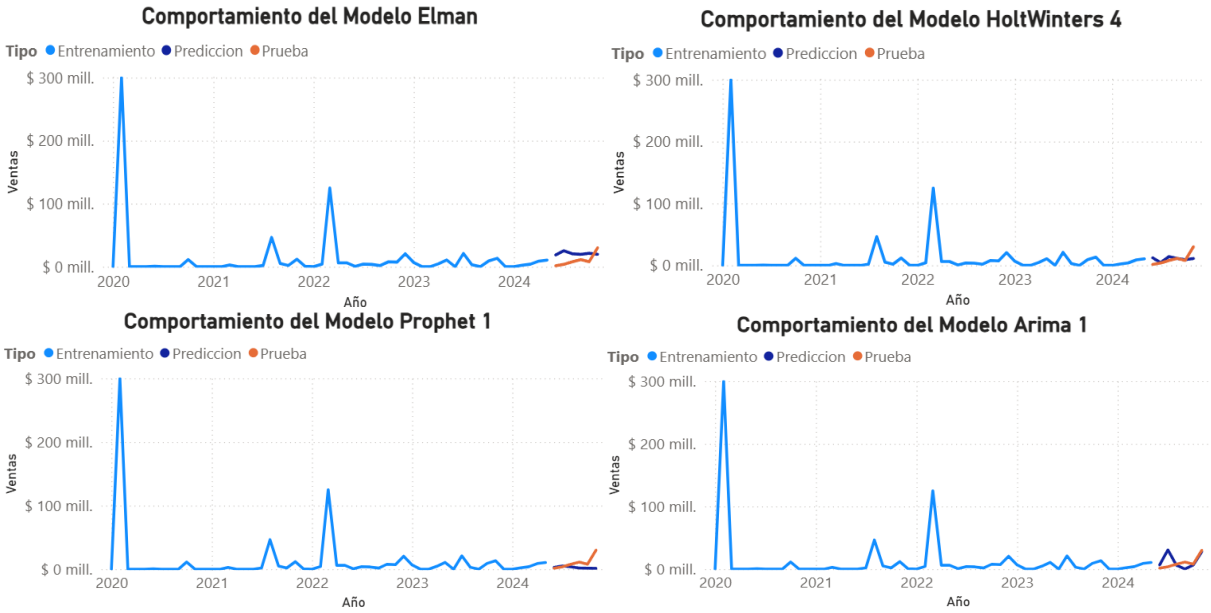

Ilustración 22. Mejores modelos para la estacion Acevedo (ACE)

En la estación Acevedo (ACE), los resultados evidencian que el modelo con mejor desempeño en términos del índice compuesto de desempeño (CPM) fue Holt-Winters 4, con un valor de 1.4, el más bajo entre todos los modelos evaluados para esta estación. Esto indica una mayor capacidad predictiva del modelo en cuanto a precisión general y estabilidad ante variaciones. Esta conclusión se respalda visualmente en la gráfica correspondiente al comportamiento de Holt-Winters 4, donde se observa un buen ajuste entre la predicción (línea negra) y los datos reales del periodo de prueba (línea naranja), manteniendo la coherencia con los patrones históricos sin sobreajustes evidentes.

En comparación, aunque el modelo ARIMA 1 también logra un ajuste razonable tanto visual y estadísticamente con un CPM de 1.9. Por su parte, el modelo Prophet 1, con un CPM de 2.1, aunque también competitivo, no logra igualar la precisión alcanzada por Holt-Winters 4. Finalmente, el modelo Elman contiene el CPM más alto con un valor de 2.5 y visualmente su curva muestra ligeros desfaces respecto a la serie real.

En conclusión, el análisis cuantitativo y visual confirma que Holt-Winters 4 es el modelo más adecuado para predecir el comportamiento de ventas en la estación Acevedo, dada su alta capacidad de captura de estacionalidades y tendencias locales en la serie temporal.

