Estimación de la demanda en aerolíneas con operación en Colombia mediante análisis predictivo basado en modelos de series de tiempo

Alba Gómez Arango, Francisco Javier Moya, Juan Felipe Burbano Zapata

***Resumen*—Este estudio examina el comportamiento del tráfico aéreo de las aerolíneas que operan en Colombia. El objetivo principal es desarrollar un modelo predictivo para estimar la cantidad mensual de pasajeros que demandarán el servicio para cada aerolínea. Para lograrlo, se recopilaron datos de vuelos históricos de la página web de la Aeronáutica Civil de Colombia y se entrenaron modelos de predicción supervisada. Los resultados obtenidos proporcionarán información valiosa tanto para las aerolíneas como para los pasajeros al momento de planificar sus viajes.**

I. Introducción

El éxito de las aerolíneas depende en gran medida de su capacidad para transportar pasajeros de manera efectiva y satisfacer sus necesidades de sus uruarios. Este estudio, se enfoca en examinar el comportamiento del tráfico aéreo de las aerolíneas que operan en Colombia, y se destaca la importancia de generar conocimiento a partir de los datos disponibles.

La estimación de la demanda mensual de pasajeros en una aerolínea es crucial para la planificación operativa, la gestión de ingresos, la eficiencia financiera y la mejora de la experiencia del cliente. Para lograrlo, es necesario analizar y comprender los datos históricos de vuelos y emplear técnicas avanzadas de análisis. Al utilizar métodos como el análisis de series de tiempo, se puede predecir con mayor precisión la cantidad de pasajeros que utilizarán el servicio, lo que permite a la aerolínea tomar decisiones informadas sobre la asignación de recursos, la fijación de precios y la optimización de la capacidad. Esto conduce a una operación más eficiente, una mayor rentabilidad y una mejor satisfacción de los pasajeros [2] .

El objetivo principal de este estudio es desarrollar un modelo predictivo para estimar la cantidad mensual de pasajeros que demandarán el servicio en cada aerolínea. Para lograrlo, se recopilaron datos de vuelos históricos proporcionados por la Aeronáutica Civil de Colombia y se emplearon técnicas de análisis de datos avanzadas. Estos modelos permitirán analizar y predecir el tráfico aéreo de las aerolíneas estudiadas, generando conocimiento valioso a partir de la información disponible.

Los resultados obtenidos serán de gran utilidad tanto para las aerolíneas, al ayudarles en la planificación de su capacidad, como para los pasajeros, al permitirles tomar decisiones informadas sobre sus viajes. Además, este estudio resalta la importancia de utilizar los datos disponibles de manera inteligente y generar conocimiento a partir de ellos, lo que puede contribuir al crecimiento y la eficiencia del sector aéreo en Colombia.

II. Pregunta de investigación

¿Cómo podemos utilizar el análisis de series de tiempo para predecir la cantidad de pasajeros mensuales en aerolíneas con operación en Colombia y considerando datos disponibles históricos?

III. Objetivos

- **Objetivo general:** Desarrollar un modelo predictivo que analice y prediga el tráfico aéreo mensual en aerolíneas con operación en Colombia.

- **Objetivos específicos:**
  1. Construir un conjunto de datos estandarizado que recoja el histórico del tráfico aéreo en Colombia.
  2. Analizar el comportamiento e interacción entre las variables presentes en los conjuntos de datos.
  3. Entrenar modelos de predicción supervisados utilizando las variables relevantes identificadas en el análisis exploratorio de datos y aplicando técnicas de validación cruzada para evaluar su capacidad de predicción.
  4. Seleccionar el mejor modelo de predicción basándose en la(s) métrica(s) de evaluación elegidas previamente.

IV. Marco conceptual

IV-A. Industria aérea

La industria aérea es uno de los sectores más importantes del transporte en todo el mundo. Desde sus comienzos, ha revolucionado la manera en que las personas viajan y se conectan entre sí. La industria ha experimentado una gran evolución, desde los primeros vuelos comerciales hasta la actualidad, con aerolíneas que conectan ciudades de todo el mundo. Sin embargo, la industria aérea también enfrenta grandes desafíos, como la necesidad de reducir su impacto ambiental y hacer frente a las fluctuaciones del mercado. A pesar de estos desafíos, la industria aérea sigue siendo un elemento clave en el transporte y en la economía mundial, conectando personas y lugares como nunca antes.

En particular, en Colombia, la industria aérea ha experimentado una evolución significativa a lo largo del tiempo y

desempeña un papel fundamental en la conexión territorial, así como en el desarrollo económico y turístico del país. La compleja geografía del país, atravesado por las cordilleras de los Andes, las deficiencias en la infraestructura de conexión vial terrestre y la falta de desarrollo del transporte ferroviario de pasajeros, hacen que la industria aérea en el país tenga una alta demanda, especialmente para distancias medias y largas [3].

En Colombia, la industria aérea ha experimentado un gran crecimiento en los últimos años; especificamente para el periodo comprendido entre 1992 y 2016 se estima para todos los aeropuertos del país un crecimiento de más del 800 % y en la última década este crecimiento se ha sostenido, en gran medida, gracias a la inversión pública y privada en la industria, la mejora en la infraestructura, la modernización tecnológica de la red de aeropuertos, la entrada de nuevos operadores, especialmente las aerolíneas de bajo costo, y a la incorporación de nuevos destinos y un mayor número de frecuencias por parte de las líneas aéreas [3]. Con la creciente demanda de viajes aéreos y el aumento del turismo en Colombia, las aerolíneas han tenido que adaptarse a las necesidades del mercado para ofrecer mejores servicios y precios competitivos.

Avianca es la aerolínea más grande de Colombia y ha sido un actor clave en la industria aérea del país durante muchos años. Sin embargo, en los últimos años, han surgido una nueva generación de aerolíneas de bajo costo, como Viva Air, Wingo y EasyFly, que han revolucionado el mercado con tarifas más accesibles y un enfoque en la eficiencia y la simplicidad [4].

A pesar de los retos que ha enfrentado la industria aérea colombiana, como la falta de infraestructura en algunos aeropuertos, el sector sigue creciendo y evolucionando. Se espera que la industria continúe expandiéndose en el futuro, a medida que más colombianos tengan acceso a viajar en avión y se abran nuevas rutas y destinos en el país.

El crecimiento de la industria aérea en Colombia conlleva el desafío de aprovechar y generar conocimiento a partir de la gran cantidad de datos generados en las operaciones. En este contexto, la analítica y la minería de datos juegan un papel fundamental al permitir a las aerolíneas recopilar, analizar y comprender estos volúmenes masivos de información. Al combinar diversas técnicas de análisis avanzadas, como el aprendizaje automático y la inteligencia artificial, las aerolíneas pueden extraer información valiosa y descubrir patrones ocultos en los datos. Esto, a su vez, les brinda la capacidad de tomar decisiones más informadas y estratégicas para impulsar su rendimiento. En resumen, la analítica y la minería de datos son herramientas indispensables que permiten a las aerolíneas colombianas aprovechar al máximo su información para tomar decisiones más efectivas y obtener una ventaja competitiva.

