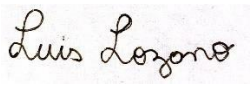
Santiago de Cali, 07 de 06 de 2024

Ingeniero:
Juan Carlos Martínez Arias Director
Posgrados de IngenieríaFacultad de
Ingeniería y Ciencias
Pontificia Universidad Javeriana - Cali

Con el fin de cumplir con los requisitos exigidos por la Universidad para llevar a cabo el Trabajo de Grado y posteriormente optar por el título de Magíster en Ciencia de Datos, nos permitimos presentar a su consideración el proyecto de Trabajo de Grado denominado "Pronóstico del mercado aéreo en las principales rutas comerciales domésticas de pasajeros en Colombia mediante algoritmos de machine learning", el cual serárealizado por los estudiantes Luis Lozano, Juan Restrepo y Natalia Mayorga con código 8980396, 8976156 y 8975847 perteneciente al énfasis en N/A, bajo la dirección de la profesora María Costanza Pabón Burbano.

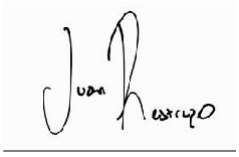
El suscrito director del Trabajo de Grado autoriza para que se proceda a hacer la evaluación de este Proyecto ante el Tribunal que para el efecto se designe, toda vez que ha revisado cuidadosamente el documento y avala que ya se encuentra listo para ser presentado oficialmente.

Atentamente,

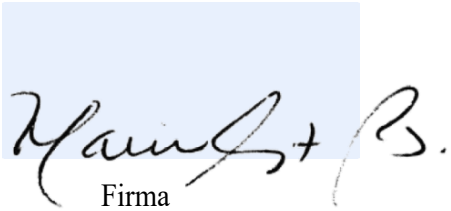


Firma
Luis Carlos Lozano Caballero
C.C. 1.043.877.350 de Palmar de Varela

Firma
Natalia Mayorga Ballesteros
C.C. 1.057.601.680 de Sogamoso



Firma
Juan Carlos Restrepo Blanco
C.C. 1.143.261.537 de Barranquilla



Firma
María Costanza Pabón Burbano
C.C. 34.559.226 de. Popayán

FICHA RESUMEN

TRABAJO DE GRADO DE MAESTRÍA

TÍTULO: PRONÓSTICO DEL MERCADO AÉREO EN LAS PRINCIPALES RUTAS COMERCIALES DOMÉSTICAS DE PASAJEROS EN COLOMBIA MEDIANTE ALGORITMOS DE MACHINE LEARNING

1. ÉNFASIS: N/A
2. TIPO DE PROYECTO: Aplicado
3. ÁREA DE TRABAJO: Series de tiempo
4. ESTUDIANTE(S): Luis Carlos Lozano, Juan Carlos Restrepo, Natalia Mayorga
5. CORREO ELECTRÓNICO: luislozanoc@javerianacali.edu.co juankrloswhite@javerianacali.edu.co natalia96@javerianacali.edu.co
6. DIRECCIÓN Y TELEFONO: calle 3#5-60 Palmar de Varela 3003486458, calle 112#42-19 Barranquilla 3004304522, carrera 28#7ª-05 Sogamoso 3123320702
7. DIRECTOR: María Constanza Pabón Burbano
8. VINCULACIÓN DEL DIRECTOR: Planta
9. CORREO ELECTRÓNICO DEL DIRECTOR: mcpabon@javerianacali.edu.co
10. PALABRAS CLAVE: Pax, hub, Mercado Aéreo, Capacidad, Demanda, Pronostico.
11. ODS QUE APLICA EL PROYECTO (Agenda 2030): ODS 9: Industria, innovación e infraestructura
12. FECHA DE INICIO: 15/05/2023
13. RESUMEN: El mercado doméstico aéreo en Colombia juega un papel fundamental en la economía y el turismo nacional. Sin embargo, la partida de algunas aerolíneas como Viva Air en el año 2023 ha generado una disminución en la capacidad de vuelos en rutas de alta demanda, impactando tanto a las compañías aéreas restantes como a los pasajeros. En este proyecto, se han desarrollado modelos de Machine Learning que utilizan datos históricos para pronosticar la evolución del tráfico de pasajeros en el mercado aéreo. Estos pronósticos están destinados a ayudar a las compañías aéreas a tomar decisiones estratégicas informadas y mejorar la calidad del servicio ofrecido a los pasajeros.

PRONÓSTICO DEL MERCADO AÉREO EN LAS PRINCIPALES RUTAS COMERCIALES DOMÉSTICAS DE PASAJEROS EN COLOMBIA MEDIANTE ALGORITMOS DE MACHINE LEARNING

Luis Carlos Lozano Caballero

Natalia Mayorga Ballesteros

Juan Carlos Restrepo Blanco

Proyecto Aplicado para optar al título de

Magíster en Ciencia de Datos

Director(a) María Constanza Pabón

FACULTAD DE INGENIERÍA Y CIENCIAS

MAESTRÍA EN CIENCIA DE DATOS

SANTIAGO DE CALI, MAYO 30 DE 2024

TABLA DE CONTENIDO

INTRODUCCIÓN 9
1. DEFINICIÓN DEL PROBLEMA 10
1.1. PLANTEAMIENTO DEL PROBLEMA 10
1.2. FORMULACIÓN DEL PROBLEMA 11
2. OBJETIVOS DEL PROYECTO 12
2.1. OBJETIVO GENERAL 12
2.2. OBJETIVOS ESPECÍFICOS 12
3. MARCO TEÓRICO Y ANTECEDENTES 13
3.1. INDUSTRIA DE LA AVIACIÓN 13
3.1.1. MERCADO AÉREO COLOMBIANO 13
3.1.2. INSTITUCIONES DEL TRANSPORTE AÉREO COLOMBIANO 14
3.1.3. MÉTRICAS DE LA AVIACIÓN 17
3.2. MACHINE LEARNING 17
3.2.1. APLICACIONES DEL MACHINE LEARNING 18
3.2.2. FUNDAMENTOS DEL APRENDIZAJE AUTOMÁTICO 18
3.2.3. ALGORITMOS Y TÉCNICAS DE APRENDIZAJE AUTOMÁTICO 19
3.2.4. PREPROCESAMIENTO DE DATOS 19
3.2.5. SERIES DE TIEMPO 20
3.2.6. REGRESIÓN LINEAL MÚLTIPLE (RLM) 21
3.2.7. REDES NEURONALES 23
3.2.8. MÉTRICAS DE LOS MODELOS 24
3.3. ANTECEDENTES 26
4. ANÁLISIS EXPLORATORIO DE LAS VARIABLES DEL MERCADO DOMÉSTICO AÉREO DE COLOMBIA (OBJETIVO ESPECÍFICO 1) 30
4.1. RECOPILACIÓN DE DATOS 30
4.2. PREPARACIÓN DE LOS DATOS 33
4.3. ANÁLISIS EXPLORATORIO DE DATOS 37
5. SELECCIÓN DE TÉCNICAS DE MACHINE LEARNING PARA PRONÓSTICOS DE UNA VARIABLE A LO LARGO DEL TIEMPO (OBJETIVO ESPECÍFICO 2) 41
6. ENTRENAMIENTO DE MODELOS Y AJUSTE DE PARÁMETROS PARA MEJORAR

LAS MÉTRICAS (OBJETIVO ESPECÍFICO 3) 42
6.1. RESULTADOS DE LA VALIDACIÓN DE LA RLM 42
6.2. RESULTADOS DE LA VALIDACIÓN DE LA ARIMAX 46
6.3. RESULTADOS RED NEURONAL RECURRENTE 57
6.4. RESULTADOS RED NEURONAL SECUENCIAL 59
7. RESULTADOS OBTENIDOS DE EVALUAR Y COMPARAR DIFERENTES ALGORITMOS DE MACHINE LEARNING PARA DETERMINAR CUÁL OFRECE EL MEJOR RENDIMIENTO EN TÉRMINOS DE PRECISIÓN (OBJETIVO ESPECÍFICO 4) 64
7.1. COMPARATIVA DE MODELOS 64
7.1.1. RUTA BOGMDE 64
7.1.2. RUTA BOGCLO 65
7.1.3. RUTA BOGCTG 65
7.1.4. RUTAS GENERALES 66
7.2. PRONÓSTICOS 66
7.3. PROPUESTA DE EFICIENCIA DE CAPACIDAD 67
7.4. TABLERO POWER BI 69
8. CONCLUSIONES Y TRABAJOS FUTUROS 70
8.1. CONCLUSIONES 70
8.2. TRABAJOS FUTUROS 71
9. REFERENCIAS BIBLIOGRÁFICAS 72

LISTA DE FIGURAS

Fig. 1 Growth of airline passenger traffic by world region. 13
Fig. 2 Aplicaciones de machine learning 18
Fig. 3 Ejemplo de bases de datos recopiladas de la página de Aerocivil. 31
Fig. 4 Ejemplo de bases de datos de la TRM. 32
Fig. 5 Ejemplo de bases de datos del BRENT 32
Fig. 6 Consultas en Access 34
Fig. 7 Datos ajustados T01_Final_MS 35
Fig. 8 Dataset con datos de TRM y BRENT mensual 36
Fig. 9 Matriz de correlación 37
Fig. 10 Boxplot 38
Fig. 11 Relaciones y Distribuciones de Variables 39
Fig. 12 Nombre de las 3 principales rutas domésticas de Colombia. 40
Fig. 13 Resumen del modelo RLM RUTA BOGMDE y validación de supuestos 42
Fig. 14 Resumen del modelo RLM RUTA BOGCLO y validación de supuestos 43
Fig. 15 Resumen del modelo RLM RUTA BOGCTG y validación de supuestos 44
Fig. 16 Resumen del modelo RLM MODELO GENERAL y validación de supuestos 45
Fig. 17 Tendencia, estacionalidad y residuo de ARIMAX ruta BOGMDE 47
Fig. 18 estacionalidad ruta BOGMDE 47
Fig. 19 Resume modelo (1.1.1.) ARIMAX ruta BOGMDE 48
Fig. 20 Resume modelo (0.1.1.) ARIMAX ruta BOGMDE 48
Fig. 21 Pronóstico ruta BOGMDE 49
Fig. 22 Tendencia, estacionalidad y residuo de ARIMAX ruta BOGCLO 50
Fig. 23 estacionalidad ruta BOGCLO 50
Fig. 24 Resume modelo (1.1.1.) ARIMAX ruta BOGCLO 51
Fig. 25 Pronóstico ruta BOGCLO 51
Fig. 26 Tendencia, estacionalidad y residuo de ARIMAX ruta BOGCTG 52
Fig. 27 estacionalidad ruta BOGMDE BOGCTG 52
Fig. 28 Resume modelo (1.1.1.) ARIMAX ruta BOGCTG 53
Fig. 29 Resume modelo (0.1.1.) ARIMAX ruta BOGCTG 53
Fig. 30 Pronóstico ruta BOGCTG 54
Fig. 31 Tendencia, estacionalidad y residuo de ARIMAX modelo general 55
Fig. 32 estacionalidad modelo general 55
Fig. 33 Resume modelo (1.1.1.) ARIMAX modelo general 56
Fig. 34 Resume modelo (0.1.1.) ARIMAX modelo general 56
Fig. 35 Pronóstico modelo general 57
Fig. 36 vs real de todas las rutas modelo Red neuronal recurrente 58
Fig. 37 Pérdida durante el entrenamiento, comparación de pronóstico vs real y métricas de evaluación para la ruta BOGMDE 60
Fig. 38 Pérdida durante el entrenamiento, comparación de pronóstico vs real y métricas de evaluación para la ruta BOGCLO 60
Fig. 39 Pérdida durante el entrenamiento, comparación de pronóstico vs real y métricas de

evaluación para la ruta BOGCTG 62
Fig. 40 Pérdida durante el entrenamiento, comparación de pronóstico vs real y métricas de evaluación para la ruta GENERAL 62
Fig. 41 Tablero Power BI 69

LISTA DE TABLAS

Tabla 1 Dataset resultante de la preparación de datos....35
Tabla 2 Dataset usado para entrenamiento y evaluación de modelos 36
Tabla 3. Resultados de las Pruebas de Diagnóstico RML RUTA BOGMDE 43
Tabla 4. Resultados de las Pruebas de Diagnóstico RML RUTA BOGCLO 43
Tabla 5. Resultados de las Pruebas de Diagnóstico RML RUTA BOGCTG 44
Tabla 6. Resultados de las Pruebas de Diagnóstico RML MODELO GENERAL 45
Tabla 7 métricas de los modelos de red neuronal recurrentes para todas las rutas. 58
Tabla 8 comparativa de modelos BOGMDE....64
Tabla 9 comparativa de modelos BOGCLO....65
Tabla 10 comparativa de modelos BOGCTG....65
Tabla 11 comparativa de modelos RUTAS GENERALES....66
Tabla 12 Pronósticos 67
Tabla 13 LF HISTÓRICO 2023 67
Tabla 14 CÁLCULO DE SILLAS PROPUESTAS 68
Tabla 15 PROPUESTA DE EFICIENCIA DE CAPACIDAD 68

INTRODUCCIÓN

El mercado doméstico aéreo colombiano constituye un pilar fundamental en la economía y el turismo del país. En el año 2023, la salida de dos importantes aerolíneas del sector causó una gran pérdida de capacidad en rutas de alta demanda. Esta situación generó un desequilibrio en el mercado, afectando tanto a las aerolíneas restantes como a los pasajeros que dependen de estas rutas para sus desplazamientos. La incertidumbre respecto a la demanda futura representa un desafío significativo para las aerolíneas y otros actores en el mercado aéreo. Ante esta situación, contar con herramientas que permitan prever la evolución del tráfico de pasajeros es fundamental. Este proyecto aplicado, titulado "Pronóstico del Mercado Aéreo en las principales rutas Comerciales domésticas de pasajeros en Colombia mediante algoritmos de Machine Learning", tuvo como objetivo abordar esta problemática y desarrollar un modelo de Machine Learning que permita pronosticar la evolución a un mes de la demanda del mercado aéreo doméstico en Colombia.

La industria aeronáutica en Colombia está creciendo y evolucionando constantemente. El incremento de la población, el desarrollo económico, el turismo y otras actividades han contribuido a un aumento significativo en el número de personas que optan por el transporte aéreo para realizar sus actividades diarias. Las aerolíneas que operan en el sector se enfrentan al desafío de tomar decisiones estratégicas fundamentadas para maximizar su uso, adaptar su oferta de servicios y proporcionar una experiencia eficiente y satisfactoria a los pasajeros.

La aplicación de técnicas de ciencia de datos como machine learning ofrece soluciones prometedoras al mercado aéreo. Poder aprovechar la cantidad de datos históricos de la Aerocivil, permite crear modelos de pronósticos que identifiquen patrones y tendencias para ayudar a los actores principales del mercado a tomar las decisiones necesarias. Al desarrollar modelos de pronóstico, se consigue una herramienta valiosa para planificar de manera más efectiva la capacidad, ajustar la oferta de vuelos y mejorar la calidad del servicio.

1. DEFINICIÓN DEL PROBLEMA

El mercado doméstico aéreo de Colombia es un sector crucial para el desarrollo económico y turístico del país. El crecimiento y la eficiencia de las aerolíneas que operan en este mercado son factores clave para el éxito y la competitividad de la industria, así como las políticas que desarrolle el gobierno para fortalecer dicho mercado. Sin embargo, comprender el comportamiento del mercado aéreo es un desafío complejo debido a la interacción de múltiples variables, como la capacidad de las aerolíneas, el número de pasajeros transportados, la inflación, las tasas de cambio etc. Por lo tanto, es necesario desarrollar modelos y técnicas que permitan pronosticar la evolución futura del mercado aéreo doméstico en Colombia.

1.1. PLANTEAMIENTO DEL PROBLEMA

En 2022, el sector aéreo en Colombia creció transportando 48 millones de pasajeros, un aumento significativo del 57% respecto al 2021. En contraste, en el 2023 dos Aerolíneas nacionales, que tenían el 22 % y el 8 % respectivamente de participación del mercado, se declararon en quiebra, dejando una importante pérdida de capacidad en las rutas en las que operaban. Estas situaciones demuestran lo frágil y complejo que es el mercado aéreo colombiano. De ahí la importancia de saber interpretar y pronosticar el nuevo rumbo que puede tomar dicho mercado.

Es crucial que la Aerocivil y el HUB[1] más importante de Colombia, el Aeropuerto el Dorado, cuenten con herramientas suficientes para tomar decisiones de cara a un futuro lleno de incertidumbres. El crecimiento de la demanda puede saturar la capacidad, por eso es necesario desarrollar modelos y técnicas que permitan pronosticar la evolución futura de las principales rutas del mercado aéreo doméstico en Colombia.

En este contexto, la problemática de ciencia de datos abordada radicó en la utilización de

[1] "Se utiliza el término HUB para denominar los centros de conexión que las aerolíneas establecen en los aeropuertos y que permiten la distribución del tráfico de carga y pasajeros hacia otros destinos en el mundo." [1]

técnicas de machine learning para modelar y prever el mercado aéreo en las principales rutas comerciales domésticas de pasajeros en Colombia.

1.2. FORMULACIÓN DEL PROBLEMA

¿En qué medida los algoritmos de machine learning pueden utilizarse para pronosticar de manera efectiva la demanda del mercado doméstico aéreo de Colombia?

2. OBJETIVOS DEL PROYECTO

2.1. OBJETIVO GENERAL

Modelar mediante técnicas de machine learning la demanda del mercado doméstico aéreo de Colombia utilizando datos históricos de pasajeros transportados por las aerolíneas colombianas con el fin de proponer una mejora de eficiencia de la disposición de capacidad en las principales rutas domésticas de Colombia.

2.2. OBJETIVOS ESPECÍFICOS

- Realizar un análisis exploratorio de las variables del mercado doméstico aéreo de Colombia.
- Identificar técnicas de machine learning para realizar pronósticos de una variable a lo largo del tiempo.
- Entrenar modelos y ajustar parámetros para mejorar las métricas.
- Evaluar y comparar diferentes algoritmos de machine learning para determinar cuál ofrece el mejor rendimiento en términos de precisión.

3. MARCO TEÓRICO Y ANTECEDENTES

En este capítulo se abordan los conceptos fundamentales relacionados con el desarrollo del proyecto. Dado que su objetivo es la creación de un modelo de machine learning para prever la demanda del mercado aéreo doméstico de Colombia, se brindará una explicación detallada sobre los aspectos clave del mercado aéreo colombiano y del machine learning.

3.1. INDUSTRIA DE LA AVIACIÓN

En un mundo globalizado es importante poder realizar conexiones rápidas entre distintas ciudades y países, gracias a los distintos avances tecnológicos, la industria área ha podido darle solución a los largos e incómodos viajes. Durante el 2013 [2], la industria aérea transportó 3.1 mil millones de personas en más de 36 millones de vuelos comerciales. El transporte aéreo ha tenido un crecimiento aproximado del 5% anual. A continuación, vemos el incremento de las aerolíneas por regiones en RPK (definido en la siguiente sección):

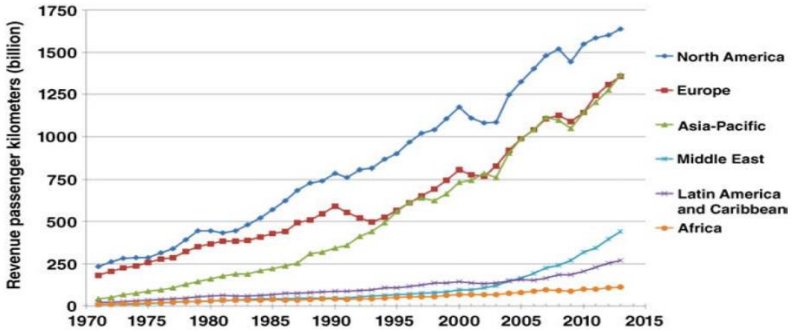


Fig. 1 Growth of airline passenger traffic by world region.

Fuente: tomado de (Data sources: ICAO, 1971–2005; IATA, 2006–2013)

3.1.1. MERCADO AÉREO COLOMBIANO

El incremento del tráfico aéreo en Colombia ha sido impulsado desde los primeros años de la década de 1990 por una serie de políticas públicas que han liberalizado el espacio aéreo tanto a nivel nacional como internacional. Esta liberalización ha sido respaldada por una reorientación de la inversión pública hacia la modernización de las infraestructuras aeroportuarias, incluyendo la concesión de los aeropuertos más transitados del país. La apertura económica de 1991 marcó un

cambio significativo, pasando de una regulación restrictiva a un esquema de libertad vigilada, lo que ha permitido la entrada de nuevos operadores aéreos en el mercado colombiano en los últimos veinte años. Este cambio también ha dado lugar a la participación significativa de capital extranjero en el mercado aéreo colombiano, a través de diversas formas de reorganización empresarial. A nivel internacional, la liberalización se ha llevado a cabo mediante la firma de varios acuerdos bilaterales y acuerdos de cielos abiertos con diversos países. Además, se implementaron concesiones aeroportuarias desde mediados de los 90, acompañadas de reformas institucionales que reforzaron la autoridad aeroportuaria nacional en cuanto a concesiones y descentralización.

