Análisis de datos para la optimización de la gestión de flotas vehiculares: Impacto en los costos operativos y rendimiento empresarial

Ronald Akerman Ortiz Garcia

Monografía presentada para optar al título de Especialista en Analítica y Ciencia de Datos

Director
Fernando Ceballos, Doctor (PhD) en Ingeniería
Asesor
David Manuel Villanueva Valdés (MSc) em Ingeniería de Software

Universidad de Antioquia
Facultad de Ingeniería
Especialización en Analítica y Ciencia de Datos
Medellín, Antioquia, Colombia
2023

Cita	(Ortiz Garcia & Ceballos, 2023)
Referencia **Estilo APA 7 (2020)**	Ortiz Garcia, Ronald A. & Ceballos, Yony. F. (2023). *Análisis de datos para la optimización de la gestión de flotas vehiculares: Impacto en los costos operativos y rendimiento empresarial - 2023* [Trabajo de grado especialización]. Universidad de Antioquia, Medellín, Colombia.

Especialización en Analítica y Ciencia de Datos, Cohorte IV.

Grupo de Investigación Ingeniería y Sociedad (I&S).

Centro de Investigación Ambientales y de Ingeniería (CIA).

Dedicatoria

A mis padres Alba Rosa García de Ortiz y Juan Lucas Evangelista Ortiz Vásquez, quienes siempre me han apoyado en proceso de formación como ser humano y profesional.

Agradecimientos

Agradezco a mi asesor de Tesis, Fernando Ceballos por su paciencia y recomendaciones con el presente trabajo realizado.

También a mi colega, compañero y amigo David Manuel Villanueva Valdés quien también supo guiarme en todas las etapas de este trabajo.

Tabla de contenido

Resumen 7

Abstract 8

Introducción 9

2 Justificación 12

3 Objetivos 14

3.1 Objetivo general 14

3.2 Objetivos específicos 14

4 Marco teórico 15

5 Metodología 17

5.1 Metodología CRISP-DM 18

Comprensión del negocio: 19

Comprensión de los datos: 20

Preparación de los datos: 24

Modelización: 25

Evaluación del modelo: 28

MSE (Mean Squared Error): 28

RMSE (Root Mean Squared Error): 28

MAE (Mean Absolute Error): 28

R2 (Coefficient of Determination): 28

6 Resultados 29

7 Discusión 31

8 Conclusiones 33

9 Recomendaciones 35

Referencias 36

Lista de tablas

Tabla 1 Métricas de evaluación para los datos de entrenamiento del modelo 28

Tabla 2: Métricas de evaluación para los datos de prueba.....29

Lista de figuras

Figura 1 Diagrama proceso metodología CRISP-DM 18

Figura 2 Correlación de las variables numéricas del dataset 22

Figura 3 Distribución de los datos en Boxplot.....23

Figura 4 Datos normalizados 24

Figura 5 Análisis de correlación 25

Figura 6 Proceso del método de aprendizaje supervisado 26

Figura 7 Código fuente del modelo aplicando Regresión Lineal.....27

Resumen

En este estudio, se aplicó el análisis de datos a la gestión de flotas vehiculares en el sector del transporte de mercancía, con el objetivo de mejorar la eficiencia y rentabilidad de las operaciones logísticas. Se utilizó un conjunto de datos recolectados de una empresa en Medellín, Colombia, y se aplicaron técnicas estadísticas y modelos de regresión para realizar predicciones.

Las conclusiones revelaron que la implementación de herramientas basadas en el análisis de datos puede contribuir a mejorar el rendimiento y la eficiencia en la gestión de flotas vehiculares. Los resultados obtenidos demostraron una relación significativa entre los costos operativos y las variables estudiadas.

Se recomienda ampliar el conjunto de variables, explorar técnicas de machine learning avanzadas, realizar análisis de sensibilidad, establecer un sistema de monitoreo continuo y aplicar técnicas de optimización. Estas recomendaciones abren futuras líneas de investigación para abordar de manera más completa los desafíos en la gestión de flotas vehiculares y el transporte de mercancía.

