Predicción Dinámica del Valor del Flete de Mercado para Vehículos 3S3 del Puerto de Buenaventura a Bogotá: Un Modelo Integrado con Variables Exógenas Económicas y del Sector Logístico

Estudiante: Camilo Alejandro Vélez Medina

Director: Johan Felipe García Vargas

Área de computación y analítica

Ciencias aplicadas e ingeniería

jfgarcaav@eafit.edu.co

Palabras clave: Time Series Forecasting; ARIMA; Random Forest; LSTM, Costo Flete Transporte, Sector Logístico, Transporte Terrestre Colombia.

Contenido

1. Resumen

La logística, en especial el transporte terrestre como parte fundamental de la cadena de abastecimiento, afecta directamente los costos y la disponibilidad de los productos en las ciudades, este proyecto desarrolla un modelo predictivo para estimar el valor de mercado en el transporte de carga para vehículos tipo 3S3 desde el puerto de Buenaventura, Colombia, destino a la ciudad de Bogotá, Colombia, variable que llamaremos FP_mean correspondiente al promedio diario del flete de producción; la innovación del modelo radica en su capacidad para integrar variables exógenas críticas, como la cotización del petróleo Brent, la tasa de cambio del dólar, factores específicos del sector recogidos en el SICE TAC (combustible, peaje, llantas, lubricantes, filtros, mantenimiento, personal), RNDC (registro nacional despachos de carga por carretera) y llegada de buques al puerto con su respectivo tipo de mercancía.

Se evaluaron múltiples enfoques avanzados de modelado, incluidos Arima, Sarima, Random Forest y LSTM y combinaciones entre los mismos, destacándose el modelo basado en Random Forest con variables exógenas (random_forest_exogen) por su desempeño superior, logrando un RMSE de 211,395.42 y un MAPE de 3.20%, siendo el más preciso para la estimación del FP_mean; Adicionalmente, los modelos LSTM y SARIMA también mostraron resultados competitivos con un equilibrio entre precisión y estabilidad en escenarios diversos; estos hallazgos subrayan la importancia de combinar técnicas avanzadas de aprendizaje automático con el conocimiento del dominio logístico.

2. Abstract

Logistics, especially road transportation as a fundamental part of the supply chain, directly impacts the costs and availability of products in cities. This project develops a predictive model to estimate the market value of freight transportation for 3S3-type vehicles from the port of Buenaventura, Colombia, to Bogotá, Colombia. The variable of interest, referred to as FP_mean, corresponds to the daily average freight production cost. The innovation of the model lies in its ability to integrate critical exogenous variables, such as Brent crude oil prices, the exchange rate of the dollar, sector-specific factors collected in the SICE TAC (fuel, tolls, tires, lubricants, filters, maintenance, personnel), RNDC (National Road Cargo Dispatch Registry), and the arrival of ships at the port with their respective types of cargo.

Multiple advanced modeling approaches were evaluated, including ARIMA, SARIMA, Random Forest, and LSTM, with the Random Forest model incorporating exogenous variables (random_forest_exogen) standing out for its superior performance, achieving an RMSE of 211,395.42 and a MAPE of 3.20%, making it the most accurate for estimating FP_mean. Additionally, the LSTM and SARIMA models also demonstrated competitive results, striking a balance between accuracy and stability across various scenarios. These findings highlight the importance of combining advanced machine learning techniques with domain expertise in logistics.

2. Planteamiento del Problema

En el contexto del transporte de carga en Colombia, los puertos juegan un papel crucial en la dinámica económica y logística del país. Particularmente, el puerto de Buenaventura se destaca por su relevancia estratégica en la ruta Buenaventura – Bogotá. Durante una evaluación inicial, se identificó que este puerto exhibe variaciones significativas en los fletes, lo que impacta directamente en los costos operacionales y la planificación logística a Eduardo Botero Soto S.A.

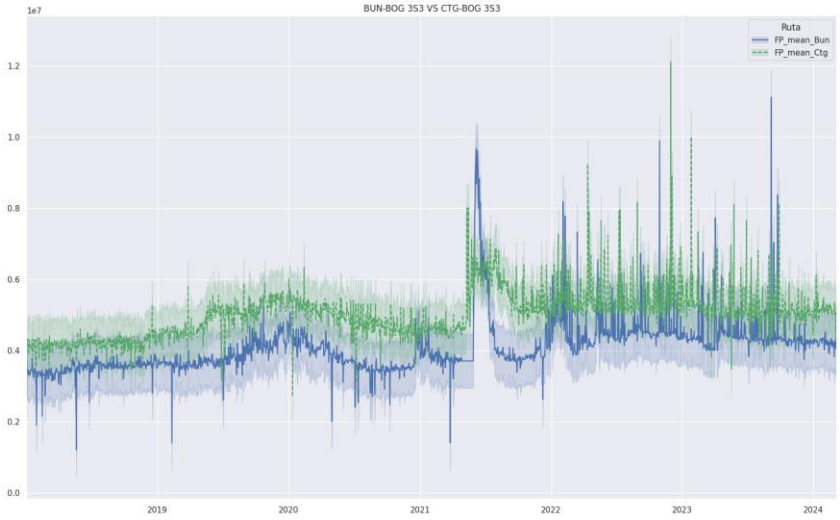
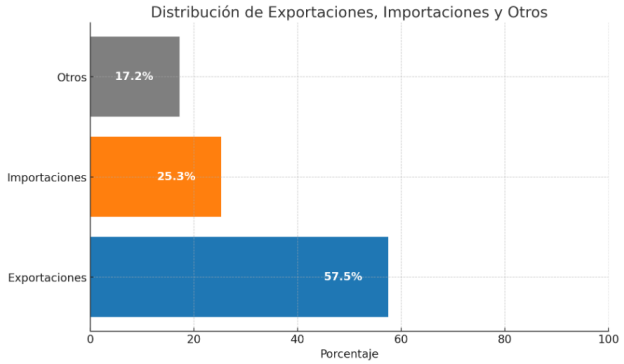


Gráfico de Variación de Fletes Promedio mercado con Destino a Bogotá - 1 desviación estándar

Estas variaciones en los fletes se intensifican debido a la descompensación que se observa a continuación en los corredores de carga de transporte, lo que añade un nivel adicional de complejidad al manejo de las operaciones logísticas en la región.



Participación del tráfico portuario movilizado por Colombia (%) 2022 [1]

Dado el escenario descrito, es imprescindible desarrollar un modelo predictivo que permita estimar con precisión los fletes de producción para los vehículos tipo 3S3 en esta ruta crítica. Tal modelo facilitaría una toma de decisiones más informada y eficiente, tanto para los actores del sector privado, como para los reguladores.

La importancia de un modelo así se ve subrayada por regulaciones como la Resolución 20213040034405 [2], expedida por el Ministerio de Transporte el 6 de agosto de 2021, que establece los costos eficientes de operación publicados en el SICE-TAC como de obligatorio cumplimiento; esta normativa asegura que los pagos no deben ser inferiores a los costos operativos establecidos y es aplicable a todas las tipologías vehiculares especificadas en la resolución.

Además, el Índice de Costos del Transporte de Carga por Carretera (ICTC) [3] ofrece una métrica esencial para el seguimiento de las variaciones en los precios de los insumos necesarios para la operación de transporte. Este indicador es una herramienta valiosa para la formulación de políticas y estrategias, tanto por parte del gobierno como por el sector privado.

Este estudio busca utilizar estos datos e indicadores para desarrollar un modelo predictivo robusto, ofreciendo así una herramienta estratégica para optimizar las operaciones de transporte y cumplir con las regulaciones vigentes, ayudando a los operadores logísticos a reducir la incertidumbre, mantener la eficiencia y competitividad de la empresa.

Los resultados del proyecto tienen un impacto directo en la optimización de la toma de decisiones para Eduardo Botero Soto S.A., permitiendo responder con anticipación a las fluctuaciones del sector, negociar de manera estratégica con proveedores y mejorar la eficiencia operativa y financiera. Además, este enfoque predictivo aporta valor en uno de los puertos más relevantes de Colombia, ayudando a fortalecer la competitividad logística del país, un RMSE de 211.395 se puede decir que el modelo logro capturar la tendencia y leves fluctuaciones en la variable FP_mean.

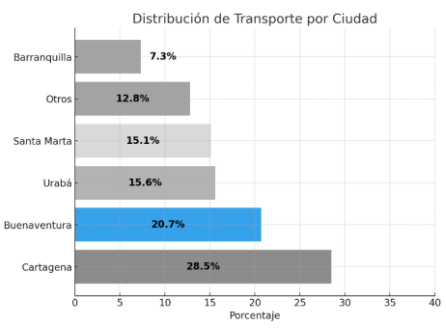
Realizar predicciones con una ventana de 30 días permitió capturar tendencias generales, pero presentó dificultades al predecir picos abruptos en los valores de fletes, esenciales en la planificación logística, aunque la incorporación de variables exógenas enriqueció el modelo, también introdujo una dependencia

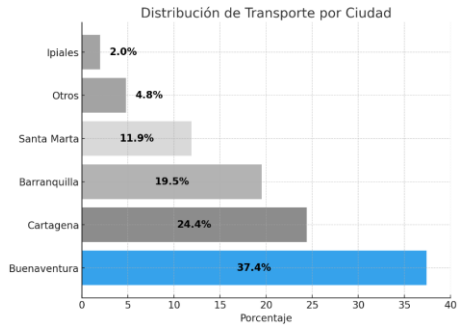
en la predicción futura de estas variables, incrementando el margen de error en escenarios a largo plazo.

Para abordar las limitaciones identificadas y ampliar las capacidades del modelo, se propone explorar a futuro herramientas avanzadas, como los modelos de AutoML ofrecidos por Azure y Databricks. Estas plataformas podrían optimizar la precisión y eficiencia del proceso de predicción, además de facilitar la integración de enfoques más sofisticados que permitan capturar eventos atípicos y tendencias emergentes en el sector. Asimismo, se identificó una aproximación interesante al replantear el problema como una tarea de clasificación; en este enfoque, variables como el peso, el cliente y la distancia se destacaron como las más relevantes para lograr una clasificación precisa que funcione como una predicción del flete. Continuar explorando esta metodología podría generar resultados prometedores. Además, se podría modificar el modelo de predicción para incorporar factores adicionales, como la diferencia entre el costo del flete y el costo real del día. Más adelante, en la sección de resultados, se complementarán estas ideas con hallazgos específicos derivados de este enfoque.

2. Justificación

El puerto de Buenaventura se destaca como el más significativo en Colombia, manejando el 37,4% del tráfico de carga sin minería, petróleo y sus derivados en toneladas. Esta capacidad destaca su importancia para las empresas del sector logístico, haciendo esencial su competitividad y eficiencia para la gestión y planificación operativa. A continuación, se muestran unos gráficos extraídos del informe enfoque competitivo #135 de la cámara de comercio de Cali [1], que revela los porcentajes de participación de los puertos y principales países de origen:

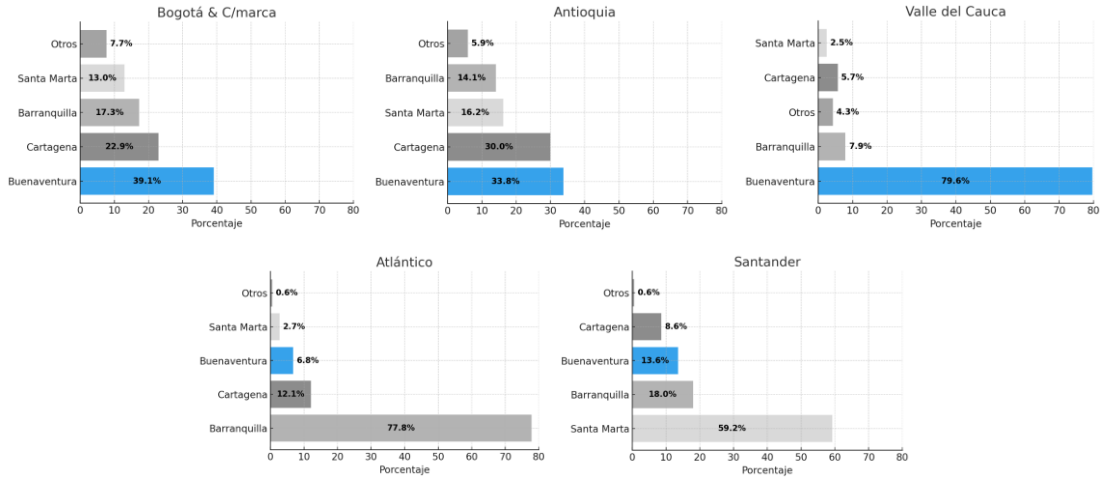




Distribución de las exportaciones [1] *Distribución de las importaciones* [1]

La creación de un modelo predictivo que integre tanto variables exógenas como otras variables críticas adicionales a las disponibles para Eduardo Botero Soto S.A, es fundamental para afinar las estimaciones del valor de mercado de vehículos tipo 3S3. Este modelo predictivo permitirá a los operadores logísticos prever variaciones en los costos y ajustar proactivamente sus estrategias operativas y financieras al optimizar la asignación de recursos, mejorar la toma de decisiones y la presentación de cotizaciones, esta herramienta no solo beneficia la operatividad interna, sino que también eleva la competitividad de Botero Soto con avance tecnológico e innovación en el sector logístico colombiano.