3.1.2. INSTITUCIONES DEL TRANSPORTE AÉREO COLOMBIANO

MINISTERIO DEL TRANSPORTE

Es una entidad encargada de coordinar y supervisar las operaciones relacionadas con la infraestructura, el transporte y el tránsito a nivel nacional. En términos de infraestructura, esta responsabilidad abarca la gestión de carreteras, puertos marítimos y fluviales, aeropuertos y sistemas ferroviarios. El Ministerio de Transporte, conforme a lo estipulado en la Ley 489 de 1998, tiene tanto responsabilidades generales como específicas. No obstante, para llevar a cabo eficientemente sus funciones, el Ministerio delega muchas de estas competencias en entidades especializadas bajo su dependencia. En el ámbito del transporte aéreo, por ejemplo, el Ministerio de Transporte confía gran parte de sus competencias y responsabilidades a dos entidades: la Unidad Administrativa Especial de Aeronáutica Civil (Aerocivil) y la Agencia Nacional de Infraestructuras (ANI), tal como se establece en el documento del Ministerio de Transporte del año 2011 [3].

AEROCIVÍL

La Aeronáutica Civil - Aerocivil tiene como objetivo garantizar el desarrollo de la aviación civil y de la administración del espacio aéreo en condiciones de seguridad y eficiencia, en concordancia con las políticas, planes y programas gubernamentales en materia económico-social y de relaciones internacionales. Aerocivil también tiene a su cargo la gestión y administración de los aeropuertos a lo largo del territorio nacional, lo que incluye su mantenimiento, adaptación y construcción según las necesidades del país. La Aerocivil expone un documento regulatorio

conocido como Reglamentos Aeronáuticos de Colombia (RAC) (Aerocivil, 2014). Este documento establece que los RAC deben ser aplicados de manera universal por todos los agentes aeronáuticos, tanto colombianos como extranjeros, que operen en el espacio aéreo colombiano. Los RAC comprenden un conjunto de normativas distribuidas en 20 partes que abarcan todos los aspectos relacionados con la aeronáutica, proporcionando la guía necesaria para cumplir con los requisitos establecidos [4].

AGENCIA NACIONAL DE INFRAESTRUCTURAS (ANI)

La Agencia Nacional de Infraestructura - ANI, es una Agencia Nacional Estatal de Naturaleza Especial, del sector descentralizado de la Rama Ejecutiva del Orden Nacional. Se encarga de planear, coordinar, estructurar, contratar, ejecutar, administrar y evaluar los proyectos de infraestructura que necesita el país como carreteras, aeropuertos puertos y líneas férreas, darles viabilidad y entregarlos a través de asociaciones público-privadas - APP, para el diseño, construcción, mantenimiento, operación, administración y/o explotación de la infraestructura pública de transporte en todos sus modos. Su misión es desarrollar infraestructura a través de Asociaciones Público-Privadas, para generar conectividad, servicios de calidad y desarrollo sostenible. La agencia se basa en el trabajo en equipo y el crecimiento personal y profesional de su talento humano [5].

CONCEPTOS AÉREOS IMPORTANTES

- **Low – Cost:** Aerolínea de bajo costo que ofrece tarifas bajas competitivas, en Colombia podemos encontrar por ejemplo Wingo y JetSMART.
- **Legacy:** Las aerolíneas legacy se caracterizan por su amplia red de rutas, servicios y comodidades ofrecidos a bordo, y una mayor historia y experiencia en la industria de la aviación. Sin embargo, también suelen tener precios más altos en comparación con las aerolíneas de bajo costo debido a los servicios y beneficios adicionales que proporcionan, por ejemplo, Fly Emirates.
- **Hub:** Se refiere a un aeropuerto o una ciudad que sirve como centro de conexión principal para una aerolínea, como por ejemplo el Aeropuerto Internacional El Dorado en Bogotá
- **Point to Point:** Es un modelo de servicio de transporte aéreo en el cual los vuelos operan

directamente desde un punto de origen a un punto de destino sin realizar escalas o conexiones en otros aeropuertos. Un ejemplo en Colombia es Satena, la cual es una aerolínea regional colombiana que opera vuelos directos desde diferentes ciudades a lo largo del país, conectando puntos de origen y destino sin escalas intermedias.

- **Pax:** Término referido a pasajeros.
- **Seats:** Término referido a las sillas.
- **OD RT:** "OD" es una sigla que se utiliza en la industria del transporte para referirse a "Origin-Destination" (Origen-Destino). Se utiliza para indicar la combinación específica de un punto de origen y un punto de destino en una ruta de transporte.

  Por otro lado, "RT" puede referirse a "Round Trip" (Viaje de Ida y Vuelta). Indica que se trata de un viaje en el que el pasajero regresa al punto de origen después de llegar a su destino.
- **OD OW:** Es una abreviatura que se utiliza en la industria del transporte para referirse a "Origin-Destination One-Way" (Origen-Destino de Solo Ida). Se utiliza para indicar un viaje de un solo trayecto desde un punto de origen a un punto de destino, sin incluir el vuelo de regreso.
- **Carrier IATA code:** Son códigos asignados por la Asociación Internacional de Transporte Aéreo (IATA, por sus siglas en inglés) a las aerolíneas de todo el mundo.
- **Rutas aéreas:** Las rutas del transporte aéreo en Colombia se refieren a los itinerarios o trayectorias que los aviones comerciales siguen para transportar pasajeros y carga entre diferentes ciudades y aeropuertos del país.
- **Principales rutas aéreas de Colombia:** Bogotá – Medellín (BOGMED), Bogotá – Cali (BOGCLO) y Bogotá – Cartagena (BOGCTG)
- **Slot:** Se refiere a un intervalo de tiempo asignado a una aerolínea por un aeropuerto para el despegue o aterrizaje de una aeronave.
- **Block Hours:** Se refieren al tiempo de vuelo real, es decir, el tiempo que la aeronave está en el aire, así como el tiempo adicional en tierra para maniobras de rodaje, carga y descarga de pasajeros, mantenimiento, abastecimiento de combustible, preparación de la tripulación,

entre otros.

- **Disruption:** Se refiere a una interrupción o perturbación en el programa o planificación normal de los vuelos. Las interrupciones pueden ser causadas por diversos factores, como condiciones climáticas adversas, problemas técnicos en las aeronaves, huelgas, cierres de aeropuertos, demoras en el tráfico aéreo, entre otros.

3.1.3. MÉTRICAS DE LA AVIACIÓN

- $\boldsymbol{ASK = Available\ Seats\ x\ Kilometers}$

Esta fórmula calcula la capacidad total de un avión o una flota de aviones. Se multiplica el número de asientos disponibles en un vuelo por la distancia total en kilómetros que cubre ese vuelo

- $\boldsymbol{RPK = Pax\ x\ Kilometers}$

Esta fórmula calcula la cantidad total de kilómetros recorridos por los pasajeros en un vuelo o en una serie de vuelos. Se multiplica el número de pasajeros transportados por la distancia total en kilómetros que han volado

- $\boldsymbol{Load\ Factor\ (LF) = \frac{Pax}{Seats} x100}$

Load Factor (LF) o Seat Factor (SF) es la métrica que indica qué tan eficientemente se están utilizando los asientos disponibles en un avión. Se calcula dividiendo los RPK (kilómetros recorridos por pasajeros) por los ASK (kilómetros disponibles para asientos). Luego se multiplica por 100 para obtener un porcentaje.

3.2. MACHINE LEARNING

El machine learning se ha convertido en un pilar fundamental de la informática moderna, destacando la importancia de dotar a las máquinas de inteligencia para análisis predictivos en diversos campos. Para ello, se han desarrollado diversas técnicas que abarcan desde los algoritmos tradicionales de aprendizaje automático, ampliamente aplicados en diversos campos, hasta la emergencia del aprendizaje profundo como una dimensión adicional [6].

3.2.1. APLICACIONES DEL MACHINE LEARNING

Dentro del contexto de nuestro proyecto, destacamos las siguientes aplicaciones del machine learning:

Predicción: El machine learning nos permite realizar pronósticos sobre el comportamiento del mercado aéreo en términos de demanda y oferta en las rutas comerciales domésticas de pasajeros en Colombia.

Análisis de datos: A través de algoritmos de aprendizaje automático, podemos analizar grandes volúmenes de datos históricos para identificar patrones y tendencias en el mercado aéreo.

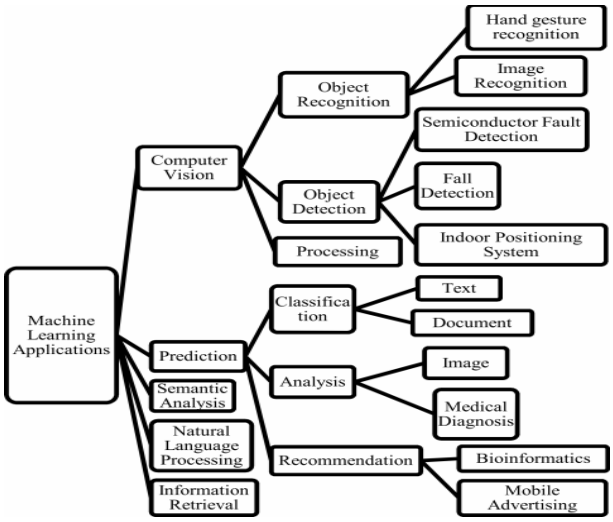


Fig. 2 Aplicaciones de machine learning

Fuente: tomado de [6]

3.2.2. FUNDAMENTOS DEL APRENDIZAJE AUTOMÁTICO

El aprendizaje automático es una subcategoría de la inteligencia artificial que se encarga en desarrollar algoritmos y modelos capaces de aprender y tomar decisiones a partir de los datos [7].

Los datos son lo principal en el aprendizaje automático, ya que para desarrollar algoritmos de aprendizaje automático se necesita de un conjunto de datos para de esta manera lograr aprender patrones y realizar predicciones o tomar decisiones. Estos datos pueden ser numéricos, textuales, imágenes, entre otros [8].

Los algoritmos de aprendizaje automático son procedimientos que se ejecutan en conjuntos

de datos para reconocer patrones y reglas, y los modelos de aprendizaje automático son el resultado del algoritmo. Los modelos actúan como un programa que se puede ejecutar con datos para hacer predicciones [9].

3.2.3. ALGORITMOS Y TÉCNICAS DE APRENDIZAJE AUTOMÁTICO

Los métodos y tareas de aprendizaje automático se clasifican generalmente en tres categorías principales:

- **Supervisado**: En el aprendizaje supervisado, el sistema recibe una serie de ejemplos de entrada y sus correspondientes salidas asignadas. Los algoritmos de aprendizaje supervisado analizan los datos de entrenamiento para generar una función inferida que puede predecir las salidas para nuevas instancias. Idealmente, el algoritmo debe ser capaz de asignar correctamente las etiquetas de clase a instancias no vistas previamente. Para lograr esto, el algoritmo debe generalizar de manera efectiva a partir de los datos de entrenamiento a situaciones no observadas [10].
- **No supervisado:** En este tipo de aprendizaje, el sistema recibe entradas de muestra sin ninguna salida asociada. Dado que no hay resultados deseados proporcionados, la tarea principal consiste en que el algoritmo clasifique o agrupe adecuadamente los conjuntos de datos. El objetivo es definir una función que describa la estructura subyacente de los datos no etiquetados. Como las muestras o conjuntos de entrenamiento no están etiquetados, no hay una señal de recompensa clara para evaluar la efectividad de una solución propuesta [10].

3.2.4. PREPROCESAMIENTO DE DATOS

La etapa de preprocesamiento de datos es esencial en el análisis de datos, particularmente en el contexto del aprendizaje automático. Engloba una serie de técnicas y pasos destinados a mejorar la calidad de los datos crudos y prepararlos adecuadamente para su análisis. Durante esta etapa, se realizan diversas tareas importantes, como la limpieza de datos para eliminar valores faltantes o inconsistentes, la integración de datos de múltiples fuentes, la selección de características relevantes y la normalización de datos para asegurar una representación uniforme. Estas acciones ayudan a evitar errores y mejorar la coherencia de los datos, minimizando la

presencia de ruido y garantizando una mayor consistencia en su análisis posterior.

3.2.5. SERIES DE TIEMPO

Los algoritmos de series de tiempo son métodos analíticos que se utilizan para pronosticar valores futuros con base a patrones históricos de una serie de tiempo. Estos algoritmos, como el modelo ARIMA (Autoregressive Integrated Moving Average) o los modelos de suavizado exponencial, utilizan las observaciones pasadas de la serie para estimar valores futuros. Los modelos ARIMA combinan componentes autorregresivos y de media móvil para capturar patrones y estacionalidad en los datos, mientras que los modelos de suavizado exponencial asignan pesos a diferentes observaciones pasadas para pronosticar el futuro. Estos algoritmos se basan en principios estadísticos y matemáticos para generar pronósticos precisos [11].

Ecuación para un AR:

$$Z_t = \delta + \emptyset_1 Z_{t-1} + \emptyset_2 Z_{t-2} + \cdots + \emptyset_p Z_{t-p} + \epsilon_t$$

Donde:

- Z_t es el valor de la serie temporal en el tiempo t.
- δ es la media del proceso ϵ_t.
- $\emptyset_1$, $\emptyset_2$, $\emptyset_\mathrm{p}$ son los coeficientes de autorregresión.
- ϵ_t es el término del error en el tiempo t.

Además, se introduce el concepto de ARIMAX (Autoregressive Integrated Moving Average with Exogenous Variables), el cual es una extensión lógica del modelado ARIMA puro que incorpora variables exógenas que agregan valor explicativo. Cuando los términos AR y MA en un modelo ARIMA puro no son suficientes para proporcionar un poder explicativo aceptable de un modelo, es natural buscar otros fenómenos impulsores cuya influencia a lo largo del tiempo no está suficientemente arraigada en los valores históricos de las series de tiempo dependientes.

Ecuación para un ARIMAX:

$$Z_t = \delta + \emptyset_1 Z_{t-1} + \emptyset_2 Z_{t-2} + \cdots + \emptyset_p Z_{t-p} + \beta_1 \mathrm{X}_{1,\mathrm{t}} + \beta_2 \mathrm{X}_{2,\mathrm{t}} + \beta_\mathrm{q} \mathrm{X}_{\mathrm{q},\mathrm{t}} + \epsilon_t$$

Donde:

- Z_t es el valor de la serie temporal en el tiempo t.
- δ es la media del proceso ϵ_t.
- $\emptyset_1$, $\emptyset_2$, $\emptyset_p$ son los coeficientes de autorregresión.
- $X_{1,t}$, $X_{2,t}$, $X_{q,t}$ son las variables exógenas en el tiempo t.
- β_1, β_2, β_q son los coeficientes asociados a las variables exógenas.
- ϵ_t es el término del error en el tiempo t.

En las series de tiempo ARIMAX se pueden validar los siguientes supuestos:

- Correlación de residuos: La correlación de residuos implica verificar si los residuos del modelo ARIMAX están correlacionados entre sí. Si los residuos muestran correlación significativa, esto puede indicar que el modelo no ha capturado completamente la estructura de autocorrelación en los datos.
- Residuos distribuidos normalmente: Este supuesto implica verificar si los residuos del modelo ARIMAX siguen una distribución normal. La normalidad de los residuos es importante para realizar inferencias válidas sobre los parámetros del modelo.
- Estacionalidad: La estacionalidad implica verificar si hay patrones estacionales o periódicos en los residuos del modelo ARIMAX. Esto se puede hacer visualmente al observar los residuos en diferentes retrasos para detectar cualquier patrón estacional. La presencia de estacionalidad en los residuos puede indicar que el modelo no ha capturado completamente la estructura temporal de los datos, y puede ser necesario ajustar el modelo para tener en cuenta estas variaciones estacionales.

3.2.6. REGRESIÓN LINEAL MÚLTIPLE (RLM)

Es un modelo estadístico que busca establecer una relación lineal entre una variable dependiente con dos o más variables independientes. Es una extensión de la regresión lineal simple, que solo considera una variable independiente. En la regresión lineal múltiple, se asume que la variable dependiente puede ser explicada por una combinación lineal de las variables independientes, ponderadas por coeficientes [12]. El modelo trata de encontrar los valores óptimos

para estos coeficientes de tal manera que la suma de los residuos al cuadrado (diferencias entre los valores observados y los predichos por el modelo) sea mínima. La RLM se puede expresar de la siguiente forma:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$$

Donde:

- Y es la variable dependiente que se está tratando de pronosticar.
- β_0 es la intersección o el término de sesgo.
- $\beta_1, \beta_2, \beta_{\mathrm{n}}$ son los coeficientes de regresión asociados a cada variable independiente.
- X_1, X_2, X_{n} son las variables independientes o predictoras.
- ε es el término del error.

En la regresión lineal múltiple, se validan los siguientes supuestos:

- Linealidad: Este supuesto establece que la relación entre las variables independientes y la variable dependiente se puede representar de manera lineal. En otras palabras, los valores de la variable dependiente están determinados por un modelo lineal que puede ser expresado mediante una combinación lineal de las variables independientes [13].
- Homocedasticidad: Este supuesto indica que las perturbaciones aleatorias tienen una varianza constante en todos los niveles de las variables independientes. Esto implica que la dispersión de los errores alrededor de la línea de regresión es constante en todos los valores de las variables predictoras [13].
- Independencia: Este supuesto establece que las perturbaciones aleatorias asociadas con cada observación son independientes entre sí. En otras palabras, no hay relación sistemática entre los errores de diferentes observaciones en el modelo de regresión [13].
- Normalidad: Este supuesto supone que la distribución de los residuos, que son las diferencias entre los valores observados y los valores predichos por el modelo, sigue una distribución normal. Esto implica que los residuos se distribuyen simétricamente alrededor de cero y que la mayoría de los residuos se concentran cerca de cero, con una dispersión

decreciente a medida que nos alejamos de cero. La normalidad de los residuos es importante para la inferencia estadística y para la precisión de los intervalos de confianza y pruebas de hipótesis en el análisis de regresión [13].

3.2.7. REDES NEURONALES

Una red neuronal es un modelo simplificado que emula el modo en que el cerebro humano procesa la información: Funciona simultaneando un número elevado de unidades de procesamiento interconectadas que parecen versiones abstractas de neuronas.

Las unidades de procesamiento se organizan en capas. Hay tres partes normalmente en una red neuronal: una capa de entrada, con unidades que representan los campos de entrada; una o varias capas ocultas; y una capa de salida, con una unidad o unidades que representa el campo o los campos de destino. Las unidades se conectan con fuerzas de conexión variables (o ponderaciones). Los datos de entrada se presentan en la primera capa, y los valores se propagan desde cada neurona hasta cada neurona de la capa siguiente al final, se envía un resultado desde la capa de salida.

La red aprende examinando los registros individuales, generando una predicción para cada registro y realizando ajustes a las ponderaciones cuando realiza una predicción incorrecta. Este proceso se repite muchas veces y la red sigue mejorando sus predicciones hasta haber alcanzado uno o varios criterios de parada [14]. Se puede expresar de la siguiente forma:

$$Y = f\left(\sum_{i=1}^{n} w_{\mathrm{i}} * X_i + \mathrm{b}\right)$$

Donde:

- Y es la salida de la neurona.
- f es la función de activación, que introduce no linealidad en la red.
- X_i son las entradas.
- w_{i} son los pesos asociados a cada entrada.
- b es el sesgo.

3.2.7.1 RED NEURONAL RECURRENTE (RNN):

Las redes neuronales recurrentes (RNN) son un tipo especializado de red neuronal que tiene

conexiones retroalimentadas entre las neuronas, lo que permite que la información se propague en ambas direcciones a través del tiempo. Esto les permite capturar patrones temporales complejos en los datos, haciéndolas especialmente útiles para modelar secuencias temporales o datos con dependencias temporales. Las RNN son capaces de recordar información de eventos pasados y utilizarla en el procesamiento de eventos futuros, lo que las hace adecuadas para una amplia gama de aplicaciones, como el procesamiento de lenguaje natural, la traducción automática, la generación de texto y la predicción de series temporales. Sin embargo, las RNN también pueden sufrir de problemas de desvanecimiento y explosión de gradientes, lo que ha llevado al desarrollo de variantes más avanzadas, como las redes neuronales LSTM (Long Short-Term Memory) y GRU (Gated Recurrent Unit), que abordan estos problemas específicos.