5.2.5. LÍNEA H (LH)

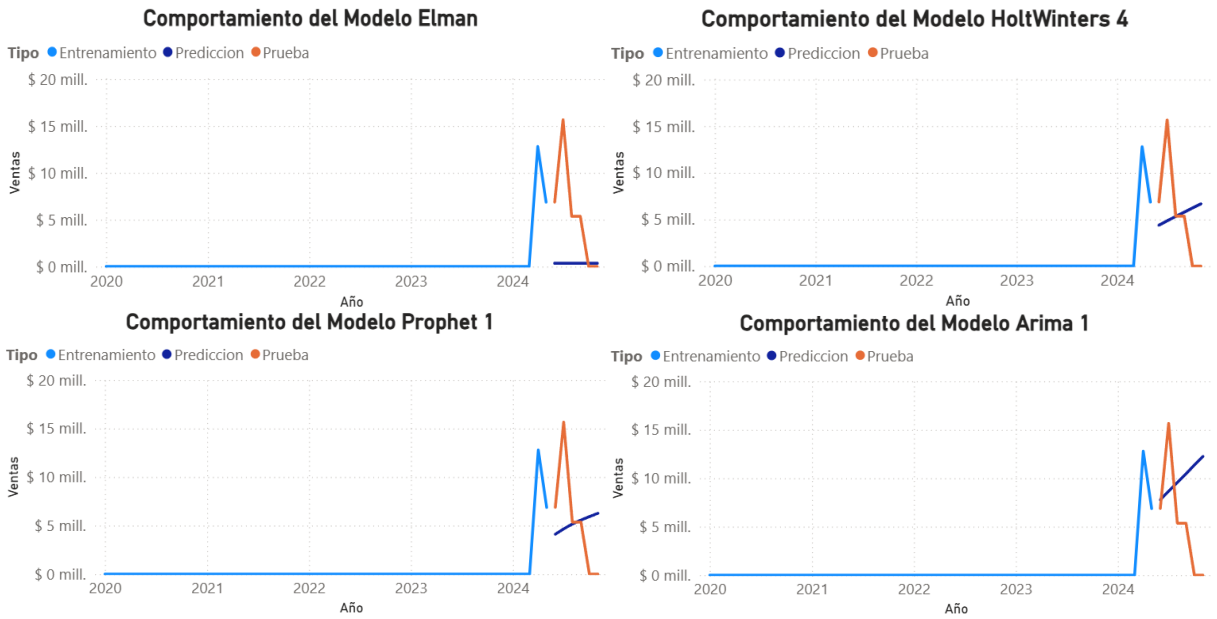

Ilustración 23. Mejores modelos para la estacion Línea H (LH)

En la estación Línea H (LH), el modelo que obtuvo el mejor desempeño según la métrica CPM fue Prophet 1, con un valor de 1.8, ligeramente inferior al segundo mejor modelo cuantitativamente hablando el cual es Holt Winters 4, que obtuvo un CPM de 1.9, esto sugiere que ambos modelos son competitivos, pero el modelo Prophet logra un mejor equilibrio entre precisión y estabilidad. Finalmente, para el modelo Arima 1 se obtuvo un CPM con valor de 2.5 siendo uno de los resultados más altos para esta métrica en comparación con los modelos anteriores, y el modelo Elman es el que mayor valor obtiene con un valor de 2.7.

Sin embargo, al observar la gráfica de comportamiento se puede concluir que para todos los modelos no se logra un ajuste optimo en general, pero si tomamos en cuenta solo el primer datos de la predicción podemos observar que el que mejor se ajusta es el modelo Arima 1 en donde la predicción tiende a sobreestimar los valores reales, pero el valor predicho se encuentra mucho más cercano visualmente que los modelos Prophet y Holt Winters que predicen la tendencia descendente de los valores con una mayor subestimación de los datos reales. Por su parte, el modelo Elman presenta una desviación más marcada, mostrando una predicción que no logra capturar adecuadamente la caída abrupta de las ventas.

En resumen, Prophet 1 es el modelo más adecuado para la estación LH, ya que proporciona el menor error compuesto y presenta una forma de predicción más coherente con la dinámica observada en los datos reales del periodo de prueba.