En general, este estudio resalta la importancia del análisis de datos en la gestión de flotas vehiculares, mostrando cómo las técnicas estadísticas y los modelos de regresión pueden utilizarse para predecir costos y gastos, optimizando así las operaciones logísticas en el sector del transporte de mercancía. Los resultados obtenidos respaldan la relevancia de aplicar estas herramientas en empresas de transporte y ofrecen una base sólida para futuras investigaciones en este campo.

Palabras clave: transporte terrestre, análisis de datos, modelos de regresión, predicción.

Abstract

In this study, data analysis was applied to the management of vehicle fleets in the merchandise transport sector, with the aim of improving the efficiency and profitability of planning operations. A data set collected from a company in Medellín, Colombia is used, and statistical techniques and regression models were applied to make predictions.

The conclusions revealed that the implementation of tools based on data analysis can contribute to improving performance and efficiency in vehicle fleet management. The results obtained supported the alternate hypothesis, demonstrating a meaningful relationship between operating costs and the variables studied.

Is recommended to expand the set of variables, explore advanced Machine Learning techniques, perform sensitivity analysis, establish a continuous monitoring system, and apply optimization techniques. These recommendations clarify future lines of research to fully address the challenges in vehicle fleet management and merchandise transport.

In general, this study highlights the importance of data analysis in vehicle fleet management, showing how statistical techniques and regression models can be used to predict costs and expenses, thus optimizing planning operations in the freight transport sector. The results obtained support the relevance of applying these tools in transport companies and offer a solid foundation for future research in this field.

*Keywords***:** Land transportation, data analysis, regression models, prediction.

Introducción

El transporte de mercancía es un sector fundamental en la economía global, ya que permite el intercambio de bienes entre diferentes regiones del mundo. Este sector es esencial para garantizar el abastecimiento de materias primas y productos terminados en los mercados, lo que a su vez promueve el crecimiento económico y el desarrollo en diferentes países (Sarri et al., 2023). Además, el transporte de mercancía también juega un papel crucial en la cadena de suministro, permitiendo que las empresas puedan recibir y distribuir sus productos de manera eficiente y efectiva. Por estas razones, el transporte de mercancía es una actividad estratégica que afecta a múltiples sectores de la economía global y tiene un impacto significativo en la vida de las personas en todo el mundo (Forigua & Lyons, 2016).

En el caso de Colombia, el transporte de mercancía es un sector clave para la economía del país, ya que es un país con una importante influencia a nivel de exportación y una producción de materias primas significativa. Además, cuenta con una ubicación geográfica estratégica en la región andina, lo que le permite ser un importante corredor de transporte para el comercio entre América del Sur, América Central y América del Norte. En este sentido, el transporte de mercancía se ha convertido en un pilar fundamental para el desarrollo de la economía colombiana y ha sido objeto de importantes inversiones en infraestructura y tecnología en los últimos años, con el fin de mejorar la eficiencia y la seguridad de las operaciones logísticas en el país (Bonilla, 2019; Forigua & Lyons, 2016).

Dentro del contexto colombiano, Medellín es una ciudad que destaca por su importante papel como centro logístico y de transporte de mercancía en el país. Gracias a su posición geográfica estratégica, la ciudad se ha convertido en un punto clave para la distribución de productos en el territorio nacional, así como para la conexión con otros países de la región. En los últimos años, Medellín ha experimentado un importante crecimiento en su infraestructura de transporte y logística, lo que ha permitido mejorar la eficiencia de las operaciones y la competitividad de las empresas en el mercado. Además, la ciudad cuenta con una amplia oferta de servicios especializados en logística y transporte, lo que la convierte en un destino atractivo para las inversiones y el desarrollo de proyectos en este sector (Bonilla, 2019). En este sentido, Medellín se ha consolidado como una de las principales ciudades de Colombia en materia de transporte de mercancía, y continúa trabajando en el fortalecimiento de su infraestructura y servicios para seguir

impulsando el crecimiento económico del país (Adarme Jaimes et al., 2015; Martínez-Jaramillo et al., 2017).