3.2.7.2 RED NEURONAL SECUENCIAL:

Las redes neuronales secuenciales son un tipo de modelo de redes neuronales en el que las capas de neuronas se organizan de manera secuencial, donde cada capa se conecta a la siguiente en una sola dirección. Estas redes son adecuadas para procesar datos que tienen una estructura secuencial o temporal, como series temporales, texto o secuencias de datos. En una red neuronal secuencial, la información fluye en una dirección, desde la capa de entrada hasta la capa de salida, sin la existencia de conexiones retroalimentadas entre las capas. Esto hace que las redes neuronales secuenciales sean eficaces para problemas de predicción, clasificación y regresión en los que la estructura secuencial de los datos es importante para el análisis y la toma de decisiones.

3.2.8. MÉTRICAS DE LOS MODELOS

RLM (Regresión Lineal Múltiple):

- **R-squared (R^2):** Mide la proporción de la variabilidad de la variable dependiente que es explicada por las variables independientes del modelo.
- **Durbin-Watson:** Estadístico que detecta la presencia de autocorrelación en los residuos del modelo.

- **RMSE (Root Mean Square Error):** Raíz cuadrada del promedio de los cuadrados de los errores, mide la diferencia entre los valores predichos por el modelo y los valores observados.

ARIMAX (Autoregressive Integrated Moving Average with Exogenous Variables):

- **Log Likelihood:** Medida de la bondad del ajuste del modelo.
- **AIC (Akaike Information Criterion):** Criterio para la selección de modelos que penaliza la complejidad del modelo. Este criterio no solo evalúa la bondad de ajuste del modelo a los datos, sino que también tiene en cuenta la cantidad de parámetros utilizados, penalizando aquellos modelos que son excesivamente complejos. Así, se busca un balance óptimo entre la calidad del ajuste y la simplicidad del modelo.
- **BIC (Bayesian Information Criterion):** Similar al AIC, pero con una penalización mayor para la complejidad del modelo. El BIC también considera tanto el ajuste del modelo a los datos como la simplicidad del modelo, pero aplica una penalización más estricta a la cantidad de parámetros. Esto significa que el BIC tiende a preferir modelos más simples en comparación con el AIC, especialmente cuando el tamaño de la muestra es grande. De esta forma, se asegura que la calidad del modelo no se vea comprometida por un sobreajuste debido a un exceso de parámetros.
- **R-squared (R^2):** Evalúa qué tan bien se ajusta el modelo a los datos observados. Es una medida de la proporción de la varianza en la variable de salida que es predecible a partir de las variables de entrada.

Red Neuronal Recurrente:

- **Pérdida durante el entrenamiento:** Se monitorea la función de pérdida a lo largo de las épocas de entrenamiento para evaluar el ajuste del modelo y detectar problemas de sobreajuste.
- **R-squared (R^2):** Evalúa qué tan bien se ajusta el modelo a los datos observados. Es una medida de la proporción de la varianza en la variable de salida que es predecible a partir de las variables de entrada.

- **RMSE:** Medida de la diferencia entre los valores predichos por un modelo y los valores observados

Red Neuronal Secuencial:

- **R-squared (R^2):** Evalúa qué tan bien se ajusta el modelo a los datos observados. Es una medida de la proporción de la varianza en la variable de salida que es predecible a partir de las variables de entrada.
- **RMSE:** Medida de la diferencia entre los valores predichos por un modelo y los valores observados

3.3. ANTECEDENTES

- **Forecasting Air Passenger Demand: A Comparison of LSTM and SARIMA:** Este estudio propone un modelo de pronóstico de la demanda de pasajeros aéreos basado en la Red Neural Profunda (Deep Neural Network DNN), específicamente, el algoritmo de Memoria a Largo Corto Plazo (LSTM), también se aplica al problema de investigación el método de Promedio Móvil Integrado Auto Regresivo Estacional (SARIMA), todo ello evaluando diferentes criterios de rendimiento para evaluar la precisión del pronóstico. En este documento se abordan temas como el análisis de series de tiempo, redes neuronales y las diferentes técnicas empleadas. También se describe la metodología empleada para llegar a los resultados. Los resultados obtenidos en este estudio son el desarrollo de modelos de predicción a corto y mediano plazo basados en LSTM. El modelo a mediano plazo realiza pronósticos mensuales, mientras que el de corto plazo se centra en predicciones semanales. La precisión del modelo LSTM fue alta, como lo confirmó el RMSE de los datos de validación. Comparado con el modelo SARIMA, ambos mostraron buen rendimiento a corto plazo, pero el LSTM fue superior a mediano plazo debido a su capacidad para aprender datos no lineales y su memoria a largo plazo. Sin embargo, el LSTM requiere más tiempo y datos para alcanzar alta precisión y necesita ajustar muchos parámetros. Por otro lado, el modelo SARIMA es menos preciso, pero más sencillo y eficiente con pocos datos[15].

- **Forecasting Air Travel Demand for Selected Destinations Using Machine Learning Methods:** En este estudio, se introdujo un método para determinar la capacidad óptima de asientos que puede proporcionar el factor de ocupación más alto para la operación de vuelo entre dos países cualesquiera. Aquí se describen los diferentes métodos de aprendizaje automático que se desarrollaron para la investigación, como lo es la Red Neuronal Artificial (ANN), Regresión Lineal (LR), Gradient Boosting (GB) y Random Forest (RF). Los resultados mostraron que el método GB produjo la menor MAPE (Error Porcentual Absoluto Medio) para el período de 2014-2019, destacando su eficacia. Se identificó la importancia del período de análisis justo antes de la pandemia de COVID-19 en la industria de viajes aéreos. Se sugirió investigar más a fondo los métodos RF y ANN para mejorar los resultados, así como analizar los efectos globales de la pandemia en diferentes modos de transporte. Se propuso considerar variables de temporada y tarifas de billetes en futuros estudios para obtener una comprensión más completa de la demanda de pasajeros. Este estudio es una muy buena fuente de información para el presente proyecto, ya que puede ser un referente sobre las técnicas que se pueden utilizar para el pronóstico del mercado aéreo en Colombia [16].
- **Estimación de la demanda del tráfico de pasajeros en vuelos domésticos de los 5 aeropuertos más importantes de Colombia:** En este trabajo de grado se encuentra cómo se logra estimar la demanda de pasajeros en vuelos domésticos en Colombia. De igual manera, se reúne información de 5 aeropuertos y se observa en cuánto afectan las variables escogidas con sustento teórico. Debido a limitaciones de datos, el análisis simplificó la homogeneidad de los pasajeros y tarifas aéreas, lo que podría haber afectado la precisión de los resultados. Sin embargo, estos resultados proporcionan una orientación sobre las tendencias del mercado y sugieren áreas para futuras investigaciones.

  Aunque las simplificaciones redujeron el alcance de los resultados, se aplicó la mejor metodología posible con los datos disponibles. A medida que mejore la calidad y disponibilidad de los datos, será importante considerar una modelización más realista que incluya la heterogeneidad de los consumidores y las características diferenciadas de los bienes. Esta información brinda gran ayuda para el presente proyecto, puesto que puede

brindar información relevante sobre las variables que se deben tener en cuenta para realizar el pronóstico de mercado aéreo en Colombia [17].

- **Pronóstico de la demanda de pasajeros aéreos en nueve aeropuertos regionales colombianos:** En este trabajo se describe el desarrollo de un análisis de los pronósticos de tráfico aéreo a corto plazo (5 años) de pasajeros nacionales e internacionales para nueve Aeropuertos regionales de Colombia. Se desarrollaron dos modelos de series de tiempo, el primero empleando la metodología de Box–Jenkins el modelo ARIMA y el segundo un modelo lineal dinámico (MLD). Se realizó una comparativa entre los resultados de dichos pronósticos y el tráfico realmente observado, y su correspondiente análisis. Los resultados mostraron la comparación entre los modelos ARIMA y el Modelo Lineal Dinámico y que ambos tienen ventajas y desventajas, pero el Modelo Lineal Dinámico demostró ser más dinámico y adaptable a diferentes series de datos. También se concluyó que, aunque ambos modelos son útiles para pronosticar la demanda de pasajeros, el Modelo Lineal Dinámico mostró una mayor flexibilidad y facilidad de ajuste a las series analizadas [18].

  La descripción del desarrollo de este trabajo es de gran utilidad como soporte y guía para el presente trabajo, ya que se tiene un referente de estos dos modelos de series de tiempo.

- **Análisis del desarrollo reciente del transporte aéreo en Colombia:** El transporte aéreo colombiano está experimentando un vertiginoso y dinámico desarrollo, nunca visto en toda su historia, que dio inicio aproximadamente hace dos décadas. Este periodo coincide con el inicio y la implementación continuados de políticas públicas elaboradas específicamente para el sector del transporte aéreo orientadas en dos líneas bien definidas, por un lado, a la desregulación y liberalización de la industria, por otro lado, a la privatización de las infraestructuras aeroportuarias. Por ello, este trabajo tiene el objetivo de describir y analizar en qué medida las políticas públicas fijadas e implementadas por las instituciones públicas colombianas contribuyeron al desarrollo reciente de la industria del transporte aéreo [19].

Los antecedentes revisados proporcionan un marco sólido para abordar el pronóstico de la demanda del mercado aéreo en Colombia. Se ha demostrado que el uso de técnicas de aprendizaje

automático, como las Redes Neuronales Recurrentes (RNN), los modelos de series temporales (ARIMA, ARIMAX) y otras técnicas como Regresión Lineal, Gradient Boosting y Random Forest, ha sido efectivo en la predicción de la demanda de pasajeros en diferentes contextos.

El análisis del desarrollo del transporte aéreo en Colombia destaca la importancia de las políticas públicas en este sector, y cómo estás han impactado en su crecimiento dinámico en las últimas décadas. Es esencial considerar estos factores externos al modelar la demanda futura, ya que pueden influir significativamente en los patrones de viaje y en la capacidad de las aerolíneas para satisfacer la demanda del mercado.

Este estudio se propone como un ejercicio académico para mejorar las técnicas de pronóstico y una herramienta práctica para contribuir al desarrollo estratégico y la toma de decisiones en la industria del transporte aéreo en Colombia.

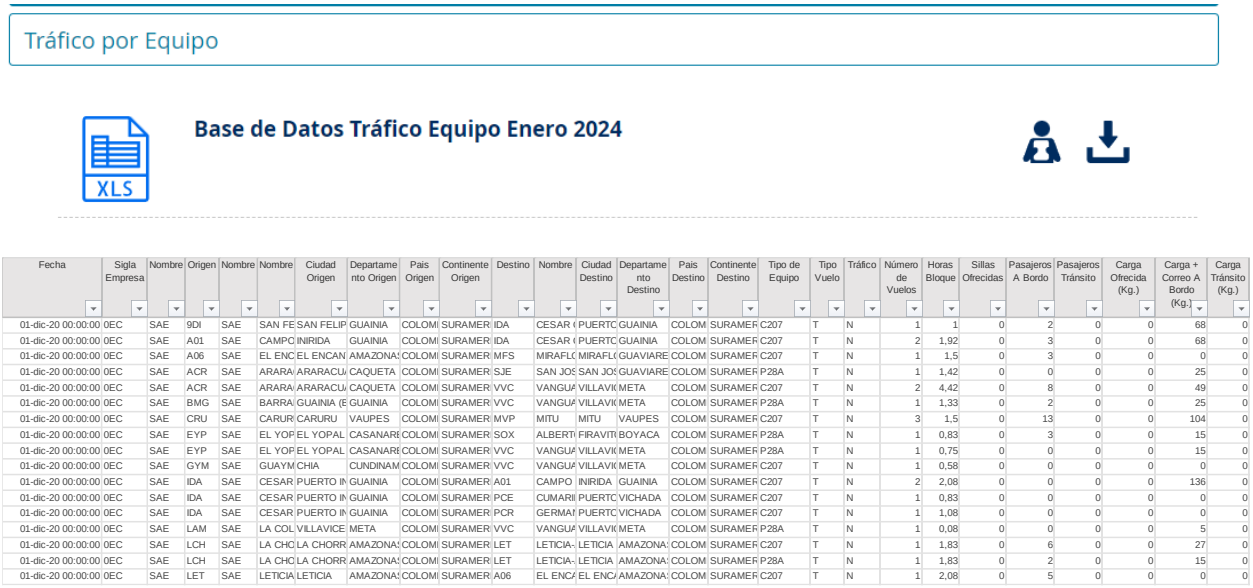
4. ANÁLISIS EXPLORATORIO DE LAS VARIABLES DEL MERCADO DOMÉSTICO AÉREO DE COLOMBIA (OBJETIVO ESPECÍFICO 1)

4.1. RECOPILACIÓN DE DATOS

Esta fase se enfocó en la obtención de los datos que publica la Aerocivil sobre las rutas y vuelos de las aerolíneas. Estos datos se descargaron directamente de la página oficial de la Aerocivil[2], ya que son datos públicos. Se descargaron de la base tráfico por equipo[3] un total de 170 archivos correspondientes al tráfico mes a mes, un archivo por cada mes, desde enero del 2010 hasta febrero del 2024, estas bases de datos tenían 28 columnas, que incluía información como fecha, sigla de la empresa, nombre, nombre y sigla del aeropuerto de origen y destino, ciudad de origen y destino del vuelo, continente - departamento y país de origen y destino, información más específica sobre el tipo de vuelo[4], equipo (modelo del avión), horas bloqueadas[5], número de vuelo, carga ofrecida y transitada y el total de sillas ofrecidas y pasajeros transitados, cada uno de estos

[2] https://www.aerocivil.gov.co/atencion/estadisticas-de-las-actividades-aeronauticas/bases-de-datos
[3] La palabra equipo hace referencia a los diferentes tipos de Aeronaves usadas en el mercado comercial
[4] Tipos de vuelos: A (adicional), C (charter), T (taxi), R (regular)
5 Se refieren al tiempo total transcurrido desde que una aeronave comienza a moverse por su propio impulso para despegar hasta que se detiene por completo al finalizar el vuelo

archivos tiene aproximadamente 2600 registros para cada mes. A continuación se muestra un ejemplo de nuestras bases de datos recopiladas.



Fecha	Sigla Empresa	Nombre	Origen	Nombre	Nombre	Ciudad Origen	Departamento Origen	Pais Origen	Continente Origen	Destino	Nombre	Ciudad Destino	Departamento Destino	Pais Destino	Continente Destino	Tipo de Equipo	Tipo Vuelo	Tráfico	Número de Vuelos	Horas Bloque	Sillas Ofrecidas	Pasajeros A Bordo	Pasajeros Tránsito	Carga Ofrecida (Kg.)	Carga + Correo A Bordo (Kg.)	Carga Tránsito (Kg.)
01-dic-20 00:00:00	0EC	SAE	9DI	SAE	SAN FE	SAN FELIP	GUAINIA	COLOM	SURAMER	IDA	CESAR (	PUERTC	GUAINIA	COLOM	SURAMER	C207	T	N	1	1	0	2	0	0	68	0
01-dic-20 00:00:00	0EC	SAE	A01	SAE	CAMPO	INIRIDA	GUAINIA	COLOM	SURAMER	IDA	CESAR (	PUERTC	GUAINIA	COLOM	SURAMER	C207	T	N	2	1,92	0	3	0	0	68	0
01-dic-20 00:00:00	0EC	SAE	A06	SAE	EL ENC	EL ENCAN	AMAZONA	COLOM	SURAMER	MFS	MIRAFLO	MIRAFLO	GUAVIARE	COLOM	SURAMER	C207	T	N	1	1,5	0	3	0	0	0	0
01-dic-20 00:00:00	0EC	SAE	ACR	SAE	ARARA	ARARACU	CAQUETA	COLOM	SURAMER	SJE	SAN JOS	SAN JOS	GUAVIARE	COLOM	SURAMER	P28A	T	N	1	1,42	0	0	0	0	25	0
01-dic-20 00:00:00	0EC	SAE	ACR	SAE	ARARA	ARARACU	CAQUETA	COLOM	SURAMER	VVC	VANGUA	VILLAVI	META	COLOM	SURAMER	C207	T	N	2	4,42	0	8	0	0	49	0
01-dic-20 00:00:00	0EC	SAE	BMG	SAE	BARRA	GUAINIA (E	GUAINIA	COLOM	SURAMER	VVC	VANGUA	VILLAVI	META	COLOM	SURAMER	P28A	T	N	1	1,33	0	2	0	0	25	0
01-dic-20 00:00:00	0EC	SAE	CRU	SAE	CARUR	CARURU	VAUPES	COLOM	SURAMER	MVP	MITU	MITU	VAUPES	COLOM	SURAMER	C207	T	N	3	1,5	0	13	0	0	104	0
01-dic-20 00:00:00	0EC	SAE	EYP	SAE	EL YOP	EL YOPAL	CASANARE	COLOM	SURAMER	SOX	ALBERT	FIRAVIT	BOYACA	COLOM	SURAMER	P28A	T	N	1	0,83	0	3	0	0	15	0
01-dic-20 00:00:00	0EC	SAE	EYP	SAE	EL YOP	EL YOPAL	CASANARE	COLOM	SURAMER	VVC	VANGUA	VILLAVI	META	COLOM	SURAMER	P28A	T	N	1	0,75	0	0	0	0	15	0
01-dic-20 00:00:00	0EC	SAE	GYM	SAE	GUAYM	CHIA	CUNDINAM	COLOM	SURAMER	VVC	VANGUA	VILLAVI	META	COLOM	SURAMER	C207	T	N	1	0,58	0	0	0	0	0	0
01-dic-20 00:00:00	0EC	SAE	IDA	SAE	CESAR	PUERTO IN	GUAINIA	COLOM	SURAMER	A01	CAMPO	INIRIDA	GUAINIA	COLOM	SURAMER	C207	T	N	2	2,08	0	0	0	0	136	0
01-dic-20 00:00:00	0EC	SAE	IDA	SAE	CESAR	PUERTO IN	GUAINIA	COLOM	SURAMER	PCE	CUMARI	PUERTC	VICHADA	COLOM	SURAMER	C207	T	N	1	0,83	0	0	0	0	0	0
01-dic-20 00:00:00	0EC	SAE	IDA	SAE	CESAR	PUERTO IN	GUAINIA	COLOM	SURAMER	PCR	GERMA	PUERTC	VICHADA	COLOM	SURAMER	C207	T	N	1	1,08	0	0	0	0	0	0
01-dic-20 00:00:00	0EC	SAE	LAM	SAE	LA COL	VILLAVICE	META	COLOM	SURAMER	VVC	VANGUA	VILLAVI	META	COLOM	SURAMER	P28A	T	N	1	0,08	0	0	0	0	5	0
01-dic-20 00:00:00	0EC	SAE	LCH	SAE	LA CHO	LA CHORR	AMAZONA	COLOM	SURAMER	LET	LETICIA-	LETICIA	AMAZONA	COLOM	SURAMER	C207	T	N	1	1,83	0	6	0	0	27	0
01-dic-20 00:00:00	0EC	SAE	LCH	SAE	LA CHO	LA CHORR	AMAZONA	COLOM	SURAMER	LET	LETICIA-	LETICIA	AMAZONA	COLOM	SURAMER	P28A	T	N	1	1,83	0	2	0	0	15	0
01-dic-20 00:00:00	0EC	SAE	LET	SAE	LETICIA	LETICIA	AMAZONA	COLOM	SURAMER	A06	EL ENC/	EL ENC/	AMAZONA	COLOM	SURAMER	C207	T	N	1	2,08	0	5	0	0	0	0

Fig. 3 Ejemplo de bases de datos recopiladas de la página de Aerocivil.

Fuente: tomado de [4]

Puesto que la correlación entre el TRM (Tasa Representativa del Mercado) y el precio del Brent con el pronóstico del mercado aéreo tienen una influencia directa en los costos operativos y la demanda del transporte aéreo. El TRM afecta el costo de los insumos importados y los precios de los boletos aéreos para los pasajeros, mientras que el precio del Brent influye en los costos del combustible, que es uno de los principales gastos operativos de las aerolíneas. Por lo tanto, variaciones en estos indicadores económicos impactan significativamente la rentabilidad y la competitividad de las aerolíneas, así como la demanda de los pasajeros, afectando directamente la previsión del mercado aéreo, es por esto que estas dos variables fueron incluidas en nuestros modelos.

Se descargó el histórico de la Tasa Representativa del Mercado (TRM)[6], Estas bases de datos contienen la información diaria de la TRM.

Precio del dólar TRM por día

Click en los encabezados para ordenar.

Fecha	Valor
viernes 1 de enero de 2010	2044.23
sábado 2 de enero de 2010	2044.23
domingo 3 de enero de 2010	2044.23
lunes 4 de enero de 2010	2044.23
martes 5 de enero de 2010	2021.21
miércoles 6 de enero de 2010	1992.78

Fig. 4 Ejemplo de bases de datos de la TRM.

Fuente: tomado de [20]

Los datos históricos del precio del petróleo (BRENT)[7] se logró obtener una base de datos con el valor mensual, en este caso la base de datos contaba con las columnas de: Fecha. Último, Apertura, Máximo, Mínimo, Vol., % var; para nuestro ejercicio se tuvo en cuenta el promedio del primer valor registrado y el último de cada mes.

