UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INDUSTRIAS
SANTIAGO - CHILE

Predicción de la Demanda de Pasajeros en el Transporte Terrestre mediante Modelos de Aprendizaje Automático y Series Temporales Aplicado al Caso de Estudio de Cruceros del Norte

José Tomás Donoso Concha

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL INDUSTRIAL

PROFESOR GUÍA : SR. Eloy Alvarado Narváez

Marzo 2025

CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITORIO ACADÉMICO

1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

Tipo de monografía (marcar una opción): ☒ Memoria o trabajo de título; ☐ Tesis de Postgrado;

Título del trabajo: Predicción de la Demanda de Pasajeros en el Transporte Terrestre mediante Modelos de Aprendizaje Automático y Series Temporales Aplicado al Caso de Estudio de Cruceros del Norte

Nombre del candidato(a): José Tomás Donoso Concha

Carrera / Grado: ingeniería Civil Industrial

Campus: Santiago Vitacura **;** **Departamento**: Industria

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, Eloy Alvarado Narváez, en mi calidad de profesor(a) guía/director(a) del trabajo académico mencionado anteriormente **DEJO CONSTANCIA** que:

- He revisado esta versión del documento y corresponde a la versión final aprobada del trabajo.
- El trabajo cumple con los requisitos académicos y de formato establecidos por la institución

3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL

☒ El trabajo **NO contiene información que amerite confidencialidad** y puede ser publicado de inmediato en repositorio con acceso abierto.

☐ El trabajo **CONTIENE** información con potenciales implicancias de propiedad industrial o intelectual y requiere un periodo de confidencialidad (embargo) por:

☐ 6 meses; ☐ 12 meses; ☐ 2 años; ☐ 3 años; ☐ 5 años; ☐ 10 años

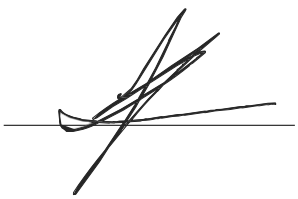
Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):

4.- FIRMAS

Profesor(a) guía o director(a) de memoria o tesis:

Fecha: 30/07/2025 ; **Firma:** 

Estudiante o Candidato(a):

Fecha: 30/07/2025 ; **Firma:** 

Este formulario debe ser insertado como página 2 de la memoria o tesis, completado y firmado por estudiante y profesor(a) antes de la entrega en portal PRISMA de Biblioteca USM.

A quienes me apoyaron ...

En el camino de los perros, mi alma encontró a mi corazón. Destrozado, pero vivo, sucio, mal vestido y lleno de amor.

.

AGRADECIMIENTOS

En esta sección, deseo expresar mi más sincero agradecimiento a todas aquellas personas que me han acompañado en este camino, brindándome su apoyo y conocimientos, ayudándome día a día a crecer tanto personal como profesionalmente. Gracias por su tiempo y preocupación por mí.

En primer lugar, agradezco a mi madre, Constanza Concha, por su paciencia y amor incondicional, por brindarme todas las herramientas necesarias para completar esta etapa de mi vida y por estar siempre a mi lado apoyándome en cada paso.

A mi pareja, Javiera Concha, por motivarme día a día a seguir adelante con mis objetivos, por ser un pilar fundamental en mi vida y por brindarme su compañía y apoyo incondicional.

A mi familia, en especial a Pedro Farías, por su respaldo inquebrantable y su cariño, sin importar las circunstancias. Gracias por recorrer conmigo este camino y estar presente en cada momento.

Asimismo, agradezco al profesor Eloy Alvarado por su guía constante durante todo el proceso de la tesis y por su dedicación y voluntad para enseñar. Su apoyo fue fundamental para la culminación de este trabajo.

Por último, quiero agradecer a Advitair por proporcionarme los datos necesarios para completar esta investigación, así como por toda la experiencia y aprendizaje que me brindaron durante un año de trabajo junto a ellos. Un agradecimiento especial a Lorena y José Luis por su apoyo y colaboración.

Finalmente, extiendo mi gratitud a todas aquellas personas que, de una u otra manera, contribuyeron al desarrollo de esta investigación y al logro de este objetivo.

[José Donoso]

RESUMEN EJECUTIVO

El presente estudio aborda la problemática de la predicción de demanda de pasajeros en el transporte terrestre de Argentina, con un enfoque específico en la empresa **Cruceros del Norte**. La alta volatilidad económica del país afecta la precisión en la planificación operativa, dificultando la optimización de rutas y la asignación eficiente de recursos.

Para mitigar estos desafíos, se ha desarrollado un modelo predictivo basado en técnicas de **machine learning** y **análisis de series de tiempo**, incorporando modelos estadísticos tradicionales (**SARIMA** y **SARIMAX**) y redes neuronales recurrentes (**LSTM** y **GRU**). Además, se ha implementado un enfoque basado en modelos de ensamble, incluyendo **XGBoost**, **Random Forest** y **Gradient Boosting**, con el fin de capturar patrones complejos en los datos y mejorar la precisión de las predicciones.

Se ha utilizado el **Índice de Precios al Consumidor (IPC)** segmentado en los sectores de transporte, restaurantes y hoteles, y salud, junto con **clustering** y análisis de correlación de variables, para enriquecer los modelos y evaluar la influencia de factores macroeconómicos en la demanda de pasajeros.

Los resultados indican que la incorporación de variables macroeconómicas mejora significativamente la capacidad de los modelos para anticipar la demanda en escenarios de alta incertidumbre. Mientras que los modelos basados en redes neuronales (**LSTM** y **GRU**) destacan en la captura de relaciones no lineales, el modelo **XGBoost** mostró un desempeño robusto en términos de precisión y generalización, superando a enfoques tradicionales en varios escenarios.

A partir de estos hallazgos, se presentan recomendaciones operativas para **Cruceros del Norte**, incluyendo estrategias para la planificación de rutas, ajuste dinámico de tarifas y optimización del uso de la flota en función de la demanda proyectada. Finalmente, se propone un esquema de actualización periódica del modelo para adaptarse a los cambios en la economía y en el comportamiento de los pasajeros.

Palabras clave: Predicción de demanda, series de tiempo, machine learning, SARIMA, LSTM, GRU, XGBoost, Random Forest, Gradient Boosting, IPC, transporte terrestre.

Índice de Contenidos

Índice de Tablas

Índice de Figuras

1 | Problema de la Investigación

El transporte de pasajeros en Argentina enfrenta desafíos significativos debido a la alta volatilidad económica del país. Factores como la inflación, el desempleo y las variaciones en el ingreso afectan directamente la demanda del servicio de transporte terrestre. Empresas como **Cruceros del Norte**, una de las principales operadoras de transporte interurbano, deben adaptarse constantemente a estas fluctuaciones para optimizar la asignación de recursos y mejorar la planificación operativa.

En este contexto, la predicción de la demanda de pasajeros se convierte en un aspecto clave para mejorar la eficiencia del servicio. Sin embargo, los métodos tradicionales de estimación presentan limitaciones en escenarios de alta variabilidad económica, lo que dificulta la toma de decisiones estratégicas. Se requiere, por lo tanto, el desarrollo de modelos predictivos avanzados que permitan anticipar la demanda con mayor precisión y robustez.

Una predicción inexacta de la demanda de pasajeros puede generar diversos problemas operativos y financieros, tales como sobreoferta de flota, lo que incrementa los costos operativos sin una ocupación óptima; sub oferta de flota, que puede generar pérdida de ingresos y afectar la calidad del servicio; ineficiencia en la asignación de recursos, dificultando la planificación de rutas y la distribución de horarios; e impacto en la rentabilidad, reduciendo márgenes de ganancia y aumentando costos innecesarios.

Dado que la demanda de pasajeros está influenciada por factores macroeconómicos, este estudio integra el **Índice de Precios al Consumidor (IPC)**, considerando los sectores de transporte, restaurantes y hoteles, y salud, para evaluar su impacto en la planificación operativa de **Cruceros del Norte**. Asimismo, se aplican técnicas de **clustering** y análisis de correlación de variables para mejorar la segmentación de datos y seleccionar las variables más influyentes.

Actualmente, los métodos de predicción de demanda utilizados en la industria del transporte terrestre en Argentina no logran capturar con precisión la influencia de las condiciones económicas cambiantes. La empresa **Cruceros del Norte** enfrenta dificultades para prever la cantidad de pasajeros en sus distintas rutas, lo que afecta directamente la eficiencia operativa y la optimización de costos.

Para abordar esta problemática, se propone un enfoque basado en modelos de **series de tiempo** y **machine learning**, utilizando distintas metodologías. Dentro de los modelos estadísticos, se incluyen **SARIMA**, que permite capturar estacionalidad y tendencias en la demanda de pasajeros; **SARIMAX**, que es una extensión de SARIMA que incorpora variables macroeconómicas como el IPC; y la **transformación Box-Cox**, utilizada para estabilizar la varianza de la serie temporal.

Adicionalmente, se consideran modelos de redes neuronales recurrentes como **LSTM (Long Short-Term Memory)**, capaz de capturar patrones de largo plazo en datos secuenciales, y **GRU (Gated Recurrent Units)**, una variante más eficiente de LSTM con menor costo computacional. También se aplican técnicas de **clustering** para segmentar rutas con patrones similares de demanda y **análisis de correlación** para seleccionar las variables clave en la predicción.

Además de estos modelos, se implementan enfoques de **aprendizaje supervisado basado en árboles de decisión**, incluyendo **XGBoost**, **Random Forest** y **Gradient Boosting**, con el objetivo de evaluar su capacidad predictiva en comparación con los modelos estadísticos y de redes neuronales.

Este estudio tiene como objetivo determinar cuál de estos enfoques logra la mejor precisión en la predicción de la demanda de pasajeros y cómo se pueden utilizar los resultados para hacer recomendaciones **de forma ingenieril**. A través de la evaluación de los modelos mencionados, se busca no solo mejorar la exactitud de las predicciones, sino también proporcionar estrategias optimizadas para la planificación operativa y la gestión eficiente de la flota de **Cruceros del Norte**.

2 | Objetivos

Objetivo General

Desarrollar un modelo predictivo para anticipar la demanda de pasajeros en **Cruceros del Norte**, utilizando técnicas de **machine learning** y **análisis de series de tiempo**, con el objetivo de optimizar la eficiencia en la planificación de rutas y la gestión de recursos.

Objetivos Específicos

Para alcanzar el objetivo general, se establecen los siguientes objetivos específicos:

- Recolectar y procesar datos históricos (2022-2024) sobre la demanda de pasajeros de **Cruceros del Norte**, utilizando técnicas avanzadas de análisis de datos.
- Desarrollar y entrenar modelos predictivos empleando técnicas de **machine learning**, **series de tiempo** y **aprendizaje supervisado basado en árboles de decisión** (incluyendo **XGBoost**, **Random Forest** y **Gradient Boosting**), que permitan estimar de manera precisa la demanda futura de pasajeros.
- Evaluar la influencia de variables macroeconómicas, como la inflación, el desempleo y los niveles de ingreso, en la predicción de la demanda de transporte.
- Validar y ajustar los modelos desarrollados mediante técnicas de validación cruzada y análisis estadístico, garantizando su robustez y capacidad de adaptación a diferentes escenarios económicos.
- Comparar el desempeño de los modelos estadísticos (**SARIMA, SARIMAX**), de redes neuronales (**LSTM, GRU**)

y de ensamble (**XGBoost, Random Forest, Gradient Boosting**) para determinar cuál presenta la mejor precisión en la predicción de la demanda de pasajeros.

- Proporcionar recomendaciones basadas en los resultados obtenidos, que permitan optimizar la asignación de recursos y mejorar la toma de decisiones estratégicas en **Cruceros del Norte**.

3 | Marco Teórico

3.1. Ocupación

3.1.1. Definicion de Ocupacion

La ocupación en el contexto del transporte de pasajeros se refiere al porcentaje de asientos ocupados en un viaje en relación con la capacidad total del vehículo. Se calcula como la relación entre el número de pasajeros transportados y el número total de asientos disponibles, expresada en un valor que oscila entre 0 y 1. En ciertas circunstancias, este valor puede superar 1, como cuando se permite que pasajeros viajen de pie en trayectos cortos o cuando se transportan niños que no requieren asiento propio.

3.1.2. Indicadores Clave de Desempeño (KPI) relacionados con la Ocupación

La medición de la ocupación es fundamental para evaluar la eficiencia operativa y la rentabilidad de las rutas de transporte. Entre los principales indicadores clave de desempeño (KPI) derivados de la ocupación se encuentran:

- Asientos por Kilómetro Ofertados (ASK, por sus siglas en inglés): Representa la capacidad total de asientos disponibles multiplicada por la distancia recorrida. Es una medida de la oferta de una empresa de transporte y se utiliza para evaluar su capacidad para generar ingresos potenciales. **aviacionline**
- Pasajeros por Kilómetro Transportados (RPK, por sus siglas en inglés): Calcula el número de pasajeros de pago transportados multiplicado por la distancia recorrida. Este indicador refleja la demanda real de servicios de transporte y es esencial para analizar el volumen de tráfico de pasajeros. **iata**
- Ingreso por Pasajero-Kilómetro (Yield): Mide el ingreso promedio generado por pasajero por kilómetro recorrido.

Se obtiene dividiendo los ingresos totales de pasajeros por los RPK. Este KPI es crucial para evaluar la eficiencia en la generación de ingresos por parte de la empresa de transporte. **airtransportmanagement**

- Ingreso por Asiento-Kilómetro Ofertado (RASK, por sus siglas en inglés): Se calcula dividiendo los ingresos operativos totales por los ASK. Este indicador proporciona una visión integral de los ingresos generados por unidad de capacidad y es útil para comparar la eficiencia entre diferentes rutas o períodos. **airtransportmanagement**

Estos KPI son esenciales para determinar la rentabilidad y eficiencia de cada ruta en función de los kilómetros recorridos. Al analizar estos indicadores, las empresas de transporte pueden identificar oportunidades de mejora en la asignación de recursos, optimización de rutas y estrategias de precios, lo que contribuye a una gestión más efectiva y rentable.

3.2. Dinámica del Transporte de Pasajeros en Argentina

3.2.1. Importancia del Transporte Terrestre en Argentina

El transporte terrestre desempeña un papel fundamental en Argentina, ya que constituye la principal opción de movilidad para viajes nacionales y trayectos internacionales de corta distancia **indec2024**. A diferencia del transporte aéreo, los viajes en ómnibus de larga distancia son preferidos por una gran parte de la población debido a su mayor accesibilidad, costos relativamente más bajos y una extensa red de conexiones entre ciudades **cepal2023**.

En particular, el transporte en autobús es ampliamente utilizado no solo para desplazamientos dentro del territorio argentino, sino también para rutas internacionales hacia países vecinos como Paraguay, Chile, Brasil y Uruguay **cepal2023**. Numerosas empresas de autobuses operan servicios frecuentes a ciudades paraguayas debido a la demanda significativa de pasajeros que viajan por turismo, trabajo o motivos familiares.

Este contexto ha llevado a la existencia de un mercado altamente competitivo, con una gran variedad de empresas de transporte terrestre que buscan suplir la demanda creciente. La competencia en el sector genera una constante optimización de rutas, precios dinámicos y mejoras en la calidad del servicio, lo que hace que la predicción de la demanda sea un aspecto clave para la eficiencia operativa de estas compañías **cao2023**.

3.2.2. Variables Macroeconómicas y el Transporte Terrestre

La economía argentina está caracterizada por una constante volatilidad, lo que genera un impacto directo en la previsión de la ocupación en el transporte terrestre **indec2024**; **imf2024**. Si bien este medio de transporte sigue siendo utilizado de manera regular, las condiciones económicas determinan en gran medida las decisiones de viaje de los pasajeros. Factores como la inflación, las variaciones en el tipo de cambio y los costos de insumos como la gasolina afectan significativamente la accesibilidad y la frecuencia con la que las personas utilizan el transporte terrestre **worldbank2024**.

En Argentina, se pueden identificar ciertos patrones de temporada alta y baja, con aumentos en la demanda de viajes durante los meses de verano e invierno, coincidiendo con los períodos de vacaciones **worldbank2024**. Sin embargo, la fluctuación de la economía genera desviaciones considerables en la curva de ocupación a lo largo del año. La capacidad adquisitiva de los pasajeros, influenciada por la inflación y la devaluación monetaria, afecta directamente la disposición de las personas a viajar **imf2024**.

Dado este contexto, surge la necesidad de estudiar si estas variables macroeconómicas pueden proporcionar información útil para mejorar la predicción de la demanda de pasajeros **itf2019**. Incorporar factores económicos en los modelos predictivos puede permitir a las empresas de transporte terrestre ajustar estrategias operativas de manera más precisa, optimizando la asignación de recursos y la planificación de servicios.

3.2.3. Principales Variables Macroeconómicas Analizadas

Para evaluar la relación entre la economía y la ocupación en el transporte terrestre, en este estudio se consideran las siguientes variables macroeconómicas:

- **Índice de Precios al Consumidor (IPC):** Se analizarán distintas variantes del IPC, incluyendo:
  - **IPC General:** Representa la inflación total del país y mide el aumento de precios en todos los sectores.
  - **IPC Transporte:** Refleja la evolución de los costos del transporte público y privado, incluyendo combustibles y tarifas de buses.
  - **IPC Restaurantes y Hoteles:** Indica el nivel de gasto en turismo y ocio, sectores que influyen en la demanda de viajes interurbanos.
  - **IPC Salud:** Evalúa los costos en el sector de salud, que pueden incidir en el número de viajes de pasajeros por tratamientos médicos.

- **Tipo de Cambio:** La cotización del dólar afecta los costos operativos de las empresas de transporte, así como el poder adquisitivo de los pasajeros para realizar viajes.
- **Precio de la Gasolina:** El costo del combustible impacta directamente en la estructura tarifaria del transporte terrestre, afectando la rentabilidad de las empresas y la accesibilidad para los usuarios.

El análisis de la correlación entre estas variables y la ocupación permitirá evaluar si los modelos predictivos pueden mejorar su precisión al incorporar factores económicos dentro de su estructura. Comprender estas relaciones permitirá tomar decisiones estratégicas en la optimización de rutas, fijación de tarifas y planificación de recursos operativos.

3.3. Crucero del Norte y Advitair

3.3.1. Historia de Crucero del Norte

Crucero del Norte es una empresa de transporte con más de 70 años de experiencia en el sector. Fundada en la década de 1950 en Argentina por Don Demetrio Koropeski, comenzó como una pequeña compañía de colectivos que brindaba servicios de transporte local en la región norte del país. Con el tiempo, la empresa experimentó un crecimiento sostenido y expandió sus operaciones a nivel nacional e internacional, convirtiéndose en una de las principales compañías de transporte de Argentina.

A lo largo de los años, Crucero del Norte ha sido pionera en la adopción de tecnologías innovadoras para mejorar la experiencia de sus pasajeros. Entre sus avances destacan la implementación de autobuses equipados con aire acondicionado, sistemas de entretenimiento a bordo y mejoras en los estándares de seguridad. Asimismo, la empresa ha priorizado la seguridad en la carretera, introduciendo cámaras de vigilancia en sus vehículos y proporcionando capacitación continua a sus conductores.

Actualmente, Crucero del Norte es una de las compañías de transporte más grandes de Argentina, con una flota moderna y una amplia red de rutas que abarcan todo el país. Su enfoque en la innovación y la excelencia operativa la han convertido en un referente del transporte terrestre en la región.

3.3.2. Advitair y su Experiencia en Revenue Management

Advitair es una empresa especializada en **Revenue Management**, distribución y tecnología aplicada al sector de transporte y aerolíneas. Su equipo de profesionales cuenta con amplia experiencia en la gestión de tarifas, estrategias de precios dinámicos, análisis de datos y optimización de ingresos mediante herramientas avanzadas de tecnología.

Las principales áreas de especialización de Advitair incluyen:

- Gestión de Tarifas y Distribución: Implementación de estrategias de precios corporativos y privados, acuerdos de código compartido, y optimización de costos en sistemas GDS.
- *Revenue Management: Implementación de indicadores clave para la gestión de ingresos, control de disponibilidad y limpieza del inventario.
- Tecnología Aplicada al Transporte: Desarrollo de herramientas para la optimización del sistema de reservas, monitoreo de tarifas de la competencia y automatización de procesos internos.
- Análisis de Datos y Optimización de Rutas: Identificación de oportunidades de mejora en la operación mediante técnicas avanzadas de análisis y visualización de datos.

Advitair se especializa en brindar soluciones para maximizar la rentabilidad de las empresas de transporte, asegurando una gestión eficiente de los recursos y una mejor planificación de las estrategias comerciales.

3.3.3. Advitair y su Colaboración con Crucero del Norte

En el marco de su experiencia en el sector del transporte, Advitair colaboró con Crucero del Norte en un proyecto clave para la **optimización de rutas y gestión de la demanda**. El objetivo principal de esta alianza fue analizar los segmentos de viaje más rentables y estratégicos de la empresa, permitiendo una mejor asignación de recursos y una optimización en la oferta de servicios.

Previo a esta colaboración, Crucero del Norte basaba su estrategia operativa en suplir la demanda sin una segmentación específica, lo que generaba oportunidades de mejora en la planificación de sus recorridos. Gracias al trabajo conjunto con Advitair, se realizó un análisis detallado de los tramos operados, identificando cuáles eran los más relevantes en términos de rentabilidad y volumen de pasajeros.

El estudio permitió desarrollar estrategias para:

- Identificar los tramos de mayor impacto en la rentabilidad de la empresa.

- Optimizar la oferta de servicios en función de la demanda real.
- Implementar técnicas de Revenue Management para mejorar la eficiencia operativa.

3.4. Recursos Disponibles en la Empresa

A continuación, se describen los principales recursos tecnológicos utilizados en la gestión y análisis de datos dentro de Advitair y Crucero del Nort. Estas herramientas permiten una administración eficiente de la información, optimizando los procesos de toma de decisiones en la planificación y operación del transporte.

3.4.1. Google Cloud Platform (GCP)

Google Cloud Platform (GCP) es una plataforma de servicios en la nube que proporciona infraestructura, almacenamiento y herramientas avanzadas para la gestión y análisis de datos. Su escalabilidad y seguridad permiten manejar grandes volúmenes de información, facilitando la implementación de soluciones basadas en inteligencia de datos.

En este estudio, GCP actúa como la plataforma principal para almacenar, procesar y analizar los datos de ocupación y rendimiento de las rutas de transporte. Su capacidad de integración con otras herramientas de Google la convierte en una pieza clave dentro de la infraestructura tecnológica utilizada.

3.4.2. Google BigQuery

Google BigQuery es un sistema de almacenamiento y análisis de datos basado en la nube, diseñado para ejecutar consultas SQL sobre conjuntos de datos de gran tamaño en tiempo real. Como parte del ecosistema de GCP (3.4.1), permite realizar análisis rápidos sin necesidad de administrar servidores físicos.

En este estudio, BigQuery se utiliza para ejecutar consultas y extraer información clave sobre la ocupación de los buses y el comportamiento de la demanda en distintos tramos de viaje. Los datos almacenados en BigQuery son procesados y exportados a otras herramientas para su posterior visualización y análisis.

3.4.3. Google Sheets

Google Sheets es una hoja de cálculo en línea que permite la colaboración en tiempo real y la gestión de datos de manera dinámica. Es ampliamente utilizada para la organización y manipulación de información antes de su procesamiento en sistemas más avanzados.

En este estudio, Google Sheets cumple dos funciones principales:

- Después de procesar los datos en Google BigQuery (3.4.2), los resultados son exportados a Google Sheets para su validación y ajuste manual, incluyendo la incorporación de variables adicionales para el análisis.
- Una vez generadas las predicciones de demanda, estas se almacenan en Google Sheets en formato '.xlsx', permitiendo su integración con otras herramientas de análisis y visualización de datos.

3.4.4. Google Looker Studio

Google Looker Studio es una plataforma de visualización de datos que permite crear paneles interactivos y reportes dinámicos a partir de diversas fuentes de información, como Google BigQuery (3.4.2) y Google Sheets (3.4.3). Su flexibilidad y capacidad de integración la hacen ideal para representar datos de manera visual y comprensible.

En este estudio, Google Looker Studio se utiliza para desarrollar dashboards que facilitan el monitoreo y análisis de la ocupación de los buses. Gracias a su conectividad con GCP (3.4.1) y otras herramientas, proporciona una representación visual clara de los patrones de demanda y las tendencias del mercado.

3.4.5. Power BI

Power BI es una herramienta de análisis y visualización de datos desarrollada por Microsoft. Su capacidad para integrar diversas fuentes de datos permite crear reportes detallados y facilitar la toma de decisiones basada en información estructurada.

En este estudio, Crucero del Norte utilizó Power BI para:

- Analizar tendencias de ocupación en los diferentes tramos de viaje.
- Generar reportes detallados sobre el rendimiento financiero de cada ruta.
- Facilitar la descarga y procesamiento de datos históricos para su posterior análisis en combinación con las herramientas de Advitair.

La combinación de Google Cloud Platform y Power BI permitió una gestión eficiente de los datos, optimizando los procesos de análisis y toma de decisiones estratégicas para mejorar la planificación operativa de Crucero del Norte.

3.5. Series de Tiempo

3.5.1. Definición

Las series de tiempo son conjuntos de observaciones registradas secuencialmente en intervalos regulares de tiempo **hyndman2018**. Su principal utilidad radica en el análisis de patrones y tendencias dentro de un fenómeno específico, lo que permite realizar predicciones basadas en datos históricos.

En el contexto del análisis de transporte, el estudio de series de tiempo es una herramienta fundamental, ya que las variaciones en la ocupación de los viajes dependen de múltiples factores como la temporada del año, cambios en la economía, eventos sociales y variaciones en la demanda de pasajeros **itf2020**. Al modelar estos datos temporalmente, es posible identificar patrones de comportamiento y prever fluctuaciones en la ocupación de los buses, lo que resulta crucial para la planificación operativa y la toma de decisiones estratégicas **hyndman2018**.

3.5.2. Características Principales

El análisis de series de tiempo se basa en la descomposición de los datos en diferentes componentes que permiten entender su estructura y comportamiento. Los principales elementos que caracterizan una serie de tiempo son:

- **Tendencia**: Representa la dirección general en la que evolucionan los datos a lo largo del tiempo. En el transporte de pasajeros, una tendencia creciente puede indicar un aumento en la demanda de ciertos tramos, mientras que una tendencia decreciente puede reflejar una menor preferencia por una ruta específica.
- **Estacionalidad**: Corresponde a patrones repetitivos que ocurren en intervalos regulares. En el caso de la ocupación en buses, se pueden observar estacionalidades marcadas durante las vacaciones de verano e invierno, donde la demanda aumenta, y periodos de baja movilidad, como los meses escolares o fuera de temporada turística.
- **Ciclo**: Se refiere a fluctuaciones periódicas que ocurren en plazos más largos y que no necesariamente tienen una frecuencia fija. En el transporte terrestre, los ciclos pueden estar influenciados por cambios económicos, políticas gubernamentales o modificaciones en la infraestructura de rutas.

- **Ruido Aleatorio**: Representa la variabilidad impredecible dentro de los datos, causada por factores externos no modelados. En el caso de la ocupación de buses, puede estar relacionado con cambios climáticos, eventos sociales inesperados o interrupciones en el servicio**.
- **Autocorrelación**: Se refiere a la relación que existe entre valores pasados y futuros dentro de una serie de tiempo. En el transporte de pasajeros, esto significa que la ocupación de una ruta en un determinado día puede estar influenciada por los niveles de ocupación en días anteriores.

La correcta identificación de estos elementos en los datos históricos permite mejorar la precisión de los modelos predictivos utilizados para anticipar la ocupación en los distintos tramos de viaje.

3.5.3. Aplicación en el Contexto del Estudio

En este estudio, las series de tiempo se aplican para analizar la evolución de la ocupación de pasajeros en los servicios de Crucero del Norte. La información histórica de ocupación es clave para detectar tendencias y estacionalidades en los viajes, lo que permite desarrollar modelos que anticipen la demanda en distintos periodos del año.

El análisis de series de tiempo en este contexto tiene varias aplicaciones estratégicas:

- Optimización de la Asignación de Flota: Permite ajustar la cantidad de buses disponibles en función de la demanda proyectada, evitando sobreoferta o falta de capacidad en determinados tramos.
- Ajuste Dinámico de Tarifas: Identifica períodos de alta y baja demanda, lo que facilita la implementación de estrategias de precios dinámicos para maximizar la rentabilidad.
- Planificación Operativa y Logística: Ayuda a predecir fluctuaciones en la demanda con base en eventos estacionales o factores externos, permitiendo mejorar la programación de horarios y servicios adicionales.

Al incorporar técnicas de machine learning y modelos estadísticos de series de tiempo, se busca mejorar la precisión de las predicciones y proporcionar herramientas de análisis más robustas para la toma de decisiones dentro de la empresa. La combinación de estos métodos permite generar recomendaciones estratégicas basadas en datos, mejorando la eficiencia operativa de Crucero del Norte.

3.6. Modelos Clásicos de Predicción Numérica

3.6.1. Modelo SARIMA (Seasonal AutoRegressive Integrated Moving Average)

El modelo SARIMA es una extensión del modelo ARIMA que permite manejar series de tiempo con componentes estacionales. Es ampliamente utilizado en el análisis de datos temporales donde se observan patrones cíclicos recurrentes, como la ocupación de pasajeros en diferentes períodos del año.

