Modelo de Machine Learning para la caracterización de carga y estimación de costos de transporte, a partir de histórico de manifiestos de carga en Colombia entre 2015 y 2023

David Alexander Urrego Higuita

Asesora

Danitza Maria Cortes Perez

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Especialización en Ciencia de Datos y Analítica

2024

Dedicatoria

Esta es una dedicatoria para mi esposa Luisa Fernanda Álvarez Monsalve y mi hija Maria Clara Urrego Álvarez, por su apoyo y amor incondicional, que permitieron afrontar este maravilloso reto.

Agradecimientos

Un agradecimiento especial a la profesora Danitza Maria Cortes Perez por su acompañamiento y ayuda para finalizar este texto y a la jurado Sandra Patricia Barreto por sus comentarios que permiten dejar un texto de consulta para toda la comunidad académica.

A la docente Alba Maribel Sanchez Galvez, de la Benemérita Universidad de Puebla por su acompañamiento en la pasantía que me permitió construir un gran modelo para este trabajo.

Finalmente, a todos los docentes y directivos que hicieron posible de esta especialización, de la cual me siento extremadamente orgulloso y alegre de haber podido participar.

Resumen

El transporte de carga en Colombia presenta grandes desafíos debido a la variabilidad en la demanda y los costos asociados en los que se incurre, este trabajo propone el diseño de un modelo de Machine Learning que permita la caracterización de la carga y estimar los costos de transporte basándose en el histórico de manifiestos de carga recopilados entre 2015 y 2023 en Colombia a través de la plataforma RNDC del Ministerio de Transporte, se busca mejorar la precisión en la estimación de costos y optimizar la logística del transporte de carga.

Para alcanzar este objetivo, se establecerá un análisis exploratorio de los datos que permita una caracterización del transporte de carga, se implementará diversas técnicas de Machine Learning, incluyendo regresión lineal, árboles de decisión y redes neuronales, entre otros. Los datos fueron preprocesados para manejar la variabilidad y las inconsistencias, y se aplicaron métodos de validación cruzada para asegurar la robustez del modelo. Se espera que los resultados muestren un modelo que se ajuste que permita su implementación en las plataformas que faciliten la visualización de los datos y permitir interactuar con los datos, se espera que el modelo seleccionado ofrezca la mayor precisión en la estimación de costos, con una reducción significativa en el error promedio comparado con los métodos probados.

Estos hallazgos demuestran el potencial del Machine Learning para transformar la gestión del transporte de carga, proporcionando herramientas más precisas para la toma de decisiones y la planificación logística. La implementación de este modelo puede resultar en ahorros considerables y en una mejora general en la eficiencia del transporte de carga en Colombia, además de permitir dejar unas bases sólidas para mejorar el modelo para predicción de costos, incluyendo nuevos datos, como características de la región, tamaño económico, vocación productiva, variación de precios de fletes y combustible, esto subrayaría la importancia

de considerar una gama más amplia de variables para capturar con precisión la dinámica del costo de transporte. Finalmente, en un futuro desarrollar otro tipo de estudios como clasificación de tipo de carga, segmentación de rutas de transporte, análisis de tendencias temporales y predicción de la demanda de transporte en Colombia.

***Palabras clave*:** Machine Learning, Transporte de Carga, Manifiesto de Carga, Grandient Boosting Regressor, Costos de Transporte.

Abstract

Cargo transportation in Colombia presents significant challenges due to the variability in demand and the associated costs incurred. This work proposes the design of a Machine Learning model that enables the characterization of cargo and the estimation of transportation costs based on the historical cargo manifests collected between 2015 and 2023 in Colombia through the RNDC platform of the Ministry of Transportation. The aim is to improve the accuracy of cost estimation and optimize cargo transportation logistics.

To achieve this objective, an exploratory data analysis will be conducted to characterize cargo transportation. Various Machine Learning techniques will be implemented, including linear regression, decision trees, and neural networks, among others. The data was preprocessed to manage variability and inconsistencies, and cross-validation methods were applied to ensure the model's robustness. The expected results will yield a model suitable for implementation in the Power BI platform for data visualization and interaction. The selected model is expected to offer the highest precision in cost estimation, with a significant reduction in the average error compared to the tested methods.

These findings demonstrate the potential of Machine Learning to transform cargo transportation management by providing more accurate tools for decision-making and logistical planning. Implementing this model could result in considerable savings and a general improvement in transportation efficiency in Colombia. Additionally, it will lay a solid foundation for improving cost prediction models by incorporating new data, such as regional characteristics, economic size, productive vocation, and fluctuations in freight and fuel prices. This would highlight the importance of considering a broader range of variables to accurately capture the dynamics of transportation costs.

Finally, future studies could focus on other areas such as cargo type classification, transportation route segmentation, temporal trend analysis, and transportation demand forecasting in Colombia.

Keywords**:** Machine Learning, Cargo Transportation, Cargo Manifest, Gradient Boosting Regressor, Transportation Costs..

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Lista de Apéndices

Introducción

El sector transporte en Colombia desempeña un papel fundamental en el desarrollo económico del país, ya que conecta las regiones productivas con los centros de consumo y los mercados internacionales. Sin embargo, este sector enfrenta diversos retos, como la falta de infraestructura adecuada, altos costos logísticos, y una limitada capacidad para monitorear y gestionar eficientemente las operaciones de transporte. En este contexto, el uso de datos se presenta como una herramienta clave para abordar estos desafíos y transformar la manera en que se diseñan y ejecutan las políticas públicas. A pesar de la importancia de los datos, su integración en los procesos de toma de decisiones ha sido limitada, subutilizando un recurso que podría proporcionar información valiosa para mejorar la eficiencia y competitividad del sector.

Desde la implementación de la Resolución 377 de 2013 del Ministerio de Transporte, se estableció la obligatoriedad para las empresas de transporte de registrar todas sus operaciones mediante manifiestos de carga. Este marco normativo buscaba generar un histórico de datos que permitiera un mejor entendimiento de las dinámicas del transporte de carga en Colombia. Sin embargo, su aplicación práctica ha sido desigual, con una baja adopción en la formulación y evaluación de políticas públicas. Gran parte de esta información sigue sin ser explotada de manera efectiva, lo que representa una oportunidad desaprovechada para optimizar procesos, reducir costos y diseñar estrategias más informadas en el sector.

El presente trabajo tiene como objetivo principal aprovechar esta vasta base de datos generada en la última década para estimar los costos de transporte de carga en Colombia. Este análisis no solo contribuirá a una mejor comprensión de los factores que influyen en estos costos, sino que también establecerá una base sólida para investigaciones futuras en áreas como la segmentación de rutas, caracterización de carga y predicción de la demanda de transporte. De

esta manera, se busca demostrar cómo el uso adecuado de datos puede generar información estratégica que impacte positivamente la gestión del sector transporte.

Además, este enfoque tiene implicaciones significativas para el diseño de políticas públicas en Colombia. Al proporcionar un modelo basado en datos reales y actualizados, los formuladores de políticas podrán identificar patrones, tendencias y problemáticas específicas que afectan el transporte de carga. Esto permitirá la creación de normativas más efectivas, programas de incentivo adecuados, y estrategias que impulsen el desarrollo de cadenas logísticas sostenibles. En resumen, este trabajo se posiciona como un puente entre la generación de datos y su uso práctico para transformar el sector transporte, promoviendo una cultura de decisiones informadas que beneficien tanto al sector público como privado.

Planteamiento del Problema

El problema central que enfrenta el sector de transporte en Colombia se centra en la baja capacidad para caracterizar adecuadamente la carga y estimar con precisión los costos, esto se evidenció en la última década y en el proceso de implementación de la Resolución 377 de 2013, del Ministerio de Transporte (Resolución 377 de 2013, 2013), donde se obliga a las empresa de transporte a registrar todas sus operaciones de carga a través de manifiestos de carga, creando un histórico de información con baja aplicación en la toma de decisiones. Esta limitación se traduce en una serie de ineficiencias operativas y estratégicas que afectan directamente la rentabilidad y competitividad de las empresas dentro de este sector. La caracterización inadecuada de la carga y la estimación errónea de costos no solo dificultan la planificación logística, sino que también impiden una asignación óptima de recursos y activos de transporte, generando sobrecostos y afectando la calidad del servicio ofrecido (Gonzalez, 2022).

La ineficiencia en la planificación logística conlleva a aumentos en los tiempos de tránsito de la carga, lo que repercute negativamente en la satisfacción del cliente y en la capacidad de cumplir con los tiempos de entrega prometidos (Martinez, 2022). A su vez, el uso ineficiente de recursos y activos de transporte incrementa los costos operativos, reduciendo los márgenes de beneficio y limitando la capacidad de inversión en innovación y mejora continua. Estas consecuencias resaltan la necesidad urgente de optimizar los procesos logísticos y de estimación de costos para mantenerse competitivos en el mercado (Boyer et al., 2017).

Por otro lado, los errores en la estimación de costos de transporte llevan a pérdidas financieras significativas debido a cotizaciones inexactas, lo que puede resultar en la incapacidad para competir efectivamente en precios de transporte. Esta situación se agrava aún más por la dependencia en decisiones basadas en datos limitados o inexactos, lo que dificulta la

adaptabilidad y respuesta a los cambios en la demanda del mercado. La falta de adaptabilidad no solo impide una inversión eficaz en infraestructura logística y tecnología, sino que también limita la capacidad de ofrecer servicios personalizados y adaptados a las necesidades específicas de los clientes (Plakandaras et al., 2019).

Además, la baja capacidad en la caracterización de carga y estimación de costos de transporte tiene un impacto directo en la competitividad del mercado, evidenciado por la pérdida de clientes a favor de competidores más eficientes y el deterioro de la imagen y reputación de la empresa. Este panorama subraya la importancia de desarrollar estrategias innovadoras y soluciones tecnológicas avanzadas para abordar eficazmente estos desafíos.

En este contexto, surge la propuesta de desarrollar e implementar un modelo de Machine Learning como solución al problema identificado. Este modelo permitiría una caracterización avanzada de la carga y una estimación de costos de transporte más precisa, aprovechando los datos históricos de manifiestos de carga. La implementación de este modelo de Machine Learning no solo mejoraría la eficiencia y precisión en la planificación logística y la estimación de costos, sino que también fortalecería la toma de decisiones basada en datos, aumentando la competitividad y sostenibilidad de las empresas de transporte en Colombia.

Es así que se define como pregunta de investigación, *¿Cuál es el modelo de Machine Learing que permite estimar los costos de transporte con mayor precisión, con base en variables provenientes de los manifiestos de carga en Colombia?*

Justificación

La necesidad de desarrollar e implementar un modelo de Machine Learning para la caracterización de carga y estimación de costos de transporte en Colombia se justifica por varias razones críticas que afectan tanto a la eficiencia operativa como a la competitividad de las empresas del sector. Primero, la caracterización inadecuada de la carga y las estimaciones inexactas de costos han resultado en una planificación logística deficiente, uso ineficiente de los recursos y, en última instancia, en una disminución en la satisfacción del cliente. Esto no solo afecta la rentabilidad de las empresas debido a los aumentos en los costos operativos, sino que también limita su capacidad de respuesta ante las demandas del mercado y la implementación de estrategias de mejora continua (Colfecar, 2022).

En segundo lugar, la incapacidad para proporcionar estimaciones precisas de costos representa un desafío significativo en un entorno competitivo, donde la precisión en la cotización puede ser un factor determinante para ganar o perder clientes. Las empresas que se basan en datos históricos limitados o inexactos para tomar decisiones estratégicas encuentran dificultades para adaptarse a los cambios del mercado, lo que puede resultar en decisiones de inversión equivocadas y en la pérdida de oportunidades clave (Martinez, 2022).

Además, el entorno empresarial actual exige una mayor flexibilidad y capacidad de adaptación a las tendencias emergentes. La implementación de un modelo de Machine Learning permitiría a las empresas analizar grandes volúmenes de datos históricos de manera eficiente, identificando patrones y tendencias que pueden informar decisiones estratégicas más efectivas. Esta tecnología no solo mejora la precisión en la caracterización de carga y en las estimaciones de costos, sino que también facilita una planificación más estratégica y una gestión de recursos más eficiente.

La justificación para este proyecto también se basa en la oportunidad de mejorar la competitividad de las empresas colombianas en el sector de transporte. Al adoptar soluciones tecnológicas avanzadas, las empresas pueden mejorar significativamente su oferta de servicios, aumentando su atractivo para los clientes actuales y potenciales (Gonzalez, 2022). La capacidad de ofrecer estimaciones de costos rápidas y precisas, junto con una planificación logística optimizada, puede ser un diferenciador clave en el mercado.

Finalmente, el desarrollo de un modelo de Machine Learning para abordar estos desafíos no solo tiene el potencial de transformar la operación individual de las empresas, sino que también puede contribuir al avance de la industria del transporte en Colombia en su conjunto. Al mejorar la eficiencia, reducir costos y aumentar la satisfacción del cliente, este proyecto puede servir como un modelo para la adopción de tecnologías innovadoras en el sector, promoviendo un enfoque más basado en datos para la toma de decisiones estratégicas y operativas.

Objetivos

Objetivo General

Desarrollar un modelo de Machine Learning que permita la estimación de los costos del transporte de carga en Colombia a partir de datos históricos de manifiestos de carga entre 2015 y 2023.

Objetivos Específicos

Establecer un análisis exploratorio de los manifiestos de carga para la caracterización del transporte en Colombia.

Diseñar y entrenar un modelo que permita la estimación de los costos de transporte de carga en Colombia

Validar y optimizar modelo de Machine Learning que permita la toma decisiones basadas en datos en los costos de transporte de carga en Colombia.

Marco Referencia

Marco Teórico

Machine Learning

El Machine Learning (ML) es un tipo de técnica de Inteligencia Artificial (IA) que permite que el sistema obtenga conocimiento sin programación explícita. El objetivo principal de la técnica ML es permitir que las computadoras aprendan sin ayuda humana, ML se divide principalmente en tres categorías, enfoques de aprendizaje supervisado, no supervisado y semisupervisado (Saravanan & Sujatha, 2019), aunque a primera vista pueden parecer conceptos complejos, su comprensión es esencial para desentrañar cómo las máquinas aprenden de los datos.

Los algoritmos supervisados necesitan humanos para proporcionar la entrada y la salida requerida, además de proporcionar comentarios sobre la precisión de la predicción en el proceso de entrenamiento, por otro lado, los enfoques de aprendizaje no supervisado contrastan con los enfoques de aprendizaje supervisado, ya que no requiere ningún proceso de capacitación (Amanpreet et al., 2016).

En la implementación se destaca que los enfoques de aprendizaje supervisado son más simples que los enfoques de aprendizaje no supervisado, ya que se identifican de forma más sencillas los mecanismos inherentes en el proceso. En cuanto a los mecanismos semi supervisados, son lo que en su desarrollo utilizan herramientas de ambas aproximaciones, siendo su implementación más adaptada a las ventajas de ambos, sin embargo, con las debilidades que tiene cada una de ellas (van Engelen & Hoos, 2020).

Aprendizaje Supervisado

En el vasto conocimiento de la inteligencia artificial, el Machine Learning o aprendizaje automático emerge como una de sus ramas más fascinantes, prometiendo revolucionar desde la forma en que interactuamos con la tecnología hasta cómo entendemos el mundo que nos rodea. Dentro de este campo, dos conceptos fundamentales son el aprendizaje supervisado y el no supervisado, cada uno con sus metodologías, aplicaciones y desafíos únicos (Alloghani et al., 2020).

El aprendizaje supervisado, como su nombre lo indica, funciona bajo la tutela de un mecanismo de supervisión, donde existe un conjunto de datos etiquetados que alimentan el algoritmo, es decir, cada ejemplo o referencia de entrenamiento viene con la respuesta correcta. El objetivo es que, a través del entrenamiento, el algoritmo aprenda a predecir la salida correcta para nuevos datos (Sindhu Meena & Suriya, 2020). Se podría asimilar a como enseñarle a un niño a distinguir entre diferentes tipos de frutas mostrándole ejemplos concretos de cada una; de manera similar, en el aprendizaje supervisado, el modelo se entrena con ejemplos etiquetados hasta que puede hacer predicciones por sí mismo (Sindhu Meena & Suriya, 2020).

Aprendizaje No Supervisado

El aprendizaje no supervisado, no se tiene esta respuesta correcta, sin una ruta clara a seguir, en este enfoque entonces, los algoritmos se enfrentan a datos no etiquetados, lo que significa que no se les dice cuál es la respuesta correcta durante el entrenamiento. Su tarea es identificar patrones, agrupaciones o anomalías dentro de los datos por sí mismos (Sindhu Meena & Suriya, 2020). Si hacemos nuevamente la analogía con un niño, sería como dejarlo en un cuarto lleno de juguetes de diferentes formas y colores y observar cómo los organiza intuitivamente en grupos sin que nadie le diga cómo hacerlo. Este tipo de aprendizaje es

poderoso para descubrir la estructura oculta en los datos y es ampliamente utilizado en la detección de anomalías, segmentación de mercado y sistemas de recomendación.

La elección entre aprendizaje supervisado y no supervisado depende en gran medida del problema específico a resolver y del tipo de datos disponibles. El aprendizaje supervisado es ideal cuando el objetivo es predecir o clasificar resultados basándose en ejemplos pasados. Por ejemplo, podría utilizarse para predecir el precio de una casa basándose en características como su tamaño y ubicación, o para diagnosticar enfermedades a partir de imágenes médicas. En contraste, el aprendizaje no supervisado es la herramienta de elección cuando se desconoce la estructura de los datos o cuando se busca explorar los datos para encontrar patrones o agrupaciones intrínsecas (Rajoub, 2020). A medida que avanzamos hacia un futuro cada vez más impulsado por datos, la importancia de comprender estos dos enfoques de aprendizaje automático solo puede aumentar. Las aplicaciones de ambos métodos ya están transformando industrias, desde el desarrollo de sistemas de recomendación personalizados hasta la mejora de la detección temprana de enfermedades, pasando por la optimización de operaciones logísticas. A medida que la tecnología continúa evolucionando, el aprendizaje supervisado y no supervisado seguirán siendo pilares fundamentales en el desarrollo de sistemas inteligentes capaces de aprender de la complejidad del mundo real (Verma et al., 2022).

Áreas de Trabajo del Machine Learning

El aprendizaje autónomo ha demostrado tener capacidades para resolver gran cantidad de problemas en todos campos del saber, hasta la fecha la página (Taylor, 2020) ha podido identificar un total de 12.444 comparativas de rendimiento, 4.700 tareas y un total de 119.050 artículos científicos con código, que utilizan el Machine Learning, en la siguiente Tabla 1, se

hace una selección de dos tareas en 33 áreas de desarrollo, contabilizando el número de artículos por tarea.

Tabla 1

Documentos Consultados y Hallazgos

Área	Tarea	Artículos
Computer Vision	Semantig Segmentation	4949
Computer Vision	Image Classification	3624
Natural Language Processing	Language Modelling	3976
Natural Language Processing	Translation	3108
Medical	Medical Image Segmentation	704
Medical	EEG	369
Miscellaneous	Transfer Learning	2695
Miscellaneous	BIG-bench Machine Learning	2313
Methodology	General Classification	3925
Methodology	Reinforcement Learning (RL)	3805
Time Series	Time Series Analysis	1863
Time Series	Computational Efficiency	771
Graphs	Link Prediction	783
Graphs	Node Classification	743
Speech	Speech Recognition	1057
Speech	Automatic Speech Recognition (ASR)	469
Audio	Speech Synthesis	413
Audio	Voice Conversion	144
Reasoning	Decision Making	1897
Reasoning	Navigate	377
Computer Code	Semantic Segmentation	4949
Computer Code	Test	3565
Playing Games	Sentence	3284
Playing Games	Continuous Control	398

Adversarial	Adversarial Robustness	579
Adversarial	Adversarial Attack	574
Robots	Benchmarking	1369
Robots	Motion Planning	184
Knowledge Base	Knowledge Graphs	914
Knowledge Base	Causal Inference	395
Knowledge Base	Explainable Artificial Intelligence (XAI)	197
Music	Music Generation	123
Music	Music Information Retrieval	92

Nota. Tomado de la fuente mencionada en (Taylor, 2020)

Aplicaciones del Machine Learning en Logística y Transporte

El machine learning (ML) ha revolucionado numerosos sectores, incluido el de la logística y el transporte. Estas tecnologías no solo optimizan las operaciones, sino que también contribuyen a la reducción de costos y mejoran la toma de decisiones. Algunas de estas aplicaciones en el sector se centran en la predicción de la predicción de demanda, donde el ML permite analizar grandes volúmenes de datos históricos para predecir la demanda futura de servicios de transporte y logística. Modelos predictivos como las redes neuronales y las máquinas de soporte vectorial son utilizados para estimar las variaciones estacionales, diarias o incluso horarias en la demanda de envíos, lo que ayuda a las empresas a prepararse adecuadamente y optimizar sus recursos.

En las aplicaciones de optimización de rutas, los algoritmos de ML ayudan a desarrollar rutas de entrega más eficientes, reduciendo el tiempo de viaje y el consumo de combustible. Algoritmos como el de optimización de enjambre de partículas o algoritmos genéticos pueden analizar factores como el tráfico en tiempo real, las condiciones meteorológicas, y los plazos de entrega para ajustar y mejorar las rutas en tiempo real (de la Torre et al., 2021). En la gestión de

inventarios, el ML mejora la precisión de los sistemas de gestión de inventarios al predecir los niveles óptimos de stock basados en tendencias históricas, cambios de estacionalidad, y otros factores predictivos. Esto reduce el exceso de inventario y minimiza los costos asociados con el almacenamiento y la posible obsolescencia de los productos.

En el mantenimiento predictivo de los vehículos es crucial para evitar interrupciones y retrasos costosos. El ML permite implementar estrategias de mantenimiento predictivo, utilizando datos de sensores y registros históricos para predecir cuándo es probable que una pieza del vehículo falle. Esto permite realizar mantenimientos proactivos antes de que ocurran averías, mejorando la fiabilidad de la flota (Boukerche & Wang, 2020). La estimación de costos de transporte se utilizan modelos de regresión avanzados analizan una variedad de factores, como la distancia, el tipo de carga, el consumo de combustible, y las tarifas de mano de obra para proporcionar estimaciones precisas de los costos de transporte. Estas estimaciones ayudan a las empresas a formular estrategias de precios más efectivas y a mejorar la negociación de contratos con clientes y proveedores (Farchi et al., 2023; Singh et al., 2021).

En la automatización de procesos, el ML facilita la automatización de tareas administrativas repetitivas en la logística, como la facturación, el seguimiento de envíos y la gestión de documentos. Los modelos de aprendizaje profundo, por ejemplo, pueden analizar documentos de envío para extraer información relevante automáticamente, reduciendo errores y mejorando la eficiencia. Finalmente se encuentra también aplicaciones donde se hace análisis de sentimientos, los modelos de procesamiento de lenguaje natural (NLP) se utilizan para analizar las opiniones y sentimientos de los clientes a partir de comentarios y reseñas en línea. Esto proporciona valiosas perspectivas sobre la satisfacción del cliente y ayuda a las empresas a mejorar sus servicios y atención al cliente

Marco Normativo

La base de datos de información se basa en la consolidación de los manifiestos de carga que las empresas de transporte en el país han declarado desde 2015 en Colombia gracias a la Resolución 377 de 2013, donde se implementa el Registro Nacional de Despacho de Carga RNDC (Resolución 377 de 2013, 2013), otras disposiciones y el marco normativo se pueden ver en la siguiente tabla, donde se reconoce hasta la actualidad, todos los procesos de actualización y sobre la normatividad, además de disposiciones general, anexos, resoluciones, decretos, entre otros que muestran la evolución de la normatividad.