Plazo

Monthly

Descargar datos

01.01.2010 - 31.12.2023

Fecha	Último	Apertura	Máximo	Mínimo	Vol.	% var.
01.12.2023	77,04	80,67	81,72	72,29	4,66M	-4.72%
01.11.2023	80,86	84,78	87,12	76,71	5,82M	-4.89%
01.10.2023	85,02	90,48	92,32	82,20	6,35M	-10.80%

Fig. 5 Ejemplo de bases de datos del BRENT

Fuente: tomado de [21]

[6] https://dolar.wilkinsonpc.com.co/2010
[7] https://es.investing.com/commodities/brent-oil-historical-data

4.2. PREPARACIÓN DE LOS DATOS

Todos los datos históricos descargados se integraron en un archivo de Access donde se ejecutaron consultas para el análisis, limpieza y mejoramiento de atributos. En cuanto al análisis de los datos suministrados por la Aerocivil este se realizó para comprender su estructura y características. Se encontró que no todos los archivos descargados tenían los mismos atributos. Desde el año 2018 hacia atrás, tenían varias diferencias con respecto a los archivos de los años más recientes. Algunos archivos incluían carga ofrecida, carga tránsito, distancia, mientras que otros no. También se evidenció que el nombre de los atributos variaba de posición según el año. En este caso se analizó la relevancia de cada uno de los atributos para el proyecto y se decidió eliminar las columnas de carga ofrecida, carga tránsito, distancia, sigla de la empresa, nombre, nombre del aeropuerto de origen y destino, tipo de vuelo, equipo (modelo del avión), horas bloqueadas, número de vuelo. Recordemos que gran parte de la operación comercial en Colombia se lleva a cabo en equipos similares, Airbus A320 o Boeing 737, ambos con configuraciones similares, lo que nos permite despreciar el número de vuelos y el tipo de equipo. Las horas bloque no son significativas debido a que no se realiza un análisis de costos. Los atributos que hacen referencia a las aerolíneas se eliminaron porque el proyecto analizó los datos del total de pasajeros en las rutas relevantes. Se eliminan los nombres de “aeropuerto de origen y destino”, ya que se decide emplear los códigos IATA de las rutas.

También se cambiaron los nombres de las columnas que estaban diferentes, ya que en algunos archivos el atributo “sillas” aparecía como "sillas ofrecidas", “origen” en algunos archivos aparecía como “ciudad origen”, y se organizaron las columnas que estaban en posiciones diferentes para que todos los archivos quedaran en un mismo formato, todo lo anterior se realizó mediante consultas en Access. Además de esto, se halló que el mes de mayo de 2019 no tenía datos, por lo que se decidió analizar la línea de tendencia de ese año y de los años anteriores, ya que en los meses de abril y mayo se sigue una tendencia similar para los demás años, por lo cual se decidió tomar el valor de abril de 2019 para completar el dato faltante.

Se unificaron los archivos para crear el histórico y se realizaron diferentes consultas para

entender mejor los datos y determinar qué atributos eran necesarios para el proyecto. En este caso, la consulta A04_OD nos permitió hacer agrupamientos por OD OW y OD RT[8], ya que en esta consulta guardamos la información de origen y destino de los vuelos. Otra consulta, como la A01_YYMM, nos permitió extraer el mes y año de los datos, y las demás consultas sirvieron para el análisis de los datos. La figura 6 muestra las consultas que se desarrollaron en Access para la limpieza de los datos.

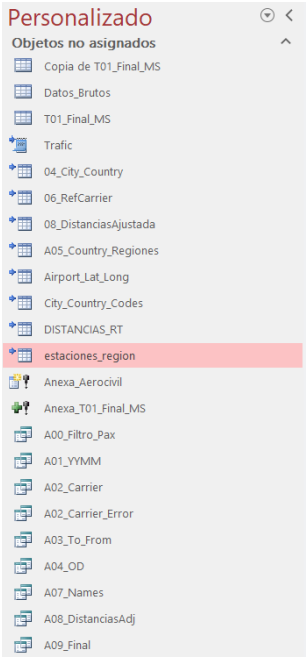


Fig. 6 Consultas en Access

Fuente: Elaboración propia

Luego de realizar la preparación de los datos de la Aerocivil se consolidó la información en la tabla T01_Final_MS (figura 7).

- [8] **OD RT:** Se refiere a "Origin-Destination" (Origen-Destino). Se utiliza para indicar la combinación específica de unpunto de origen y un punto de destino en una ruta de transporte. Por otro lado, "RT" puede referirse a "Round Trip" (Viaje de Ida y Vuelta). Indica que se trata de un viaje en el que el pasajero regresa al punto de origen después de llegar a su destino.

T01_Final_MS_1

YYYY	MM	Org	Des	OD_OW	OD_RT	City_Org	City_Des	Ctry_Org	Ctry_Des	Reg_Org	Reg_Dst	Tipo de Vuel
2019	5	BOG	YYZ	BOGYYZ	BOGYYZ	BOGOTA	TORONTO	CO	CA	COL	NAM	R
2019	5	YYZ	BOG	YYZBOG	BOGYYZ	TORONTO	BOGOTA	CA	CO	NAM	COL	R
2019	5	BOG	MAD	BOGMAD	BOGMAD	BOGOTA	MADRID	CO	ES	COL	EUR	R
2019	5	MAD	BOG	MADBOG	BOGMAD	MADRID	BOGOTA	ES	CO	EUR	COL	R
2019	5	BOG	CDG	BOGCDG	BOGCDG	BOGOTA	PARIS	CO	FR	COL	EUR	R
2019	5	CDG	BOG	CDGBOG	BOGCDG	PARIS	BOGOTA	FR	CO	EUR	COL	R
2019	5	BOG	CUN	BOGCUN	BOGCUN	BOGOTA	CANCUN	CO	MX	COL	MEX	A
2019	5	BOG	CUN	BOGCUN	BOGCUN	BOGOTA	CANCUN	CO	MX	COL	MEX	R
2019	5	BOG	CUN	BOGCUN	BOGCUN	BOGOTA	CANCUN	CO	MX	COL	MEX	R
2019	5	BOG	MEX	BOGMEX	BOGMEX	BOGOTA	MEXICO	CO	MX	COL	MEX	R

Fig. 7 Datos ajustados T01_Final_MS

Fuente: Elaboración propia.

Acá se muestra una descripción detallada del dataset resultante que cuenta con 368.626 registros.

Tabla 1 Dataset resultante de la preparación de datos

Atributo	Descripción	Tipo de dato
YYYY	Año del vuelo	Numérico
MM	Mes del vuelo	Numérico
Org	Código IATA del aeropuerto de origen	Categórico
Des	Código IATA del aeropuerto de destino	Categórico
OD_OW	Combinación de los códigos IATA de origen y destino para un solo trayecto	Categórico
OD_RT	Combinación de los códigos IATA de origen y destino para un trayecto redondo	Categórico
City_Org	Ciudad de origen	Categórico
City_Des	Ciudad de destino	Categórico
Ctry_Org	Código de país de origen	Categórico
Ctry_Des	Código de país de destino	Categórico
Reg_Org	Región de origen	Categórico
Reg_Dst	Región de destino	Categórico
Tipo de Vuelo	Tipo de vuelo (R: Regular, C: Charter)	Categórico
Trafico	Tipo de tráfico (I: Internacional, D: Doméstico)	Categórico
Sillas	Número de asientos disponibles en el vuelo	Numérico
Pasajeros A Bordo	Número de pasajeros a bordo	Numérico
COL_To	Código de tráfico desde Colombia	Categórico
COL_From	Código de tráfico hacia Colombia	Categórico
Flight_Region	Región de vuelo (e.g., NAMALL para todos los vuelos en América del Norte)	Categórico
Distance_Adj	Distancia ajustada del vuelo en millas	Numérico

Fuente: Elaboración propia.

Para las bases de datos del histórico de la Tasa Representativa del Mercado (TRM), dado que los datos de la Aerocivil se registran mensualmente, se decidió calcular el promedio de la TRM de todos los días del mes para obtener el valor mensual. Y como los datos históricos del precio del petróleo (BRENT), se lograron obtener una base de datos con el valor mensual, se tomó el promedio del primer valor registrado y el último de cada mes, estas dos bases de datos se unieron resultando una tabla como muestra la figura 8.

YYYY	MM	TRM	BRENT
2010	1	$ 1.983,43	$ 77,99
2010	2	$ 1.951,72	$ 75,69
2010	3	$ 1.908,99	$ 80,73
2010	4	$ 1.937,45	$ 86,44
2010	5	$ 1.983,59	$ 78,65
2010	6	$ 1.926,85	$ 76,74
2010	7	$ 1.874,41	$ 76,31

Fig. 8 Dataset con datos de TRM y BRENT mensual

Fuente: Elaboración propia.

Para el dataset usado en el entrenamiento y evaluación de los modelos, se consolidaron los atributos de año, mes, pasajeros, BRENT y TRM, aunque en la sesión de análisis exploratorio se tuvieron en cuenta las sillas para confirmar tendencias, patrones y correlaciones. El resto de las variables de georreferenciación mencionadas en la tabla 1 se descartaron al momento de filtrar por rutas o consolidar los datos para el modelo general. Al final se obtienen 144 registros correspondientes a datos desde enero del 2010 hasta diciembre del 2023, exceptuando 2020 y 2022, ya que estos dos años por tema de pandemia el mercado aéreo estuvo paralizado y los únicos vuelos que se realizaron fueron vuelos humanitarios que no son parte de nuestro estudio.

Tabla 2 Dataset usado para entrenamiento y evaluación de modelos

YYYY	MM	PAX	BRENT	TRM
2010	1	135.872	$ 77,99	$ 1.983,43
2010	2	133.046	$ 75,69	$ 1.951,72
2010	3	166.853	$ 80,73	$ 1.908,99
2010	4	154.121	$ 86,44	$ 1.937,45
2010	5	166.196	$ 78,65	$ 1.983,59
2010	6	174.293	$ 76,74	$ 1.926,85
2010	7	187.947	$ 76,31	$ 1.874,41

Fuente: Elaboración propia.

4.3. ANÁLISIS EXPLORATORIO DE DATOS

En este análisis exploratorio de datos, examinamos la relación entre las variables YYYY (año), MM (mes), PAX (pasajeros), Seats (sillas), BRENT (precio del petróleo) y TRM (tasa de cambio). Aquí identificamos patrones y relaciones significativas que influyen en el número de pasajeros.

A continuación, en la figura 9 se presenta una matriz de correlación entre las variables, acompañada de gráficos de boxplot y de tendencias. Esto permite visualizar tanto las relaciones cuantitativas entre las variables como la distribución y dispersión de cada una. Hemos utilizado la ruta general para este análisis, ya que, al realizarlo individualmente por rutas, se observa una similitud en los resultados obtenidos, lo que justifica el uso del modelo general para una interpretación más amplia.

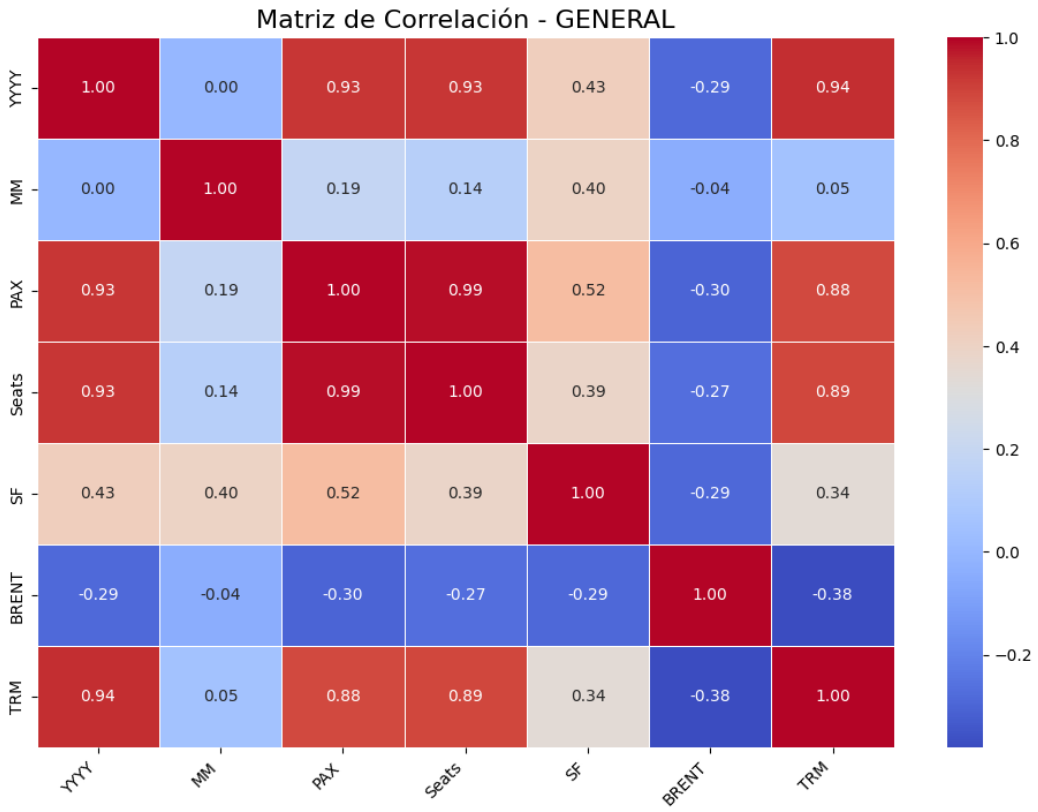


Fig. 9 Matriz de correlación

Fuente: Elaboración propia.

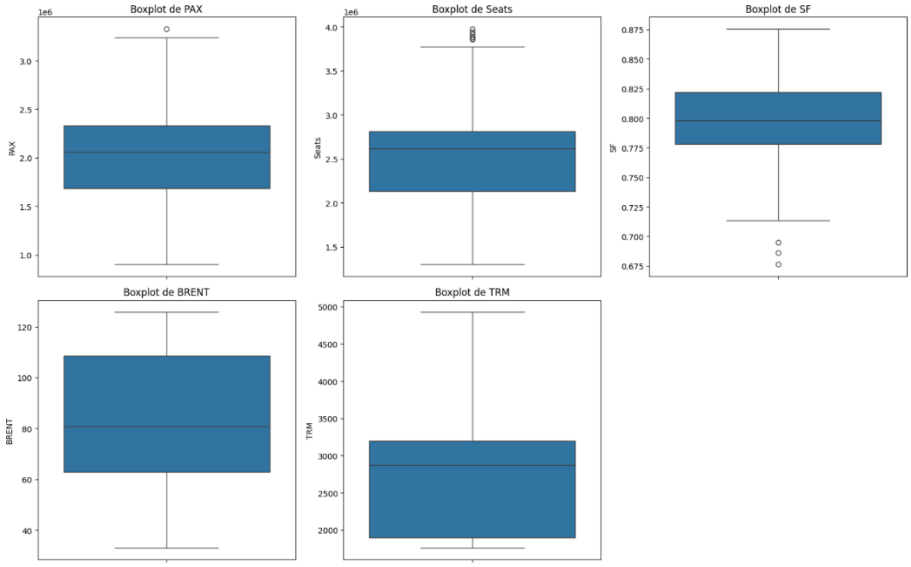


Fig. 10 Boxplot

Fuente: Elaboración propia.

En la figura 10 podemos identificar que datos atípicos en nuestro gráfico de SF y de Seats. Esto lo podemos analizar cómo nuevas rutas que se agregan a la operación, por lo cual el factor de llenado de sillas al inicio es inferior y se normaliza a lo largo del tiempo.

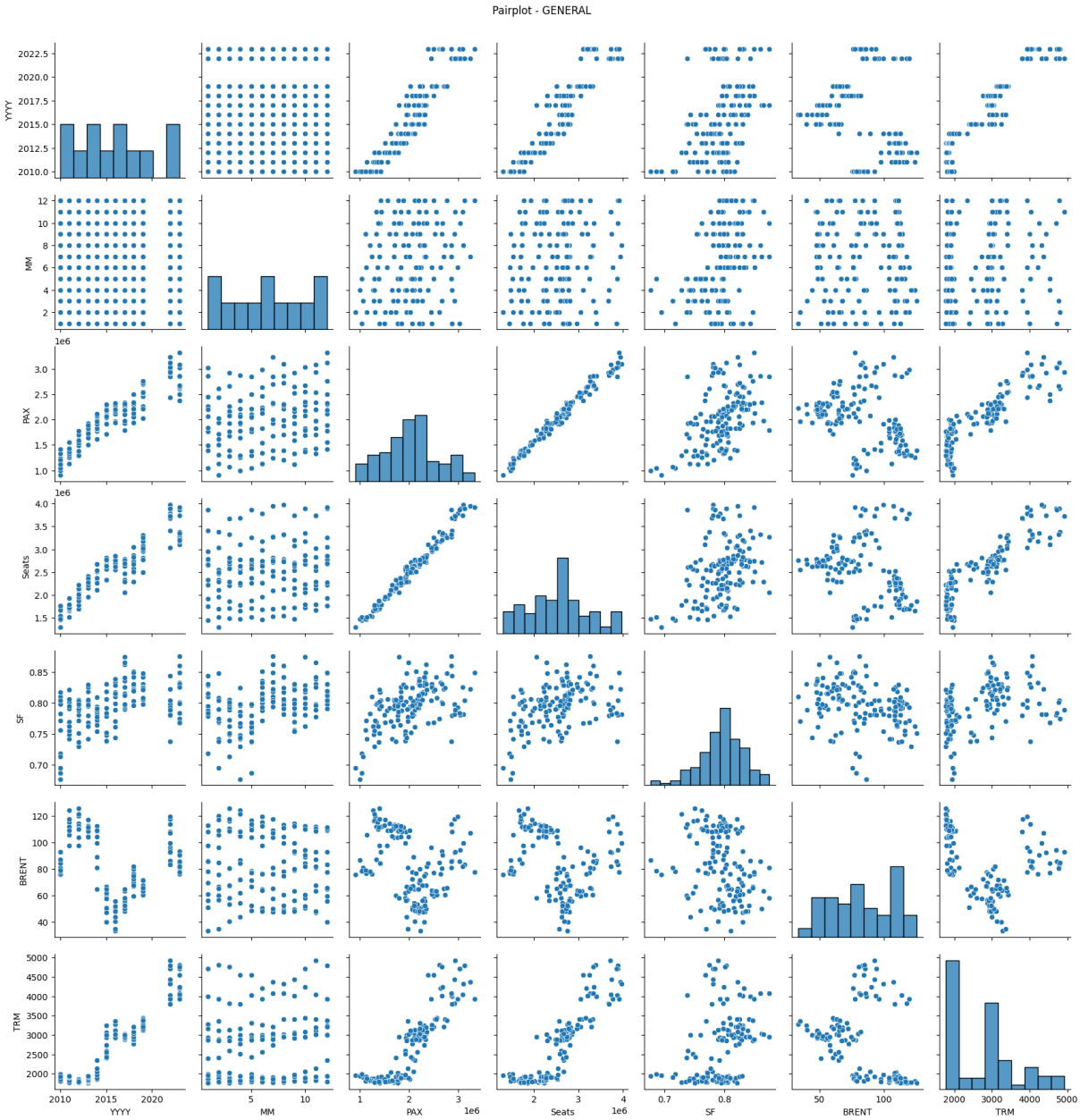


Fig. 11 Relaciones y Distribuciones de Variables

Fuente: Elaboración propia.

En la figura 11 podemos observar el análisis de correlación para la ruta general que revela varias relaciones importantes entre las variables estudiadas. La variable 'PAX' (pasajeros) muestra una alta correlación positiva con 'Seats' (Sillas), indicando que a medida que aumenta el número

de pasajeros, también lo hace la cantidad de asientos. 'PAX' también está fuertemente correlacionada con 'YYYY' (año) y 'TRM' (tasa de cambio), lo que sugiere una tendencia temporal y una posible influencia económica en el número de pasajeros. Por otro lado, el 'BRENT' (precio del petróleo) tiene una correlación negativa débil con la mayoría de las variables, destacando que los costos de combustible, aunque afectan el precio de los tiquetes, no tienen una relación directa fuerte con las demás variables

A continuación, en la figura 12 se presentan las principales rutas en transporte de pasajeros dentro de Colombia Bogotá – Medellín (BOGMED), Bogotá – Cali (BOGCLO) y Bogotá – Cartagena (BOGCTG).

Fig. 12 Nombre de las 3 principales rutas domésticas de Colombia.

Fuente: Elaboración propia

5. SELECCIÓN DE TÉCNICAS DE MACHINE LEARNING PARA PRONÓSTICOS DE UNA VARIABLE A LO LARGO DEL TIEMPO (OBJETIVO ESPECÍFICO 2)

Se diseñó un conjunto de experimentos que incluye regresión lineal múltiple (RLM), series temporales y Redes Neuronales secuenciales y recurrentes. Esta elección se basó en la comprensión de las fortalezas y limitaciones individuales de cada enfoque, con el propósito de generar pronósticos más robustos y precisos.