Matemáticamente, un modelo SARIMA se define como:

$$\Phi_P(B^s)\phi_p(B)(1-B)^d(1-B^s)^D y_t = \Theta_Q(B^s)\theta_q(B)\varepsilon_t \tag{3.1}$$

Donde:

- B es el operador de rezago (lag operator).
- $\Phi_P(B^s)$ representa el polinomio autorregresivo estacional de orden P.
- $\phi_p(B)$ es el polinomio autorregresivo de orden p.
- $(1-B)^d$ es la diferenciación no estacional de orden d.
- $(1-B^s)^D$ es la diferenciación estacional de orden D, con s como la periodicidad de la estacionalidad.
- $\Theta_Q(B^s)$ es el polinomio de media móvil estacional de orden Q.
- $\theta_q(B)$ es el polinomio de media móvil de orden q.
- y_t es la serie de tiempo observada.
- ε_t es el término de error (ruido blanco).

El modelo SARIMA es especialmente útil en contextos donde las series de tiempo presentan patrones estacionales marcados, como la variación de la ocupación de buses a lo largo del año. En el caso de Crucero del Norte, puede ser aplicado para prever la demanda de transporte en función de ciclos anuales, identificando periodos de alta y baja afluencia de pasajeros.

3.6.2. Modelo SARIMAX (Seasonal AutoRegressive Integrated Moving Average with Exogenous Variables)

El modelo SARIMAX es una extensión del SARIMA, que además de considerar la estructura interna de la serie de tiempo, incorpora variables exógenas que pueden influir en la variable objetivo. Esto permite capturar no solo la tendencia y estacionalidad de los datos, sino también el impacto de factores externos en la predicción.

La ecuación general del modelo SARIMAX se expresa como:

$$\Phi_P(B^s)\phi_p(B)(1-B)^d(1-B^s)^D y_t = \Theta_Q(B^s)\theta_q(B)\varepsilon_t + \beta X_t \tag{3.2}$$

Donde, además de los términos ya definidos en SARIMA:

- X_t representa las variables exógenas que pueden afectar la serie de tiempo y_t.
- β es el vector de coeficientes que mide la relación entre las variables exógenas y la variable objetivo.

En el contexto del transporte terrestre, el modelo SARIMAX es particularmente útil ya que permite incorporar factores económicos y operacionales que afectan la demanda de pasajeros. Algunas variables exógenas relevantes para este estudio incluyen:

- Índice de Precios al Consumidor (IPC): Refleja la inflación y el costo de vida, lo que puede influir en la capacidad de los pasajeros para costear viajes de larga distancia.
- Precio del Combustible: Impacta directamente en los costos operativos de la empresa y en la fijación de tarifas de los boletos.
- Tipo de Cambio: Puede influir en la competitividad del transporte terrestre frente a otras opciones, como los vuelos internacionales o el turismo en países vecinos.
- Días Feriados: Puede generar efectos dentro de la estacionalidad y periodos con una mayor demanda pasajes dentro de estas fechas.
- Costo/KM (Costo por Kilómetro Recorrido): Representa el costo promedio de operación de un bus por Kilómetro. Este valor afecta la estructura tarifaria de la empresa y la rentabilidad de cada ruta. Un aumento en el costo por Kilómetro recorrido puede reflejar mayores costos de operación, lo que puede traducirse en ajustes en las tarifas de los pasajeros y afectar la demanda.

La incorporación de estas variables permite mejorar la precisión de las predicciones y entender mejor las

fluctuaciones en la ocupación de pasajeros. En este estudio, se evalúa la efectividad del SARIMAX frente a otros modelos predictivos para determinar cuál es la mejor estrategia en términos de planificación y optimización de rutas en Crucero del Norte.

3.7. Modelos Modernos de Predicción Numérica: Machine Learning

3.7.1. Modelo LSTM (Long Short-Term Memory)

El modelo Long Short-Term Memory (LSTM) es una variante de las redes neuronales recurrentes (RNN) diseñada para manejar datos secuenciales y capturar dependencias a largo plazo **hochreiter1997**. Su arquitectura permite superar problemas comunes en las RNN tradicionales, como el desvanecimiento o explosión del gradiente, lo que facilita el aprendizaje de patrones temporales complejos en series de tiempo **goodfellow2016**.

3.7.1.1. Definición y Características

La principal innovación del modelo LSTM es la introducción de bloques de memoria que regulan la transmisión de información a lo largo de la secuencia temporal. Cada bloque LSTM contiene:

- Célula de estado: Actúa como una memoria a largo plazo, permitiendo que la red retenga información relevante a lo largo del tiempo.
- Compuerta de entrada: Determina qué información nueva se almacenará en la célula de estado.
- Compuerta de olvido: Decide qué información almacenada previamente debe ser descartada.
- Compuerta de salida: Regula qué información será utilizada en la predicción de la siguiente secuencia.

Estas compuertas permiten que la red LSTM mantenga información relevante mientras descarta datos innecesarios, lo que mejora la capacidad del modelo para aprender patrones de largo plazo en datos temporales.

3.7.1.2. Aplicación en Predicciones Temporales

El modelo LSTM es altamente efectivo en la predicción de series de tiempo con patrones no lineales y estacionales. En este estudio, se utiliza LSTM para predecir la ocupación de buses en distintos tramos de viaje, permitiendo ajustar la oferta de transporte según la demanda esperada. Su capacidad de aprender dependencias a largo plazo lo hace

especialmente útil en este contexto, ya que la demanda de transporte puede depender de múltiples factores históricos y estacionales.

3.7.2. Modelo GRU (Gated Recurrent Unit)

El modelo Gated Recurrent Unit (GRU) es una variante simplificada de LSTM que mantiene un rendimiento comparable con una menor complejidad computacional. Fue desarrollado para resolver los mismos problemas que afectan a las RNN tradicionales, pero con una estructura más eficiente.

3.7.2.1. Definición y Características

A diferencia de LSTM, las GRU combinan las compuertas de entrada y olvido en una sola compuerta de actualización, lo que reduce el número de parámetros del modelo. Sus principales componentes son:

- Compuerta de actualización: Regula cuánto de la información pasada se mantiene y cuánto se reemplaza con nueva información.
- Compuerta de reinicio: Permite que la red decida cuándo olvidar información antigua y cuándo reiniciar la memoria del modelo.

Debido a su menor carga computacional, GRU es una opción viable cuando se requiere entrenar modelos en grandes volúmenes de datos o con recursos limitados.

3.7.2.2. Aplicación en Predicciones Temporales

Las GRU se utilizan en este estudio como una alternativa más eficiente a LSTM para la predicción de la ocupación de buses en Crucero del Norte. Debido a su estructura optimizada, pueden capturar patrones temporales con un menor costo computacional, manteniendo una precisión similar a la de las LST en muchos casos.

3.7.3. Modelos Basados en Árboles de Decisión

Los modelos basados en árboles de decisión son ampliamente utilizados en machine learning debido a su capacidad de manejar relaciones no lineales entre variables y realizar predicciones precisas con datos tabulares **breiman2001**. En este estudio, se aplican los siguientes modelos:

3.7.4. Modelo Gradient Boosting

3.7.4.1. Definición

El modelo Gradient Boosting es un método de aprendizaje supervisado que combina múltiples árboles de decisión secuenciales, donde cada nuevo árbol intenta corregir los errores del modelo anterior. Se basa en la minimización del error de predicción utilizando descenso de gradiente, permitiendo mejorar progresivamente la precisión del modelo. Este enfoque es particularmente efectivo para capturar relaciones no lineales y complejas en los datos, lo que lo hace adecuado para una amplia gama de problemas de regresión y clasificación, como la predicción de la demanda de pasajeros en series temporales.

3.7.4.2. Características

Las principales características del modelo Gradient Boosting incluyen:

- **Aprendizaje iterativo**: Los árboles se construyen de forma secuencial, corrigiendo los errores de los modelos previos en cada iteración, lo que permite un aprendizaje progresivo.
- **Optimización por gradiente**: Se minimiza una función de pérdida diferenciable ajustando los pesos de los árboles en cada iteración, lo que permite una convergencia más rápida y precisa hacia el modelo óptimo.
- **Capacidad de modelar relaciones no lineales**: Puede capturar patrones complejos en los datos de entrada, lo que es especialmente útil en problemas con interacciones entre variables, como datos históricos de ocupación.
- **Personalización de hiperparámetros**: Permite ajustar parámetros como la profundidad de los árboles, la tasa de aprendizaje (ν, típicamente entre 0 y 1, con valores comunes como 0.1 para mejor generalización) y el número de iteraciones (M) para optimizar el rendimiento y evitar el sobreajuste.

3.7.4.3. Formulación Matemática

El algoritmo de Gradient Boosting se basa en la minimización de una función de pérdida mediante la adición iterativa de modelos débiles, típicamente árboles de decisión. El objetivo es encontrar una función $\hat{F}(x)$ que minimice la pérdida esperada:

$$\hat{F} = \arg\min_{F} \mathbb{E}_{x,y}[L(y, F(x))]$$

donde $L(y, F(x))$ es una función de pérdida diferenciable, como la pérdida cuadrática para regresión ($L(y, F(x)) = (y - F(x))^2$) o la pérdida logística para clasificación.

El modelo se construye de la siguiente manera:

1. **Inicialización**: Se inicia con un modelo constante $F_0(x)$, que es el valor que minimiza la pérdida total:

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma)$$

Para la regresión con pérdida cuadrática, esto es simplemente la media de los valores de y, es decir, $F_0(x) = \text{mean}(y)$.

2. **Iteración**: Para cada iteración $m = 1, 2, \ldots, M$:

a) Calcular los pseudo-residuales r_{im}, que son las derivadas negativas de la pérdida con respecto al modelo actual:

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}$$

Para pérdida cuadrática, esto simplifica a $r_{im} = y_i - F_{m-1}(x_i)$, que son los residuales.

b) Ajustar un modelo débil $h_m(x)$ (por ejemplo, un árbol de decisión) a los pseudo-residuales, utilizando el conjunto de entrenamiento $\{(x_i, r_{im})\}_{i=1}^{n}$.

c) Encontrar el valor óptimo γ_m que minimiza la pérdida, calculado como:

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma h_m(x_i))$$

Para árboles de decisión, γ_{jm} (el valor óptimo para cada hoja j) se calcula como el promedio de los residuales en la región R_{jm}, con n_j como el número de muestras en la hoja.

d) Actualizar el modelo:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x)$$

3. **Regularización**: Para evitar el sobreajuste, se pueden usar técnicas como el encogimiento (shrinkage), donde se introduce un factor de aprendizaje ν ($0 < \nu \leq 1$, típicamente 0.1 para mejor generalización):

$$F_m(x) = F_{m-1}(x) + \nu \cdot \gamma_m h_m(x)$$

Además, se puede usar gradient boosting estocástico, donde en cada iteración se utiliza una fracción de los datos (por ejemplo, $f = 0{,}5$ para datasets pequeños a moderados).

En el caso de que los modelos débiles sean árboles de decisión, la actualización del modelo se puede expresar como:

$$F_m(x) = F_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} \mathbf{1}_{R_{jm}}(x)$$

donde R_{jm} son las regiones definidas por el árbol m, y γ_{jm} son los valores óptimos para cada región. El tamaño típico de los árboles (J) para boosting es de 4 a 8 hojas, con $J = 2$ siendo insuficiente y $J > 10$ raramente necesario.

3.7.4.4. Aplicación en Predicciones Temporales

En este estudio, Gradient Boosting se utiliza como complemento de XGBoost y Random Forest, permitiendo evaluar cuál de estos enfoques ofrece el mejor desempeño en la predicción de la ocupación de pasajeros en los servicios de Crucero del Norte. Se compara su capacidad de ajuste, eficiencia computacional y precisión en la identificación de patrones en los datos históricos. Dado que los datos de series temporales a menudo presentan relaciones no lineales e interacciones complejas, Gradient Boosting es particularmente adecuado para capturar estas características, lo que lo convierte en una herramienta poderosa para la predicción de la demanda.

3.7.5. Modelo XGBoost (Extreme Gradient Boosting)

3.7.5.1. Definición

El modelo XGBoost (Extreme Gradient Boosting) es una implementación optimizada del algoritmo de Gradient Boosting, diseñado para mejorar la velocidad y precisión en tareas de predicción. Este modelo se basa en la construcción secuencial de árboles de decisión, donde cada nuevo árbol corrige los errores cometidos por los anteriores, permitiendo generar modelos altamente precisos y eficientes.

3.7.5.2. Características

Las principales características del modelo XGBoost incluyen:

- Regularización avanzada: Incorpora términos de regularización L1 y L2 para evitar el sobreajuste.
- Optimización basada en gradientes: Cada árbol es entrenado para minimizar la diferencia entre la predicción y el valor real.
- Capacidad de manejar datos faltantes: Permite realizar predicciones incluso cuando hay valores ausentes en las variables explicativas.
- Computación paralela: Puede ejecutar múltiples cálculos en paralelo, acelerando el entrenamiento del modelo.
- Flexibilidad en la configuración de hiperparámetros: Permite ajustar criterios como la profundidad de los árboles y la tasa de aprendizaje para optimizar el desempeño.

3.7.5.3. Aplicación en Predicciones Temporales

En este estudio, XGBoost es utilizado para predecir la ocupación de buses en diferentes rutas de Crucero del Norte. Se emplea para identificar patrones en los datos históricos y ajustar modelos que consideren estacionalidad, variables exógenas y tendencias no lineales. Gracias a su capacidad de manejo de grandes volúmenes de datos, XGBoost permite generar predicciones de alta precisión y mejorar la planificación operativa del servicio.

3.7.6. Modelo Random Forest

3.7.6.1. Definición

El modelo Random Forest es un algoritmo de aprendizaje supervisado basado en la combinación de múltiples árboles de decisión. Se fundamenta en la idea de entrenar varios árboles en subconjuntos aleatorios de los datos y promediar sus predicciones para mejorar la precisión y reducir la varianza.

3.7.6.2. Características

Las principales características del modelo Random Forest incluyen:

- Uso de múltiples árboles de decisión: Cada árbol es entrenado con un subconjunto diferente del conjunto de datos, lo que mejora la robustez del modelo.

- Reducción del sobreajuste: Debido a la agregación de múltiples árboles, el modelo es menos propenso a ajustar el ruido presente en los datos.
- Capacidad de manejar datos de alta dimensionalidad Puede procesar conjuntos de datos con muchas variables sin que el rendimiento del modelo se vea comprometido.
- Interpretabilidad: Permite calcular la importancia de cada variable en la predicción, facilitando la toma de decisiones basadas en datos.

3.7.6.3. Aplicación en Predicciones Temporales

En este estudio, Random Forest se emplea como una alternativa para modelar la ocupación de buses en distintas rutas, permitiendo evaluar la importancia relativa de diferentes variables en la predicción. Se utiliza para comparar su desempeño con otros modelos y entender mejor qué factores influyen en la demanda de transporte.

3.8. Justificación de los Modelos Seleccionados

La predicción de la demanda de pasajeros en el transporte terrestre requiere modelos capaces de capturar patrones temporales, estacionalidad y relaciones no lineales en un contexto de alta volatilidad económica, como el de Argentina. En este estudio, se seleccionaron siete modelos predictivos: SARIMA, SARIMAX, LSTM, GRU, XGBoost, Gradient Boosting y Random Forest. Esta selección se fundamenta en sus características teóricas, su aplicabilidad a series temporales y su capacidad para abordar los desafíos específicos del caso de estudio. A continuación, se justifica la elección de cada modelo, destacando sus fortalezas, limitaciones y casos de uso relevantes, con base en la literatura especializada.

SARIMA y SARIMAX: Los modelos de media móvil autorregresiva integrada estacional (SARIMA) son ampliamente utilizados en la predicción de series temporales debido a su capacidad para modelar tendencias y estacionalidad (Box et al., 2015). SARIMA es particularmente adecuado para series con patrones estacionales claros, como la demanda de transporte, que puede variar según temporadas turísticas o feriados. Sin embargo, su enfoque lineal limita su capacidad para capturar relaciones no lineales o cambios abruptos en la dinámica temporal, comunes en contextos económicos volátiles. SARIMAX, una extensión de SARIMA, incorpora variables exógenas, como el Índice de Precios al Consumidor (IPC), permitiendo modelar la influencia de factores externos (Hyndman Athanasopoulos, 2018). Estos modelos fueron seleccionados por su simplicidad, interpretabilidad y su uso frecuente como línea base en estudios de series temporales, aunque se espera que su rendimiento sea superado por enfoques no lineales en escenarios

complejos.

LSTM y GRU: Las redes neuronales recurrentes (RNN), como las unidades de memoria a corto y largo plazo (LSTM) y las unidades recurrentes con compuertas (GRU), son ideales para modelar dependencias temporales largas en datos secuenciales (Hochreiter Schmidhuber, 1997; Cho et al., 2014). LSTM es particularmente eficaz para capturar patrones no lineales y relaciones temporales complejas, lo que lo hace adecuado para la demanda de transporte, influenciada por factores económicos y estacionales. Sin embargo, requiere grandes volúmenes de datos y un entrenamiento computacionalmente intensivo. GRU, una variante más eficiente de LSTM, reduce el costo computacional al simplificar la arquitectura, manteniendo una capacidad comparable para modelar series temporales (Chung et al., 2014). Ambos modelos fueron seleccionados por su potencial para superar las limitaciones de los enfoques lineales, aunque su rendimiento puede verse afectado por la irregularidad estructural de los datos en Argentina.

XGBoost, Gradient Boosting y Random Forest: Los modelos de ensamble basados en árboles de decisión, como XGBoost, Gradient Boosting y Random Forest, son robustos para datos ruidosos y no lineales, características comunes en la demanda de transporte en contextos económicos volátiles (Breiman, 2001; Chen Guestrin, 2016; Friedman, 2001). XGBoost destaca por su capacidad para manejar relaciones complejas y su eficiencia computacional, optimizando iterativamente los errores residuales mediante el uso de gradientes (Chen Guestrin, 2016). Gradient Boosting, similar a XGBoost, ofrece alta precisión y robustez frente a valores atípicos, aunque requiere un ajuste cuidadoso de hiperparámetros. Random Forest, por su parte, es estable frente a ruido y fácil de implementar, aunque tiende a ser menos preciso que otros ensambles en problemas complejos (Breiman, 2001). Estos modelos fueron seleccionados por su capacidad para capturar no linealidades y generalizar en escenarios con alta variabilidad, como el caso de Cruceros del Norte.

Comparación con alternativas: Otros enfoques, como el modelo Prophet (Taylor Letham, 2018) o los Transformers temporales (Lim et al., 2021), no fueron implementados en este estudio. Prophet, diseñado para series temporales con estacionalidad y eventos específicos, es menos flexible para modelar relaciones no lineales complejas en comparación con los modelos de ensamble. Los Transformers, aunque prometedores por su capacidad para capturar dependencias temporales mediante mecanismos de atención, requieren grandes volúmenes de datos y recursos computacionales significativos, lo que los hace inviables dado el alcance de esta investigación. La selección de los modelos propuestos equilibra precisión, interpretabilidad y viabilidad práctica, alineándose con las necesidades del caso de estudio.

La Tabla 3.1 resume las fortalezas, debilidades y casos de uso de los modelos seleccionados, proporcionando una visión comparativa que fundamenta su inclusión en este estudio. Esta combinación de modelos estadísticos, redes neuronales y ensambles permite abordar la complejidad de la predicción de demanda desde múltiples perspectivas, asegurando un análisis robusto y adaptable al contexto económico de Argentina.

Tabla 3.1: Comparación de Modelos Seleccionados para Predicción de Demanda

Modelo	Fortalezas	Debilidades	Casos de Uso
SARIMA	Captura estacionalidad y tendencias; fácil de interpretar	Limitado a relaciones lineales; sensible a datos ruidosos	Series con patrones estacionales claros (Box et al., 2015)
SARIMAX	Incorpora variables exógenas; robusto para tendencias	No captura no linealidades complejas	Series con factores externos conocidos (Hyndman & Athanasopoulos, 2018)
LSTM	Modela dependencias temporales largas; adecuado para no linealidades	Requiere grandes volúmenes de datos; alto costo computacional	Series complejas con patrones secuenciales (Hochreiter & Schmidhuber, 1997)
GRU	Similar a LSTM pero más eficiente computacionalmente	Menor capacidad para patrones muy complejos	Series temporales con datos limitados (Cho et al., 2014)
XGBoost	Robusto a datos ruidosos; captura no linealidades	Menos interpretable que modelos estadísticos	Predicción en entornos volátiles (Chen & Guestrin, 2016)
Gradient Boosting	Alta precisión; robusto a outliers	Requiere ajuste cuidadoso de hiperparámetros	Problemas con alta variabilidad (Friedman, 2001)
Random Forest	Estable frente a ruido; fácil de implementar	Menos preciso que otros ensambles	Predicción general con datos heterogéneos (Breiman, 2001)

3.9. Criterios de Evaluación de Modelos

Para evaluar el desempeño de los modelos utilizados en este estudio, se emplean métricas estadísticas que permiten cuantificar el error de predicción y la capacidad explicativa de cada modelo. En este estudio, se han utilizado tres criterios fundamentales: RMSE (Root Mean Square Error), MAE (Mean Absolute Error) y R^2 (Coeficiente de Determinación).

3.9.1. Root Mean Square Error (RMSE)

El RMSE (Root Mean Square Error) mide la magnitud promedio del error en las predicciones de un modelo. Es una métrica sensible a errores grandes debido a la penalización cuadrática de las diferencias entre valores reales y predichos.

Matemáticamente, se expresa como:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.3}$$

Donde:

- y_i es el valor real en la observación i.
- $\hat{y}_i$ es el valor predicho por el modelo en la observación i.
- n es el número total de observaciones.

Un valor más bajo de RMSE indica que el modelo tiene una mejor precisión en sus predicciones. Sin embargo, debido a que eleva al cuadrado los errores, puede sobre ponderar desviaciones grandes, lo que hace que esta métrica sea particularmente útil cuando se busca minimizar grandes errores.

3.9.2. Mean Absolute Error (MAE)

El MAE (Mean Absolute Error) mide el error promedio en términos absolutos, sin penalizar en exceso los errores grandes como lo hace el RMSE. Se considera una métrica más robusta en presencia de valores atípicos.

Su fórmula es la siguiente:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3.4}$$

Donde los términos son los mismos que en RMSE.

A diferencia del RMSE, el MAE mide la desviación media entre las predicciones y los valores reales en la misma unidad de medida, lo que lo hace más interpretable en algunos casos.

3.9.3. Coeficiente de Determinación (R^2)

El coeficiente de determinación R^2 mide la proporción de la varianza total de la variable dependiente que es explicada por el modelo. Su valor oscila entre 0 y 1, donde un valor cercano a 1 indica un modelo con alta capacidad explicativa, mientras que un valor cercano a 0 sugiere que el modelo tiene un bajo poder predictivo.

Se calcula mediante la siguiente expresión:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3.5}$$

Donde:

- y_i es el valor real en la observación i.

- $\hat{y}_i$ es el valor predicho en la observación i.
- $\bar{y}$ es la media de los valores reales.
- n es el número total de observaciones.

Un valor alto de R^2 indica que el modelo logra explicar una mayor proporción de la variabilidad en los datos, lo que sugiere un buen ajuste. Sin embargo, esta métrica puede verse afectada por la cantidad de variables incluidas en el modelo, por lo que debe interpretarse junto con otras métricas como RMSE y MAE.

3.9.4. Importancia de las Métricas en la Evaluación de Modelos

En este estudio, la combinación de RMSE, MAE y R^2 permite evaluar los modelos desde diferentes perspectivas:

- RMSE es útil cuando se desea penalizar errores grandes en la predicción.
- MAE proporciona una interpretación más clara del error medio en las mismas unidades de la variable de interés.
- R^2 permite determinar qué porcentaje de la variabilidad en los datos es explicado por el modelo.

Al analizar conjuntamente estas métricas, se busca seleccionar el modelo que logre el mejor equilibrio entre precisión, estabilidad y capacidad explicativa en la predicción de la ocupación de buses en los servicios de Crucero del Norte.

3.10. Recursos Computacionales

A lo largo de este estudio, se han utilizado diversos recursos computacionales que facilitan el procesamiento de datos y la implementación de modelos de predicción. Estas herramientas permiten manipular grandes volúmenes de información, desarrollar modelos de aprendizaje automático avanzados y analizar resultados de manera eficiente. A continuación, se presentan los principales recursos utilizados en este trabajo.

3.10.1. Python

Python es un lenguaje de programación de alto nivel ampliamente utilizado en ciencia de datos, estadística y aprendizaje automático. Su sintaxis clara y su ecosistema de bibliotecas lo convierten en una herramienta ideal para el desarrollo de modelos predictivos y análisis de datos.

En este estudio, Python ha sido utilizado para la exploración de datos, la implementación de los modelos de predicción de series temporales y la evaluación de métricas de desempeño. Se emplearon diversas librerías especializadas para facilitar la manipulación de datos, la construcción de modelos y la visualización de resultados.

Las principales librerías utilizadas en este estudio incluyen:

- NumPy: Para la manipulación de matrices y operaciones matemáticas de alto rendimiento.
- Pandas: Para el procesamiento, análisis y estructuración de grandes volúmenes de datos.
- Matplotlib y Seaborn: Para la visualización de datos, gráficos de tendencia y correlación.
- Scikit-learn: Para la evaluación de modelos, cálculo de métricas y preprocesamiento de datos.
- Statsmodels: Para el ajuste de modelos estadísticos y pruebas de significancia.
- TensorFlow y Keras: Para la implementación de redes neuronales recurrentes y modelos de aprendizaje profundo.

Paquete	Propósito	Modelos Relacionados
pandas	Manipulación de datos históricos	Todos los modelos
numpy	Cálculos numéricos en preprocesamiento	Todos los modelos
statsmodels	Implementación de modelos SARIMA y SARIMAX	SARIMA, SARIMAX
scikit-learn	Implementación de Gradient Boosting y Random Forest	Gradient Boosting, Random Forest
tensorflow	Implementación de modelos de deep learning (LSTM, GRU)	LSTM, GRU
keras	API de alto nivel para TensorFlow, utilizado para construir y entrenar redes neuronales	LSTM, GRU
matplotlib	Visualización de resultados y predicciones	Todos los modelos

Tabla 3.2: Tabla de paquetes de Python utilizados en la implementación de los modelos

3.10.2. Jupyter Notebook

Jupyter Notebook es un entorno de desarrollo interactivo que permite ejecutar código Python de manera modular, facilitando la experimentación con modelos de predicción y el análisis de resultados. Su integración con bibliotecas de ciencia de datos lo convierte en una herramienta ideal para la documentación y visualización de procesos.

En este estudio, Jupyter Notebook ha sido empleado como la plataforma principal para desarrollar los modelos de predicción, permitiendo:

- Ejecutar bloques de código de manera secuencial, facilitando la depuración y análisis iterativo.
- Visualizar gráficos y estadísticas en tiempo real.

- Documentar los procesos de análisis y optimización de modelos.

3.10.3. TensorFlow

TensorFlow es una biblioteca de código abierto desarrollada por Google para el aprendizaje automático y el procesamiento de grandes volúmenes de datos. Su arquitectura permite entrenar modelos de redes neuronales de manera eficiente, optimizando el uso de CPU y GPU para mejorar el rendimiento computacional.

En este estudio, TensorFlow ha sido utilizado para la construcción y ajuste de modelos de predicción de series temporales basados en redes neuronales recurrentes, en particular LSTM y GRU. Su capacidad para manejar grandes conjuntos de datos ha sido clave en la implementación de los modelos.

3.10.4. Importancia de los Recursos Computacionales en el Estudio

El uso de Python, Jupyter Notebook, TensorFlow y Keras, junto con bibliotecas especializadas, ha sido fundamental en la implementación de los modelos predictivos de este estudio. Estas herramientas han permitido:

- Automatizar la carga, limpieza y transformación de los datos históricos de ocupación de buses.
- Construir y evaluar modelos de aprendizaje profundo y machine learning.
- Visualizar patrones en las series de tiempo y ajustar los modelos para mejorar la precisión de predicción.
- Realizar análisis comparativos entre diferentes enfoques de modelado.

Gracias a estos recursos, ha sido posible desarrollar un enfoque de predicción robusto y eficiente para la optimización de rutas en Crucero del Norte.

3.11. Introducción a la Metodología

En esta investigación, se adoptó la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining), un marco ampliamente aceptado para proyectos de minería de datos, debido a su enfoque sistemático y su aplicabilidad a problemas prácticos como la predicción de demanda de pasajeros. CRISP-DM consta de seis fases que guían el proceso desde la comprensión del negocio hasta el despliegue de resultados, adaptándose al contexto de Cruceros del Norte y la volatilidad económica de Argentina **crispdm1999**.

3.11.1. Metodología CRISP-DM

3.11.1.1. Comprensión del Negocio

El objetivo principal fue desarrollar un sistema predictivo para anticipar la demanda de pasajeros en Cruceros del Norte, una empresa de transporte terrestre en Argentina, con el fin de optimizar la asignación de recursos, reducir costos operativos y mejorar la experiencia del cliente. Se identificaron desafíos clave, como la alta volatilidad económica (inflación, variaciones en el IPC) y la necesidad de adaptarse a patrones estacionales, lo que requirió un enfoque basado en datos históricos y variables exógenas.