5.2.6. LÍNEA J (LJ)

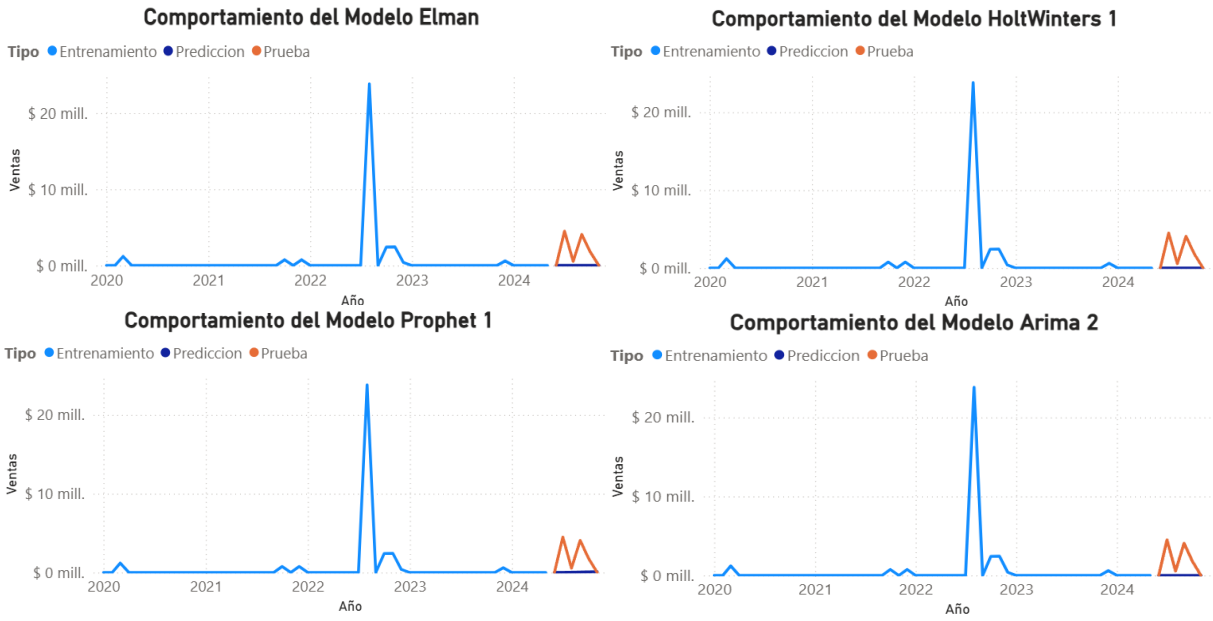

Ilustración 24. Mejores modelos para la estacion Línea J (LJ)

En la estación Línea J (LJ), el modelo con mejor desempeño de acuerdo con el índice CPM fue ARIMA 2, con un valor de 1.9, sumamente mejor al alcanzado por otros modelos, ya que como se pudo observar en la tabla 18 al inicio de esta sección el modelo con el segundo mejor desempeño de acuerdo con el índice CPM fue Prophet, con un valor de 2.4. Luego los siguen los modelos Holt Winters y Elman con valores de 2.5 y 2.6 respectivamente.

Sin embargo, visualmente se observa que ninguno de los modelos logra un ajuste perfecto a la forma de los datos reales del periodo de prueba. Todos los modelos presentan una subestimación marcada de los valores reales, lo que se evidencia en predicciones significativamente por debajo de los picos observados en las observaciones de prueba. Esta consistencia en el comportamiento predicho frente a los patrones reales sugiere que ARIMA 2 es el modelo más adecuado estadísticamente hablando para la estación LJ.

6. PROYECCIÓN

6.1. TABLERO

Luego de los resultados obtenidos y ya expuestos en los anteriores numerales, se desarrolló un tablero dinámico por medio de la herramienta Power BI, con el objetivo de que El Metro de Medellín pueda emplear todo lo desarrollado en el proyecto para el cumplimiento de su propósito, potenciando las estrategias del negocio por medio del uso de los datos y la información que este contiene.

El tablero se trabajó en 4 componentes principales:

- Afluencia: Donde se puede analizar el comportamiento de la afluencia de usuarios en la Red Metro.
- Ventas: Permite analizar el comportamiento de las ventas en los diferentes servicios publicitarios
- Ventas-Afluencia: Permite analizar en paralelo como se comportan ambas series de tiempo
- Proyección: Se emplea para analizar el comportamiento de las ventas con base en los modelos elegidos en el proyecto.