Por otra parte, la logística empresarial en el sector transporte se ha beneficiado del uso de tecnologías avanzadas, como el análisis de datos y la inteligencia artificial, para mejorar la eficiencia y rentabilidad de las operaciones. En particular, el análisis de datos del ruteo de vehículos se ha convertido en una herramienta clave para pronosticar costos y otras variables relevantes en el transporte de mercancías. Al conocer de forma precisa y en tiempo real las condiciones del tráfico, las características de las vías y otros factores relevantes, las empresas pueden ajustar sus rutas y tiempos de entrega para minimizar costos y tiempos de espera. Diversas empresas en el mundo ya han implementado estas tecnologías en sus operaciones logísticas, y han obtenido importantes beneficios en términos de eficiencia y competitividad. De esta forma, el análisis de datos del ruteo de vehículos se ha convertido en una herramienta clave para la logística empresarial en el sector del transporte de mercancía, permitiendo a las empresas mejorar sus operaciones y aumentar su rentabilidad (Montoya-Torres et al., 2021; Vidya & Deepa, 2019).

En este contexto, se hace relevante destacar la importancia de llevar a cabo estudios de análisis de datos en empresas de transporte con el fin de mejorar la eficiencia y rentabilidad de sus operaciones logísticas. En particular, se presenta un estudio de análisis de datos de una empresa de Medellín la cual cuenta con su flota vehicular propia que podría beneficiarse significativamente de este tipo de análisis para pronosticar costos y gastos asociados a dicha flota con el objetivo de optimizar su operación logística.

En el presente trabajo se llevará a cabo un análisis exhaustivo de los datos recolectados con el fin de comprender su naturaleza y establecer las acciones necesarias para mejorar la consistencia de estos. En este sentido, se buscará identificar posibles datos faltantes, errores o inconsistencias, para luego aplicar las correcciones necesarias y asegurar la calidad del dataset.

A partir de este análisis previo, se establecerán los parámetros necesarios para determinar los mejores métodos y técnicas estadísticas de análisis de datos que permitan pronosticar los costos y gastos futuros de la flota vehicular de la empresa, es posible que sólo se haga en términos de una variable del dataset proporcionado con el fin de planificar para el siguiente periodo los costos asociados a la actividad de transporte de la empresa.

1 Antecedentes y planteamiento del problema

En la actualidad, el análisis de datos se ha convertido en una herramienta fundamental para la toma de decisiones empresariales. Las empresas pueden aprovechar los datos generados por sus procesos de negocio para identificar patrones, tendencias y oportunidades de mejora en su desempeño. Al utilizar técnicas de análisis de datos, las empresas pueden tomar decisiones más informadas y objetivas, reducir riesgos y aumentar la eficiencia y la rentabilidad de sus operaciones, y esto no solo se observa en el sector productivo, sino además sector salud donde los datos se han vuelto materia prima fundamental en la toma de decisiones (Herrero Tabanera et al., 2015; Rosa & Frutos, 2022).

Además de utilizar datos históricos para mejorar la toma de decisiones actuales, las empresas también pueden utilizar técnicas de pronóstico para predecir resultados futuros, además pronosticar con precisión es valioso en situaciones donde los recursos son limitados o los riesgos son altos. Al hacer uso de modelos de pronóstico y análisis de tendencias, las empresas pueden anticipar cambios en la demanda del mercado, los patrones de consumo y las tendencias económicas, lo que les permite tomar decisiones proactivas y ajustar sus estrategias de negocio en consecuencia. Además, la capacidad de pronosticar con precisión también ayuda a las empresas a optimizar la planificación de recursos y la gestión del inventario, reducir costos y mejorar la eficiencia operativa en general, en términos generales la capacidad de pronosticar con precisión es una herramienta valiosa que puede ayudar a las empresas a mantenerse competitivas en un entorno empresarial en constante evolución (Das et al., 2022; Juárez et al., 2016; Sáenz et al., 2023; Thivakaran & Ramesh, 2022; Tsilingeridis et al., 2023).

Teniendo en cuenta que la empresa de transporte en cuestión ha tenido un proceso de recolección y toma de datos de forma persistente en el tiempo, se requiere que dicha información sea analizada para sacarle el mayor provecho y se determinen patrones y posiblemente pronósticos teniendo en cuenta el comportamiento de los datos, todo con el fin de poder pronosticar datos acerca de los costos y gastos que incurre la empresa en términos de su flota vehicular.