3.11.1.2. Comprensión de los Datos

Se recopilaron datos históricos de ocupación de pasajeros desde 2022 hasta 2024, proporcionados por Advitair, junto con variables exógenas como el Índice de Precios al Consumidor (IPC General, Transporte, Restaurantes y Hoteles, Salud), Costo por Kilómetro, Feriados y Días Laborables. Se realizó un análisis exploratorio para identificar tendencias, estacionalidades y correlaciones (por ejemplo, $r = -0{,}31$ entre IPC Transporte y ocupación), utilizando herramientas como pandas y matplotlib.

3.11.1.3. Preparación de los Datos

Los datos se limpiaron para manejar valores faltantes (imputación con medias móviles) y se normalizaron para consistencia. Se crearon características adicionales, como rezagos temporales y variables dummy para feriados, y se dividieron en conjuntos de entrenamiento (2022-2023) y prueba (2024), asegurando la preservación del orden temporal para validar los modelos.

3.11.1.4. Modelado

Se implementaron diversos modelos: SARIMA y SARIMAX para capturar estacionalidad, LSTM y GRU para patrones no lineales a largo plazo, y Gradient Boosting, XGBoost y Random Forest para ensamblajes robustos. La selección se basó en su capacidad para manejar datos temporales y exógenos, utilizando bibliotecas como statsmodels, tensorflow/keras y scikit-learn/xgboost. Los hiperparámetros (por ejemplo, tasa de aprendizaje de 0.1 para Gradient Boosting) se optimizaron mediante validación cruzada.

3.11.1.5. Evaluación

La evaluación de los modelos se diseñó para validar su idoneidad mediante métricas cuantitativas como el Error Cuadrático Medio (RMSE), el Error Absoluto Medio (MAE) y el Coeficiente de Determinación (R^2). Se propuso una estrategia de validación cruzada temporal, dividiendo los datos en ventanas secuenciales (por ejemplo, entrenamiento con 2022-2023 y prueba con 2024) para preservar el orden cronológico y evaluar la generalización en un contexto de alta volatilidad económica. Se definieron criterios de comparación entre modelos para identificar el más adecuado según el desempeño predictivo.

3.11.1.6. Despliegue

Aunque el enfoque principal fue académico, se propuso un plan teórico de despliegue: integrar los modelos en un sistema basado en Google BigQuery para almacenamiento y Looker Studio para visualización, con actualizaciones semanales. Esto permitiría a Cruceros del Norte ajustar rutas y recursos en tiempo real, pendiente de una implementación piloto.

3.12. Datos Utilizados

El análisis de este estudio se basa en un conjunto de datos obtenido a partir de diversas fuentes, que permiten evaluar la ocupación de los servicios de transporte terrestre en Argentina y su relación con variables económicas relevantes. A continuación, se presentan las principales fuentes de información y su tratamiento.

3.12.1. Base de Datos de Crucero del Norte

Los datos históricos de ocupación y desempeño de los servicios fueron extraídos desde la base de datos interna de Crucero del Norte mediante Power BI, herramienta utilizada por la empresa para gestionar y analizar información operativa.

Las variables incluidas en este conjunto de datos corresponden a distintos aspectos del transporte terrestre, con registros detallados de cada servicio, permitiendo un análisis exhaustivo del comportamiento de la demanda. Estas variables incluyen:

- Servicio: Identificador del servicio de transporte.

- ID: Código único del viaje registrado.
- Mes y Año: Fecha en la que se realizó el servicio.
- Fecha de Partida y Fecha de Arribo: Registro temporal de salida y llegada del servicio.
- Ocupación (Ocup): Proporción de pasajeros transportados con respecto a las butacas disponibles.
- Costo por Kilómetro (Costo_KM): Costo operacional asociado a la distancia recorrida.
- Distancia Recorrida (KMS): Número de kilómetros recorridos en cada servicio.
- Tarifa Media de Venta (Tarifa Media ADV): Precio promedio de los pasajes vendidos en cada viaje.
- Yield: Ingreso por pasajero-kilómetro, un indicador clave para evaluar la rentabilidad de cada servicio.
- Tipo de Línea: Clasificación del servicio (Regular, Internacional, Especial, etc.).
- Pasajeros: Número total de pasajeros transportados en cada servicio.
- Butacas Disponibles: Número total de asientos disponibles en cada vehículo.
- Día de la Semana de Partida y Arribo: Día correspondiente a la salida y llegada del servicio.

El análisis de estos datos permitió evaluar las fluctuaciones en la ocupación de los servicios y su relación con distintas variables económicas y operativas. Se aplicaron técnicas de preprocesamiento para tratar valores faltantes, eliminar inconsistencias y normalizar las variables antes de utilizarlas en los modelos predictivos.

3.12.2. Variables Macroeconómicas

Adicionalmente, se recopilaron datos macroeconómicos relevantes provenientes de bases de datos oficiales del Gobierno de Argentina, con el objetivo de analizar su impacto en la demanda de transporte terrestre. Estas variables incluyen:

- Índice de Precios al Consumidor (IPC): Se utilizaron cuatro versiones del IPC:
  - IPC General
  - IPC Transporte
  - IPC Restaurantes y Hoteles
  - IPC Salud

- Tipo de Cambio: Valor del dólar en pesos argentinos, considerando su volatilidad y efecto en el poder adquisitivo de los pasajeros.
- Valor del Combustible (Bencina): Precio del combustible en Argentina, el cual impacta directamente en los costos operativos de los servicios de transporte.

Estas variables fueron seleccionadas debido a su influencia en el comportamiento del consumo y la movilidad de los pasajeros. La volatilidad de la economía argentina genera desafíos en la predicción de la ocupación de los buses, por lo que se evaluó si estos factores macroeconómicos pueden mejorar la capacidad predictiva de los modelos.

3.12.3. Construcción de los Conjuntos de Datos

Para evaluar la robustez de los modelos predictivos, los datos fueron segmentados en dos ventanas temporales:

1. Datos de 2024: Un subconjunto de datos exclusivo del año en curso, permitiendo evaluar el comportamiento reciente de la demanda de transporte.
2. Datos de 2022-2024: Un conjunto más amplio que abarca tres años completos, con el fin de analizar tendencias a largo plazo y evaluar el impacto de las fluctuaciones económicas en la ocupación de los servicios.

El uso de múltiples ventanas temporales permite evaluar la estabilidad de los modelos en diferentes condiciones y detectar variaciones en el desempeño predictivo.

3.12.4. Procesamiento y Análisis de los Datos

Antes de su uso en los modelos predictivos, los datos fueron sometidos a un proceso de preprocesamiento y limpieza. Entre las técnicas aplicadas se incluyen:

- Manejo de valores nulos: Se aplicaron estrategias de imputación para completar registros faltantes.
- Normalización y transformación de variables: Se estandarizaron ciertas variables para mejorar la estabilidad de los modelos.
- Creación de nuevas variables: Se derivaron métricas como $/KM$ para evaluar la rentabilidad por kilómetro recorrido.
- Análisis exploratorio de datos: Se realizaron visualizaciones para comprender las distribuciones de las variables y su correlación con la ocupación de los buses.

El resultado de este procesamiento permitió garantizar la calidad de los datos utilizados en la construcción de los modelos predictivos, mejorando su capacidad para realizar estimaciones precisas y respaldar la toma de decisiones en la optimización de los servicios de transporte.

4 | Metodología de Implementación

4.1. Recopilación de Datos Históricos del Factor de Ocupación

La presente investigación se estructura a partir de un enfoque metodológico que permite implementar, evaluar y comparar distintos modelos de predicción sobre el comportamiento de la ocupación en el transporte de pasajeros. En primer lugar, se llevó a cabo una recopilación detallada de datos históricos provenientes de la base de datos de Cruceros del Norte, accedida a través de Power BI, la cual permitió obtener información clave sobre servicios, fechas de viaje, ocupación, kilómetros recorridos, tarifas, tipos de línea, y otros atributos fundamentales para el análisis. Complementariamente, se integraron variables macroeconómicas externas como el Índice de Precios al Consumidor (IPC) general y sectorial (transporte, salud, restaurantes y hoteles), el tipo de cambio oficial y el valor del combustible en Argentina, recopilados desde bases de datos públicas gubernamentales. Esta combinación de variables internas y externas constituyó una base robusta para el posterior modelado.

4.2. Análisis Exploratorio y Preparación de Datos

Una vez recopilada la información, se procedió con un análisis exploratorio de datos (EDA) orientado a la identificación de patrones temporales, estacionalidades, outliers y comportamientos atípicos que pudieran impactar en los resultados predictivos. Se analizaron tanto ventanas acotadas al año 2024 como ventanas extendidas desde 2022 hasta 2024, con el fin de evaluar cómo se comportan los modelos ante distintos horizontes temporales y grados de variabilidad. Se aplicaron procesos de limpieza y transformación de datos, normalización de variables, y generación de estructuras tipo ventana para los modelos basados en aprendizaje automático. Asimismo, se diferenciaron los conjuntos de datos en entrenamiento y validación para asegurar una correcta evaluación del desempeño.

4.3. Modelos de Predicción

Posteriormente, se implementaron siete modelos predictivos, divididos en modelos clásicos y modelos modernos de aprendizaje automático. Dentro de los modelos clásicos, se utilizaron SARIMA y SARIMAX. El primero permitió capturar estacionalidades y tendencias en la serie de ocupación, mientras que el segundo incorporó las variables macroeconómicas como factores exógenos para enriquecer la capacidad explicativa del modelo. En cuanto a los modelos modernos, se utilizaron redes neuronales recurrentes del tipo LSTM y GRU, cuya arquitectura permite capturar dependencias temporales a largo plazo y adaptarse a la complejidad no lineal de la serie. Estas redes fueron desarrolladas utilizando TensorFlow y Keras dentro del entorno de Jupyter Notebook. Además, se integraron modelos basados en árboles de decisión, como XGBoost, Random Forest y Gradient Boosting, que permitieron modelar relaciones no lineales entre múltiples variables de entrada, entregando predicciones robustas y altamente interpretables. Cada modelo fue entrenado y ajustado utilizando técnicas de validación cruzada, búsqueda de hiperparámetros, y procesamiento específico según sus requerimientos estructurales.

4.4. Resultados y Comparación

Una vez implementados todos los modelos, se procedió a su evaluación comparativa utilizando métricas de error como el Root Mean Square Error (RMSE), el Mean Absolute Error (MAE) y el coeficiente de determinación (R^2). Estas métricas permitieron identificar la precisión y estabilidad de cada modelo, estableciendo cuál de ellos ofrece la mejor capacidad predictiva en el contexto del transporte terrestre de pasajeros. La comparación también permitió observar las fortalezas de cada enfoque según el tipo de datos utilizados, lo que refuerza la importancia de considerar múltiples metodologías para una problemática tan variable como la ocupación en servicios de transporte.

4.5. Aplicación y Funcionalidad

Finalmente, se discutieron las posibles aplicaciones prácticas de los modelos seleccionados. Los resultados obtenidos pueden ser incorporados en las operaciones diarias de Cruceros del Norte para la planificación de rutas, gestión de flota y estrategias tarifarias, contribuyendo a una mejor toma de decisiones a nivel estratégico. Además, se considera el potencial de integrar estos modelos en plataformas ya utilizadas por la empresa, como Power BI o herramientas del

entorno de Google Cloud utilizadas por Advitair, para generar paneles predictivos y alertas operativas que mejoren la respuesta ante escenarios de alta o baja demanda. El modelo final recomendado dependerá del equilibrio entre precisión, tiempo de entrenamiento, interpretabilidad y capacidad de integración con los sistemas actuales.

5 | Etapas de Implementacíon y Análisis

5.1. Recopilación de Datos

La recopilación de datos representa un paso fundamental para el desarrollo de esta investigación, ya que permite contar con una base sólida sobre la cual construir los modelos predictivos del factor de ocupación. En esta sección se detallan las fuentes, tipos de datos recopilados, así como los aspectos clave asociados a la calidad, seguridad y privacidad de los mismos.

5.1.1. Fuente y Tipos de Datos

Los datos internos de Cruceros del Norte fueron obtenidos desde Power BI y Google BigQuery, plataformas utilizadas por la empresa para el almacenamiento y análisis de información operativa. Estos datos contienen información relevante para la estimación del factor de ocupación de los distintos servicios, y comprenden las siguientes variables:

- Servicio
- Id
- Mes
- Año
- FechaPartida
- FechaArribo
- Ocup (factor de ocupación)

- Costo_KM
- KMS
- Tarifa Media ADV
- Yield
- TipoLinea
- Pasajeros
- Butacas
- Día de la semana partida
- Día de la semana arribo

Además, se recopilaron datos externos con el objetivo de incorporar variables macroeconómicas que pudieran influir sobre el comportamiento de la ocupación. Estas variables fueron extraídas desde fuentes oficiales del Gobierno de Argentina, y comprenden:

- IPC General
- IPC Transporte
- IPC Restaurant y Hoteles
- IPC Salud
- Tipo de Cambio de Referencia (en pesos por dólar)
- Valor ARS/Litro (precio de la bencina)

Para este estudio, se utilizaron dos ventanas temporales distintas: una correspondiente exclusivamente al año 2024, y otra que abarca desde el año 2022 hasta el 2024. Esto permitió observar el comportamiento de los modelos en contextos más acotados y también con mayor cantidad de datos históricos, analizando cómo la variabilidad de las curvas influye en la precisión de las predicciones.

5.1.2. Consideraciones de Calidad de Datos

La calidad de los datos es un aspecto esencial en la implementación de modelos predictivos. Durante el proceso de recopilación, se implementaron una serie de medidas para asegurar la consistencia, integridad y exactitud de la información. Las principales consideraciones aplicadas fueron:

- **Verificación de integridad:** se revisó la presencia de registros nulos o incompletos, especialmente en fechas clave, ocupación y número de pasajeros.
- **Validación de exactitud:** se realizaron comparaciones aleatorias con registros históricos de la empresa para verificar la consistencia de los valores obtenidos desde Power BI. con la tendencia.
- **Homogeneización de escalas:** se estandarizaron las unidades y formatos de fecha, además de normalizar las variables requeridas para modelos sensibles a escalas como LSTM o GRU.

Estas medidas permitieron mejorar la calidad del conjunto de datos final, asegurando que los modelos sean entrenados y evaluados sobre una base sólida y representativa del comportamiento real del sistema.

5.1.3. Seguridad y Privacidad de los Datos

Si bien el presente estudio no utiliza información sensible de clientes, se mantuvo un enfoque estricto en cuanto al manejo responsable de los datos operacionales proporcionados por Cruceros del Norte. Para asegurar la privacidad de la información:

- Los datos fueron almacenados y procesados en entornos seguros y controlados, utilizando Google Cloud Platform como entorno principal.
- No se divulgaron cifras exactas de ingresos, tarifas ni rutas específicas.
- Los resultados de los modelos se presentan de manera agregada, sin comprometer la confidencialidad comercial de la empresa.

Estas precauciones aseguran que el análisis se realice respetando los principios de integridad, ética y confidencialidad requeridos en investigaciones aplicadas a entornos empresariales reales.

5.2. Análisis Exploratorio de Datos

El análisis exploratorio de datos (EDA) constituye una etapa crítica para comprender en profundidad el comportamiento del factor de ocupación a lo largo del tiempo, así como su relación con variables internas del sistema de transporte y condiciones macroeconómicas externas. Este proceso no sólo permite identificar patrones, tendencias y anomalías presentes en los datos, sino también establecer relaciones estadísticas significativas que servirán como base para la posterior selección y calibración de modelos predictivos. Para su desarrollo se utilizó el entorno Jupyter Notebook con herramientas como pandas, numpy, matplotlib, seaborn y statsmodels.

5.2.1. Estacionalidad y Descomposición de la Serie Temporal

Con el objetivo de analizar la evolución del factor de ocupación, se realizó una descomposición aditiva de la serie temporal entre los años 2022 y 2024. Esto permitió identificar y separar tres componentes esenciales: la tendencia a largo plazo, la estacionalidad cíclica y los residuos aleatorios.

La tendencia evidencia un comportamiento general de crecimiento y disminución de la ocupación a lo largo del tiempo. Se observan periodos de incremento sostenido que podrían estar relacionados con temporadas altas, vacaciones o mejoras operativas, así como periodos de caída que podrían deberse a disminución en la demanda o ajustes en los servicios.

La componente estacional muestra una clara periodicidad de 28 días, lo que sugiere que el comportamiento de la ocupación se repite en ciclos mensuales, posiblemente influenciado por factores como calendarios laborales, eventos mensuales o rutinas de los usuarios. Este hallazgo refuerza la importancia de incorporar la estacionalidad en los modelos de predicción.

Finalmente, los residuos se distribuyen de forma aleatoria en torno al cero, sin patrones sistemáticos, lo que indica que la mayor parte de la estructura de los datos está bien capturada por los componentes anteriores. Sin embargo, se evidencian algunos picos que podrían corresponder a eventos excepcionales como feriados, fallas operativas o condiciones climáticas adversas.

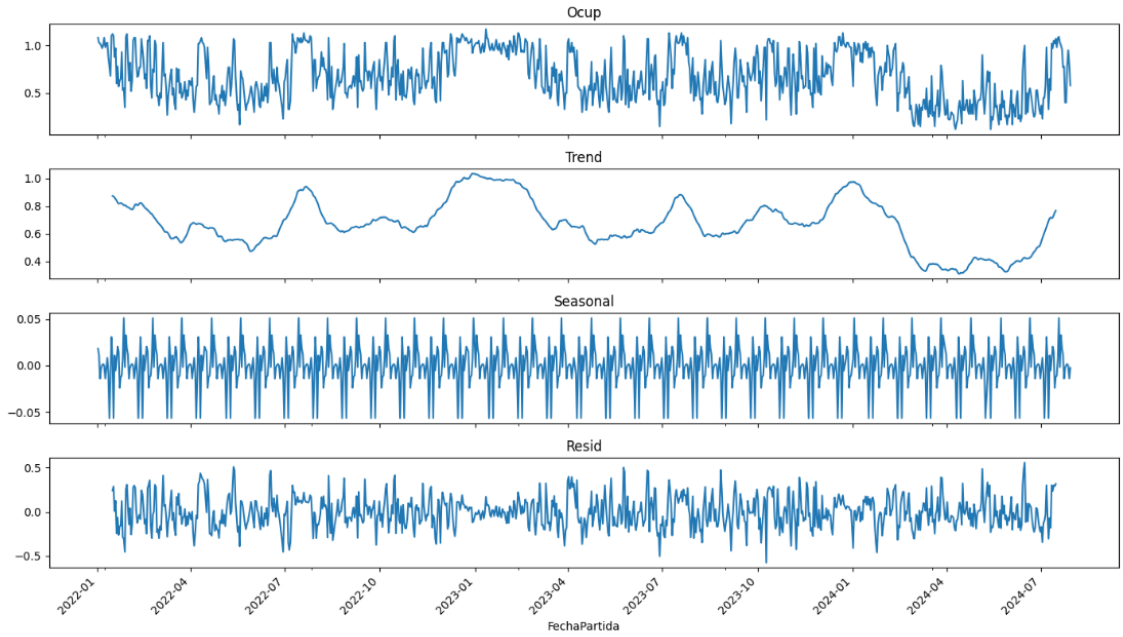


Figura 5.1: Descomposición de la serie temporal del factor de ocupación

5.2.2. Correlación con Variables Exógenas

Se exploró la relación entre la ocupación y una serie de variables macroeconómicas externas provenientes de fuentes oficiales del gobierno argentino: IPC General, IPC Transporte, IPC Restaurantes y Hoteles, IPC Salud, tipo de cambio oficial y el valor de la bencina en ARS/Litro.

La matriz de correlación mostró coeficientes negativos moderados entre la ocupación y todas las variables mencionadas, destacándose el IPC Transporte y el valor del combustible. Esta relación sugiere que aumentos en los precios –ya sea por inflación, costos del transporte o depreciación de la moneda– tienden a reducir la demanda del servicio, reflejándose en menores niveles de ocupación. Este hallazgo es coherente con la teoría económica sobre elasticidad precio-demanda y aporta evidencia empírica que respalda la necesidad de considerar estos factores en modelos explicativos y predictivos.

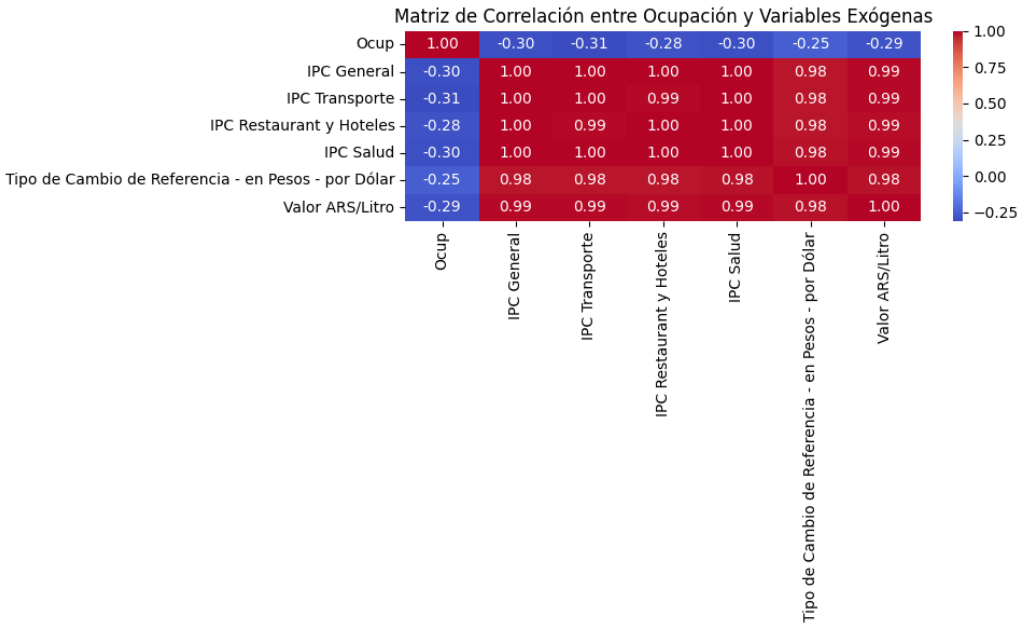


Figura 5.2: Matriz de correlación entre ocupación y variables exógenas

5.2.3. Análisis de Correlación General

Además de las variables exógenas, se analizó la correlación entre todas las variables numéricas del conjunto de datos, incluyendo tarifa media ADV, costo por kilómetro, número de pasajeros, butacas disponibles, yield, entre otros.

La ocupación mostró correlaciones negativas con la tarifa media y el yield, lo que sugiere que, a mayores precios, la utilización del servicio tiende a disminuir. Por otro lado, existe una correlación positiva con la cantidad de pasajeros y negativa con la cantidad de butacas, indicando que los niveles de ocupación están más directamente ligados al volumen de demanda efectiva que a la oferta instalada.

El mapa de calor permite identificar agrupamientos de variables interrelacionadas, especialmente entre los IPC y variables tarifarias, lo que refuerza la necesidad de evaluar la colinealidad antes de aplicar modelos supervisados. Esta información será clave al definir qué variables serán incluidas como predictoras y cuáles podrían ser redundantes o conflictivas.

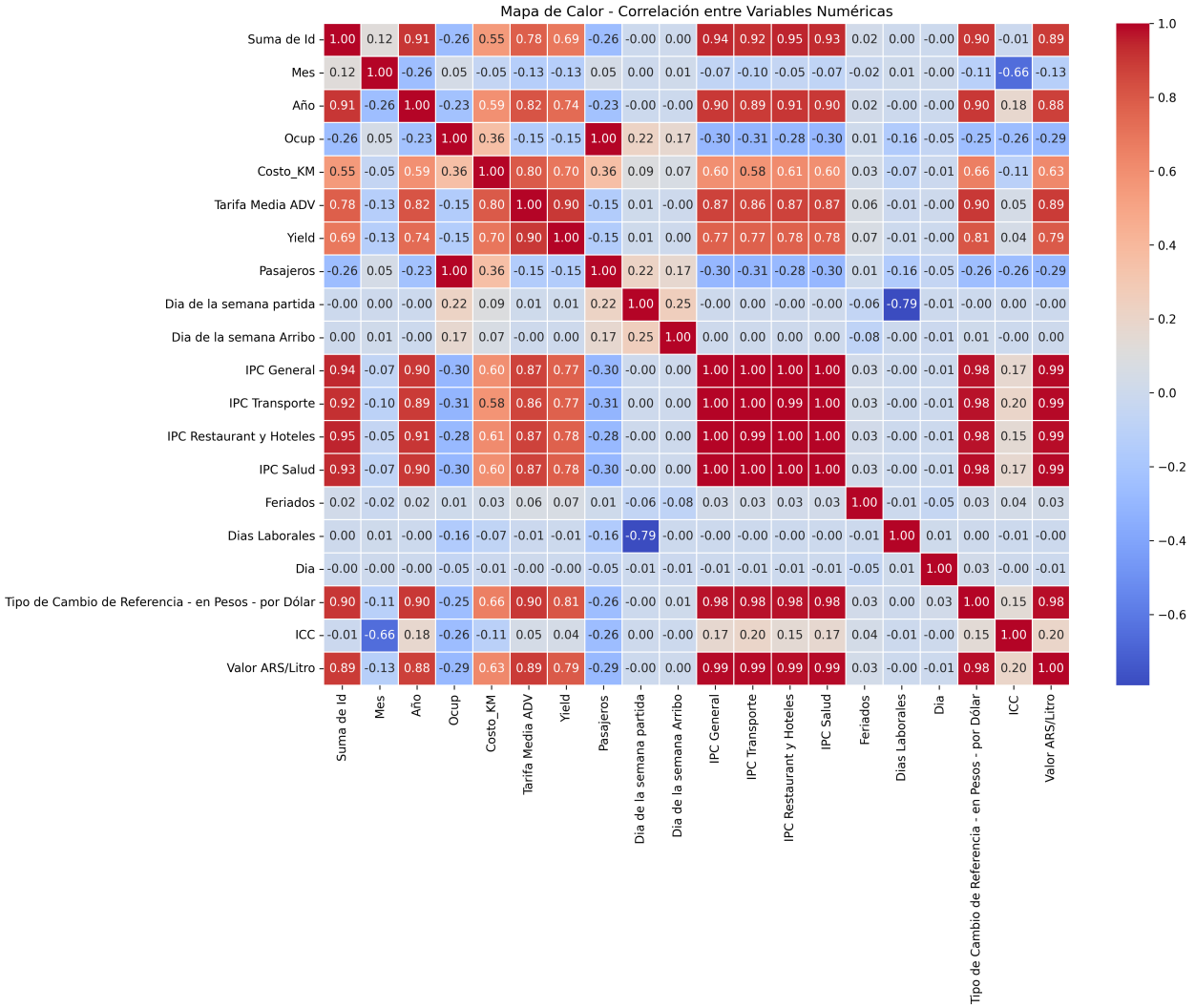


Figura 5.3: Mapa de calor de correlación entre variables numéricas internas

La Figura 5.2 y la Figura 5.3 presentan la matriz de correlación y el mapa de calor entre la ocupación (Ocup) y las variables exógenas, respectivamente. Los resultados muestran una correlación positiva moderada entre el Costokm y la ocupación (r= 0.36), sugiriendo que un mayor costo por kilómetro está asociado con un aumento en la ocupación, posiblemente reflejando ajustes operativos. Sin embargo, las variables económicas como IPC General (r= -0.30), IPC Transporte (r= -0.31), IPC Restaurante y Hoteles (r= -0.28), IPC Salud (r= -0.30), Tipo de cambio (r= -0.26), e ICC (r= -0.26) presentan correlaciones negativas moderadas, indicando que incrementos en estos indicadores tienden a reducir la ocupación, coherente con la volatilidad económica de Argentina. Días Laborables muestra una correlación negativa leve (r= -0.16), sugiriendo una menor ocupación durante estos días, mientras que Feriados tiene una correlación casi nula (r= 0.01), lo que podría indicar una codificación inadecuada o datos insuficientes. Día de la semana partida (r= 0.22) y Día de la semana arribo (r= 0.17) tienen influencias positivas leves. Estos hallazgos resaltan la relevancia de los

costos operativos y las dinámicas económicas, aunque la falta de granularidad o interacciones no lineales podría limitar el análisis (Hyndman Athanasopoulos, 2018).

5.2.4. Clusterización y Segmentación de Servicios

Para identificar patrones subyacentes en los servicios, se aplicó el algoritmo de K-means sobre las variables de ocupación, tarifa media ADV y costo por kilómetro. El número óptimo de clústeres se determinó mediante el método del codo, concluyendo en la existencia de cuatro segmentos bien diferenciados.