6.1.1. AFLUENCIA

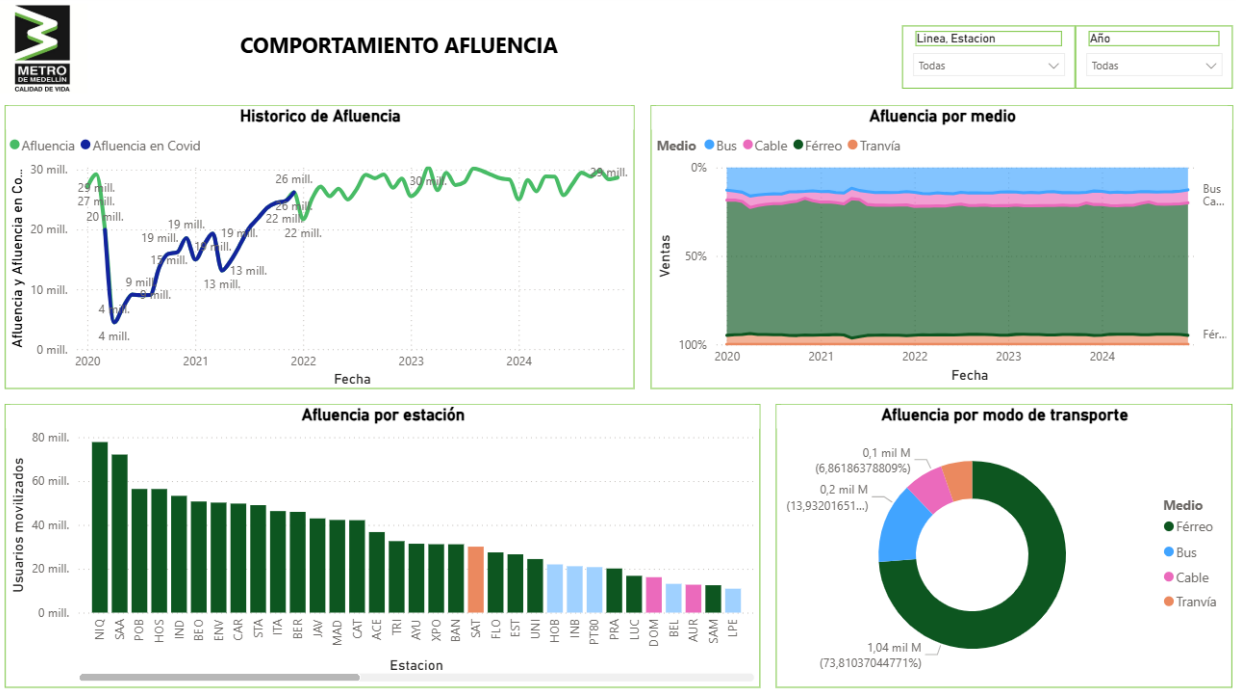

Ilustración 25. Tablero Afluencia

El tablero permite realizar un análisis integral de la evolución y distribución del número de usuarios movilizados en el sistema Metro de Medellín. En primer lugar, el gráfico de histórico de afluencia destaca el crecimiento progresivo de los usuarios desde el 2020, evidenciando una recuperación sostenida posterior al impacto de la pandemia de COVID-19, representado en azul. Esta visualización permite observar cómo la afluencia ha retornado e incluso superado los niveles prepandemia, alcanzando picos de hasta 30 millones de usuarios mensuales.

Además, el panel de afluencia por medio y el gráfico circular de afluencia por modo de transporte muestran claramente que el sistema férreo concentra la mayor proporción de los viajes, seguido en menor medida por buses, cables y tranvía. Esta dominancia del modo férreo se mantiene estable en el tiempo, como se evidencia en la gráfica de distribución porcentual por fecha. Por otro lado, el gráfico de afluencia por estación permite identificar las estaciones con mayor volumen de usuarios movilizados, destacándose Niquía, San Antonio y Poblado como puntos estratégicos del sistema. En conjunto, este tablero facilita la comprensión de los patrones de movilidad por modo, estación y tiempo, sirviendo como base para decisiones operativas y estratégicas orientadas a la mejora del servicio y asignación eficiente de recursos.