Inicialmente el proyecto se centra en la ruta Buenaventura - Bogotá con vehículos tipo 3S3, lo que servirá como piloto para evaluar la viabilidad del proyecto. Dada su importancia estratégica, se anticipa que el éxito de este modelo podría extenderse para adaptarse a otras rutas y tipos de vehículos en el futuro.



Distribución de las importaciones sin minería, petróleo y sus derivados – principales departamentos, según aduanas (%) 2022 [1]

Con un RMSE de 211,395, el modelo logró una precisión destacable, que se traduce en una desviación aproximada de $6,039 pesos por tonelada para un vehículo 3S3 con capacidad de hasta 35 toneladas. Este enfoque predictivo aporta valor en uno de los puertos más relevantes de Colombia, fortaleciendo la competitividad logística del país.

3. Objetivos

3.1. Objetivo general:

Desarrollar un modelo predictivo que estime el valor de mercado para el transporte de carga en vehículos 3S3 desde el puerto de Buenaventura a Bogotá, integrando variables exógenas económicas y del sector logístico.

3.1.1. Objetivos específicos:

- Identificar y analizar cuáles variables, endógenas y exógenas, serán útiles para predecir el valor del mercado para transporte de carga en vehículos 3S3 de Buenaventura a Bogotá.
- Implementar técnicas de ciencia de datos desde modelos tradicionales de series de tiempo hasta modelos de aprendizaje profundo para la predicción de valor de FP_mean.
- Validar la capacidad predictiva del modelo realizando pruebas con datos históricos, cuantificando la mejora frente al statu quo.

4. Estado del arte

El pronóstico de series temporales es un área crítica de investigación que encuentra aplicaciones tanto en la industria como en la academia y ha utilizado tradicionalmente modelos estadísticos como ARIMA y sus variantes estacionales (SARIMA) debido a su capacidad para capturar patrones lineales en datos secuenciales. ARIMA, propuesto por Box y Jenkins en 1970, combina componentes autorregresivos y de media móvil con diferenciación para lograr la estacionariedad, mientras que SARIMA extiende este enfoque a datos con patrones periódicos; además, modelos como ARIMAX y Modelos de Suavizamiento Exponencial (ETS) incorporan variables exógenas, mejorando la precisión de las predicciones. Aunque son efectivos en la captura de relaciones lineales, estos modelos tienen limitaciones al abordar comportamientos no lineales, lo que ha llevado al desarrollo de métodos híbridos que combinan la interpretabilidad de los modelos estadísticos con la capacidad de los algoritmos de aprendizaje automático para manejar dinámicas complejas.

En el campo logístico, los modelos predictivos se han centrado principalmente en áreas como tasas de accidentalidad, flujo vehicular, costos de mantenimiento, consumo de combustible, demanda y costos generales. No obstante, una revisión exhaustiva de 1,035 documentos indexados en Scopus, utilizando palabras clave como "cost AND prediction", "supply AND chain", "blockchain", y "cost AND logistic AND prediction", revela una notable escasez de investigaciones enfocadas en la predicción de costos

específicos del transporte terrestre en mercados particulares. Esta laguna investigativa se destaca aún más en el contexto colombiano, donde las peculiaridades logísticas locales difieren significativamente de las tendencias globales, limitando hasta ahora el interés en su estudio.

Aunque algunos estudios no tienen exactamente el mismo alcance, pueden servir como referencia y guía para el propósito de esta investigación. Tal es el caso del sistema de apoyo a decisiones (Decision Support System, DSS) desarrollado en Brasil por una empresa de autopartes, desde la perspectiva del generador de carga. Este estudio demostró que, mediante la implementación del modelo DSS, la empresa logró ahorrar 2,621.28 USD anuales en la contratación de servicios externos de transporte [4].

En cuanto a pronóstico de series temporales financieras, las Máquinas de Vectores de Soporte (SVM) han demostrado ser una herramienta prometedora, superando a las redes neuronales tradicionales en precisión y eficiencia. Das y Padhy [4] aplicaron SVM para predecir los precios futuros en el mercado de valores indio, demostrando que las SVM proporcionan resultados precisos, particularmente en el pronóstico de precios de futuros negociados en la Bolsa Nacional de India (NSE). Este hallazgo es consistente con el trabajo de Lindemann [5], quienes realizaron un extenso análisis sobre las redes LSTM para la predicción de series temporales, enfatizando la capacidad de las técnicas de aprendizaje automático para modelar dinámicas de sistemas complejos con precisión. Ambos estudios destacan el potencial de las metodologías avanzadas de aprendizaje automático en el pronóstico financiero, marcando un avance significativo sobre los métodos estadísticos convencionales y ofreciendo una nueva perspectiva en el análisis predictivo de mercados financieros volátiles.

Además, la investigación de Aamer [6] resalta los algoritmos más utilizados en diversas áreas y del sector logístico, proporcionando un marco valioso para futuras investigaciones que busquen abordar la predicción de costos de transporte de carga desde nuevas perspectivas.

Este panorama sugiere una oportunidad significativa para desarrollar modelos predictivos que aborden las necesidades específicas del transporte terrestre en mercados como el colombiano, marcando un paso adelante en la optimización de la cadena de suministro a nivel local y global. Para el sector de interés de este estudio vemos que los modelos más utilizados son Natural Network, Support vector Regression:

Sector	Detalle Sector	Algoritmo	# papers	% de Algoritmo
Agriculture Sector	Agriculture	Support Vector Machine	3	75%
Industry Sector	Energy Demand	Artificial Neural Network	3	25%
	Electricity Demand	Artificial Neural Network	6	30%
	Water Demand	Artificial Neural Network	2	22%
	Natural Gas Demand	Extreme Learning Machine	1	33%
	Cellular Network Demand	Deep Learning	1	100%
	Apparel Industry Demand	Artificial Neural Network	2	100%
	Heat Demand	Neural Network	1	100%
	Electronics Demand	Neural Network	1	100%
	Residential Demand	Neural Network	1	100%
	Coal Demand	Artificial Neural Network	1	100%
Services Sector	Tourism Demand	Neural Network	4	36%
	Transportation Demand	Support Vector Regression	2	22%
		Adaptive-neuro-fuzzy classifier	1	11%
		Back Propagation Network	1	11%
		Deep Learning	1	11%
		Neural Network	3	33%
		Random Forest	1	11%
	Healthcare Service Demand	Neural Network	2	67%
	Banking Service Demand	Neural Network	1	100%
	Service-Oriented Manufacturing Demand	Support Vector Machine	1	100%

Data analytics in the supply chain management: Review of machine learning applications in demand forecasting [6]

Como ejemplo, en un proyecto de predicción del clima basado en datos recolectados cada hora por sensores, propusieron un modelo híbrido RF-LSTM para el pronóstico del clima local [7]. Este modelo aprovecha las ventajas del algoritmo Random Forest, que maneja relaciones complejas y patrones no lineales, junto con la capacidad del modelo LSTM para capturar dependencias secuenciales; los resultados de la investigación demostraron que el modelo híbrido RF-LSTM supera a otros modelos individuales y combinados, como ARIMA, SVR y LSTM, en términos de precisión de predicción. Las métricas utilizadas para evaluar la precisión fueron el error absoluto medio (MAE), la raíz del error cuadrático medio (RMSE) y el coeficiente de determinación (R^2). Los resultados se presentan en la siguiente tabla:

Modelo / Métrica	**MAE**	**RMSE**	**R^2**
ARIMA	0.326	0.544	0.910
SVR	0.430	0.565	0.904
Híbrido ARIMA-SVR	0.320	0.545	0.910
LSTM	0.313	0.513	0.922
Random Forest	0.279	0.380	0.956
Híbrido RF-LSTM	0.269	0.358	0.961

Comparación de métricas de los modelos [7]

Se destaca la efectividad del enfoque híbrido para mejorar la precisión del pronóstico meteorológico a corto plazo y sugiere que la combinación de diferentes técnicas de aprendizaje automático puede ofrecer ventajas significativas en aplicaciones prácticas.

Con la poderosa capacidad de representación de las redes neuronales, los modelos de pronóstico profundo han experimentado un rápido desarrollo. Dos métodos ampliamente utilizados para el pronóstico de series temporales son las redes neuronales recurrentes (RNN) y las redes neuronales convolucionales (CNN). Las RNN modelan puntos de tiempo sucesivos basándose en la suposición de Markov, mientras que las CNN extraen información de variación a lo largo de la dimensión temporal utilizando técnicas como las redes convolucionales temporales (temporal convolutional networks) por sus siglas en inglés (TCN). Sin embargo, debido a la suposición de Markov en las RNN y a la propiedad de recepción local en las TCN, ambos modelos son incapaces de capturar las dependencias a largo plazo en los datos secuenciales; recientemente, el potencial de los modelos Transformer para tareas de pronóstico de series temporales a largo plazo ha ganado atención debido a su capacidad para extraer dependencias a largo plazo mediante el mecanismo de atención[8].

El pronóstico multivariante a largo plazo de series temporales es cada vez más significativo en los procesos de toma de decisiones. Aunque los Transformers han demostrado una eficacia notable en varios dominios, su complejidad plantea desafíos en escenarios de pronóstico a largo plazo. Los esfuerzos para adaptar los modelos basados en Transformers para series temporales con una complejidad reducida incluyen el Informer, que utiliza una estrategia de subsampling probabilístico para mecanismos de atención más eficientes [8], el Autoformer, que emplea autocorrelación y transformaciones rápidas de Fourier para acelerar los cálculos [9], y de manera similar, el FEDformer aplica atención dentro del dominio de frecuencia utilizando componentes seleccionados para mejorar el rendimiento [10]. A pesar de estas innovaciones, los modelos que mezclan canales en series multivariantes a menudo muestran una menor robustez para adaptarse a cambios en la distribución y logran un rendimiento inferior.

En consecuencia, algunos investigadores han adoptado un enfoque independiente de canales, simplificando la arquitectura del modelo y obteniendo resultados robustos también. Sin embargo, ignorar las interacciones entre variables puede limitar los avances adicionales.

Las tendencias recientes se han orientado hacia el aprovechamiento de mecanismos de atención para

capturar correlaciones de canales. Aunque el rendimiento es prometedor, su escalabilidad es limitada en conjuntos de datos grandes. Otra corriente de investigación se centra en modelar dependencias de tiempo y canal a través de estructuras más simples como Multi-Layer Perceptron (MLP). No obstante, generalmente logran un rendimiento subóptimo en comparación con los métodos basados en Transformer de última generación, especialmente cuando el número de canales es grande [11].

5. Marco teórico

El valor flete mercado FP_mean es un fenómeno complejo influenciado por variedad de factores, como políticas gubernamentales, economía global y eventos geopolíticos, es decir, que para este ejercicio estaremos enfocados en la predicción de series de tiempo.

Las series de tiempo pueden definirse como un conjunto de datos recopilados o registrados a intervalos regulares, donde el orden de los datos es crucial por la dependencia temporal entre las observaciones. Se utiliza en economía, meteorología, e ingeniería para predecir cambios futuros basándose en patrones pasados (autorregresivo) o incorporando otras variables externas.