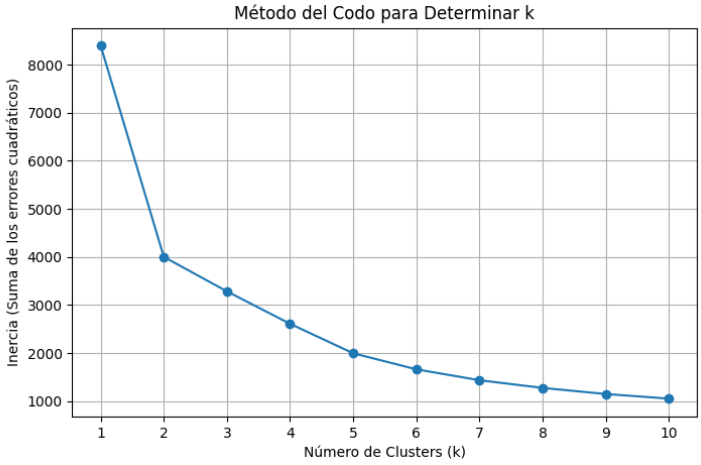


Figura 5.4: Método del codo para determinar el número de clústeres

Cada clúster refleja un grupo de servicios con características similares. Por ejemplo, uno de los grupos incluye servicios con baja ocupación, bajas tarifas y altos costos por kilómetro, lo que podría representar rutas poco eficientes o con baja demanda. Otro grupo incluye servicios con alta ocupación y tarifas más elevadas, probablemente asociados a rutas principales o de alta rentabilidad.

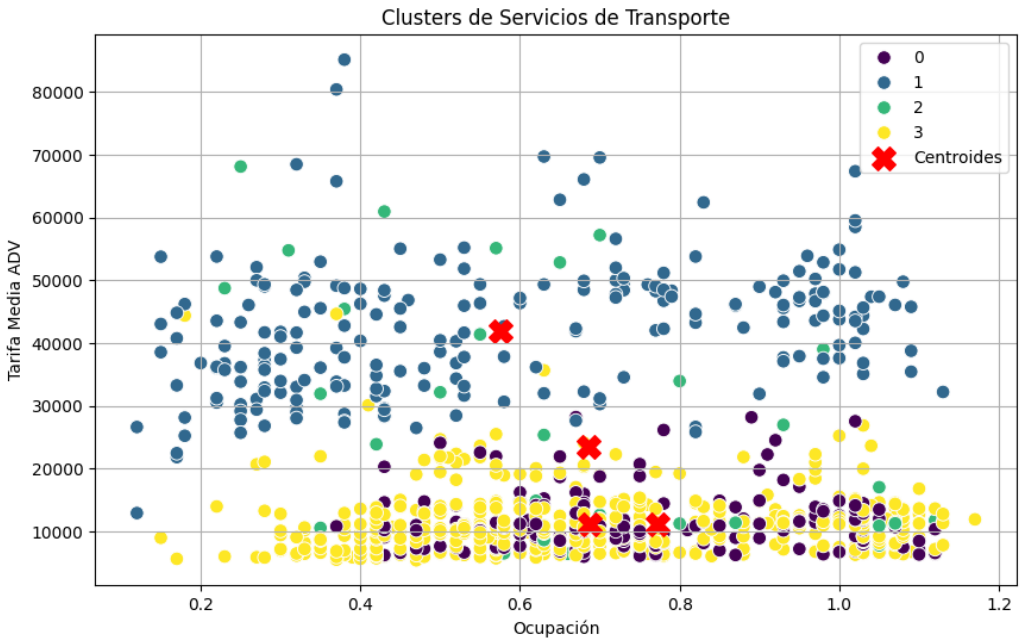


Figura 5.5: Clústeres según ocupación y tarifa media ADV

La visualización en 2D y 3D de estos grupos facilita la interpretación operativa, permitiendo identificar oportunidades de mejora, optimización de flota, o ajustes tarifarios según el perfil del servicio. Además, los centroides de los clústeres pueden emplearse como perfiles representativos para análisis posteriores.

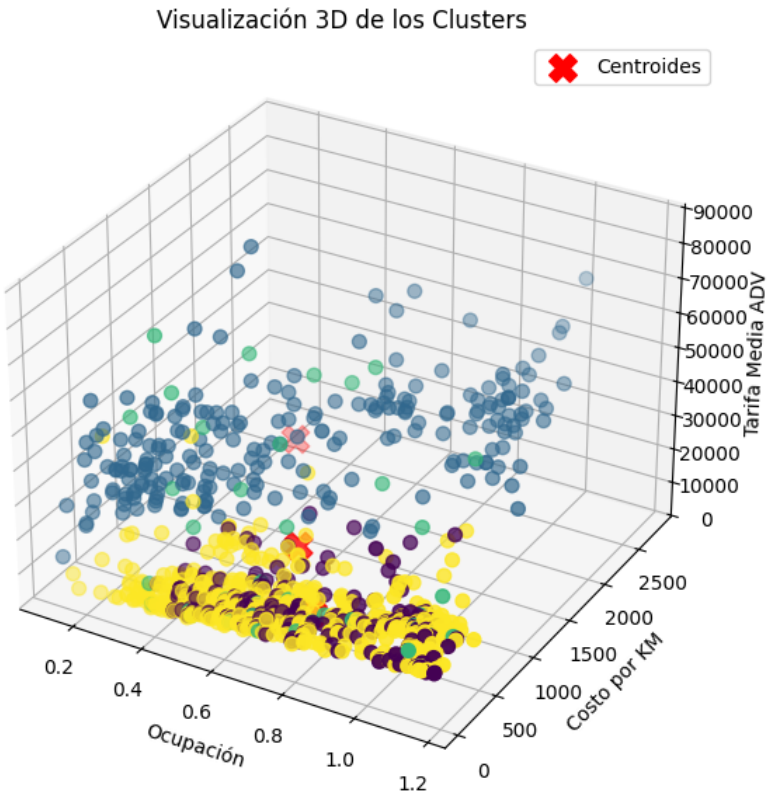


Figura 5.6: Visualización 3D de los clústeres de servicios

Los clusters identificados reflejan perfiles de servicios diferenciados en cuanto a su eficiencia y rentabilidad. La Tabla 5.1 resume las principales características de cada uno.

Tabla 5.1: Descripción general de los clusters de servicios

Cluster	Ocupación Promedio	Tarifa Media ADV	Costo por KM
0	Baja (< 0,5)	Baja (menor a $15.000)	Medio (~ $900)
1	Media (0,5–0,8)	Alta (sobre $30.000)	Medio-Alto
2	Alta (> 0,9)	Alta (mayor a $40.000)	Bajo (~ $600)
3	Variable (0,4–0,9)	Baja a media	Alto (sobre $1.200)

Este proceso de segmentación permite identificar servicios eficientes y rentables (como los del cluster 2), así como aquellos con potencial de mejora (como los del cluster 3, con alto costo y baja tarifa). Estos resultados son de gran utilidad para la toma de decisiones estratégicas en la planificación de rutas, definición de precios, asignación de recursos

y evaluación de la sostenibilidad operativa.

En resumen, el análisis exploratorio ha permitido comprender en detalle la estructura y comportamiento del sistema, revelando tanto patrones esperados como aspectos críticos para la mejora del servicio y la construcción de modelos predictivos robustos.

5.3. Modelos de Predicción

Con el análisis previo de los datos, se procede a la modelación de estos mediante una combinación de métodos estadísticos clásicos y modelos modernos de aprendizaje automático. Aunque los datos han sido limpiados y estructurados previamente, cada modelo presenta requerimientos específicos en cuanto a tratamiento y preparación de los datos, por lo que fue necesario adaptar las entradas según la arquitectura de cada modelo para asegurar un rendimiento óptimo.

La implementación de los modelos se realizó en el entorno de desarrollo integrado Spyder, utilizando el lenguaje de programación Python. El código correspondiente a cada modelo se encuentra documentado y almacenado en el repositorio de GitHub desarrollado durante este proyecto. Los modelos seleccionados fueron evaluados en términos de precisión predictiva, capacidad de generalización y su utilidad práctica en contextos reales de toma de decisiones en el sector del transporte.

A continuación, se detallan los siete modelos utilizados:

SARIMA (Seasonal Autoregressive Integrated Moving Average): Modelo estadístico que amplía ARIMA al incorporar componentes estacionales. Es particularmente útil para series temporales con patrones regulares de estacionalidad y tendencia. Fue uno de los primeros modelos aplicados para establecer una línea base de comparación.

SARIMAX (SARIMA with Exogenous Regressors): Variante extendida de SARIMA que permite incorporar variables exógenas como predictores adicionales. En este estudio, se incluyeron indicadores macroeconómicos como el IPC y el tipo de cambio, permitiendo capturar efectos externos sobre la ocupación.

LSTM (Long Short-Term Memory): Tipo de red neuronal recurrente diseñada para aprender dependencias a largo plazo en secuencias de datos. Se utilizó una arquitectura profunda con varias capas ocultas y mecanismos de regularización, entrenada con el conjunto de series temporales de ocupación. Este modelo mostró un buen comportamiento al capturar patrones complejos y no lineales.

GRU (Gated Recurrent Unit): Arquitectura de red similar a LSTM, pero con menor complejidad. Su implementación permitió una reducción en el tiempo de entrenamiento sin sacrificar precisión de manera significativa. Resultó una alternativa eficiente frente al modelo LSTM.

XGBoost (Extreme Gradient Boosting): Algoritmo de boosting basado en árboles de decisión, conocido por su alto rendimiento y capacidad de generalización. Se alimentó con variables derivadas del dataset y transformaciones estadísticas, logrando resultados competitivos incluso en comparación con modelos basados en redes neuronales.

Random Forest: Ensamble de múltiples árboles de decisión entrenados sobre subconjuntos aleatorios de los

datos y características. Su principal fortaleza radica en la robustez frente al overfitting y en su capacidad para manejar datasets con muchas variables explicativas.

Gradient Boosting: Método de boosting que construye árboles secuenciales optimizando el error en cada iteración. Comparado con Random Forest, este modelo tiende a ofrecer mejores resultados en datasets bien estructurados pero requiere un mayor ajuste de hiperparámetros.

Cada uno de estos modelos fue entrenado y evaluado utilizando una partición temporal de los datos, validando su rendimiento con métricas como el error cuadrático medio (RMSE), el error absoluto medio (MAE) y el coeficiente de determinación (R^2). La variedad de enfoques permite contrastar la capacidad predictiva de modelos clásicos frente a algoritmos modernos en el contexto del transporte terrestre, ofreciendo una base sólida para seleccionar la solución más adecuada a implementar.

5.3.1. SARIMA

Una vez completado el análisis exploratorio y entendidas las características estacionales y tendencias de la variable *ocupación*, se procedió a la implementación del modelo SARIMA (Seasonal Autoregressive Integrated Moving Average), ampliamente reconocido por su capacidad para capturar tanto dinámicas temporales como patrones estacionales recurrentes. El modelo fue aplicado a dos subconjuntos de la serie temporal: uno correspondiente al periodo completo 2022–2024, y otro enfocado exclusivamente en el año 2024, con el fin de evaluar la estabilidad del modelo ante distintas longitudes de series y cambios en los patrones de comportamiento.

Para ambos casos, se trabajó con datos previamente suavizados mediante media móvil de ventana 7, con el objetivo de reducir ruido estocástico sin perder la estructura de tendencia. La variable objetivo fue indexada por fecha de partida y, posteriormente, la serie fue dividida en conjunto de entrenamiento (80 %) y de prueba (20 %). Los modelos fueron entrenados con la librería `statsmodels`, utilizando como método de optimización `powell` y con un máximo de 5000 iteraciones, lo que permitió una convergencia robusta incluso en presencia de posibles multicolinealidades o estructuras complejas.

Los parámetros óptimos definidos para el modelo SARIMA sobre la serie completa (2022–2024) fueron: `order=(3, 0, 2)` y `seasonal_order=(2, 0, 2, 7)`. Estos parámetros fueron seleccionados tras un proceso de ajuste iterativo basado en la evaluación de criterios como el AIC y BIC, así como mediante validación de las predicciones generadas sobre el conjunto de prueba. Para el modelo centrado en el año 2024, los parámetros se mantuvieron constantes a fin de observar su capacidad de generalización en un escenario de corto plazo con fuerte variabilidad exógena.

Las figuras a continuación ilustran las predicciones generadas por el modelo SARIMA frente a los datos reales,

diferenciando entre conjunto de entrenamiento, valores reales observados y predicciones realizadas por el modelo:

El modelo SARIMA (Seasonal Autoregressive Integrated Moving Average) es una extensión del modelo ARIMA que incorpora componentes estacionales, permitiendo capturar patrones tanto de tendencia como de estacionalidad en series temporales. Este enfoque es particularmente útil en contextos donde los datos presentan una estructura cíclica clara, como es el caso del transporte de pasajeros.

En esta investigación, se desarrollaron dos modelos SARIMA distintos: uno abarcando el periodo completo 2022–2024 y otro exclusivamente sobre el año 2024. Ambos modelos fueron implementados en Python utilizando las bibliotecas `pandas`, `numpy`, `statsmodels` y `sklearn`. El proceso general consistió en la carga y limpieza de datos, suavizado mediante media móvil, ajuste del modelo, predicción sobre el conjunto de prueba y posterior evaluación mediante métricas de error.

Modelo SARIMA para el periodo 2022–2024

Para este modelo, los parámetros óptimos encontrados fueron: `order=(3,0,2)` y `seasonal_order=(2,0,2,7)`. Se utilizó un enfoque de optimización `Powell` con un número máximo de iteraciones de 5000. Los resultados obtenidos muestran un desempeño aceptable del modelo, con un buen ajuste visual a los datos reales.

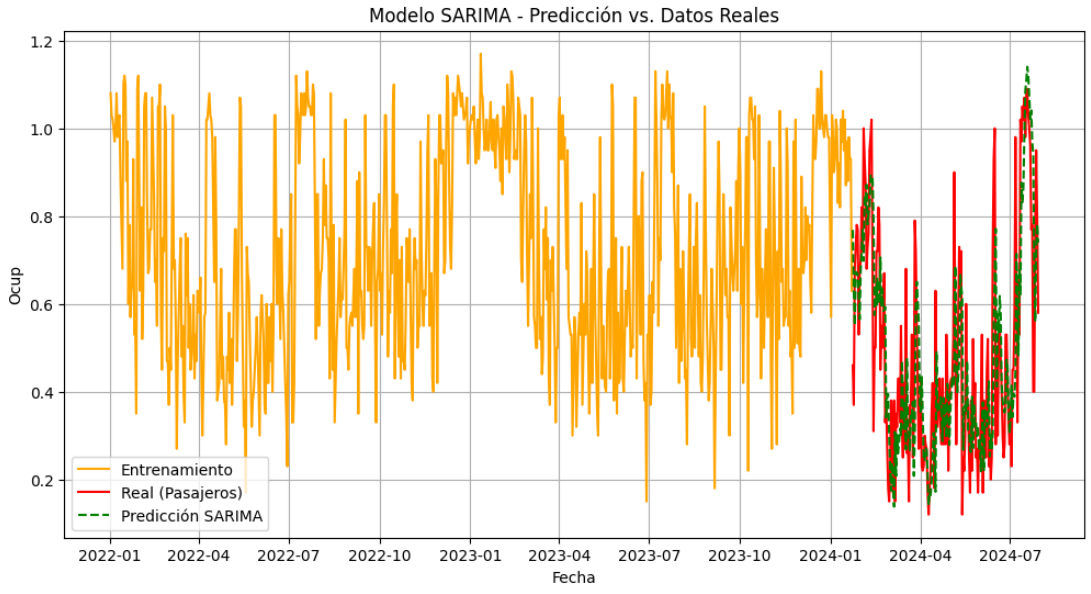


Figura 5.7: Predicción del Modelo SARIMA (2022–2024) vs. Datos Reales

Tabla 5.2: Métricas de Evaluación del Modelo SARIMA (2022–2024)

Métrica	**Valor**
Raíz del Error Cuadrático Medio (RMSE)	0.1790
Error Absoluto Medio (MAE)	0.1366
Coeficiente de Determinación (R^2)	0.4956

Estos valores sugieren que el modelo SARIMA fue capaz de capturar adecuadamente las variaciones en la ocupación a lo largo del periodo evaluado. En particular, el R^2 cercano a 0.5 indica que aproximadamente el 50 % de la varianza en la ocupación es explicada por el modelo. Aunque existen diferencias puntuales entre las predicciones y los valores reales, especialmente en los extremos de la serie, la tendencia general es bien replicada.

Modelo SARIMA para el año 2024

Para capturar la dinámica específica del año 2024, se entrenó un segundo modelo SARIMA utilizando únicamente datos de dicho año. Este modelo utilizó los mismos parámetros estructurales que el anterior, pero mostró un rendimiento considerablemente más bajo, lo que evidencia la complejidad de predecir ocupación con datos limitados a un solo año.

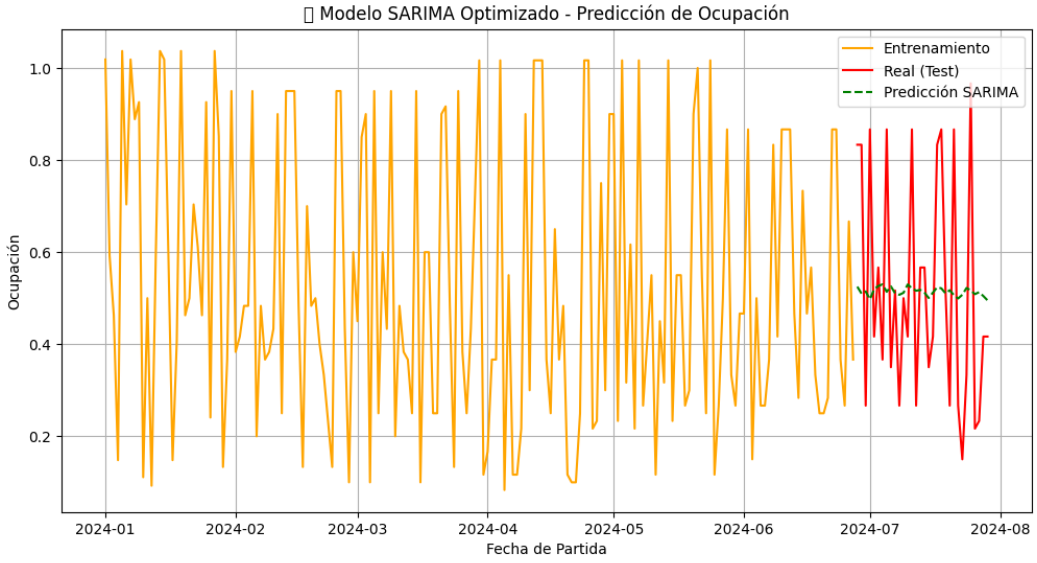


Figura 5.8: Predicción del Modelo SARIMA (2024) vs. Datos Reales

Tabla 5.3: Métricas de Evaluación del Modelo SARIMA (2024)

Métrica	**Valor**
Raíz del Error Cuadrático Medio (RMSE)	0.2428
Error Absoluto Medio (MAE)	0.2074
Coeficiente de Determinación (R^2)	0.0087

El bajo valor de R^2 indica que el modelo tiene escasa capacidad predictiva al utilizar únicamente los datos del año 2024, probablemente debido a una menor cantidad de datos, mayor volatilidad o presencia de eventos atípicos no considerados explícitamente en la modelación.

Análisis Comparativo y Consideraciones

Los resultados obtenidos demuestran diferencias significativas entre los modelos entrenados con distintos periodos de datos. El modelo SARIMA ajustado al intervalo 2022–2024 mostró un desempeño notoriamente superior en todas las métricas: menor error absoluto medio (MAE), menor raíz del error cuadrático medio (RMSE) y un coeficiente de determinación (R^2) de 0.4956, lo que implica que cerca del 50 % de la varianza de la ocupación es explicada por el modelo. Este comportamiento evidencia que una serie de mayor longitud, al incorporar más patrones históricos y estacionales, facilita una mejor modelación de las tendencias subyacentes.

En contraste, el modelo ajustado exclusivamente sobre el año 2024 presenta un RMSE y MAE considerablemente más altos, junto con un valor de R^2 de apenas 0.0087, lo cual indica una capacidad predictiva prácticamente nula. Esto sugiere que la reducción de la ventana temporal implicó una pérdida de patrones relevantes, probablemente debido a una mayor volatilidad o cambios estructurales en la demanda observada durante 2024. Además, la falta de un histórico robusto impidió al modelo capturar adecuadamente la dinámica de la serie, generando predicciones más erráticas y menos alineadas con los datos reales.

Estas diferencias entre ambos modelos permiten concluir que, en contextos donde se cuenta con series temporales extensas y estables, los modelos SARIMA pueden desempeñarse de manera sólida y confiable. Sin embargo, en escenarios más acotados temporalmente, donde las series presentan alta variabilidad o eventos disruptivos, el modelo puede resultar insuficiente, siendo recomendable evaluar alternativas más flexibles como modelos basados en aprendizaje profundo o técnicas con manejo explícito de variables exógenas.

5.3.2. SARIMAX

El modelo SARIMAX constituye una extensión directa del modelo SARIMA presentado previamente, incorporando variables exógenas que permiten capturar variaciones adicionales que no pueden explicarse únicamente a partir de la estructura autorregresiva y estacional de la serie. En este caso, se han considerado tres variables explicativas: *IPC General*, *Costo por Kilómetro* y *Feriados*, seleccionadas por su relevancia contextual en la industria del transporte y su correlación con la ocupación de pasajeros en los servicios. Esta inclusión permite modelar el impacto de factores macroeconómicos y estacionales externos sobre la variable objetivo.

Para la implementación de este modelo se utilizó el entorno Python mediante la biblioteca `statsmodels`, junto con `Pandas`, `Numpy` y `Scikit-learn` para la manipulación de datos y cálculo de métricas. Los datos fueron particionados en una proporción 80 % para entrenamiento y 20 % para testeo, manteniendo el orden cronológico para evitar fuga de información.

Se estimaron dos modelos SARIMAX con distintas ventanas temporales: uno con el rango completo entre los años 2022 y 2024, y otro centrado exclusivamente en el año 2024. Esta doble aproximación tiene por objetivo evaluar el comportamiento del modelo en dos escenarios: uno con mayor cantidad de datos históricos y otro que representa una visión más reciente y condensada, posiblemente más representativa de las condiciones actuales del mercado.

En ambos casos, los modelos fueron ajustados con órdenes $(1, 0, 1)$ para la componente autorregresiva y de media móvil, y una componente estacional de orden $(1, 0, 1, 12)$, respetando la periodicidad anual esperada en este tipo de datos. La restricción de estacionariedad y invertibilidad fue aplicada para asegurar la estabilidad del modelo.

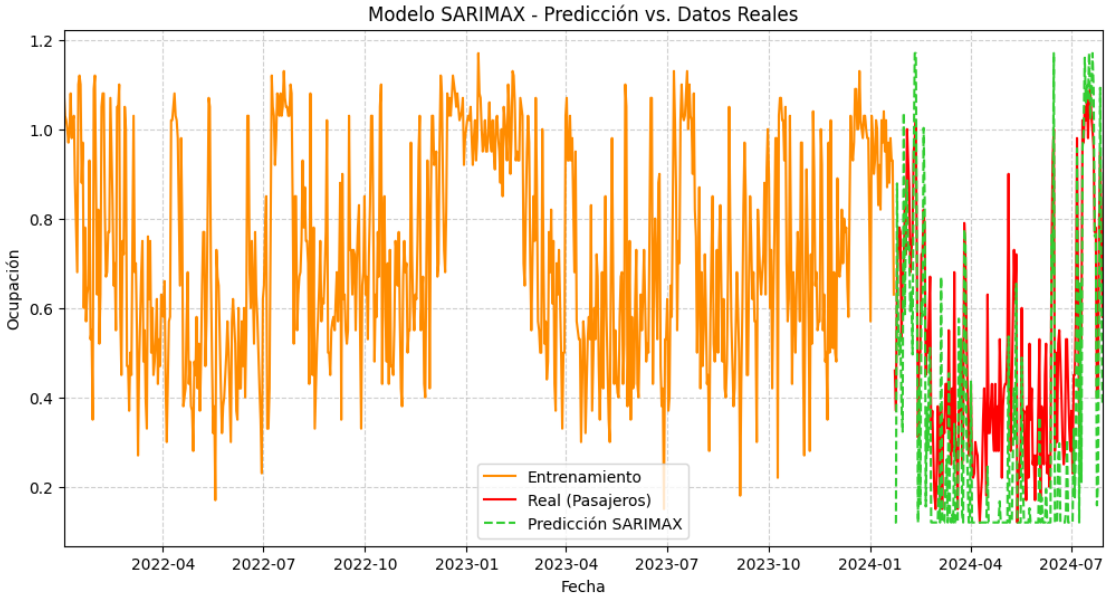


Figura 5.9: Predicción del modelo SARIMAX para el periodo 2022–2024.

Tabla 5.4: Métricas de Evaluación del Modelo SARIMAX (2022–2024)

Métrica	**Valor**
RMSE	0.2004
MAE	0.1817
R^2 Score	0.3675

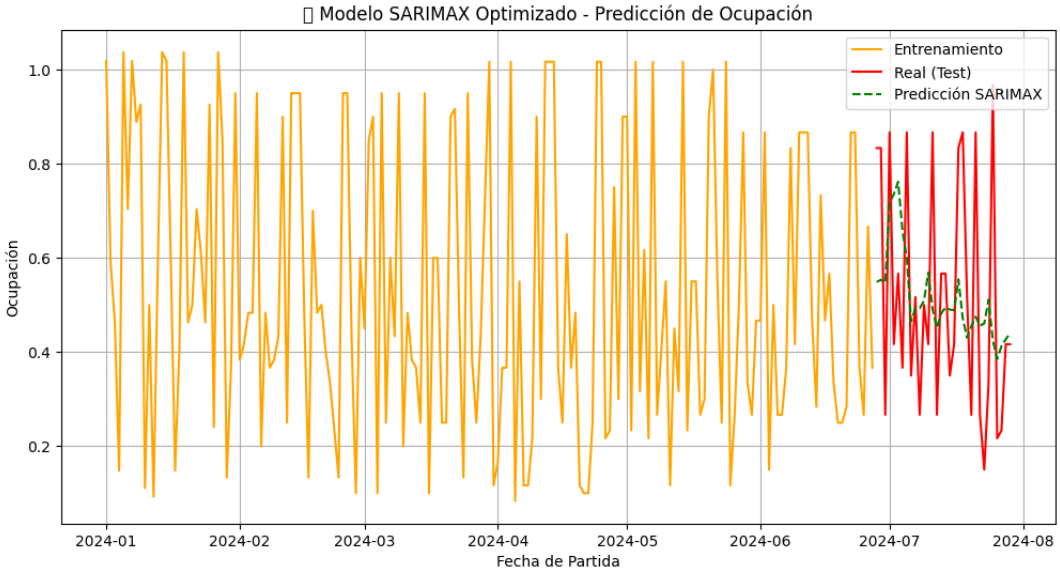


Figura 5.10: Predicción del modelo SARIMAX para el año 2024.

Tabla 5.5: Métricas de Evaluación del Modelo SARIMAX (2024)

Métrica	Valor
RMSE	0.2375
MAE	0.2019
R^2 Score	0.0512

Análisis y Comparación de Resultados

En la Figura 5.9 se observa el desempeño del modelo SARIMAX entrenado sobre el conjunto extendido de datos desde 2022 a 2024. Visualmente, las predicciones (línea verde punteada) logran seguir de forma moderadamente acertada la tendencia general de la serie de ocupación durante el conjunto de prueba (en rojo), aunque con ciertas fluctuaciones alrededor de los valores reales. Esto se confirma con un RMSE de 0.2004 y un MAE de 0.1817 (Tabla 5.5), valores bajos que sugieren un buen desempeño general. Además, el coeficiente de determinación R^2 de 0.3675 indica que el modelo logra explicar aproximadamente un 36.75 % de la variabilidad observada en los datos, lo cual resulta razonable considerando la complejidad de la serie y la inclusión de variables exógenas.

Por otro lado, el modelo entrenado exclusivamente con datos del año 2024, ilustrado en la Figura 5.10, presenta un comportamiento más errático y menos ajustado a la realidad. Si bien sigue la tendencia base en ciertos tramos, se observan divergencias notorias entre las predicciones y los valores reales. Esto se traduce en un RMSE superior de 0.2375 y un MAE de 0.2019 (Tabla 5.5), acompañados de un R^2 significativamente más bajo (0.0512), lo que sugiere una capacidad explicativa muy limitada del modelo en este contexto.

La diferencia en desempeño entre ambos modelos puede explicarse por la cantidad y calidad de la información utilizada para su entrenamiento. El modelo basado en un rango más amplio (2022–2024) dispone de mayor diversidad estacional y comportamiento histórico, lo que permite capturar mejor los patrones subyacentes. En cambio, el modelo basado únicamente en 2024 podría estar sobreajustado a un contexto más restringido, con menor representación de variabilidad estacional y menor volumen de datos, lo que limita su capacidad de generalización.

Asimismo, la inclusión de variables exógenas demostró ser útil en ambos escenarios, aunque con mayor impacto en el modelo extendido. Las variables como el IPC General y los Feriados pueden haber aportado información valiosa en contextos de mayor variabilidad histórica. En el caso del modelo 2024, es posible que la influencia de estas variables haya sido menos significativa debido a la menor diversidad temporal.

En conclusión, el modelo SARIMAX entrenado con datos desde 2022 hasta 2024 muestra un desempeño superior y más robusto, tanto en métricas como en representación visual. Este resultado respalda la importancia de utilizar

contextos temporales amplios para el entrenamiento de modelos predictivos cuando se trabaja con series temporales de naturaleza estacional y dependientes de factores externos.