Tabla 2

Marco Normativo RNDC – Ministerio de Transporte

Título	Descripción	Categoría
No.20223040045515 de 2022	Resolución "Por la cual se actualiza el sistema del Registro Nacional de Despachos de Carga -RNDC y se dictan otras disposiciones" Agosto 5 del 2022- Pag 1 a 12	Resolución
No.20223040045515 Anexo 1- Parte 1	ANEXO 1- PARTE 1 MANUAL DE DESCRIPCIÓN E INSTRUCCIONES PARA LA OPERACIÓN GENERAL DEL REGISTRO NACIONAL DE DESPACHOS DE CARGA RNDC - Pag 13 a 56	Resolución
No.20223040045515 Anexo 1 - Parte 2	ANEXO 1- PARTE 2- MANUAL DE DESCRIPCIÓN E INSTRUCCIONES PARA LA OPERACIÓN GENERAL DEL REGISTRO NACIONAL DE DESPACHOS DE CARGA RNDC - Pag 57 a 70	Resolución
377 de 2013	Se adopta e implementa el Registro Nacional de Despacho de Carga RNDC	Resolución
1079 de 2015	Se expide el Decreto Único Reglamentario del Sector Transporte	Decreto
2228 de 2013	Se definen las relaciones económicas entre las empresas de transporte, generadores y tenedores de vehículos usando referencia Sice-Tac	Decreto
2044 de 1988	Decreto de Contratación Directa 2044 de Sep 30 de 1988. Lista de productos excluidos para contratar empresas de transporte.	Decreto
1766 de 2016	Decreto 1766 por el cual se definen los requisitos para la movilización de Transporte de Ganado	Decreto
20213040005875 de 2021	Resolución de Puertos que reglamenta el envío de la información al modulo RNDC - INSIDE Información del Sistema de Enturnamiento Portuario. Feb 15 de 2021	Resolución

3958 de 2018	Se modifica el artículo 2° y 3° y se prorroga el término establecido en el artículo 4° de la resolución 0000540 de 2018	Resolución
3956 de 2018	Se adiciona parágrafo al artículo 2° se prórroga el término establecido en el artículo 4° de la resolución 0000541 de 2018	Resolución
2534 de 2018	Se realiza la Intervención de las rutas Medellín - -Sincelejo, Medellín - Cartagena, Medellín - Barranquilla, Medellín - Santa Marta	Resolución
541 de 2018	Se establece una tarifa diferencial del 50% en las estaciones de peaje La Loma, El Copey y Tucarinca para transporte de Carbón mineral	Resolución
540 de 2018	Se establece una tarifa diferencial del 50% en las estaciones de peaje La Loma, El Copey y Tucarinca para transporte de Carbón mineral	Resolución
4256 de 2016	Se prorroga término concedido en artículo 4 de las resoluciones 3376 y 3377 de 2015, prorrogados por las resoluciones 1134, 1136 y 3900 de 2016	Resolución
3443 de 2016	Se dictan lineamientos para el control del cumplimiento de las normas que rigen la actividad transportadora	Resolución
3442 de 2016	Se realiza la intervención de las rutas Manizales - Bogotá, Manizales - Barranquilla, Manizales - Medellín, Manizales - Buenaventura, Pasto - Buenaventura, Buenaventura - Pitalito	Resolución
3441 de 2016	Se realiza la intervención de las rutas Buenaventura - Cali, Cali - Barranquilla, Bogotá - Cali	Resolución
3440 de 2016	Se realiza la intervención de las rutas Bogotá - Buenaventura, Bogotá - Ipiales, Bogotá - Cartagena, Barrancabermeja - Bogotá	Resolución
3439 de 2016	Se realiza la intervención de las rutas Duitama - Cartagena, Duitama - Buenaventura, Duitama - Barranquilla	Resolución
3438 de 2016	Se realiza la intervención de las rutas Santa Marta - Bucaramanga, Cúcuta - Barranquilla, Barranquilla - Barrancabermeja, Barrancabermeja - Rubiales	Resolución
3437 de 2016	Se realiza la intervención de las rutas Medellín - Buenaventura, Medellín - Cartegena, Medellín Barranquilla	Resolución
757 de 2015	Se establece la aplicación del artículo 2 del decreto 2228 de 2013. En ningún caso se pueden efectuar pagos por debajo de los costos publicados en el SICE TAC	Resolución
1499 de 2009	Se modifica y se derogan algunas disposiciones de los Decretos 173 del 5 de febrero de 2001 y 1842 del 25 de mayo de 2007	Decreto
4959 de 2006	Se fijan los requisitos y procedimientos para conceder los permisos para el transporte de cargas indivisibles extrapesadas y extra dimensionadas	Resolución
4100 de 2004	Se adoptan los límites de pesos y dimensiones en los vehículos de transporte terrestre automotor de carga por carretera	Resolución

No.20223040045515 de 2022	Resolución" Por la cual se actualiza el sistema del Registro Nacional de Despachos de Carga -RNDC y se dictan otras disposiciones" Agosto 5 del 2022- Pag 1 a 12	Resolución
No.20223040045515 Anexo 1- Parte 1	ANEXO 1- PARTE 1 MANUAL DE DESCRIPCIÓN E INSTRUCCIONES PARA LA OPERACIÓN GENERAL DEL REGISTRO NACIONAL DE DESPACHOS DE CARGA RNDC - Pag 13 a 56	Resolución
No.20223040045515 Anexo 1 - Parte 2	ANEXO 1- PARTE 2- MANUAL DE DESCRIPCIÓN E INSTRUCCIONES PARA LA OPERACIÓN GENERAL DEL REGISTRO NACIONAL DE DESPACHOS DE CARGA RNDC - Pag 57 a 70	Resolución
377 de 2013	Se adopta e implementa el Registro Nacional de Despacho de Carga RNDC	Resolución
1079 de 2015	Se expide el Decreto Único Reglamentario del Sector Transporte	Decreto
2228 de 2013	Se definen las relaciones económicas entre las empresas de transporte, generadores y tenedores de vehículos usando referencia Sice-Tac	Decreto
2044 de 1988	Decreto de Contratación Directa 2044 de Sep 30 de 1988. Lista de productos excluidos para contratar empresas de transporte.	Decreto

Nota. Basada en el contenido de la página (Mintransporte, 2024)

Metodología

Tipo de Estudio

La metodología de desarrollo de este proyecto aplicado es una aproximación mixta (descriptiva y experimental), en la cual se utiliza el diseño de algoritmos para extraer y procesar la información, procesos descriptivos para reconocer los datos, la información que contienen y procesos experimentales en el diseño e implementación de los modelos de Machine Learning para encontrar un candidato que se pueda validar y permita modelar la estimación los costos.

Método

La metodología de desarrollo de este proyecto aplicado es una aproximación mixta (descriptiva y experimental), en la cual se utiliza el diseño de algoritmos para extraer y procesar la información, procesos descriptivos para reconocer los datos, la información que contienen y procesos experimentales en el diseño e implementación de los modelos de Machine Learning para encontrar un candidato que se pueda validar y permita modelar la estimación los costos.

Para esto se definen una serie de actividades a desarrollar por cada uno de los objetivos específicos propuestos en donde para el primer objetivo definido como establecer un análisis exploratorio de los manifiestos de carga para la caracterización del transporte en Colombia, donde se hará una recolección y consolidación de datos, basados en la plataforma destinada por el Ministerio de Transporte, donde se agruparán los datos históricos de manifiestos de carga, asegurando que incluyan todas las variables clave como tipos de carga, rutas, volúmenes, y costos asociado, esta información se hará una limpieza de datos, se reconocerá cuales se pueden eliminar basado en técnicas de tratamiento de datos, tales como información duplicada, incompleta, valores faltantes y finalmente validar la integridad de los datos.

Se desarrollará un análisis estadístico descriptivo, donde se reconocerá la distribución de las variables, promedios, variaciones, y otros datos estadísticos relevantes que permitan detallar el contenido de la base de datos y resumir los resultados, con esto se crearán crear gráficos para visualizar las tendencias en los tipos de carga, rutas más frecuentes, y temporadas de alta actividad. Esto facilitará la identificación de patrones y correlaciones, donde se utilizarán técnicas de análisis para identificar patrones recurrentes o correlaciones significativas que podrían influir en los costos de transporte.

Para el segundo objetivo de diseñar y entrenar un modelo que permita la estimación de los costos de transporte de carga en Colombia, se plantean las siguientes actividades, donde se hace una selección de características, para identificar las variables más influyentes que afectan los costos de transporte a partir del análisis exploratorio, la selección de un modelo de ML, basado en la naturaleza de los datos y los resultados de pruebas con modelos tipo regresión lineal, árboles de decisión y máquinas de soporte vectorial; adiciona hace una división de datos, para lograr crear un grupo de entrenamiento y validación para asegurar la generalización del modelo, luego una etapa de entrenamiento del modelo, para ajustar los parámetros del mismo, y una validación de la precisión del modelo en el conjunto de validación, y así hacer un nuevo ajuste de los parámetros según sea necesario para mejorar el rendimiento.

Para el tercer objetivo definido como validar y optimizar el modelo de Machine Learning que permita la toma de decisiones basadas en datos en los costos de transporte de carga en Colombia, a través de técnicas como hiper parámetros, que facilite ajustar los hallazgos en los procesos de entrenamiento, así finalmente se determina la predictibilidad y se analizan los residuos del modelo

Este modelo se espera que logre una precisión suficiente en la predicción de costos de transporte, demostrando ser una herramienta valiosa para las empresas de logística. Esto indicaría que las variables seleccionadas y el algoritmo utilizado son adecuados para capturar la complejidad del problema (Morabit et al., 2023) . Además, podría facilitar la identificación de ciertas variables, que sean predictores fuertes en la determinación de los costos de transporte de carga en Colombia y su selectividad en las diferentes regiones del país y el movimiento en los territorios. Esto no solo validaría la selección de características, sino que también proporciona perspectivas valiosas sobre los factores que más influyen en los costos de transporte en el país.

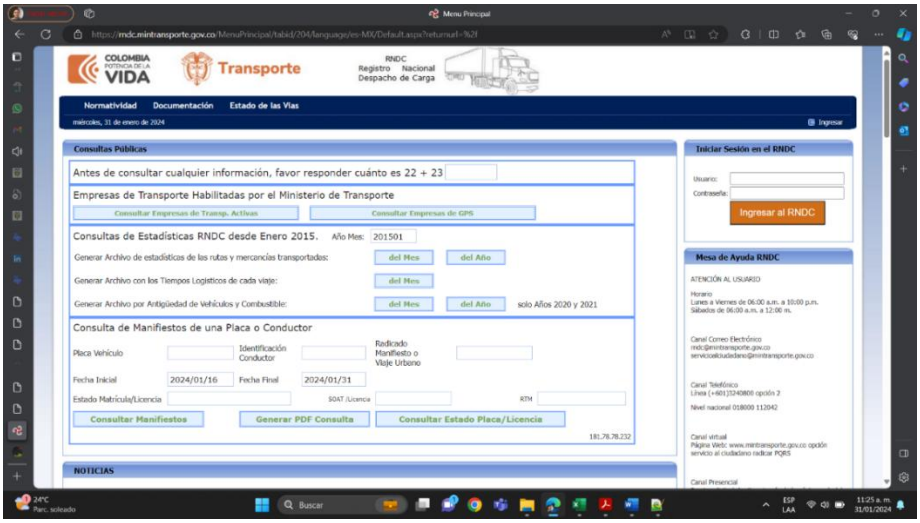
Esto también permitiría dejar unas bases sólidas para mejorar el modelo para predicción de costos, incluyendo nuevos datos, como características de la región, tamaño económico, vocación productiva, variación de precios de fletes y combustible, esto subrayaría la importancia de considerar una gama más amplia de variables para capturar con precisión la dinámica del costo de transporte. Además de sentar bases para otro tipo de estudios como clasificación de tipo de carga (Nama et al., 2021), segmentación de rutas de transporte (Akbari & Do, 2021), análisis de tendencias temporales (Moscoso-López et al., 2021) y predicción de la demanda de transporte en Colombia (Plakandaras et al., 2019).

Recolección de Datos

Se obtiene una base de datos para esta propuesta se define a través de los datos publicados por el Ministerio de Transporte Información Registro Nacional de Despachos Carga (RNDC), en el cual las empresas de carga están obligados a registrar todos los despachos de carga a través de un manifiesto de carga (MINTRANSPORTE, 2024), en la Figura 1, se puede ver una imagen de la plataforma donde se puede extraer la información, donde se accede a través de cumplir un validador tipo suma, y con este se da acceso a la información solicitada.

Figura 1

Imagen de Plataforma para Recolección de Información RNDC



Nota. Tomado de (MINTRANSPORTE, 2024)

Figura 2

Viajes por Año Registrados RNDC 2017-2023

Nota. Tomado de (MINTRANSPORTE, 2024)

Los datos que se pueden extraer de la plataforma son desde el año 2015 hasta la fecha, para hacer una correcta delimitación del proceso se hace haría una segmentación desde el año 2015 hasta 2023, en la Figura 2 se puede ver la totalidad de viajes registrados desde el 2017 hasta 2023, el año 2023 se tuvieron un total de 10’940.686 viajes registrados, esto permite un volumen adecuado de datos para la construcción de un modelo adecuado de ciencia de datos para la caracterización y modelado de los costos y carga en Colombia.

Resultados

Resultado Objetivo 1

Para el análisis exploratorio de los datos, se desarrolla varias etapas de recolección, limpieza y análisis de los datos para ello, las variables que contienen la base de datos con el total 29, en la siguiente tabla se detalla su nombre, tipo de dato y su descripción.

Tabla 3

Descripción de Variables de la Base de Datos

Variable	Tipo	Descripción
MES	Texto	El texto se compone por el año y por el mes en el cual se esta registrado el viaje, por ejemplo 201910, viaje en el año 2019 en el mes de octubre
COD_CONFIG_VEHICULO	Texto	Código configuración vehículo
CONFIG_VEHICULO	Texto	Tipo de configuración del vehículo
CODOPERACIONTRANSPORTE	Texto	Código de la operación de transporte
OPERACIONTRANSPORTE	Texto	Tipo de operación de transporte
CODTIPOCONTENEDOR	Texto	Código del tipo de contenedor utilizado
TIPOCONTENEDOR	Texto	Tipo de contenedor en la operación
CODMUNICIPIOORIGEN	Número	Código DANE del municipio origen
MUNICIPIOORIGEN	Texto	Nombre del municipio de origen
CODMUNICIPIODESTINO	Número	Código DANE del municipio destino
MUNICIPIODESTINO	Texto	Nombre del municipio de destino
CODMERCANCIA	Texto	Código del tipo de mercancía transportada
MERCANCIA	Texto	Detalle de la mercancía transportada
NATURALEZACARGA	Texto	Naturaleza de la carga
VIAJESTOTALES	Número	Número de viajes totales declarados
KILOGRAMOS	Número	Kilogramos declarados en el manifiesto
GALONES	Número	Galones de líquidos transportados
VIAJESLIQUIDOS	Número	Viajes con transporte de mercancía líquida
VIAJESVALORCERO	Número	No se reconoce la variable, en investigación
KILOMETROS	Número	No se reconoce la variable, en investigación
VALORESPAGADOS	Número	Valores en pesos colombianos como costo de transporte

CODMUNICIPIOINTERMEDIO	Número	Código del DANE intermedio, en caso de que no se tenga municipio intermedio el valor es 0
MUNICIPIOINTERMEDIO	Texto	Nombre del Municipio intermedio, en caso de que no se tenga municipio intermedio el valor es nulo
DEPARTAMENTOINTERMEDIO	Texto	Nombre del Departamento intermedio, en caso de que no se tenga municipio intermedio el valor es nulo
KILOMETROSREGRESO	Número	Declaración de kilómetros de regreso
KILOGRAMOSREGRESO	Número	Declaración de peso de regreso
GALONESREGRESO	Número	Declaración de galones de regreso

Recolección de Datos

Para la recolección de datos se hace un análisis del perfil de la plataforma, en el cual se identifica que se hace una pregunta matemática y se puede extraer por mes los datos, es así que se hace una extracción manual, en el cual se va respondiendo la pregunta se extraen los datos mes a mes desde 201501 hasta 202312. Por cada mes se obtiene un archivo, en total se tiene 108 meses, de un total de 9 años, los datos se bajan en un formato .csv, estos se consolidan al final en un solo archivo a través de un código.

En la extracción se establece el orden de las columna que se nombra como ANOMES, Cod_Config_Vehiculo, CODMUNICIPIOORIGEN, CODMUNICIPIOINTERMEDIO, CodOperacionTransporte, las cuales se aprecia que los valores nulos o no validos, se agrega un valor numérico 0 en las variables Kilometros,valorespagados, CodMunicipioIntermedio,KilometrosRegreso,GalonesRegreso y en el caso de variables tipo texto, se agrega el caracter coma o punto en las variables MunicipioIntermedio y DepartamentoIntermedio. En este punto se identifica que estas variables no son consistentes en todos los meses disponibles dentro de la base de datos.

Limpieza de Datos

Los archivos se bajan de la base de datos en formato XLSX, en este caso es necesario pasarlos a un archivo CSV con el fin de lograr un adecuado manejo de los datos. En este caso se encuentra dos elementos principales:

1. La posición de los datos desde el año 2015 hasta 2018 varía con los años posteriores, esto hace que una consolidación de todos los años, se corran los datos y se ubiquen en una fila diferentes

2. A partir del año 2022 se agregaron 6 columnas más: (CODMUNICIPIOINTERMEDIO, MUNICIPIOINTERMEDIO, DEPARTAMENTOINTERMEDIO, KILOMETROSREGRESO, KILOGRAMOSREGRESO y GALONESREGRESO) , esto también debe considerarse al momento de juntar los datos

La solución que se ha tomado para esto implica

1. Juntar los datos en CSV en las franjas de año 2015 hasta 2018, 2019 hasta 2021, y 2022 hasta 2023, esto pasarse a una base de datos SQLite en tres tablas diferentes.

2. Luego juntar las tres tablas en una consolidado, basado en el nombre de cada una de las columnas.

Para la identificación de datos faltantes se observa que en las variables CODMUNICIPIOINTERMEDIO, MUNICIPIOINTERMEDIO, DEPARTAMENTOINTERMEDIO, KILOMETROSREGRESO, KILOGRAMOSREGRESO y GALONESREGRESO, esta solo a partir del año 2022, donde se agregaron, luego esto pueden ser candidatos para eliminarse. En el siguiente conteo se hace solo con los años 2022 y 2023, donde se tiene un total de 3.825.473 registros

Tabla 4

Identificación de Datos Faltantes

Variable	% datos	Valores cero, nulos, vacíos
MES	100%	0%
COD_CONFIG_VEHICULO	100%	0%
CONFIG_VEHICULO	100%	0%
CODOPERACIONTRANSPORTE	100%	0%
OPERACIONTRANSPORTE	100%	0%
CODTIPOCONTENEDOR	2.948 %	97.052% - 3.712.682 datos
TIPOCONTENEDOR	2.948 %	97.052% - 3.712.682 datos
CODMUNICIPIOORIGEN	100%	0%
MUNICIPIOORIGEN	100%	0%
DEPARTAMENTOORIGEN	100%	0%
CODMUNICIPIODESTINO	100%	0%
MUNICIPIODESTINO	100%	0%
DEPARTAMENTODESTINO	100%	0%
CODMERCANCIA	99.3%	0.69% - 26497 datos
MERCANCIA	99.3%	0.69% - 26497 datos
NATURALEZACARGA	99.3%	0.69% - 26497 datos
VIAJESTOTALES	100%	0%
KILOGRAMOS	94,690%	5,31% - 203150 datos
GALONES	99,142%	0.858 %- 32836 datos
VIAJESLIQUIDOS	99,986%	0.014% - 548 datos
VIAJESVALORCERO	99,982%	0.018% - 679 datos
KILOMETROS	99,957%	0.043% - 1640 datos
VALORESPAGADOS	84,169%	15.831% - 605608 datos
CODMUNICIPIOINTERMEDIO	0.395%	99.605% - 3.810.357 datos
MUNICIPIOINTERMEDIO	0.395%	99.605% - 3.810.357 datos
DEPARTAMENTOINTERMEDIO	0.395%	99.605% - 3.810.357 datos
KILOMETROSREGRESO	0.338%	99.662% - 3.812.530 datos

KILOGRAMOSREGRESO	0.284%	99.712% - 3.812.530 datos
GALONESREGRESO	0%	100% - 3825480 datos

Dado estos datos se hace eliminación y agregado de variables para reconocer algunos elementos de valor, En la siguiente tabla se detalla los cambios sobre las variables de la base de datos, en este caso se eliminan las variables CODOPERACIONTRANSPORTE, CODTIPOCONTENEDOR, CODMUNICIPIOORIGEN, CODMUNICIPIODESTINO, CODMUNICIPIOINTERMEDIO, MUNICIPIOINTERMEDIO, DEPARTAMENTOINTERMEDIO, KILOMETROSREGRESO, KILOGRAMOSREGRESO y GALONESREGRESO, por ser redundantes o no contener información que se puede utilizar en el análisis final.

Además, se agregan otras variables que facilitan la segmentación y clasificación de los datos: DEPARTAMENTOORIGEN, DEPARTAMENTODESTINO, Año y Mes.