La regresión lineal múltiple se aplicó para capturar relaciones lineales entre las variables independientes y dependientes, ofreciendo una interpretación clara a través de coeficientes fácilmente comprensibles. Este método demostró ser efectivo en la modelización de relaciones directas entre las variables.

Por otra parte, se integró el modelo ARIMAX de series temporales para abordar patrones temporales más complejos, tales como tendencias y estacionalidades. Esto permitió una mayor capacidad para analizar y pronosticar el comportamiento del mercado a lo largo del tiempo.

Además, se implementaron redes neuronales recurrentes y secuenciales para capturar relaciones no lineales y patrones a largo plazo en los datos, considerando la interdependencia de las observaciones a lo largo del tiempo. La red neuronal secuencial proporcionó una flexibilidad y capacidad de adaptación que complementó los enfoques más tradicionales de RLM y series temporales.

6. ENTRENAMIENTO DE MODELOS Y AJUSTE DE PARÁMETROS PARA MEJORAR LAS MÉTRICAS (OBJETIVO ESPECÍFICO 3)

6.1. RESULTADOS DE LA VALIDACIÓN DE LA RLM

Se programaron en Python los modelos de Regresión Lineal Múltiple (RLM) para las rutas Bogotá-Medellín (BOGMDE), Bogotá-Cali (BOGCLO), Bogotá-Cartagena (BOGCTG), y un modelo general que englobó todas las rutas, sumando un total de cuatro modelos de RLM.

Los datos se separaron en un conjunto de entrenamiento y otro de prueba. Dado que las series presentan tendencia, se decidió dividir el conjunto de datos de manera que se mantuvieran años completos. Para el entrenamiento, se utilizaron datos mensuales desde enero de 2010 hasta diciembre de 2022, excluyendo los años 2020 y 2021 debido a la pandemia. Para la prueba, se emplearon datos desde enero hasta diciembre del 2023. Esta división se aplicó a los cuatro modelos de RLM, en resumen 92% (132 registros) de los datos para entrenamiento y 8% (12 registros) para prueba. A continuación, se presentan los resultados para las rutas BOGMDE, BOGCLO, BOGCTG y modelo general.

RUTA BOGMDE:

	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.878e+07	6.27e+06	-6.188	0.000	-5.12e+07	-2.64e+07
YYYY	1.943e+04	3128.721	6.209	0.000	1.32e+04	2.56e+04
MM	7572.3810	1143.671	6.621	0.000	5305.893	9838.869
TRM	-44.8394	15.139	-2.962	0.004	-74.841	-14.838
BRENT	-372.3763	176.900	-2.105	0.038	-722.950	-21.802

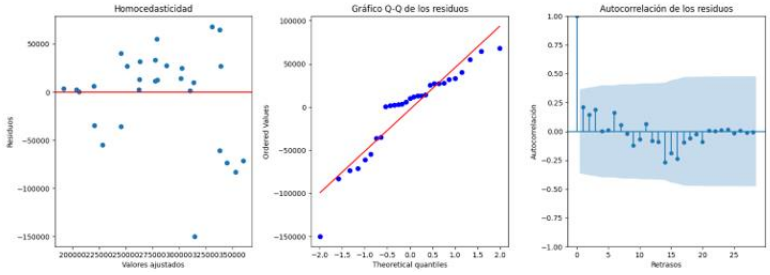


Fig. 13 Resumen del modelo RLM RUTA BOGMDE y validación de supuestos

Fuente: Elaboración propia

Tabla 3. Resultados de las Pruebas de Diagnóstico RML RUTA BOGMDE

	p-valor	n	Decisión
Homocedasticidad (Breusch-Pagan test)	0,104	29	Los residuos tienen varianza constante.
Normalidad (Shapiro-Wilk test)	0,198	29	Los residuos siguen una distribución normal.
Independencia (Durbin-Watson test)	1,553	29	No hay una autocorrelación significativa en los residuos del modelo

Fuente: Elaboración propia

RUTA BOGCLO:

	coef	std err	t	P>|t|	[0.025	0.975]
const	-9.913e+06	5.18e+06	-1.913	0.058	-2.02e+07	3.55e+05
YYYY	5045.8531	2587.084	1.950	0.054	-81.141	1.02e+04
MM	4643.9488	945.681	4.911	0.000	2769.830	6518.067
TRM	-8.8313	12.518	-0.705	0.482	-33.639	15.977
BRENT	-503.7929	146.275	-3.444	0.001	-793.676	-213.909

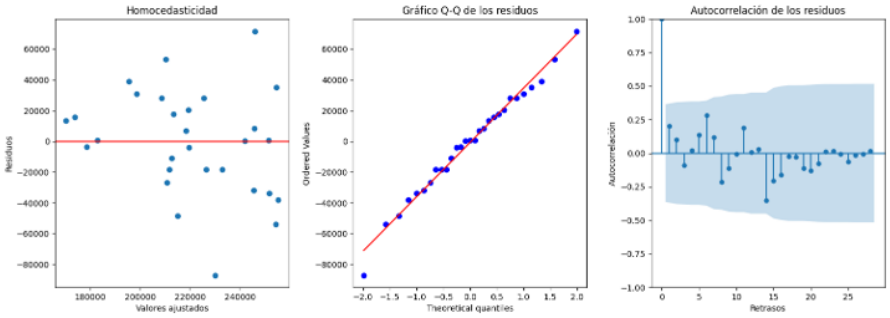


Fig. 14 Resumen del modelo RLM RUTA BOGCLO y validación de supuestos

Fuente: Elaboración propia

Tabla 4. Resultados de las Pruebas de Diagnóstico RML RUTA BOGCLO

	p-valor	n	Decisión
Homocedasticidad (Breusch-Pagan test)	0,298	29	Los residuos tienen varianza constante.
Normalidad (Shapiro-Wilk test)	0,996	29	los residuos siguen una distribución normal
Independencia (Durbin-Watson test)	1,561	29	No hay una autocorrelación significativa en los residuos del modelo

Fuente: Elaboración propia

RUTA BOGCTG:

	coef	std err	t	P>|t|	[0.025	0.975]
const	-2.471e+07	4.33e+06	-5.701	0.000	-3.33e+07	-1.61e+07
YYYY	1.237e+04	2163.935	5.715	0.000	8079.551	1.67e+04
MM	3380.2916	791.004	4.273	0.000	1812.708	4947.875
TRM	-10.5709	10.471	-1.010	0.315	-31.321	10.179
BRENT	-253.3871	122.350	-2.071	0.041	-495.857	-10.918

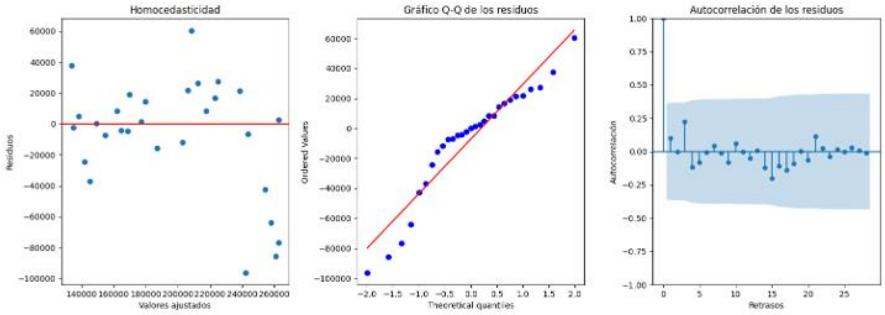


Fig. 15 Resumen del modelo RLM RUTA BOGCTG y validación de supuestos

Fuente: Elaboración propia

Tabla 5. Resultados de las Pruebas de Diagnóstico RML RUTA BOGCTG

	p-valor	n	Decisión
Homocedasticidad (Breusch-Pagan test)	0,063	29	Los residuos tienen varianza constante.
Normalidad (Shapiro-Wilk test)	0,126	29	Los residuos siguen una distribución normal.
Independencia (Durbin-Watson test)	1,712	29	No hay una autocorrelación significativa en los residuos del modelo

Fuente: Elaboración propia

MODELO GENERAL:

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.799e+08	4.3e+07	-4.181	0.000	-2.65e+08	-9.46e+07
YYYY	8.989e+04	2.15e+04	4.185	0.000	4.73e+04	1.32e+05
MM	3.655e+04	7851.510	4.656	0.000	2.1e+04	5.21e+04
TRM	134.6598	103.931	1.296	0.198	-71.307	340.627
BRENT	634.5856	1214.450	0.523	0.602	-1772.169	3041.341

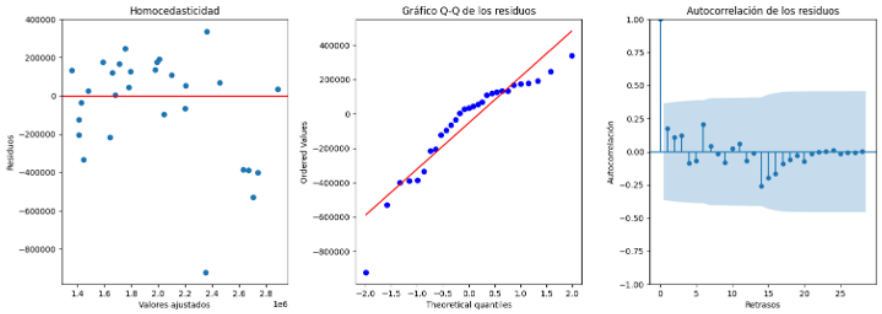


Fig. 16 Resumen del modelo RLM MODELO GENERAL y validación de supuestos

Fuente: Elaboración propia

Tabla 6. Resultados de las Pruebas de Diagnóstico RML MODELO GENERAL

	p-valor	n	Decisión
Homocedasticidad (Breusch-Pagan test)	0,239	29	Los residuos tienen varianza constante.
Normalidad (Shapiro-Wilk test)	0,256	29	Los residuos siguen una distribución normal.
Independencia (Durbin-Watson test)	1,575	29	No hay una autocorrelación significativa en los residuos del modelo

Fuente: Elaboración propia

En el modelo BOGMDE, todas las variables son estadísticamente significativas y se cumple los supuestos de homocedasticidad, normalidad e independencia, además presenta un RMSE de 48.535 y un R^2 = 0,59. En el modelo BOGCLO, las variables año y TRM no son significativas, pero los residuos cumplen con la homocedasticidad, normalidad e independencia, además presenta un RMSE de 33.660 y un R^2 = 0,36. En el modelo BOGCTG, la variable TRM no es significativa, aunque los residuos cumplen con los supuestos de homocedasticidad, normalidad e independencia, además presenta un RMSE de 36.836 y un R^2 = 0,69. Finalmente, en el modelo GENERAL, las

variables BRENT y TRM no son significativas, pero se cumplen los supuestos de la regresión lineal múltiple, lo que permite su uso para pronósticos, además presenta un RMSE de 275.979 y un $R^2 =$ 0,73.

6.2. RESULTADOS DE LA VALIDACIÓN DE LA ARIMAX

Se programaron en Python los modelos de Arimax para las rutas Bogotá-Medellín (BOGMDE), Bogotá-Cali (BOGCLO), Bogotá-Cartagena (BOGCTG), y un modelo general que englobó todas las rutas, sumando un total de cuatro modelos de Arimax.

Los datos se separaron en un conjunto de entrenamiento y otro de prueba. Dado que las series presentan tendencia, se decidió dividir el conjunto de datos de manera que se mantuvieran años completos. Para el entrenamiento, se utilizaron datos mensuales desde enero de 2010 hasta diciembre de 2022, excluyendo los años 2020 y 2021 debido a la pandemia. Para la prueba, se emplearon datos desde enero hasta diciembre del 2023. Esta división se aplicó a los cuatro modelos de Arimax, en resumen 92% (132 registros) de los datos para entrenamiento y 8% (12 registros) para prueba. A continuación, se presentan los resultados para las rutas BOGMDE, BOGCLO, BOGCTG y modelo general.

RUTA BOGMDE

Como primer paso, se realizó un gráfico de los datos para identificar tendencia, estacionalidad y residuos, descomponiendo la serie en estas tres partes tal como se muestra a continuación:

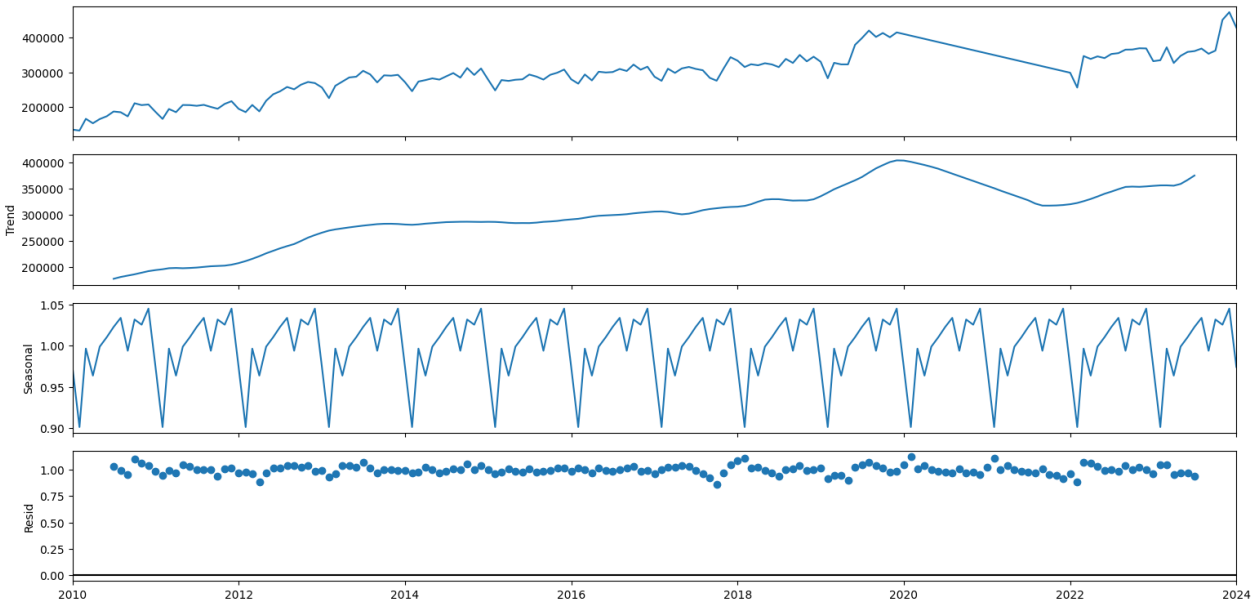


Fig. 17 Tendencia, estacionalidad y residuo de ARIMAX ruta BOGMDE

Fuente: Elaboración propia

Sin considerar los datos del 2020 y 2021 debido a la pandemia, se observó que los valores tenían una tendencia creciente con el paso de los años, además de presentar estacionalidad, lo cual era de esperar dado que más personas se transportan en temporadas vacacionales. Se verificó la estacionalidad de la serie de forma gráfica:

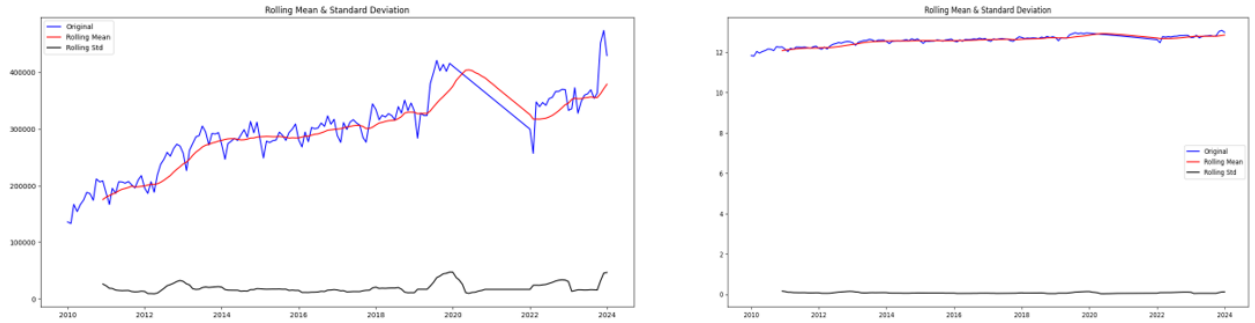


Fig. 18 estacionalidad ruta BOGMDE

Fuente: Elaboración propia

Se observó que la serie no era estacionaria debido a que su varianza no era constante a lo largo del tiempo (gráfica izquierda), por lo que se decidió aplicar una transformación logarítmica. Se evidencio que la varianza era constante a lo largo del tiempo después de la transformación

(gráfica derecha) y los datos estaban listos para generar el modelo.

SARIMAX Results

Dep. Variable:	Pax	No. Observations:	132
Model:	ARIMA(1, 1, 1)	Log Likelihood	152.993
Date:	Sun, 26 May 2024	AIC	-295.986
Time:	16:20:29	BIC	-281.610
Sample:	0	HQIC	-290.144
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
TRM	-0.3572	0.158	-2.260	0.024	-0.667	-0.047
BRENT	-0.1607	0.059	-2.735	0.006	-0.276	-0.046
ar.L1	0.3569	0.231	1.548	0.122	-0.095	0.809
ma.L1	-0.6355	0.206	-3.084	0.002	-1.039	-0.232
sigma2	0.0057	0.001	8.151	0.000	0.004	0.007

Fig. 19 Resume modelo (1.1.1.) ARIMAX ruta BOGMDE

Fuente: Elaboración propia

Se generó un modelo Arimax (1,1,1) tal como se muestra en la imagen, el cual indica que las variables exógenas del modelo son significativas, pero el componente Autorregresivo no lo es. Se decidió ejecutar un modelo (0,1,1) para comparar los resultados.

SARIMAX Results

Dep. Variable:	Pax	No. Observations:	132
Model:	ARIMA(0, 1, 1)	Log Likelihood	152.293
Date:	Sun, 26 May 2024	AIC	-296.586
Time:	19:10:23	BIC	-285.085
Sample:	0	HQIC	-291.913
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
TRM	-0.4277	0.162	-2.642	0.008	-0.745	-0.110
BRENT	-0.1836	0.061	-3.017	0.003	-0.303	-0.064
ma.L1	-0.3007	0.076	-3.948	0.000	-0.450	-0.151
sigma2	0.0057	0.001	8.312	0.000	0.004	0.007

Fig. 20 Resume modelo (0.1.1.) ARIMAX ruta BOGMDE

Fuente: Elaboración propia

Se evidencio que ambos modelos tienen AIC y BIC similares, pero con el fin de garantizar que todos los componentes de la serie fueran estadísticamente significativos se deja el modelo

Arimax (0,1,1) para realizar los pronósticos.

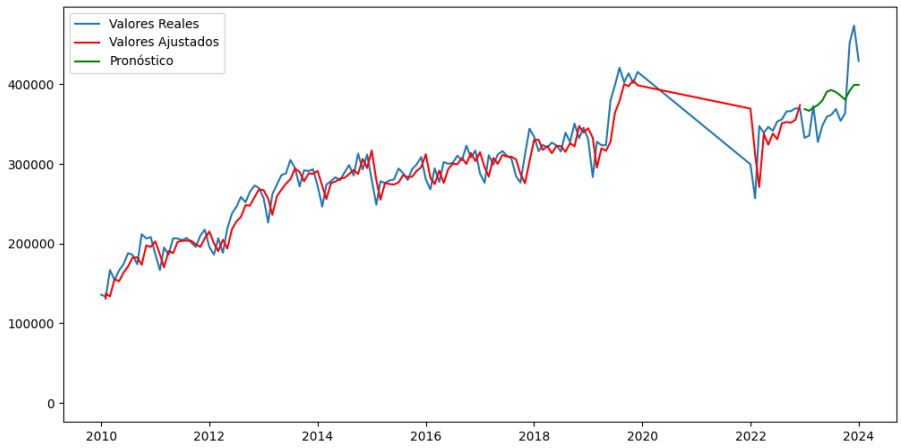


Fig. 21 Pronóstico ruta BOGMDE

Fuente: Elaboración propia

Cabe recordar que los datos de pandemia no fueron tenidos en cuenta para el entrenamiento, y la gráfica realizó una interpolación con el fin de evitar datos nulos. Al realizar el pronóstico para los datos de entrenamiento, se observó un buen ajuste de la serie, pero en los datos de prueba la serie no se asemejaba a los valores reales. Esto podría explicarse por la quiebra de las aerolíneas Viva y Ultra, lo que ocasionó una contracción momentánea del mercado. Al calcular el coeficiente de determinación R^2 para el conjunto de entrenamiento, se obtuvo un valor de 0.85 y un RMS de 38.411.