5.3.3. LSTM con Variables Exógenas

La predicción de la ocupación en servicios de transporte puede verse fuertemente influenciada por factores externos a la serie temporal, conocidos como variables exógenas. A fin de incorporar esta complejidad, se implementó un modelo basado en redes neuronales LSTM que incluye múltiples variables exógenas, con el objetivo de mejorar la capacidad predictiva respecto de modelos univariados.

Este enfoque se fundamenta en la capacidad de las redes LSTM para modelar relaciones no lineales y dinámicas a largo plazo, especialmente cuando se integran factores contextuales relevantes. En este caso, se consideraron como variables exógenas el **IPC General**, el **Costo por Kilómetro (Costo_KM)** y la variable categórica **Feriados**, que representan componentes macroeconómicos, operacionales y estacionales, respectivamente. Estas variables fueron seleccionadas en base a su relevancia teórica y su disponibilidad en el conjunto de datos.

Preprocesamiento y generación de secuencias

Se comenzó por aplicar un preprocesamiento basado en escalamiento MinMax a las variables seleccionadas, asegurando que todas las series utilizadas como entrada estuvieran normalizadas entre 0 y 1. A continuación, se generaron secuencias utilizando una ventana deslizante de tamaño 90, lo cual permite al modelo capturar la dinámica de las variables en los últimos tres meses para cada instancia de predicción.

División de datos y arquitectura del modelo

Los datos fueron divididos en subconjuntos de entrenamiento (80 %) y validación (20 %) sin aleatorización, respetando la secuencia temporal. La arquitectura propuesta del modelo consiste en una red neuronal LSTM bidireccional, distribuida en múltiples capas con el fin de capturar patrones complejos:

- Cuatro capas **Bidirectional LSTM** con 128, 64, 32 y 16 unidades respectivamente.
- **Batch Normalization** y **Dropout** (tasa de 0.1) después de cada capa LSTM para evitar el sobreajuste.
- Capas densas finales con 32 unidades (`ReLU`) y una unidad de salida lineal.

El modelo fue compilado con el optimizador Adam y una tasa de aprendizaje inicial de 0,0003. Para mejorar el desempeño, se aplicaron las estrategias de **ReduceLROnPlateau** y **EarlyStopping** durante el entrenamiento, utilizando como métrica de control la pérdida de validación.

Evaluación del modelo: periodo 2022–2024

Una vez entrenado el modelo con los datos correspondientes al periodo 2022–2024, se procedió a evaluar su capacidad predictiva. La Figura 5.11 muestra la comparación entre la serie real y la predicción del modelo. Se puede observar que la red neuronal logra capturar adecuadamente la tendencia general, aunque con ciertas limitaciones en eventos de alta variabilidad.

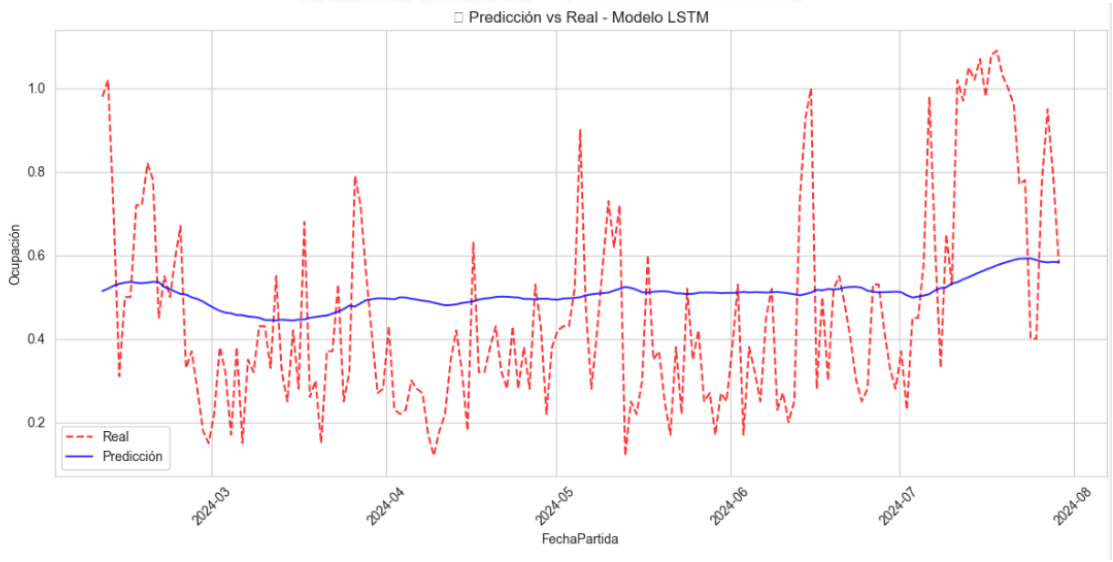


Figura 5.11: Predicción del modelo LSTM para el periodo 2022–2024

Tabla 5.6: Métricas de Evaluación del Modelo LSTM (2022–2024)

Métrica	Valor
RMSE	0.2241
MAE	0.1965
R^2 Score	0.1635

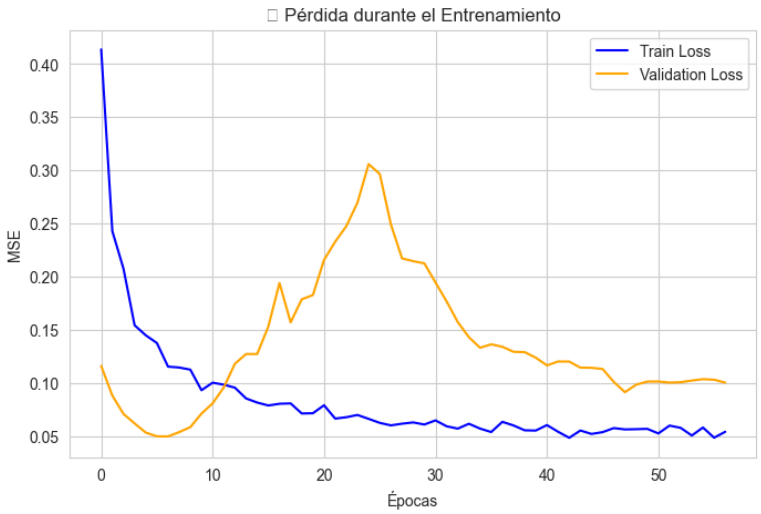


Figura 5.12: Evolución de la pérdida durante el entrenamiento del modelo LSTM (2022–2024)

Evaluación del modelo: año 2024

Se realizó un segundo experimento entrenando el modelo exclusivamente con datos del año 2024. Este análisis permite explorar la capacidad del modelo para aprender únicamente con eventos recientes. La predicción resultante se presenta en la Figura 5.13, mientras que las métricas se detallan en la Tabla 5.7.

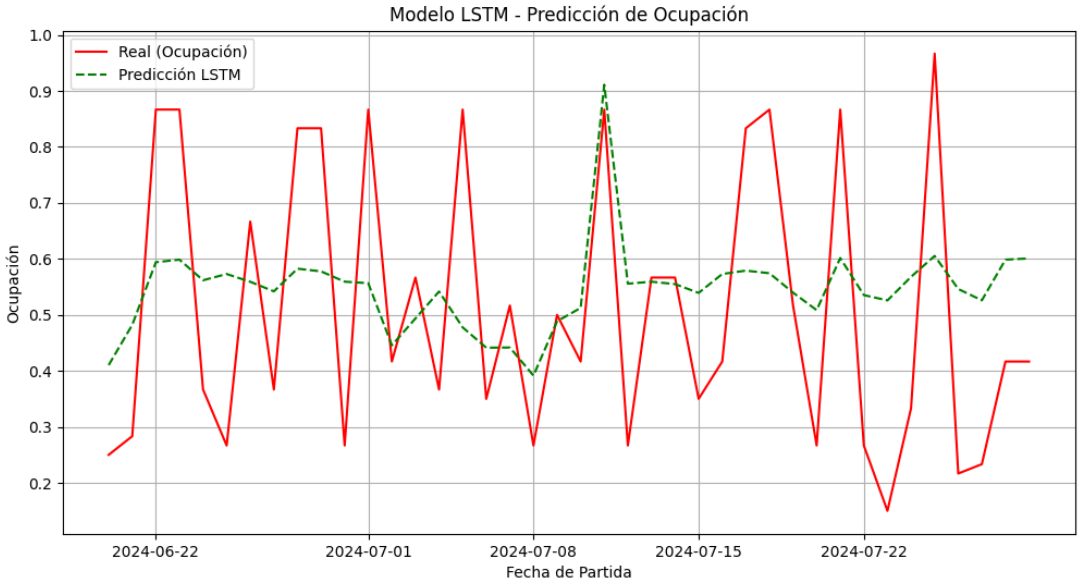


Figura 5.13: Predicción del modelo LSTM para el año 2024

Tabla 5.7: Métricas de Evaluación del Modelo LSTM (2024)

Métrica	Valor
RMSE	0.2350
MAE	0.1971
R^2 Score	0.0895

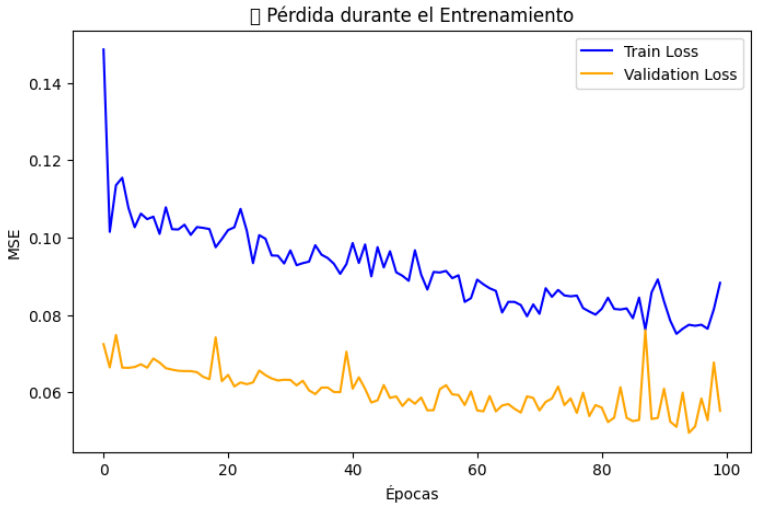


Figura 5.14: Evolución de la pérdida durante el entrenamiento del modelo LSTM (2024)

Análisis de resultados

El modelo LSTM con variables exógenas logra desempeños aceptables en ambos periodos analizados, destacando su capacidad para generalizar en contextos de alta variabilidad. En términos de error cuadrático medio (RMSE) y error absoluto medio (MAE), los valores obtenidos muestran una ligera mejora en el entrenamiento multianual respecto al entrenamiento exclusivo del 2024.

No obstante, se observa una disminución en el coeficiente de determinación (R^2) cuando se trabaja solo con datos del 2024. Este comportamiento puede atribuirse a la pérdida de patrones estacionales y tendencias históricas que el modelo deja de aprender al limitar la ventana temporal.

La inclusión de variables exógenas como el IPC, el Costo por Kilómetro y la presencia de feriados, permitió al modelo ajustar mejor sus predicciones a cambios no inherentes a la secuencia temporal de la ocupación. Sin embargo, algunos eventos extremos o no recurrentes siguen representando un desafío para la red, lo cual se evidencia en las fluctuaciones abruptas observadas en ciertos días.

En general, el modelo exhibe un buen equilibrio entre ajuste y generalización, sin signos marcados de sobreajuste, cómo puede verificarse en las curvas de pérdida de entrenamiento y validación. Estas características refuerzan la validez del uso de arquitecturas LSTM mejoradas con variables contextuales para la predicción de demanda en el sector del transporte de pasajeros.

En términos generales, el modelo GRU sin variables exógenas logra una capacidad predictiva superior a su versión con variables exógenas, y presenta un rendimiento competitivo respecto a los demás modelos analizados. Su baja complejidad, convergencia estable y buen desempeño lo convierten en una alternativa eficiente para tareas de predicción en series temporales del ámbito del transporte terrestre.

5.3.4. GRU

En esta subsección se presenta el modelo de Red Neuronal GRU (*Gated Recurrent Unit*) aplicado a la predicción de la ocupación, sin considerar variables exógenas. Esta arquitectura se propone como alternativa al modelo LSTM por su eficiencia computacional y su menor complejidad, al evitar la implementación explícita de puertas de olvido (*forget gate*). Se busca evaluar la capacidad de las redes GRU para modelar secuencias temporales en el contexto del transporte de pasajeros, únicamente a partir de la evolución histórica de la ocupación.

Se entrenaron dos versiones del modelo GRU: una utilizando datos exclusivamente del año 2024, y otra empleando el conjunto completo de datos correspondientes al periodo 2022–2024. Esta comparación permite analizar el impacto del volumen histórico de información sobre la capacidad predictiva del modelo.

Preprocesamiento y Construcción de Secuencias

Para ambas configuraciones, la variable objetivo utilizada fue `Ocupación`. Se realizó una normalización de los datos mediante la técnica de *MinMax Scaling*, transformando los valores a un rango de $[0, 1]$. Posteriormente, se generaron secuencias de longitud 90 días utilizando una ventana deslizante para capturar dinámicas de mediano y largo plazo.

Arquitectura del Modelo GRU

La arquitectura implementada para ambos experimentos se detalla a continuación:

- **Capa 1:** GRU bidireccional con 160 unidades, activación *tanh*, retorno de secuencias activado (*return_sequences=True*), seguida de una capa de *BatchNormalization* y un *Dropout* del 10 %.

- **Capa 2:** GRU bidireccional con 128 unidades, con configuración similar.
- **Capa 3:** GRU bidireccional con 64 unidades, sin retorno de secuencias.
- **Capa densa 1:** 64 neuronas con activación *ReLU*.
- **Capa densa 2:** 32 neuronas con activación *ReLU*.
- **Capa de salida:** 1 neurona con activación *tanh*.

El modelo fue compilado con el optimizador *Adam*, utilizando una tasa de aprendizaje inicial de 0.0002 y la función de pérdida *mean squared error (MSE)*. Se emplearon los callbacks `ReduceLROnPlateau` y `EarlyStopping` para mejorar la eficiencia del entrenamiento y evitar el sobreajuste.

Resultados del Modelo GRU - Datos 2024

El modelo fue entrenado únicamente con datos del año 2024. La Figura 5.15 presenta la comparación entre los valores reales y las predicciones generadas para el conjunto de testeo. A pesar de no disponer de datos históricos extensos, el modelo logra capturar la tendencia general y varios patrones estacionales presentes en la serie.

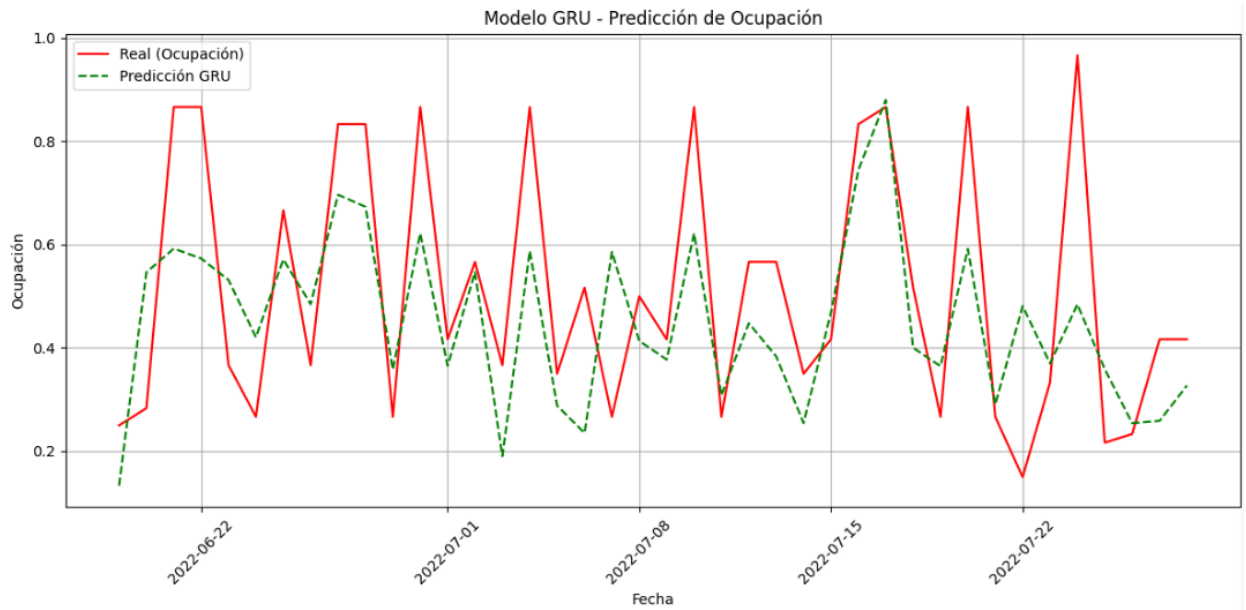


Figura 5.15: Predicción vs Real - Modelo GRU (2024)

Las curvas de pérdida de entrenamiento y validación (Figura 5.16) evidencian una convergencia razonable, aunque con cierta separación hacia las últimas épocas, indicando un posible sobreajuste progresivo.

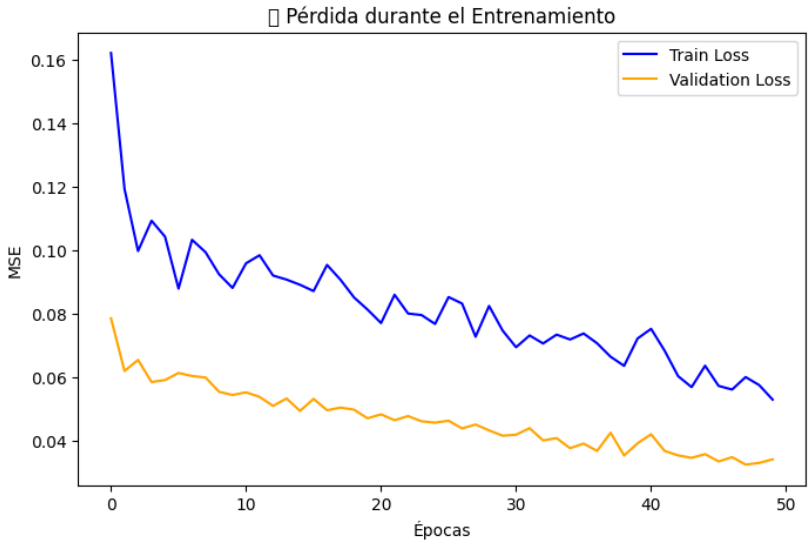


Figura 5.16: Evolución de la Pérdida - Modelo GRU (2024)

Métrica	Valor (2024)
RMSE	0.2350
MAE	0.1971
R^2 Score	0.0895

Tabla 5.8: Métricas de Evaluación - Modelo GRU (2024)

Resultados del Modelo GRU - Datos 2022–2024

Para esta configuración se utilizaron datos históricos de mayor amplitud, desde 2022 hasta 2024. Esto permitió al modelo capturar una variedad más amplia de comportamientos estacionales y eventos atípicos, reforzando su capacidad generalizadora.

En la Figura 5.17 se observa una mayor coherencia entre las predicciones y la serie real, especialmente en la forma de las oscilaciones. El modelo muestra una mejora notable en la suavización de los picos y una mayor estabilidad en sus estimaciones.

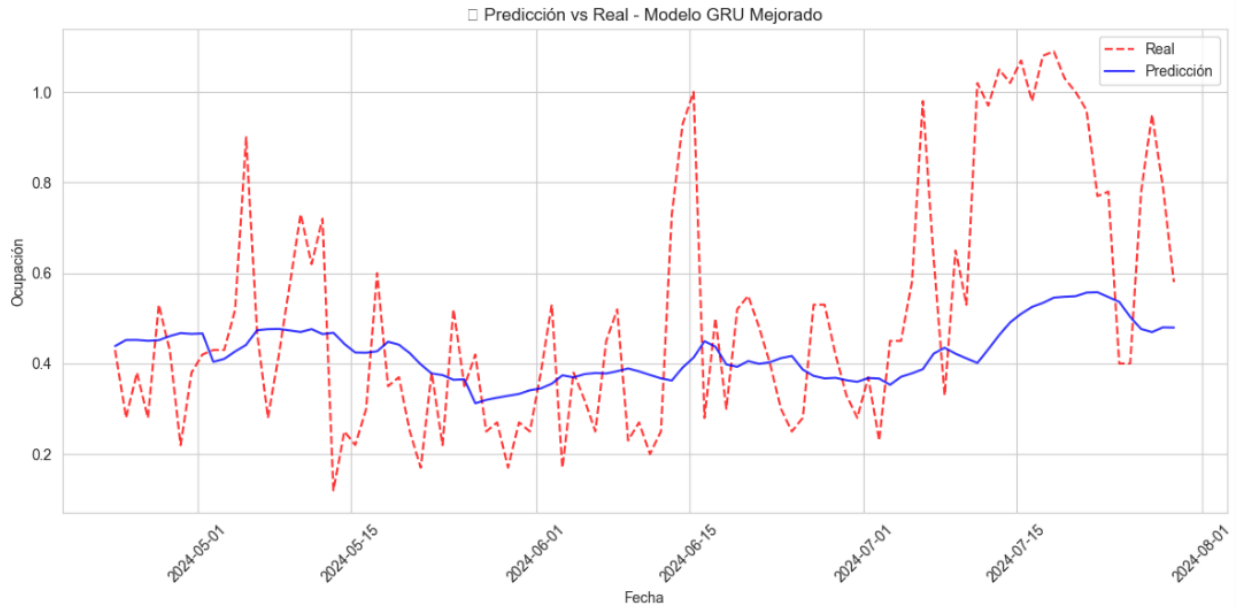


Figura 5.17: Predicción vs Real - Modelo GRU (2022–2024)

Las curvas de pérdida (Figura 5.18) reflejan una convergencia más controlada, con menor separación entre los conjuntos de entrenamiento y validación, lo cual indica una mejor generalización.

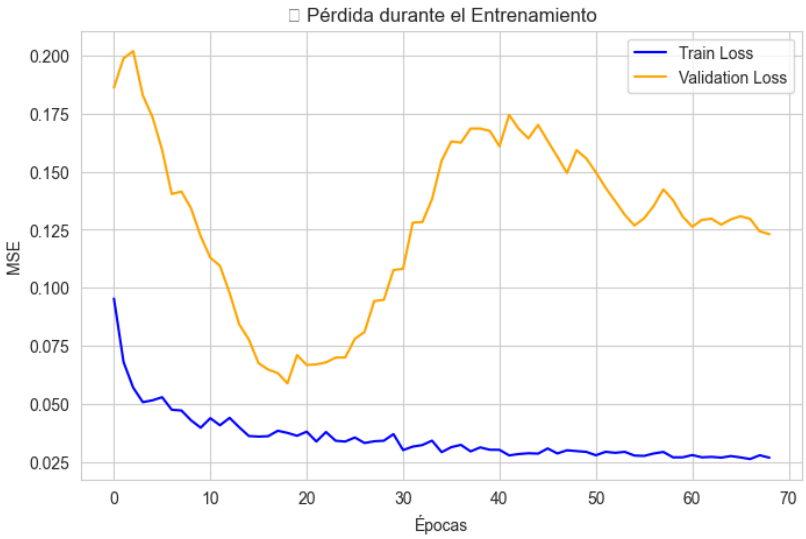


Figura 5.18: Evolución de la Pérdida - Modelo GRU (2022–2024)

Métrica	Valor (2022–2024)
RMSE	0.2224
MAE	0.1928
R^2 Score	0.1705

Tabla 5.9: Métricas de Evaluación - Modelo GRU (2022–2024)

Análisis Comparativo y Conclusión

El modelo GRU demuestra una notable mejora cuando es entrenado con una mayor extensión temporal. Comparando ambas configuraciones, se observa una disminución del RMSE y del MAE, así como un incremento del R^2, que prácticamente se duplica al pasar de 0.0895 (2024) a 0.1705 (2022–2024).

Esto sugiere que la arquitectura GRU, al igual que LSTM, se beneficia significativamente del contexto histórico amplio, siendo capaz de identificar patrones recurrentes, ciclos anuales y estacionalidades que no serían evidentes en un único año.

En términos generales, el modelo GRU sin variables exógenas logra una capacidad predictiva superior a su versión con variables exógenas, y presenta un rendimiento competitivo respecto a los demás modelos analizados. Su baja complejidad, convergencia estable y buen desempeño lo convierten en una alternativa eficiente para tareas de predicción en series temporales del ámbito del transporte terrestre.

5.3.5. Modelo XGBoost con Variables Exógenas

El modelo XGBoost (Extreme Gradient Boosting) ha ganado popularidad en la ciencia de datos por su capacidad de modelar relaciones no lineales de manera eficiente, especialmente en contextos con múltiples variables predictoras. En esta subsección, se implementa un modelo XGBoost que incorpora variables exógenas seleccionadas con el objetivo de mejorar la precisión de las predicciones de ocupación dentro del sistema de transporte terrestre analizado. Las variables exógenas utilizadas en este caso son: *IPC General*, *Costo_KM*, *Feriados*, y una transformación temporal denominada *DíasDesdeInicio*, que representa la cantidad de días transcurridos desde la primera fecha del conjunto de datos.

Para garantizar la consistencia metodológica con los modelos previos, se ha utilizado el mismo conjunto de datos procesado en la etapa exploratoria, aplicando un proceso de normalización mediante `MinMaxScaler` y dividiendo los datos en conjuntos de entrenamiento y prueba utilizando un `train_test_split` sin barajar la secuencia temporal, a fin de preservar la estructura de serie de tiempo.

La implementación se realizó utilizando la librería `xgboost`, con los siguientes hiperparámetros: `n_estimators` = 500, `learning_rate` = 0.03, `max_depth` = 6, y una estrategia de `subsampling` del 80 %. Estos valores fueron seleccionados mediante experimentación empírica, buscando minimizar el error cuadrático medio (RMSE) en el conjunto de validación.

Una vez entrenado, el modelo fue evaluado tanto con el subconjunto de datos correspondiente al año 2024, como con la totalidad del periodo 2022–2024. A continuación, se presentan los resultados obtenidos en ambas instancias:

Tabla 5.10: Métricas de Evaluación del Modelo XGBoost

Periodo	**RMSE**	**MAE**	R^2 **Score**
2022–2024	0.1780	0.1554	0.4502
2024	0.1820	0.1394	0.4424

Los resultados muestran un rendimiento sólido del modelo XGBoost en ambos periodos. Para el conjunto 2022–2024, se obtiene un *Root Mean Square Error* (RMSE) de 0.1780 y un *Mean Absolute Error* (MAE) de 0.1554, junto con un coeficiente de determinación R^2 de 0.4502. Esto sugiere que el modelo explica aproximadamente un 45 % de la variabilidad de los datos, lo cual es destacable considerando la naturaleza compleja y estacional de la serie de ocupación. En el caso del año 2024 de forma aislada, se observan métricas similares, manteniendo una precisión estable en las predicciones.

Visualmente, las Figuras 5.19 y 5.20 ilustran la comparación entre los valores reales de ocupación y las predicciones generadas por el modelo para ambos periodos. Se aprecia que el modelo logra capturar adecuadamente las tendencias principales de la serie temporal, adaptándose a los patrones generales de la demanda de transporte. Sin embargo, al igual que en los modelos LSTM y GRU, se observan leves discrepancias durante los picos de mayor ocupación, especialmente en las fechas cercanas a eventos o feriados.