IV-B. Análisis de grandes volúmenes de información

En un mundo cada vez más globalizado, el volumen de datos producidos crece de manera desorbitante día a día. Si bien estos datos son útiles para cada una de las industrias que los generan debido a la necesidad misma que los crea, por sí solos no producen información. Para ello, se requiere de procesos de análisis que permitan extraer valor y generar información.

Las compañías aéreas generan y procesan grandes cantidades de información relacionada con la programación de vuelos, el rendimiento de los aviones, la gestión de reservas y la experiencia del cliente, entre otros aspectos. Este aumento en los volúmenes de datos ha llevado a las compañías aéreas a adoptar soluciones a gran escala y análisis de datos tales como predecir la demanda de vuelos, lo que les permite optimizar el uso de sus recursos y maximizar sus ingresos, lo que redunda en mayor eficiencia y rentabilidad.

Existen diversas herramientas para el análisis de grandes volúmenes de información, entre ellas se encuentran:

- **Apache Hadoop**
  Plataforma de software de código abierto que permite el almacenamiento y procesamiento distribuido de grandes conjuntos de datos en clusters de hardware. Fue creado para manejar conjuntos de datos extremadamente grandes (terabytes o petabytes) y se utiliza ampliamente en la industria para el análisis de datos y la minería de datos. Hadoop está basado en el concepto de MapReduce, que divide un conjunto de datos en subconjuntos más pequeños y los procesa de forma paralela en diferentes nodos de un cluster. También utiliza el sistema de archivos distribuido Hadoop Distributed File System (HDFS) para almacenar los datos de forma distribuida en los nodos del cluster [5].
- **Apache Spark**
  Herramienta de procesamiento de datos de código abierto y distribuida que se utiliza para el análisis de grandes conjuntos de datos. Spark es más rápido que otras herramientas de procesamiento de datos distribuidos porque utiliza la memoria RAM para el almacenamiento en caché de datos y la planificación de tareas. Spark se basa en el concepto de Resilient Distributed Datasets (RDD), que permite el procesamiento distribuido y paralelo de datos. Además de RDDs, Spark también proporciona una API de programación para el procesamiento de datos llamada DataFrames, que es una estructura de datos tabulares similar a un DataFrame en Python o R. Spark se utiliza en una amplia variedad de aplicaciones, desde el análisis de datos en tiempo real hasta el procesamiento de datos de grandes volúmenes de datos[6].

IV-B1. Tecnologías para el procesamiento de grandes volumenes de datos: En este proyecto, se utilizaron diferentes tecnologías para el procesamiento de grandes volúmenes de datos y la construcción de modelos de predicción. A continuación, se describen las principales tecnologías empleadas y su relevancia en el proyecto.

- **Python**
  Lenguaje de programación ampliamente utilizado en el ámbito del análisis de datos y la construcción de modelos de machine learning. Proporciona una amplia gama de librerías y herramientas especializadas que facilitan la manipulación, transformación y análisis de datos. En este proyecto, Python fue la base principal para el desarrollo del código y la implementación de los modelos de predicción [7].

- **Polars**
  Biblioteca de código abierto para la manipulación y análisis de datos que está diseñada para ofrecer una alternativa rápida y eficiente a pandas. Está construida sobre Apache Arrow y Rust, lo que le permite aprovechar el procesamiento en múltiples hilos y las operaciones SIMD (Instrucción Única, Datos Múltiples) para acelerar el procesamiento [8].
  Polars proporciona una API de DataFrame similar a pandas, lo que facilita la transición de pandas a Polars. Admite varios tipos de datos, incluyendo numéricos, booleanos, fechas, horas y cadenas, y ofrece una amplia gama de operaciones para la transformación de datos, filtrado, agregación, fusión y más.
  Una de las características destacadas de Polars es su capacidad para manejar conjuntos de datos grandes con facilidad, gracias a su gestión optimizada de la memoria y capacidades de procesamiento paralelo. También admite cómputo distribuido mediante la integración con Apache Arrow Flight.
- **Pyspark**
  Biblioteca de Python que permite interactuar con Apache Spark, un potente motor de procesamiento distribuido diseñado para manejar grandes volúmenes de datos en paralelo. Pyspark proporciona una interfaz de programación en Python para trabajar con Spark, lo que facilita el desarrollo y la implementación de aplicaciones de big data [9].
  Spark se basa en el concepto de Resilient Distributed Datasets (RDD), que son colecciones distribuidas de objetos inmutables. Pyspark permite crear RDDs y realizar operaciones de transformación y acción en ellos. Las transformaciones son operaciones que se aplican a los RDDs para modificar su contenido o estructura, mientras que las acciones son operaciones que generan un resultado o devuelven valores.
  Además de las RDDs, Pyspark también ofrece DataFrames, que son estructuras de datos tabulares similares a las de pandas. Los DataFrames en Pyspark ofrecen una interfaz más familiar y amigable para trabajar con datos estructurados, lo que los hace ideales para tareas de análisis y manipulación de datos.
  Una de las principales ventajas de Pyspark es su capacidad para distribuir y procesar datos en clústeres de computadoras, lo que permite realizar análisis y operaciones en paralelo para mejorar el rendimiento. Spark también ofrece módulos y bibliotecas adicionales para tareas específicas, como procesamiento de datos en tiempo real (Spark Streaming), aprendizaje automático (Spark MLlib) y procesamiento de grafos (GraphX).

IV-C. Técnicas de analítica de datos para series de tiempo

La predicción de pasajeros en vuelos es de gran importancia para cualquier aerolínea, ya que permite planificar de manera efectiva la capacidad de sus aviones y optimizar la utilización de recursos como tripulaciones y combustible. Además, la predicción precisa de la demanda de pasajeros también permite a las aerolíneas ajustar sus precios y ofertas promocionales para maximizar sus ingresos y mejorar su rentabilidad. Al utilizar modelos avanzados de análisis de datos y algoritmos de aprendizaje automático, las aerolíneas pueden obtener pronósticos más precisos sobre la cantidad de pasajeros que volarán en un determinado vuelo, lo que les permitirá tomar decisiones más informadas y reducir el riesgo de pérdidas financieras.