Para más información sobre series de tiempo puede consultarse: *Forecasting: principles and practice*, 3rd edition [12]

5.1. Ecuaciones Básicas de Series de Tiempo

Una forma simple de representar una serie de tiempo es:

$$Y_t = T_t + S_t + E_t$$

- Y_t: valor de la serie en el tiempo t,
- T_t: tendencia en el tiempo t,
- S_t: estacionalidad en el tiempo t,

- E_t: componente de error (o ruido) en el tiempo t.

5.2. Herramientas y Referencias

- SciPy y StatsModels: Implementaciones de ARIMA y otros modelos estadísticos.
- Scikit-learn: RandomForestRegressor y otros algoritmos de ML.
- Keras y TensorFlow: Facilitan la construcción de modelos de deep learning.
- skforecast: Facilita la manipulación de los datos con herramientas de multi-step forecasting compatible con las librerías de sckit-learn.

La principal dificultad en la aplicación de la ciencia de datos a la logística radica en la selección e integración efectiva de variables exógenas relevantes. Sin embargo, este desafío representa también una oportunidad para innovar en el campo, desarrollando modelos más robustos y precisos que respondan mejor a las dinámicas del mercado.

Para la predicción de series de tiempo del valor flete mercado FP_mean, es crucial desarrollar una arquitectura de datos robusta que soporte el eficiente procesamiento y manejo de grandes volúmenes de datos temporales. Esta infraestructura debe incorporar tecnologías escalables que permitan la integración continua y la actualización de modelos mediante prácticas de MLops, facilitando ajustes dinámicos en respuesta a nuevas tendencias y cambios en los datos. La implementación de estas tecnologías asegura que las predicciones sean replicables y consistentes, optimizando así el rendimiento del modelo y reforzando la confiabilidad y transparencia necesarias para la toma de decisiones en logística [13].

5.3. Procesamiento y validación de los datos:

5.3.1. Normalización de la data:

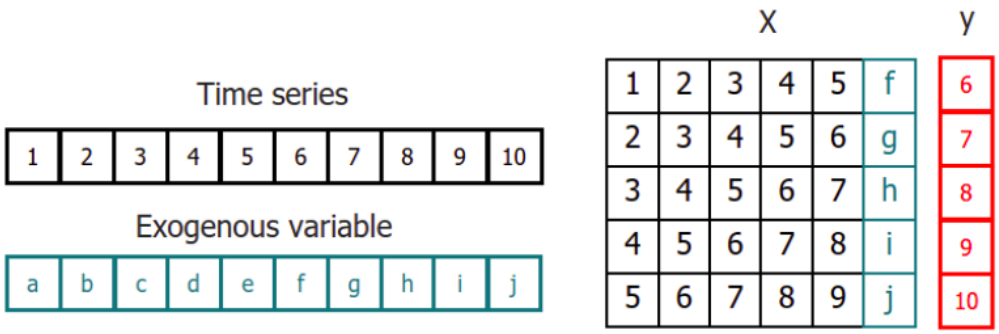
La normalización es una técnica común para calibrar la distribución de los datos de entrada. En el pronóstico de series temporales, las estadísticas locales del historial generalmente se eliminan para estabilizar la predicción del pronosticador base y se restauran estas estadísticas en la predicción del modelo. Siguiendo la práctica común en muchos modelos de última generación, aplicamos la normalización de instancias reversible que centra las series a medias cero, las escala a varianza unitaria y revierte la normalización en las series pronosticadas. Cabe destacar que la normalización puede perder

información estadística y perjudicar el rendimiento [11].

5.3.2. Representación matricial de datos:

Para aplicar modelos de aprendizaje automático a problemas de pronóstico, la serie temporal debe transformarse en una matriz donde cada valor esté asociado con una ventana temporal específica (conocida como retrasos) que lo precede. En el contexto de series temporales, un retraso con respecto a un paso de tiempo t se define como el valor de la serie en pasos de tiempo anteriores. Por ejemplo, el retraso 1 representa el valor en el paso de tiempo t-1, mientras que el retraso m representa el valor en el paso de tiempo t-m [14].

Esta transformación es esencial para que los modelos de aprendizaje automático capturen las dependencias y patrones que existen entre los valores pasados y futuros en una serie temporal. Al usar retrasos como características de entrada, los modelos de aprendizaje automático pueden aprender del pasado y hacer predicciones sobre valores futuros. El número de retrasos utilizados como características de entrada en la matriz es un hiperparámetro importante que debe ajustarse cuidadosamente para obtener el mejor rendimiento del modelo [14].



Transformación de time series con variable exógenas integrando lags [14]

Para el manejo de los datos a la aplicabilidad de los modelos serán utilizados tres conjuntos training, validation y test, esto con el fin de garantizar un ajuste de hiperparámetros en el conjunto de validación y tener una medición del error más acertada, lo cual se reforzará con la biblioteca skforecast que permitirá aplicar:

5.3.3. *Multi-step forecasting:*

Dado que el valor t(n-1) es necesario para predecir t(n), y t(n+1) es desconocido, se aplica un proceso recursivo en el cual cada nueva predicción se basa en la anterior. Este proceso es conocido como pronóstico recursivo o pronóstico multistep recursivo y se puede generar fácilmente con las clases ForecasterAutoreg y ForecasterAutoregCustom.

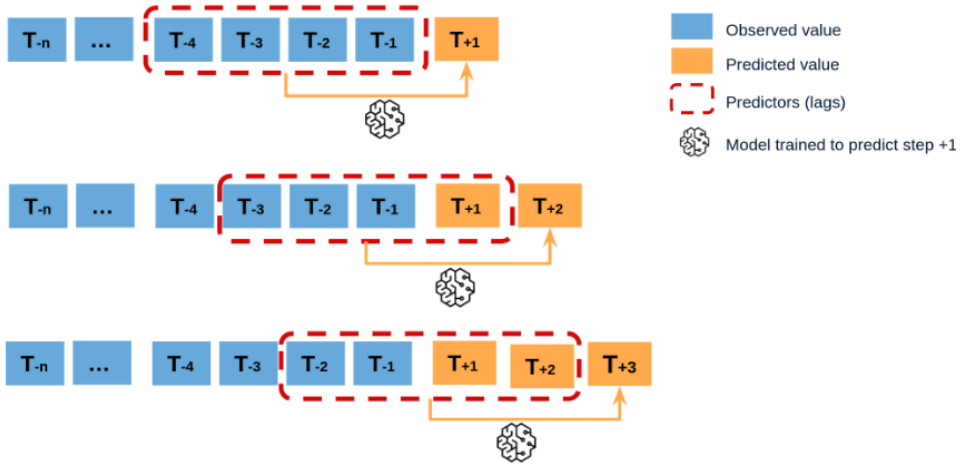
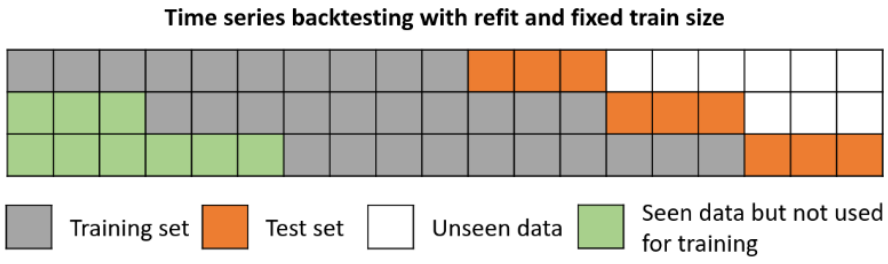


Diagrama de Multi-step forecasting [14]

5.3.4. *Backtesting with refit and fixed training size (rolling origin):*

En este enfoque, el modelo se entrena utilizando una ventana fija de observaciones pasadas, y la prueba se realiza de manera continua, moviendo la ventana de entrenamiento hacia adelante en el tiempo. El tamaño de la ventana de entrenamiento se mantiene constante, lo que permite que el modelo se pruebe en diferentes secciones de los datos. Esta técnica es particularmente útil cuando hay una cantidad limitada de datos disponibles, o cuando los datos no son estacionarios y el rendimiento del modelo puede variar con el tiempo. También se conoce como validación cruzada de series temporales o validación continua [14].



Time series backtestin con refit y tamaño fijo de train [14]

Para obtener una evaluación más robusta sobre el rendimiento del modelo con mejor desempeño se puede realizar backtesting dentro de los conjuntos de entrenamiento y validación para ajustar los hiperparámetros, lo que permite utilizar los datos de manera eficiente. Después de realizar el backtesting y la optimización dentro de estos conjuntos, se puede utilizar el modelo en el conjunto test por separado para evaluar el rendimiento final del modelo y una idea de su desempeño en producción.

5.3.5. Imputación de Datos Faltantes:

En el tratamiento de datos faltantes se implementarán técnicas de imputación que aseguren la integridad y la continuidad de la serie temporal:

Forward Fill: Esta técnica propaga la última observación válida hacia los valores faltantes subsiguientes. Es especialmente útil para datos que no varían significativamente entre periodos consecutivos[15].

Ventana Promedio de 7 Días: Calcular el promedio de los valores de los 7 días anteriores. Esta técnica suaviza las fluctuaciones de corto plazo y es efectiva para datos con patrones estacionales o ciclos semanales.

Imputación por Constante: En casos donde las técnicas anteriores puedan no ser aplicables, se utilizará una constante predeterminada, que será definida basándose en el conocimiento del dominio específico de los datos.

Dependiendo de la necesidad y rendimiento de los modelos podremos evaluar otras técnicas más sofisticadas como las siguientes:

Regresión Dinámica: Esta técnica avanzada utiliza modelos de regresión que incorporan variables adicionales y dinámicas temporales para estimar los valores faltantes. Es útil cuando los datos presentan dependencias complejas que no pueden ser adecuadamente capturadas por métodos más simples[12].

Imputación STL: La técnica STL se usa para descomponer las series temporales en sus componentes estacionales, de tendencia y residuales, seguido de la imputación de los componentes faltantes. Esto

permite una reconstrucción más precisa de la serie, manteniendo las características estacionales y de tendencia intactas[16].

La selección final de las técnicas de imputación dependerá de su efectividad en preservar las características estadísticas cruciales de los datos, así como de su capacidad para integrarse de manera coherente en el marco general de análisis de la serie temporal.

5.4. Evaluación:

En esta fase, se evaluará la efectividad y la robustez de los modelos de filtro colaborativo utilizando técnicas de validación cruzada y pruebas en conjuntos de datos independientes. Analizando los resultados y ajustando los enfoques según sea necesario. Como métrica de selección para este proyecto se utilizará el Root Mean Square Error (RMSE) el cual corresponde a la norma Euclidiana, lo que nos da una idea de cuanto error tiene el sistema en sus predicciones, con un mayor peso en los errores grandes [17]:

$$RMSE(X,h) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(h(x^{(i)}) - y^{(i)})^2}$$

- m es el número de instancias en el conjunto de datos que está midiendo el RMSE.
- $x^{(i)}$ es un vector de todos los valores de características (excluyendo la etiqueta) de la instancia i^{th} en el conjunto de datos, donde y^{i} es su etiqueta (el valor de salida deseado para esa instancia).
- X es una matriz que contiene todos los valores de las características (excluyendo las etiquetas) de todas las instancias del conjunto de datos. Hay una fila por instancia, con la i^{th} fila es igual a la transposición de $x^{(i)}$, anotado $(x^{(i)})^{T}$.
- h es la función de predicción de su sistema, también llamada hipótesis. Cuando a su sistema se le da el vector de característica de una instancia $x^{(i)}$ genera un valor predicho $\Delta y^{(i)} = h\left(x^{(i)}\right)$ para esa instancia.