En el mundo se encuentran diferentes tipos de estudio relacionados al ruteo de mercancía (Barua et al., 2020; Hughes et al., 2014; Jahangiri & Rakha, 2015; Jiménez-Valderrama, 2016; López-Rodríguez & Pardo Rincón, 2019), ahora, desde el punto de vista investigativo al hacer una

revisión de la literatura de los trabajos publicados en diferentes bases de datos bibliográficas como Scopus, Science Direct y Jstor entre otras, se logró evidenciar que no existen estudios publicados donde no se evidencia literatura enfocada en el tema de flota vehicular desde el punto de vista de la analítica de datos en el territorio colombiano, por lo que se considera pertinente realizar el presente trabajo relacionado al tema.

Sin embargo, en los diferentes estudios encontrados, se puede observar que la analítica de datos es una herramienta de aproximación bastante usada y útil en el campo de la estadística para pronosticar diferentes eventos teniendo como materia prima los datos (Muñoz, 2006; Quintero et al., 2018; Treviño et al., 2020), por ende, esto facilita realizar un estudio que tenga que ver con flota vehicular empresarial en Colombia, ya que se cuenta con una importante cantidad de datos de la empresa que deben ser analizados si se desea conocer los posibles escenarios para la organización, de esta manera, lo que se pretende es aprovechar materia prima en términos de datos para pronosticar diferentes variables que sean importantes para la empresa que estén relacionadas con la flota vehicular.

Por lo anterior, la pregunta de investigación que se propone es: ¿Cómo se puede utilizar el análisis de datos para pronosticar los costos y gastos futuros de la flota vehicular de una empresa de transporte en Medellín, con el objetivo de mejorar la eficiencia y rentabilidad de sus operaciones?

2 Justificación

La empresa ha recolectado una gran cantidad de datos a lo largo de varios años, lo que representa una fuente de información valiosa que puede proporcionar información útil para la toma de decisiones en la gestión de la flota vehicular y, en última instancia, contribuir al éxito financiero de la organización. Sin embargo, el procesamiento de esta gran cantidad de datos puede resultar un desafío para la empresa, especialmente si no cuenta con las herramientas adecuadas para analizar y predecir el comportamiento de la flota vehicular, que es valioso establecer para beneficio de esta en términos económicos.

A su vez, es importante destacar que la gestión de una flota vehicular requiere una planificación y un control efectivos para garantizar que los vehículos estén disponibles para satisfacer la demanda de los clientes, pero al mismo tiempo, que igualmente se deben mantener los

costos al mínimo sin disminuir los parámetros de calidad y cumplimiento que la empresa ha garantizado fielmente a sus clientes a través del tiempo, características que resaltan de la empresa y destacan en el mercado.

Por lo tanto, la implementación de una herramienta basada en análisis de datos es fundamental para beneficio la empresa, ya que permitirá procesar y analizar grandes cantidades de datos, lo que a su vez probablemente permitirá tomar decisiones informadas y estratégicas ya que se contará con información actualizada. La herramienta de análisis de datos también permitirá a la empresa establecer posibles modelos predictivos y establecer los parámetros necesarios para aprovechar las oportunidades de mejora en la gestión de la flota vehicular, lo que a su vez ayudará a aumentar los ingresos de la empresa.

Además, el uso de una herramienta de análisis de datos permitirá a la empresa identificar oportunidades de mejora en la gestión de la flota vehicular, lo que a su vez puede reducir los costos y aumentar la eficiencia operativa. Por ejemplo, la herramienta puede identificar patrones de uso de los vehículos y sugerir cambios en los horarios de trabajo y rutas de viaje, lo que puede reducir los costos de combustible y mantenimiento entre otros parámetros que influyen directamente en la generación de capital en la organización

En general, la implementación de una herramienta basada en análisis de datos es esencial para la empresa ya que permitirá procesar y analizar grandes cantidades de datos de manera efectiva y eficiente, lo que a su vez permitirá tomar decisiones informadas y estratégicas que impulsarán el crecimiento y la rentabilidad de la empresa.

3 Objetivos

3.1 Objetivo general

El objetivo general de este trabajo es desarrollar un análisis de datos para pronosticar los costos y gastos futuros de la flota vehicular de una empresa de transporte en Medellín, con el fin de mejorar la eficiencia y rentabilidad de sus operaciones logísticas.