Existen varios tipos de modelos analíticos para la predicción de series de tiempo, a continuación, te menciono algunos de los más comunes:

- **SARIMA**

  El modelo SARIMA es una técnica de pronóstico utilizada para analizar y predecir datos temporales. Es una extensión del modelo ARIMA que incorpora términos de tendencia estacional para capturar patrones periódicos en los datos. El SARIMA se basa en el análisis de tendencias y patrones históricos para predecir el comportamiento futuro. Este modelo se aplica en diversos campos como la economía, la meteorología y la ingeniería. Al igual que otros métodos de análisis de series temporales, el SARIMA se construye combinando componentes como la autocorrelación, la diferenciación y la media móvil estacional. Estos componentes permiten capturar las características específicas de los datos y mejorar la precisión de las predicciones [10]. Así como otros métodos tradicionales de análisis de series temporales, SARIMA se construye con la combinación de uno o más de los siguientes componentes [11]:

  - **Componente estacional (S)**: El parámetro S en el modelo SARIMA (Seasonal Autoregressive Integrated Moving Average) indica el número de períodos de tiempo necesarios para que el patrón se repita. En otras palabras, S representa la periodicidad o los comportamientos cíclicos que pueden observarse en las series temporales.
  - **Componente autorregresivo (AR)**: en este caso, el modelo de la serie de tiempo es la representación de un proceso aleatorio, en el que la variable a pronosticar depende de sus observaciones pasadas. En particular, si una serie de tiempo está representada por un modelo AR(4), indica que el pronóstico de un dato depende de las cuatro(4) observaciones anteriores de la serie.
  - **Componente de integración**: Para que los datos secuenciales se puedan modelar mediante modelos estadísticos de series de tiempo, se requiere que esta sea estacionaria; es decir; que la media de los datos sea igual a cero (no presente ningún tipo de tendencia) y que la varianza sea constante. En los casos donde los datos no cumplen estas dos condiciones, se puede realizar un proceso de diferenciación de la serie. Que básicamente consiste en consiste en calcular la diferencia entre cada dato de la serie y el anterior.
  - **Componente de media móvil (MA)**: Este tipo de modelo se centra en la acumulación de términos de

error en el modelo autorregresivo. Por ejemplo, en un modelo de tipo MA(2) indica que el valor actual de lo observado es la acumulación de ruido blanco de las dos últimas observaciones.

A modo de resumen se presenta en la siguiente imagen las tipologías de modelos de series de tiempo tradicionales.

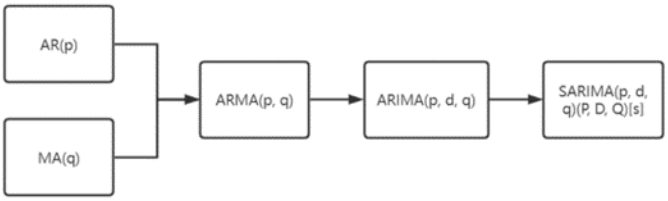

Figura 1. Diferentes tipologias de modelos de series de tiempo

- **Modelos de Regresión**
  La regresión es parte de las técnicas de Machine Learning supervisado, con esta se quiere encontrar una función que recoja la relación que hay entre la variable objetivo y conjunto de características [1]. El ejemplo más sencillo de un algoritmo de regresión es la regresión lineal, en la que se asume una relación lineal entre los predictores y la variable objetivo y se construye una línea media que se utiliza para predecir el valor de la variable objetivo.
  Tradicionalmente en Machine Learning, las características (predictores) usadas en la regresión son estáticas y no se tienen en cuenta relaciones con el tiempo, por lo que no se espera que cambien sus valores, por ejemplo, la predicción del valor de una vivienda usando predictores como: el número de cuartos, el número de baños, metros cuadrados, accesibilidad y polución. En estos casos se usan algoritmos de Machine Learning como Random Forest [12], XGBoost [13] incluso regresión lineal. Para los métodos tradicionales de series de tiempo, la regresión es usada para ajustar modelos autoregresivos AR [14] con estos se crean funciones matemáticas con las que se modelan los valores recientes y/o estacionalidad de la serie de tiempo, volviendo al ejemplo de la predicción del precio de la vivienda ya no se usan variables como el número de baños y el número de habitaciones, sino que se modela usando el precio de venta anterior de la casa, pero esto es viable si se tiene el valor del precio de la vivienda periódicamente.
  Existe otra forma de modelar las series de tiempo con regresión, esta requiere que los datos de la serie de tiempo se traten de manera tabular, aquí los valores pasados son las características (predictores) de valores futuros. Se transforma la serie de tiempo de un vector largo de dimensión L x 1 a una matriz L x C, donde D es el largo de la serie de tiempo y C es la cantidad de características seleccionadas de la serie de tiempo. La forma de seleccionar estas características es muy variada, Kang et al. [-], proponen usar los valores pasados de la serie, estadísticas de la serie mensuales, trimestrales, semestrales, anuales, tendencia, estacionalidad, entre otras. Fulcher et al. [15] introduce el algoritmo Highly Comparative Time Series Analysis (HCTSA) con el que se pueden obtener hasta 7000 características de la serie de tiempo y en [16] se introduce el algoritmo The Canonical Time Series Characteristics (Catch22) que es una mejora de HCTSA, con el cual reduce el número de características a 22, las más discriminantes.
  Una vez seleccionadas las características, se procede a usar cualquier modelo de regresión de la manera usual. Algunas de las más usadas son:

  - Support vector regression (SVR) [17]: Este algoritmo es usado para predecir valores discretos, usa el mismo principio de Support Vector Machine, se busca encontrar hiperplanos que clasifiquen/separen muy bien los puntos de datos, esto lo hace usando una función kernel que puede ser lineal o sigmoidal o polinomial. La línea que modele el hiperplano que contenga la mayor cantidad de datos es la seleccionada. En el SVR, los hiper parámetros que se deben ajustar son: el hiperplano, el kernel y las líneas límite. Esta técnica de Machine Learning es robusta a datos atípicos [8].
  - Random Forest [12]: Es un algoritmo muy robusto, este es una agregación bootstrap, también conocida como bagging, la data es separada aleatoriamente con reemplazo, en cada pedazo se entrena un árbol de decisión, luego se agrega el resultado de todos los modelos con la media. Bagging ayuda a reducir la gran varianza de la predicción de cada árbol. En el random forest, los dos hiper parámetros que hay que ajustar son: el número de árboles y el número de características en cada nodo.
  - Extreme Gradient Boosting XGBoost [13]: Similar al Random Forest XGBoost es un ensamble de árboles de decisión, pero el ensamble se realiza con gradient boosting con regularización, esto ayuda que se reduzca el sesgo en la predicción. Booting consiste en construir modelos consecutivos que se entrenan con los errores que quedan después de cada predicción. Regresiones Lasso y Ridge: Estos son unas técnicas de regularización, consisten en encoger el peso de los coeficientes del modelo, es decir penalizarlos. Cuando se usa la regularización L1 la técnica se llama Lasso (Leat Absolute Shrinkage and Selecto Operator) Donde, λ es el término de penalización, determina el tamaño del encogimiento que se va a realizar al modelo, cuando se elige un λ grande, las características menos importantes serán eliminadas. Cuando el valor de λ es 0, tenemos una ecuación equivalente a la de una regresión por mínimos cuadrados.