- $RMSE(X, h)$ es la función de coste medida en el conjunto de ejemplos utilizando su hipótesis h

Como respaldo adicional se utilizará el Mean Absolute Percentaje Error (MAPE) el cual se calcula mediante una comparación término a término del error relativo en la predicción con respecto al valor real de la variable. Por lo tanto, el MAPE es una estadística imparcial para medir la capacidad predictiva de un modelo. Es una medida de la precisión en un valor de serie temporal ajustado en estadística y se ha utilizado para la evaluación de la predicción de series temporales de flujo de ríos [18]. Generalmente expresa la precisión como un porcentaje y se define como:

$$MAPE(X, h) = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{\left(h(x^{(i)}\right) - y^{(i)}}{y^{(i)}}\right|$$

- m es el número de instancias en el conjunto de datos que está midiendo el MAPE.
- $x^{(i)}$ es un vector de todos los valores de características (excluyendo la etiqueta) de la instancia i^{th} en el conjunto de datos, donde y^{i} es su etiqueta (el valor de salida deseado para esa instancia).
- X es una matriz que contiene todos los valores de las características (excluyendo las etiquetas) de todas las instancias del conjunto de datos. Hay una fila por instancia, con la i^{th} fila es igual a la transposición de $x^{(i)}$, anotado $(x^{(i)})^{T}$.
- h es la función de predicción de su sistema, también llamada hipótesis. Cuando a su sistema se le da el vector de característica de una instancia $x^{(i)}$ genera un valor predicho $\Delta y^{(i)} = h\left(x^{(i)}\right)$ para esa instancia.
- $MAPE(X, h)$ es la función de coste medida en el conjunto de ejemplos utilizando su hipótesis h

5.5. Análisis de estacionalidad:

5.5.1. *Autocorrelación (ACF)*

La autocorrelación mide la relación lineal entre valores rezagados de una serie temporal. Cuando los datos tienen una tendencia, las autocorrelaciones para rezagos pequeños tienden a ser grandes y positivas porque las observaciones cercanas en tiempo también están cercanas en valor. Por lo tanto, la función de autocorrelación

(ACF, por sus siglas en inglés) de una serie temporal con tendencia tiende a tener valores positivos que disminuyen lentamente a medida que aumentan los rezagos [12].

Cuando los datos son estacionales, las autocorrelaciones serán mayores para los rezagos estacionales (en múltiplos del período estacional) que para otros rezagos.

$$r_k = \frac{\sum_{t=k+1}^{T}(y_t - \overline{y})(y_{t-k} - \overline{y})}{\sum_{t=1}^{T}(y_t - \overline{y})^2}$$

- Donde T es la longitud de la serie temporal. Los coeficientes de autocorrelación constituyen la función de autocorrelación o ACF.
- Los coeficientes de autocorrelación para los datos de flete de producción (FP_mean) se pueden calcular utilizando la función ACF().

5.5.2. Dickey-Fuller Aumentada (ADF)

La prueba de Dickey-Fuller aumentada (ADF, por sus siglas en inglés) es una prueba estadística utilizada para determinar si una serie temporal es estacionaria. Esta prueba es una extensión de la prueba de Dickey-Fuller original, y está diseñada para manejar series temporales más complejas que pueden mostrar una estructura de dependencia temporal más elaborada [19].

La hipótesis nula de la prueba ADF es que existe una raíz unitaria en la serie temporal, lo que implica que la serie no es estacionaria. La hipótesis alternativa es que la serie temporal no tiene raíz unitaria y por lo tanto es estacionaria.

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \sum_{i=1}^{p} \phi_i \Delta y_{t-i} + \epsilon_t$$

- Δy_t es el cambio en la serie temporal entre dos períodos consecutivos.
- α es el término constante (intercepto).
- βt es el término de tendencia lineal.
- γ es el coeficiente para y_{t-1} (retardo de la serie).
- ϕ_i son los coeficientes para los retardos de los cambios en la serie temporal.
- ϵ_t es el término del error.

5.5.3. Descomposición de serie temporal – Seasonal-Trend decomposition using LOESS (STL)

STL es un método versátil y robusto para descomponer series temporales. Loess, a su vez, es una técnica para estimar relaciones no lineales en los datos. El método STL permite descomponer una serie temporal en tres componentes principales: tendencia, estacionalidad y residuo. Esta descomposición es crucial para analizar y entender la estructura subyacente de la serie temporal, proporcionando información valiosa para los modelos predictivos que se aplicarán en este estudio. Al identificar claramente la tendencia y la estacionalidad, STL facilita la construcción de modelos más precisos y fiables para predecir el valor del flete en el mercado de transporte de mercancías entre Buenaventura y Bogotá [12].

5.6. Modelos estadísticos

5.6.1. Modelos autorregresivos

Un modelo autoregresivo de orden p, denotado como AR(p), modela el valor actual de la serie X_t como la suma de los p valores anteriores de la serie más un término de error que es generalmente asumido como ruido blanco.

5.6.2. SARIMA: Modelo Autoregresivo Integrado de Media Móvil

El ARIMA estacional (SARIMA) es una técnica de ARIMA, donde el componente estacional se puede manejar en datos de series temporales univariadas. Añade tres nuevos hiperparámetros para establecer AR(P), I(D) y MA(Q) para el componente de estacionalidad de una serie temporal. SARIMA permite la ocurrencia de estacionalidad en una serie. El modelo ARIMA estacional combina componentes no estacionales y estacionales en un modelo multiplicativo. La notación se puede definir de la siguiente manera [20]:

ARIMA $(p, d, q)X(P, D, Q)_m$

(P, D, Q) es un componente estacional. Hay cuatro componentes estacionales que no forman parte del

modelo ARIMA que son esenciales configurar:

P: Orden autorregresivo estacional

D: Orden de diferenciación estacional

Q: Orden de promedio móvil estacional

m: Periodicidad de una sola temporada

Un ARIMA $(1,1,1)X(1,1,1)_m$ puede definirse:

$$\left(1-\phi_1 B\right)\left(1-\phi_1 B^m\right)(1-B)(1-B^m)y_{\ t} = (1+\theta_1 B)(1+\Theta_1 B^m)\epsilon_t$$

Componente no estacionario:

- AR: $\phi(B) = 1-\phi_1 B^p$
- MA: $\theta(\mathrm{B}) = 1+\theta_1 B^q$

Componente estacionario

- AR: $\phi(B) = 1-\phi_1 B^{pm}$
- MA: $\Theta(\mathrm{B}) = 1+\Theta_1 B^{qm}$

Termino de error ϵt

5.7. Modelos de aprendizaje de maquina

5.7.1. Random Forest:

El Random Forest es un algoritmo de Machine Learning de uso común registrado por Leo Breiman y Adele Cutler, que combina la salida de múltiples árboles de decisión para alcanzar un solo resultado. Su facilidad de uso y flexibilidad han impulsado su adopción, ya que maneja problemas de clasificación y regresión[21].

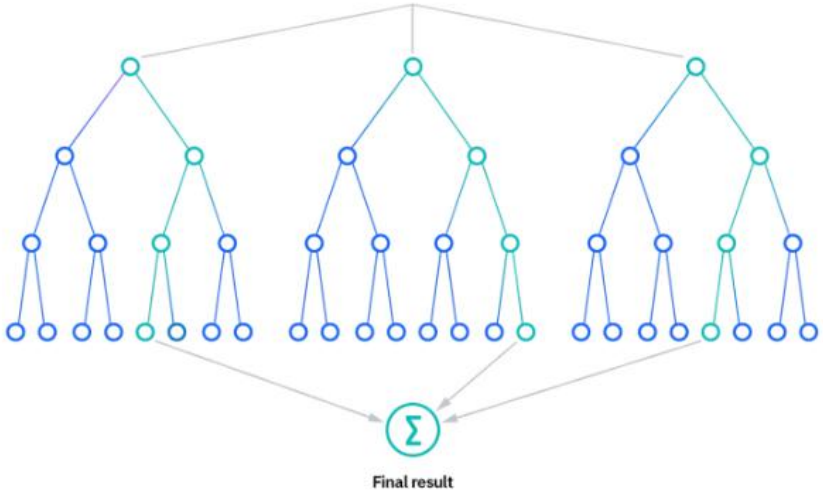
El Random Forest tiene una serie de ventajas sobre otros algoritmos de aprendizaje automático, entre las que se incluyen:

Precisión: El Random Forest es capaz de capturar relaciones complejas entre las variables de entrada y la variable de salida. Esto le permite obtener predicciones más precisas que otros algoritmos.

Robusto: El Random Forest es relativamente robusto a los datos ruidosos y a los valores atípicos. Esto significa que puede producir predicciones precisas incluso cuando los datos de entrenamiento contienen errores o anomalías.

Interpretabilidad: El Random Forest es relativamente interpretable. Esto significa que es posible comprender cómo el modelo llega a sus predicciones.

En general, el Random Forest es un algoritmo de aprendizaje automático versátil y potente que puede utilizarse para una variedad de problemas de predicción [21].



Random Forest [21]

El algoritmo de bosque aleatorio introduce una mayor aleatoriedad al crecer los árboles; en lugar de buscar la mejor característica al dividir un nodo, busca la mejor característica entre un subconjunto aleatorio de características. Por defecto, toma una muestra de √n características (donde n es el número total de características). El algoritmo resulta en una mayor diversidad de árboles, que (de nuevo) intercambia un sesgo más alto por una menor varianza, en general, produciendo un modelo generalmente mejor [21].

5.7.2. Long short term memory LSTM:

Las LSTMs mantienen un estado de celda que actúa como una “memoria” de la red, capaz de transportar información relevante a través de secuencias largas. Este estado se modula con tres puertas distintas:

- La puerta de olvido (forget gate), que decide qué parte de la información previa se mantiene o descarta.
- La puerta de entrada (input gate), que actualiza el estado de la celda con nueva información de la entrada actual.
- La puerta de salida (output gate), que determina qué parte del estado de la celda contribuye a la salida en el momento actual.

Este diseño permite que las LSTMs mitiguen el problema de desvanecimiento o explosión de los gradientes que es común en las RNNs estándar, haciendo posible el aprendizaje en secuencias de datos con dependencias complejas y a largo plazo. Por tanto, las LSTMs son particularmente útiles en tareas de procesamiento de lenguaje natural, predicción de series temporales, reconocimiento de voz y más, donde la secuencialidad y la contextualización a largo plazo son críticas [17].

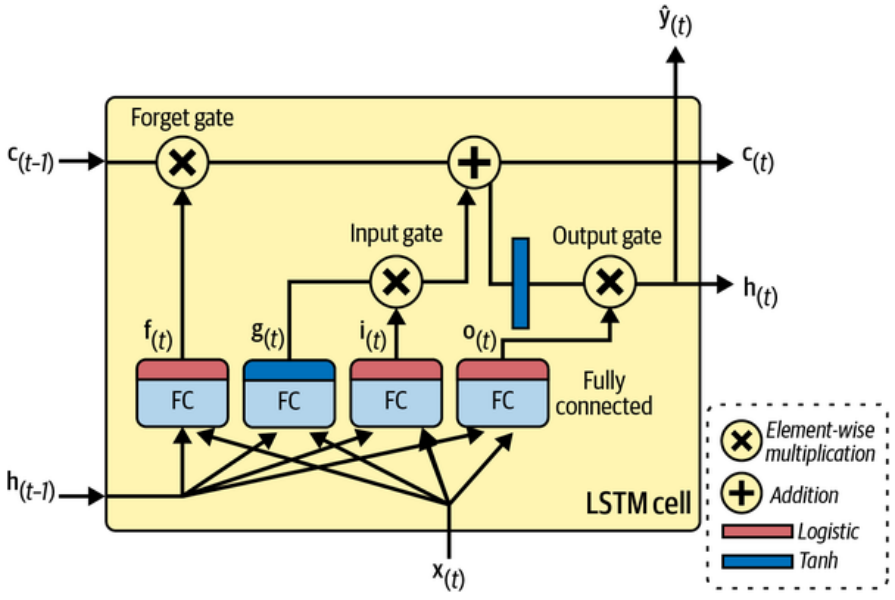


Figure 15-12 an LSTM cell [17]

$$i_{(t)} = \sigma\left(W_{xi}{}^{T} x_{(t)} + W_{hi}{}^{T} h_{(t-1)} + b_i\right)$$
$$f_{(t)} = \sigma\left(W_{xf}{}^{T} x_{(t)} + W_{hf}{}^{T} h_{(t-1)} + b_f\right)$$
$$o_{(t)} = \sigma\left(W_{xo}{}^{T} x_{(t)} + W_{ho}{}^{T} h_{(t-1)} + b_o\right)$$
$$g_{(t)} = \tanh\left(W_{xg}{}^{T} x_{(t)} + W_{hg}{}^{T} h_{(t-1)} + b_g\right)$$
$$c_{(t)} = f_{(t)} \otimes c_{(t-1)} + i_{(t)} \otimes g_{(t)}$$
$$y_{(t)} = h_{(t)} = o_{(t)} \otimes \tanh\left(c_{(t)}\right)$$

Donde:

- W_{xi}, W_{xf}, W_{xo} y W_{xg} son las matrices de pesos de las cuatro capas para su conexión con el vector de entrada $x_{(t)}$.
- W_{hi}, W_{hf}, W_{ho} y W_{hg} son las matrices de pesos de cada una de las capas para su conexión con el estado corto plazo anterior $h_{(t-1)}$.
- b_i, b_f, b_o y b_g son los términos de sesgo de cada una de las cuatro capas. Note que TensorFlow inicializa b_f a un vector lleno de 1s en lugar de 0s. Esto previene el olvido de todo al principio del entrenamiento.