RUTA BOGCLO

Se identificaron las tendencias y estacionalidades de los datos para la ruta:

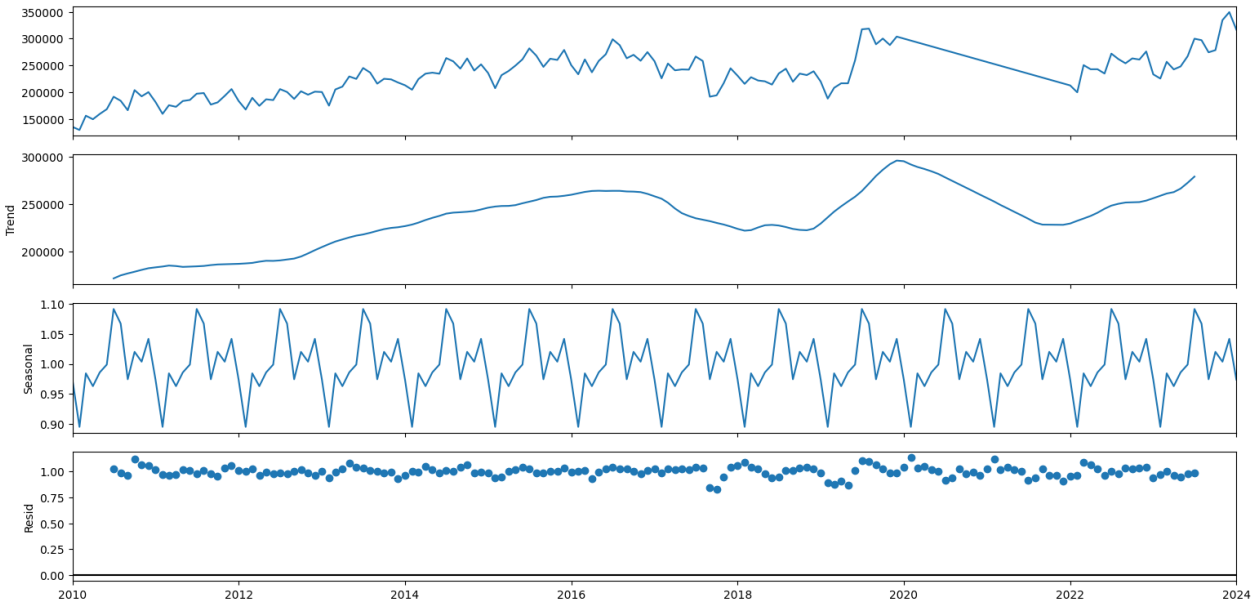


Fig. 22 Tendencia, estacionalidad y residuo de ARIMAX ruta BOGCLO

Fuente: Elaboración propia

Se evidenció una tendencia positiva en los datos, exceptuando los años de pandemia. La estacionalidad de los datos era acorde a la naturaleza del negocio, con un mayor transporte de pasajeros en temporada de vacaciones, especialmente en junio y diciembre.

A continuación, se verificó la estacionalidad de la serie de forma gráfica.

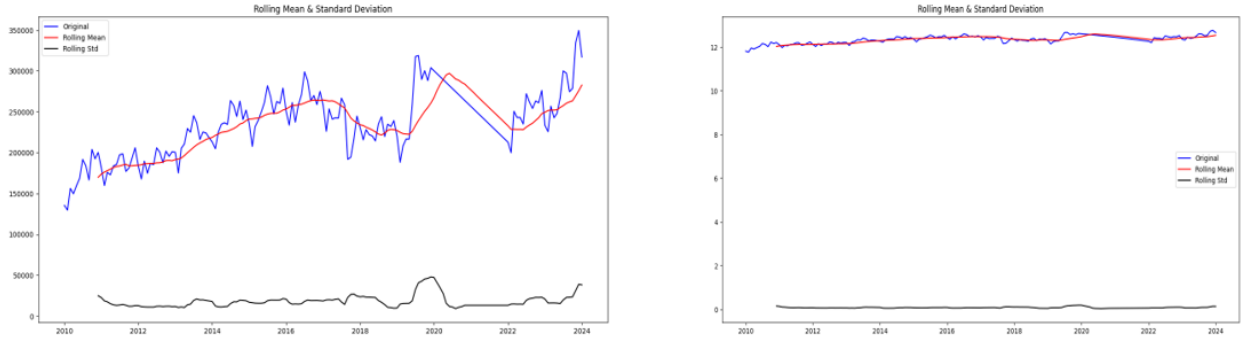


Fig. 23 estacionalidad ruta BOGCLO

Fuente: Elaboración propia

Se observó que la serie no era estacionaria debido a que su varianza no era constante a lo

largo del tiempo (gráfico izquierdo). Por esta razón, se decidió aplicar una transformación logarítmica y se observó que la varianza se volvió constante a lo largo del tiempo (gráfico derecho) y los datos estaban listos para generar el modelo.

SARIMAX Results

Dep. Variable:	Pax	No. Observations:	132
Model:	ARIMA(1, 1, 1)	Log Likelihood	138.841
Date:	Mon, 27 May 2024	AIC	-267.681
Time:	21:38:42	BIC	-253.305
Sample:	0	HQIC	-261.839
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
TRM	-0.2481	0.166	-1.496	0.135	-0.573	0.077
BRENT	-0.1776	0.070	-2.519	0.012	-0.316	-0.039
ar.L1	0.5250	0.161	3.263	0.001	0.210	0.840
ma.L1	-0.7985	0.135	-5.907	0.000	-1.063	-0.534
sigma2	0.0070	0.001	9.400	0.000	0.006	0.008

Fig. 24 Resume modelo (1.1.1.) ARIMAX ruta BOGCLO

Fuente: Elaboración propia

Se observó que para el modelo Arimax(1,1,1), la variable exógena TRM no era estadísticamente significativa. Se graficó el pronóstico para los valores de entrenamiento y de prueba.

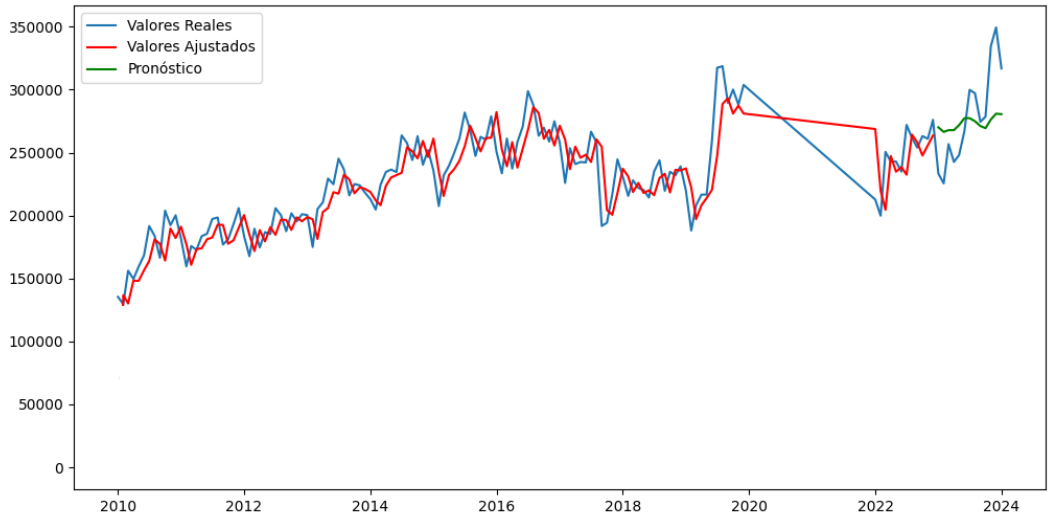


Fig. 25 Pronóstico ruta BOGCLO

Fuente: Elaboración propia

Tal como se evidenció en el modelo Arimax para la ruta BOGMDE, el pronóstico de los datos de prueba se vio afectado por la quiebra de las aerolíneas Viva y Ultra. Al realizar el cálculo del coeficiente de determinación R^2, se encontró un valor de 0.65 y un RMS de 33.925.

RUTA BOGCTG

Se identificaron las tendencias y estacionalidades de los datos para la ruta:

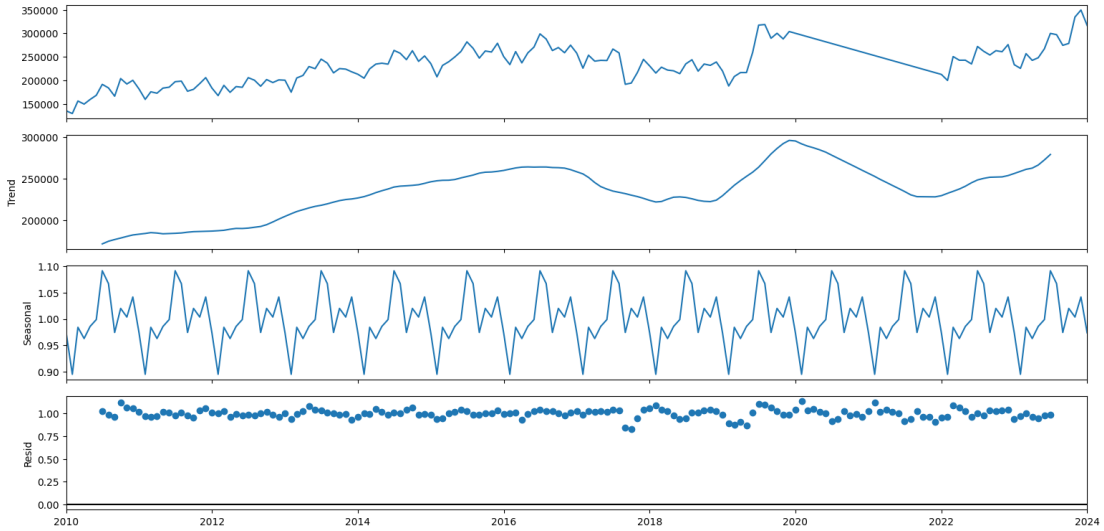


Fig. 26 Tendencia, estacionalidad y residuo de ARIMAX ruta BOGCTG

Fuente: Elaboración propia

La estacionalidad de los datos era acorde a la naturaleza del negocio, donde se esperaba un mayor transporte de pasajeros en temporada de vacaciones, especialmente en junio y diciembre. A continuación, se verificó la estacionalidad de la serie de forma gráfica:

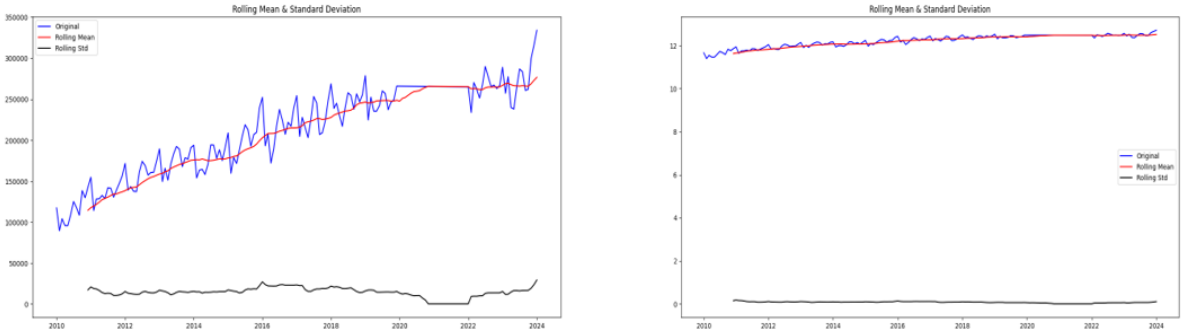


Fig. 27 estacionalidad ruta BOGMDE BOGCTG

Fuente: Elaboración propia

Se observó que la serie no era estacionaria debido a que su varianza no era constante a lo largo del tiempo. Por esta razón, se decidió aplicar una transformación logarítmica. Se observó que la varianza se vuelve constante a lo largo del tiempo y los datos se encuentran listos para generar el modelo.

SARIMAX Results

Dep. Variable:	Pax	No. Observations:	132
Model:	ARIMA(1, 1, 1)	Log Likelihood	128.842
Date:	Mon, 27 May 2024	AIC	-247.685
Time:	21:54:41	BIC	-233.309
Sample:	0	HQIC	-241.843
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
TRM	0.1584	0.221	0.716	0.474	-0.275	0.592
BRENT	-0.0397	0.087	-0.454	0.650	-0.211	0.131
ar.L1	0.2532	0.183	1.382	0.167	-0.106	0.612
ma.L1	-0.7586	0.112	-6.777	0.000	-0.978	-0.539
sigma2	0.0082	0.001	7.700	0.000	0.006	0.010

Fig. 28 Resume modelo (1.1.1.) ARIMAX ruta BOGCTG

Fuente: Elaboración propia

Se observó que para el modelo Arimax(1,1,1) las variables exógenas TRM y BRENT, así como el componente autorregresivo *p* no eran significativo para el modelo. Se procedió a correr un modelo Arimax(0,1,1).

SARIMAX Results

Dep. Variable:	Pax	No. Observations:	132
Model:	ARIMA(0, 1, 1)	Log Likelihood	126.695
Date:	Mon, 27 May 2024	AIC	-245.391
Time:	22:03:45	BIC	-233.890
Sample:	0	HQIC	-240.718
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
TRM	0.1320	0.220	0.600	0.548	-0.299	0.563
BRENT	-0.0662	0.083	-0.796	0.426	-0.229	0.097
ma.L1	-0.6159	0.062	-9.964	0.000	-0.737	-0.495
sigma2	0.0084	0.001	7.543	0.000	0.006	0.011

Fig. 29 Resume modelo (0.1.1.) ARIMAX ruta BOGCTG

Fuente: Elaboración propia

Se observó que el AIC del modelo mejoraba un poco, pero las variables exógenas seguían no siendo significativas para el modelo. Al calcular el R^2 para el conjunto de entrenamiento se obtuvo un valor de 0.83 y un RMS de 31.363.

A continuación, se muestra de forma gráfica la serie original con el pronóstico realizado por el modelo para los datos de entrenamiento y prueba.

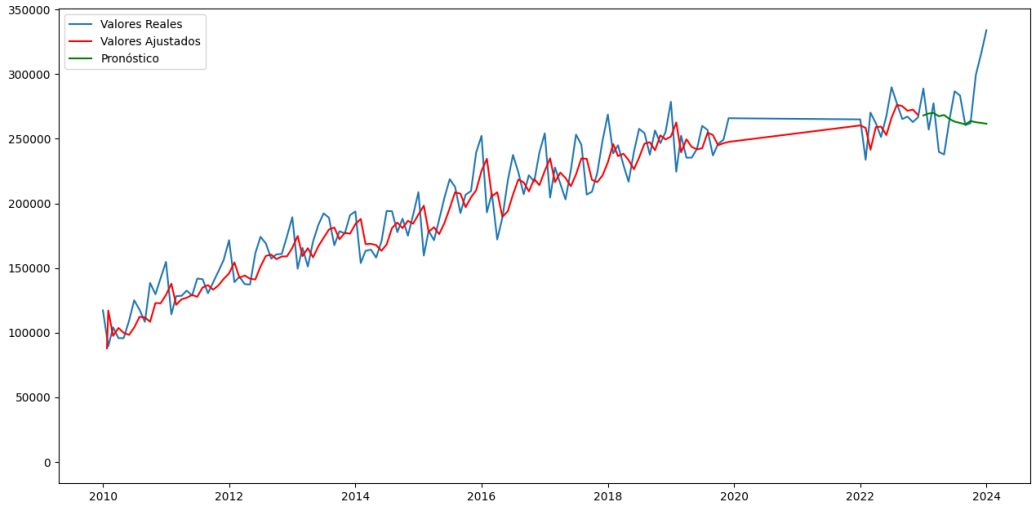


Fig. 30 Pronóstico ruta BOGCTG

Fuente: Elaboración propia

La ruta BOGCTG fue una de las más afectadas por la salida de Viva, de hecho, se puede observar que el pronóstico de los datos de prueba presentó una tendencia decreciente.

MODELO GENERAL

Se identificaron las tendencias y estacionalidades de los datos para la ruta:

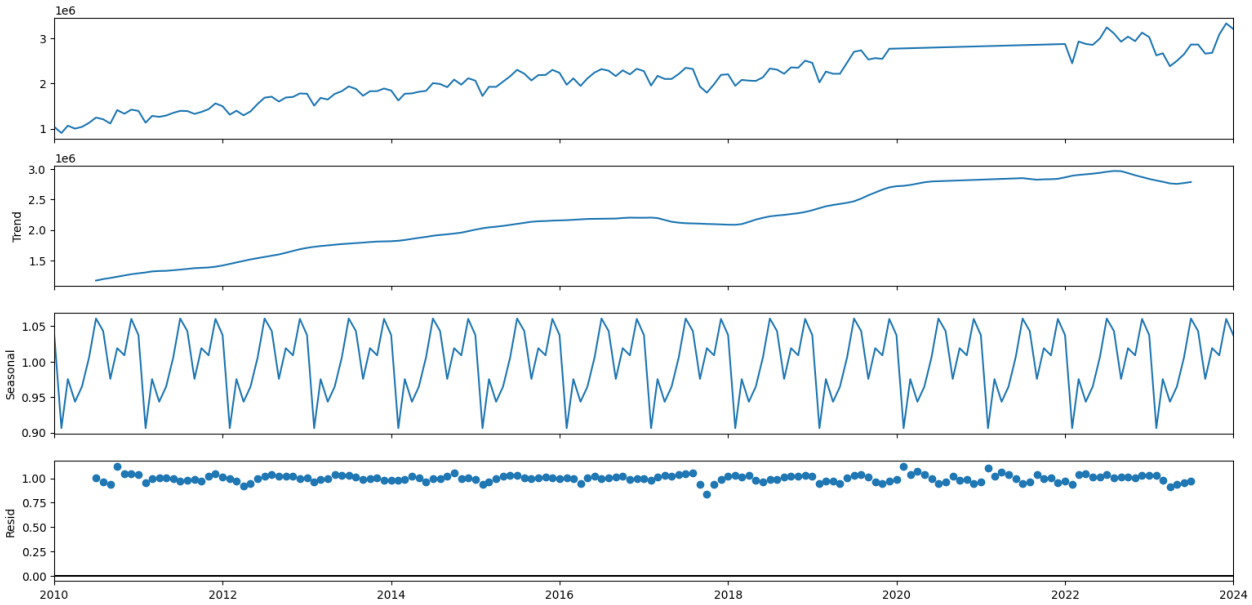


Fig. 31 Tendencia, estacionalidad y residuo de ARIMAX modelo general

Fuente: Elaboración propia

Se evidenció una tendencia positiva exceptuando los años de pandemia. La estacionalidad de los datos fue acorde a la naturaleza del negocio, en el que se esperaba mayor trasporte de pasajeros en temporada de vacaciones, junio y diciembre. Se verificó la estacionalidad de la serie de forma gráfica:

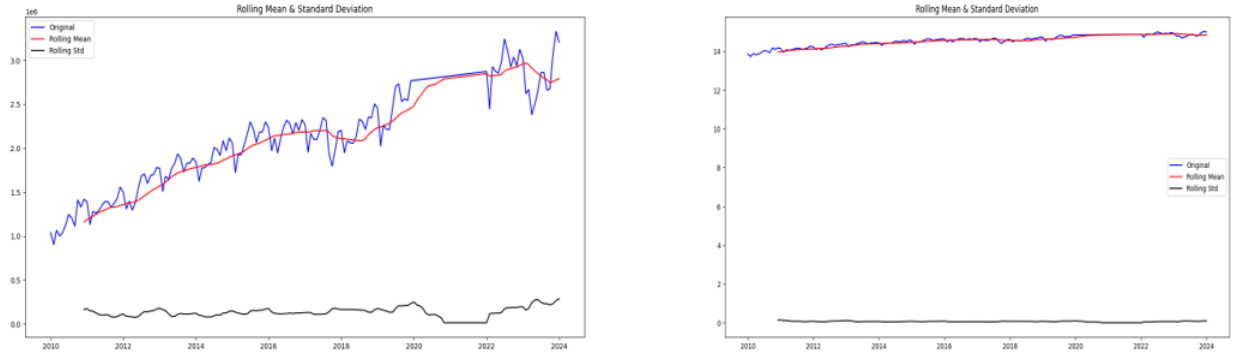


Fig. 32 estacionalidad modelo general

Fuente: Elaboración propia

Se observó que la serie no es estacionaria debido a que su varianza no es constante a lo largo del tiempo, se procedió a aplicar una transformación logarítmica. Una vez se aplicó la

transformación se observó que la varianza se vuelve constante a lo largo del tiempo y los datos estaban listos para generar el modelo.

SARIMAX Results

Dep. Variable:	Pax	No. Observations:	132
Model:	ARIMA(1, 1, 1)	Log Likelihood	157.199
Date:	Mon, 27 May 2024	AIC	-304.397
Time:	22:14:19	BIC	-290.021
Sample:	0	HQIC	-298.556
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
TRM	0.2082	0.203	1.024	0.306	-0.190	0.607
BRENT	-0.0095	0.078	-0.122	0.903	-0.162	0.143
ar.L1	0.3161	0.176	1.791	0.073	-0.030	0.662
ma.L1	-0.6907	0.129	-5.363	0.000	-0.943	-0.438
sigma2	0.0053	0.001	8.516	0.000	0.004	0.007

Fig. 33 Resume modelo (1.1.1.) ARIMAX modelo general

Fuente: Elaboración propia

Se observó que para el modelo Arimax(1,1,1) las variables exógenas BRENT y TRM no eran representativas, así como la parte autorregresiva del modelo. Se procedió a ejecutar un Modelo Arimax(0,1,1).