- **TSLM**
  Las redes neuronales recurrentes son una clase de aprendizaje profundo basada en los trabajos de David Rumelhart en 1986. La característica principal de las redes recurrentes es que la información puede mantenerse en el modelo introduciendo bucles en el diagrama de la red, permitiendo que la serie recuerde los estados previos y

utilice esta información para predecir. Esta característica las hace muy adecuadas para manejar series cronológicas con memoria de corto plazo [18].
Un caso particular de redes neuronales recurrentes son los modelos LSTM (Long Short Term Memory), que tienen la particularidad de poder "recordar" un dato relevante en la secuencia y de preservarlo por varios instantes de tiempo, por lo que el modelo puede tener una memoria tanto de corto plazo como también de largo plazo.

- **Técnicas estadísticas robustas y no paramétricas**
  El análisis no paramétrico es una técnica que no requiere asumir una distribución específica para los datos y que se basa en observar los datos tal y como son, sin hacer suposiciones sobre la naturaleza de los mismos. En el contexto de series de tiempo, las técnicas no paramétricas pueden ser útiles para analizar datos que presentan una distribución no normal o para identificar patrones no lineales en la dinámica temporal.
  La técnica de datos funcionales, por su parte, se enfoca en tratar las observaciones como funciones continuas en vez de valores discretos en cada punto en el tiempo. De esta manera, se pueden hacer comparaciones entre diferentes curvas o series de tiempo, identificar patrones comunes, y aplicar técnicas de interpolación y extrapolación más precisas.
  En el análisis de series de tiempo, la técnica de datos funcionales se utiliza para modelar la dinámica temporal de una serie de tiempo como una función continua que varía con el tiempo. Esta técnica se basa en la idea de que los datos de una serie de tiempo pueden ser vistos como funciones continuas, en vez de valores discretos aislados. Esta técnica también permite hacer comparaciones entre diferentes series de tiempo, ya que se pueden representar de manera visual las diferencias y similitudes entre ellas.

IV-C1. Tecnologías para la construcción de modelos predictivos: Para la construcción de modelos de predicción en el desarrollo del proyecto, se utilizaron tecnologías que facilitaron el ejercicio, tales cómo:

Statsforecast de Nixtla
Statsforecast es una biblioteca específica para el modelado y la predicción de series de tiempo en Python. Esta biblioteca proporciona una variedad de modelos de series de tiempo que fueron utilizados en este proyecto. Statsforecast simplificó la implementación y evaluación de los modelos, permitiendo explorar diferentes enfoques y seleccionar el modelo más adecuado para cada caso. Esta librería está optimizada para la creación de modelos de series de tiempo y su desempeño y eficiencia es mayor a otras tecnologías en el mercado como Prophet de Meta o statsmodels [19] .

A continuacipon se relaciona cada una de las tipologías de modelos implementados en el proyecto, gracias a la librería Statsforecast en Python:

- AutoARIMA: Esta herramienta permite de manera automática parametrizar y seleccionar el mejor modelo de serie de tiempo tipo SARIMA tradiciona. Son especialemente útiles cuando se trabaja con grandes conjuntos de series temporales univariadas
- HoltWinters: Esta herramienta permite en series de tiempo que no presentan tendencia o estacionalidad clara, realizar un susvizado exponencial utilizando el promedio ponderado de las observaciones históricas.
- CrostonClassic: Herramienta potente para pronosticar series de tiempo con datos intermitentes o dispersos. A modo general, el método consiste en dividir la serie de tiempo en dos componentes distintos: la primera parte corresponde a una serie que contiene valores positivos de demanda, mientras que la segunda parte representa los intervalos de tiempo entre demandas consecutivas no nulas.
- HistoricAverage: Modelo que utiliza el promedio histórico para la generación de pronósticos futuros. Una de las principales ventajas del modelo HistoricAverage es su simplicidad y facilidad de implementación. No requiere de un análisis complejo ni de la consideración de factores adicionales.
- DynamicOptimizedTheta: Modelo de serie de tiempo más complejo, que Ss basa en el concepto de las líneas theta, que son técnicas utilizadas para desestacionalizar una serie de tiempo y pronosticar sus valores futuros. Para ello, el modelo ajusta dos líneas theta a la serie y que representan dos componentes principales: la tasa de nivel; que estima la tendencia general de la serie; y la tasa de cambio; que captura la variabilidad y las fluctuaciones de corto plazo.
- SeasonalNaive: Modelo de pronóstico que utiliza el último valor observado para la misma temporada y mes del año como la predicción para el período futuro. Este enfoque supone que los patrones estacionales se repetirán de manera similar en cada temporada.

La combibación de tecnologías (Python, Polars, Pyspark y Statsforecast de Nixtla), proporcionaron una sólida base para el procesamiento eficiente de grandes volúmenes de datos y la construcción de modelos de predicción precisos en este proyecto. Su utilización permitió realizar tareas de limpieza, agregación, modelado y evaluación de manera efectiva, proporcionando resultados valiosos para la predicción del tráfico aéreo en Colombia.

V. Metodología

Para llevar a cabo el proyecto se utilizó la metodología CRISP-DM, que es considerada como la guía de referencia más ampliamente utilizada en proyectos de analítica y minería de datos [1]. Esta metodología comprende seis fases (ver Figura 1) que guiarán todo el proceso del proyecto:

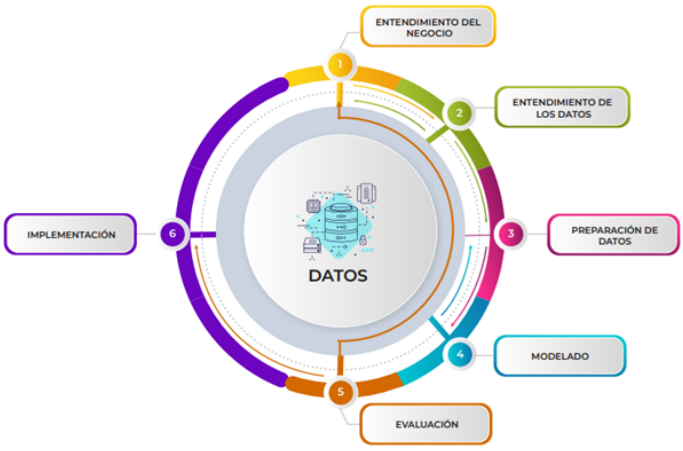

Figura 2. Fases empleadas en la metodología CRISP-DM.