Tabla 5

Detalle de Eliminación y Agregación de Variables

Variable	Tipo	Descripción	Justificación	Elemento
CODOPERACION TRANSPORTE	Texto	Código de la operación de transporte	Este valor está duplicado con la columna OPERACIONTRANSPORTE	Eliminado
CODTIPOCONTEN EDOR	Texto	Código del tipo de contenedor utilizado	Este valor está duplicado con la columna TIPOCONTENEDOR	Eliminado
CODMUNICIPIOO RIGEN	Número	Código DANE del municipio origen	Información duplicada con la columna MUNICIPIOORIGEN	Eliminado
CODMUNICIPIOD ESTINO	Número	Código DANE del municipio destino	Información duplicada con la columna MUNICIPIODESTINO	Eliminado
DEPARTAMENTO ORIGEN	Texto	Se extrae de la MUNICIPIOORIGEN, con el fin de establecer el departamento de la carga	Se agrega para facilitar el proceso de análisis de información con una capa adicional determinada por el departamento de origen	Agregado
DEPARTAMENTO DESTINO	Texto	Se extrae de la MUNICIPIODESTINO, con el	Se agrega para facilitar el proceso de análisis de información con una capa	Agregado

		fin de establecer el departamento de la carga	adicional determinada por el departamento de destino	
CODMUNICIPIOINTERMEDIO	Número	Código municipio intermedio del manifiesto	Se elimina ya que se considerará que el origen de la carga y destino es único	Eliminado
MUNICIPIOINTERMEDIO	Texto	Municipio intermedio del manifiesto	Se elimina ya que se considerará que el origen de la carga y destino es único	Eliminado
DEPARTAMENTOINTERMEDIO	Texto	Departamento intermedio del manifiesto	Se elimina ya que se considerará que el origen de la carga y destino es único	Eliminado
KILOMETROSREGRESO	Número	Declaración de kilómetros de regreso	No se tendrá en cuenta debido a la falta de datos en los años pasados a 2022	Eliminado
KILOGRAMOSREGRESO	Número	Declaración de peso de regreso	No se tendrá en cuenta debido a la falta de datos en los años pasados a 2022	Eliminado
GALONESREGRESO	Número	Declaración de galones de regreso	No se tendrá en cuenta debido a la falta de datos en los años pasados a 2022	Eliminado
Año	Número	Basado en la variable MES	Se estable en un valor de cuatro dígitos el año del registro	Agregado
Mes	Número	Basado en la variable MES	Se estable en un valor de dos dígitos el mes del registro	Agregado

Se agrega una tabla llamada CodificacionDANE, con el fin de facilitar la identificación de los lugares, el cual se extrae del Geoportal del DANE - Codificación Divipola, En la siguiente tabla se detallan los datos

Tabla 6

Detalle Tabla Codificaciondane

Variable	Descripción	Datos	Tipo
Código Departamento	Código del Departamento es un número de dos cifras se tiene un total de 33 datos, corresponde los 32 departamentos más la capital Bogotá D.C.	33	INT
Código Municipio	Código del Municipio es un número de 5 cifras, que representan los 1122 municipios de Colombia	1122	INT
Código Centro Poblado	Código del Centro Poblado, en este caso se tiene un total 8158 Centros Poblados dentro los municipios a lo largo de todo el país	8158	INT
Nombre Departamento	Es el nombre en Texto de la variable Código Departamento	33	TEXT
Nombre Municipio	Es el nombre de Texto del Código Municipio, en este caso hay menos datos ya que existe casos de homonimia	1040	TEXT

Nombre Centro Poblado	Es el nombre en Texto de la variable Código Centro Poblado, en este caso se tiene un total de 5714 variables distintas, este caso se presenta ya que se conversa el nombre y se tiene diferentes puntos dentro del centro poblado que están codificados o se presenta el caso de homonimia entre centros poblados, se distingue ya es por diferente departamento, municipio y posición GPS	5714	TEXT
Tipo Centro Poblado	Se tiene dos y corresponde a CABECERA MUNICIPAL O CENTRO POBLADO	2	TEXT
Longitud	Corresponde a longitud de la posición geoespacial, este caso de los 7652 valores se tiene 497 NULL	7652	REAL
Latitud	Corresponde a latitud de la posición geoespacial en este caso de los 7660 valores se tiene 497 NULL	7660	REAL
Nombre Distrito	Corresponde a nombre distrito especial, en el cual se tiene un total de 12 distritos especial a lo largo del país, en estos datos NULL hay 7670, el resto en alguno de los 12 distritos especiales	13	TEXT
ANM	Corresponde a las AREA NO MUNICIPALIZADAS con 27, ISLA con 4 y MUNICIPIO con 8127	3	TEXT
Nombre Área Metropolitana	Corresponde a nombre área metropolitana con un total de 6, y un total de 7756 datos NULL	7	TEXT

Dada la importancia de este tipo de datos en el cálculo de la distancia, se hacer un pequeño algoritmo que permita el calculo la distancia entre ciudad, para eso se parte de la suposición que es la misma distancia entre dos puntos, esto no necesariamente es cierto, pero es un supuesto que racionaliza las consultas ya que:

Si tomamos como punto de partida que todos los municipios son homogéneos y todos se comunican entre sí, para 1122 municipios, se tendría un total de par origen destino de 1.258.884, sin embargo el par origen/destino de los datos es de 120.782 y si se trabaja bajo el supuesto es la misma distancia, el par origen destino es de 56.843, adicional se encuentran par origen/destino iguales, luego son manifiestos de carga de Ida y Regreso o entrega la misma ciudad, si estos no se consideran, se tiene finalmente un total de par origen/destino de 56.760, esto demuestra que hay esta estructura no es homogénea y las operaciones de carga se concentran en varios municipios

Adicional dentro de la base se tiene una condición de Viaje de Ida y Regreso en este caso se tiene un total de 13396, luego estas no se tendrán en cuenta para facilitar el proceso de análisis.

Análisis Exploratorio de Datos

Ahora para el EDA propuesto se hace una descripción final de las variables, donde se detalla las variables, tipo y el valor más frecuente, además de la cantidad de datos diferentes que tiene, esta descripción se hace sobre los años 2015 a 2023, en el Anexo A, se puede ver el detalle de todos los cálculos.

Tabla 7

Descripción de Variables y Tipo

Variable	Tipos de datos	Datos diferentes	Tipo
MES	202311 - dato más frecuente con 174860 datos	108	INT
COD_CONFIG_VEHICULO	2 - dato más frecuente con 3599304 datos	47	TEXT
CONFIG_VEHICULO	Camión Rígido de 2 ejes - dato más frecuente con 6515656 datos	43	TEXT
CODOPERACIONTRANSPORTE	G - dato más frecuente con 12267067 datos	9	TEXT
OPERACIONTRANSPORTE	General - dato más frecuente con 12267067 datos	9	TEXT
CODTIPOCONTENEDOR	. - dato más frecuente con 13179325 datos	5	TEXT
TIPOCONTENEDOR	. - dato más frecuente con 13218122 datos	3	TEXT
CODMUNICIPIOORIGEN	11001000 - dato más frecuente con 1407649 datos	3817	INT

MUNICIPIOORIGEN	BOGOTA BOGOTA D. C. - dato más frecuente con 1407649 datos	3817	TEXT
DEPARTAMENTOORIGEN	VALLE DEL CAUCA - dato más frecuente con 1542453 datos	32	TEXT
CODMUNICIPIODESTINO	11001000 - dato más frecuente con 1029090 datos	5231	INT
MUNICIPIODESTINO	BOGOTA BOGOTA D. C. - dato más frecuente con 1029090 datos	5229	TEXT
DEPARTAMENTODESTINO	ANTIOQUIA - dato más frecuente con 1316697 datos	33	TEXT
CODMERCANCIA	9980 - dato más frecuente con 1234648 datos	2691	TEXT
MERCANCIA	PRODUCTOS VARIOS - dato más frecuente con 1234464 datos	1260	TEXT
NATURALEZACARGA	Carga Normal - dato más frecuente con 12678521 datos	8	TEXT
VIAJESTOTALES	1 - dato más frecuente con 7212041 datos	1603	INT
KILOGRAMOS	10000 - dato más frecuente con 529025 datos	413317	INT
GALONES	0 - dato más frecuente con 13249470 datos	86360	INT
VIAJESLIQUIDOS	0 - dato más frecuente con 13249470 datos	698	INT
VIAJESVALORCERO	0 - dato más frecuente con 11654550 datos	896	INT
KILOMETROS	0 - dato más frecuente con 3338431 datos	32806	INT
VALORESPAGADOS	0 - dato más frecuente con 1399270 datos	1830959	INT

Finalmente, la selección de las variables se hace sobre: MES, COD_CONFIG_VEHICULO, CODOPERACIONTRANSPORTE, CODTIPOCONTENEDOR, CODMUNICIPIOORIGEN, CODMUNICIPIODESTINO, CODMERCANCIA, NATURALEZACARGA, VIAJESTOTALES, KILOGRAMOS, GALONES, VIAJESLIQUIDOS, VIAJESVALORCERO, KILOMETROS, VALORESPAGADOS.

En estas se tiene varias que son texto por lo tanto dentro del algoritmos se transforman a valores numéricos, a través de un LabelEncoder, para el análisis descriptivo, se hace un análisis de cada una de las variables de los años 2015 a 2023 en el cual se define el conteo, el promedio, la desviación estándar, el valor mínimo, Q1, Q2, Q3 y el valor máximo, se muestra en la siguiente tabla

Tabla 8

Análisis Estadísticos de las Variables

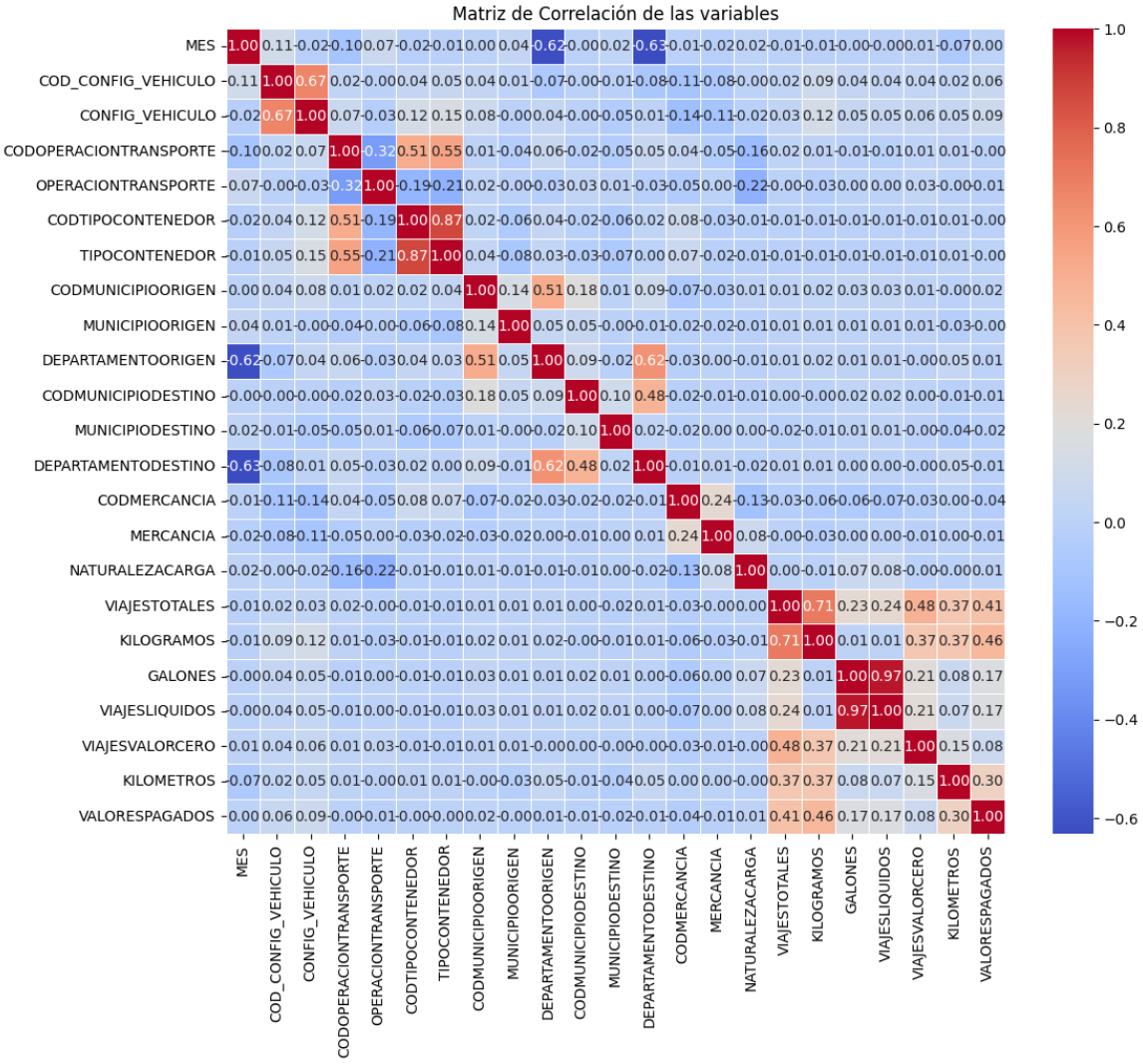
Variable	count	mean	std	min	25%	50%	75%	max
MES	1362730	201954,080	254,410644	201501	201711	202003	202203	202312
COD_CONFIG_VEHICULO	1362730	10,904955 7	12,073820 1	0	0	8	24	46
CONFIG_VEHICULO	1362730	12,534542 1	13,590232 2	0	1	1	30	43
CODOPERACIONTRANSPORTE	1362730	1,4826744 2	1,5427684 2	0	1	1	1	8
OPERACIONTRANSPORTE	1362730	2,9688081 4	0,5521479 1	0	3	3	3	8
CODTIPOCONTENEDOR	1362730	0,0442616 4	0,2866346 7	0	0	0	0	4
TIPOCONTENEDOR	1362730	0,0327987 0	0,1930453 9	0	0	0	0	2
CODMUNICIPIOORIGEN	13627308	34885212, 88	28365552, 18	0	11001000	25214000	68081000	99999999
MUNICIPIOORIGEN	13627308	1536,755641	1203,714855	0	405	1294	2372	3817
DEPARTAMENTOORIGEN	13627308	20,263122 4	12,209911 17	0	5	26	32	32
CODMUNICIPIODESTINO	13627308	36072378, 13	27504490, 55	0	11001000	25286000	66400000	99999999
MUNICIPIODESTINO	13627308	2343,043986	1649,532447	0	679	2028	3650	5229
DEPARTAMENTODESTINO	13627308	21,291453 53	12,014514 16	0	10	26	33	33

CODMERCANCIA	13627308	1739,314854	819,5855806	0	1081	1873	2482	2690
MERCANCIA	13627308	631,4537242	381,7491875	0	276	664	1031	1260
NATURALEZACARGA	13627308	3,077484269	0,48871336	0	3	3	3	7
VIAJESTOTALES	13627308	5,625777226	25,62246163	1	1	1	3	4005
KILOGRAMOS	13627308	76041,98625	545351,1667	0	4500	14000	34500	300054322
GALONES	13627308	2598,293404	57602,11181	0	0	0	0	16132972
VIAJESLIQUIDOS	13627308	0,297864699	5,97011813	0	0	0	0	1664
VIAJESVALORCERO	13627308	0,954330892	9,190964117	0	0	0	0	1649
KILOMETROS	13627308	730,9734385	5064,728539	0	10	238	571	1014701
VALORESPAGADOS	13627308	7297767,351	56582348,6	-25497	726220	1900000	4680000	115643520000

Finalmente se hace una estimación de la matriz de correlación de las variables numéricas, en esta caso la variable objetivo es VALORESPAGADOS, la cual tiene un correlación positiva con KILOGRAMOS, VIAJESTOTALES y KILOMETROS, también es importante reconocer que los VIAJESTOTALES también guarda una relación positiva KILOGRAMOS, VIAJESVALORCERO, KILOMETROS, no se toman otros tipos de correlación ya que no tiene una buena correlación con VALORESPAGADOS, en la siguiente anterior (Figura 3) se muestra el resultado de la correlación en el rango de los años 2015 a 2023.

Figura 3

Matriz de Correlación de Todas las Variables

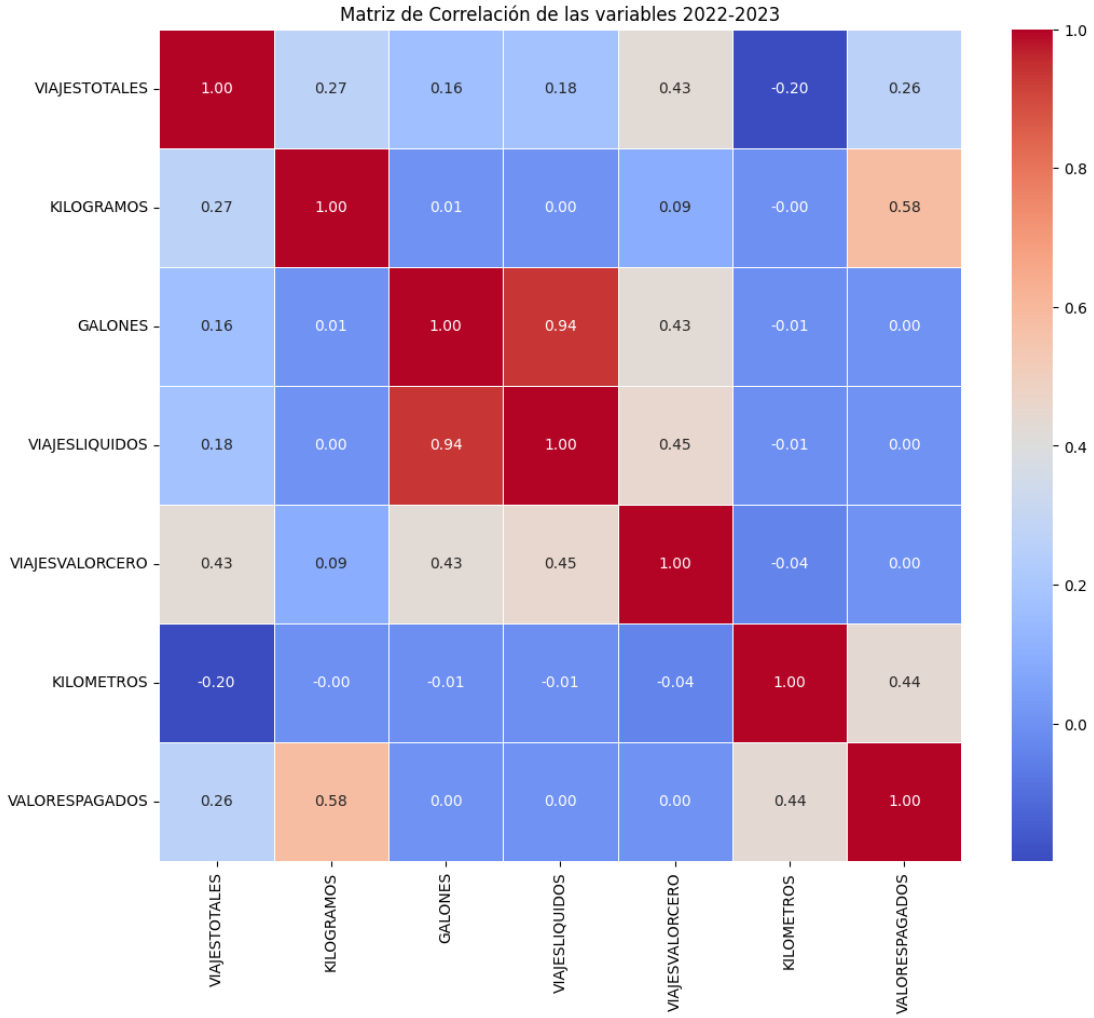


Es interesante es reconocer que no importa el tipo de carga o código de mercancía, lo que muestra que no son variables representativas para la estimación de la correlación, adicional la mayor correlación se encuentra entre las variables VIAJESLIQUIDOS y GALONES que tiene una correlación de 0.97, por lo tanto se toma para el análisis las variables de VIAJESTOTALES, KILOMETROS, GALONES, VIAJESLIQUIDOS, VIAJESVALORCERO, KILOGRAMOS,

VALORESPAGADOS, adicional sobre las variables VALORESPAGADOS, KILOMETROS, KILOGRAMOS y VIAJESTOTALES se eliminan los valores cero, nulos y negativos, pasando de 13.627.308 registros a 9.056.738, en la siguiente figura 4. se muestra la matriz, calculada nuevamente, con las variables de interés, adicional a esto, dado que se reconoce como grandes cambios en los costos asociados a lo largo de los años, se registre los datos a los años 2022 y 2023, bajando a un total de registro de 3191001.

Figura 4

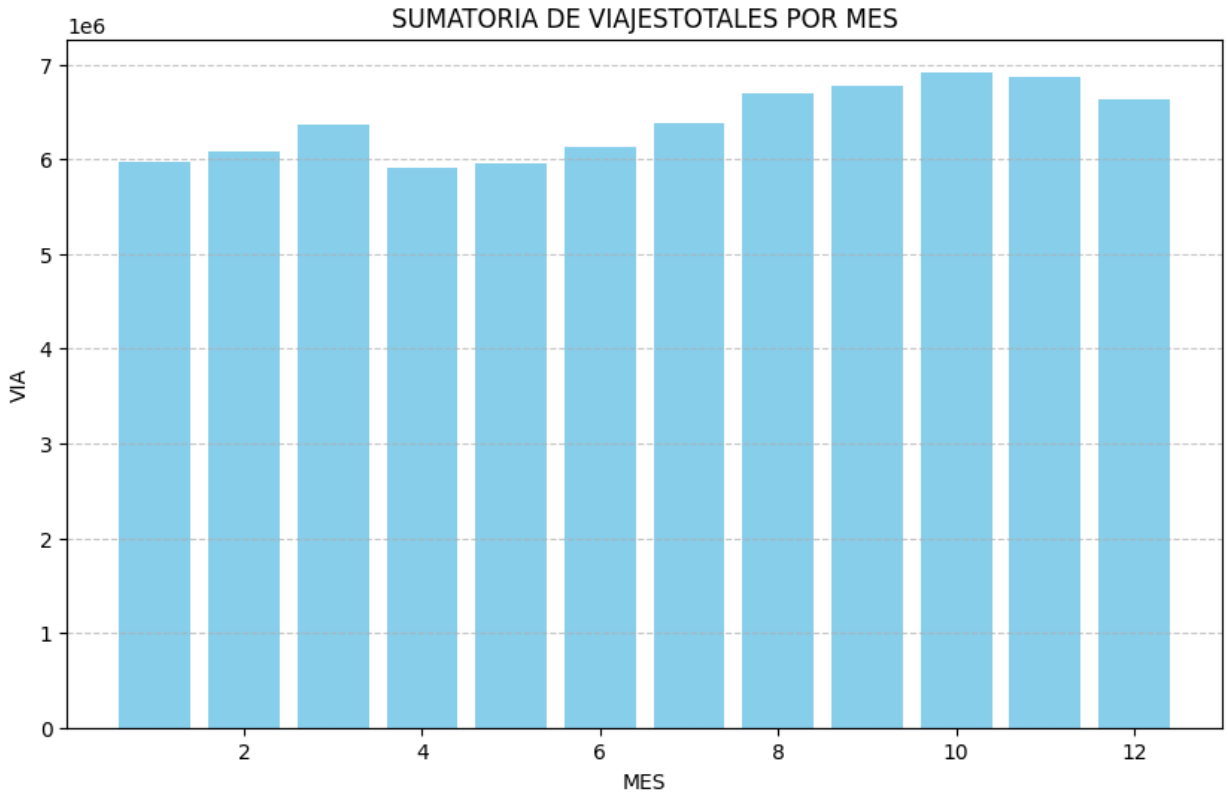
Correlación entre las Variables de Interés



Adicional a esto se quitan los outliers, utilizando el rango intercuartil (IQR), bajando a un total de datos de 2.563.749, registros obteniendo así la matriz de correlación de la figura 4, se observa una mayor correlación entre los valores de VALORESPAGADOS, con KILOGRAMOS y KILOMETROS, siendo esto coherente para un proceso de logística, donde a mayor peso y mayor distancia se obtiene finalmente un mayor costo por el viaje.