5.7.3. Neural network con variables exógenas:

El modelo implementado en el código tiene como objetivo entrenar y realizar predicciones en series temporales utilizando una red neuronal que combina tanto la variable objetivo como variables exógenas. Este enfoque es particularmente útil para capturar la dinámica interna de la serie temporal y la influencia de factores externos representados por las variables exógenas; el preprocesamiento de datos asegura que la frecuencia temporal sea consistente y normaliza tanto la variable objetivo como las variables exógenas, lo que mejora la estabilidad y el desempeño del modelo durante el entrenamiento.

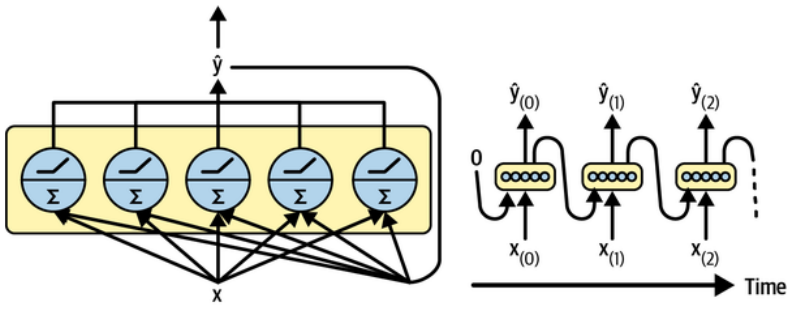
Para procesar la secuencia temporal, el modelo genera segmentos de longitud fija (seq_length) a partir de los datos históricos. Las secuencias incluyen tanto la variable objetivo como las variables exógenas, las cuales se combinan para formar un conjunto tridimensional con dimensiones [batch_size,

seq_length, num_features]. Este formato asegura que el modelo reciba la información necesaria para capturar las relaciones temporales y las influencias exógenas.

El modelo es un perceptrón multicapa (MLP) diseñado para manejar datos secuenciales. Su arquitectura incluye una capa de entrada adaptada al tamaño de las secuencias, seguida de una capa de aplanado para convertirlas en vectores y varias capas densas con activación ReLU para capturar relaciones no lineales. Una capa de salida devuelve un único valor por predicción, y el modelo está optimizado mediante el algoritmo Adam con una función de pérdida de error cuadrático medio (MSE), adecuada para tareas de regresión.

Durante el entrenamiento, el modelo utiliza las secuencias combinadas como entrada y los valores correspondientes de la variable objetivo como etiquetas. Una vez entrenado, el modelo predice valores iterativamente: cada predicción generada se usa como parte de la entrada para calcular la siguiente, utilizando las últimas observaciones del conjunto de entrenamiento como punto de partida. Las variables exógenas del conjunto futuro también se incorporan en las predicciones, tras ser escaladas de forma consistente con los datos de entrenamiento.

Finalmente, las predicciones se desescalan para devolver los resultados en la escala original de la variable objetivo, facilitando su interpretación. Este modelo permite incorporar factores externos relevantes y ofrece flexibilidad para adaptarse a diferentes contextos mediante parámetros personalizables. Sin embargo, su diseño basado en perceptrones multicapa puede ser limitado en series temporales con dependencias a largo plazo, donde modelos como LSTMs podrían ofrecer mejores resultados.



[20]

6. Metodología

De acuerdo con el artículo de Michael [22], es indispensable comprender completamente el problema y el impacto que se busca tener para evitar caer en las trampas de "resolver el problema incorrecto" y de la "sobreingeniería", por lo tanto, se utilizará una metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) [23] para guiar nuestro proceso de investigación y garantizar que se tengan claros los objetivos y el alcance desde el inicio.

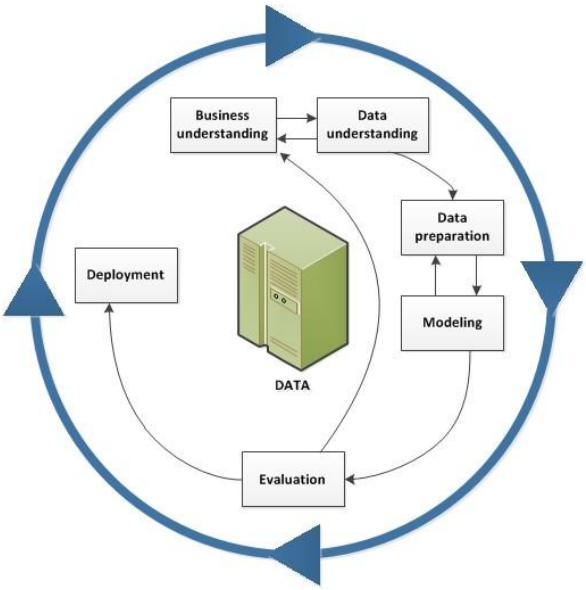


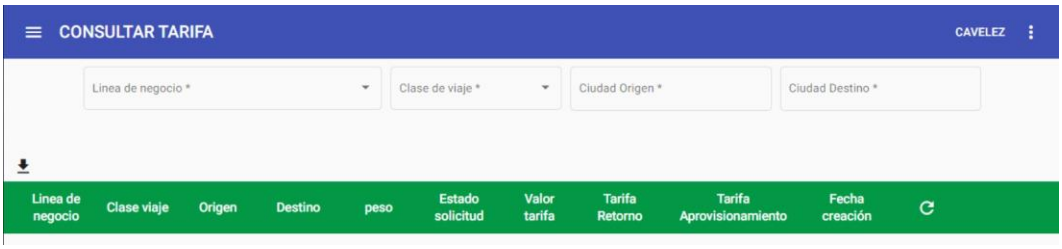
Ilustración 1. Metodología de Crisp-Dm [23]

Esta metodología consta de seis fases interrelacionadas:

Comprensión del Negocio:

Este segmento se centra en comprender los objetivos del proyecto y los requisitos comerciales, incluyendo un análisis exhaustivo de los datos disponibles y las limitaciones existentes. La metodología vigente para la fijación de tarifas se basa en análisis de mercado y la colaboración con conductores de confianza, lo que facilita la valoración precisa del mercado para rutas específicas. Sin embargo, estas tarifas no siempre pueden garantizarse a largo plazo y dependen en gran medida de la experiencia y habilidad del personal encargado. Es importante notar que la valoración del mercado no solo depende de factores como el SICE-TAC o la estructura de costos, sino también de la dinámica de oferta y demanda en los puertos.

Actualmente, la metodología para establecer tarifas se apoya en análisis de mercado colaborando con conductores de confianza, por medio de una app llamada tarifario donde los comerciales registran las tarifas que desean conocer, en promedio se reciben 30 solicitudes por día, con un tiempo de respuesta entre 2 y 24 horas:



6.1. Comprensión de los Datos:

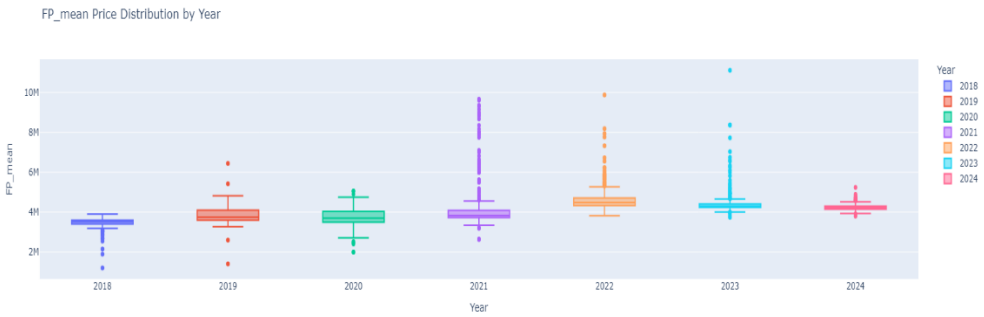
Se explorará el conjunto de datos históricos de Eduardo Botero Soto S.A, Kpler – Marine traffic data histórica de port calls en Buenaventura, registros del RNDC, TRM, días feriados y disponibilidad histórica de vehículos 3S3 en puerto Buenaventura, para comprender su estructura, calidad y relevancia para nuestros objetivos de investigación. Se realizará análisis descriptivos y visualizaciones para identificar patrones iniciales y posibles áreas de interés.

Los datos históricos provienen de las siguientes fuentes, una vez depurados y unificadas las tablas el conjunto de datos tiene 1.237 KB, para ver el diccionario de los datos véase Anexo 1. Diccionario de datos [24].

Nombre	Fuente	Descripción
Base data -Data historica Botero Soto	On-premise MYSQL, 1 DW (Datawarehose) DWCubosOperacionesEBS (FactFWOFUFO, DimBPSolicitante, DimVehiculo) esta extracción comprende del 2018/01/01 hasta el 2024-03-06 con 1.182.893 observaciones con 307 KB.	Esta es la base da datos principal para el proyecto, la cual cotiene los registros historicos de la empresa
SICE-TAC (Sistema de Información de Costos Eficientes para el Transprote Automotor de Carga SICE-TAC)	On-premise MYSQL, 1 DW (Datawarehose) DWSice(FactSice) esta extracción comprende del 2018/01/01 hasta el 2024-03-06 con 1.182.893 observaciones con 307 KB. https://plc.mintransporte.gov.co/Runtime/empresa/ctl/SiceTAC/mid/417	Esta base de datos se extrae unida con la Base data historica de Botero Soto, ya que tiene un vinculo directo con cada despacho para garantizar el valor minimo de flete impuesto por el SICE-TAC
RNDC (Registro Nacional Despachos de Carga por Carretera)	On-premise MYSQL, 1 DW (Datawarehose) DWCubosRNDC(FactRNDC) esta extracción comprende del 2018/01/01 hasta el 2024-03-06 con 212781 observaciones con 45.5 MB. https://rndc.mintransporte.gov.co/MenuPrincipal/tabid/204/language/es-MX/Default.aspx?returnurl=%2f	En esta base de datos quedan registrados todos los movimientos de carga del país, sin embargo, no contiene la información discriminada por día, solo se encuentra disponible a nivel año/mes.
TRM (Tasa representativa del mercado)	TRM (Tasa representativa del mercado) es extraída de la API de datos abiertos www.datos.gov.co desde el 2015-09-29 hasta el 2024-03-06 con 117 KB	Al realizar la conexión a esta API se extrae la tasa de cambio de COP a USD
Marine traffic	Portcall es una API de la empresa Marine traffic especializada en el seguimiento de buques, de aquí se extrajo la data histórica del arrivo de buques a Buenaventura como variable exógena con 29.425 Kb desde el 2016-01-01 hasta 2024-03-01 https://servicedocs.marinetraffic.com/tag/Port-Events#operation/portcalls_	Al realizar la conexión a esta API se extrae la data historica de arrivo de buques al puerto de Buenaventura, Colombia.
Holidays	Portcall es una API gratuita que proporciona los días festivos de un país https://date.nager.at/Api	Al realizar la conexión a esta API se extrae la data historica de días festivos en Colombia.