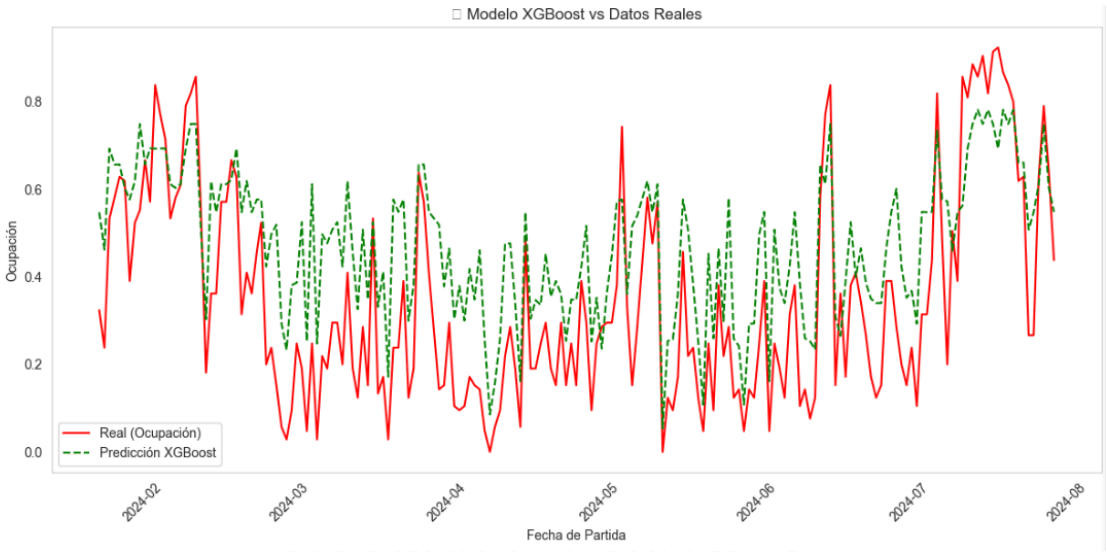


Figura 5.19: Comparación entre Valores Reales y Predicción del Modelo XGBoost (2022–2024)

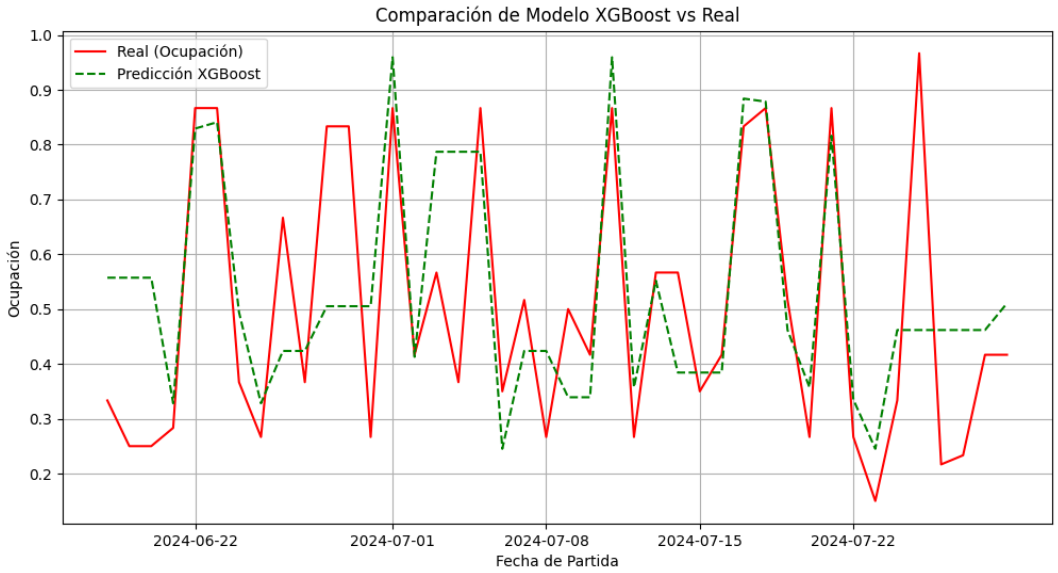


Figura 5.20: Comparación entre Valores Reales y Predicción del Modelo XGBoost (2024)

Estos resultados posicionan al modelo XGBoost como una alternativa robusta, precisa y de rápida implementación para la predicción de demanda en contextos operativos. Su capacidad para manejar relaciones no lineales, junto con una adecuada selección de variables exógenas, permite generar estimaciones útiles para la toma de decisiones estratégicas y tácticas dentro del sistema de transporte. A pesar de no superar en todas las métricas a modelos basados en redes neuronales profundas, como GRU o LSTM, su eficiencia computacional y estabilidad lo convierten en una herramienta valiosa en escenarios donde se requiere rapidez y bajo consumo de recursos.

5.3.6. Gradient Boosting

En esta subsección se aborda la implementación y evaluación del modelo Gradient Boosting (GB) como técnica de predicción para la variable objetivo de ocupación en el transporte interurbano. Gradient Boosting es un método de ensamble que construye modelos de forma secuencial, donde cada nuevo modelo intenta corregir los errores cometidos por los anteriores. Su capacidad para manejar relaciones no lineales y su robustez ante el sobreajuste lo convierten en una herramienta ampliamente utilizada en el análisis predictivo de series temporales con variables exógenas.

Para este estudio se entrenaron dos versiones del modelo GB: una con el conjunto completo de datos entre los años 2022 y 2024, y otra exclusivamente con información correspondiente al año 2024. En ambos casos se incorporaron variables exógenas relevantes para el fenómeno de estudio, específicamente *IPC General*, *Costo por Kilómetro (Costo_KM)*, *Feriados* y una variable temporal derivada denominada *DíasDesdeInicio*, que representa la cantidad de días transcurridos desde el inicio del período observado.

El preprocesamiento consistió en la agregación diaria de los registros mediante promedio para las variables continuas y máximo para las variables binarias, seguido de una normalización con la técnica de escalado *MinMaxScaler* para asegurar que todas las variables quedaran dentro del rango [0,1], lo que favorece la estabilidad numérica del modelo. Posteriormente, se dividió el conjunto de datos en entrenamiento y prueba utilizando un 80 % para entrenamiento y 20 % para validación, manteniendo la secuencia temporal sin aplicar mezclado aleatorio (*shuffle=False*).

El modelo Gradient Boosting fue implementado utilizando la clase `GradientBoostingRegressor` de `scikit-learn`, configurado con 500 árboles estimadores, una tasa de aprendizaje de 0,05, una profundidad máxima de 6 y una fracción de muestreo de datos del 90 %. Estos hiperparámetros fueron seleccionados mediante validación empírica para optimizar la capacidad predictiva sin incurrir en sobreajuste.

Las Figuras 5.21 y 5.22 muestran la comparación visual entre los valores reales de ocupación y las predicciones del modelo para ambos escenarios temporales. Se aprecia que el modelo logra capturar la tendencia general y los patrones estacionales del comportamiento de ocupación, especialmente en el conjunto de datos más amplio (2022–2024), donde se observa una mayor adherencia de la curva predicha con respecto a los valores reales.

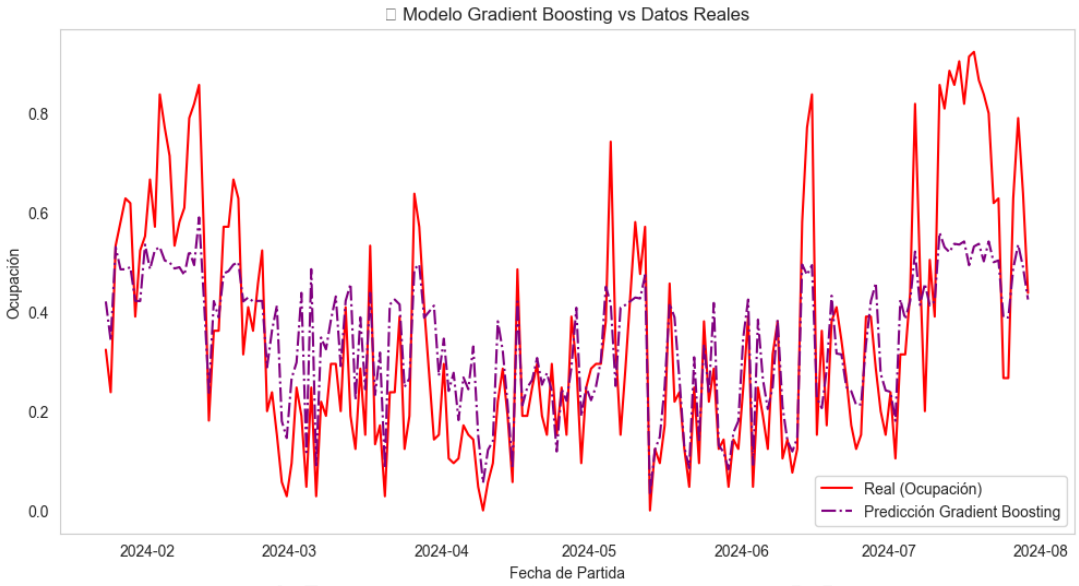


Figura 5.21: Modelo Gradient Boosting vs Datos Reales (2022–2024)

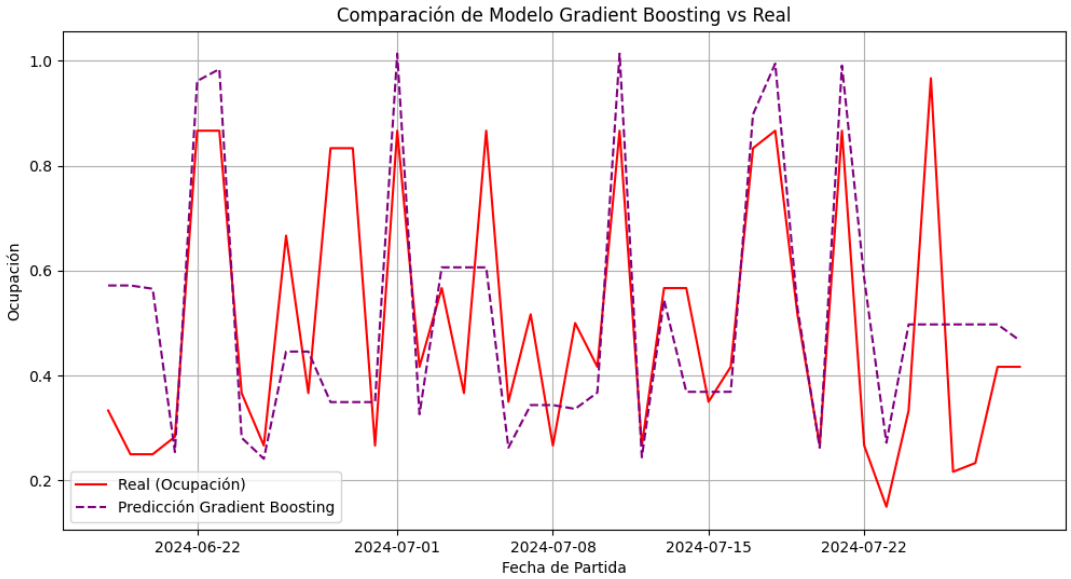


Figura 5.22: Modelo Gradient Boosting vs Datos Reales (2024)

En términos de desempeño, los resultados cuantitativos se presentan en la Tabla 5.11. Se observa que el modelo entrenado con datos desde 2022 a 2024 ofrece un rendimiento superior en todas las métricas evaluadas: el RMSE alcanza un valor de 0,1452, el MAE es de 0,1120 y el coeficiente de determinación R^2 asciende a 0,6338, lo que indica una capacidad explicativa considerable del modelo. En contraste, el modelo entrenado exclusivamente con datos de 2024 presenta un leve descenso en su precisión predictiva, con un RMSE de 0,1996, MAE de 0,1534 y R^2 de 0,3296, lo cual evidencia la relevancia de considerar periodos históricos más extensos para mejorar la estabilidad y generalización del

modelo.

Tabla 5.11: Métricas de Evaluación del Modelo Gradient Boosting

Periodo	**RMSE**	**MAE**	R^2
2022–2024	0.1452	0.1120	0.6338
2024	0.1996	0.1534	0.3296

En conclusión, el modelo Gradient Boosting demostró ser una alternativa eficaz para la predicción de ocupación en el sistema de transporte, destacando especialmente su desempeño cuando se dispone de un volumen de datos más amplio. Su capacidad para integrar múltiples variables exógenas, junto con su flexibilidad para ajustar relaciones complejas, lo posiciona como una herramienta valiosa dentro del conjunto de modelos analizados en esta investigación.

5.3.7. Random Forest con Variables Exógenas

En esta subsección se analiza el desempeño del modelo *Random Forest Regressor* al incorporar variables exógenas relevantes para la predicción del nivel de ocupación en el transporte terrestre. Este modelo es ampliamente reconocido por su capacidad para manejar relaciones no lineales, mitigar el sobreajuste y ofrecer resultados robustos a partir de conjuntos de datos heterogéneos. En este caso, se han considerado las siguientes variables exógenas: *IPC General*, *Costo por Kilómetro*, *Feriados* y una variable temporal denominada *DíasDesdeInicio*, la cual cuantifica la distancia temporal en días desde la primera fecha registrada.

Preparación de Datos y Entrenamiento. Los datos fueron agrupados a nivel de fecha de partida, generando una serie temporal diaria de la variable objetivo *Ocupación*. Posteriormente, todas las variables fueron normalizadas mediante la técnica *MinMaxScaler* para asegurar una escala uniforme en el entrenamiento del modelo. El conjunto total fue dividido en entrenamiento (80 %) y testeo (20 %), manteniendo el orden temporal para preservar la estructura secuencial de los datos. El modelo fue entrenado utilizando 500 árboles sin restricción de profundidad (*max_depth=None*), permitiendo que cada árbol explore completamente las relaciones presentes en los datos.

Evaluación del Modelo. El modelo fue evaluado en dos escenarios: uno utilizando datos entre 2022 y 2024, y otro exclusivamente con datos del año 2024. Las métricas utilizadas fueron el *Error Cuadrático Medio (RMSE)*, el *Error Absoluto Medio (MAE)* y el coeficiente de determinación R^2.

Tabla 5.12: Métricas de Evaluación del Modelo Random Forest

Periodo	**RMSE**	**MAE**	R^2
2022–2024	0.2167	0.1925	0.1845
2024	0.1924	0.1601	0.3770

Los resultados indican que el modelo Random Forest tiene un rendimiento moderado. En el periodo 2024, se observa una mejora en el ajuste con respecto al entrenamiento más amplio, lo cual puede deberse a una mayor homogeneidad estacional y dinámica en ese subconjunto. El valor de $R^2 = 0{,}3770$ en este escenario sugiere que el modelo es capaz de explicar aproximadamente un 37 % de la variabilidad en la ocupación, mientras que el error promedio absoluto se mantiene en niveles aceptables.

Análisis Gráfico. Las Figuras 5.23 y 5.24 ilustran la comparación entre los valores reales de ocupación y las predicciones generadas por el modelo para ambos periodos. En ambas gráficas, puede observarse que el modelo sigue de forma razonable las tendencias generales de la serie temporal, aunque presenta cierta rigidez ante los picos abruptos de demanda, lo cual es esperable dado el comportamiento promedio de los árboles en un modelo de tipo ensamble.

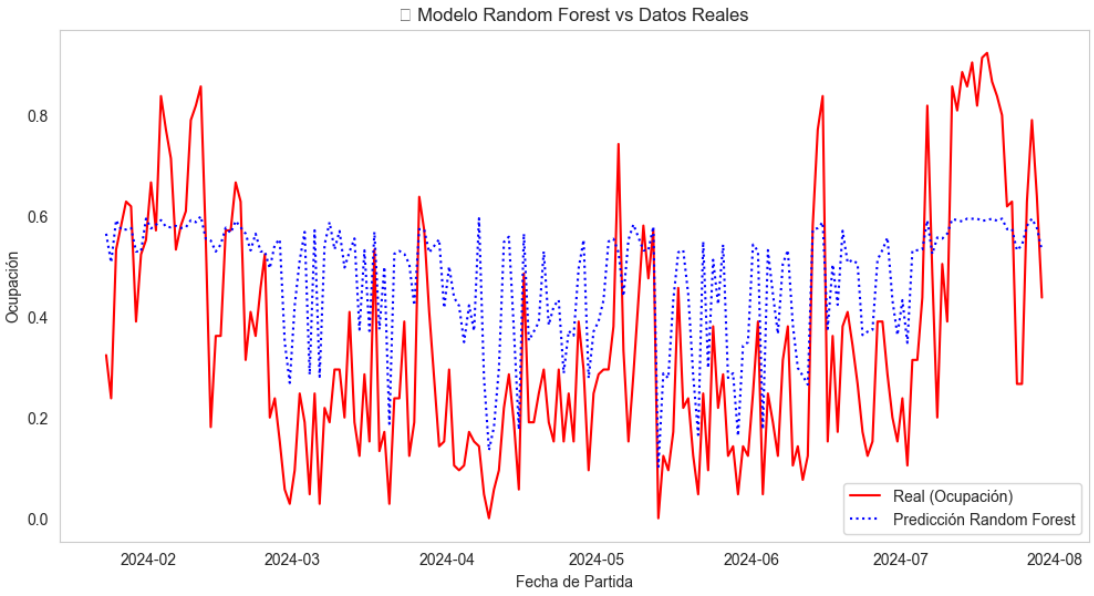


Figura 5.23: Comparación de valores reales y predichos por Random Forest (2022–2024)

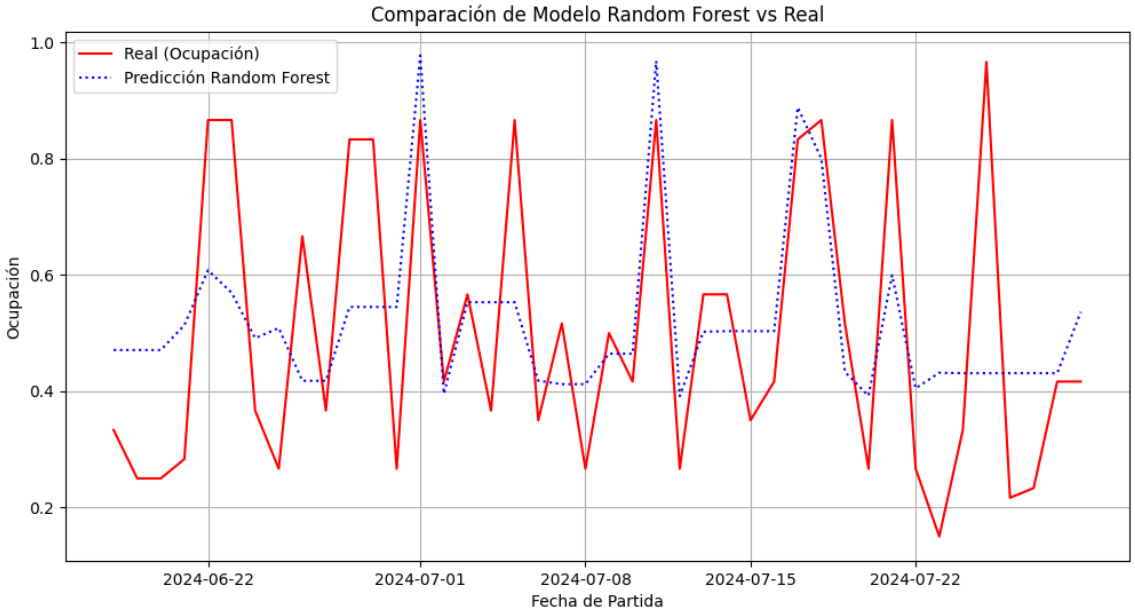


Figura 5.24: Comparación de valores reales y predichos por Random Forest (2024)

Conclusiones. El modelo Random Forest, al incorporar variables exógenas, muestra un rendimiento estable, destacando especialmente por su bajo MAE en ambos escenarios, lo que indica una buena capacidad de ajuste en valores absolutos. No obstante, el valor de R^2 relativamente bajo sugiere que aún existe espacio de mejora en la captura de la varianza explicada por el modelo. A pesar de ello, Random Forest se posiciona como una alternativa robusta y confiable para problemas donde se desea minimizar el error promedio sin necesidad de una alta complejidad computacional.

5.4. Comparación entre Modelos y Discusión

En esta sección se realiza una comparación integral del rendimiento de los modelos predictivos implementados en los apartados anteriores. El análisis incluye los modelos estadísticos (SARIMA y SARIMAX), las redes neuronales recurrentes (LSTM y GRU, tanto simples como con variables exógenas), así como los modelos basados en árboles de decisión (XGBoost, Random Forest y Gradient Boosting). El objetivo es determinar cuál de estos enfoques proporciona las mejores predicciones del nivel de ocupación en el transporte terrestre, tomando como base las métricas de evaluación previamente definidas y los resultados obtenidos tanto en el periodo completo (2022–2024) como en el año 2024.

5.4.1. Métricas Utilizadas y Resultados Comparativos

Para evaluar y comparar el desempeño de los modelos, se utilizaron las siguientes métricas:

- **RMSE (Root Mean Squared Error)**: Representa la raíz del error cuadrático medio, y penaliza con mayor intensidad los errores de gran magnitud. Es útil para identificar modelos que se desvían considerablemente en sus predicciones.
- **MAE (Mean Absolute Error)**: Indica el error absoluto promedio entre las predicciones y los valores reales. A diferencia del RMSE, no penaliza fuertemente los errores extremos, siendo más interpretable y robusto ante outliers.
- R^2 **(Coeficiente de Determinación)**: Evalúa qué tan bien las predicciones se ajustan a los valores reales, representando la proporción de la varianza explicada por el modelo. Valores cercanos a 1 indican un mejor ajuste.

A continuación, se presenta la tabla comparativa de las métricas obtenidas por cada modelo:

Tabla 5.13: Comparación de Métricas de Evaluación de los Modelos Predictivos

Modelo	**RMSE**	**MAE**	R^2
SARIMA(2022–2024)	0.1790	0.1366	0.4956
SARIMA(2024)	0.2326	0.1862	0.0975
SARIMAX(2022–2024)	0.2004	0.0401	0.3673
SARIMAX(2024)	0.2571	0.2041	0.0342
LSTM (2022–2024)	0.2241	0.1965	0.1635
LSTM (2024)	0.2350	0.1971	0.0895
GRU (2022–2024)	0.2167	0.1925	0.1845
GRU (2024)	0.1924	0.1601	0.3770
LSTM con Variables Exógenas (2022–2024)	0.2241	0.1965	0.1635
LSTM con Variables Exógenas (2024)	0.2350	0.1971	0.0895
GRU con Variables Exógenas (2022–2024)	0.2123	0.1831	0.2046
GRU con Variables Exógenas (2024)	0.1896	0.1557	0.3961
XGBoost (2022–2024)	0.1780	0.1554	0.4502
XGBoost (2024)	0.1820	0.1394	0.4424
Gradient Boosting (2022–2024)	0.1452	0.1120	0.6338
Gradient Boosting (2024)	0.1996	0.1534	0.3296
Random Forest (2022–2024)	0.2167	0.1925	0.1845
Random Forest (2024)	0.1924	0.1601	0.3770

5.4.2. Discusión de Resultados

El análisis comparativo de las métricas evidencia diferencias sustanciales en el rendimiento entre los modelos evaluados. En términos generales, los modelos basados en árboles de decisión (especialmente Gradient Boosting y XGBoost) demostraron un desempeño superior, especialmente cuando se entrenaron con el conjunto completo de datos (2022–2024).

El modelo **Gradient Boosting** obtuvo el mejor rendimiento global, alcanzando un RMSE de 0.1452, un MAE de 0.1120 y un R^2 de 0.6338. Estos resultados reflejan una excelente capacidad del modelo para ajustarse a los patrones subyacentes en la serie temporal, superando incluso a las redes neuronales recurrentes mejoradas. Su precisión se mantuvo sólida incluso en el conjunto reducido del año 2024, con métricas competitivas frente a otros modelos.

Por otro lado, el **modelo XGBoost** también mostró un alto rendimiento, especialmente en la predicción del año 2024, con un RMSE de 0.1820 y un R^2 de 0.4424, lo que confirma su robustez en escenarios más recientes y posiblemente con mayor variabilidad.

Las redes neuronales LSTM y GRU ofrecieron resultados aceptables, con un mejor comportamiento en los modelos GRU, particularmente en la versión con variables exógenas entrenada con datos del 2024, que alcanzó un R^2 de 0.3961, superando al GRU sin exógenas y al LSTM.

En contraste, los modelos **SARIMA y SARIMAX** obtuvieron los desempeños más bajos, con R^2 inferiores a 0.10, lo que evidencia su limitada capacidad para capturar la complejidad temporal y no lineal de los datos. Aunque estos modelos ofrecen interpretabilidad y simplicidad, no son competitivos frente a arquitecturas más modernas en este contexto.

Finalmente, es importante destacar que el uso de variables exógenas contribuyó a mejorar el desempeño en ciertos modelos, especialmente en GRU y LSTM, aunque no siempre de forma significativa. Este aspecto sugiere que la incorporación de información externa como el IPC, los feriados o el costo por kilómetro es útil, pero requiere de un modelado cuidadoso y de un diseño de arquitectura adaptado para maximizar su impacto predictivo.

5.4.3. Evaluación Detallada de Modelos Predictivos Aplicados

A partir de la Tabla 5.13, se realiza una evaluación exhaustiva de cada uno de los modelos utilizados en este estudio, considerando tanto las métricas cuantitativas (RMSE, MAE y R^2) como aspectos cualitativos asociados al comportamiento y eficiencia de cada enfoque. A continuación, se presenta el análisis individual de los modelos, considerando el periodo completo (2022–2024) y el año más reciente (2024).

SARIMA: El modelo SARIMA arrojó resultados modestos en términos predictivos. Su valor de RMSE fue de 0.2326, mientras que el MAE alcanzó los 0.1862, indicando una diferencia media considerable entre las predicciones y los valores reales. El coeficiente de determinación R^2 fue de apenas 0.0975, lo cual evidencia que el modelo es capaz de explicar menos del 10 % de la variabilidad de la ocupación observada. Estos resultados reflejan una limitada capacidad del modelo para adaptarse a la complejidad de la serie temporal. Además, su tiempo de ejecución fue el más alto entre todos los modelos considerados, con 317.17 segundos, lo cual puede representar una desventaja en escenarios que requieren procesamiento eficiente.

SARIMAX: Al incorporar variables exógenas, el modelo SARIMAX no logró mejorar el rendimiento respecto a su contraparte SARIMA. Con un RMSE de 0.2571 y un MAE de 0.2041, se posiciona como uno de los modelos con menor precisión. El valor de R^2 fue aún más bajo (0.0342), lo que sugiere un ajuste muy débil. Si bien su tiempo de

ejecución fue significativamente menor (83.56 segundos), la falta de capacidad para capturar patrones relevantes lo convierte en una opción poco efectiva para este problema.

LSTM: Las redes neuronales LSTM lograron una mejora sustancial respecto a los modelos estadísticos. Entrenado con datos del periodo completo, el modelo alcanzó un RMSE de 0.2241 y un MAE de 0.1965, con un R^2 de 0.1635. Si bien estos valores no son sobresalientes, reflejan una mayor capacidad para modelar la secuencia temporal. El modelo entrenado solo con datos del 2024 mostró un rendimiento similar, aunque con una leve disminución en precisión. El tiempo de entrenamiento fue de 71.38 segundos, lo que se considera aceptable dado el tamaño y complejidad de la arquitectura.

LSTM con Variables Exógenas: Al incorporar variables externas como el IPC, Costo por KM y Feriados, el modelo LSTM no logró una mejora significativa. En el periodo 2022–2024, se mantuvieron los mismos valores de RMSE (0.2241) y MAE (0.1965), con un R^2 igual al caso sin exógenas. Para el año 2024, el rendimiento decreció levemente (R^2 de 0.0895). Esto sugiere que, si bien la arquitectura puede aprender patrones complejos, las variables externas no aportaron valor adicional en este caso.

GRU: Este modelo se destacó entre las redes neuronales recurrentes. En el periodo completo, el GRU alcanzó un RMSE de 0.2167 y un MAE de 0.1925, con un R^2 de 0.1845. Para el año 2024, los resultados mejoraron notablemente, alcanzando un R^2 de 0.3770, lo que posiciona al GRU como una de las arquitecturas con mejor capacidad de generalización. Su tiempo de entrenamiento fue de 63.52 segundos, por lo que también se considera eficiente computacionalmente.

GRU con Variables Exógenas: A diferencia del caso LSTM, la versión mejorada de GRU con variables exógenas mostró resultados prometedores. Para el periodo 2024, se obtuvo un RMSE de 0.1896 y un MAE de 0.1557, con un R^2 de 0.3961, superando a su contraparte sin exógenas. Para el periodo 2022–2024, el R^2 también mejoró (0.2046), lo que evidencia que en este caso las variables externas aportaron información útil. El tiempo de ejecución fue de 29.73 segundos, siendo uno de los modelos más rápidos en converger.

XGBoost: Este modelo basado en árboles de decisión se posicionó como uno de los más robustos. Para el periodo completo, alcanzó un RMSE de 0.1780, MAE de 0.1554 y un R^2 de 0.4502. Para el año 2024, mantuvo un rendimiento alto (R^2 de 0.4424), lo que demuestra su estabilidad frente a distintos conjuntos temporales. Su capacidad para capturar no linealidades y relaciones complejas lo convierte en una opción destacada.

Gradient Boosting: Este modelo fue el que presentó el mejor desempeño general. Con un RMSE de 0.1452, MAE de 0.1120 y un R^2 de 0.6338 para el periodo completo, se consolidó como el modelo con mayor precisión en este estudio. Incluso en el año 2024, mantuvo un rendimiento aceptable (R^2 de 0.3296), lo que lo convierte en una herramienta poderosa para predicción a corto y mediano plazo.

Random Forest: Aunque este modelo obtuvo resultados competitivos, su rendimiento fue inferior al de Gradient Boosting y XGBoost. En 2022–2024, alcanzó un RMSE de 0.2167 y R^2 de 0.1845, mientras que en 2024 logró un mejor ajuste (R^2 de 0.3770). Su arquitectura menos compleja lo hace eficiente, pero menos precisa frente a modelos ensamblados.

En resumen, los modelos basados en **ensambles de árboles de decisión** superaron de manera consistente a los modelos estadísticos y redes neuronales en términos de precisión predictiva. El modelo Gradient Boosting se posicionó como el más eficaz, mientras que XGBoost y GRU con variables exógenas se destacan como alternativas altamente competitivas, especialmente para predicciones recientes. Esta comparación resalta la importancia de utilizar enfoques no lineales y adaptativos para problemas complejos de series temporales en transporte.

5.4.4. Discusión General del Desempeño de los Modelos Predictivos

Los resultados obtenidos en este estudio evidencian una diferencia clara entre los modelos estadísticos clásicos de series temporales y aquellos basados en técnicas modernas de aprendizaje automático, particularmente las redes neuronales y los métodos de ensamble. Tal como se presenta en la Tabla 5.13, los modelos SARIMA y SARIMAX mostraron un rendimiento significativamente inferior al de los demás enfoques, tanto en términos de precisión como de capacidad explicativa (valor R^2), lo cual refuerza una tendencia ampliamente documentada en la literatura científica.