Figura 5

Viajes Totales por Mes 2015 a 2023

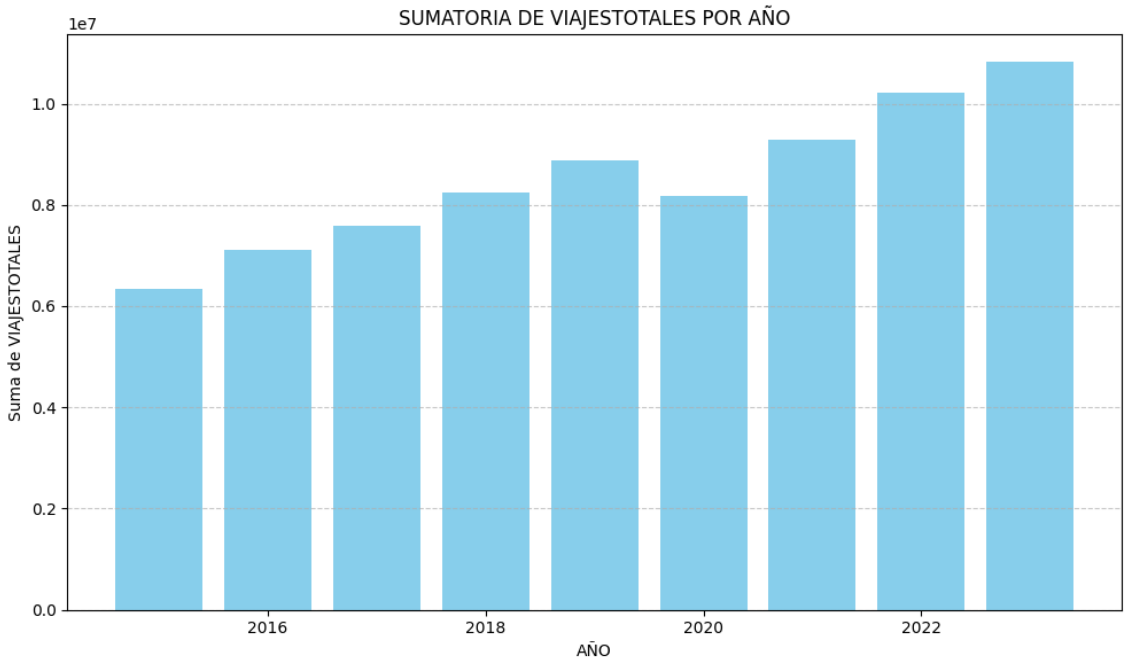


Con los valores total sin la eliminación de los valores nulos, se quiere explorar la representación de los datos en este caso se grafica la totalidad de los viajes por mes, desde Enero(1) hasta Diciembre(12), con un total de 75.642.468 de viajes, en la figura 5 se observa el resultado por mes , en la cual se puede deducir que un comportamiento particular en el primer

semestre con menor cantidad total de viajes que el segundo semestre, se observa una caída importante en el mes de Marzo – Abril, esto se podría explicar por la semana santa que usualmente se ubica en esta temporada, el mes con mayor cantidad de viajes es octubre, presumiblemente por dinámicas como preparatorio para abastecer el mes de Diciembre, el detalle se puede ver en el Anexo B.

Figura 6

Viajes Totales por Año Entre 2015 a 2023



De igual se busca la distribución por años, en este caso en la siguiente imagen, se observa el número creciente de viajes, esto por el perfeccionamiento de las empresas de transporte en el reporte de viajes, adicional al crecimiento económico que se ha experimentado en los últimos años, de igual forma se puede apreciar una caída en el año 2020, consecuencia de la pandemia de Covid-19.

Otro análisis que se plantea es determinar la frecuencia de viajes entre ciudades, a través de un mapa de calor para determinar el intercambio entre los diferentes departamentos, en la siguiente tabla se puede observar el total de viajes por departamento desde el origen y el conteo desde el destino.

Tabla 9

Total de Viajes en los Principales Departamentos Origen Destino

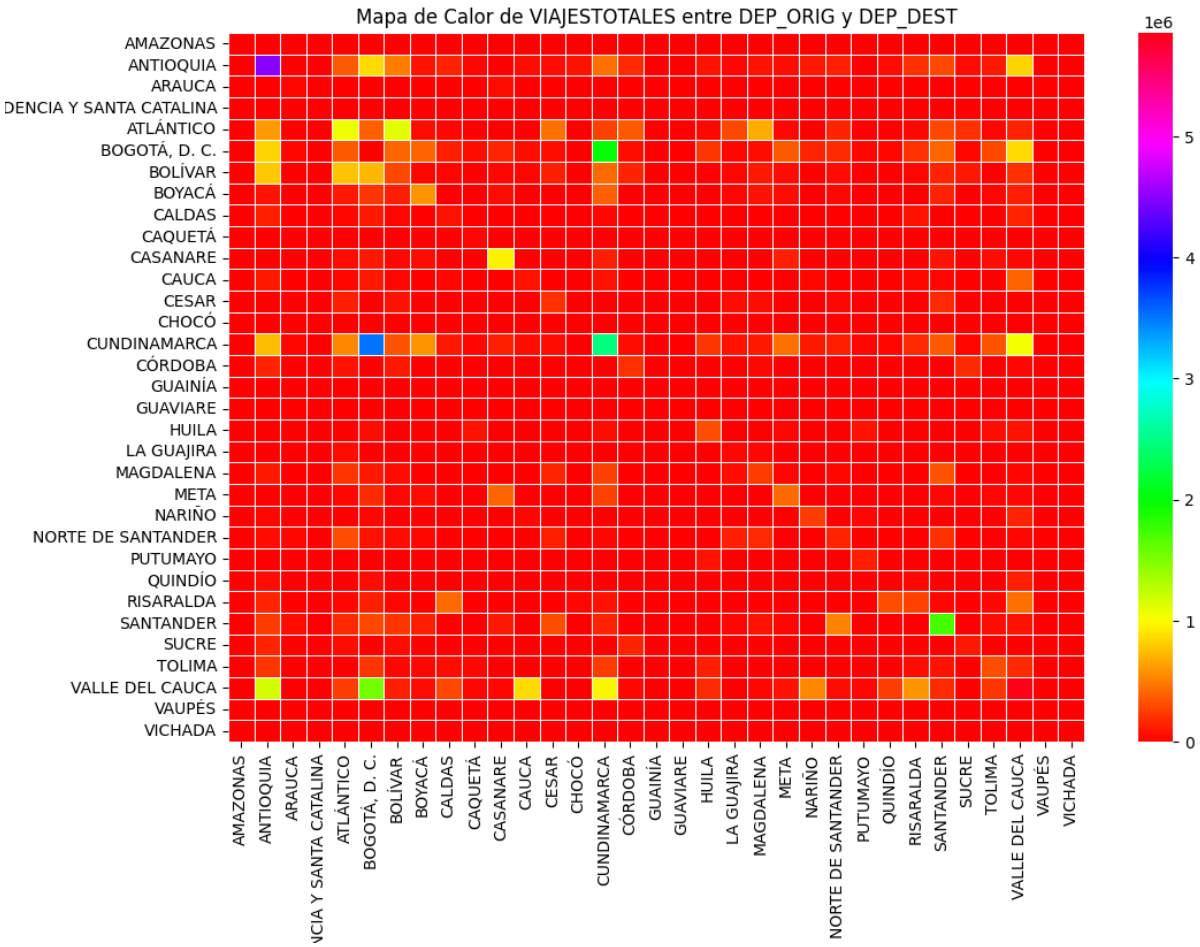
Departamento Origen	Total viajes departamento origen	Departamento Destino	Total viajes departamento destino
VALLE DEL CAUCA	13406933	VALLE DEL CAUCA	10845932
CUNDINAMARCA	11777908	ANTIOQUIA	10161808
ANTIOQUIA	9330727	BOGOTÁ, D. C.	8784968
BOGOTÁ, D. C.	7360401	CUNDINAMARCA	8320896
ATLÁNTICO	6284093	ATLÁNTICO	4646906
SANTANDER	4285983	SANTANDER	4554658
BOLÍVAR	4147649	BOLÍVAR	3631088
BOYACÁ	2155075	BOYACÁ	2221628
RISARALDA	2031693	CASANARE	2008264
TOLIMA	1719922	META	1842502

En el origen se puede determinar los principales polos de manufactura y desarrollo, los cuales abastecen a los demás departamentos, en el caso del departamento de destino, puede ser el mismo departamento y aparecen otros como CASANARE y META, donde llegan gran cantidad de viajes, en la figura siguiente se observa el mapa de calor de los viajes entre departamentos.

En la figura 7 se observa la frecuencia de viajes entre departamentos, así se determina que hay mayor frecuencia en viajes intradepartamental, esto muestra que estas cadenas circulación, priorizan la atención del departamento y lo demás se hace para fuera del departamento, para mejorar el análisis donde el origen destino sea a un departamento diferente, obtenido la siguiente gráfica.

Figura 7

Mapa de Calor Viajes entre Departamentos, Destino/Origen de 2015 a 2023

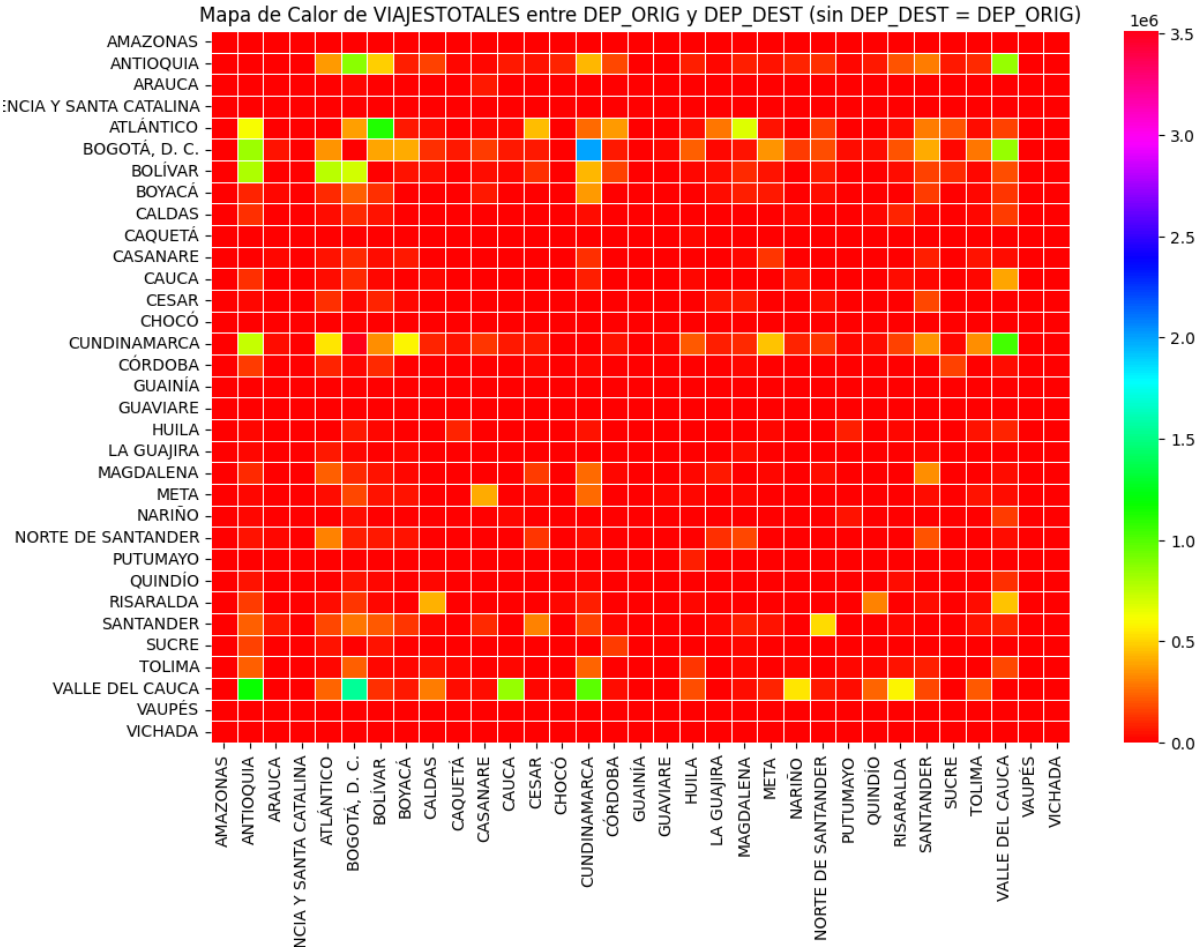


Para esta nueva figura 8 se observan otras dinámicas muy interesantes, siendo los departamentos con mayor frecuencia en Valle del Cauca, Antioquia, Bogotá, Cundinamarca,

Bolívar y Atlántico, aquí se confirma los principales polos de desarrollo y los lugares donde ubican los 4 principales puertos de Colombia (Puerto de Buenaventura, Puerto de Cartagena, Puerto de Barranquilla, Puerto de Santa Marta).

Figura 8

Mapa de Calor Viajes entre Departamentos Diferentes, Destino/Origen sin Repetir entre 2015 a 2023



Otro elemento de interés para reconocer es el tipo de carga que se transporta, en la tabla 10 se observa los códigos y su descripción, además en la gráfica de la frecuencia, siendo los productos varios la mayor frecuencia y que esto mostraría la necesidad en la norma de aumentar

la descripción de los productos, le sigue el paqueteo, siendo este muy relacionado con los correos y el tránsito de productos y el tercero se ubica el contenedor vacíos, siendo este una oportunidad como indicador de reducir este número, al establecer y una mayor conexión con los originadores de la carga para aprovechar esta capacidad ociosa del sistema.

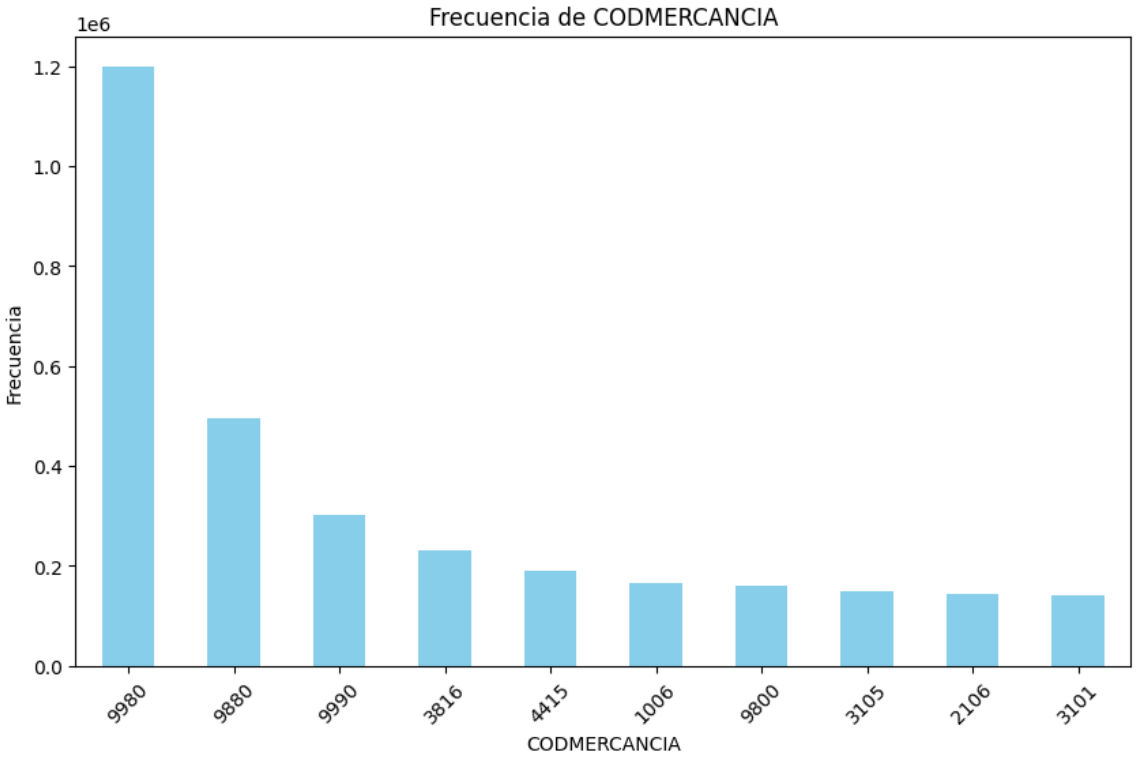
Tabla 10

Descripción Códigos de 10 Principales de Mercancía

Código	Descripción
9980	PRODUCTOS VARIOS
9889	MISCELANEOS CONTENIDOS EN PAQUETES (PAQUETEO)
9990	CONTENEDOR VACIO
3816	CEMENTOS; MORTEROS; HORMIGONES Y PREPARACIONES SIMILARES; REFRACTARIOS;
4415	CAJONES; CAJAS; JAULAS; TAMBORES Y ENVASES SIMILARES; DE MADERA; CARRETES
1006	ARROZ
9800	CONJUNTOS INDUSTRIALES EXPORTADOS DE CONFORMIDAD CON EL REGLAMNETO (CEE) N? 518/
3105	ABONOS MINERALES O QUIMICOS; CON DOS O TRES DE LOS ELEMENTOS FERTILIZANTES NI
2106	PREPARACIONES ALIMENTICIAS NO EXPRESADAS NI COMPRENDIDAS EN OTRAS PARTIDAS
3101	ABONOS DE ORIGEN ANIMAL O VEGETAL; INCLUSO MEZCLADOS ENTRE SI O TRATADOS QUIMICAMENTE

Figura 9

Frecuencia de los 10 Principales Códigos de Mercancía



Se observa en la distribución de la figura 9, confirmado el hallazgo de la gran cantidad de contenedores vacíos, siendo esto una oportunidad de mejora futura, además de análisis en detalle para reconocer entre el ciudad y departamentos que elementos se pueden dinamizar para reducir estos contendores, adicional también como mejorar detallar más en profundidad el contenido.

Resultado Objetivo 2

Dado los resultados del EDA se establece una estrategia para reconocer el mejor modelo, en este caso se hace pruebas por tres tipos de modelos, ya que no es concluyente la correlación encontrada, el detalle se puede ver en el Anexo A.

- Regresión Lineal: Se puede observar una relación lineal entre varias variables con la variable objetivo VALORESPAGADOS, una regresión lineal podría dar una aproximación sencilla para el modelo.
- Árboles de Decisión o Random Forest: Dado la gráfica de correlación, las relaciones entre las variables no son tan lineales, y es posible que se tengan interacciones complejas, es por esto que los árboles de decisión pueden facilitar la construcción de un modelo más preciso, ya que pueden capturar, relaciones no lineales y son robustos a las interacciones entre variables.
- Gradient Boosting Regresor (GBR): Este se prueba con la posibilidad de mejorar la precisión del modelo, gracias a su potencia para mejorar relaciones complejas entre las variables

Para la separación de los datos en train y test se selecciona la librería *sklearn.model_selection*, donde se importa los métodos *train_test_split*, estos facilitan la separación y las pruebas del modelo. En la fase de entrenamiento se pueden comprobar los tres tipos de modelo.

Como condición de mejora para manejar datos se hace una normalización de todos los valores que participan en la construcción de los modelos.

Modelo ee Regresión Lineal

En este se utiliza la librería *LinearRegression* de *sklearn.linear_model* y para validar el resultado se utiliza la librería *mean_squared_error* de sklearn.metrics, con un separación de entrenamiento y prueba de 80% y 20% correspondientemente. Donde el resultado del modelo es el siguiente

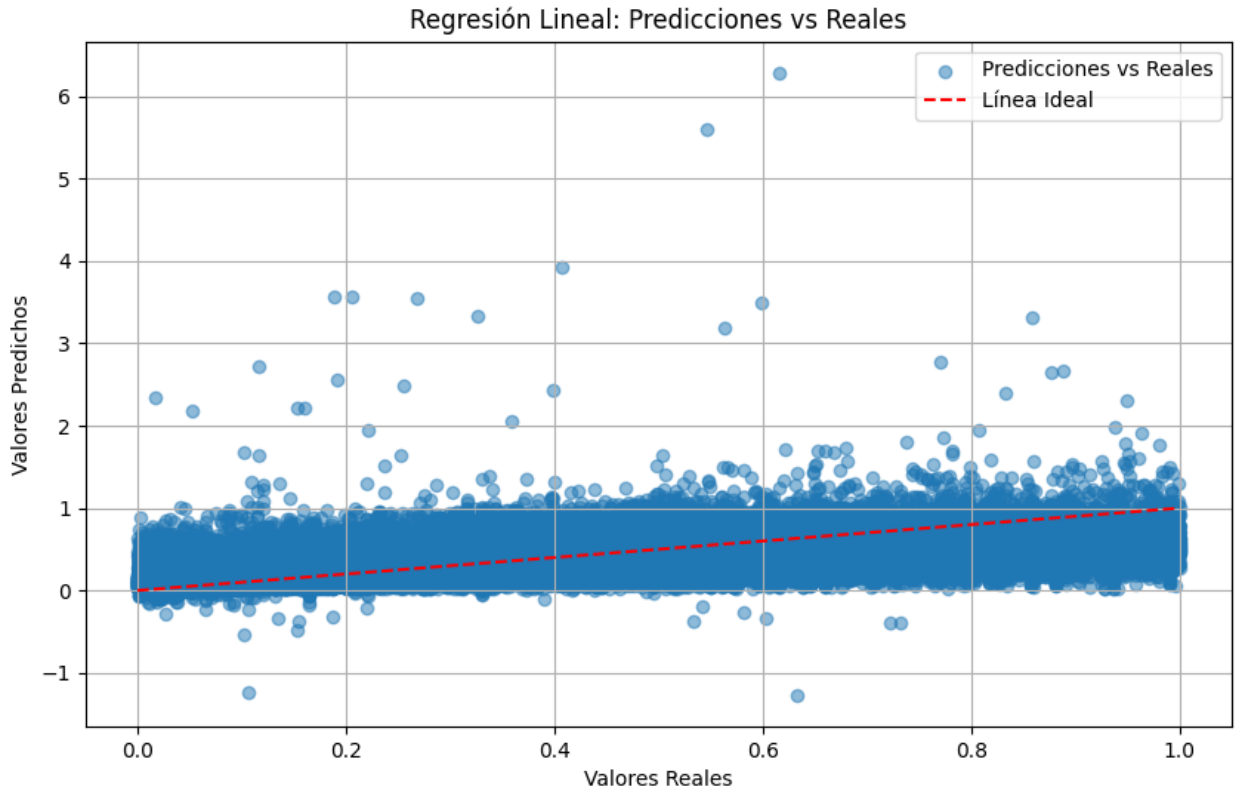
- R2 score: 0.58809953

- Mean Absolute Error (MAE): 0.09846416
- Mean Squared Error (MSE): 0.02040918
- Root Mean Squared Error (RMSE): 0.14286071. Es así se que obtendría la siguiente ecuación

*y =0.01213088 + (6.93680192*x1) + (0.59642706*x2) + (0.43355027*x3) + (2.39448328*x4) + (-8.95068901*x5) + (0.44611243*x6))*

Figura 10

Resultado Modelo de Regresión Lineal: Predicciones vs Valores Reales



En la figura 10, se observa el resultado donde se reconoce una gran distribución y una pequeña tendencias, sin embargo esta no logra capturar toda la variabilidad de los datos, en este caso este modelo no es óptimo para modela, ya que en el contraste de valores reales vs predichos

existe una gran distribución, con varios valores fuera de la agrupación principal dejando grandes dudas de su capacidad para explicar los datos, se confirma esto con el R2 Score, de tan solo el 58.8%, siendo un valor que no serviría para predecir los valores, siendo así que tipo de modelo de regresión lineal no es el adecuado para este trabajo, es por eso que se hacen otras pruebas con modelos más complejos que pueden recoger toda la variabilidad.

Modelo Random Forest Regressor

Para este modelo se utilizan las librerías de RandomForestRegressor de sklearn.ensemble y nuevamente *mean_squared_error* de sklearn.metrics, con los mismos niveles de entrenamiento y prueba, donde se obtiene el siguiente resultado

- R-squared (R^2): 0.76777886
- Mean Absolute Error (MAE): 0.07038581
- Mean Squared Error (MSE): 0.01150628
- Root Mean Squared Error (RMSE): 0.10726734

Un buen resultado para el R2, con MSE ajustado, con respecto al modelo anterior, se observa una mejora una posible mejora es revisar los outliers y la normalización una normalización de la variable objetivo.