6.2. Preparación de los Datos:

En esta fase, limpiaremos y preprocesaremos los datos para eliminar valores atípicos, datos faltantes o redundantes que puedan afectar la precisión del análisis. También realizaremos transformaciones y selección de características según sea necesario para alimentar nuestros algoritmos de detección de anomalías. Dado que estamos trabajando con series de tiempos, se realizarán análisis de estacionariedad y estacionalidad para la aplicación de los modelos de aprendizaje estadístico y por consiguiente profundizar en modelos de machine learning, a continuación, la distribución de la variable FP_mean, el modelo entidad relación, y la correlación de la variable FP_mean con las demás variables (umbral del 0.5):



Ruta en el proyecto: notebooks/EDA/EDA.ipynb

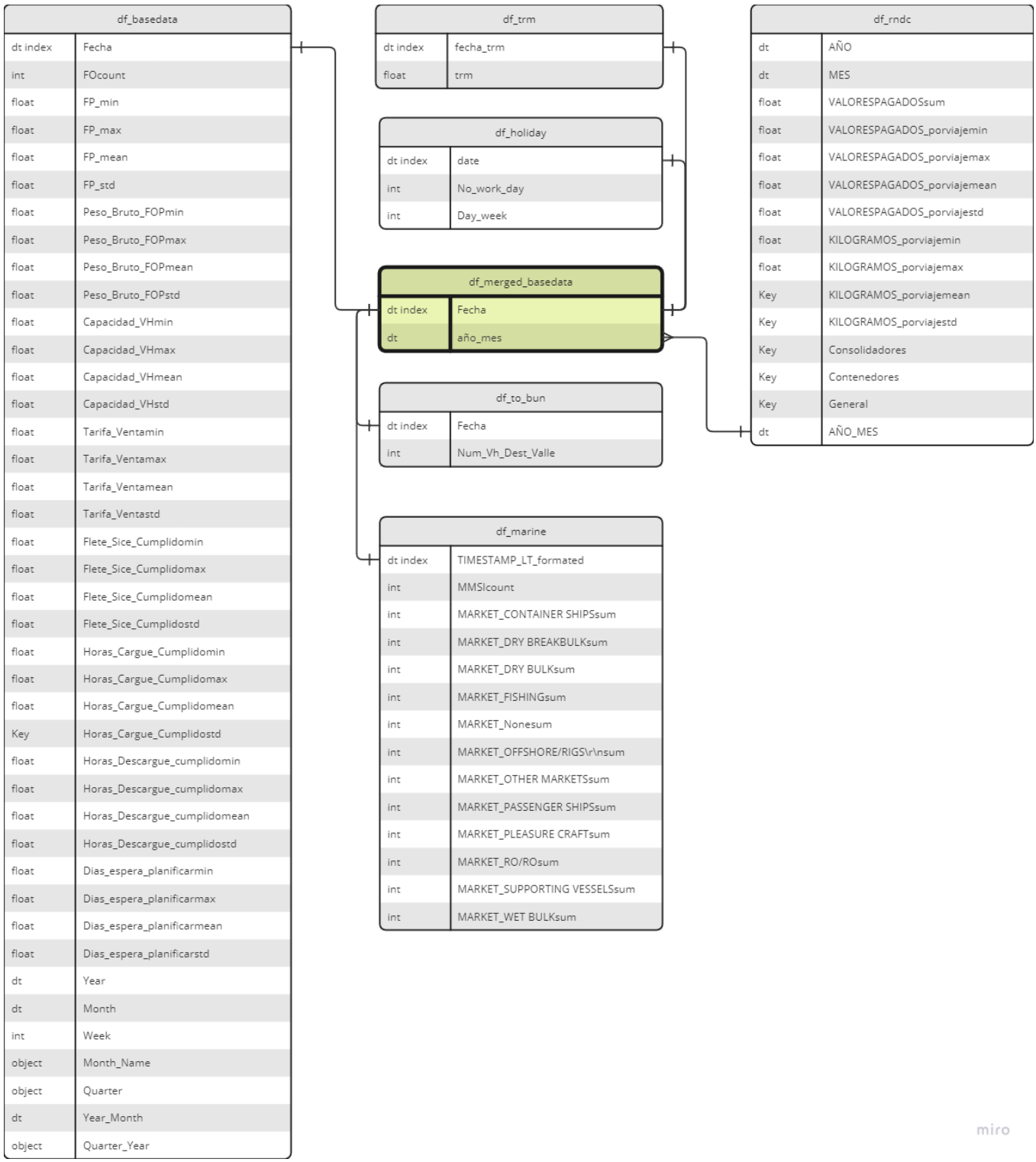
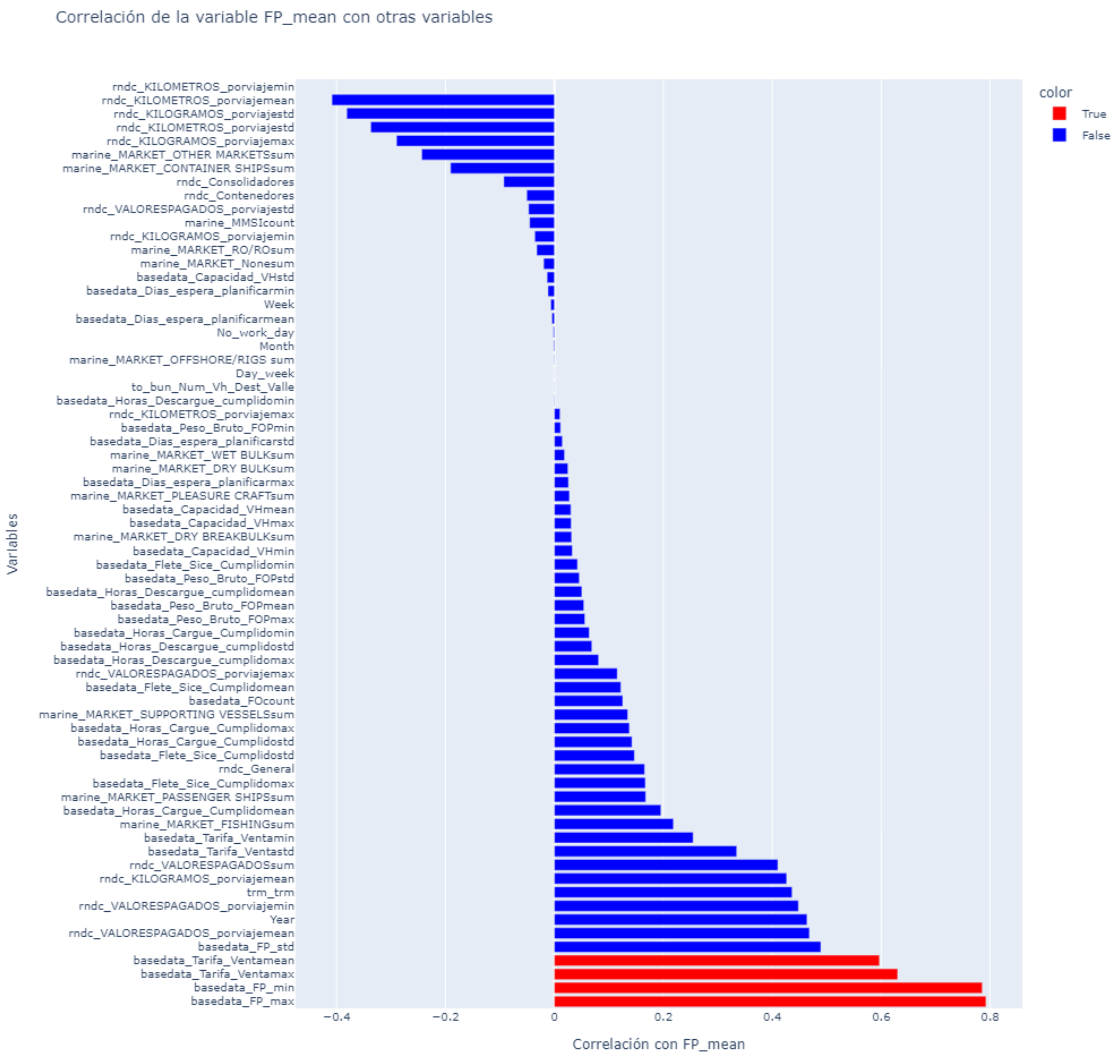


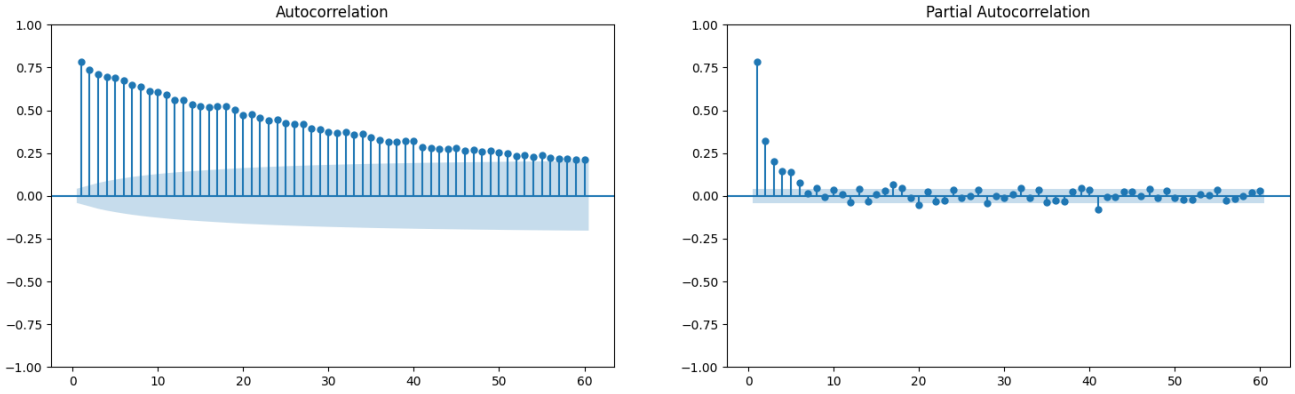
Diagrama entidad relación



Ruta en el proyecto: notebooks/EDA/EDA.ipynb

6.2.1. Análisis de estacionalidad:

6.2.1.1. Autocorrelación (ACF) y autocorrelación parcial (PACF)



El gráfico ACF muestra una autocorrelación inicialmente alta que decrece con más rezagos, indicativo de influencia significativa de valores previos cercanos en la serie temporal. La presencia de picos notables en rezagos múltiples de siete sugiere patrones estacionales, reflejando tendencia y estacionalidad en la serie.

El gráfico PACF exhibe que los primeros seis lags poseen una correlación parcial significativa con la variable FP_mean, lo que implica que estos rezagos aportan información única para la predicción de la serie. La correlación parcial se atenúa considerablemente después del sexto rezago, lo que sugiere que los valores más allá no contribuyen de manera sustancial a la explicación de la variable en cuestión.

6.2.1.2. Dickey Fuller Aumentada (ADF)

ADF Statistic: -4.703368
p-value: 0.000083
Critical Values:
1%: -3.433
5%: -2.863
10%: -2.567

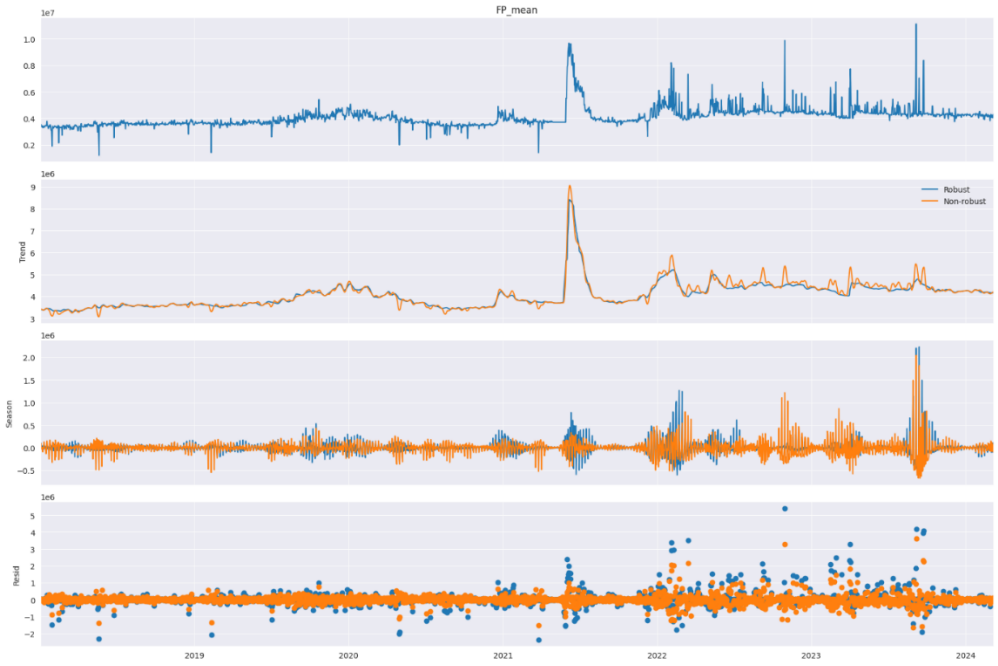
El valor del estadístico ADF es -4.703368, lo cual es menor que el valor crítico para el 1% (-3.433). Significa que se puede rechazar la hipótesis nula con un nivel de confianza superior al 99%, sugiriendo que la serie de tiempo es estacionaria.