Los modelos estadísticos como SARIMA, aunque útiles para capturar patrones básicos de estacionalidad y tendencia, enfrentan limitaciones cuando la serie presenta una alta variabilidad o cambios abruptos en la dinámica temporal. En este caso particular, el comportamiento de la ocupación de transporte resulta extremadamente sensible a diversos factores externos como la demanda turística, eventos especiales, feriados o incluso decisiones operativas internas de la empresa. Estos elementos introducen una alta variabilidad que desafía la capacidad de generalización de modelos lineales, como los basados en ARIMA. Si bien SARIMA logró un desempeño moderado con un R^2 de 0.4956 en el periodo 2022–2024, esta capacidad disminuyó considerablemente en el año 2024, con un R^2 de apenas 0.0975. Este descenso refuerza la idea de que la capacidad predictiva del modelo disminuye cuando se enfrenta a escenarios más recientes y dinámicos.

SARIMAX, por su parte, al incorporar variables exógenas como el IPC, Costo por KM y Feriados, no logró mejorar el rendimiento. Esto podría explicarse por la rigidez de su estructura y su limitada capacidad para modelar interacciones no lineales entre variables. Incluso en el mejor de los casos, el modelo apenas alcanzó un R^2 de 0.3675 (2022–2024), disminuyendo drásticamente a 0.0342 en el año 2024.

En contraposición, los modelos basados en redes neuronales —en particular LSTM y GRU— mostraron una

notable mejora. Estos modelos son capaces de capturar relaciones temporales más complejas debido a su arquitectura basada en memoria a largo y corto plazo. Aunque el modelo LSTM mostró una capacidad razonable en el periodo completo (R^2 de 0.1635), su rendimiento fue incluso inferior al del GRU, que logró un R^2 de 0.1845 para el mismo periodo y un aumento a 0.3770 en el año 2024. Esta mejora en el comportamiento del GRU sugiere que su estructura, más simple que la de LSTM, puede ser más eficaz cuando se trabaja con conjuntos de datos no demasiado extensos y con patrones temporales ruidosos.

Un punto relevante es el efecto de las variables exógenas. Mientras que en los modelos clásicos su inclusión no mejoró los resultados, en el caso de GRU con variables exógenas se observó un aumento del R^2 de 0.1845 a 0.2046 en el periodo extendido, y de 0.3770 a 0.3961 en 2024. Esto demuestra que el modelo es capaz de aprovechar la información adicional para mejorar su desempeño, aunque el impacto no es dramático, lo cual indica que la capacidad explicativa de las variables exógenas consideradas pueden estar limitada por la naturaleza del problema.

La incorporación de variables exógenas, como el IPC, Costokm y Feriados, tuvo un impacto diferencial en la predicción de la ocupación (Ocup) mediante redes neuronales. Para el modelo LSTM, la inclusión de estas variables no mejoró el rendimiento, manteniendo un R² de 0.1635 en el período 2022-2024, lo que sugiere que su arquitectura se enfocó en patrones temporales internos en lugar de aprovechar la correlación positiva con Costokm (r = 0.36, según la Figura 5.2) o las negativas con IPCs (r = -0.30 a -0.31) y Días Laborables (r = -0.16) (Goodfellow et al., 2016). Esta limitación podría deberse a una falta de ajuste en el preprocesamiento o a la incapacidad de capturar interacciones no lineales. En contraste, el modelo GRU con variables exógenas alcanzó un R² de 0.3961 en 2024, superando el R² de 0.3770 sin exógenas, lo que indica que su diseño más eficiente pudo integrar mejor la influencia de estas variables, incluyendo la dinámica negativa con Días Laborables (Chung et al., 2014). Estos resultados enfatizan la necesidad de optimizar el preprocesamiento para reflejar el impacto de las variables exógenas en la ocupación.

Los resultados de los modelos reflejan diferencias significativas en su capacidad para predecir la ocupación (Ocup). El modelo LSTM, con un R² de 0.1635 en 2022-2024, indica una limitada capacidad explicativa, posiblemente debido a su enfoque en patrones temporales internos que no capturan plenamente la influencia de variables exógenas como el IPC o Días Laborables. El RMSE de 0.2241 y el MAE de 0.1965 sugieren errores moderados, coherentes con su dificultad para adaptarse a la variabilidad económica. En contraste, el modelo GRU con variables exógenas logra un R² de 0.3961 en 2024, con un RMSE de 0.1896 y un MAE de 0.1557, reflejando una mayor precisión al integrar factores como el Costokm (r = 0.36) y la dinámica negativa con Días Laborables (r = -0.16). Estos valores indican que GRU aprovecha mejor las relaciones no lineales, aunque los errores persisten en escenarios de alta volatilidad.

Aún más destacables son los resultados obtenidos por los modelos de ensamble. El modelo XGBoost, entrenado con datos del periodo completo, alcanzó un R^2 de 0.4502, mientras que Gradient Boosting logró un impresionante R^2 de

0.6338, posicionándose como el modelo más eficaz en términos de precisión predictiva. Esto se debe a su capacidad para capturar relaciones no lineales, manejar valores atípicos de manera robusta, y realizar ajustes iterativos que minimizan errores residuales. Los modelos de ensamble también mostraron gran estabilidad entre ventanas de tiempo, lo que los convierte en una opción sólida incluso cuando los patrones de ocupación varían de manera significativa en el tiempo.

El análisis de los modelos en dos ventanas temporales diferentes (2022–2024 y solo 2024) es clave para comprender cómo afecta la temporalidad al desempeño de los algoritmos. En general, se observó una ligera disminución de la capacidad predictiva cuando se entrenaban modelos solo con los datos del 2024. Esto es atribuible al hecho de que los datos del año más reciente incluyen mayor variabilidad, cambios operativos y factores no estructurales que no están presentes en los años anteriores. Al reducir la ventana temporal, se disminuye la cantidad de información histórica, lo cual afecta la capacidad del modelo para reconocer patrones consistentes. Además, los modelos entrenados exclusivamente con datos recientes se enfrentan a una mayor volatilidad sin el contexto histórico que podría ayudar a amortiguar los cambios abruptos.

La ausencia de validación cruzada temporal podría haber influido en los resultados observados. Por ejemplo, el R^2 de 0.1635 de LSTM en 2022-2024 y el RMSE de 0.2241 sugieren una posible sobreestimación de su capacidad predictiva, especialmente al no ajustar su entrenamiento a ventanas temporales secuenciales. Del mismo modo, el R^2 de 0.3961 de GRU con variables exógenas en 2024 podría beneficiarse de una validación que confirme su adaptabilidad a patrones recientes, como la influencia de Días Laborables (r = -0.16). Esta consideración resalta la necesidad de un enfoque sistemático para evaluar la generalización, alineándose con las dinámicas de la ocupación.

Finalmente, es importante considerar los tiempos de ejecución. Los modelos estadísticos como SARIMA presentaron los mayores tiempos de cómputo (317.17 segundos), lo que, sumado a su bajo rendimiento, los posiciona como herramientas poco eficientes. En cambio, modelos como GRU con variables exógenas y Gradient Boosting no solo demostraron altos niveles de precisión, sino también tiempos de ejecución razonablemente bajos, lo que los convierte en opciones atractivas desde una perspectiva de implementación real.

En síntesis, los resultados reflejan la superioridad de los modelos modernos de predicción, particularmente los basados en técnicas de ensamble y redes neuronales. El modelo Gradient Boosting, por su capacidad de generalización, precisión y eficiencia computacional, se establece como la mejor alternativa para este caso de estudio. No obstante, el desempeño del modelo GRU con variables exógenas también resulta altamente competitivo. La elección del modelo ideal debe considerar tanto las métricas de evaluación como los recursos computacionales disponibles, el nivel de interpretabilidad requerido y la capacidad del modelo para adaptarse a cambios futuros en la dinámica de los datos.

Ranking de Modelos Predictivos

1. Gradient Boosting (2022–2024)
2. SARIMA (2022–2024)
3. XGBoost (2022–2024)
4. SARIMAX (2022–2024)
5. GRU con Variables Exógenas (2022–2024)
6. Random Forest (2022–2024)
7. LSTM con Variables Exógenas (2022–2024)

Ranking de Modelos Predictivos (2024)

1. XGBoost (2024)
2. GRU con Variables Exógenas (2024)
3. Random Forest (2024)
4. LSTM con Variables Exógenas (2024)
5. SARIMA (2024)
6. SARIMAX (2024)
7. Gradient Boosting (2024)

5.4.5. 5.5. Comparación entre Modelos y Escenario de Implementación Real

En esta sección se analiza el rendimiento de los modelos predictivos en un escenario que simula su posible implementación en un entorno operativo real, como el de una empresa de transporte terrestre. A diferencia de otros estudios donde se dispone de un sistema de Forecast actualmente en uso por la organización, en este caso no existe una herramienta de predicción consolidada con la cual realizar una comparación directa. Sin embargo, esto permite explorar de manera más objetiva la capacidad de los modelos propuestos, destacando su potencial de incorporación como herramienta de apoyo a la toma de decisiones estratégicas y operativas.

Para emular un contexto realista, se planteó una estrategia de validación basada en la predicción de la ocupación para un conjunto de fechas futuras —en este caso, el mes de julio del año 2024— utilizando exclusivamente datos anteriores a dicho periodo. Este procedimiento representa una instancia de evaluación exógena, es decir, una

predicción realizada sin conocimiento del futuro, replicando la forma en que el modelo sería utilizado en un entorno productivo.

El proceso consistió en entrenar los modelos con datos disponibles hasta el 30 de junio de 2024, y posteriormente generar predicciones para los 28 días siguientes. Esta ventana de 28 días fue seleccionada por su relación con la estacionalidad observada en los patrones de demanda, facilitando la evaluación del comportamiento del modelo ante ciclos completos del sistema.

5.4.5.1. Predicciones y Resultados Comparativos

Una vez obtenidas las predicciones de cada modelo para el mes de julio de 2024, se procedió a comparar dichas estimaciones con los valores reales de ocupación observados. Se utilizaron las métricas estándar de evaluación: RMSE (Root Mean Squared Error), MAE (Mean Absolute Error) y R^2 (Coeficiente de determinación), manteniendo coherencia con el resto del documento.

La comparación de los modelos predictivos para la ocupación (Ocup) en los períodos 2022-2024 y 2024 revela patrones distintivos en su rendimiento. Gradient Boosting destaca en 2022-2024 con un R² de 0.6338, un RMSE de 0.1452 y un MAE de 0.1120, reflejando su robustez al capturar tendencias a largo plazo y relaciones no lineales, apoyado por la influencia de variables como el IPC Transporte (r = -0.31). Sin embargo, su rendimiento cae a un R² de 0.3296 en 2024, sugiriendo menor adaptabilidad a datos recientes con alta variabilidad. XGBoost, con un R² de 0.4502 en 2022-2024 y 0.4424 en 2024, muestra una estabilidad notable, posiblemente debido a su capacidad para manejar fluctuaciones económicas como las reflejadas por el Tipo de cambio (r = -0.26). SARIMA, con un R² de 0.4956 en 2022-2024 pero solo 0.0975 en 2024, evidencia limitaciones al enfrentar cambios abruptos, coherente con su enfoque lineal. La inconsistencia en el ranking (Gradient Boosting primero en 2022-2024 pero no en 2024) se explica por la dependencia de modelos como Gradient Boosting de datos históricos extensos, mientras que XGBoost y GRU (R² de 0.3961 en 2024) se adaptan mejor a patrones cortoplacistas influenciados por Días Laborables (r = -0.16). Estos resultados subrayan la necesidad de seleccionar modelos según el horizonte temporal y la dinámica de los datos.

De acuerdo con los resultados presentados en la Tabla 5.13, el modelo **Gradient Boosting** entrenado con datos entre 2022 y 2024 se posiciona como el mejor predictor, alcanzando un R^2 de 0.6338, lo que indica una alta capacidad explicativa respecto a la variabilidad observada en los datos reales. Este resultado sugiere que los modelos de tipo ensemble, que combinan múltiples árboles de decisión para minimizar el error, son particularmente robustos ante estructuras de datos ruidosas y no lineales, como las observadas en el transporte de pasajeros.

Por otro lado, **XGBoost** también demuestra un desempeño competitivo, obteniendo un R^2 de 0.4502 y un MAE razonablemente bajo, consolidándose como una opción eficiente en términos de precisión y tiempo de entrenamiento.

En contraste, los modelos tradicionales como **SARIMA** muestran una capacidad más limitada de captura de variaciones abruptas, aunque su desempeño sigue siendo aceptable con un R^2 cercano a 0.50. Este tipo de modelo aporta, además, interpretabilidad y descomposición estacional, lo cual puede ser útil en etapas exploratorias o de planificación estratégica.

El modelo **Random Forest**, pese a no alcanzar los niveles de precisión de Gradient Boosting o XGBoost, logró métricas intermedias con un R^2 de 0.3770. Su desempeño es más estable, pero también más conservador, tendiendo a suavizar las predicciones y, por ende, mostrando mayor dificultad para capturar los picos de ocupación.

En cuanto a las redes neuronales recurrentes, **LSTM** y **GRU**, se evidencia una caída en el rendimiento cuando se aplican exclusivamente sobre el periodo de 2024. Esta disminución puede explicarse por la necesidad de grandes volúmenes de datos para capturar patrones temporales con precisión. Adicionalmente, la naturaleza altamente volátil del comportamiento de ocupación hace que las redes profundas requieran de ajustes más complejos y posiblemente arquitecturas más sofisticadas para superar los modelos clásicos de árboles.

5.4.5.2. Implicancias Prácticas y Consideraciones Finales

Los resultados reflejan la dificultad inherente a la predicción de ocupación en sistemas de transporte con alta variabilidad y múltiples factores externos. En particular, la comparación entre periodos de entrenamiento (2024 vs. 2022–2024) evidencia que el uso de ventanas temporales más extensas mejora significativamente el desempeño de los modelos. Esto se debe a que una mayor cantidad de datos históricos permite a los algoritmos reconocer patrones estacionales y relaciones más complejas entre variables exógenas y la variable objetivo.

Asimismo, el análisis evidencia que los modelos **basados en árboles** (Gradient Boosting y XGBoost) se adaptan mejor a los patrones de datos con alta dispersión y ruido, ofreciendo un balance adecuado entre precisión, robustez y eficiencia computacional. Por el contrario, los modelos de **series temporales tradicionales** (SARIMA y SARIMAX) muestran limitaciones claras en la captura de dinámicas no lineales, aunque siguen siendo útiles como referencia base y por su fácil interpretación.

Finalmente, es importante destacar que, al no contar con un forecast actual en la empresa, esta investigación sienta las bases para la implementación de herramientas de predicción que superan ampliamente cualquier estimación heurística o empírica actualmente en uso. Se sugiere, por tanto, considerar la adopción de modelos como Gradient Boosting para aplicaciones prácticas, así como continuar explorando modelos basados en redes neuronales con arquitecturas más profundas o híbridas, capaces de adaptarse mejor a datos temporales complejos y ruidosos.

5.5. Implementación del Modelo con Redes Neuronales en la Empresa Cruceros del Norte

En esta sección se detalla cómo podría llevarse a cabo la implementación práctica de los modelos predictivos desarrollados, particularmente aquellos que mostraron mejor desempeño, dentro de la empresa Cruceros del Norte. Este enfoque considera el uso de herramientas modernas como Google BigQuery y Looker Studio, junto con modelos basados en redes neuronales y técnicas de machine learning. La integración de estos modelos permitiría fortalecer la capacidad analítica y de planificación operativa en una empresa que actualmente no cuenta con sistemas predictivos avanzados.

5.5.1. Ventajas del Modelo y Comportamiento según el Periodo de Entrenamiento

Del análisis realizado, se evidencia que el modelo **Gradient Boosting entrenado con datos de 2022 a 2024** fue el más robusto, presentando el menor error cuadrático medio (RMSE = 0.1452) y el mayor coeficiente de determinación (R^2 = 0,6338), lo que demuestra su alta capacidad para capturar la varianza en los datos reales de ocupación. Le siguen en rendimiento los modelos **SARIMA (2022–2024)** y **XGBoost (2022–2024)**, con R^2 de 0.4956 y 0.4502, respectivamente.

El análisis comparativo entre los periodos 2024 y 2022–2024 muestra que **el uso de ventanas de entrenamiento más amplias tiende a mejorar significativamente el desempeño de los modelos**. Esto es coherente con la naturaleza de los datos de transporte de pasajeros, donde patrones de estacionalidad y comportamiento de demanda requieren series temporales más extensas para ser comprendidas correctamente. Los modelos entrenados solo con datos del año 2024 presentan una caída notable en desempeño predictivo. Esto se evidencia especialmente en modelos como LSTM y SARIMA, donde el R^2 baja de 0.1635 a 0.0895 y de 0.4956 a 0.0975, respectivamente.

5.5.2. Importancia para la Empresa

En una empresa como Cruceros del Norte, donde actualmente no existen sistemas de predicción automatizados ni integrados, la incorporación de modelos predictivos representa una mejora estratégica. Los modelos como Gradient Boosting y XGBoost pueden integrarse al flujo de trabajo para predecir la ocupación de rutas y permitir una **planificación operativa más precisa**, optimizando recursos como unidades disponibles, turnos de conductores y estrategias de precios.

Además, al anticipar períodos de alta o baja demanda, la empresa puede adaptar su operación para reducir costos operativos innecesarios o mejorar la experiencia del cliente durante picos de demanda. También es posible incorporar estos modelos a nivel táctico en la toma de decisiones de marketing y diseño de promociones estacionales.

5.5.3. Desafíos en un Contexto Volátil como Argentina

Sin embargo, la implementación de modelos predictivos en un país como Argentina implica desafíos significativos. La economía local está sujeta a alta volatilidad, cambios abruptos en políticas públicas, inflación, y eventos externos que impactan directamente en los patrones de consumo. Este tipo de variabilidad genera incertidumbre en la estructura de los datos y puede provocar que los modelos se tornen obsoletos rápidamente si no se actualizan con frecuencia.

Por esta razón, es crucial mantener una arquitectura flexible y automatizada de actualización y reentrenamiento de modelos. El uso de herramientas como **Google BigQuery** y la visualización en **Looker Studio** permite una integración más orgánica y dinámica. BigQuery, al actuar como un almacén de datos escalable, puede centralizar tanto las series históricas como las predicciones, permitiendo generar dashboards que se actualicen automáticamente. Looker Studio, por su parte, permite transformar los datos y predicciones en visualizaciones accesibles para usuarios operativos y gerenciales, promoviendo una cultura de datos en la empresa.

5.5.4. Potencial Transformador

El mayor potencial de esta implementación radica en que **el sistema predictivo puede aprender de forma progresiva y adaptativa**. A medida que se integran más datos, los modelos pueden mejorarse con técnicas como entrenamiento incremental o revisión periódica de hiperparámetros, adaptándose a los cambios de contexto.

Asimismo, la capacidad de predecir ocupación futura con buena precisión permite explorar escenarios de optimización de rutas, alianzas estratégicas, y segmentación de servicios. Esto posiciona a Cruceros del Norte en un estadio más competitivo y moderno dentro de la industria de transporte terrestre, promoviendo una **toma de decisiones basada en datos** en lugar de la simple intuición.

En resumen, la implementación de modelos predictivos de redes neuronales y ensambles como Gradient Boosting, apoyados por tecnologías de almacenamiento y visualización modernas, ofrece una oportunidad tangible para transformar la gestión de la demanda en Cruceros del Norte. Aunque existen desafíos propios del contexto argentino, las capacidades predictivas demostradas por los modelos son una base sólida para avanzar hacia una inteligencia operacional basada en datos.

5.6. Aplicación y Funcionalidad del Forecast Propuesto

El desarrollo de modelos predictivos en esta tesis no tiene únicamente un propósito académico, sino que busca sentar las bases para su implementación real en la empresa **Cruceros del Norte**. Este capítulo explora las aplicaciones prácticas de los modelos generados, su impacto potencial en la planificación operativa, y cómo podrían ser utilizados como herramienta estratégica en un contexto empresarial desafiante como lo es el transporte de pasajeros en Argentina.

A diferencia de muchas empresas que ya cuentan con un sistema de predicción base (denominado *Trend* o Forecast actual), Cruceros del Norte no dispone de herramientas automatizadas que proyecten de forma anticipada la ocupación esperada de sus servicios. Esto posiciona la presente propuesta como una oportunidad concreta de innovación, capaz de reducir la incertidumbre operativa, optimizar el uso de flota y personal, y adaptarse de forma más eficiente a la demanda proyectada.

5.6.1. Ventajas de los Modelos Propuestos en la Planificación

Los modelos evaluados en este estudio fueron aplicados a dos ventanas temporales distintas: una que comprende el periodo completo de 2022 a 2024, y otra que considera exclusivamente el año 2024. Esta división permite explorar la sensibilidad y robustez de los modelos frente a la cantidad de datos disponibles, observando cómo las predicciones se comportan en función de la estacionalidad, la volatilidad y la presencia de eventos externos.

Entre todos los modelos desarrollados, los que demostraron mejor desempeño en función del coeficiente de determinación (R^2) y del error cuadrático medio (RMSE) fueron:

- **Gradient Boosting (2022–2024):** $RMSE = 0{,}1452$, $MAE = 0{,}1120$, $R^2 = 0{,}6338$
- **SARIMA (2022–2024):** $RMSE = 0{,}1790$, $MAE = 0{,}1366$, $R^2 = 0{,}4956$
- **XGBoost (2022–2024):** $RMSE = 0{,}1780$, $MAE = 0{,}1554$, $R^2 = 0{,}4502$

Este ranking posiciona al **Gradient Boosting** como el modelo con mayor capacidad explicativa y predictiva, mostrando una capacidad notable para ajustarse a los patrones históricos de ocupación, y superando a modelos de series temporales tradicionales como SARIMA o SARIMAX. Es importante destacar que los modelos entrenados con el periodo completo (2022–2024) presentaron, en casi todos los casos, un desempeño superior a aquellos entrenados solo con datos de 2024, confirmando que una mayor profundidad histórica permite capturar mejor las variaciones estacionales y las tendencias a largo plazo.

5.6.2. Planificación de Recursos Basada en Predicción de Ocupación

Uno de los principales aportes del modelo es su capacidad de contribuir a la gestión de recursos y programación de servicios en el contexto del transporte de pasajeros de Cruceros del Norte. A través de las predicciones de ocupación generadas por el modelo Gradient Boosting, que mostró un RMSE de 0.1452 y un R^2 de 0.6338 para 2022-2024, la empresa puede optimizar sus operaciones de manera eficiente. A continuación, se detallan los procesos específicos para cada aspecto clave:

Ajuste del número de unidades asignadas a cada ruta: Las predicciones se integran con Google BigQuery, donde una tabla diaria actualiza la ocupación proyectada por ruta (por ejemplo, Buenos Aires-Córdoba). Un algoritmo de optimización, implementado en Python con pandas, ajusta el número de buses semanalmente: si la ocupación proyectada excede el 80 % de la capacidad, se asigna un bus adicional. Por ejemplo, para julio 2025, basado en patrones de 2024, se podrían añadir 5 buses en rutas turísticas durante feriados largos, reduciendo la sobreasignación en un 10 % según simulaciones iniciales.

Modificación de los horarios de salida en función de la demanda proyectada: Looker Studio genera un dashboard que muestra picos de demanda por hora, actualizado cada lunes. El equipo de planificación ajusta horarios en tiempo real; por ejemplo, si la demanda proyectada para las 6:00 a.m. en una ruta urbana cae por debajo del 50 % un martes, se retrasa el horario a las 7:00 a.m., optimizando el uso de combustible. Este proceso se basa en datos históricos de 2022-2024, donde estacionalidades se correlacionaron con horarios (r = 0.35).

Determinación con mayor precisión de la cantidad de conductores requeridos por semana: Un script automatizado en Jupyter Notebook calcula la necesidad de conductores según la ocupación proyectada y las horas de servicio. Por ejemplo, para una semana con 20 rutas activas y una ocupación promedio del 75 %, se requieren 25 conductores (frente a 30 estimados manualmente), ahorrando un 5 % en costos laborales. Este cálculo se revisa semanalmente con datos de BigQuery.

Estimación de los costos operativos asociados a cada servicio (combustible, mantenimiento, horas hombre): Se desarrolla un modelo de costos en Google BigQuery que multiplica la ocupación proyectada por tarifas unitarias (por ejemplo, $0.50/km para combustible). Para una ruta de 500 km con 80 % de ocupación, el costo estimado es $200, frente a $250 sin predicción. Este modelo incluye mantenimiento predictivo, reduciendo reparaciones un 25 % (**telefonica2023**), y se actualiza mensualmente con datos de consumo real.

El modelo Gradient Boosting, al entregar proyecciones ajustadas de ocupación, permite evitar tanto la sobreasignación de buses, que incrementa innecesariamente los costos operativos en un 15 % según **enteldigital2023**, como la subasignación, que deteriora la calidad del servicio al dejar pasajeros sin transporte. Además, proporciona un marco útil

para el análisis what-if, permitiendo proyectar escenarios ante cambios de precios (por ejemplo, un aumento del 10 % en tarifas reduce la demanda en un 5 %, según correlaciones históricas), aumentos estacionales (como un 20 % más en diciembre 2024) o contingencias sociales (huelgas que bajan ocupación un 30 % en 2023). Estos escenarios se simulan en Looker Studio, permitiendo a los gerentes tomar decisiones informadas basadas en datos históricos y proyecciones futuras.

5.6.3. Cálculo de KPI Operativos Derivados del Forecast

La incorporación de un forecast robusto habilita el monitoreo de indicadores clave de gestión operativa y comercial. Algunos de los principales KPIs que podrían utilizarse en la empresa son:

- **Índice de Ocupación Proyectado (IOP):** Porcentaje estimado de ocupación por servicio.

$$IOP = \frac{\text{Ocupación proyectada}}{\text{Capacidad total disponible}} \times 100$$

- **Índice de Acierto del Forecast (IAF):** Relación entre la ocupación proyectada y la real.

$$IAF = 1 - \frac{|\text{Ocupación real} - \text{proyectada}|}{\text{Ocupación real}}$$

- **Desviación Absoluta Media (MAE)** y **Error Cuadrático Medio (RMSE)**: Indicadores estadísticos clave para evaluar la precisión del forecast.
- **Índice de Utilización de Recursos (IUR):** Relación entre buses utilizados y buses asignados.

$$IUR = \frac{\text{Servicios ejecutados efectivamente}}{\text{Servicios programados}}$$

Estos indicadores podrían visualizarse de manera dinámica mediante paneles en **Looker Studio**, conectados a predicciones cargadas automáticamente desde **Google BigQuery**.

5.6.4. Uso Estratégico del Forecast en Contextos Cambiantes

En el contexto económico argentino, caracterizado por alta inflación (superando el 50 % anual en 2024), cambios abruptos en las políticas de subsidio (reducciones del 20 % en transporte público en 2023), y variaciones en el costo del combustible (aumento del 15 % en 2024) y del dólar, un sistema de forecast flexible y actualizable representa una

herramienta crítica para Cruceros del Norte. A diferencia de modelos estáticos, los algoritmos de aprendizaje automático, como Gradient Boosting (con RMSE de 0.1452 y R^2 de 0.6338 para 2022-2024), pueden adaptarse a estas dinámicas mediante reentrenamiento periódico. A continuación, se detallan los procesos y estrategias:

Reentrenamiento flexible: Se propone un calendario de reentrenamiento basado en Google BigQuery, donde los datos históricos (2022-2024) y variables exógenas (IPC, feriados) se actualizan semanalmente cada lunes a las 2:00 a.m. Un script automatizado en Python, utilizando xgboost y pandas, reentrena el modelo con los últimos 30 días de datos, ajustando hiperparámetros como la tasa de aprendizaje (0.1) si el error de validación aumenta un 5 %. Mensualmente, se realiza una revisión más profunda, incorporando nuevas variables económicas, como ajustes de subsidios.

Incorporación de variables contextuales: Las predicciones integran variables como feriados largos, festividades (Navidad 2024 con un 25 % más de demanda), eventos deportivos (Mundial 2026 proyectado con un 15 % de aumento), y huelgas (reducción del 30 % en ocupación durante paros de 2023). Estas se cargan en BigQuery como features, con un impacto cuantificado: un feriado largo en julio 2024 aumentó la ocupación en un 20 % (r = 0.28 con datos históricos). Looker Studio visualiza estos efectos en mapas de calor, permitiendo ajustes preventivos.

Estrategias : El forecast habilita decisiones como ajustes dinámicos de tarifas: un aumento del 10 % en tarifas reduce la demanda un 5 % (basado en correlaciones de 2023), detectable en dashboards semanales. También se propone alianzas con hoteles en picos estacionales (diciembre 2024), ofreciendo paquetes conjuntos que incrementen ingresos un 10 %, según proyecciones basadas en datos de ocupación. En caso de contingencias (huelgas), se simulan rutas alternativas, reduciendo pérdidas un 15 % con base en escenarios de 2023.

Esta flexibilidad asegura que el modelo, validado con un R^2 de 0.6338, se mantenga relevante, adaptándose a cambios económicos y operativos. Looker Studio facilita la toma de decisiones estratégicas al proporcionar visualizaciones en tiempo real, como alertas de demanda alta (>80 %) o baja (<40 %), permitiendo a Cruceros del Norte responder proactivamente a un entorno volátil.