Modelo Gradient Boosting Regressor

Para este modelo se utilizan las librerías GradientBoostingRegressor, de sklearn.ensemble, donde logra obtener los siguiente resultados

- R-squared (R^2): 0.75418800
- Mean Absolute Error (MAE): 0.07452025
- Mean Squared Error (MSE): 0.01217969
- Root Mean Squared Error (RMSE): 0.11036165

Estos resultados no son tan prometedores, tiene MSE bajo, requier mejorar los ajustes y revisar. El valor es similar con el modelo de Random Forest Regressor, sin embargo estes un poco mas liviana y conveniente para implementar.

Resultado Objetivo 3

El resultado y detalle del código se puede ver en el Anexo B, donde se detalla el análisis el reentrenamiento del modelo basado en los resultados anteriores, que permite su optimización.

Reentrenamiento

Ahora para la validación de los modelos se establece dos criterios iniciales para mejorar el resultado del error, no se quitan los outliers con el método intercuartil (IQR). Donde se obtiene los siguientes resultados, en el cual se obtiene le siguiente resultado

Tabla 11

Nuevos Modelos con Resultados Incluyendo los Outliers

Parámetro	Regresión Lineal	Random Forest	Gradient Boosting
R-squared (R^2)	0.61527719	0.96246783	0.93711103
Mean Absolute Error (MAE)	0.00069182	0.00022897	0.00035985
Mean Squared Error (MSE)	0.00001345	0.00000131	0.00000220
Root Mean Squared Error (RMSE)	0.00366785	0.00114562	0.00148295

Se mejoran los resultados significativamente, gracias a la consideración de incluir valores outliers, luego se concluye que estos cuentan con gran cantidad de información para el modelo que no debe ignorarse, a su vez el mejor desempeño lo tiene el modelo Random Forest Regressor sin embargo el resulto de este es de gran peso y puede dificultarse su aplicación, en su lugar el

modelo Grandient Boosting Regressor es más liviano y fácil de utilizar, siendo este modelo el elegido para su optimización.

Optimización

Dado que el modelo de Gradient Boosting Regressor tuvo el mejor desempeño, se busca hacer una optimización a través de la variación de los hiperparámatros con el fin de encontrar un modelo optimizado, en este el parámetros se definen antes del entrenamiento, estos se conocen como hiperparámetros, que controlan aspectos como arquitectura, la regulación y la tasa de aprendizaje, el modelo se controla con el error cuadrático medio (MSE), en la siguiente tabla se establecen los parámetros y su descripción

Tabla 12

Descripción Parámetros de la Optimización

Parámetro	Rangos	Random Forest
n_estimators	100 - 500	Número de árboles en el modelo. Un valor más alto generalmente mejora el ajuste, pero aumenta el tiempo de entrenamiento y el riesgo de sobreajuste.
learning_rate	0.01 - 0.3	Factor de aprendizaje que controla cuánto contribuye cada árbol al modelo final. Valores más pequeños hacen que el entrenamiento sea más lento pero más preciso.
max_depth	3-10	Profundidad máxima de los árboles. Controla la complejidad de los árboles. Valores más altos pueden capturar más relaciones, pero también aumentar el sobreajuste.

min_samples_split	2-20	Número mínimo de muestras requeridas para dividir un nodo. Valores más altos hacen que el modelo sea menos flexible, reduciendo el riesgo de sobreajuste.
min_samples_leaf	1-20	Número mínimo de muestras requeridas en una hoja. Ayuda a suavizar el modelo y evita que los árboles se ajusten demasiado a los datos de entrenamiento.
subsample	0.5 - 1.0	Proporción de muestras utilizadas para entrenar cada árbol. Un valor menor a 1.0 introduce aleatoriedad, lo que puede mejorar la generalización.
max_features	0.5 - 1.0	Proporción de características consideradas al buscar la mejor división. Valores más pequeños introducen aleatoriedad y reducen el riesgo de sobreajuste.
min_samples_leaf	0.5 - 1.0	Número mínimo de muestras requeridas en una hoja. Ayuda a suavizar el modelo y evita que los árboles se ajusten demasiado a los datos de entrenamiento.

Con esto se obtiene el siguiente resultado

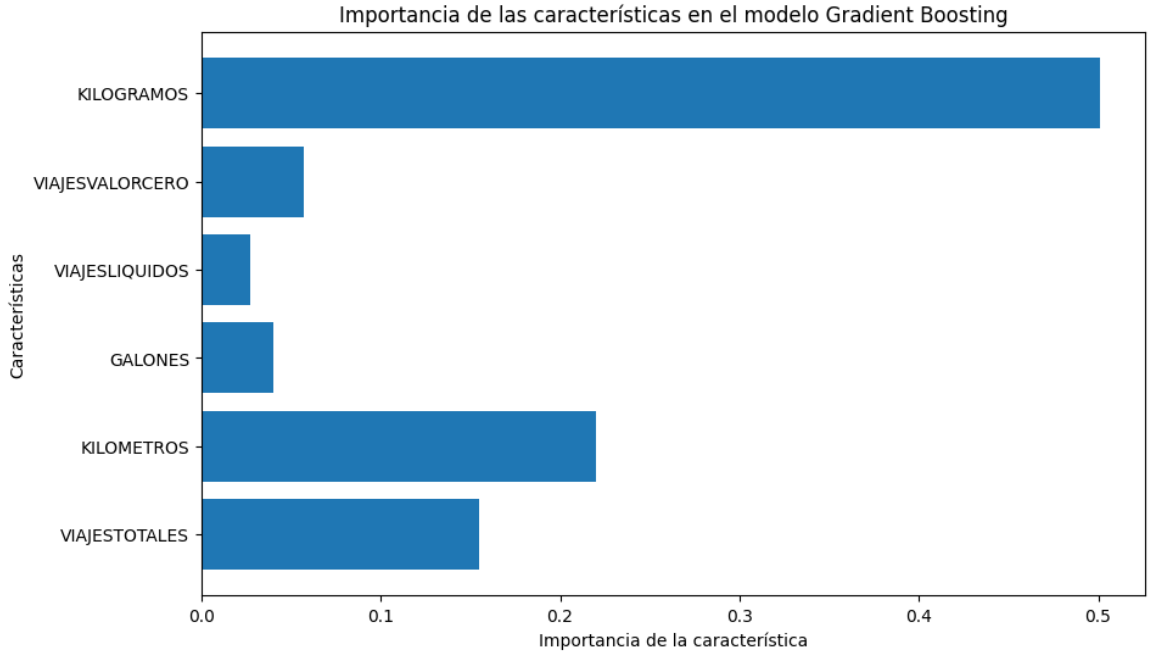
- R-squared (R^2): 0.9685597030
- Mean Absolute Error (MAE): 0.0002408971
- Mean Squared Error (MSE): 0.0000010994
- Root Mean Squared Error (RMSE): 0.0010485308

Ahora se detallan algunas gráficas para ver como mayor precisión el resultado, en el caso de la figura 11, en cuanto a las características, la que tiene mayor peso es los KILOGRAMOS,

luego este es un factor determinante para predecir el costo, lo que tiene sentido en un contexto de transporte de carga, donde influye directamente, ya que a mayor peso se requiere unas características superiores en cuanto al vehículo.

Figura 11

Importancia de las Variables en el Modelo Final

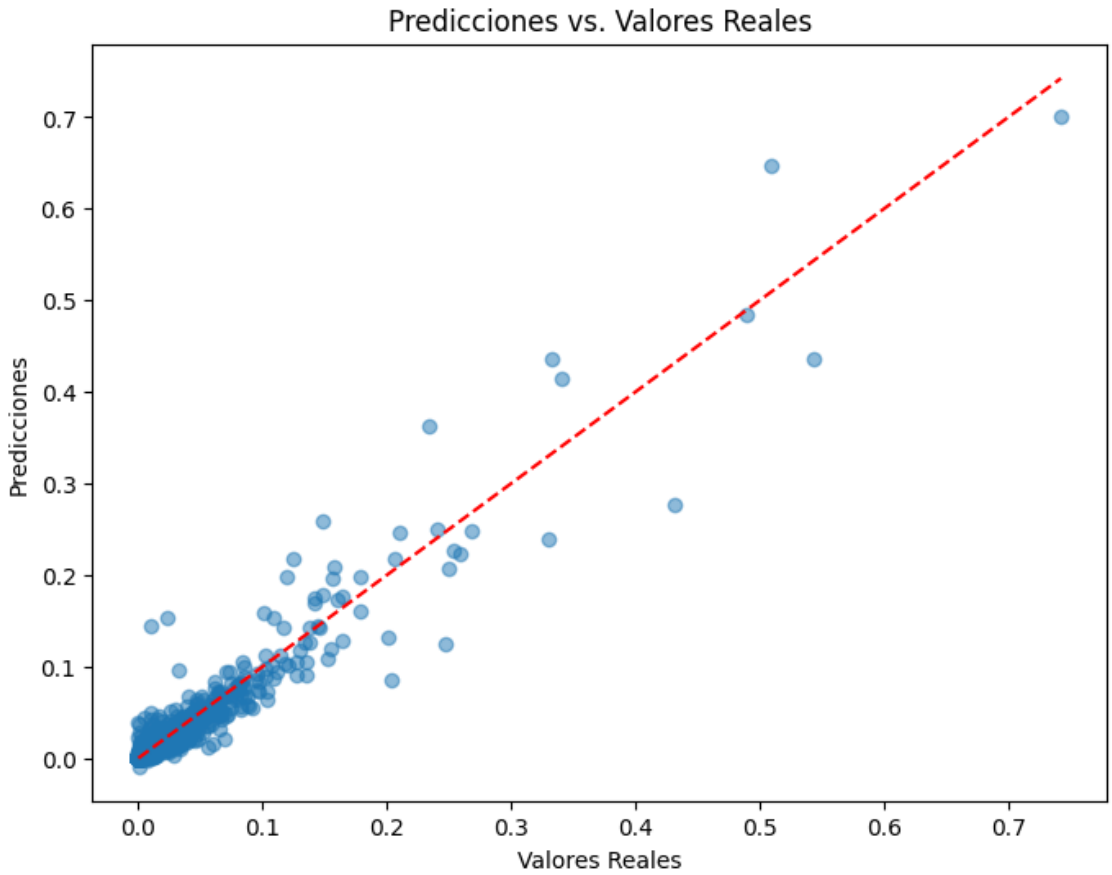


Los KILOMETROS, también es una característica relevante, siendo también un factor crítico que puede aumentar los costos, que se relacionan con mayor cantidad de tiempo, combustible, viatico y fletes, por otro lado las variables VIAJESTOTALES y VIAJESVALORCERO, tiene su influencia y aportar a los costos, pero no son tan relevantes, en el caso final de VIAJESLIQUIDOS, GALONES, la influencia es menor y aporta como las demás variables para explicar el comportamiento del sistema. Se estable así que el costo del viaje depende en su gran mayoría del peso, aumentando el valor de la carga y de la distancia, que implica mayores costos y tiempo para su entrega.

En cuanto a la predictibilidad del modelo, en la figura 12, se reconoce que los valores de distribución están cercanos a la diagonal, siendo este un excelente resultado, la cercanía sugiere que el modelo es preciso para lo datos con los que se está probando. Se puede observar que, en valores más altos, la desviación es mayor, no hay un ajuste tan claro en esta zona, esto debido a que seguramente se tienen outliers que modifican el resultado.

Figura 12

Predictibilidad del Modelo Final

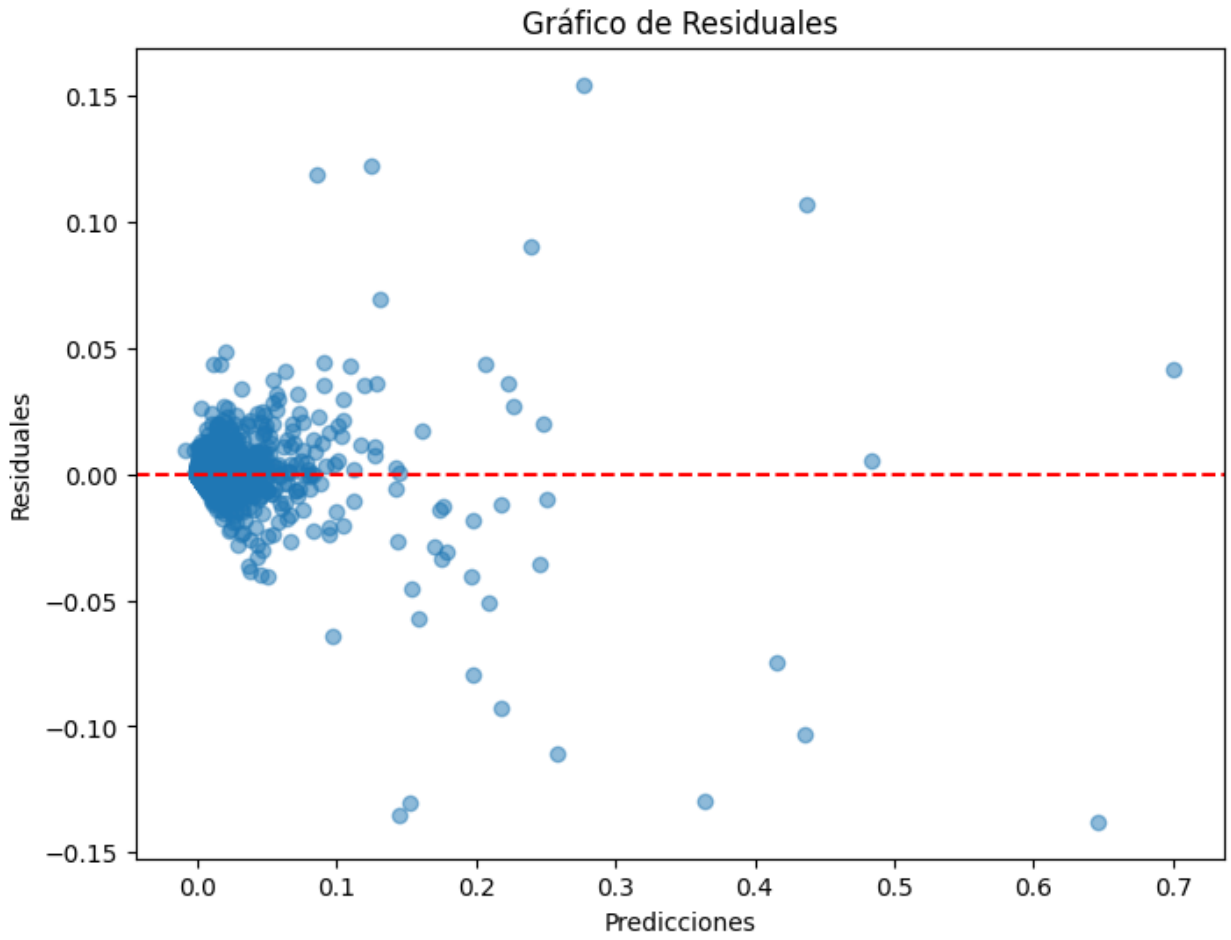


La mayoría de los puntos están en una tendencia ascendente, lo que puede sugerir que el modelo captura la relación general entre las características de entrada y la variable objetivo. En valores bajos los errores son pequeños, dada la concentración alrededor de la línea diagonal, en

valores más pequeños el modelo es más preciso, cual el valor aumento el modelo pierde asertividad sin embargo el lo suficientemente buena para considera el modelo.

Figura 13

Gráfico de Residuos del Modelo Final



Para los residuos se observa en la gráfica una concentración de residuos en los valores bajos, en su mayoría se concentran cerca de cero, lo que indica que el modelo tiene predicciones razonables para valores bajos de la variable objetivo. La concentración es una buena señal, ya que se puede estimar que el modelo no tiene un sesgo sistémico en los valores bajos. A medida que las predicciones aumentan, se observa una mayor dispersión en los residuos, tanto para

valores positivos como negativos, esto muestra que el modelo tiene dificultad para predecir valores altos, resultando en errores grandes. Puede establecerse que el modelo no captura toda la variabilidad de los datos en rangos altos. El residuo aumenta con la predicción, lo que puede sugerir una heteroscedasticidad, el error no es constante a lo largo de las predicciones. No se reconoce un patrón claro en los residuos, lo que se puede considerar positivo, esto indica que no hay un sesgo direccional fuerte y que estos parecen aleatorios.

Como conclusión del modelo, se logra reconocer un modelo preciso, las métricas muestran que la predicción es precisa, con un R2 de 0.92, indica que el modelo explica la variabilidad en los valores a pagar. Las variables más relevantes son KILOGRAMOS y KILOMETROS, el modelo esta alineado con la lógica del transporte de carga, donde el peso y la distancia, son factores determinantes en los precios y es consistente con la realidad del sector de carga en el país.

Se pueden establecer mejoras, en cuanto a los valores a pagar altos, se puede ajustar el modelo para mejorar la precisión a través de un análisis más exhaustivo y considerar elementos como la transformación de los datos y características para aplicar otros tipos de análisis. El modelo es eficiente en predecir valores bajos, la dispersión de los residuos indica que el modelo es bueno para valores a pagar bajos, que en el sentido práctico es la mayoría de los viajes que se puede encontrar en los datos.

Conclusiones

Este trabajo logra aprovechar información pública disponible histórica para crear un Modelo de Machine Learning que permite predecir con precisión los costos de transporte en Colombia, gracias a el desarrollo de un flujo de procesamiento de información que facilita el alistamiento para el entrenamiento, optimización y validación. Estos datos históricos permitieron a su vez una caracterización detallada de la carga en Colombia, identificando características claves y relaciones significativas que permiten reconocer la importancia de la información para el sector transporte.

La identificación de variables cable como el peso de la carga, la distancia recorrida y la cantidad de viajes, demuestran ser determinantes en la estimación de los costos de carga, siendo estos los que se deben considerar con mayor peso a la hora de reconocer nuevos elementos y el análisis de nuevos datos para mejorar las estimaciones.

Para el desarrollo de proyectos de ciencias de datos y analítica, es necesario crear un proceso adecuado para el tratamiento, limpieza y transformación de los datos permite desarrollar un modelo eficiente para la estimación de los costos de transporte en Colombia, dada la gran variabilidad de datos y elementos a considerar que se han registrado desde el 2015 hasta 2023.

Se probaron diferentes modelos tales como regresión línea, Random Forest y Gradient Boosting Regressor, con resultados de carácter medio, donde fue necesario realizar diversidad de pruebas tales como normalización y outliers para llegar a un modelo con unas características interesantes sobre el cual se puedo realizar la optimización.

El mejor modelo encontrado fue el Gradient Boosting Regressor con un nivel de precisión del 96.8% en la clasificación de la información, este modelo logra una precisión suficiente en la predicción de costos de transporte, demostrando ser una herramienta valiosa para

las empresas de logística. Las variables de peso en kilogramos y la distancia entre ciudades, son las más adecuadas para comprensión de la complejidad del problema.

Este trabajo permite sentar las bases para otro tipo de estudios como clasificación de tipo de carga (Nama et al., 2021), segmentación de rutas de transporte (Akbari & Do, 2021), análisis de tendencias temporales (Moscoso-López et al., 2021) y predicción de la demanda de transporte en Colombia (Plakandaras et al., 2019).

Finalmente, con este tipo de propuestas se demuestra el impacto que puede tener el Machine Learning en el diseño de Políticas Públicas, en el aprovechamiento de información disponible que permite mejorar la toma de decisiones para las empresas, entidades territoriales y organismos de orden nacional, para mejorar el sector de carga en el país.

Recomendaciones

Se recomienda utilizar los resultados y los datos recolectados para avanzar en el entendimiento del sector transporte a través de estudios tipo caracterización de carga, comprender los patrones del transporte de carga en Colombia, como qué tipos de carga se transportan más, cuáles son las rutas más utilizadas para ciertas cargas y cómo se distribuyen en las diferentes regiones. Su utilización se puede aplicar para planificación de infraestructura vial, optimización de recursos logísticos, identificación de sectores con alta demanda en transporte.

La segmentación de rutas de transporte permite, dividir el país en rutas claves, con cierta homogeneidad para el análisis específico que permitan la optimización de los recursos de transporte, facilita así para los tomadores de decisiones, el diseño de rutas óptimas, gestión de tráfico pesado y mejoras logística en cadenas de abastecimiento regionales.

Aplicar estudios de tendencias temporales, permitiría a largo plazo, predecir cambios en los tipos de carga, pesos, rutas preferidas, y demanda temporal, el análisis a corto y largo plazo, permite tomar decisiones informadas y anticipar la evolución del sector transporte. Su uso se puede aplicar en alistamiento de temporadas altas y bajas, maximizar el uso de contenedores vacíos, predicción de picos de demanda y adaptación de las empresas de transporte a los patrones económicos en cada región.

También se puede utilizar para la predicción de demanda de transporte, incluyendo variables económicas y sociales, ayuda a fomentar políticas económicas, inversión en infraestructura, ajustes a tarifas de asociadas al transporte, capacidad de crecimiento y proyección sectorial.

Identificación de cadenas productivas regionales, conexiones agrícolas, mercados, zonas industriales, puertos y aeropuertos para establecer cadenas de valor, optimizar su

funcionamiento, facilita el comercio y aprovechamiento de la cadena, establecer alianzas, búsqueda de alternativas de proveedores de insumos, y con esto fortalecer la competitividad regional, la promoción de inversión en sectores estratégicos, y reducción del costo logístico.

Este tipo de proyectos facilita y fomenta el uso de datos públicos, como caso particular se participó en concurso Datos a la U (Resultados Convocatoria "Datos a la U" | Datos Abiertos Colombia), donde se logro avanzar a la segunda etapa, obteniendo un puntaje final de 93 puntos, y logrando publicar el modelo desarrollado davidUrr/DatosalaU_R3_G17: Repositorio del repositorio para el concurso Datos a la U https://www.datos.gov.co/stories/s/Actualidades-del-concurso-Datos-a-la-U/wn73-87k7/, esta divulgación facilita afrontar otros tipo de retos con información pública, esto facilita la toma de decisiones del gobierno y prioriza los recursos para sectores y proyectos claves.

Este trabajo pudo ser divulgado en el XXIX Verano de la Investigación Científica y Tecnológica del Pacífico, en Puerto Vallarta México, en el Anexo C, se puede ver una copia del poster, adicional en el Anexo D, se muestra un video con la presentación de este trabajo.

Finalmente, este tipo de proyecto ayuda a las entidades gubernamentales a la formulación, evaluación y actualización de Políticas Públicas, para diseñar, evaluar y ajustar las iniciativas y hacer seguimiento, promoviendo políticas más justas, basadas en evidencia, alineadas con las necesidades actuales y futuras del sector.