VI. Plan de Trabajo

Para el desarrollo de este análisis se contó con un total de ocho (8) semanas y desde la propuesta se programaron unas actividades generales para alcanzar los objetivos propuestos. Sin embargo, como en cualquier ejercicio, existen dificultades y contratiempos que afectan la ejecución del ejercicio y, para este caso putual, representó un retraso de una semana en las actividades. En la Figura 3 se puede observar tanto el cronograma propuesto como el ejecutado.

Durante la programación del análisis, se encontraron retrasos en la obtención de comentarios de Twitter (datos requeridos, considerando la propuesta inicial) debido al desconocimiento de que ahora se requería un registro y pago para el acceso a la API de la red social. Ante este obstáculo, se decidió buscar otras alternativas para obtener los comentarios deseados a través de páginas de viajes y otras fuentes de información que redujeran costo y tiempo y que posibilitara no depender exclusivamente de la API de Twitter.

No obstante, a pesar de los esfuerzos realizados, no se logró encontrar una alternativa que de fácil accceso que no implicara mayor tiempo en la ejecución del proyecto. En consecuencia, se plantea considerar dentro de lo pasos a seguir en el proyecto, con el fin de buscar nuevas estrategias que aborden el análisis de sentimientos de manera eficiente y económicamente viable.

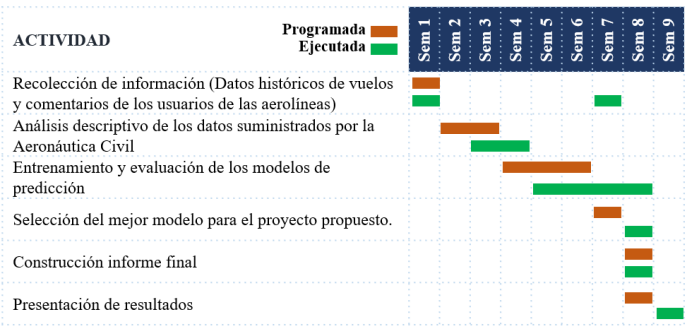

Figura 3. Cronograma propuesto y ejecutado

VII. Resultados

Para el desarrollo metodológico del proyecto se consideró la metodología CRISP-DM relacionada en el capítulo anterior. En este sentido, a continuación, se relacionan una a una las fases del ciclo del proyecto.

Es de anotar que para el desarrollo del presente informe solo se abordará de manera descriptiva cada una de las técnicas, modelos y acciones realizadas en el marco del desarrollo del proyecto, por lo que no se presentará el código generado para lograrlo. Si el lector desea mayor detalle a nivel de programación puede consultar el siguiente repositorio: https://github.com/fjmoyao/trafico-aereo-col

- **Fase 1. Entendimiento del negocio**
  La Aeronáutica Civil de Colombia (Aerocivil) es una entidad gubernamental clave en el ámbito de la aviación civil en Colombia. Como regulador y supervisor, la Aerocivil tiene la responsabilidad de asegurar la seguridad y eficiencia de las operaciones aéreas en el país. Además de supervisar el cumplimiento de los estándares internacionales de seguridad, la Aerocivil también gestiona el espacio aéreo y promueve el desarrollo aeronáutico sostenible.
  En el contexto de este proyecto de investigación, nos enfocamos en desarrollar un modelo de predicción de series de tiempo para predecir el tráfico aéreo de las diferentes aerolíneas con actividad en Colombia. Esta iniciativa busca brindar a la Aerocivil y a otras partes interesadas información precisa y confiable sobre el flujo de tráfico aéreo, lo que les permitirá tomar decisiones informadas y planificar estratégicamente la gestión del espacio aéreo.
  Con la creciente demanda y el continuo crecimiento del sector de la aviación civil en Colombia, es fundamental contar con herramientas y modelos que permitan una gestión eficiente y segura del tráfico aéreo. Al comprender y predecir los patrones de tráfico, la Aerocivil podrá tomar medidas proactivas para optimizar la capacidad y la utilización del espacio aéreo, garantizando así la seguridad de los vuelos, la reducción de retrasos y la mejora general de la experiencia de los pasajeros.
- **Fase 2. Entendimiento de los datos**
  Los datos utilizados en este proyecto fueron extraídos mediante técnicas de *web scraping* del sitio web oficial de la Aeronáutica Civil de Colombia (Aerocivil) [20]. Estos datos comprenden información detallada sobre el tráfico aéreo origen-destino de las diferentes aerolíneas que operan en Colombia desde el año 1992 hasta el año 2023.
  El conjunto de datos se encuentra estructurado en archivos individuales de hojas de cálculo, al concatenar los archivos se cuenta con aproximadamente 1,820,258 registros. Los archivos individuales se generan con una periodicidad mensual que permite un análisis temporal preciso. Cada archivo corresponde a un mes específico y contiene información detallada sobre el tráfico aéreo, incluyendo las siguientes variables:
  - **Fecha**: Mes de operación del vuelo.
  - **Sigla empresa**: Sigla OACI utilizada para identificar a la empresa ante las Autoridades Aeronáuticas.
  - **Origen**: Sigla IATA del aeropuerto de donde se

embarcan los pasajeros, la carga y/o el correo.
- **Destino**: Sigla IATA del aeropuerto donde culmina el viaje del pasajero, la carga y/o el correo.
- **Pasajeros**: Número de pasajeros por cuyo transporte la línea aérea percibe remuneración comercial.
- **Tráfico**: Categoría que indica el tipo de tráfico del vuelo, representado por las siguientes etiquetas: N (Tráfico Doméstico), I (Tráfico Internacional) y E (Tráfico entre dos aeropuertos fuera de Colombia).
- **TipoVuelo**: Clasificación del tipo de operación entre el par correspondiente de ciudades origen-destino. Las etiquetas utilizadas son: R (Operación Regular), A (Vuelos Adicionales), C (Vuelos chárter) y T (Taxi Aéreo).
- **Ciudad Origen**: Nombre de la ciudad de origen del vuelo.
- **Ciudad Destino**: Nombre de la ciudad de destino del vuelo.
- **País Origen**: Nombre del país de origen del vuelo.
- **País Destino**: Nombre del país de destino del vuelo.
- **Nombre Empresa**: Sigla OACI con la cual se identifica la empresa ante las Autoridades Aeronáuticas.
- **Apto_Origen**: Nombre completo del aeropuerto de origen.
- **Apto_Destino**: Nombre completo del aeropuerto de destino.

Al realizar un análisis exploratorio sobre los datos se identificaron las aerolineas con mayor participación en Colombia, así como también la tendencia del tráfico aéreo a traves de los años.