6.1.2. VENTAS

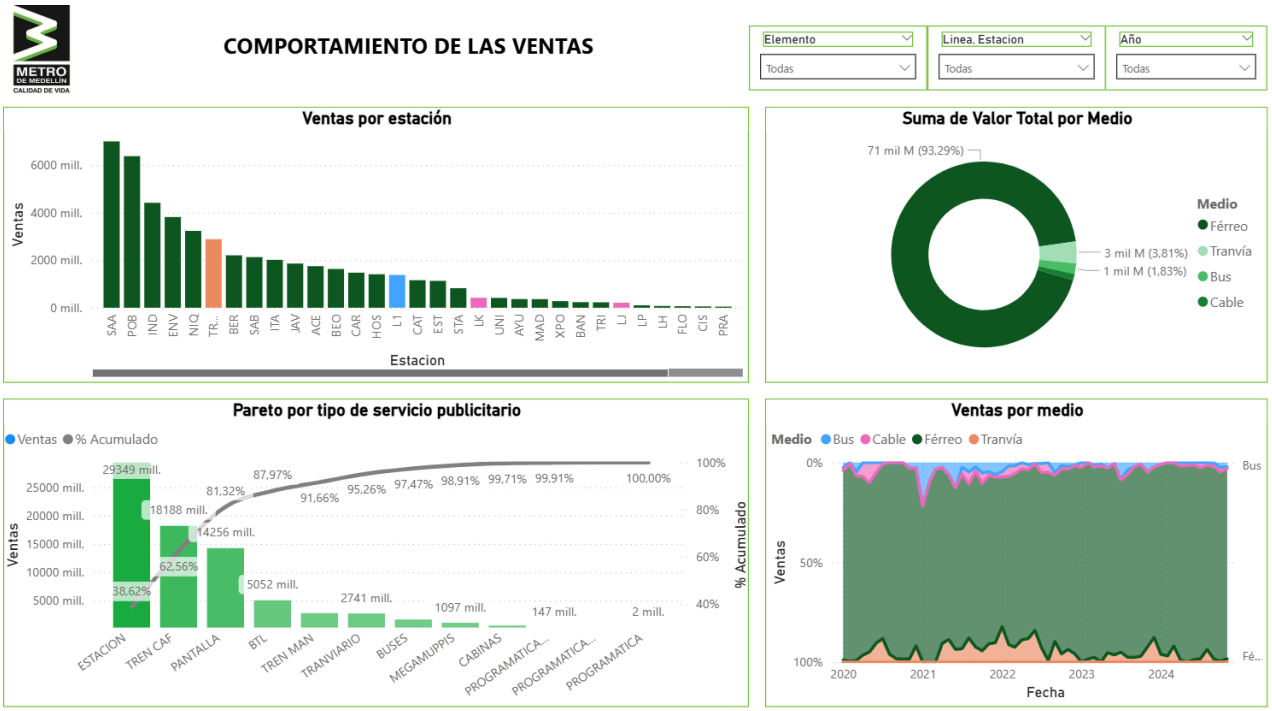

Ilustración 26. Tablero Ventas

El tablero ofrece una visualización clara del desempeño económico generado por los servicios publicitarios distribuidos en las estaciones del Metro de Medellín. En el gráfico de ventas por estación se evidencia que las estaciones SAA, POB y ENV concentran la mayor parte del valor comercial, lo que indica su relevancia estratégica como puntos de alto tráfico y potencial publicitario. Este patrón se refuerza en el gráfico de suma del valor total por medio, donde el sistema férreo representa más del 93% de los ingresos, consolidándose como el principal canal de generación de ventas, muy por encima de otros medios como tranvía, bus o cable.

Adicionalmente, el análisis del Pareto por tipo de servicio publicitario revela que las categorías “Estación”, “Tren CAF” y “Pantalla” acumulan cerca del 82% de los ingresos, lo que permite priorizar estos elementos en futuras estrategias comerciales. El gráfico de ventas por medio en el tiempo permite observar que esta distribución por tipo de transporte se mantiene relativamente estable, siendo el sistema férreo el dominante durante todo el periodo analizado. En conjunto, este tablero proporciona una visión estratégica de los focos de ingresos, permitiendo identificar las estaciones y formatos publicitarios más rentables, así como orientar decisiones en torno a la optimización del portafolio comercial.

6.1.3. VENATS - AFLUENCIA

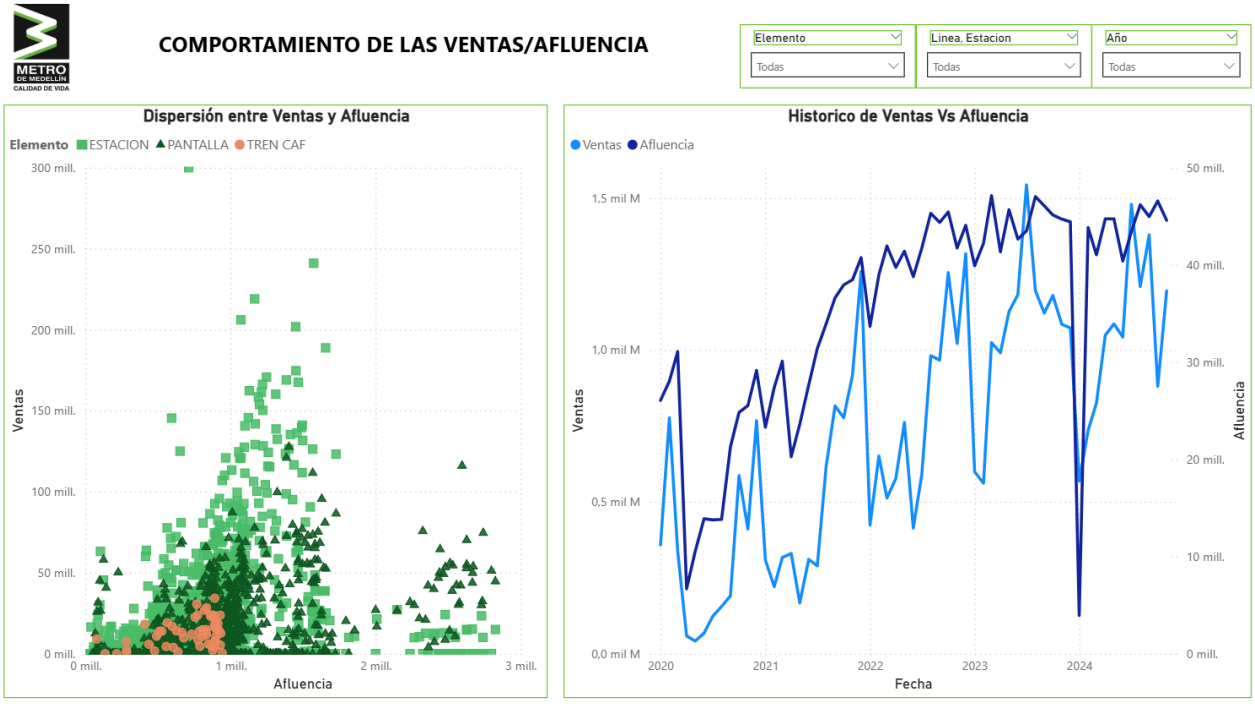

Ilustración 27. Tablero del comportamiento en Ventas y Afluencia

El tablero permite analizar la relación entre la cantidad de usuarios movilizados y los ingresos por servicios publicitarios en el sistema Metro de Medellín. A través del gráfico de dispersión entre