Referencias Bibliográficas

Akbari, M., & Do, T. N. A. (2021). A systematic review of machine learning in logistics and supply chain management: current trends and future directions. *Benchmarking: An International Journal*, *28*(10), 2977–3005. https://doi.org/10.1108/BIJ-10-2020-0514

Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. J. (2020). *A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science* (pp. 3–21). https://doi.org/10.1007/978-3-030-22475-2_1

Amanpreet, S., Narina, T., & Aakanksha, S. (2016). A review of supervised machine learning algorithms. *3rd International Conference on Computing for Sustainable Global Development (INDIACom)*. https://ieeexplore.ieee.org/abstract/document/7724478

Boukerche, A., & Wang, J. (2020). Machine Learning-based traffic prediction models for Intelligent Transportation Systems. *Computer Networks*, *181*, 107530. https://doi.org/10.1016/j.comnet.2020.107530

Boyer, R. C., Scherer, W. T., & Smith, M. C. (2017). Trends Over Two Decades of Transportation Research. *Transportation Research Record: Journal of the Transportation Research Board*, *2614*(1), 1–9. https://doi.org/10.3141/2614-01

Colfecar. (2022). *Análisis del sector transporte de carga: Productividad, eficiencia y principales cifras económicas*. Colfecar. https://www.colfecar.org.co/wp-content/uploads/An%C3%A1lisis%20del%20sector%20transporte%20de%20carga.pdf

de la Torre, R., Corlu, C. G., Faulin, J., Onggo, B. S., & Juan, A. A. (2021). Simulation, Optimization, and Machine Learning in Sustainable Transportation Systems: Models and Applications. *Sustainability*, *13*(3), 1551. https://doi.org/10.3390/su13031551

Farchi, F., Farchi, C., Touzi, B., & Mabrouki, C. (2023). A Comparative Study on AI-Based Algorithms for Cost Prediction in Pharmaceutical Transport Logistics. *Acadlore Transactions on AI and Machine Learning*, *2*(3), 129–141. https://doi.org/10.56578/ataiml020302

Gonzalez, L. C. (2022). *Retos y tendencias empresariales con relación a la optimización de costos logísticos para una efectiva logística integral* [Contaduría Pública]. Universidad Abierta y a Distancia UNAD.

Martinez, C. (2022). *Colombia y los retos en el transporte de carga*. The Logistics World. https://thelogisticsworld.com/actualidad-logistica/646219

Mintransporte. (2024). *Normatividad Ministerio de Transporte - RNDC*. Página Oficial Ministerio de Transporte Colombia. https://plc.mintransporte.gov.co/RNDC/Normatividad

MINTRANSPORTE. (2024). *Registro Nacional de Despachos de Carga*. Página Oficial Ministerio de Transporte. https://rndc.mintransporte.gov.co/MenuPrincipal/tabid/204/language/es-MX/Default.aspx?returnurl=%2f

Morabit, M., Desaulniers, G., & Lodi, A. (2023). Machine-Learning–Based Arc Selection for Constrained Shortest Path Problems in Column Generation. *INFORMS Journal on Optimization*, *5*(2), 191–210. https://doi.org/10.1287/ijoo.2022.0082

Moscoso-López, J. A., Urda, D., Ruiz-Aguilar, J. J., González-Enrique, J., & Turias, I. J. (2021). A machine learning-based forecasting system of perishable cargo flow in maritime transport. *Neurocomputing*, *452*, 487–497. https://doi.org/10.1016/j.neucom.2019.10.121

Nama, M., Nath, A., Bechra, N., Bhatia, J., Tanwar, S., Chaturvedi, M., & Sadoun, B. (2021). Machine learning-based traffic scheduling techniques for intelligent transportation system:

Opportunities and challenges. *International Journal of Communication Systems*, *34*(9). https://doi.org/10.1002/dac.4814

Plakandaras, V., Papadimitriou, T., & Gogas, P. (2019). Forecasting transportation demand for the U.S. market. *Transportation Research Part A: Policy and Practice*, *126*, 195–214. https://doi.org/10.1016/j.tra.2019.06.008

Rajoub, B. (2020). Supervised and unsupervised learning. In *Biomedical Signal Processing and Artificial Intelligence in Healthcare* (pp. 51–89). Elsevier. https://doi.org/10.1016/B978-0-12-818946-7.00003-2

Resolución 377 de 2013, Pub. L. No. Resolución 377, Página Oficial Ministerio de Transporte (2013). https://plc.mintransporte.gov.co/Portals/0/Documentos/Resolucion0000377-2013.pdf?ver=2018-09-20-185238-000

Saravanan, R., & Sujatha, P. (2019). A State of Art Techniques on Machine Learning Algorithms: A Perspective of Supervised Learning Approaches in Data Classification. *Proceedings of the 2nd International Conference on Intelligent Computing and Control Systems, ICICCS 2018*, 945–949. https://doi.org/10.1109/ICCONS.2018.8663155

Sindhu Meena, K., & Suriya, S. (2020). A Survey on Supervised and Unsupervised Learning Techniques. In *Proceedings of International Conference on Artificial Intelligence, Smart Grid and Smart City Applications* (pp. 627–644). Springer International Publishing. https://doi.org/10.1007/978-3-030-24051-6_58

Singh, A., Das, A., Bera, U. K., & Lee, G. M. (2021). Prediction of Transportation Costs Using Trapezoidal Neutrosophic Fuzzy Analytic Hierarchy Process and Artificial Neural Networks. *IEEE Access*, *9*, 103497–103512. https://doi.org/10.1109/ACCESS.2021.3098657

Taylor, R. (2020). *Papers with Code is Expanding to More Sciences!* Página Oficial Medium. https://medium.com/paperswithcode/papers-with-code-is-expanding-to-more-sciences-5d375d10ca3a

van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine Learning*, *109*(2), 373–440. https://doi.org/10.1007/s10994-019-05855-6

Verma, K. K., Singh, B. M., & Dixit, A. (2022). A review of supervised and unsupervised machine learning techniques for suspicious behavior recognition in intelligent surveillance system. *International Journal of Information Technology*, *14*(1), 397–410. https://doi.org/10.1007/s41870-019-00364-0

Apéndices

Apéndice A

Análisis EDA de la Base de Datos

```
# %%
import joblib
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.preprocessing import MinMaxScaler
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
import optuna
from sklearn.metrics import mean_squared_error
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.preprocessing import LabelEncoder


# %%
#El archivo se puede bajar directamente del conjunto de datos
# Carga de la data
#test_data_path = r'C:\Users\durrego\Downloads\UNAD\DatosalaU\Registro_Nacional_de_Despachos_de_Carga_por_Carretera_20241111.csv'
path = 'C:\\Users\\durrego\\Downloads\\UNAD\\DatosalaU\\'
df = pd.read_csv(path+'TODOCARGA.csv',delimiter='|')
#df = pd.read_csv(test_data_path)

# %%
df.info()

# %%
#Se identifica el tipo de variable y cuantos tipos de datos únicos hay por variable, el valor mas frecuente y la cantidad de los valores que hay
unique_values = df.nunique()
data_types = df.dtypes
most_frequent = df.mode().iloc[0]
frequencies = df.apply(lambda col: col.value_counts().iloc[0])

# Crear el resumen
summary = pd.DataFrame({
    'Unique Values': unique_values,
    'Data Type': data_types,
    'Most Frequent': most_frequent,
    'Frequency': frequencies
})
summary.to_csv('C:\\Users\\durrego\\Downloads\\UNAD\\Semestre 2\\Semestre02\\PROYECTO DE GRADO II\\Fase 1 - Documentación\\summary.csv')

# %%
summary

# %% [markdown]
# ### Reconocimiento de variables
#
# Se reconoce en la base de datos un total de 22 variables, donde se reconoce
#
# * Información del vehículo (e.g., configuración)
# * Tipo de operación de transporte
```

```
# * Origen y destino de la carga
# * Tipo y naturaleza de la carga
# * Variables numéricas como el peso en kilogramos, kilómetros recorridos y el valor pagado (VALORESPAGADOS)
#
# Para esto se lleva todas las variables a numéricas, en el caso de las variables tipo texto se hace un labelEncoder para reconocer la relación entre todas estas a través de una correlación

# %%
categorical_columns = df.select_dtypes(include='object').columns
#Se confirma que todas la variables sean tipo string de forma consistente
for col in categorical_columns:
    df[col] = df[col].astype(str)

# Se aplica LabelEncoder a cada columna categórica
label_encoders = {}
for col in categorical_columns:
    le = LabelEncoder()
    df[col] = le.fit_transform(df[col])
    label_encoders[col] = le
df.dtypes

# %% [markdown]
# ## Correlación
#
# Ahora con estas variables se hace un estimación de la correlación

# %%
matriz_corr=df.corr()

plt.figure(figsize=(12, 10))
#Mapa de calor de la correlación
sns.heatmap(matriz_corr, annot=True, cmap="coolwarm", fmt=".2f", linewidths=0.5)
plt.title("Matriz de Correlación de las variables")
plt.show()

# %% [markdown]
# Dado que la variable objetivos es VALORESPAGADOS, se analiza sobre esta la correlación, donde se establece que la guarda mayor relación son las variable de
#
# VIAJESTOTALES y KILOGRAMOS, en este se es de interes la baja correlación de los KILOMETROS, se quiere observa ahora los diferentes datos presentes en esta base de datos y como es su distribución
#
#
#
#

# %%
#df_describe = df.describe()
#df_describe.applymap(lambda x: f"{x:.6f}" if isinstance(x, float) else x)
pd.set_option('display.float_format', '{:.6f}'.format)
df_describe = df.describe()
print(df.describe())
#df_describe.to_csv('C:\\Users\\durrego\\Downloads\\UNAD\\Semestre 2\\Semestre02\\PROYECTO DE GRADO II\\Fase 1 - Documentación\\tabla_describe.csv',float_format='%.6f',sep='|')
#df_describe.to_excel('C:\\Users\\durrego\\Downloads\\UNAD\\Semestre 2\\Semestre02\\PROYECTO DE GRADO II\\Fase 1 - Documentación\\tabla_describe.xlsx')

# %% [markdown]
# ### Limpieza y outliers
# Se observa en las variables KILOGRAMOS, KILOMETROS y VALORESPAGADOS valores en cero, los que no son de interés, por lo tanto se eliminan.
#
# Ademas se observa una gran variabilidad en estos, datos por ende se hace un estimación de los outliers y se quitan en de los datos
```

```
111
112 # %%
113 ### Crear las nuevas columnas MES_2 y ANIO
114 df['MES_2'] = df['MES'].astype(str).str[4:6]
115 df['ANIO'] = df['MES'].astype(str).str[0:4]
116
117 df['ANIO'] = df['ANIO'].astype(int)
118 df['MES_2'] = df['MES_2'].astype(int)
119 filtered_df = df[df['ANIO'].isin([2022, 2023])]
120
121 filtered_df = filtered_df[(filtered_df[['VALORESPAGADOS', 'KILOMETROS', 'KILOGRAMOS',
    'VIAJESTOTALES']] > 0).all(axis=1)]
122 filtered_df.shape
123
124 # %%
125 df1 =
    filtered_df[['VIAJESTOTALES','KILOGRAMOS','GALONES','VIAJESLIQUIDOS','VIAJESVALORCERO','KILOMETROS',
    'VALORESPAGADOS']]
126 columns_to_filter = ['KILOGRAMOS', 'KILOMETROS', 'VALORESPAGADOS']
127
128 # Función para eliminar outliers usando el método IQR
129 def remove_outliers(df, columns):
130     for col in columns:
131         Q1 = df[col].quantile(0.25)  # Primer cuartil
132         Q3 = df[col].quantile(0.75)  # Tercer cuartil
133         IQR = Q3 - Q1  # Rango intercuartílico
134         lower_bound = Q1 - 1.5 * IQR  # Límite inferior
135         upper_bound = Q3 + 1.5 * IQR  # Límite superior
136         # Filtrar valores dentro de los límites
137         df = df[(df[col] >= lower_bound) & (df[col] <= upper_bound)]
138     return df
139
140 # Aplicar la función para eliminar outliers
141 df1_cleaned = remove_outliers(df1, columns_to_filter)
142 df1_cleaned.shape
143
144 # %%
145 #matriz_corr=df1.corr()
146 matriz_corr=df1_cleaned.corr()
147 plt.figure(figsize=(12, 10))
148 #Mapa de calor de la correlación
149 sns.heatmap(matriz_corr, annot=True, cmap="coolwarm", fmt=".2f", linewidths=0.5)
150 plt.title("Matriz de Correlación de las variables 2022-2023")
151 plt.show()
152
153 # %%
154 df1_cleaned.describe()
155
156
157 # %% [markdown]
158 # ## Escalamiento de datos
159 #
160 # Dada la variabilidad de los datos tanto de entrenamiento y prueba se define un escalamiento para
    los datos, esto con el fin de mejorar las condiciones para las siguientes estimaciones del modelo.
161 #
162 # Esta decisión se basa en varios elementos ya que se hicieron multiples pruebas de outliers y los
    resulatos finales no mejoraron, por lo tanto, se esta estrategia fue la seleccionada.
163
164 # %%
165 from sklearn.preprocessing import MinMaxScaler
166
167 # Supongamos que df es tu DataFrame original
168 # Seleccionar las columnas a escalar
169 columns_to_scale = ['VIAJESTOTALES','KILOMETROS', 'GALONES', 'VIAJESLIQUIDOS', 'VIAJESVALORCERO',
    'KILOGRAMOS','VALORESPAGADOS']
```

```
#columns_to_scale = ['VIAJESTOTALES','KILOMETROS', 'GALONES', 'VIAJESLIQUIDOS', 'VIAJESVALORCERO', 'KILOGRAMOS']

# Inicializar el escalador
scaler = MinMaxScaler()

# Escalar solo las columnas seleccionadas
df1_cleaned[columns_to_scale] = scaler.fit_transform(df1_cleaned[columns_to_scale])

df1_cleaned.head()

# %% [markdown]
# ## Prueba de Modelos
#
# Se seleccionan 3 tipos de Regresión Lineal, RandomForest y Gradient Boosting Regresor

# %%
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
# Variables predictoras (X) y variable objetivo (y)
X = df1_cleaned.drop(columns=['VALORESPAGADOS'])
y = df1_cleaned['VALORESPAGADOS']

# Dividir en conjuntos de entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

modelo_RL = LinearRegression()
modelo_RL.fit(X_train, y_train)

# Predecir y evaluar
y_pred = modelo_RL.predict(X_test)
# Calcular las métricas de rendimiento
mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
rmse = mean_squared_error(y_test, y_pred, squared=False)  # Raíz del MSE
r2 = r2_score(y_test, y_pred)

# Mostrar las métricas
print(f"Mean Absolute Error (MAE): {mae:.8f}")
print(f"Mean Squared Error (MSE): {mse:.8f}")
print(f"Root Mean Squared Error (RMSE): {rmse:.8f}")
print(f"R-squared (R²): {r2:.8f}")

# %%
coef = modelo_RL.coef_
intercepto = modelo_RL.intercept_
equation = f"y={intercepto:.8f}"
for i, c in enumerate(coef):
    equation+=f" + ({c:.8f}*x{i+1})"
equation

# %%
plt.figure(figsize=(10, 6))
plt.scatter(y_test, y_pred, alpha=0.5, label="Predicciones vs Reales")
plt.plot([y_test.min(), y_test.max()], [y_test.min(), y_test.max()], color='red', linestyle='--', label="Línea Ideal")
plt.title("Regresión Lineal: Predicciones vs Reales")
plt.xlabel("Valores Reales")
plt.ylabel("Valores Predichos")
plt.legend()
plt.grid(True)
plt.show()

# %%
# Variables predictoras (X) y variable objetivo (y)
```

```
X = df1_cleaned.drop(columns=['VALORESPAGADOS'])
y = df1_cleaned['VALORESPAGADOS']

# Dividir en conjuntos de entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

modelo_GBR = GradientBoostingRegressor(random_state=42)
modelo_GBR.fit(X_train, y_train)

# Predecir y evaluar
y_pred = modelo_GBR.predict(X_test)

# Calcular las métricas de rendimiento
mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
rmse = mean_squared_error(y_test, y_pred, squared=False)  # Raíz del MSE
r2 = r2_score(y_test, y_pred)

# Mostrar las métricas
print(f"Mean Absolute Error (MAE): {mae:.8f}")
print(f"Mean Squared Error (MSE): {mse:.8f}")
print(f"Root Mean Squared Error (RMSE): {rmse:.8f}")
print(f"R-squared (R²): {r2:.8f}")


# %%
from sklearn.ensemble import RandomForestRegressor
# Variables predictoras (X) y variable objetivo (y)
X = df1_cleaned.drop(columns=['VALORESPAGADOS'])
y = df1_cleaned['VALORESPAGADOS']

# Dividir en conjuntos de entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

modelo_RF = RandomForestRegressor(random_state=42)
modelo_RF.fit(X_train, y_train)

# Predecir y evaluar
y_pred = modelo_RF.predict(X_test)

# %%
# Calcular las métricas de rendimiento
mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
rmse = mean_squared_error(y_test, y_pred, squared=False)  # Raíz del MSE
r2 = r2_score(y_test, y_pred)

# Mostrar las métricas
print(f"Mean Absolute Error (MAE): {mae:.8f}")
print(f"Mean Squared Error (MSE): {mse:.8f}")
print(f"Root Mean Squared Error (RMSE): {rmse:.8f}")
print(f"R-squared (R²): {r2:.8f}")

# %% [markdown]
# # Reetrenamiento
# En esta se considera loa outliers para comparar los resultados
#
#

# %%
#escalado
columns_to_scale = ['VIAJESTOTALES','KILOMETROS', 'GALONES', 'VIAJESLIQUIDOS', 'VIAJESVALORCERO',
'KILOGRAMOS','VALORESPAGADOS']
scaler = MinMaxScaler()
# Escalar solo las columnas seleccionadas
```

```
df1[columns_to_scale] = scaler.fit_transform(df1[columns_to_scale])
#Regresión lineal con outliers
#modelo_RL_WO
X = df1.drop(columns=['VALORESPAGADOS'])
y = df1['VALORESPAGADOS']

# Dividir en conjuntos de entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

modelo_RL_WO = LinearRegression()
modelo_RL_WO.fit(X_train, y_train)
print("Modelo Regresión Lineal")
# Predecir y evaluar
y_pred = modelo_RL_WO.predict(X_test)
# Calcular las métricas de rendimiento
mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
rmse = mean_squared_error(y_test, y_pred, squared=False)  # Raíz del MSE
r2 = r2_score(y_test, y_pred)

# Mostrar las métricas
print(f"Mean Absolute Error (MAE): {mae:.8f}")
print(f"Mean Squared Error (MSE): {mse:.8f}")
print(f"Root Mean Squared Error (RMSE): {rmse:.8f}")
print(f"R-squared (R²): {r2:.8f}")

print("Modelo Random Forest Regressor")
##Ramdom Forest
modelo_RF_WO = RandomForestRegressor(random_state=42)
modelo_RF_WO.fit(X_train, y_train)
# Predecir y evaluar
y_pred = modelo_RF_WO.predict(X_test)

# Calcular las métricas de rendimiento
mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
rmse = mean_squared_error(y_test, y_pred, squared=False)  # Raíz del MSE
r2 = r2_score(y_test, y_pred)

# Mostrar las métricas
print(f"Mean Absolute Error (MAE): {mae:.8f}")
print(f"Mean Squared Error (MSE): {mse:.8f}")
print(f"Root Mean Squared Error (RMSE): {rmse:.8f}")
print(f"R-squared (R²): {r2:.8f}")

print("Modelo Gradiente Boosting Regressor")
modelo_GBR_WO = GradientBoostingRegressor(random_state=42)
modelo_GBR_WO.fit(X_train, y_train)

# Predecir y evaluar
y_pred = modelo_GBR_WO.predict(X_test)

# Calcular las métricas de rendimiento
mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
rmse = mean_squared_error(y_test, y_pred, squared=False)  # Raíz del MSE
r2 = r2_score(y_test, y_pred)

# Mostrar las métricas
print(f"Mean Absolute Error (MAE): {mae:.8f}")
print(f"Mean Squared Error (MSE): {mse:.8f}")
print(f"Root Mean Squared Error (RMSE): {rmse:.8f}")
print(f"R-squared (R²): {r2:.8f}")

# %% [markdown]
```

```
# ## Optimización
#
# Dado que el modelo de Gradient Boosting Regressor tuvo el mejor desempeño, se busca hacer una optimización a través de la variación de los hiperparámatros con el fin de encontrar un modelo optimizado

# %%
df1.describe()

# %%
# Variables predictoras (X) y variable objetivo (y)
X = df1.drop(columns=['VALORESPAGADOS'])
y = df1['VALORESPAGADOS']

# Dividir en conjuntos de entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Preprocesador para escalar características
preprocessor = ColumnTransformer(
    transformers=[('scaler', MinMaxScaler(), X.columns)]
)

# Definir la función objetivo para Optuna
def objective(trial):
    # Espacio de búsqueda de hiperparámetros para GradientBoostingRegressor
    params = {
        'n_estimators': trial.suggest_int('n_estimators', 100, 500),
        'learning_rate': trial.suggest_float('learning_rate', 0.01, 0.3, log=True),
        'max_depth': trial.suggest_int('max_depth', 3, 10),
        'min_samples_split': trial.suggest_int('min_samples_split', 2, 20),
        'min_samples_leaf': trial.suggest_int('min_samples_leaf', 1, 20),
        'subsample': trial.suggest_float('subsample', 0.5, 1.0),
        'max_features': trial.suggest_float('max_features', 0.5, 1.0)
    }

    # Crear el modelo GradientBoostingRegressor
    model = Pipeline(steps=[
        ('preprocessor', preprocessor),
        ('regressor', GradientBoostingRegressor(
            **params,
            random_state=42
        ))
    ])

    # Entrenar el modelo
    model.fit(X_train, y_train)

    # Predecir y evaluar
    y_pred = model.predict(X_test)
    error = mean_squared_error(y_test, y_pred)

    return error

# Crear el estudio y ejecutar la optimización en paralelo usando 12 núcleos
study = optuna.create_study(direction='minimize')
study.optimize(objective, n_trials=50, n_jobs=4)  # Usa 12 núcleos

# Obtener los mejores hiperparámetros
print("Mejores hiperparámetros:", study.best_params)

# Entrenar el modelo final con los mejores hiperparámetros
best_params = study.best_params
best_model = Pipeline(steps=[
    ('preprocessor', preprocessor),
    ('regressor', GradientBoostingRegressor(
```

```
        **best_params,
        random_state=42
    ))
])

best_model.fit(X_train, y_train)

# Guardar el modelo optimizado
joblib.dump(best_model, r'C:\Users\durrego\Downloads\UNAD\DatosalaU\model_GB_optuna_optimized.joblib')

# %%
# Realizar predicciones
y_pred = best_model.predict(X_test)

# Calcular métricas de rendimiento
mae = mean_absolute_error(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
rmse = mean_squared_error(y_test, y_pred, squared=False)  # Raíz del MSE
r2 = r2_score(y_test, y_pred)

# Mostrar las métricas
print(f"Mean Absolute Error (MAE): {mae:.10f}")
print(f"Mean Squared Error (MSE): {mse:.10f}")
print(f"Root Mean Squared Error (RMSE): {rmse:.10f}")
print(f"R-squared (R²): {r2:.10f}")

# %% [markdown]
# Se establece un excelente resultado el cual se almacena, ahora se hace varias gráficas para reconocer las características del modelo

# %%
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score,explained_variance_score

file = r'C:\Users\durrego\Downloads\UNAD\DatosalaU\model_GB_optuna_optimized.joblib'


model = joblib.load(file)

non_zero_indices = y_test != 0
y_test_non_zero = y_test[non_zero_indices]
y_pred_non_zero = y_pred[non_zero_indices]
# Calcular MAE
mae = mean_absolute_error(y_test, y_pred)
print(f"Mean Absolute Error (MAE): {mae}")

# Calcular MSE
mse = mean_squared_error(y_test, y_pred)
print(f"Mean Squared Error (MSE): {mse}")

# Calcular R-squared
r2 = r2_score(y_test, y_pred)
print(f"R-squared (R²): {r2}")

rmse = np.sqrt(mean_squared_error(y_test, y_pred))
print(f"Root Mean Squared Error (RMSE): {rmse}")

mape = np.mean(np.abs((y_test - y_pred) / y_test)) * 100
print(f"Mean Absolute Percentage Error (MAPE): {mape}%")

# Calcular el MAPE solo para los valores no cero
mape_non_zero = np.mean(np.abs((y_test_non_zero - y_pred_non_zero) / y_test_non_zero)) * 100

print(f"Mean Absolute Percentage Error (MAPE) excluyendo ceros: {mape_non_zero}%")
```

```
evs = explained_variance_score(y_test, y_pred)
print(f"Explained Variance Score: {evs}")

# %%
importances = best_model.steps[1][1].feature_importances_

features = ['VIAJESTOTALES','KILOMETROS', 'GALONES', 'VIAJESLIQUIDOS', 'VIAJESVALORCERO',
'KILOGRAMOS']

# Crear un gráfico de barras para visualizar las importancias
plt.figure(figsize=(10, 6))
plt.barh(features, importances)
plt.xlabel("Importancia de la característica")
plt.ylabel("Características")
plt.title("Importancia de las características en el modelo Gradient Boosting")
plt.show()

# %% [markdown]
# * La característica más importante son los kilogramos, luego este es un factor determinante para
predecir el costo, lo que tiene sentido en un contexto de transporte de carga, donde influye
directamente
# * Los kilómetros, tambien es una característica relevante, siendo tambien un factor critico que
puede aumentar los costos.
# * Para las variable VIAJESTOTALES y VIAJESVALORCERO, tiene su influencia y aportar a los costos
#
# ## Predicciones

# %%
# Suponiendo que tienes y_test (valores reales) y y_pred (predicciones)
plt.figure(figsize=(8, 6))
plt.scatter(y_test, y_pred, alpha=0.5)
plt.plot([min(y_test), max(y_test)], [min(y_test), max(y_test)], 'r--')  # Línea diagonal
plt.xlabel('Valores Reales')
plt.ylabel('Predicciones')
plt.title('Predicciones vs. Valores Reales')
plt.show()

# %% [markdown]
# * Se reconoce que los valores de distribución estan cercanos a la diagonal, siendo este un
excelente resultado, la cercanía sugiere que el modelo es preciso para lo datos con los que se esta
probando.
# * Se puede obsevar que en valores más altos, la desviación es mayor, no hay un ajuste tan claro en
esta zona, esto debido a que seguramente se tienen outliers que modifican el resultado.
# * La mayoría de los puntos estan en una tendencia ascendente, lo que puede sugerir que el modelo
captura la relación general entre las características de entrada y la variable objetivo
# * En valores bajos los errores son pequeños, dada la concentración al rededor de la línea
diagonal, en valores más pequeños el modelo es más preciso.
#
# ## Residuos
#
# Ahora para los residuos se tiene la siguente gráfica

# %%
residuals = y_test - y_pred

plt.figure(figsize=(8, 6))
plt.scatter(y_pred, residuals, alpha=0.5)
plt.axhline(y=0, color='r', linestyle='--')
plt.xlabel('Predicciones')
plt.ylabel('Residuales')
plt.title('Gráfico de Residuales')
plt.show()