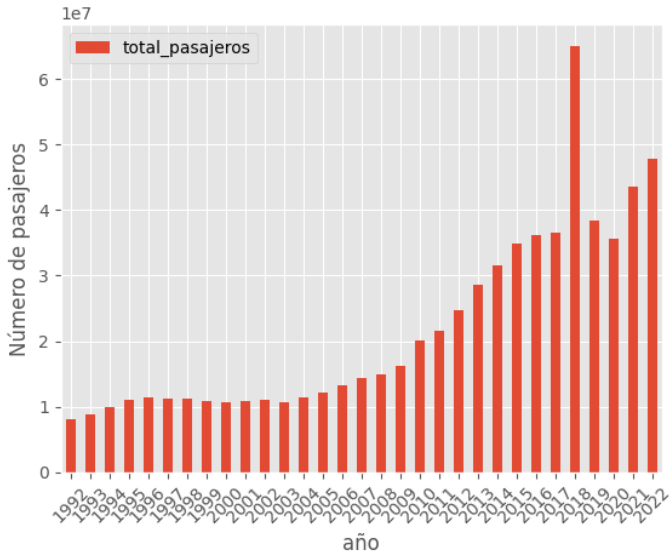

Figura 4. Tráfico aéreo en Colombia a través de los años

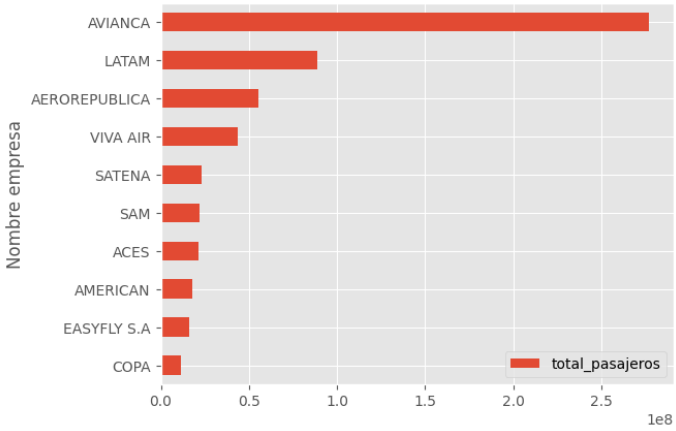

Figura 5. Tráfico aéreo en Colombia por aerolinea

Con base a lo observado se decidió realizar el entrenamiento de los modelos sobre el top 10 de aerolíneas de acuerdo a la cantidad de pasajeros que transportan. Además, se iteraron modelos con la totalidad de los datos y con datos de los ultimos 10 años. A continuación se visualizan las series de tiempo correspondientes:

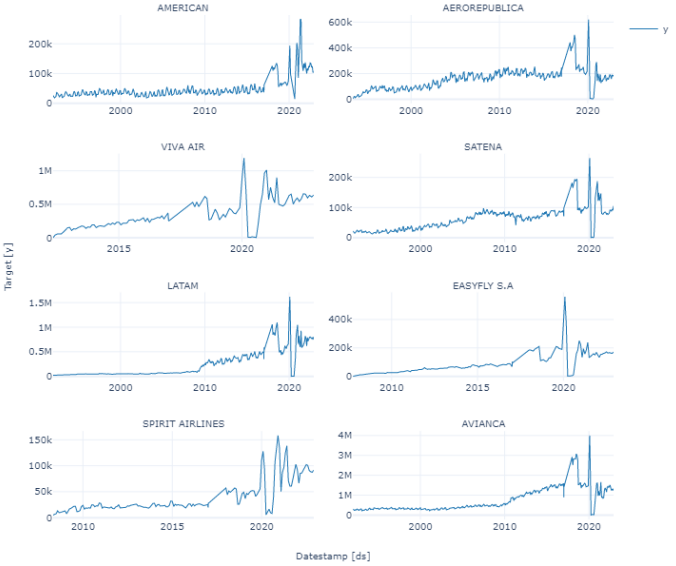

Figura 6. Series de tiempo de las top 10 aerolineas 1992-2022

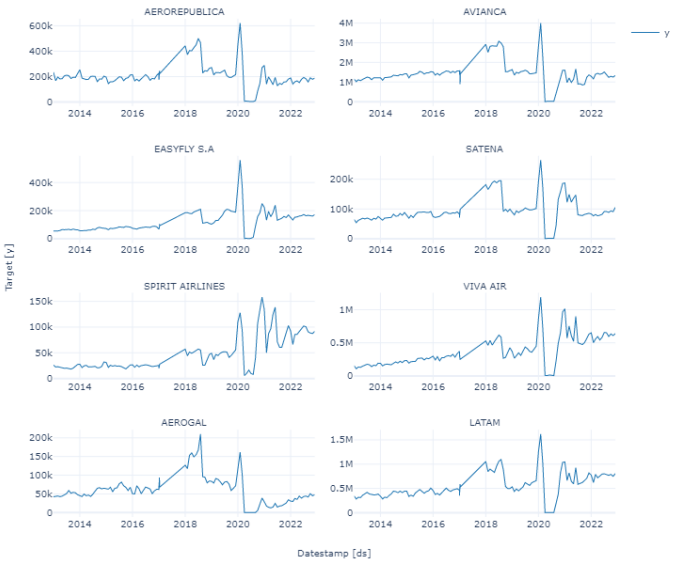

Figura 7. Series de tiempo de las top 10 aerolineas 2013-2022

- **Fase 3. Preparación de los datos**
  Se realizaron varios pasos de procesamiento de datos para prepararlos adecuadamente antes de su análisis. Los datos se organizaron en diferentes zonas según el nivel de procesamiento aplicado: 'Raw', donde se almacenaron los datos sin modificar; 'Trusted', donde se guardaron los datos preprocesados para realizar análisis preliminares; y 'Refined', donde se almacenaron los datos procesados utilizados para generar los análisis reportados y entrenar los modelos.
  En la etapa inicial, se descargaron los datos de la página web de la Aerocivil y se almacenaron en la zona Raw". Posteriormente, se empleó la librería de Python llamada Polars para leer los archivos individuales y consolidarlos en un único archivo general. Además, se estableció un esquema común para los datos, ya que los diferentes archivos seguían esquemas distintos. La elección de Polars se basó en su similitud en sintaxis con Pandas y su capacidad para leer archivos en formato .xlsx, lo cual facilitó la tarea en comparación con Pyspark, que requiere una configuración adicional para leer este tipo de archivos. El resultado de esta etapa se almacenó en la zona "Trusted".
  A continuación, se utilizó Pyspark para llevar a cabo la limpieza de los datos. Se eliminaron los valores nulos y se corrigieron errores de tipado en diversas variables. Además, se realizaron agregaciones específicas para responder a distintas interrogantes sobre los datos: ¿Cuales son las aerolineas con mayor trafico aereo? ¿Hay diferencias significativas entre el tráfico nacional e internacional?. Estas agregaciones se llevaron a cabo tanto a nivel mensual como por aerolínea, realizando dos iteraciones: una para el tráfico nacional y otra para cualquier tipo de tráfico. Los resultados finales de estas agregaciones se guardaron en la zona Refined".
  A continuación se visualiza la arquitectura de datos empleada en el proyecto:

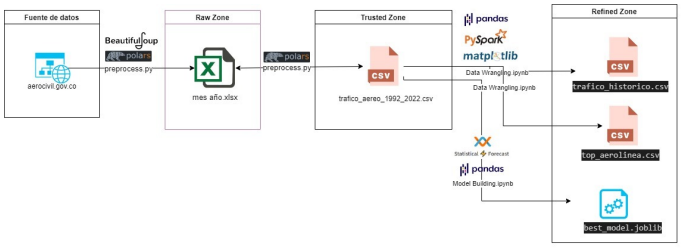

Figura 8. Arquitectura de consumo y procesamiento de datos

- **Fase 4. Modelado**
  En la etapa de modelado, se utilizó la biblioteca de Python statsforecast. Se seleccionaron las 10 aerolíneas con mayor tráfico aéreo en Colombia, que aún estaban activas hasta el año 2023, y se realizaron iteraciones con diferentes modelos de series de tiempo. El objetivo era evaluar el ajuste y el rendimiento de los modelos para determinar el más adecuado en cada caso.
  Se llevaron a cabo dos iteraciones: una utilizando la totalidad de los datos y otra utilizando los últimos 10 años de información. Esta segunda iteración tenía como propósito validar si la información más antigua seguía siendo relevante para las predicciones futuras.
  Los modelos que se iteraron fueron los siguientes:

  - AutoARIMA
  - HoltWinters
  - CrostonClassic
  - HistoricAverage
  - DynamicOptimizedTheta
  - SeasonalNaive.

  A continuación se observan las series de tiempo con sus predicciones correspondientes usando los diferentes modelos.

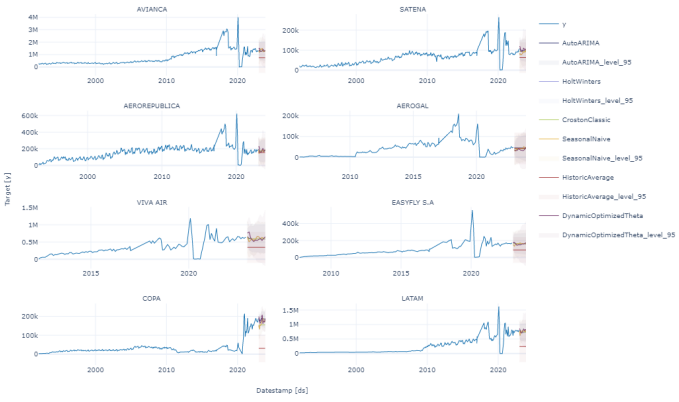

Figura 9. Predicción del tráfico aereo para las top 10 aerolineas con diferentes modelos

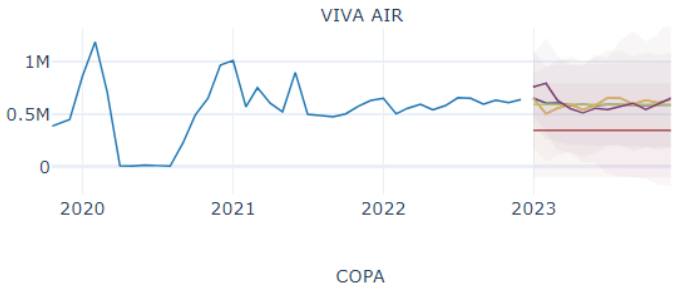

Figura 10. Predicción para una aerolinea con diferentes modelos

- **Fase 5. Evaluación de los modelos**
  Se utilizó el MAPE para la evaluación del desempeño de los diferentes modelos, adicionalmente se realizó crossvalidation para la evaluación. Con este proiceso se seleccionó el mejor modelo para cada serie de tiempo.
  A continuación, se visualizan los resultados para las series con la totalidad de los datos y solo los ultimos 10 años de información.

Cuadro I
EVALUACIÓN DE MODELOS DE PREDICCIÓN (ULTIMOS 10 AÑOS DE INFORMACION)

Aerolínea	MAPE	Modelo
AEROGAL	0.47	DynamicOptimizedTheta
AEROREPUBLICA	0.13	HistoricAverage
AMERICAN	0.26	CrostonClassic
AVIANCA	0.16	AutoARIMA
COPA	0.28	DynamicOptimizedTheta
EASYFLY S.A	0.10	CrostonClassic
LATAM	0.18	AutoARIMA
SATENA	0.22	HistoricAverage
SPIRIT AIRLINES	0.22	CrostonClassic
VIVA AIR	0.20	CrostonClassic

Cuadro II
EVALUACIÓN DE MODELOS DE PREDICCIÓN (ULTIMOS 10 AÑOS DE INFORMACION)

Aerolínea	MAPE	Modelo
AEROGAL	0.63	DynamicOptimizedTheta
AEROREPUBLICA	0.13	CrostonClassic
AMERICAN	0.19	AutoARIMA
AVIANCA	0.18	CrostonClassic
COPA	0.30	DynamicOptimizedTheta
EASYFLY S.A	0.10	CrostonClassic
LATAM	0.17	AutoARIMA
SATENA	0.18	HistoricAverage
SPIRIT AIRLINES	0.22	CrostonClassic
VIVA AIR	0.16	AutoARIMA

VIII. CONCLUSIONES

En este proyecto, se han construido modelos de predicción de series de tiempo para predecir el tráfico aéreo en diferentes aerolíneas con actividad en Colombia. A lo largo del desarrollo del proyecto, se han aplicado diversas técnicas y tecnologías de procesamiento de grandes volúmenes de datos y modelos de predicción.

En cuanto al procesamiento de datos, se utilizó Python como lenguaje de programación principal, aprovechando su amplia gama de bibliotecas y herramientas para manipular y procesar datos. La biblioteca Polars se destacó por su rendimiento y capacidad para manejar conjuntos de datos más grandes que la memoria disponible. Con Polars, se logró una manipulación eficiente de DataFrames, lo que facilitó el procesamiento y análisis de los datos del tráfico aéreo.

Además, se utilizó PySpark, una biblioteca de Apache Spark, para el procesamiento y la limpieza de los datos. Spark permitió realizar tareas de manera distribuida y escalable, lo que resultó especialmente útil para trabajar con grandes volúmenes de datos. Con PySpark, se eliminaron los valores nulos, se corrigieron los errores de tipado y se generaron agregaciones necesarias para el análisis de los datos.