ventas y afluencia, se puede observar que existe una tendencia positiva entre ambas variables, es decir, a mayor afluencia de usuarios, tienden a registrarse mayores niveles de ventas. No obstante, esta relación no es completamente lineal, lo que sugiere que otros factores como el tipo de elemento publicitario (estación, pantalla, tren CAF), influyen de manera significativa en los ingresos obtenidos. Este análisis resulta clave para identificar oportunidades de optimización en la asignación del portafolio publicitario, priorizando puntos de alta visibilidad y concentración de usuarios.

Por su parte, el gráfico de histórico de ventas vs afluencia permite observar la evolución temporal de ambas variables. Se evidencia un comportamiento correlacionado en la mayoría del periodo analizado, con una recuperación sostenida luego de la pandemia y ciertos momentos de desacople temporal, probablemente asociados a cambios operativos, eventos especiales o estrategias comerciales puntuales. Esta herramienta resulta fundamental para analizar el impacto directo de la movilidad sobre los ingresos, permitiendo identificar patrones estacionales, evaluar la efectividad de campañas y planificar estrategias comerciales con base en el comportamiento real del sistema.

6.1.4. PROYECCIÓN DE VENTAS

Ilustración 28. Tablero del resultados de los modelos

Este tablero permite visualizar de manera detallada el comportamiento de los modelos de predicción aplicados a cada estación, facilitando una comparación visual entre los valores históricos reales y los valores proyectados por cada modelo. En el gráfico se distinguen claramente

las fases de entrenamiento (línea azul), predicción (línea morada) y datos reales de prueba (línea naranja), lo que permite evaluar el grado de ajuste que logra cada modelo respecto a los datos observados en el periodo de prueba.

6.2. RESULTADOS

Finalmente, para realizar la validación de los resultados obtenidos con base a las necesidades de negocio, se tuvieron en cuenta 4 criterios para la selección de los mejores modelos para la operación, los criterios son:

- Métrica de desempeño CPM para los modelos evaluados se promedia el CPM de todas las estaciones a continuación se muestran los resultados

Tabla 2020. Consolidado metrica de desempeño CPM

	Modelos			
Estación	AR2	HW2	FP1	ELMAN
POB	2,7	2,3	2,0	2,7
SAA	2,1	1,5	2,7	2,2
ACE	2	2,3	1,7	1,9
HOS	2,1	1,4	2,1	2,5
LH	2,5	1,9	1,9	2,7
LJ	1,9	2,6	2,5	2,6
CPM promedio	2,22	2,00	2,15	2,43

- Timing o ajuste visual, la cual se califica de la siguiente manera:

Tabla 2121. Definicion criterio de validacion Ajuste visual

Criterio	Calificación
Bueno	1
Medio	2
Malo	3

Tabla 2222. Ccalificacion del ajuste visual

	Modelos			
Estación	**AR2**	**HW2**	**FP1**	**ELMAN**
POB	3	2	1	3
SAA	3	1	3	3
ACE	2	2	2	3
HOS	3	2	1	3

	Modelos			
Estación	**AR2**	**HW2**	**FP1**	**ELMAN**
LH	3	3	3	3
LJ	3	3	3	3
Timing promedio	2,83	2,17	2,17	3,00

- Tiempo de ejecución en minutos

Tabla 2323. Tiempos de ejecución por modelo

Tiempo de ejecución			
AR2	**HW2**	**FP1**	**ELMAN**
4,5	5	27	40

- Facilidad de implementación, la cual se califica de la siguiente manera:

Tabla 2424. Definición criterio Facilidad de implementación

Criterio	Calificación
Fácil	1
Medio	2
Difícil	3

Tabla 2525. Calificacion criterio Facilidad de implementación

Facilidad de implementación por modelo			
AR2	**HW2**	**FP1**	**ELMAN**
1	1	1	3

- Cantidad de estaciones, este criterio evalúa que modelo tuvo los mejores resultados en diferentes estaciones, se cuantifica contando la cantidad de estaciones con mejor desempeño por modelo, como este criterio es inversamente proporcional al resto, se calcula su inverso de la siguiente forma.