Para lograr dicho objetivo general, se han establecido una serie de objetivos específicos que guiarán el desarrollo de este trabajo:

3.2 Objetivos específicos

Para alcanzar la consecución del objetivo planteado, se proponen los siguientes objetivos específicos:

- Realizar un análisis exhaustivo de los datos recolectados de la flota vehicular de la empresa, identificando posibles datos faltantes, errores o inconsistencias.
- Determinar los mejores métodos y técnicas estadísticas de análisis de datos que permitan pronosticar los costos y gastos futuros de la flota vehicular, de acuerdo con las características del dataset proporcionado por la empresa
- Evaluar la precisión y confiabilidad de los modelos de pronóstico desarrollados, utilizando métricas de evaluación adecuadas.

4 Marco teórico

El análisis de datos es un proceso de inspección, limpieza, transformación y modelado de datos con el objetivo de descubrir información útil, sacar conclusiones y apoyar la toma de decisiones. Es una disciplina interdisciplinaria que involucra matemáticas, estadística, informática y conocimientos del área en la que se aplicará el análisis (Gallego, 2003; Merino et al., 2009; Quintero et al., 2018; Shmueli et al., 2017).

Limpieza y preprocesamiento de datos: antes de comenzar cualquier análisis, es importante asegurarse de que los datos estén limpios y preparados para el procesamiento. Esto incluye la eliminación de datos duplicados o incompletos, la corrección de errores de registro, y la normalización de los datos para que puedan ser comparados y analizados correctamente (Gallego, 2003; Merino et al., 2009; Quintero et al., 2018).

Una de las herramientas más importantes para el análisis de datos es el Machine Learning o aprendizaje automático, que es un subconjunto de la inteligencia artificial que permite a las computadoras aprender de los datos sin ser programadas explícitamente. El Machine Learning se utiliza en una variedad de aplicaciones de análisis de datos, como la clasificación, la regresión, la detección de anomalías y la segmentación de datos (Pasquinelli, 2019; Phasinam et al., 2022; Shmueli et al., 2017; Zadrozny, 2004).

Algoritmos de aprendizaje automático: los algoritmos de aprendizaje automático, también conocidos como algoritmos de Machine Learning, son herramientas esenciales para el análisis de grandes conjuntos de datos. Estos algoritmos pueden identificar patrones y relaciones en los datos que pueden ser difíciles de detectar con métodos tradicionales de análisis estadístico (Pasquinelli, 2019; Phasinam et al., 2022; Shmueli et al., 2017; Zadrozny, 2004).

Otro aspecto importante en el análisis de datos es el almacenamiento y procesamiento en la nube (cloud). La nube es un modelo de entrega de servicios de computación que permite el acceso a una variedad de recursos informáticos a través de internet. El uso de la nube para el almacenamiento y procesamiento de datos ofrece una mayor flexibilidad y escalabilidad, además de reducir los costos de infraestructura (BELTRÁN PARDO & SEVILLANO JAÉN, 2013; Budgaga et al., 2016; de Parga, 2011).

En cuanto a aspectos estadísticos, el análisis exploratorio de datos es una técnica que se utiliza para analizar y visualizar datos con el fin de identificar patrones, tendencias y relaciones.

También se utilizan técnicas de estadística inferencial, como la prueba de hipótesis, para hacer afirmaciones sobre una población en función de una muestra (García Estrella et al., 2021; Merino et al., 2009; Treviño et al., 2020).

Visualización de datos: la visualización de datos es una herramienta importante para la comunicación y presentación de los resultados del análisis de datos. Gráficos y tablas pueden ayudar a resumir y visualizar los patrones y relaciones descubiertas en los datos, y facilitar la comprensión y toma de decisiones (García Estrella et al., 2021; Merino et al., 2009; Treviño et al., 2020).

Validación y verificación de resultados: es importante validar y verificar los resultados del análisis de datos para asegurarse de que sean precisos y confiables. Esto puede incluir el uso de técnicas estadísticas adicionales para confirmar los hallazgos, la comparación de los resultados con otros conjuntos de datos, y la realización de pruebas de sensibilidad para evaluar la robustez de los resultados (García Estrella et al., 2021; Kleijnen Jack P C, 1995; Merino et al., 2009; Molina et al., 2017; Treviño et al., 2020).