En cuanto al modelado, se empleó la biblioteca statsforecast de Nixtla, que proporcionó implementaciones eficientes de varios modelos de series de tiempo. Se iteraron diferentes modelos, considerando las características específicas de cada aerolínea y evaluando su ajuste y desempeño en la predicción del tráfico aéreo. Se realizó una validación comparando el desempeño de los modelos entrenados con la totalidad de los datos y los últimos 10 años de información para determinar la relevancia de los datos más antiguos en la predicción futura.

Se observa que los modelos ajustados con los ultimos 10 años de información logran un desempeño similar o superior a los modelos entrenados con la totalidad de los datos, por lo cual se infiere que las características y tendencia de los datos antiguos tienen poco efecto sobre los datos mas recientes y su capacidad prediciva es limitada.

Es importante destacar que se identificó heterogeneidad en el mercado aéreo colombiano, ya que los modelos con mejor desempeño variaron según la aerolínea evaluada. Esto sugiere que las aerolíneas responden a dinámicas distintas y tienen características únicas en cuanto a su tráfico aéreo.

En conclusión, este proyecto ha demostrado la eficacia de las tecnologías utilizadas en el procesamiento de grandes volúmenes de datos y modelos de predicción. La combinación de Python, Polars, PySpark y statsforecast ha permitido realizar un análisis exhaustivo del tráfico aéreo en Colombia y construir modelos de predicción precisos. Estas herramientas y técnicas son altamente aplicables en otros proyectos de análisis y predicción de datos, ofreciendo resultados precisos y escalables.

IMPLICACIONES ÉTICAS

El presente proyecto se centra en el análisis de datos de acceso público relacionados con el tráfico aéreo de las aerolíneas que operan en Colombia. Estos datos son recopilados y proporcionados por la autoridad aeroportuaria del país. Aunque no se han identificado riesgos evidentes de falta de ética en el procesamiento de la información, se han tomado las siguientes medidas para minimizar cualquier posible impacto negativo en la privacidad de los usuarios y garantizar su protección:

1. Se utilizaron datos anonimizados y de acceso público que no contienen información personal identificable.
2. Se garantizó el procesamiento y análisis de los datos de acuerdo con las políticas de privacidad y términos de servicio de la página web consultada.

3. Se adoptó una metodología de análisis responsable y se evitó el uso de los datos de manera perjudicial o discriminatoria.
4. Se garantizó la propiedad intelectual de los datos y que cualquier publicación o divulgación de los resultados se realice de manera ética y responsable.

En resumen, se tuvo en cuenta la privacidad y protección de los datos en todas las etapas del proyecto, asegurándose de cumplir con los estándares éticos apropiados para preservar la integridad de la investigación.

ASPECTOS LEGALES Y COMERCIALES

En el desarrollo de este proyecto y otros similares, es importante asegurarse de cumplir con las políticas de privacidad y términos de servicio de las páginas web utilizadas para la recolección de datos. Además, considerar la propiedad intelectual de los datos recopilados y utilizarlos de acuerdo a las leyes y regulaciones vigentes.

En cuanto a los aspectos comerciales, es necesario tener en cuenta la competencia en el mercado de las aerolíneas y asegurarse de cumplir con las regulaciones en cuanto a publicidad y promoción de los resultados del proyecto. También considerar la posibilidad de que los resultados del proyecto pueden ser utilizados con fines comerciales, que requieren permisos para ello.

REFERENCIAS

[1] IBM, ÏBM,"05 06 2022. [En línea]. Disponible: https://www.ibm.com/docs/es/SS3RA7_18.3.0/pdf/ModelerCRISPDM.pdf.

[2] C. Á. Fierro Torres, V. H. Castillo Pérez, y C. I. Torres Saucedo, "Análisis comparativo de modelos tradicionales y modernos para pronóstico de la demanda: enfoques y características," *RIDE. Rev. Iberoam. Investig. Desarro. Educ*, vol. 12, no. 24, ene./jun. 2022. doi: 10.23913/ride.v12i24.1203.

[3] O. Díaz-Olariaga, Çontribución del transporte aéreo a la conectividad territorial. El caso de Colombia,"*EURE (Santiago)*, vol. 47, no. 140, pp. 77-100, ene. 2021. ISSN: 0250-7161. doi: 10.7764/eure.47.140.06.

[4] J. C. Garmendia-Mora, "Satisfacción y lealtad del cliente en las operaciones domésticas de las aerolíneas colombianas,"Forum Empresarial, vol. 24, no. 1, 2019.

[5] Apache Software Foundation. (2023). Apache Hadoop. Recuperado de https://hadoop.apache.org/

[6] .Apache Software Foundation. (2023) Apache Spark. Recuperado de https://spark.apache.org/

[7] . Python Software Foundation. (2023). Tutorial de Python (Versión 3). Recuperado de https://docs.python.org/es/3/tutorial/

[8] . Pola-RS. (2023). Guía del usuario de Polars. Recuperado de https://pola-rs.github.io/polars-book/user-guide/

[9] Apache Software Foundation. (2023). Documentación de Apache Spark (Versión más reciente). Recuperado de https://spark.apache.org/docs/latest/api/python/

[10] J. Durbin and S. J. Koopman, *Time Series Analysis by State Space Methods*, Oxford University Press, 2001.

[11] F. A. Gers, J. Schmidhuber, and F. Cummins, *Learning to Forget: Continual Prediction with LSTM*, *Neural Comput*, vol. 12, pp. 2451-2471, 2000.

[12] W. G. Sammut C, Çomputationally efficient heart rate estimation during physical,"*Encyclopedia of machine learning*. Springer, Berlin, p. 2478–2481, 2017.

[13] L. Breiman, Random Forest,"*Mach Learn*, vol. 45, no. 1, pp. 5–32, 2001.

[14] G. C. Chen T, "XGBoost: a scalable tree boosting system,"*Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, p. 785–794, 2016.

[15] S. E. A. V. Makridakis S, "The M4 competition: results, findings, conclusion and,"*Int J Forecast*, vol. 34, no. 4, p. 802–808, 2018.

[16] L. M. J. N. Fulcher BD, "Highly comparative time-series analysis: the empirical structure of,"*J R Soc Interface*, vol. 10, no. 83, 2013.

[17] C. S. S. K. P. e. a. Lubba, çatch22: CAnonical Time-series CHaracteristics,"*Data Min Knowl Disc*, vol. 33, p. 1821–1852, 2019.

[18] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory,"*Neural Comput*, vol. 9, no. 8, p. 1735–1780, 1997.

[19] Nixtla, "Nixtla: Sitio web de referencia,"*Disponible en: https://www.nixtla.io/*.

[20] Aerocivil, .Estadísticas de las actividades aeronáuticas - Bases de datos,"*Disponible en: https://www.aerocivil.gov.co/atencion/estadisticas-de-las-actividades-aeronauticas/bases-de-datos*.