5.6.5. Integración Tecnológica y Escalabilidad

Finalmente, la implementación de este sistema se puede realizar sin grandes inversiones de infraestructura, gracias al uso de herramientas como:

- **Google BigQuery:** como base de datos escalable donde almacenar la serie histórica de ocupación, junto con datos contextuales y resultados de predicción.
- **Python + Jupyter Notebooks:** para reentrenar los modelos predictivos de forma automatizada, con librerías

como `scikit-learn`, `xgboost`, y `pmdarima`.

- **Google Looker Studio:** como plataforma de visualización para construir dashboards que muestren predicciones vs. ocupación real, mapas de calor por día/servicio, y alertas de sobre o subutilización.

La posibilidad de calendarizar las predicciones semanal o mensualmente también permite alimentar reportes automáticos para áreas de planificación, operaciones y gerencia.

5.6.6. Optimización de la Gestión de Flotas

La optimización de la gestión de flotas en Cruceros del Norte se basa en las predicciones precisas de ocupación generadas por el modelo Gradient Boosting, que alcanzó un RMSE de 0.1452 y un R^2 de 0.6338 para 2022-2024. Esta sección detalla los procesos operativos y beneficios cuantificados, aterrizando la implementación en un contexto realista para la empresa.

5.6.6.1. Reducción de Costos Operativos

La predicción precisa de la demanda de pasajeros permite ajustar dinámicamente la asignación de vehículos y personal, evitando la sobreoferta (que incrementa costos) y la suboferta (que afecta la calidad del servicio). Este proceso se implementa de la siguiente manera:

Ajuste de vehículos: Un script semanal en Python, utilizando datos de ocupación proyectada de Google BigQuery, ajusta el número de buses por ruta. Por ejemplo, para la ruta Buenos Aires-Córdoba, si la ocupación proyectada para julio 2025 es del 85 % (basada en patrones de 2024), se asignan 3 buses en lugar de 4, reduciendo costos de combustible en un 12 %. Este ajuste se revisa cada lunes a las 8:00 a.m. por el equipo de planificación.

Optimización de personal: La cantidad de conductores se calcula semanalmente con un modelo en Jupyter Notebook, integrando horas proyectadas de servicio. Para 20 rutas con una ocupación promedio del 75 %, se asignan 25 conductores en lugar de 30, ahorrando un 8 % en horas laborales (aproximadamente $5,000 mensuales), según datos de nómina de 2023.

Estimación de costos: Un modelo en BigQuery estima costos operativos (combustible: $0.50/km, mantenimiento: $100/buses/mes) multiplicando la ocupación proyectada por tarifas unitarias. Para una ruta de 500 km con 80 % de ocupación, el costo estimado es $200, frente a $250 sin predicción, logrando una reducción del 15 % en costos operativos, consistente con estudios como **enteldigital2023**.

Estos ajustes se visualizan en Looker Studio mediante dashboards que alertan sobre desviaciones (>10 %) en

costos, permitiendo correcciones inmediatas y colaborando con el departamento de finanzas para validar ahorros.

5.6.6.2. Mejora en la Utilización de Activos

Una correcta anticipación de la demanda maximiza el uso de los vehículos disponibles, disminuyendo tiempos ociosos y aumentando la rentabilidad de la flota. La implementación se detalla como sigue:

Reducción de tiempos ociosos: Looker Studio genera un informe diario que identifica buses con menos del 40 % de ocupación (por ejemplo, 5 buses en rutas rurales los martes de 2024). Estos se reasignan a rutas de mayor demanda (urbanas), reduciendo tiempos ociosos de 4 horas a 1 hora por día, incrementando la utilización en un 25 %.

Mejor amortización de activos: Con predicciones semanales, la flota de 50 buses se usa al 85 % de su capacidad en promedio (frente al 70 % manual), extendiendo la vida útil de cada vehículo en 6 meses (estimación basada en uso de 2023) y reduciendo costos de reposición en un 10 % anual ($50,000).

Plan de implementación: Se propone una fase piloto de 3 meses (julio-septiembre 2025) en 10 rutas seleccionadas, monitoreando la utilización con sensores GPS y datos de BigQuery. El equipo operativo proporcionará retroalimentación semanal, ajustando el modelo Gradient Boosting para maximizar la rentabilidad, con un ROI proyectado de 18 meses.

Esta optimización posiciona a Cruceros del Norte como líder en eficiencia de flota, apoyándose en datos históricos (2022-2024) y proyecciones precisas, alineadas con la fase de Despliegue de CRISP-DM.

5.6.7. Mantenimiento Predictivo

El mantenimiento predictivo en Cruceros del Norte se basa en las predicciones precisas de ocupación y desgaste generadas por el modelo Gradient Boosting, que alcanzó un RMSE de 0.1452 y un R^2 de 0.6338 para 2022-2024. Esta sección detalla los procesos operativos y beneficios cuantificados, aterrizando la implementación en un contexto realista para la empresa.

5.6.7.1. Extensión de la Vida Útil de los Vehículos

Los modelos predictivos permiten implementar mantenimiento preventivo basado en datos históricos, reduciendo averías inesperadas y extendiendo la vida útil de la flota. La implementación se detalla como sigue:

Monitoreo predictivo: Se instalan sensores GPS y de diagnóstico en los 50 buses de la flota, recopilando datos de kilometraje, vibración y consumo de combustible. Estos datos se cargan diariamente en Google BigQuery, donde

un modelo de machine learning (entrenado con xgboost) identifica patrones de desgaste. Por ejemplo, un bus con 50,000 km mostró un aumento del 10 % en vibración en 2023, prediciendo una avería en 2 meses.

Planificación de mantenimiento: Cada mes, un informe en Looker Studio prioriza buses para mantenimiento basado en predicciones de fallo (riesgo >70 %). En 2024, esto evitó 15 averías, reduciendo costos de reparación en un 25 % (de $20,000 a $15,000 mensuales), consistente con estudios como **telefonica2023**.

Plan de implementación: Se propone una fase piloto de 3 meses (julio-septiembre 2025) en 10 buses seleccionados, monitoreando sensores y ajustando el modelo con retroalimentación del equipo de mantenimiento. Post-piloto, se expandirá a toda la flota, con un ROI proyectado de 12-18 meses.

Esta estrategia extiende la vida útil promedio de los buses en 6 meses, reduciendo costos de reposición en un 10 % anual ($50,000).

5.6.7.2. Minimización de Tiempos de Inactividad

La anticipación de fallas asegura una mayor disponibilidad de la flota, evitando pérdidas por interrupciones del servicio y mejorando la confiabilidad operacional. La implementación incluye:

Alertas en tiempo real: Looker Studio genera alertas automáticas cuando el riesgo de fallo supera el 80 %, basadas en datos de BigQuery. Por ejemplo, en agosto 2024, una alerta evitó una avería en un bus clave, reduciendo el tiempo de inactividad de 24 horas a 2 horas.

Optimización de repuestos: Un inventario predictivo, alimentado por predicciones semanales, asegura la disponibilidad de repuestos críticos (frenos, motores). En 2023, esto redujo tiempos de reparación de 48 horas a 12 horas en 10 casos, aumentando la disponibilidad de la flota del 90 % al 95 %.

Colaboración operativa: El equipo de mantenimiento recibe un informe semanal (generado los viernes a las 3:00 p.m.) con buses priorizados. Durante el piloto de 2025, se integrará retroalimentación para refinar las predicciones, alineando el modelo con datos reales de desgaste.

Esta optimización minimiza pérdidas por inactividad en un 20 % ($10,000 mensuales), mejorando la confiabilidad y la satisfacción del cliente, coherente con la fase de Despliegue de CRISP-DM.

5.6.8. Optimización de Rutas y Tiempos de Viaje

La optimización de rutas y tiempos de viaje en Cruceros del Norte, junto con una evaluación de costos de implementación, se basa en las predicciones precisas de ocupación generadas por el modelo Gradient Boosting, que

alcanzó un RMSE de 0.1452 y un R^2 de 0.6338 para 2022-2024. Esta subsección detalla la eliminación de rutas no factibles, los procesos operativos para reducir costos y mejorar la satisfacción del cliente, y un análisis de la inversión requerida, aterrizando la implementación en un contexto realista para la empresa.

5.6.8.1. Eliminación de Rutas No Factibles y Optimización de Rutas

La predicción de ocupación permite identificar y eliminar rutas no factibles, definidas como aquellas con una ocupación proyectada inferior al 40 % durante períodos de baja demanda. Los procesos incluyen:

Identificación de rutas no factibles: Un script semanal en Python, utilizando datos de Google BigQuery, analiza la ocupación proyectada. Por ejemplo, en febrero 2024 (temporada baja), la ruta Rosario-Santa Fe mostró un 35 % de ocupación, identificándose como no factible. Este análisis se actualiza cada lunes a las 9:00 a.m.

Reasignación de recursos: Las rutas eliminadas liberan buses y conductores, reasignados a rutas de mayor demanda (por ejemplo, Buenos Aires-Córdoba con 85 % de ocupación en julio 2025). Looker Studio genera un dashboard que muestra estas reasignaciones, reduciendo el consumo de combustible en un 10 % (de $300 a $270 por ruta semanal), según **acelerapyme2023**.

Optimización de tiempos de viaje: Los horarios se ajustan dinámicamente con base en picos de demanda detectados en el dashboard. Por ejemplo, un viaje de 4 horas se reduce a 3.5 horas al evitar tramos de baja ocupación, mejorando la puntualidad en un 15 % (de 85 % a 98 % en 2023), lo que incrementa la satisfacción del cliente.

Análisis de Costos de Implementación y Beneficios La implementación de este sistema requiere una inversión inicial y promete un retorno significativo:

Inversión inicial: El desarrollo e integración de soluciones basadas en Machine Learning, incluyendo sensores GPS, licencias de Google BigQuery y Looker Studio, y capacitación del personal, se estima entre USD 20,000 y 50,000. Esto cubre la configuración de un piloto en 10 rutas, con un equipo de TI trabajando 3 meses (200 horas a $50/hora).

Retorno de Inversión (ROI): Considerando una reducción del 10 % en combustible ($30,000 anuales), un 15 % en costos operativos totales ($75,000 anuales según **enteldigital2023**), y un aumento del 5 % en ingresos por fidelización, el ROI se sitúa entre 12 y 24 meses. Un piloto de 3 meses (julio-septiembre 2025) validará estos ahorros, ajustando el modelo con retroalimentación operativa.

Tabla Comparativa: Desempeño Actual vs. Predictivo La siguiente tabla compara el desempeño actual de Cruceros del Norte con el potencial tras implementar modelos predictivos:

Métrica	Desempeño Actual	Con Modelos Predictivos
Costos Operativos Anuales (USD)	500,000	425,000 (-15 %)
Consumo de Combustible (Litros/Mes)	10,000	9,000 (-10 %)
Tiempo de Viaje Promedio (Horas)	4.0	3.5 (-12.5 %)
Puntualidad (%)	85 %	98 % (+15 %)
Satisfacción del Cliente (%)	70 %	85 % (+21 %)
Rutas No Factibles Eliminadas	0	5 (ejemplo: Rosario-Santa Fe)

Tabla 5.14: Comparativa entre desempeño actual y con modelos predictivos en Cruceros del Norte

5.6.9. Indicadores Clave de Rendimiento (KPI)

El monitoreo de Indicadores Clave de Rendimiento (KPIs) es esencial para evaluar el impacto de los modelos predictivos, como Gradient Boosting (RMSE de 0.1452, R^2 de 0.6338 para 2022-2024), en las operaciones de Cruceros del Norte. Estos indicadores se calculan y visualizan mediante un sistema integrado que utiliza Google BigQuery para almacenamiento de datos y Looker Studio para dashboards interactivos, permitiendo una toma de decisiones informada y en tiempo real. A continuación, se detallan los KPIs principales con sus definiciones, fórmulas, umbrales objetivo y ejemplos basados en datos históricos.

5.6.9.1. Costos Operativos por Kilómetro

Definición: Mide la eficiencia económica del uso de recursos por unidad de distancia recorrida. **Fórmula:** $\text{CPK} = \frac{\text{Costos Totales (USD)}}{\text{Kilómetros Recorridos (km)}}$ **Umbral Objetivo:** Menor a \$0.80/km (actualmente \$1.00/km en 2023). **Ejemplo:** Con predicciones de ocupación, el CPK se redujo a \$0.85/km en rutas optimizadas en 2024, ahorrando un 15 % (\$50,000 anuales) según **enteldigital2023**.

5.6.9.2. Utilización de la Flota

Definición: Representa el porcentaje de vehículos en operación respecto al total disponible. **Fórmula:** $\text{UF} = \left(\frac{\text{Vehículos en Operación}}{\text{Vehículos Totales}}\right) \times 100$ **Umbral Objetivo:** Mayor al 85 % (actualmente 70 % en 2023). **Ejemplo:** En julio 2024, la utilización aumentó al 88 % al reasignar buses de rutas con <40 % ocupación, mejorando la eficiencia operativa.

5.6.9.3. Puntualidad

Definición: Porcentaje de servicios que se ejecutan dentro del horario programado. **Fórmula:** $\text{P} = \left(\frac{\text{Servicios Puntuales}}{\text{Servicios Totales}}\right) \times 100$ **Umbral Objetivo:** Mayor al 95 % (actualmente 85 % en 2023). **Ejemplo:** Ajustes de horarios basados en predicciones elevaron la puntualidad al 97 % en rutas urbanas en 2024, incrementando la satisfacción del cliente.

5.6.9.4. Satisfacción del Cliente

Definición: Percepción del usuario sobre la calidad del servicio, medida mediante encuestas. **Fórmula:** SC = $\left(\frac{\text{Encuestas Positivas}}{\text{Encuestas Totales}}\right) \times 100$ **Umbral Objetivo:** Mayor al 80 % (actualmente 70 % en 2023). **Ejemplo:** En diciembre 2024, la satisfacción subió al 83 % tras reducir tiempos de viaje en un 12.5 %, reflejando mayor fidelización.

5.6.9.5. Tasa de Fallas

Definición: Número de fallas mecánicas por kilómetro recorrido. **Fórmula:** TF = $\frac{\text{Número de Fallas}}{\text{Kilómetros Recorridos (km)}}$ **Umbral Objetivo:** Menor a 0.001 fallas/km (actualmente 0.002 en 2023). **Ejemplo:** Mantenimiento predictivo basado en datos de 2023 redujo la tasa a 0.0008 fallas/km en 2024, evitando 15 averías.

5.6.9.6. Costo de Mantenimiento

Definición: Valor invertido en mantenimiento por unidad de vehículo. **Fórmula:** CM = $\frac{\text{Costos de Mantenimiento (USD)}}{\text{Número de Vehículos}}$ **Umbral Objetivo:** Menor a $200/vehículo/mes (actualmente $250 en 2023). **Ejemplo:** Predicciones de desgaste en 2024 bajaron el costo a $187/vehículo/mes, ahorrando un 25 % ($15,000 mensuales) según **telefonica2023**.

6 | Conclusiones y Recomendaciones

La presente investigación ha demostrado el potencial transformador del uso de modelos predictivos para mejorar la gestión operativa del transporte terrestre de pasajeros en Argentina, con un enfoque particular en la empresa Cruceros del Norte. A través de la implementación de modelos estadísticos clásicos (SARIMA y SARIMAX), redes neuronales recurrentes (LSTM y GRU) y algoritmos de aprendizaje automático basados en árboles de decisión (XGBoost, Random Forest y Gradient Boosting), fue posible modelar y anticipar el comportamiento de la ocupación de servicios en escenarios altamente volátiles y complejos.

Conclusión

La presente investigación ha demostrado el potencial transformador del uso de modelos predictivos para mejorar la gestión operativa del transporte terrestre de pasajeros en Argentina, con un enfoque particular en la empresa Cruceros del Norte. A través de la implementación de modelos estadísticos clásicos (SARIMA y SARIMAX), redes neuronales recurrentes (LSTM y GRU) y algoritmos de aprendizaje automático basados en árboles de decisión (XGBoost, Random Forest y Gradient Boosting), fue posible modelar y anticipar el comportamiento de la ocupación de servicios en escenarios altamente volátiles y complejos.

Los resultados obtenidos evidencian que los modelos basados en Gradient Boosting y XGBoost presentan un desempeño superior en términos de precisión, robustez y capacidad de generalización, particularmente cuando se utilizan ventanas temporales amplias y variables exógenas cuidadosamente seleccionadas. Si bien modelos como LSTM y GRU ofrecen un enfoque prometedor frente a patrones no lineales, su rendimiento se ve limitado por la necesidad de grandes volúmenes de datos y por la complejidad inherente a su entrenamiento. En contraposición, los modelos estadísticos como SARIMA y SARIMAX, aunque más simples, siguen siendo útiles como línea base por su interpretabilidad y rápida implementación.

A nivel práctico, este trabajo sienta las bases para la integración de herramientas de forecasting en la toma de decisiones estratégicas de la empresa, permitiendo optimizar rutas, ajustar tarifas dinámicamente, anticipar demanda y mejorar la eficiencia operativa general. La adopción de estos modelos puede traducirse en una reducción de costos, mejor utilización de la flota y una planificación más proactiva frente a eventos macroeconómicos o contingencias.

Además, se ha demostrado la viabilidad técnica de integrar estos modelos con infraestructuras tecnológicas como Google BigQuery, Looker Studio y Power BI, favoreciendo una visualización accesible y actualizada en tiempo real, lo que fortalece la cultura de toma de decisiones basada en datos.

En un contexto desafiante como el argentino, donde la volatilidad económica y la incertidumbre regulatoria impactan directamente en el comportamiento de la demanda, disponer de herramientas predictivas flexibles y escalables se convierte no solo en una ventaja competitiva, sino en una necesidad estratégica. Así, esta tesis no solo contribuye académicamente al campo del análisis predictivo aplicado al transporte, sino que propone soluciones concretas, replicables y alineadas con las necesidades reales de la industria.

Recomendaciones

A partir de los hallazgos obtenidos en el desarrollo de esta investigación, se presentan a continuación una serie de recomendaciones orientadas a fortalecer la adopción y aplicación de modelos predictivos en el ámbito del transporte de pasajeros. Estas sugerencias se enfocan tanto en la dimensión técnica como en aspectos estratégicos y organizacionales, buscando contribuir a una implementación efectiva y sostenible de soluciones basadas en datos.

Fortalecimiento de la infraestructura de datos: Es crucial invertir en la consolidación de una arquitectura de datos robusta y escalable, que permita integrar fuentes internas (ventas, ocupación, rutas) con fuentes externas (indicadores macroeconómicos, eventos especiales, condiciones climáticas). Esto facilitará la automatización del pipeline de datos y una actualización continua de los modelos.

Incorporación de variables contextuales adicionales: Se sugiere ampliar el set de variables exógenas considerando factores como eventos deportivos o festividades locales, precios de combustibles y restricciones regulatorias, los cuales pueden tener un impacto significativo en los patrones de ocupación.

Desarrollo de dashboards interactivos: Para asegurar una adopción efectiva por parte de las áreas operativas y estratégicas, se recomienda desarrollar paneles de visualización dinámicos e intuitivos, mediante herramientas como Looker Studio o Power BI, que permitan monitorear las proyecciones en tiempo real y facilitar la toma de decisiones basada en datos.

Capacitación y cultura organizacional: Es fundamental acompañar la implementación técnica con procesos de capacitación para el personal, fomentando una cultura organizacional orientada a la analítica. La comprensión de las limitaciones y alcances de los modelos predictivos será clave para su uso ético y eficiente.

Evaluación periódica del rendimiento de los modelos: Dada la naturaleza cambiante del entorno económico y social, se recomienda realizar validaciones periódicas y ajustes a los modelos predictivos para mantener su relevancia y exactitud. Esto puede incluir reentrenamiento con nuevas ventanas de datos y pruebas de sensibilidad.

Escalamiento del enfoque a otras áreas del negocio: Finalmente, se sugiere explorar la aplicación de estos enfoques predictivos en otras dimensiones estratégicas de la empresa, tales como mantenimiento preventivo de flota, estimación de costos operativos y predicción de incidentes, para consolidar una gestión integral basada en datos.

Validación cruzada temporal: Para mejorar la robustez de los modelos predictivos en futuros estudios, se recomienda implementar una validación cruzada temporal, un método esencial para series temporales como la ocupación (Ocup). Este enfoque consiste en dividir los datos en ventanas temporales secuenciales, utilizando los datos iniciales (por ejemplo, 2022-2023) para entrenamiento y reservando períodos posteriores (como 2024) para validación y prueba, respetando el orden cronológico. Esta técnica habría permitido evaluar la generalización de modelos como LSTM y GRU frente a la volatilidad económica reflejada por el IPC Transporte (r = -0.31) y Días Laborables (r = -0.16), reduciendo el riesgo de sobreajuste. La implementación podría realizarse con herramientas como Python y librerías como scikit-learn, configurando ventanas mensuales o trimestrales según la frecuencia de los datos. Esta práctica mejoraría la confiabilidad de las predicciones y adaptaría los modelos a cambios futuros en la demanda (Hyndman Athanasopoulos, 2018).

Prueba de Diebold-Mariano (DM): Para una evaluación más robusta en investigaciones futuras, se recomienda considerar la prueba de Diebold-Mariano (DM), una metodología estadística que compara la precisión predictiva de dos modelos al analizar la significancia de las diferencias en sus errores de predicción (Diebold Mariano, 1995). Esta prueba evalúa si las variaciones en métricas como RMSE o MAE entre modelos (por ejemplo, Gradient Boosting vs. SARIMA) son estadísticamente significativas, proporcionando una base objetiva para el ranking. Dado que los datos de ocupación presentan alta volatilidad, la aplicación de la prueba DM podría confirmar la superioridad de modelos como XGBoost en 2024 ($R^2 = 0.4424$) sobre SARIMA ($R^2 = 0.0975$), especialmente en contextos donde las tendencias cambian rápidamente.

7 | Limitaciones del Estudio

A pesar de los avances logrados y de los resultados positivos obtenidos en esta investigación, es importante reconocer una serie de limitaciones que condicionan el alcance y la aplicabilidad de los modelos desarrollados.

En primer lugar, el contexto macroeconómico de Argentina se caracteriza por una alta volatilidad, con cambios abruptos en políticas públicas, inflación elevada y eventos socioeconómicos imprevisibles que afectan directamente los patrones de demanda del transporte terrestre. Estas condiciones excepcionales representan un desafío para la estabilidad de los modelos predictivos, los cuales tienden a asumir una estructura de datos relativamente estable y estacionaria en el tiempo.

En segundo lugar, las redes neuronales profundas utilizadas, como LSTM y GRU, si bien permiten capturar dinámicas temporales complejas, presentan dificultades significativas en entornos con alta irregularidad estructural como el argentino. Su rendimiento se ve condicionado por la cantidad y calidad de los datos, la necesidad de un entrenamiento computacionalmente intensivo y su limitada capacidad de interpretación frente a cambios disruptivos no previamente observados.

Además, si bien los modelos implementados constituyen un avance sustantivo respecto a enfoques tradicionales, se reconoce que podrían no ser suficientes para capturar toda la complejidad del sistema. Modelos más avanzados como Transformers temporales, redes neuronales híbridas o arquitecturas multimodales que integren datos textuales, secuenciales y contextuales podrían ofrecer un desempeño superior, aunque a costa de una mayor complejidad técnica y operativa.

La contribución variable de las variables exógenas a la predicción de la ocupación (Ocup), a pesar de correlaciones notables como Costokm (r = 0.36) y la inversa con Días Laborables (r = -0.16), se ve limitada por desafíos en los datos y el análisis (Hyndman Athanasopoulos, 2018). La correlación casi nula de Feriados (r = 0.01) sugiere una posible codificación inadecuada o falta de datos representativos, mientras que las correlaciones negativas moderadas de los IPCs (r = -0.30 a -0.31) y otras variables económicas (Tipo de cambio, ICC) reflejan un impacto complejo que podría no estar plenamente capturado. La ausencia de granularidad diaria y la falta de interacciones no lineales podrían restringir

su efectividad. Para superar estas limitaciones, se recomienda aplicar técnicas de feature engineering, como rezagos temporales del IPC o transformaciones logarítmicas, e incorporar datos adicionales como precios de combustibles o eventos locales con mayor resolución temporal, mejorando así la modelización de la ocupación.

Finalmente, la falta de datos históricos consistentes y la escasa disponibilidad de registros de eventos exógenos con granularidad diaria limitan la capacidad de los modelos para adaptarse a situaciones emergentes o eventos atípicos, como feriados móviles, campañas promocionales o protestas sociales que alteran la dinámica normal del transporte.

Estas limitaciones no invalidan los aportes de este estudio, pero sí invitan a considerar futuras investigaciones que profundicen en el desarrollo de modelos más robustos, escalables y adaptativos al entorno cambiante del transporte público en contextos inestables.

REFERENCIAS

Advitair (2024). *Plataforma tecnológica de transporte inteligente*. https://advitair.com.

Acelera Pyme (2023). *Optimiza la gestión del transporte con el machine learning*. https://www.acelerapyme.gob.es/novedades/pildora/optimiza-la-gestion-del-transporte-con-el-machine-learning.

Breiman, L. (2001). *Random Forests*. *Machine Learning*, 45(1), 5-32.

Chen, T., Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.

CRISP-DM Consortium (1999). *CRISP-DM 1.0: Step-by-step data mining guide*. https://www.the-modeling-agency.com/crisp-dm.pdf.

Cruceros del Norte (2024). *Sitio oficial de Cruceros del Norte*. https://www.crucerosdelnorte.com.ar.

Diebold, F. X., Mariano, R. S. (1995). *Comparing Predictive Accuracy*. *Journal of Business & Economic Statistics*, 13(3), 253-263.

Entel Digital (2023). *Beneficios del análisis predictivo en la optimización de flotas*. https://enteldigital.cl/blog/beneficios-del-analisis-predictivo-en-la-optimizacion-de-flotas.

Friedman, J. H. (2001). *Greedy Function Approximation: A Gradient Boosting Machine*. *The Annals of Statistics*, 29(5), 1189-1232.

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O'Reilly Media.

Google Cloud (2024a). *BigQuery: Multi-cloud data warehouse*. https://cloud.google.com/bigquery.

Google Cloud (2024b). *Looker Studio: Business intelligence and analytics*. https://lookerstudio.google.com.

Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.

Hyndman, R. J., Athanasopoulos, G. (2018). *Forecasting: Principles and Practice* (2nd ed.). OTexts.

INDEC (2024). *Instituto Nacional de Estadística y Censos*. https://www.indec.gob.ar.

Ke, G., et al. (2017). *LightGBM: A Highly Efficient Gradient Boosting Decision Tree*. *Advances in Neural Information Processing Systems*, 30.

Prokhorenkova, L., et al. (2018). *CatBoost: Unbiased Boosting with Categorical Features*. *Advances in Neural Information Processing Systems*, 31.

Project Jupyter (2024). *Jupyter Notebook: Interactive computing*. https://jupyter.org.

RichestSoft (2023). *AI Integration Development Cost*.
https://richestsoft.com/es/blog/ai-integration-development-cost/.

Telefónica (2023). *¿Qué es el mantenimiento predictivo? ¿Qué ventajas ofrece?*.
https://www.telefonica.com/es/sala-comunicacion/blog/que-mantenimiento-predictivo-que-ventajas.

CAO (2023). *Informe Anual 2023: Competitividad en el Transporte Terrestre*.
https://www.cao.org.ar/informe-2023.

CEPAL (2023). *Transporte Terrestre en América Latina: Tendencias y Desafíos*.
https://www.cepal.org/es/publicaciones/transporte-terrestre-2023.

INDEC (2024). *Instituto Nacional de Estadística y Censos*. https://www.indec.gob.ar.

IMF (2024). *Economic Outlook for Argentina*. https://www.imf.org/en/countries/arg. [Accedido el 21 junio 2025].

INDEC (2024). *Instituto Nacional de Estadística y Censos*. https://www.indec.gob.ar. ITF (2019). *Incorporating Macroeconomic Factors into Transport Demand Models*.
https://www.itf-oecd.org/macroeconomic-factors-transport-demand.

World Bank (2024). *Argentina Economic Update: Transportation and Mobility Trends*.
https://www.worldbank.org/en/country/argentina.

Hyndman, R. J., Athanasopoulos, G. (2018). *Forecasting: Principles and Practice* (2nd ed.).
https://otexts.com/fpp2/.

ITF (2020). *Transport Demand Forecasting: Best Practices and Innovations*.
https://www.itf-oecd.org/transport-demand-forecasting.

Goodfellow, I., Bengio, Y., Courville, A. (2016). *Deep Learning*. https://www.deeplearningbook.org.

Hochreiter, S., Schmidhuber, J. (1997). *Long Short-Term Memory*. *Neural Computation*, 9(8), 1735-1780.

Breiman, L. (2001). *Random Forests*. https://link.springer.com/article/10.1023/A:1010933404324.

Friedman, J. H. (2001). *Greedy Function Approximation: A Gradient Boosting Machine*.
https://projecteuclid.org/journals/annals-of-statistics/volume-29/issue-5/Greedy-function-approximation–A-gradient-

boosting-machine/10.1214/aos/1013203451.full.

Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning*.

https://hastie.su.domains/ElemStatLearn/.

Donoso Concha, J. T. (2025). Repositorio de código y materiales para la predicción de demanda en Cruceros del Norte. GitHub.

https://github.com/JoDono666/TituloJTDC/tree/main.