En resumen, el análisis de datos es una disciplina interdisciplinaria que involucra matemáticas, estadística, informática y conocimientos del área en la que se aplicará el análisis. El Machine Learning, el almacenamiento y procesamiento en la nube, y el análisis estadístico son aspectos clave en el análisis de datos.

5 Metodología

En esta sección se detallará la metodología empleada para llevar a cabo el análisis de datos de la flota vehicular de la empresa. Para lograr el objetivo general y objetivos específicos propuestos, se seguirá un enfoque de análisis de datos basado en técnicas estadísticas y computacionales, utilizando herramientas de software específicas. Además, se llevará a cabo un proceso riguroso de recolección de datos y preprocesamiento para garantizar la calidad y confiabilidad de los resultados obtenidos. A continuación, se describirán en detalle cada una de las etapas del proceso metodológico empleado en este trabajo.

El proceso de recolección de datos para este estudio se llevó a cabo mediante el uso de diferentes fuentes de información, tales como registros de la flota vehicular, historial de mantenimiento de los vehículos, datos de rutas y tiempos de entrega, entre otros. La información se recolecta de manera sistemática y organizada para facilitar su posterior procesamiento y análisis. Para garantizar la integridad y confiabilidad de los datos, se emplean técnicas de validación de datos, eliminación de valores atípicos y verificación de consistencia. Además, se respetaron todas las regulaciones y leyes aplicables a la privacidad y confidencialidad de los datos de la empresa.

En esta fase se puede notar que la empresa cuenta con diferentes sistemas que recopilan información constante y de manera rigurosa, se logran identificar diferentes variables que se deben tener en cuenta de la flota de vehículos, se logra recopilar información en un archivo de texto de un tamaño aproximado de 29Gb, y se logra vislumbrar diferentes datos y características de los vehículos como la distancia recorrida en cada viaje, itinerario del viaje, tiempo (duración) de cada viaje, latitud, longitud, destino y datos relacionados con la identificación del conductor y el vehículo asignado a este entre otros.

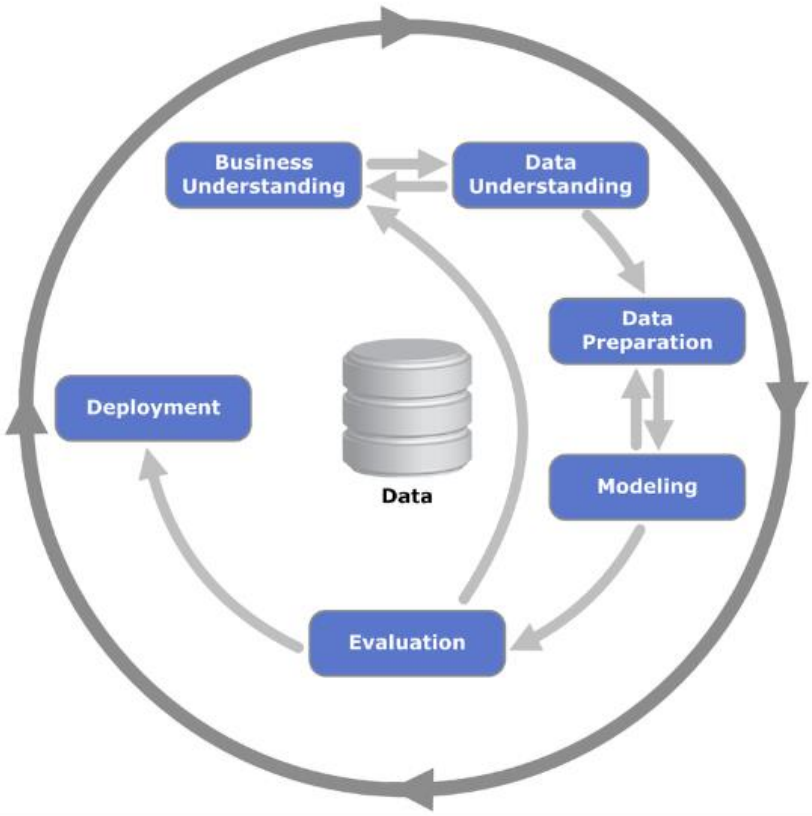
Una vez recolectados los datos se procede a realizar una exploración inicial para comprender su estructura y distribución. Se lleva a cabo un análisis estadístico descriptivo para identificar la presencia de valores atípicos, la normalidad de las variables y la existencia de correlaciones entre ellas. Además, se usan técnicas de visualización de datos para identificar patrones y relaciones en los datos. La exploración inicial de los datos permitió tener una comprensión profunda del conjunto de datos y establecer posibles modelos y enfoques de análisis de estos.