Ecuación 6. Formula utilizada para invertir la variable cantidad de estaciones

$$\frac{1}{x+1}$$

Posteriormente, se construye una matriz de criterios de priorización previamente definidos.

Tabla 2626. Consolidado criterios de validación

Modelo	Métrica (CPM)	Timing	Tiempo de ejecución (min)	Facilidad de implementación	Cantidad de estaciones
AR2	2,22	2,83	4,5	1	0,5
HW2	2,00	2,17	1	1	0,25
FP1	2,15	2,17	27	1	0,25
ELMAN	2,43	3	40	3	1

Después de consolidar los resultados, se procede a normalizar los datos mediante la siguiente formula:

Ecuación 7. Formula de normalización

$$Z = \frac{(x - \mu)}{\sigma}$$

Finalmente, se procede a calcular el Score sumando cada uno de los criterios normalizados.

Tabla 2727. Validación modelos

Modelo	Métrica (CPM)	Timing	Tiempo de ejecución (min)	Facilidad de implementación	Cantidad de estaciones	Score
AR2	0,11	0,66	-0,73	-0,50	0,00	-0,46
HW2	-1,12	-0,85	-0,92	-0,50	-0,71	-4,10
FP1	-0,28	-0,85	0,48	-0,50	-0,71	-1,86
ELMAN	1,29	1,05	1,18	1,50	1,41	6,43

Como se puede observar en la anterior tabla, el mejor modelo para implementar basado en los criterios previamente definidos es Holtwinters.

7. CONCLUSIONES Y TRABAJOS FUTUROS

7.1. CONCLUSIONES

- Inicialmente se planteó la necesidad de desarrollar un modelo, por medio del aprendizaje automático, que permitiera proyectar las ventas en los servicios publicitarios del Metro de Medellín. Sin embargo, dadas las características de la base de datos, la cantidad de variables categóricas que se tenían en la base de datos, las circunstancias del negocio y los resultados de colinealidad y cointegración, se determinó que cada estación del sistema debe contar con su propio modelo, empleando todo el rigor técnico de la Ciencia de Datos para alcanzar el objetivo planteado, lo cual permitirá una mayor precisión el en pronóstico de las ventas.
- El conocimiento en profundidad del negocio y la rigurosidad en la preparación de los datos resultó ser un factor clave en el éxito del proyecto. Durante el desarrollo, el equipo tuvo que sortear una serie de retos, los cuales se enfocaron en mejorar la robustes del proyecto, el rigor técnico y el contexto del negocio, permitiendo unos resultados con modelos lo mejor ajustados posible.
- Para este proyecto se emplearon las métricas de evaluación SMAPE, MAE y RMSE. Debido a la variedad de modelos que se probaron, para encontrar el más adecuado para el negocio en cada una de las estaciones analizadas se implementó la métrica CPM con el propósito de identificar el modelo más adecuado con las métricas elegidas por el equipo, evitando sesgos a la hora de definir la métrica que indique el mejor modelo en términos estadísticos.
- Para las estaciones Poblado (POB) y SAA (San Antonio A), las cuales tienen una alta relación entre la afluencia y las ventas, se identificó que las métricas de evaluación y el ajuste en las predicciones son las que más se ajustan en el set de prueba, lo cual corrobora el resultado del análisis realizado entre la relación de ambas variables. Para las estaciones Hospital (HOS) y Acevedo (ACE), se identifica que las métricas de evaluación tienen un resultado menos favorable que los resultados de las estaciones con alta relación. Para las estaciones con baja o nula relación entre dichas variables, se identificó que el comportamiento de las métricas, así como el ajuste de la predicción es muy desfavorable debido a la poca cantidad de datos que se tienen para el entrenamiento del modelo.
- Se generó un tablero interactivo por medio de la herramienta Power BI, el cual se subdividió en 4 pestañas, las cuales se proyectaron según las necesidades del negocio y los resultados del proyecto. En la primera pestaña se tienen unas visualizaciones que permiten analizar el comportamiento de la afluencia en la Red Metro, en la segunda el comportamiento de las ventas según los servicios publicitarios y las ubicaciones de la red,

en la tercera cómo se comporta las ventas en comparación a la afluencia y por último las proyecciones de ventas para las diferentes estaciones que hicieron parte del proyecto.

7.2. TRABAJOS FUTUROS

Como resultado del desarrollo del presente proyecto, se han identificado diversas oportunidades que podrían fortalecer y ampliar el valor generado para el Metro de Medellín. Estos posibles trabajos futuros surgen tanto de los hallazgos obtenidos como de las limitaciones enfrentadas durante el proceso, y representan líneas de acción que permitirían mejorar la calidad de los datos, optimizar los modelos actuales y explorar nuevas aplicaciones analíticas. A continuación, se presentan algunas de estas propuestas, orientadas a continuar con el aprovechamiento estratégico de los datos en la gestión de los servicios publicitarios del sistema.