SARIMAX Results

Dep. Variable:	Pax	No. Observations:	132
Model:	ARIMA(0, 1, 1)	Log Likelihood	155.618
Date:	Mon, 27 May 2024	AIC	-303.236
Time:	22:18:20	BIC	-291.735
Sample:	0	HQIC	-298.563
	- 132		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
TRM	0.1373	0.198	0.694	0.488	-0.251	0.525
BRENT	-0.0469	0.069	-0.678	0.497	-0.182	0.089
ma.L1	-0.4189	0.092	-4.577	0.000	-0.598	-0.240
sigma2	0.0054	0.001	8.566	0.000	0.004	0.007

Fig. 34 Resume modelo (0.1.1.) ARIMAX modelo general

Fuente: Elaboración propia

Se pudo observar que el AIC tuvo una leve mejora al quitar la parte autorregresiva. Se

calculó el R^2 el conjunto de entrenamiento y se obtuvo un valor de 0.89 y un RMS de 357.537. Se procedió a graficar los valores reales y el pronóstico para los datos de entrenamiento y prueba.

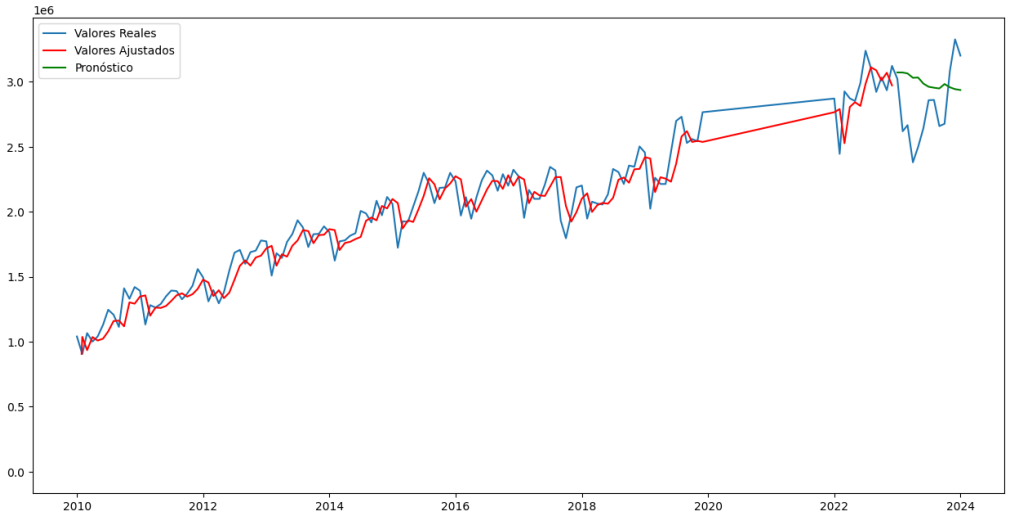


Fig. 35 Pronóstico modelo general

Fuente: Elaboración propia

6.3. RESULTADOS RED NEURONAL RECURRENTE.

La estructura utilizada en la red neuronal recurrente consistió en dividir los datos en conjuntos de entrenamiento y prueba utilizando una proporción del 92% para entrenamiento (enero 2010 – diciembre 2022, sin tener en cuenta 2020 y 2021). Con un modelo que incluye en las dos primeras capas LSTM con 64 y 32 neuronas, respectivamente. Seguidas de dos capas densas con 16 y 1 neuronas. Esta arquitectura se optimizó utilizando el algoritmo Adam con una tasa de aprendizaje de 0.001 y se entrenó durante 100 épocas con un tamaño de lote de 32, obteniendo los siguientes resultados.

GRÁFICAS REAL VS PREDICCIÓN MODELO RED NEURONAL RECURRENTE

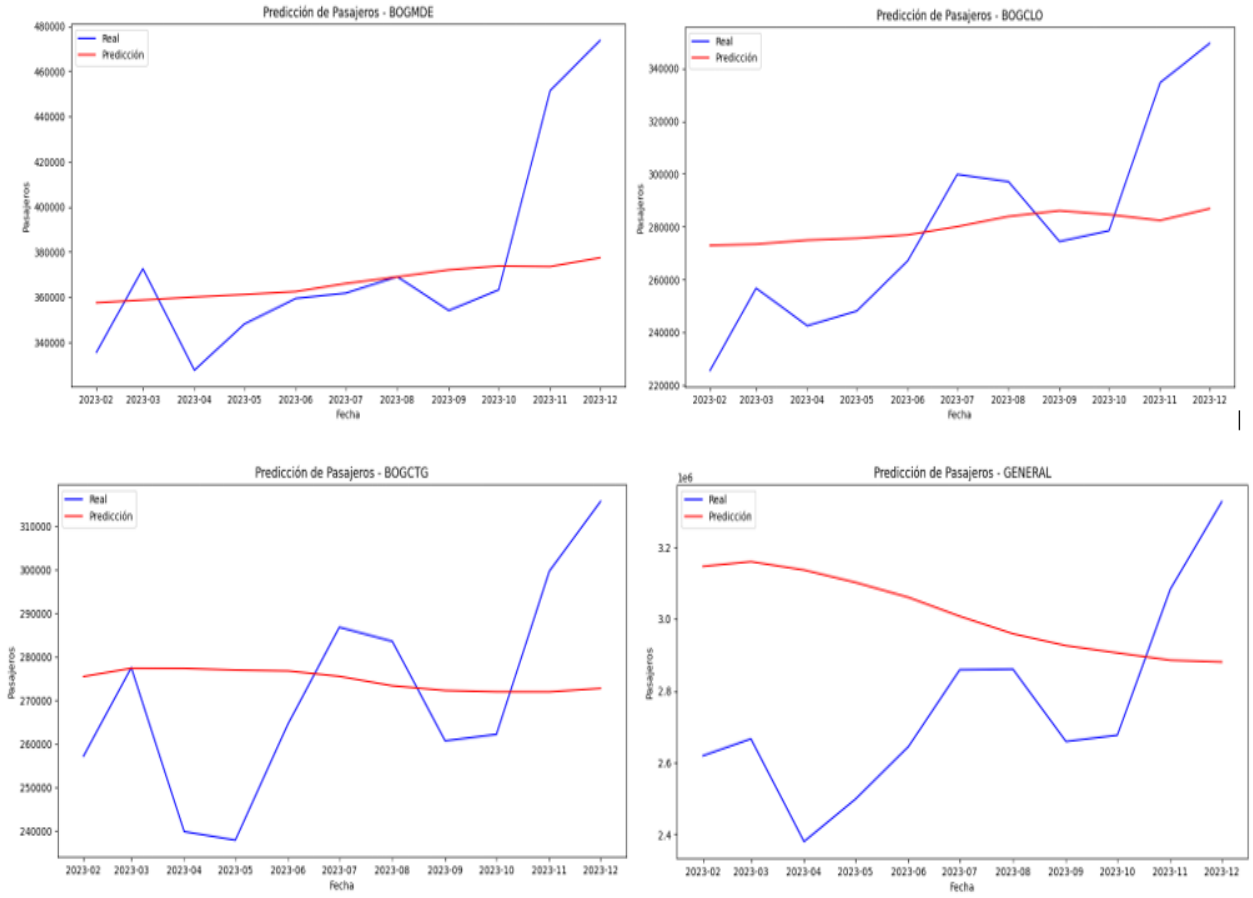


Fig. 36 vs real de todas las rutas modelo Red neuronal recurrente

Fuente: Elaboración propia

MÉTRICAS DE CADA MODELO RED NEURONAL RECURRENTE

Tabla 7 métricas de los modelos de red neuronal recurrentes para todas las rutas.

RUTA	RMSE	R^2
BOGMDE	40155.08	0.16
BOGCLO	32762.56	0.19
BOGCTG	24265.67	-0.12
GENERAL	429107.63	-1.82

Fuente: Elaboración propia

Los resultados obtenidos con la red neuronal recurrente para las distintas rutas, así como para el conjunto general, presentaron un desempeño insatisfactorio. Los R^2 revelan que los modelos no logran explicar de manera adecuada la variabilidad en los datos observados, lo cual sugiere limitaciones significativas en sus capacidades para capturar las complejas relaciones temporales inherentes a los patrones de pasajeros en el mercado aéreo colombiano en sus rutas domésticas principales. Se realizaron ajustes de hiperparámetros en el modelo de red neuronal recurrente. Se exploraron diferentes combinaciones de neuronas en cada capa LSTM y densa, como 128 y 64 neuronas en las capas LSTM respectivamente, seguidas de configuraciones de 32 y 1 neuronas en las capas densas. Además, se probaron diversas tasas de aprendizaje, incluyendo valores como 0.0005 con el algoritmo Adam, y se ajustaron las épocas de entrenamiento variando desde 50 hasta 200 épocas con un tamaño de lote de 16 presentando a veces sobre entrenamiento. A pesar de estos ajustes exhaustivos, los resultados seguían siendo no tan favorables.

6.4. RESULTADOS RED NEURONAL SECUENCIAL

Debido a la cantidad de datos con los que estamos trabajando (144), se decidió probar con una arquitectura de red neuronal secuencial. Inicialmente, se configuró la siguiente estructura: Una capa densa (Fully Connected) con 16 neuronas con activación ReLU y un Dropout de 0.2, seguida de otra capa densa con 8 neuronas y un Dropout de 0.5. Finalmente, una tercera capa densa con 4 neuronas y activación ReLU. La capa de salida consistió en una sola neurona.

Este modelo se entrenó durante 100 épocas con un tamaño de lote de 32 con una división del 80% de los datos en entrenamiento y 20% prueba. Sin embargo, los resultados obtenidos tampoco fueron favorables con respecto a los R^2, lo que llevó a ajustar la red neuronal secuencial a una estructura similar a la recurrente en términos de la cantidad de neuronas por capas, epochs y batch size.

- Capa densa 1 (Fully Connected): 64 neuronas con activación ReLU y un Dropout de 0.2.
- Capa densa 2 (Fully Connected): 32 neuronas con activación ReLU y un Dropout de 0.5.
- Capa densa 3 (Fully Connected): 16 neuronas con activación ReLU.
- Capa de salida (Fully Connected): 1 neurona.
- Épocas: 100 y batch size de 32.

- 80% entrenamiento y 20% prueba.

Esta modificación tuvo como objetivo aumentar la capacidad del modelo para capturar las relaciones complejas en los datos y producir predicciones más precisas y confiables. Este ajuste llevó a una mejora significativa en el rendimiento del modelo y a una mayor exactitud en las predicciones, las cuales se presentan a continuación.

RUTA BOGMDE

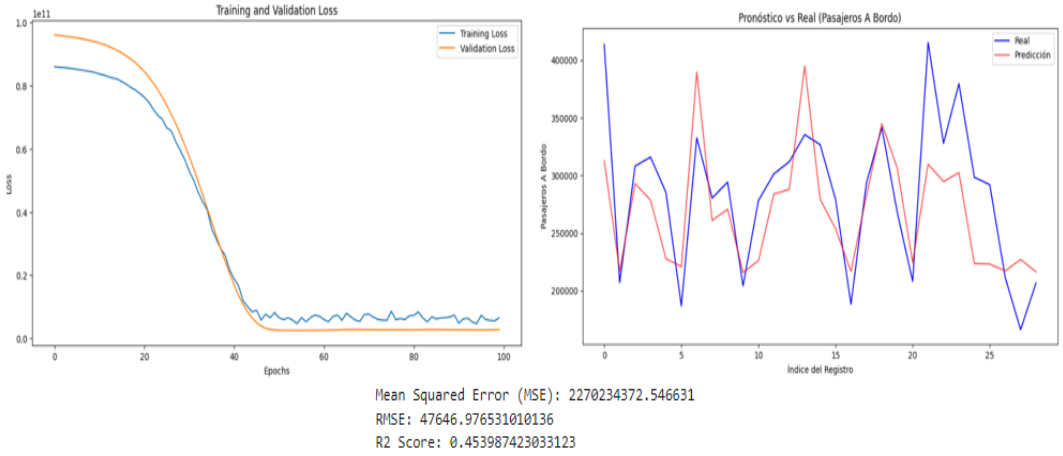


Fig. 37 Pérdida durante el entrenamiento, comparación de pronóstico vs real y métricas de evaluación para la ruta BOGMDE

Fuente: Elaboración propia

RUTA BOGCLO

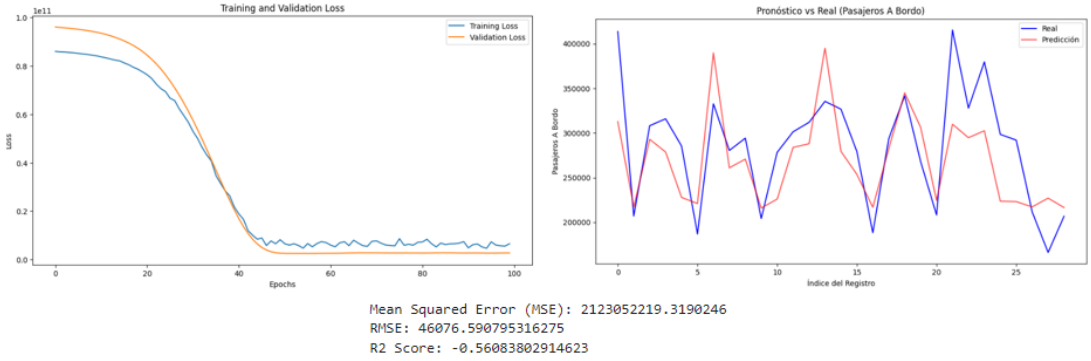


Fig. 38 Pérdida durante el entrenamiento, comparación de pronóstico vs real y métricas de evaluación para la ruta BOGCLO

Fuente: Elaboración propia

RUTA BOGCTG

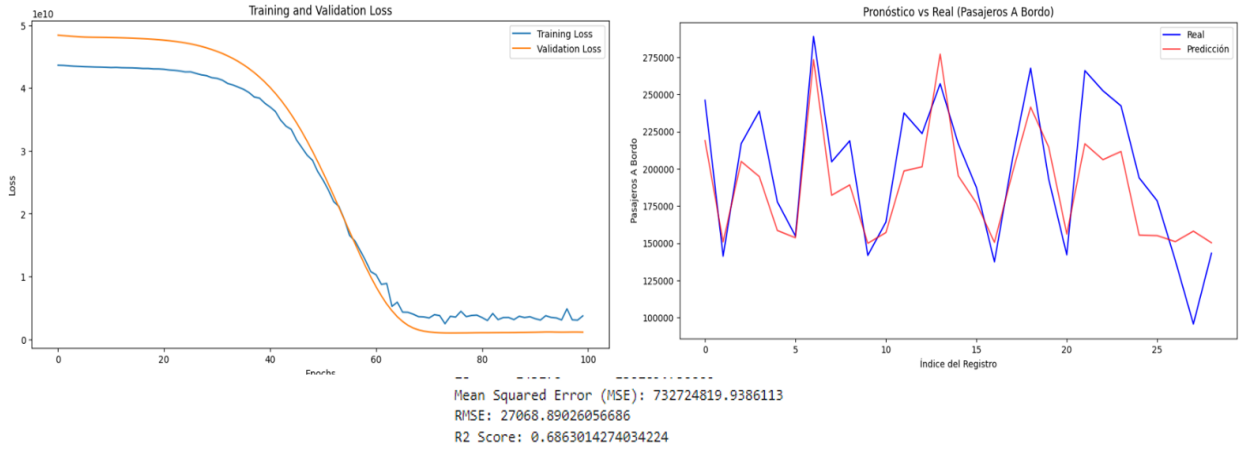


Fig. 39 Pérdida durante el entrenamiento, comparación de pronóstico vs real y métricas de evaluación para la ruta BOGCTG

Fuente: Elaboración propia

RUTAS GENERALES

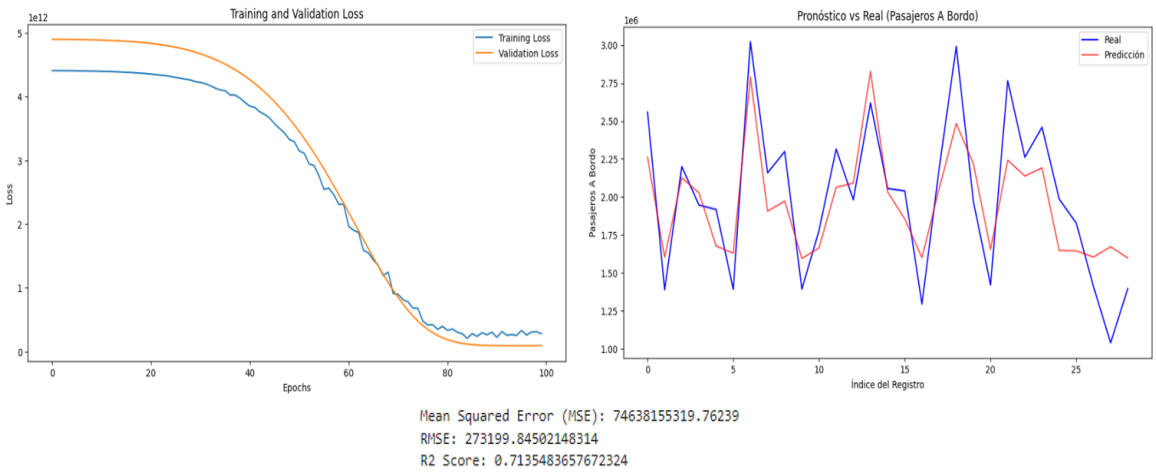


Fig. 40 Pérdida durante el entrenamiento, comparación de pronóstico vs real y métricas de evaluación para la ruta GENERAL

Fuente: Elaboración propia

La arquitectura en cantidad de neuronas por capas iguales a la red neuronal recurrente

resultó ser una decisión acertada, ya que mejoró significativamente el desempeño predictivo para todas las rutas evaluadas. Los resultados muestran puntuaciones de R^2 llamativas para la ruta general (0.71) y para la ruta BOGCTG (0.68). Esto indica que los modelos son capaces de explicar un alto porcentaje de la variabilidad observada en los datos de pasajeros.

A pesar de haber configurado tanto la red neuronal secuencial como la recurrente (LSTM) con la misma cantidad de neuronas por capa, epochs y batch size, la red secuencial produjo resultados superiores. Esto puede atribuirse a la cantidad de datos, la ausencia de los años 2020 y 2021 y/o la posible simplicidad de los patrones temporales en los datos, lo que pudo haber favorecido a la red secuencial. Además, aunque los hiperparámetros eran los mismos, es posible que la red secuencial haya sido más efectiva en aprender y capturar las relaciones entre las variables debido a su estructura más simple y directa. En este contexto específico, la red secuencial demostró ser más adecuada para este conjunto de datos, por lo cual se propone para comparar con los modelos de Arimax y RLM.

7. RESULTADOS OBTENIDOS DE EVALUAR Y COMPARAR DIFERENTES ALGORITMOS DE MACHINE LEARNING PARA DETERMINAR CUÁL OFRECE EL MEJOR RENDIMIENTO EN TÉRMINOS DE PRECISIÓN (OBJETIVO ESPECÍFICO 4)

7.1. COMPARATIVA DE MODELOS

A continuación, se detallan los resultados obtenidos mediante la aplicación de los diversos modelos, considerando que cada uno fue empleado para pronosticar el tráfico de pasajeros en las tres rutas principales del transporte aéreo, así como un pronóstico global o total del número mensual de pasajeros transportados en las rutas de Colombia. Al pronóstico de este último modelo se le aplica una técnica de prorrateo basada en las tendencias históricas para desglosar el valor total en función del comportamiento pasado de cada ruta específica.

7.1.1. RUTA BOGMDE

Tabla 8 comparativa de modelos BOGMDE

RLM				
BOGMDE				
	Coeficiente	P>|t|		Valor
Intercepto	-38.776.311,3	-	R-squared	0,59
YYYY	19.426,4	-	RMSE	48.535
MM	7.572,4	-		
TRM	-44,8	0,004		
BRENT	-372,4	0,038		

ARIMA - X				
BOGMDE (0,1,1)				
	Coeficiente	P>|t|		Valor
TRM	-0,43	0,008	R-squared	0,85
BRENT	-0,18	0,003	RMSE	38.411
MA.L1	-0,30	-		

RED NEURONAL	
BOGMDE	
Capas	4
Neuronas	64-32-16-1
R-squared	0,45
RMSE	47.646

Fuente: Elaboración propia

Para la ruta BOGMDE, el modelo de RLM mostró una moderada capacidad explicativa con un error relativamente alto. El modelo ARIMAX demostró ser superior, reflejando una alta capacidad explicativa y menor error en las predicciones, por otro lado, las Redes Neuronales Recurrentes mostraron un desempeño con menor capacidad para explicar la variabilidad en los datos. En conclusión, el modelo ARIMAX es el más adecuado para la ruta BOGMDE, ya que logra capturar mejor las relaciones subyacentes en los datos con menor error en las predicciones.