El valor p de 0.000083 confirma esta conclusión, ya que es mucho menor que 0.01 (o incluso 0.001), lo que indica que es extremadamente improbable que la serie de tiempo contenga una raíz unitaria bajo la suposición de la hipótesis nula.

En resumen, los resultados de la prueba ADF sugieren fuertemente que la serie de tiempo es estacionaria y que no es necesario diferenciarla para alcanzar la estacionariedad antes de ajustar un modelo ARIMA.

6.2.1.3. Descomposición de serie temporal - Seasonal-Trend decomposition using LOESS (STL)

Descomposición con factor de robustez y sin robustez:



La estacionalidad se está realizando con un periodo semanal

Se encuentran necesario realizar la normalización y la imputación de los datos, puesto que la variable objetivo FP_mean contiene 319 valores nulos del 2018 al 2024 correspondientes a festivos y dominicales sin operación, los cuales deben ser imputados con la data histórica teniendo cuidado de no involucrar valores futuros.

6.3. Modelado y evaluación:

Se entrenaron un total de nueve modelos de series de tiempo (ARIMA, SARIMA, Random Forest Regressor y LSTM) y uno de clasificación (Random Forest). Algunos de los modelos de series de tiempo fueron híbridos, utilizando las predicciones generadas por SARIMA como variables exógenas. Para evaluar el desempeño de los modelos, se emplearon las métricas MAPE (Error Absoluto Medio Porcentual) y RMSE (Raíz del Error Cuadrático Medio). El MAPE brinda una percepción del error en términos porcentuales, mientras que el RMSE mide el promedio de los errores absolutos, permitiendo una evaluación integral de la precisión del modelo sobre la variable FP_mean.

La evaluación se realizó en 30 iteraciones, cada una utilizando ventanas móviles de 30 días sin refit de los datos. Es decir, los últimos 60 días de la serie de tiempo se emplearon para generar y testear las predicciones de 30 días a futuro.

6.3.1. Resultados generales:

En términos generales, el modelo Random Forest con variables exógenas (random_forest_exogen) mostró el mejor desempeño. Este modelo logró capturar la tendencia de la serie de tiempo y presentó una menor desviación estándar en los errores. Las variables exógenas utilizadas para este modelo incluyeron:

- No_work_day: Indicador de días no laborales.
- Day_week: Día de la semana.
- Predicciones de SARIMA:
- sarima_prediction_FP_mean
- sarima_prediction_FP_max
- sarima_prediction_FP_min
- sarima_prediction_FP_std

Sin embargo, se observó que, a medida que avanzan las predicciones, el error se acumula debido a la ausencia de actualización de los valores reales. En otras palabras, el modelo predice 30 días consecutivos a futuro partiendo únicamente del último valor observado.

	Modelo	RMSE	MAPE	RMSE_std	MAPE_std
5	random_forest_exogen	211,395.42	3.20%	98,867.46	1.23%
3	lstm_with_arima_predictions	214,404.03	3.38%	96,996.98	1.26%
0	sarima_predictions	214,817.02	3.34%	115,494.34	1.51%
7	naive_predictions	238,290.74	3.97%	111,995.71	1.81%
4	random_forest	245,978.70	4.42%	56,338.22	0.64%
2	lstm_exogen_predictions	269,270.89	4.63%	139,337.49	2.16%
1	lstm_prediction	320,141.06	5.68%	136,771.86	1.98%
6	neural_network_exogen	330,672.95	5.95%	101,793.67	2.26%

Ruta en el proyecto tests 4.0/main_notebook.ipynb

6.3.2. Resultados por iteración:

En las iteraciones individuales, el modelo reflejó un comportamiento consistente en la captura de tendencias, con variaciones mínimas en el error entre iteraciones.

iteration	best_model	best_rmse	best_mape	rmse_std	mape_std
1	naive_predictions	109,964.92	2.06%	54,471.09	1.17%
2	sarima_predictions	118,414.23	2.12%	94,372.55	2.07%
3	naive_predictions	119,715.51	2.24%	137,152.42	3.04%
4	sarima_predictions	128,101.48	2.61%	128,763.27	2.90%
5	naive_predictions	102,919.34	2.09%	119,860.31	2.66%
6	sarima_predictions	100,392.23	2.04%	76,954.75	1.70%
7	sarima_predictions	100,091.70	2.03%	121,179.18	2.75%
8	sarima_predictions	102,355.16	2.10%	33,993.54	0.80%
9	sarima_predictions	98,488.15	1.94%	78,686.88	1.59%
10	random_forest_exogen	102,301.45	1.85%	33,567.54	0.83%
11	naive_predictions	95,048.01	1.80%	40,284.63	0.97%
12	sarima_predictions	92,700.76	1.75%	41,075.67	0.98%
13	sarima_predictions	96,369.14	1.76%	40,402.99	0.93%
14	lstm_with_arima_predictions	107,546.76	2.07%	45,158.59	1.04%
15	random_forest_exogen	155,925.53	2.34%	37,808.70	0.99%
16	random_forest_exogen	171,681.50	2.44%	35,036.06	0.88%
17	random_forest_exogen	174,230.97	2.50%	35,957.68	0.85%
18	random_forest_exogen	258,139.62	3.12%	38,073.97	0.82%
19	random_forest	261,346.94	4.38%	43,869.79	0.97%
20	random_forest	260,630.37	4.34%	51,019.44	1.15%
21	random_forest	259,135.65	4.08%	48,646.09	0.94%
22	random_forest	278,460.15	4.20%	53,372.19	1.01%
23	random_forest	281,617.52	4.52%	60,102.39	1.21%
24	random_forest	323,551.67	4.45%	67,739.16	1.63%
25	random_forest	325,233.69	4.57%	61,722.00	1.43%
26	random_forest	326,255.43	4.58%	64,485.95	1.51%
27	random_forest	288,336.84	4.52%	76,237.17	1.60%
28	naive_predictions	343,226.20	5.11%	61,085.41	1.37%
29	lstm_with_arima_predictions	344,252.84	5.24%	58,707.19	1.24%
30	random_forest	294,800.23	4.69%	66,941.72	1.33%

Ruta en el proyecto tests 4.0/main_notebook.ipynb

6.3.3. Plot de valores reales con predicciones iteración 10:

El gráfico correspondiente a la iteración 10 muestra cómo el modelo random_forest_exogen sigue de cerca la tendencia de los datos reales. Sin embargo, los errores se amplifican con el tiempo debido a la falta de realimentación con datos reales en las predicciones extendidas.

Ruta en el proyecto: tests 4.0/main_notebook.ipynb

6.3.4. Clasificación:

Se desarrolló un modelo de clasificación con el objetivo de identificar las variables clave y predecir categorías de flete: bajo (-1), normal (0) o alto (1). Las categorías se definieron utilizando la siguiente metodología:

Lower bound: Límite inferior, calculado como el primer tercil (quantile(0.33)) menos 1.5 veces la distancia intercuartílica (IQR).

Upper bound: Límite superior, calculado como el tercer tercil (quantile(0.67)) más 1.5 veces la distancia intercuartílica (IQR).

Los intervalos de clasificación son:

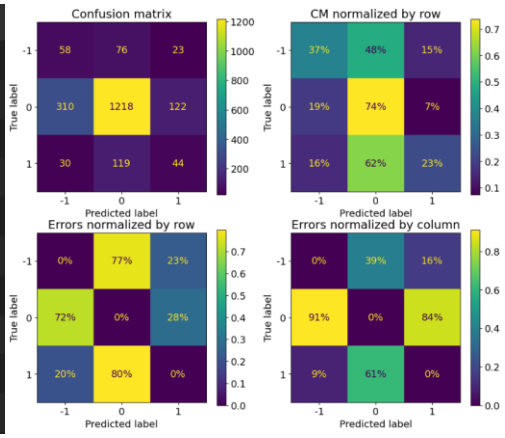
De -∞ a lower_bound: Categoría bajo (-1).

De lower_bound a upper_bound: Categoría normal (0).

De upper_bound a ∞: Categoría alto (1).

El modelo clasificatorio permite explorar el problema desde una perspectiva diferente. Como idea a futuro, se plantea la posibilidad de abordar el problema como una tarea de clasificación con un conjunto de variables predictoras, donde el modelo pueda generar directamente el valor del flete esperado. Esto está alineado con el planteamiento inicial del problema y podría abrir nuevas oportunidades para mejorar la predicción y gestión de los datos.

	Feature	Importance
0	num_a__Peso_Bruto_FOP	0.09
1	cat_a__Cliente	0.06
2	num_a__Distancia_Total_FO	0.05
3	cat_a__ZONATRANSPORTE_AGRUPADA	0.03
4	num_a__Tiempo_Planificación	0.02
5	num_a__Tiempo_Cargue_FO_EM	0.02
6	num_b__FO	0.02
7	num_b__Tiempo_Para_Planificacion	0.02
8	num_b__FU	0.02
9	num_b__FWO	0.02



Ruta en el Proyecto: notebooks/EDA/EDA_rf_new.ipynb

La intención principal del modelo de clasificación desarrollado fue no solo predecir las categorías de fletes (bajo, normal y alto), sino también identificar las variables más relevantes que influyen en la clasificación. La importancia de las características, como se muestra en la tabla, revela qué factores tienen un mayor impacto en la toma de decisiones del modelo.

En este caso, las variables con mayor importancia fueron num_a__Peso_Bruto_FOP (0.09), cat_a__Cliente (0.06) y num_a__Distancia_Total_FO (0.05), lo que indica que el peso bruto, el cliente y la distancia total son determinantes clave para clasificar los fletes. Estas variables tienen un efecto significativo en la categorización de los fletes como bajo, normal o alto.

6.4. Repositorio*:*

https://github.com/Botero-Soto/prediction-project

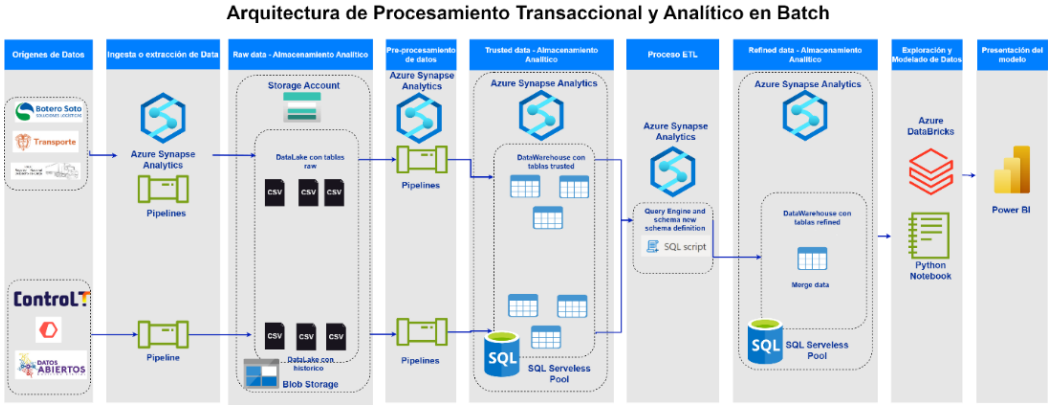
6.5. Despliegue*:*

Este modelo no fue desplegado por los siguientes motivos:

- La empresa Eduardo Botero Soto S.A. no cuenta actualmente con infraestructura en la nube, por lo que se están realizando ajustes para su implementación en un servidor local on-premises.
- Al no pertenecer ya a la organización, la implementación del modelo dependerá de futuros acuerdos entre las partes interesadas.