```

```
# %% [markdown]
# * Se observa en la gráfica una concentración de residuos en los valores bajos, en su mayoria se concentran cerca de cero, lo que indica que el modelo tiene predicciones razonables para valores bajos de la variable objetivo. La concentración es una buena señal, ya que se puede estimar que el modelo no tiene un sesgo sistémico en los valores bajos
# * A medida que las predicciones aumentan, se observa una mayor dispersión en los residuos, tanto para valores positivos como negativos, esto muestra que el modelo tiene dificultad para predicir valores altos, resultando en errores grandes. Puede establecerse que el modelo no captura toda la variabilidad de los datos en rangos altos.
# * El residuos aumenta con la predicción, lo que puede sugerir una heteocedasticidad, el error no es constante a lo largo de las predicciones.
# * No se reconoce un patrón claro en los residuos, lo que se puede considerar positivo, esto indica que no hay un sesgo direccional fuerte y que estos parecen aleatorios.
#
#
# ## Conclusiones
#
# * Se logra reconocer un modelo preciso, las métricas muestran que la predicción es precisa, con un R2 de 0.92, indica que el modelo explica la variabilidad en los valores a pagar
#
# * Las variables más relevantes son KILOGRAMOS y KILOMETROS, el modelo esta alineado con la lógica del transporte de carga, donde el peso y la distancia, son factores determinantes en los precios y es consistente con la realidad del sector de carga en el país.
#
# * Se pueden establecer mejoras, en cuanto a los valores a pagar altos, se puede ajustar el modelo para mejorar la precisión a través de un análisis más exhaustivo y considerar elementos como la transformación de los datos y características para aplicar otros tipos de análisis
#
# * El modelo es eficiente en predecir valores bajos, la dipersión de los residuos indica que el modelo es bueno para valores a pagar bajos, que en el sentido práctico es la mayoría de viajes que se puede encontra en los datos
```

Apéndice B

Limpieza, Entrenamiento y Resultados de los Modelos

```
# %%
import pandas as pd
import numpy as np
from sklearn.preprocessing import LabelEncoder
import matplotlib.pyplot as plt
import seaborn as sns

# %%
path = 'C:\\Users\\durrego\\Downloads\\UNAD\\DatosalaU\\'
df = pd.read_csv(path+'TODOCARGA.csv',delimiter='|')
df2 = pd.read_csv(path+'Cod_DANE.csv',delimiter=',')
df3 = pd.read_csv(path+'pesosMaximos.csv',delimiter=';')


# %%
df.shape

# %%
df['VIAJESTOTALES'].sum()

# %%
### Crear las nuevas columnas MES_2 y ANIO
df['MES_2'] = df['MES'].astype(str).str[4:6]
df['ANIO'] = df['MES'].astype(str).str[0:4]

df['ANIO'] = df['ANIO'].astype(int)
df['MES_2'] = df['MES_2'].astype(int)

# %% [markdown]
# Se grafica la distribución de los viajes en todo el conjunto de datos

# %%
df_grouped = df.groupby('MES_2')['VIAJESTOTALES'].sum().reset_index()

# Graficar MES_2 contra la suma de VIAJESTOTALES
plt.figure(figsize=(10, 6))
plt.bar(df_grouped['MES_2'], df_grouped['VIAJESTOTALES'], color='skyblue')
plt.title('SUMATORIA DE VIAJESTOTALES POR MES')
plt.xlabel('MES')
plt.ylabel('VIA')
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.show()

# %%
df_grouped = df.groupby('ANIO')['VIAJESTOTALES'].sum().reset_index()

# Graficar MES_2 contra la suma de VIAJESTOTALES
plt.figure(figsize=(10, 6))
plt.bar(df_grouped['ANIO'], df_grouped['VIAJESTOTALES'], color='skyblue')
plt.title('SUMATORIA DE VIAJESTOTALES POR AÑO')
plt.xlabel('AÑO')
plt.ylabel('Suma de VIAJESTOTALES')
plt.tight_layout()
```

```
plt.grid(axis='y', linestyle='--', alpha=0.7)

plt.show()

# %%
df = df.merge(df2[['Codigo Centro Poblado', 'Nombre Departamento']],
              how='left',
              left_on='CODMUNICIPIOORIGEN',
              right_on='Codigo Centro Poblado')

df.rename(columns={'Nombre Departamento': 'DEP_ORIG'}, inplace=True)
df.drop(columns=['Codigo Centro Poblado'], inplace=True)

df = df.merge(df2[['Codigo Centro Poblado', 'Nombre Departamento']],
              how='left',
              left_on='CODMUNICIPIODESTINO',
              right_on='Codigo Centro Poblado')
df.rename(columns={'Nombre Departamento': 'DEP_DEST'}, inplace=True)
df.drop(columns=['Codigo Centro Poblado'], inplace=True)

# %%
df = df.dropna(subset=['DEP_ORIG'])
df = df.dropna(subset=['DEP_DEST'])
matriz_viajes = df.pivot_table(values='VIAJESTOTALES', index='DEP_ORIG',
columns='DEP_DEST', aggfunc='sum', fill_value=0)
plt.figure(figsize=(12, 8))
sns.heatmap(matriz_viajes, annot=False, fmt=".0f", cmap="hsv", linewidths=0.5)
# Personalizar el gráfico
plt.title("Mapa de Calor de VIAJESTOTALES entre DEP_ORIG y DEP_DEST")
plt.xlabel("DEP_DEST")
plt.ylabel("DEP_ORIG")
plt.show()

# %%
df_filtered = df[df['DEP_ORIG'] != df['DEP_DEST']]
matriz_viajes = df_filtered.pivot_table(values='VIAJESTOTALES', index='DEP_ORIG',
columns='DEP_DEST', aggfunc='sum', fill_value=0)

# Configurar la figura y crear el mapa de calor sin números en las casillas
plt.figure(figsize=(12, 8))
sns.heatmap(matriz_viajes, annot=False, cmap="hsv", linewidths=0.5)

# Personalizar el gráfico
plt.title("Mapa de Calor de VIAJESTOTALES entre DEP_ORIG y DEP_DEST (sin DEP_DEST =
DEP_ORIG)")
plt.xlabel("DEP_DEST")
plt.ylabel("DEP_ORIG")
plt.show()

# %%
df_filtered = df[df['DEP_ORIG'] != df['DEP_DEST']]
matriz_viajes = df_filtered.pivot_table(values='VIAJESTOTALES', index='DEP_ORIG',
columns='DEP_DEST', aggfunc='sum', fill_value=0)

# Configurar la figura y crear el mapa de calor sin números en las casillas
plt.figure(figsize=(12, 8))
sns.heatmap(matriz_viajes, annot=False, cmap="tab20b", linewidths=0.5)

```

```
# Personalizar el gráfico
plt.title("Mapa de Calor de VIAJESTOTALES entre DEP_ORIG y DEP_DEST (DEP_DEST = DEP_ORIG)")
plt.xlabel("DEP_DEST")
plt.ylabel("DEP_ORIG")
plt.show()

# %%
# Contar la frecuencia de cada valor en CODMERCANCIA
frecuencia_codmercancia_top10 = df['CODMERCANCIA'].value_counts().nlargest(10)


# Crear el gráfico de barras
plt.figure(figsize=(10, 6))
frecuencia_codmercancia_top10.plot(kind='bar', color='skyblue')
plt.title('Frecuencia de CODMERCANCIA')
plt.xlabel('CODMERCANCIA')
plt.ylabel('Frecuencia')
plt.xticks(rotation=45)
plt.show()

# %%
top5_dep_orig_viajes = df.groupby('DEP_ORIG')['VIAJESTOTALES'].sum().nlargest(10)
top5_dep_dest_viajes = df.groupby('DEP_DEST')['VIAJESTOTALES'].sum().nlargest(10)

# Mostrar el resultado
print("Top 10 DEP_ORIG por número de viajes:")
print(top5_dep_orig_viajes)
print("Top 10 DEP_DEST por número de viajes:")
print(top5_dep_dest_viajes)
```

Apéndice C

Poster Presentado en el XXIX Verano de la Investigación Científica y Tecnológica del Pacífico, en Puerto Vallarta México

Congreso Internacional del Verano de la Investigación Científica y Tecnológica del Pacífico 2024

Modelo de Machine Learning para la caracterización de carga y estimación de costos de transporte, a partir de histórico de manifiestos de carga en Colombia

David Alexander Urrego Higuita (1), Universidad Nacional Abierta y a Distancia
John Jairo Castrillón Cardona (2), Institución Universitaria ITM
Alba Maribel Sanchez Galvez (3), Benemérita Universidad Autónoma de Puebla
ODS 9 Industria innovación e infraestructura
(1)daurregoh@unadvirtual.edu.co; (2)johncastrillon@itm.edu.co; (3)alba.sanchez@correo.buap.mx

Introducción

El sector de transporte en Colombia ha tenido diversos retos que se pueden enfrentar con el uso de datos, tales como los resultados de la última década, en la implementación de la Resolución 377 de 2013, del Ministerio de Transporte [1], donde se instruye a las empresas de transporte a registrar todas sus operaciones a través de manifiestos de carga, creando un histórico de información, que hasta el momento ha tenido una baja aplicación en la toma de decisiones en política pública, es así que este trabajo busca un aprovechamiento de esta información para estimar costos de carga en Colombia y aportar al diseño de políticas públicas. Es así que se define como pregunta de investigación, ***¿cuál es el modelo de Machine Learning óptimo, que permite estimar los costos de transporte con mayor precisión, con base en variables provenientes de los manifiestos de carga en Colombia?***

Metodología

El objetivo principal de este proyecto es desarrollar un modelo de Machine Learning que permita la estimación de los costos del transporte de carga en Colombia a partir de datos históricos de manifiestos de carga entre 2015 y 2023, basado factores determinantes, tales como origen, destino, peso carga, el tipo de carga, el tipo de producto transportado, valores pagados, entre otros, la principal fuente de información se encuentra en la plataforma del RNDC donde se puede obtener todos los registros de manifiesto de carga [2], en la figura se detalla se muestra la secuencia de la metodología.

Figura 1. Metodología propuesta

La metodología de desarrollo de este proyecto aplicado es una aproximación mixta (descriptiva y experimental), en la cual se utiliza el diseño de algoritmos para extraer y procesar la información, procesos descriptivos para reconocer los datos, la información que contienen y procesos experimentales en el diseño e implementación de los modelos de Machine Learning para encontrar un candidato que se pueda validar y permita modelar la estimación los costos.

Bibliografía

[1] Resolución 377 de 2013, Pub. L. No. Resolución 377, Página Oficial Ministerio de Transporte (2013). [2] MINTRANSPORTE. (2024). *Registro Nacional de Despachos de Carga*. Página Oficial Ministerio de Transporte. https://mdc.mintransporte.gov.co/Menu/Principal/tabid/204/language/es-MX/Default.aspx. [3]Farchi, F., Farchi, C., Touzi, B., & Mabrouki, C. (2023). A Comparative Study on AI-Based Algorithms for Cost Prediction in Pharmaceutical Transport Logistics. *Acadlore Transactions on AI and Machine Learning*, 2(3), 129–141. [4] Nama, M., Nath, A., Bechra, N., Bhatia, J., Tanwar, S., Chaturvedi, M., & Sadoun, B. (2021). Machine learning based traffic scheduling techniques for intelligent transportation system: Opportunities and challenges. *International Journal of Communication Systems*, 34(9). [5] Akbari, M., & Do, T. N. A. (2021). A systematic review of machine learning in logistics and supply chain management: current trends and future directions. *Benchmarking: An International Journal*, 28(10), 2977–3005. [6] Moscoso-López, J. A., Urda, D., Ruiz-Aguilar, J. J., González-Enrique, J., & Turias, I. J. (2021). A machine learning-based forecasting system of perishable cargo flow in maritime transport. *Neurocomputing*, 452, 487–497. [7] Plakandaras, V., Papadimitriou, T., & Gogas, P. (2019). Forecasting transportation demand for the U.S. market. *Transportation Research Part A: Policy and Practice*, 126, 195–214.

México Colombia Costa Rica Perú Nicaragua Estados Unidos Ecuador

Resultados

En cuanto al diseño metodológico, se logra construir un bot Python para acceder a la plataforma y extraer los datos, en el cual se obtiene una base de datos con 29 columnas, un total de XX millones de registros que representan de XX viajes entre los años 2015 y 2016. Luego se hace una limpieza de datos a través de datos análisis de datos faltantes, valores nulos y variables redundantes. Se hace un análisis exploratorio EDA, aquí se convierten todas las variables a numéricas, para tipo texto se utiliza un LabelEncoder, con el fin de caracterizar los datos, construyendo la matriz de correlación (Fig 3), así se reducen las 15 columnas, siendo la variable objetivo VALORESPAGADOS y encontrando una mayor correlación con KILOGRAMOS, VIAJESTOTALES y KILOMETROS. Es así como, se logra diseñar un proceso de entrenamiento, donde se dividen los datos en *train* y *test*, de la librería *scikit-learn* de Python, y se prueban tres tipos de modelos, regresión línea, Random Forest y Grandient Boosting Regressor (GBR), que generan un coeficiente de precisión (R2), como se ve en la tabla.

Parámetro	Regresión Lineal	Random Forest	Grandient Boosting Regressor
R2	0.658815	0.9795976	0.942596

La selección es el GBR, ya que genera un modelo más liviano, con una buena precisión y fácil implementación, se hace un reentrenamiento, normalizando VALORESPAGADOS, en la tabla se detalla las características del modelo, en las figuras 4, 5 y 6 se muestra elementos de interés para reconocer el modelo.

Variable	Valor	Variable	Valor
Alpha	0.9	Learning_rate	0.1
Max_depth	3	N_estimators	100
MAE	0.0003617758	MSE	2.659781919025
R2	0.9428035207	RMSE	0.001630883784

Fig. 3 Correlación

Fig. 4 Importancia características

Fig. 5 Predicciones vs valores

Fig. 6 Residuos

Conclusiones

1. El adecuado tratamiento, limpieza y transformación de los datos permite desarrollar un modelo eficiente para la estimación de los costos de transporte en Colombia, con modelo Gradient Boost con un nivel de precisión del 94.3% en la clasificación de la información
2. Permite sentar las bases para otro tipo de estudios como clasificación de tipo de carga [4], segmentación de rutas de transporte [5], análisis de tendencias temporales [6] y predicción de la demanda de transporte en Colombia [7].
3. El aporte al ODS 9 facilita el avance del país en el aprovechamiento de su infraestructura, para la evaluación y mejora de las políticas públicas del sector transporte

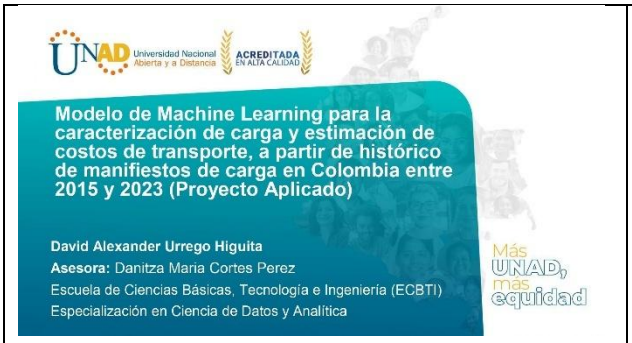
Apéndice D

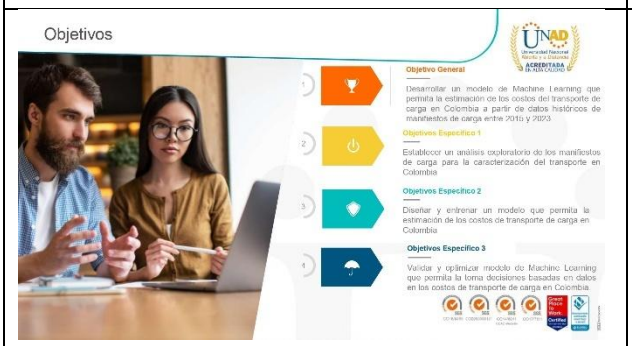
Video Presentación

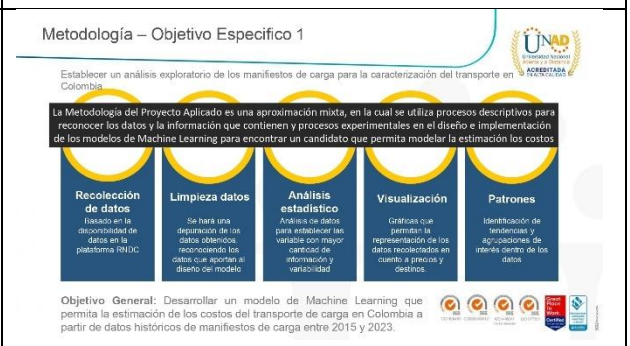
Se muestra a continuación el enlace de la presentación, https://youtu.be/o-VY1HqTdh0

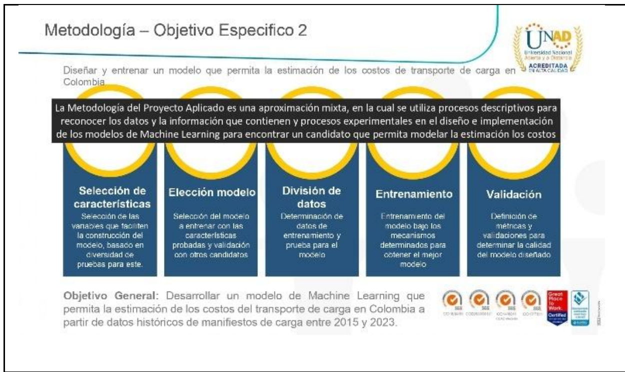
Apéndice E

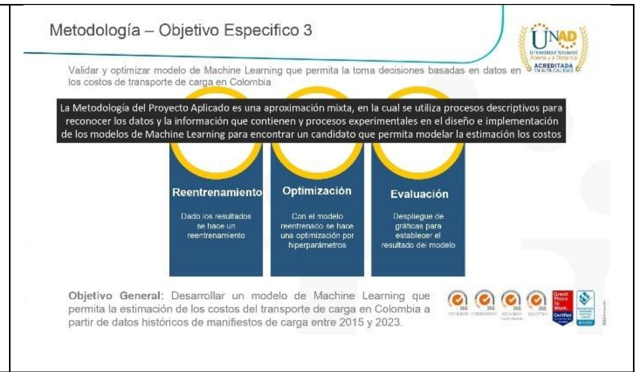
Diapositivas Presentación

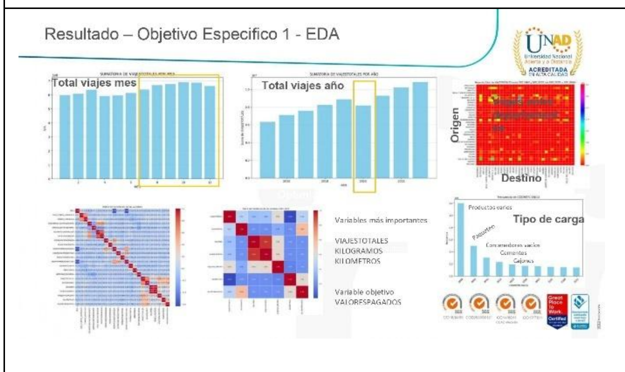






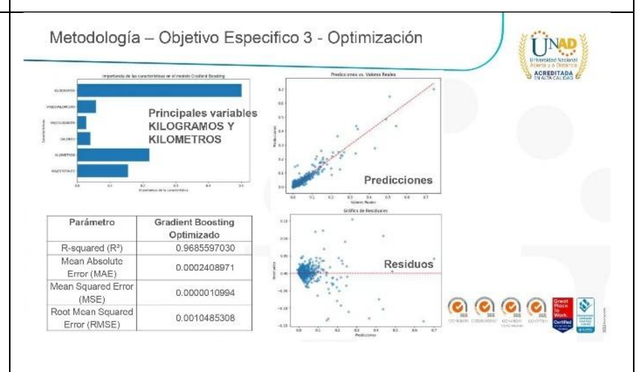




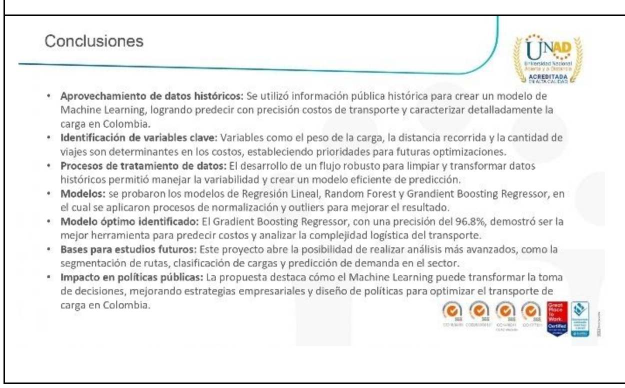


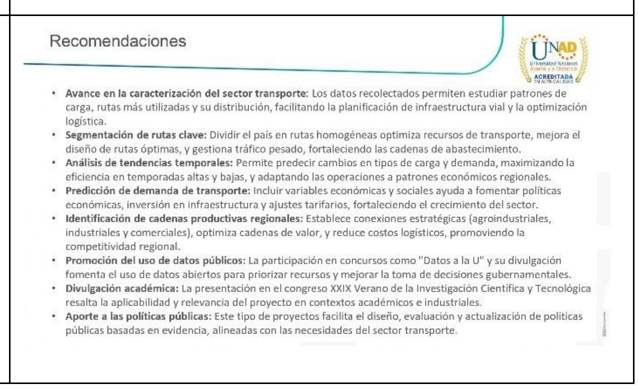
Parámetro	Regresión Lineal	Random Forest	Gradient Boosting
R-squared (R²)	0.58809953	0.76777886	0.75418800
Mean Absolute Error (MAE)	0.09846416	0.07038581	0.07452025
Mean Squared Error (MSE)	0.02040918	0.01150628	0.01217969
Root Mean Squared Error (RMSE)	0.14286071	0.10726734	0.11036165