7.1.2. RUTA BOGCLO

Tabla 9 comparativa de modelos BOGCLO

RLM				
BOGCLO				
	Coeficiente	P>|t|		Valor
Intercepto	-9.913.310,3	0,058	R-squared	0,36
YYYY	5.045,9	0,054	RMSE	33.660
MM	4.643,9	-		
TRM	-8,8	0,482		
BRENT	-503,8	0,001		

ARIMA - X				
BOGCLO (1,1,1)				
	Coeficiente	P>|t|		Valor
TRM	-0,25	0,14	R-squared	0,65
BRENT	-0,18	0,01	RMSE	33.925
AR.L1	0,53	0,00		
MA.L1	-0,80	-		

RED NEURONAL	
BOGCLO	
Capas	4
Neuronas	64-32-16-1
R-squared	-0,56
RMSE	46.076

Fuente: Elaboración propia

Para la ruta BOGCLO el modelo ARIMAX mostró el mejor desempeño con un coeficiente de determinación (R^2) más alto y un menor error cuadrático medio (RMSE), indicando una mejor capacidad para explicar la variabilidad en los datos y realizar predicciones precisas. En contraste, tanto el modelo de Regresión Lineal Múltiple como las Redes Neuronales Recurrentes obtuvieron resultados inferiores, con menor capacidad explicativa y errores más altos en las predicciones. Por lo tanto, el modelo ARIMAX es el más adecuado para la ruta BOGCLO debido a su mejor rendimiento predictivo y capacidad explicativa.

7.1.3. RUTA BOGCTG

Tabla 10 comparativa de modelos BOGCTG

RLM				
BOGCTG				
	Valor	P>|t|		Valor
Intercepto	- 24.705.638,8	-	R-squared	0,69
YYYY	12.368,0	-	RMSE	36.836
MM	3.380,3	-		
TRM	- 10,6	0,315		
BRENT	- 253,4	0,041		

ARIMA - X				
BOGCTG (0,1,1)				
	Valor	P>|t|		Valor
TRM	0,13	0,548	R-squared	0,83
BRENT	- 0,07	0,426	RMSE	31.363
MA.L1	- 0,62	-		

RED NEURONAL	
BOGCTG	
Capas	4
Neuronas	64-32-16-1
R-squared	0,68
RMSE	27.068

Fuente: Elaboración propia

Para la ruta BOGCTG, los modelos de ARIMAX y RNN superaron al modelo de Regresión

Lineal Múltiple en términos de coeficiente de determinación (R^2) más alto y menor error cuadrático medio (RMSE). Sin embargo, el modelo RNN mostró un rendimiento ligeramente superior al modelo ARIMAX, con un RMSE más bajo y un R^2 cercano, lo que sugiere una mejor capacidad predictiva y explicativa. Por lo tanto, para la ruta BOGCTG, el modelo de Redes Neuronales Recurrentes sería preferible debido a su mejor rendimiento en la predicción de pasajeros a bordo.

7.1.4. RUTAS GENERALES

Tabla 11 comparativa de modelos RUTAS GENERALES

RLM				
Rutas Generales				
	Coeficiente	P>|t|		Valor
Intercepto	- 179.859.860	-	R-squared	0,73
YYYY	89.892	-	RMSE	275.979
MM	36.553	-		
TRM	135	0,198		
BRENT	635	0,602		

ARIMA - X				
Modelo General (0,1,1)				
	Coeficiente	P>|t|		Valor
TRM	0,14	0,488	R-squared	0,89
BRENT	- 0,05	0,497	RMSE	357.537
MA.L1	- 0,42	-		

RED NEURONAL	
Modelo General	
Capas	4
Neuronas	64-32-16-1
R-squared	0,71
RMSE	273.199

Fuente: Elaboración propia

Para la ruta general, aunque el modelo ARIMAX mostró el coeficiente de determinación (R^2) más alto, el modelo RNN demostró un rendimiento comparable con un R^2 ligeramente más bajo pero un error cuadrático medio (RMSE) más bajo. Esto sugiere que el modelo RNN tiene una capacidad predictiva efectiva y una mejor capacidad para explicar la variabilidad en los datos, lo que lo convierte en una opción preferible para la predicción de pasajeros a bordo en la ruta general.

7.2. PRONÓSTICOS

Una vez desarrollados los tres modelos se realizaron pronósticos para cada mes en curso del 2024, estos pronósticos son adicionales a los realizados con los datos de prueba. Para los valores de la TRM y BRENT se asumía que todo el mes tendría el mismo valor del promedio de la primera semana debido a que se está pronosticando el mes en curso y no se conocen los valores completos de estas variables. A continuación, se presentan los pronósticos que se fueron realizando mes a mes junto con los valores reales que se fueron dando en el mercado con las respectivas diferencias que iba presentando cada modelo.

Tabla 12 Pronósticos

MES	RUTA	REAL	RNR	RLM	ARIMAX	RNR_Diferencia	RLM_Diferencia	ARIMAX_Diferencia
1	General	3.201.865	2.420.254	2.692.774	2.987.116	-24%	-16%	-7%
	BOGMDE	429.327	350.154	346.123	409.992	-18%	-19%	-5%
	BOGCLO	316.868	266.407	229.873	302.562	-16%	-27%	-5%
	BOGCGT	334.126	241.999	269.273	262.690	-28%	-19%	-21%
2	General	2.811.702	2.420.246	2.735.838	2.986.791	-14%	-3%	6%
	BOGMDE	391.831	350.109	351.139	406.094	-11%	-10%	4%
	BOGCLO	278.296	266.401	232.977	299.372	-4%	-16%	8%
	BOGCGT	281.208	241.958	271.652	262.497	-14%	-3%	-7%
3	General	3.002.480	2.424.579	2.769.798	2.978.552	-19%	-8%	-1%
	BOGMDE	424.219	350.691	359.117	405.006	-17%	-15%	-5%
	BOGCLO	302.394	266.875	236.476	298.210	-12%	-22%	-1%
	BOGCGT	309.320	242.343	274.656	261.612	-22%	-11%	-15%

Fuente: Elaboración propia

Se pudo evidenciar que el modelo ARIMAX presentaba el mejor desempeño al pronosticar el número de pasajeros en las diferentes rutas al compararlo con los valores reales del mercado que se iban dando mes a mes. Gracias a su componente de media móvil y autoregresión el modelo ARIMAX pudo captar mejor las tendencias del mercado. Debido a su superior desempeño, se decidió utilizar los pronósticos de este modelo para la propuesta de eficiencia de capacidad del mercado. En contraste, los modelos de Red Neural Recurrente (RNR) y Regresión Lineal Múltiple (RLM) mostraron grandes desviaciones con respecto a los valores reales.

7.3. PROPUESTA DE EFICIENCIA DE CAPACIDAD

Para la propuesta de eficiencia de capacidad, primero se calculó el factor de ocupación (LF[9]) promedio del año 2023 para cada ruta, tal como se muestra en la siguiente tabla:

Tabla 13 LF HISTÓRICO 2023

RUTA	LF 2023
BOGCLO	80%
BOGMDE	82%
BOGCTG	82%
MODELO GENERAL	81%

Fuente: Elaboración propia

Este LF se utilizó para calcular la cantidad de sillas necesarias en cada mes/ruta para los

[9] Factor de ocupación (LF) = $\frac{Pax}{Sillas\ disponibles}$

pasajeros pronosticados. Recordemos que la fórmula es $Sillas = \frac{Pax\ Pronosticados}{LF}$

Tabla 14 CÁLCULO DE SILLAS PROPUESTAS

		PROPUESTO			REAL		
MES	RUTA	PAX FCST	SILLAS	LF	PAX	SILLAS	LF
1	General	2.987.116	3.667.902	81%	3.201.865	3.902.737	82%
	BOGMDE	409.992	502.061	82%	429.327	539.826	80%
	BOGCLO	302.562	378.040	80%	316.868	391056	81%
	BOGCTG	262.690	320.578	82%	334.126	375.554	89%
2	General	2.986.791	3.667.502	81%	2.811.702	3.721.718	76%
	BOGMDE	406.094	497.288	82%	391.831	530.643	74%
	BOGCLO	299.372	374.054	80%	278.296	368.974	75%
	BOGCTG	262.497	320.343	82%	281.208	360.431	78%
3	General	2.978.552	3.657.386	81%	3.002.480	4.088.217	73%
	BOGMDE	405.006	495.956	82%	424.219	589.555	72%
	BOGCLO	298.210	372.602	80%	302.394	408.619	74%
	BOGCTG	261.612	319.263	82%	309.320	404.615	76%

Fuente: Elaboración propia

Una vez realizado el cálculo de las sillas propuestas para transportar a los pasajeros pronosticados utilizando el factor de ocupación (LF) histórico del 2023, se procedió a calcular el LF propuesto, considerando los pasajeros reales transportados en cada ruta y mes junto con las sillas propuestas.

Este LF propuesto se comparó con el LF real para verificar si se producía un incremento o no.

Tabla 15 PROPUESTA DE EFICIENCIA DE CAPACIDAD

MES	RUTA	PAX REALES	SILLAS PROPUESTAS	LF	LF vReal
1	General	3.201.865	3.667.902	87%	5%
	BOGMDE	429.327	502.061	86%	6%
	BOGCLO	316.868	378.040	84%	3%
	BOGCTG	334.126	320.578	104%	15%
2	General	2.811.702	3.667.502	77%	1%
	BOGMDE	391.831	497.288	79%	5%
	BOGCLO	278.296	374.054	74%	-1%
	BOGCTG	281.208	320.343	88%	10%
3	General	3.002.480	3.657.386	82%	9%
	BOGMDE	424.219	495.956	86%	14%
	BOGCLO	302.394	372.602	81%	7%
	BOGCTG	309.320	319.263	97%	20%

Fuente: Elaboración propia

Se logra evidenciar que en 11 ocasiones el factor de ocupación (LF) propuesto aumenta en

comparación con el LF real. Solamente en el mes 2, en la ruta Bogotá-Cali, el LF disminuye en un punto porcentual. En la ruta Bogotá-Cartagena en el mes 1, se observa un LF del 104%, lo que indica que la cantidad de sillas propuestas no fue suficiente para transportar a todos los pasajeros de la ruta. Esta ruta se presenta como crítica y podría requerir análisis adicionales para aumentar la cantidad de sillas propuestas. La sobreventa podría impactar las sillas propuestas, por lo que sería necesario considerar este porcentaje de sobreventa en futuras propuestas de sillas.

7.4. TABLERO POWER BI

Con el objetivo de visualizar en detalle los resultados obtenidos por nuestros modelos, se optó por desarrollar un tablero en Power BI. Este tablero muestra el historial de pasajeros transportados y sillas ofrecidas desde el año 2010 hasta la fecha pronosticada, lo que permite corroborar la tendencia de nuestro pronóstico. Además, incluye un mapa que representa las ciudades de origen o destino y permite filtrar por fecha, tipo de vuelo y ruta.

Link del tablero:

https://app.powerbi.com/view?r=eyJrIjoiNDFhYTBhZjEtODgwOS00ZGNlLTkxNmQtNzExZDgxMTY3YWFkIiwidCI6IjE2YWY2YjQ1LTAwYzUtNGJhMy05ZDRjLThiZmExNmU0MzYwMyIsImMiOjR9

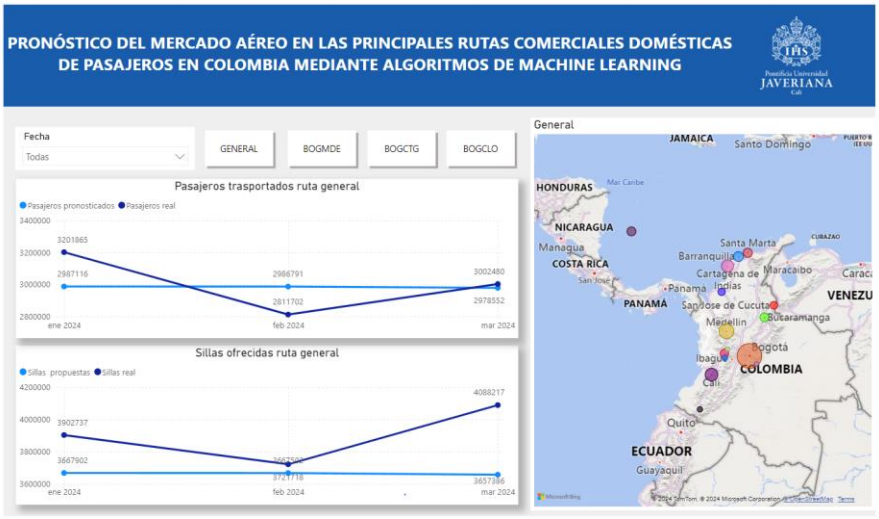


Fig. 41 Tablero Power BI

Fuente: Elaboración propia

8. CONCLUSIONES Y TRABAJOS FUTUROS

8.1. CONCLUSIONES

Nuestro proyecto de grado se centró en pronosticar el número de pasajeros transportados en tres rutas específicas: BOGCLO, BOGMDE y BOGCTG, además de desarrollar un modelo general para todas las rutas. Para cada ruta, se implementaron cuatro modelos: una red neuronal recurrente, un modelo ARIMA-X y un modelo de regresión lineal múltiple (RLM). Comparando los resultados de estos modelos, se identificó que el modelo ARIMA-X era el más preciso y, por ende, se utilizó para desarrollar la propuesta de eficiencia de capacidad.

La propuesta de eficiencia de capacidad consistió en plantear un factor de ocupación (LF) de las rutas durante los meses analizados mayor a los presentados en realidad. Para lograr esto, se calcularon las sillas necesarias para cada ruta, basándose en los pasajeros pronosticados con el LF histórico de 2023. Al comparar estos valores con los datos reales del mercado, se evidenció un incremento en el LF propuesto en 11 de los casos. Sin embargo, se observaron dos excepciones: en la ruta BOGCLO durante el mes 2, donde el LF disminuyó en un punto porcentual, y en la ruta Bogotá-Cartagena durante el mes 1, donde el LF alcanzó el 104%, indicando una sobreventa.

El trabajo realizado demuestra que los modelos de pronóstico, especialmente el modelo ARIMA-X, son herramientas valiosas para obtener una visión general del mercado y detectar tendencias en las diferentes rutas. Estos modelos permiten identificar rutas con potencial de crecimiento y ajustar la capacidad de forma más eficiente. Además, resaltan la importancia de realizar ajustes comerciales, como campañas de marketing, para mejorar el LF en rutas con baja ocupación, y la necesidad de aumentar la cantidad de slots en rutas con alta demanda.

En resumen, este estudio no solo válida la utilidad de los modelos de pronóstico en la gestión eficiente de la capacidad de transporte, sino que también proporciona una base sólida para futuras decisiones estratégicas en la optimización del servicio y la maximización del LF en las rutas analizadas.

8.2. TRABAJOS FUTUROS

Para futuros trabajos, se podría explorar el mejoramiento y la optimización de los modelos de pronóstico mediante el desarrollo de técnicas más avanzadas, como redes neuronales profundas o modelos híbridos. Además, sería beneficioso ampliar el alcance geográfico del análisis a otras rutas nacionales e internacionales, así como incorporar variables exógenas adicionales, como factores económicos y meteorológicos, para mejorar la precisión de los pronósticos. También sería relevante investigar la elasticidad de la demanda y desarrollar simulaciones de escenarios futuros para evaluar el impacto de diferentes estrategias de capacidad y marketing. Otra área de interés sería analizar la saturación de las rutas existentes y estudiar la viabilidad de abrir nuevas rutas basadas en la demanda proyectada. Asimismo, se podría trabajar en la integración de los modelos de pronóstico con los sistemas de gestión de aerolíneas y en la evaluación del impacto ambiental de las decisiones de capacidad. Por último, sería importante realizar estudios sobre el comportamiento del pasajero y colaborar con entidades reguladoras para desarrollar políticas que mejoren la eficiencia y sostenibilidad del transporte aéreo. Estas propuestas podrían contribuir significativamente al avance en la investigación y aplicación práctica de modelos de pronóstico en la industria aérea.

9. REFERENCIAS BIBLIOGRÁFICAS

[1] Desde la cabina de vuelo, "¿Qué es un hub?", 2019. https://desdelacabinadevuelo.com/2014/06/11/que-es-un-hub/ (consultado el 13 de abril de 2024).

[2] E. A. Millán Gordo, "Pronóstico de la demanda de pasajeros aéreos en nueve Aeropuertos regionales colombianos", Trabajo de grado para optar por el título de Ingeniero Civil, Universidad Santo Tomás, Facultad de Ingeniería Civil, 2019.

[3] Ministerio de Transporte, "¿Qué es el Ministerio de Transporte?", 2011. https://mintransporte.gov.co/publicaciones/259/que-es-el-ministerio-de-transporte/#:~:text=Es%20una%20organizaci%C3%B3n%20administrativa%20responsable,aeropuertos%20y%20las%20v%C3%ADas%20f%C3%A9rreas (consultado el 13 de abril de 2024).

[4] Aeronáutica Civil de Colombia, "¿Qué es la Aerocivil?". https://www.alcaldiabogota.gov.co/sisjur/listados/tematica2.jsp?subtema=33631&cadena=#:~:text=La%20Aeron%C3%A1utica%20Civil%20%2D%20Aerocivil%20tiene,social%20y%20de%20relaciones%20internacionales (consultado el 13 de abril de 2024).

[5] Agencia Nacional de Infraestructuras, "¿Qué es la ANI? [PDF]". https://www.ani.gov.co/sites/default/files/u785/1_que_es_la_ani_.pdf (consultado el 13 de abril de 2024).

[6] P.Shinde, y S.Shah, "A Review of Machine Learning and Deep Learning Applications", en 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) (pp. 1-6). Pune, India. DOI: 10.1109/ICCUBEA.2018.8697857.

[7] C. M. Bishop, "Pattern Recognition and Machine Learning", Springer, 2006.

[8] T. Hastie, R. Tibshirani, y J. Friedman, "The Elements of Statistical Learning: Data Mining,

Inference, and Prediction", Springer, 2009.

[9] K. P. Murphy, "Machine Learning: A Probabilistic Perspective", MIT Press, 2012.

[10] M.Somvanshi, P.Chavan, S.Tambade, y S.V.Shinde, "A review of machine learning techniques using decision tree and support vector machine", en 2016 International Conference on Computing Communication Control and Automation (ICCUBEA) (pp. 1-7). Pune, India. DOI: 10.1109/ICCUBEA.2016.7860040.

[11] C. Chatfield, "The analysis of time series: an introduction", 6th ed., CRC Press, 2016.
[12] W. Mendenhall, R. L. Scheaffer, D. D. Wackerly, A. De la Fuente Pantoja, y D. V. Verbeeck, "Estadística matemática con aplicaciones," California: Grupo Editorial Iberoamericana, 1986.

[13] J.M. Abuín, "Regresión lineal multiple", laboratorio de estadistica, 2007.

[14] IBM, "Redes neuronales", https://www.ibm.com/docs/es/spss-modeler/saas?topic=networks-neural-model (consultado el 13 de abril de 2024).

[15] Do, QH, Lo, S., Chen, J., Le, C. y Anh, "Forecasting Air Passenger Demand: A Comparison of LSTM and SARIMA", Journal of computer science, 2020.

[16] M. Firat, D.Yiltas-Kaplan, R.Samli , " Forecasting Air Travel Demand for Selected Destinations Using Machine Learning Methods," Journalof universal computer science, 2021.

[17] A. A. Martínez Toro, "Estimación de la demanda del tráfico de pasajeros en vuelos domésticos de los 5 aeropuertos más importantes de Colombia", Proyecto de Grado, Universidad de la Sabana, Escuela de Ciencias Económicas y Administrativas, Economía y Finanzas Internacionales, Chía, Cundinamarca, octubre 2013.

[18] E. A. Millán Gordo, "Pronóstico de la demanda de pasajeros aéreos en nueve Aeropuertos regionales colombianos", Trabajo de grado para optar por el título de Ingeniero Civil, Universidad Santo Tomás, Facultad de Ingeniería Civil, Bogotá, Colombia, 2019.

[19] O.D. Olariaga, "Análisis del desarrollo reciente del transporte aéreo en Colombia", Revista Transporte y Territorio, (14), 122-143.

[20] dólar hoy TRM Colombia. https://dolar.wilkinsonpc.com.co/2010 (consultado el 13 de abril de 2024).

[21] Datos históricos Futuros petróleo Brent. https://es.investing.com/commodities/brent-oil-historical-data (consultado el 13 de abril de 2024).