6.5.1. Arquitectura propuesta para un futuro despliegue

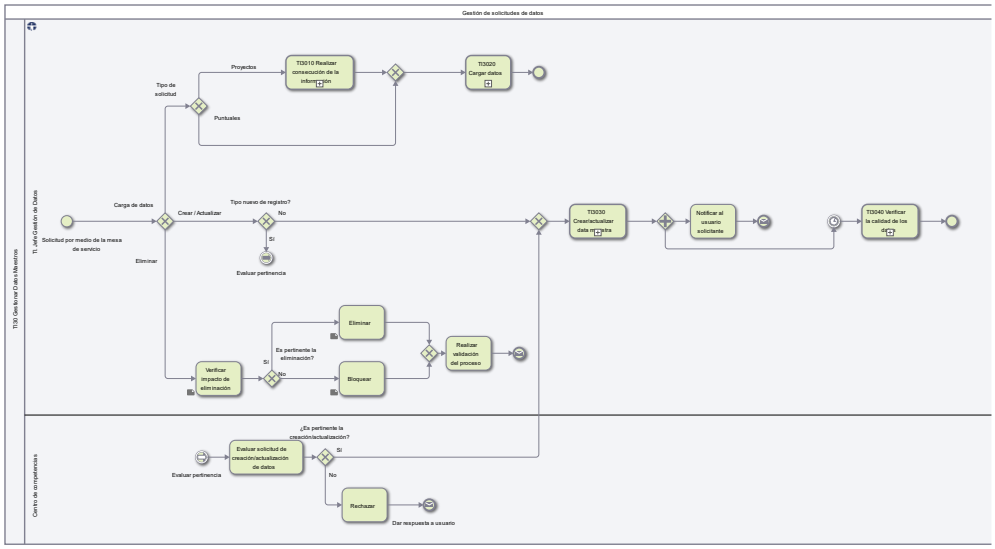
Para este proyecto por sus características se propone la siguiente arquitectura en Batch. Esta arquitectura fue diseñada e implementada en la nube de Azure, dado que actualmente no hay un ecosistema definido o implementado en Eduardo Botero Soto S.A. El flujo de datos detallado está representado por el siguiente diagrama:

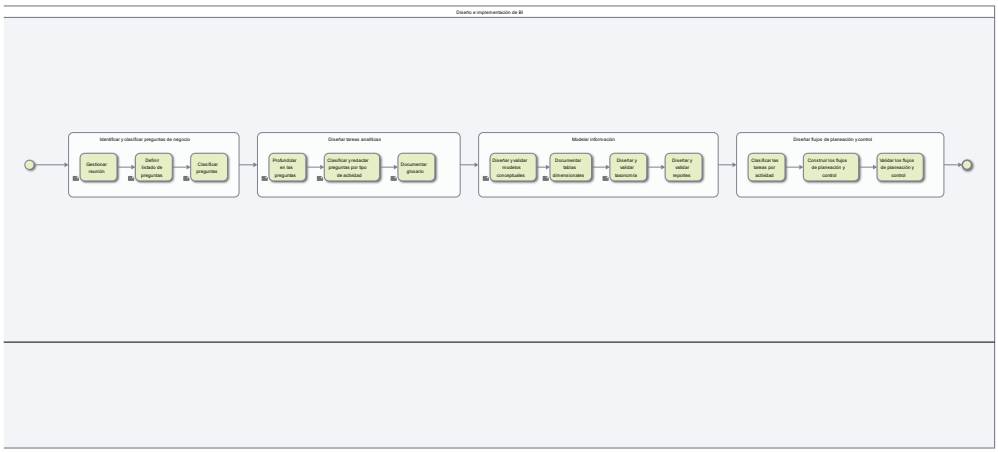


Arquitectura propuesta para un futuro despliegue

7. Plan de Gestión de Datos

Política de gestión de datos: ver Anexo 2 [25].





Para el formato y contenido de los datos y metadatos se seguirán las normativas establecidas en el Anexo 3 [26]. Por otro lado, las políticas referentes al acceso e intercambio de información, que incluyen medidas para garantizar la privacidad, confidencialidad, seguridad y protección de la propiedad intelectual, así como los derechos, requisitos, reutilización, distribución y generación de productos o derivados de investigación, se describen en el Anexo 4 [27].

8. Aspectos éticos

El proyecto de maestría focalizado en incrementar la eficiencia logística a través de la negociación de fletes se basará en el uso de datos específicos para crear un modelo avanzado. Este modelo buscará beneficiar a Eduardo Botero Soto S.A y, consecuentemente, a sus clientes, brindando ventajas competitivas dentro del sector. Botero Soto será el único encargado de manejar estos datos, comprometiéndose a un manejo ético y responsable de los mismos. La privacidad y seguridad de la información se resguardarán siguiendo las directrices del Anexo 4 [27]el cual especifica los métodos para la obtención del consentimiento por parte de las fuentes de datos.

Cualquier resultado obtenido en este proyecto estará protegido bajo un acuerdo de confidencialidad y su uso se limitará exclusivamente a fines de apoyo y evaluación del desarrollo del proyecto. La divulgación de los resultados requerirá de una autorización explícita y por escrito de parte de Botero Soto, garantizando de esta manera la seguridad y la confidencialidad de la información procesada.

9. Conclusiones

- Impacto de las Variables Exógenas, la incorporación de variables exógenas, aunque mejora el contexto predictivo, introduce una dependencia significativa que incrementa el margen de error acumulado en las predicciones, especialmente en horizontes más largos. Esto resalta la necesidad de considerar técnicas para mitigar este efecto en futuros modelos.
- Desempeño de los Modelos, Entre los enfoques evaluados, el modelo Random Forest con variables exógenas demostró ser el más preciso, con un RMSE de 211,395.42 y un MAPE de 3.20%.
- Comparación con Métodos Naïve, los modelos desarrollados mostraron un desempeño significativamente superior al enfoque naïve, incluso en una serie sin estacionalidad clara ni patrones distintivos, subrayando el valor añadido de las técnicas avanzadas implementadas.
- Limitaciones en la Captura de Picos Abruptos, a pesar de que los modelos lograron capturar tendencias generales en la variable FP_mean, enfrentaron limitaciones para predecir picos abruptos en los valores de flete, un aspecto crítico para la planificación logística.
- Evaluación con Ventanas Móviles, la estrategia de evaluación mediante ventanas móviles de 30 días permitió obtener predicciones robustas en el corto plazo. Sin embargo, la acumulación de errores debido a la falta de realimentación con datos reales en predicciones extendidas destaca como un área de mejora para desarrollos futuros.
- Reformulación como Problema de Clasificación, abordar el problema desde la perspectiva de clasificación resultó ser una aproximación prometedora. Este enfoque permitió identificar variables críticas, como el peso bruto, el cliente y la distancia total, y puede ampliarse para fortalecer la capacidad de decisión en la gestión de costos logísticos.
- Impacto Estratégico del Modelo, el modelo desarrollado constituye una herramienta estratégica para optimizar la toma de decisiones en el transporte terrestre, especialmente en la ruta Buenaventura-Bogotá. Su implementación fortalece la competitividad de las empresas logísticas al mejorar la planificación operativa y el cumplimiento normativo.

10. Referencias

[1] Camara de comercio de cali, “Informe competitivo #135. ,” Camara de comercio de cali el 02-06-2023.

[2] mintransporte, *Resolución 20213040034405*. https://mintransporte.gov.co/loader.php?lServicio=Tools2&lTipo=descargas&lFuncion=descargar&idFile=26949, 2024, pp. 0–23.

[3] DANE, “Índice de Costos del Transporte de Carga por Carretera (ICTC).” Accessed: May 20, 2024. [Online]. Available: Índice de Costos del Transporte de Carga por Carretera (ICTC)

[4] S. PrasadDas and S. Padhy, “Support Vector Machines for Prediction of Futures Prices in Indian Stock Market,” *Int J Comput Appl*, vol. 41, no. 3, pp. 22–26, Mar. 2012, doi: 10.5120/5522-7555.

[5] B. Lindemann, T. Müller, H. Vietz, N. Jazdi, and M. Weyrich, “A survey on long short-term memory networks for time series prediction,” *Procedia CIRP*, vol. 99, pp. 650–655, 2021, doi: 10.1016/j.procir.2021.03.088.

[6] A. Aamer, L. P. Eka Yani, and I. M. Alan Priyatna, “Data Analytics in the Supply Chain Management: Review of Machine Learning Applications in Demand Forecasting,” *Operations and Supply Chain Management: An International Journal*, pp. 1–13, Dec. 2020, doi: 10.31387/oscm0440281.

[7] T. K. Tran *et al.*, “Exploring Hybrid Models For Short-Term Local Weather Forecasting in IoT Environment,” *MENDEL*, vol. 29, no. 2, pp. 295–306, Dec. 2023, doi: 10.13164/mendel.2023.2.295.

[8] H. Zhou *et al.*, “Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 12, pp. 11106–11115, May 2021, doi: 10.1609/aaai.v35i12.17325.

[9] H. Wu, J. Xu, J. Wang, and M. Long, *Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting*. 2021.

[10] T. Zhou, Z. Ma, Q. Wen, X. Wang, L. Sun, and R. Jin, *FEDformer: Frequency Enhanced Decomposed Transformer for Long-term Series Forecasting*. 2022.

[11] Lu Han, Xu-Yang Chen, Han-Jia Ye, and De-Chuan Zhan, “SOFTS: Efficient Multivariate Time Series Forecasting with Series-Core Fusion,” *https://arxiv.org/abs/2404.14197*, 2024.

[12] Rob J Hyndman and George Athanasopoulos, *Forecasting: principles and practice, 3rd edition*, 3rd ed. edición., vol. 3. Australia: Otexts; (31 Mayo 2021), 2021. Accessed: May 18, 2024. [Online]. Available: https://otexts.com/fpp3/index.html

[13] Yaron Haviv and Noah Gift, *Implementing MLOps in the enterprise: A production-first approach*, 1st ed., vol. 1. O’Reilly Media, 2024.

[14] Skforecast Docs, “Introduction to forecasting.” Accessed: May 18, 2024. [Online]. Available: https://skforecast.org/0.12.0/introduction-forecasting/introduction-forecasting

[15] J. Manu, *MODERN TIME SERIES FORECASTING WITH PYTHON explore industry-ready time series forecasting using modern machine learning and deep learning*. PACKT PUBLISHING LIMITED, 2022.

[16] Ahmed Abulkhair, “Data Imputation Demystified | Time Series Data,” medium. Accessed: May 18, 2024. [Online]. Available: https://medium.com/@aaabulkhair/data-imputation-demystified-time-series-data-69bc9c798cb7

[17] Aurélien Géron, *Hands-on machine learning with scikit-learn, keras, and TensorFlow 3e: Concepts, tools, and techniques to build intelligent systems (3.a ed.).* , 3.a ed. O’Reilly Media, 2022.

[18] T. S. HU, K. C. LAM, and S. T. NG, “River flow time series prediction with a range-dependent neural network,” *Hydrological Sciences Journal*, vol. 46, no. 5, pp. 729–745, Oct. 2001, doi: 10.1080/02626660109492867.

[19] D. A. Dickey and W. A. Fuller, “Distribution of the Estimators for Autoregressive Time Series

With a Unit Root," *J Am Stat Assoc*, vol. 74, no. 366, p. 427, Jun. 1979, doi: 10.2307/2286348.
[20] B V Vishwas and Ashish Patel, "Hands-on Time Series Analysis with Python_ From Basics to Bleeding Edge Techniques-Apress," *o'really*, 2020.
[21] IBM, "¿Qué es el random forest?," IBM. Accessed: May 18, 2024. [Online]. Available: https://www.ibm.com/mx-es/topics/random-forest
[22] E. L. D. N. J. L. Michael Parzen, "An Art & A Science: How to Apply Design Thinking to Data Science Challenges," *Harvard Business School* , pp. 0–14, Apr. 2023.
[23] IBM, "Conceptos básicos de ayuda de CRISP-DM," IBM. Accessed: May 18, 2024. [Online]. Available: https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview
[24] Eduardo Botero Soto S.A, "Diccionario de datos," Itagüí, Colombia, Anexo 1, 2024.
[25] Eduardo Botero Soto S.A, "Gestionar Datos Maestros," Itagüí, Colombia, Anexo 2, 2024.
[26] Eduardo Botero Soto S.A, "Politica Estandarizacion DM," *Eduardo Botero Soto S.A*, vol. Anexo 3, 2019.
[27] Eduardo Botero Soto S.A, "Manual De Politicas De Tratamiento De Bases De Datos Personales," Itagüí, Colombia, Anexo 4, 2019.