Parámetro	Regresión Lineal	Random Forest	Gradient Boosting	Gradient Boosting Optimizado
R-squared (R²)	0.61527719	0.96246783	0.93711103	0.9685597030
Mean Absolute Error (MAE)	0.00069182	0.00022897	0.00035985	0.0002408971
Mean Squared Error (MSE)	0.00001345	0.00000131	0.00000220	0.0000010994
Root Mean Squared Error (RMSE)	0.00366785	0.00114562	0.00148295	0.0010485308



Parámetro	Gradient Boosting Optimizado
R-squared (R²)	0.9685597030
Mean Absolute Error (MAE)	0.0002408971
Mean Squared Error (MSE)	0.0000010994
Root Mean Squared Error (RMSE)	0.0010485308






UNAD
Universidad Nacional Abierta y a Distancia
ACREDITADA EN ALTA CALIDAD
Más UNAD, más equidad
¡GRACIAS!
www.unad.edu.co
Síguenos: @UniversidadUNAD

Apéndice F

Registro Analítico Educativo

RESUMEN ANALITICO EDUCATIVO

RAE

Título del texto	Modelo de Machine Learning para la caracterización de carga y estimación de costos de transporte, a partir de histórico de manifiestos de carga en Colombia entre 2015 y 2023
Nombres y Apellidos del Autor	David Alexander Urrego Higuita
Año de la publicación	**2025**
Palabras Claves	Machine Learning, Transporte de Carga, Manifiesto de Carga, Grandient Boosting Regressor, Costos de Transporte

Problema que aborda el texto:

El problema que aborda el texto se centra en las dificultades que enfrenta el sector de transporte de carga en Colombia, específicamente la incapacidad de caracterizar adecuadamente la carga y estimar con precisión los costos de transporte. A pesar de la implementación de la Resolución 377 de 2013, que obliga a las empresas de transporte a registrar sus operaciones a través de manifiestos de carga, la información generada no se ha aprovechado de manera efectiva en la toma de decisiones y en la optimización de los procesos logísticos.

Esta falta de precisión en la estimación de los costos y la planificación logística genera ineficiencias operativas y estratégicas, que afectan la rentabilidad y competitividad de las empresas de transporte. Además, la baja capacidad de adaptación a los cambios en la demanda del mercado impide a las empresas ofrecer servicios personalizados y mejorar su infraestructura logística. Esto resulta en sobrecostos y una disminución en la satisfacción del cliente.

El trabajo propone la implementación de un modelo de Machine Learning como solución para mejorar la estimación de los costos de transporte, aprovechando los datos históricos de los manifiestos de carga, con el objetivo de optimizar la logística y hacer las operaciones más eficientes

Objetivos del texto:

Objetivo General

Desarrollar un modelo de Machine Learning que permita la estimación de los costos del transporte de carga en Colombia a partir de datos históricos de manifiestos de carga entre 2015 y 2023.

Objetivos Específicos

- Establecer un análisis exploratorio de los manifiestos de carga para la caracterización del transporte en Colombia.
- Diseñar y entrenar un modelo que permita la estimación de los costos de transporte de carga en Colombia

- Validar y optimizar modelo de Machine Learning que permita la toma decisiones basadas en datos en los costos de transporte de carga en Colombia.

Hipótesis planteada por el autor:

¿Cuál es el modelo de Machine Learing que permite estimar los costos de transporte con mayor precisión, con base en variables provenientes de los manifiestos de carga en Colombia?

La hipótesis planteada es que el uso de un modelo de Machine Learning permitirá estimaciones más precisas de los costos de transporte en Colombia, a partir de los datos históricos de los manifiestos de carga. Esto se basa en la suposición de que, al aplicar técnicas avanzadas de Machine Learning, como la regresión lineal, árboles de decisión y redes neuronales, se logrará capturar patrones complejos y relaciones no lineales entre las variables que afectan los costos de transporte, lo que no es posible con los métodos tradicionales utilizados hasta el momento.

También se plantea que este modelo mejorará la eficiencia en la planificación logística y permitirá optimizar los recursos en el sector, ayudando a las empresas de transporte a reducir costos y mejorar la competitividad.

En resumen, la hipótesis central es que un modelo de Machine Learning bien diseñado y validado puede ofrecer una herramienta más precisa y eficaz para la estimación de costos de transporte, contribuyendo al desarrollo y sostenibilidad del sector en Colombia.

Tesis principal del autor:

La tesis principal es que el uso de un modelo de Machine Learning basado en los datos históricos de los manifiestos de carga en Colombia puede transformar la estimación de los costos de transporte. Al aplicar técnicas avanzadas de Machine Learning, se puede mejorar la precisión en la estimación de los costos y optimizar la logística del transporte de carga, lo que contribuiría a resolver las ineficiencias operativas y la subutilización de los datos existentes.

Se sostiene que, mediante el diseño de un modelo robusto de Machine Learning, es posible capturar patrones complejos en los datos que los métodos tradicionales no pueden identificar, lo que permitiría reducir los costos operativos, optimizar la asignación de recursos y mejorar la competitividad del sector del transporte de carga en Colombia. Esta tesis está orientada a demostrar cómo la aplicación de herramientas tecnológicas avanzadas puede generar mejoras sustanciales en la gestión y eficiencia del sector.

Argumentos expuestos por el autor:

Se exponen varios argumentos clave para apoyar su tesis sobre la implementación de un modelo de Machine Learning para la estimación de costos de transporte en Colombia:

- Desafíos en la estimación de costos de transporte, el sector de transporte de carga en Colombia enfrenta serias ineficiencias debido a la caracterización inadecuada de la carga y la estimación errónea de costos. Estas deficiencias impactan directamente en la planificación logística, el uso ineficiente de recursos y, en consecuencia, en sobrecostos operativos. Además, estas inexactitudes dificultan una asignación óptima de recursos y afectan la calidad del servicio, lo que genera una menor competitividad en el mercado.

- Subutilización de los datos históricos, a pesar de la implementación de la Resolución 377 de 2013, que exige a las empresas de transporte registrar sus operaciones mediante manifiestos de carga, gran parte de esta información sigue sin ser utilizada adecuadamente. El autor destaca que esta gran cantidad de datos históricos no ha sido aprovechada para optimizar los procesos logísticos ni mejorar la precisión en la estimación de costos. Este vacío de aprovechamiento de datos representa una oportunidad desaprovechada.

- Ventajas del Machine Learning, este puede ser la clave para mejorar la estimación de costos, dado que

puede identificar patrones complejos y relaciones no lineales en los datos que los métodos tradicionales no son capaces de detectar. A través de modelos como la regresión lineal, árboles de decisión y redes neuronales, el modelo propuesto permitirá predecir de manera más precisa los costos de transporte, optimizando así los recursos y mejorando la competitividad del sector.

- Validación y robustez del modelo, para garantizar la efectividad del modelo, el autor argumenta que se deben aplicar técnicas de validación cruzada y optimización de hiperparámetros, lo que asegurará que el modelo sea robusto y generalizable. Además, destaca la importancia de integrar el modelo en plataformas de visualización de datos, como Power BI, lo que permitiría interactuar con los resultados y facilitar la toma de decisiones.

- Impacto en la competitividad del sector, se argumenta que la implementación de este modelo de Machine Learning no solo ayudará a reducir costos y mejorar la eficiencia operativa, sino que también fortalecerá la competitividad del sector, proporcionando herramientas más precisas para la toma de decisiones estratégicas y la optimización de la logística. La adaptabilidad a cambios en la demanda del mercado también se ve como una ventaja crítica para mejorar la sostenibilidad del sector a largo plazo.

- Base para futuros estudios, se resalta que este modelo no solo tiene un impacto inmediato en la estimación de costos, sino que también sentará las bases para futuras investigaciones, tales como la segmentación de rutas de transporte, la predicción de la demanda de transporte y el análisis de tendencias temporales. Este enfoque a largo plazo permitirá mejorar aún más la precisión y eficiencia del sistema de transporte de carga en Colombia.

Los argumentos clave giran en torno a la ineficiencia actual en la estimación de costos, la subutilización de los datos históricos, y las ventajas del Machine Learning para mejorar la precisión y optimización en la logística del transporte.

Conclusiones del texto:

Este trabajo logra aprovechar información pública disponible histórica para crear un Modelo de Machine Learning que permite predecir con precisión los costos de transporte en Colombia, gracias a el desarrollo de un flujo de procesamiento de información que facilita el alistamiento para el entrenamiento, optimización y validación. Estos datos históricos permitieron a su vez una caracterización detallada de la carga en Colombia, identificando características claves y relaciones significativas que permiten reconocer la importancia de la información para el sector transporte.

La identificación de variables cable como el peso de la carga, la distancia recorrida y la cantidad de viajes, demuestran ser determinantes en la estimación de los costos de carga, siendo estos los que se deben considerar con mayor peso a la hora de reconocer nuevos elementos y el análisis de nuevos datos para mejorar las estimaciones.

Para el desarrollo de proyectos de ciencias de datos y analítica, es necesario crear un proceso adecuado para el tratamiento, limpieza y transformación de los datos permite desarrollar un modelo eficiente para la estimación de los costos de transporte en Colombia, dada la gran variabilidad de datos y elementos a considerar que se han registrado desde el 2015 hasta 2023.

Se probaron diferentes modelos tales como regresión línea, Random Forest y Gradient Boosting Regressor, con resultados de carácter medio, donde fue necesario realizar diversidad de pruebas tales como normalización y outliers para llegar a un modelo con unas características interesantes sobre el cual se puedo realizar la optimización.

El mejor modelo encontrado fue el Gradient Boosting Regressor con un nivel de precisión del 96.8% en la clasificación de la información, este modelo logra una precisión suficiente en la predicción de costos de transporte, demostrando ser una herramienta valiosa para las empresas de logística. Las variables de peso en kilogramos y la distancia entre ciudades, son las más adecuadas para comprensión de la complejidad del

problema.

Este trabajo permite sentar las bases para otro tipo de estudios como clasificación de tipo de carga (Nama et al., 2021), segmentación de rutas de transporte (Akbari & Do, 2021), análisis de tendencias temporales (Moscoso-López et al., 2021) y predicción de la demanda de transporte en Colombia (Plakandaras et al., 2019)

Finalmente, con este tipo de propuestas se demuestra el impacto que puede tener el Machine Learning en el diseño de Políticas Públicas, en el aprovechamiento de información disponible que permite mejorar la toma de decisiones para las empresas, entidades territoriales y organismos de orden nacional, para mejorar el sector de carga en el país.

- **Bibliografía citada por el autor:**

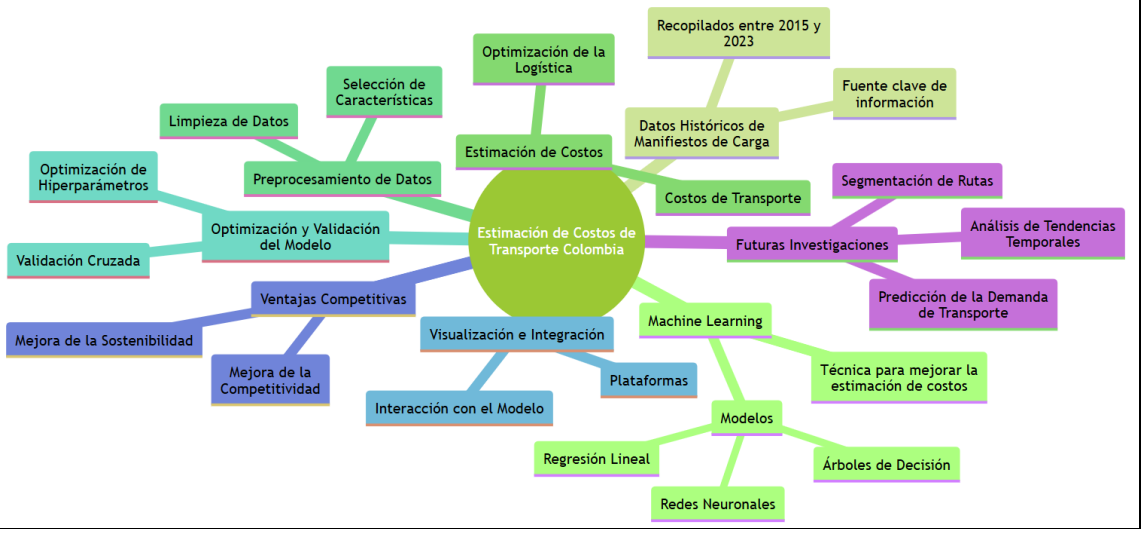
- Akbari, M., & Do, T. N. A. (2021). A systematic review of machine learning in logistics and supply chain management: current trends and future directions. Benchmarking: An International Journal, 28(10), 2977–3005. https://doi.org/10.1108/BIJ-10-2020-0514
- Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. J. (2020). A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science (pp. 3–21). https://doi.org/10.1007/978-3-030-22475-2_1
- Amanpreet, S., Narina, T., & Aakanksha, S. (2016). A review of supervised machine learning algorithms. 3rd International Conference on Computing for Sustainable Global Development (INDIACom). https://ieeexplore.ieee.org/abstract/document/7724478
- Boukerche, A., & Wang, J. (2020). Machine Learning-based traffic prediction models for Intelligent Transportation Systems. Computer Networks, 181, 107530. https://doi.org/10.1016/j.comnet.2020.107530
- Boyer, R. C., Scherer, W. T., & Smith, M. C. (2017). Trends Over Two Decades of Transportation Research. Transportation Research Record: Journal of the Transportation Research Board, 2614(1), 1–9. https://doi.org/10.3141/2614-01
- Colfecar. (2022). Análisis del sector transporte de carga: Productividad, eficiencia y principales cifras económicas. Colfecar. https://www.colfecar.org.co/wp-content/uploads/An%C3%A1lisis%20del%20sector%20transporte%20de%20carga.pdf
- de la Torre, R., Corlu, C. G., Faulin, J., Onggo, B. S., & Juan, A. A. (2021). Simulation, Optimization, and Machine Learning in Sustainable Transportation Systems: Models and Applications. Sustainability, 13(3), 1551. https://doi.org/10.3390/su13031551
- Farchi, F., Farchi, C., Touzi, B., & Mabrouki, C. (2023). A Comparative Study on AI-Based Algorithms for Cost Prediction in Pharmaceutical Transport Logistics. Acadlore Transactions on AI and Machine Learning, 2(3), 129–141. https://doi.org/10.56578/ataiml020302
- Gonzalez, L. C. (2022). Retos y tendencias empresariales con relación a la optimización de costos logísticos para una efectiva logística integral [Contaduría Pública]. Universidad Abierta y a Distancia UNAD.
- Martinez, C. (2022). Colombia y los retos en el transporte de carga. The Logistics World. https://thelogisticsworld.com/actualidad-logistica/646219
- Mintransporte. (2024). Normatividad Ministerio de Transporte - RNDC. Página Oficial Ministerio de Transporte Colombia. https://plc.mintransporte.gov.co/RNDC/Normatividad
- MINTRANSPORTE. (2024). Registro Nacional de Despachos de Carga. Página Oficial Ministerio de Transporte. https://rndc.mintransporte.gov.co/MenuPrincipal/tabid/204/language/es-MX/Default.aspx?returnurl=%2f
- Morabit, M., Desaulniers, G., & Lodi, A. (2023). Machine-Learning–Based Arc Selection for Constrained Shortest Path Problems in Column Generation. INFORMS Journal on Optimization, 5(2), 191–210. https://doi.org/10.1287/ijoo.2022.0082
- Moscoso-López, J. A., Urda, D., Ruiz-Aguilar, J. J., González-Enrique, J., & Turias, I. J. (2021). A machine learning-based forecasting system of perishable cargo flow in maritime transport. Neurocomputing, 452, 487–497. https://doi.org/10.1016/j.neucom.2019.10.121
- Nama, M., Nath, A., Bechra, N., Bhatia, J., Tanwar, S., Chaturvedi, M., & Sadoun, B. (2021). Machine learning-based traffic scheduling techniques for intelligent transportation system: Opportunities and challenges. International Journal of Communication Systems, 34(9). https://doi.org/10.1002/dac.4814
- Plakandaras, V., Papadimitriou, T., & Gogas, P. (2019). Forecasting transportation demand for the U.S. market. Transportation Research Part A: Policy and Practice, 126, 195–214.

https://doi.org/10.1016/j.tra.2019.06.008

- Rajoub, B. (2020). Supervised and unsupervised learning. In Biomedical Signal Processing and Artificial Intelligence in Healthcare (pp. 51–89). Elsevier. https://doi.org/10.1016/B978-0-12-818946-7.00003-2
- Resolución 377 de 2013, Pub. L. No. Resolución 377, Página Oficial Ministerio de Transporte (2013). https://plc.mintransporte.gov.co/Portals/0/Documentos/Resolucion0000377-2013.pdf?ver=2018-09-20-185238-000
- Saravanan, R., & Sujatha, P. (2019). A State of Art Techniques on Machine Learning Algorithms: A Perspective of Supervised Learning Approaches in Data Classification. Proceedings of the 2nd International Conference on Intelligent Computing and Control Systems, ICICCS 2018, 945–949. https://doi.org/10.1109/ICCONS.2018.8663155
- Sindhu Meena, K., & Suriya, S. (2020). A Survey on Supervised and Unsupervised Learning Techniques. In Proceedings of International Conference on Artificial Intelligence, Smart Grid and Smart City Applications (pp. 627–644). Springer International Publishing. https://doi.org/10.1007/978-3-030-24051-6_58
- Singh, A., Das, A., Bera, U. K., & Lee, G. M. (2021). Prediction of Transportation Costs Using Trapezoidal Neutrosophic Fuzzy Analytic Hierarchy Process and Artificial Neural Networks. IEEE Access, 9, 103497–103512. https://doi.org/10.1109/ACCESS.2021.3098657
- Taylor, R. (2020). Papers with Code is Expanding to More Sciences! Página Oficial Medium. https://medium.com/paperswithcode/papers-with-code-is-expanding-to-more-sciences-5d375d10ca3a
- van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. Machine Learning, 109(2), 373–440. https://doi.org/10.1007/s10994-019-05855-6
- Verma, K. K., Singh, B. M., & Dixit, A. (2022). A review of supervised and unsupervised machine learning techniques for suspicious behavior recognition in intelligent surveillance system. International Journal of Information Technology, 14(1), 397–410. https://doi.org/10.1007/s41870-019-00364-0

Nombre y apellidos de quien elaboró este RAE	David Alexander Urrego Higuita
Fecha en que se elaboró este RAE	22/02/2025

Imagen (mapa conceptual) que resume e interconecta los principalesconceptos encontrados en el texto:



Comentarios finales

Se recomienda utilizar los resultados y los datos recolectados para avanzar en el entendimiento del sector transporte a través de estudios tipo caracterización de carga, comprender los patrones del transporte de carga en Colombia, como qué tipos de carga se transportan más, cuáles son las rutas más utilizadas para ciertas cargas y cómo se distribuyen en las diferentes regiones. Su utilización se puede aplicar para planificación de infraestructura vial, optimización de recursos logísticos, identificación de sectores con alta demanda en transporte.

La segmentación de rutas de transporte permite, dividir el país en rutas claves, con cierta homogeneidad para el análisis específico que permitan la optimización de los recursos de transporte, facilita así para los tomadores de decisiones, el diseño de rutas óptimas, gestión de tráfico pesado y mejoras logística en cadenas de abastecimiento regionales.

Aplicar estudios de tendencias temporales, permitiría a largo plazo, predecir cambios en los tipos de carga, pesos, rutas preferidas, y demanda temporal, el análisis a corto y largo plazo, permite tomar decisiones informadas y anticipar la evolución del sector transporte. Su uso se puede aplicar en alistamiento de temporadas altas y bajas, maximizar el uso de contenedores vacíos, predicción de picos de demanda y adaptación de las empresas de transporte a los patrones económicos en cada región.

También se puede utilizar para la predicción de demanda de transporte, incluyendo variables económicas y sociales, ayuda a fomentar políticas económicas, inversión en infraestructura, ajustes a tarifas de asociadas al transporte, capacidad de crecimiento y proyección sectorial.

Identificación de cadenas productivas regionales, conexiones agrícolas, mercados, zonas industriales, puertos y aeropuertos para establecer cadenas de valor, optimizar su funcionamiento, facilita el comercio y aprovechamiento de la cadena, establecer alianzas, búsqueda de alternativas de proveedores de insumos, y con esto fortalecer la competitividad regional, la promoción de inversión en sectores estratégicos, y reducción del costo logístico.

Este tipo de proyectos facilita y fomenta el uso de datos públicos, como caso particular se particpo en concurso Datos a la U (Resultados Convocatoria "Datos a la U" | Datos Abiertos Colombia), donde se logro avanzar a la segunda etapa, obteniendo un puntaje final de 93 puntos, y logrando publicar el modelo desarrollado davidUrr/DatosalaU_R3_G17: Repositorio del repositorio para el concurso Datos a la U https://www.datos.gov.co/stories/s/Actualidades-del-concurso-Datos-a-la-U/wn73-87k7/, esta divulgación facilita afrontar otros tipo de retos con información pública, esto facilita la toma de decisiones del gobierno y prioriza los recursos para sectores y proyectos claves.

Este trabajo pudo ser divulgado en el XXIX Verano de la Investigación Científica y Tecnológica del Pacífico, en Puerto Vallarta México, en el Anexo C, se puede ver una copia del poster, adicional en el Anexo D, se muestra un video con la presentación de este trabajo.

Finalmente, este tipo de proyecto ayuda a las entidades gubernamentales a la formulación, evaluación y actualización de Políticas Públicas, para diseñar, evaluar y ajustar las iniciativas y hacer seguimiento, promoviendo políticas más justas, basadas en evidencia, alineadas con las necesidades actuales y futuras del sector.