- Mejorar la Calidad del etiquetado de las bases de datos de los registros contables
- Identificar otras variables que estén estrechamente relacionadas con las ventas para la optimización de los modelos.
- Identificar otras variables para el desarrollo de nuevos modelos enfocados en el pricing de los servicios publicitarios.
- Desarrollo de una aplicación para la ejecución de los modelos de pronóstico.
- Realizar proyección de ventas para las otras estaciones del sistema que cuenten con buena calidad y cantidad de datos y que no fueron incluidas en la etapa de modelado de este proyecto.

8. REFERENCIAS BIBLIOGRÁFICAS

[1] Metro de Medellín, "Memoria de Sostenibilidad", 2024. Consultado: el 25 de mayo de 2024. [En línea]. Disponible en: https://www.metrodemedellin.gov.co/hubfs/memorias-de-sostenibilidad/memoria-de-sostenibilidad-metro-de-medellin-2023.pdf

[2] "Valle de Aburrá: población por municipio, 2020 | Medellín Cómo Vamos". Consultado: el 17 de mayo de 2024. [En línea]. Disponible en: https://www.medellincomovamos.org/node/18687

[3] A. Géron, *Hands-on Machine Learning with Scikit-Learn , Keras & TensorFlow*.

[4] J. Díaz-Ramírez, "Aprendizaje Automático y Aprendizaje Profundo", *Ingeniare. Revista chilena de ingeniería*, vol. 29, núm. 2, pp. 180–181, 2021, doi: 10.4067/s0718-33052021000200180.

[5] G. J. Tellis, • I Redondo, y I. Redondo, "Estrategias de Publicidad y Promoción". [En línea]. Disponible en: www.pearsoneducacion.com

[6] IBM, "ModelerCRISPDM", Consultado: el 12 de junio de 2024. [En línea]. Disponible en: https://www.ibm.com/docs/es/SS3RA7_18.4.0/pdf/ModelerCRISPDM.pdf

[7] S. J. Taylor y B. Letham, "Forecasting at scale", el 27 de septiembre de 2017. doi: 10.7287/peerj.preprints.3190v2.

[8] "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition", https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/.

[9] G. Box, G. Jenkins, y G. Reinsel, "Time Series Analysis, Fourth Edition", 2013, pp. 137–191. doi: 10.1002/9781118619193.ch5.

[10] C. Chatfield, *Time-Series Forecasting*. Boca Raton, FL, USA: Chapman & Hall/CRC, 2000.

[11] E. 'Vivar y "Forbes México", "100 Marcas mas valiosas del mundo", https://www.forbes.com.ec/rankings/las-10-marcas-mundiales-mas-valiosas-2023-n45753.

[12] Consultores Drew, "Forecasting". Consultado: el 1 de diciembre de 2024. [En línea]. Disponible en: https://www.wearedrew.co/ss4i/produccion-inteligente/forecasting#:~:text=Permite%20a%20las%20empresas%20optimizar,reducir%20los%20costos%20de%20mantenimiento.

[13] Y. K. Elalem, S. Maier, y R. W. Seifert, "A machine learning-based framework for forecasting sales of new products with short life cycles using deep neural networks", *Int J Forecast*, vol. 39, núm. 4, pp. 1874–1894, 2023, doi: https://doi.org/10.1016/j.ijforecast.2022.09.005.

[14] M. Wan, M. A. Alshahrani, N. M. Aloraini, A. A. Alkhathami, y H. Alqahtani, "On predictive modeling of the twitter-based sales data using a new probabilistic model and machine learning methods", *Alexandria Engineering Journal*, vol. 113, pp. 661–671, 2025, doi: https://doi.org/10.1016/j.aej.2024.11.041.

[15] J. Shao, J. Hong, M. Wang, y X. Wang, “New energy vehicles sales forecasting using machine learning: The role of media sentiment”, *Comput Ind Eng*, p. 110928, 2025, doi: https://doi.org/10.1016/j.cie.2025.110928.

[16] Metro de Medellín, “Mapa esquemático del Metro de Medellín - 2021”, 2021. [En línea]. Disponible en: https://www.metrodemedellin.gov.co/hs-fs/hubfs/v1-mapa%20esquematico-2021%20(1).webp?width=1200&height=1638&name=v1-mapa%20esquematico-2021%20(1).webp

[17] Metro de Medellín, “Portal de Negocios - Metro de Medellín”, 2025. [En línea]. Disponible en: https://negocios.metrodemedellin.gov.co/

[18] R. J. Hyndman y G. Athanasopoulos, *Forecasting: Principles and Practice*, 2a ed. OTexts, 2018. [En línea]. Disponible en: https://otexts.com/fpp2/components.html