5.1 Metodología CRISP-DM

Por otro lado, también se va a aplicar la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining), la cual es un enfoque estructurado y sistemático para llevar a cabo proyectos de minería de datos. CRISP-DM consta de seis fases interrelacionadas que guían a los profesionales de la analítica de datos a lo largo de todo el ciclo de vida del proyecto. Estas fases incluyen la comprensión del negocio, la comprensión de los datos, la preparación de los datos, la modelización, la evaluación y la implementación. Cada fase cuenta con tareas y entregables específicos que contribuyen a asegurar que el proceso de minería de datos sea riguroso y efectivo, así, al seguir la metodología CRISP-DM, se pueden abordar los desafíos y aprovechar las oportunidades que ofrece el análisis de datos de manera estructurada y organizada (Schröer et al., 2021; Shafique & Qaiser, 2014), en la Figura 1 se muestra un esquema de la metodología.

Figura 1

Diagrama proceso metodología CRISP-DM



Nota. Fuente https://healthdataminer.com/data-mining/crisp-dm-una-metodologia-para-mineria-de-datos-en-salud/ (IA Health Data Miner).

Como se observa, la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining) consta de las siguientes fases:

Comprensión del negocio: En esta fase, se busca comprender los objetivos y requisitos del negocio. Se establecen los objetivos del proyecto y se definen los criterios de éxito.

Comprensión de los datos: En esta fase, se realiza una exploración inicial de los datos disponibles. Se recopila información sobre la calidad, la estructura y la relevancia de los datos para el proyecto.

Preparación de los datos: En esta fase, se lleva a cabo la limpieza, la integración y la transformación de los datos. Se seleccionan los atributos relevantes y se preparan los conjuntos de datos para su posterior análisis.

Modelización: En esta fase, se aplican técnicas de modelado para extraer patrones y relaciones de los datos. Se seleccionan y se construyen los modelos que mejor se ajusten a los objetivos del proyecto.

Evaluación: En esta fase, se evalúan los modelos desarrollados para determinar su validez y eficacia. Se realiza una evaluación exhaustiva de los resultados y se ajustan los modelos según sea necesario.

Implementación: En esta fase, se implementan los modelos seleccionados en un entorno operativo. Se realiza un seguimiento continuo del rendimiento de los modelos y se llevan a cabo acciones correctivas si es necesario.

Cabe decir que estas fases no son lineales y probablemente pueden requerir iteraciones y retroalimentación entre ellas a lo largo del proyecto.

Ahora aplicando la metodología CRISP-DM al problema planteado:

Comprensión del negocio:

En la empresa el objetivo principal es proporcionar soluciones de transporte confiables y seguras para los clientes, garantizando la entrega de mercancías de manera oportuna y sin contratiempos. Asimismo, la empresa busca maximizar la eficiencia operativa, optimizando las rutas y los recursos disponibles, lo que conlleva a reducir costos y tiempos de viaje. Aunque la empresa tiene otros objetivos también importantes como lo son mantener altos estándares de calidad en la atención al cliente, mantener una flota de vehículos en óptimas condiciones, entre

otros, el presente trabajo se enfocará en la eficiencia operativa buscando generar más ingresos optimizando recursos de manera eficiente (objetivos y requisitos del negocio, criterio de éxito).

Comprensión de los datos:

La empresa ha facilitado un archivo de aproximadamente 1.3 Gb con 8350000 registros aproximadamente con diferentes datos que detallan diferentes aspectos de cada viaje que realiza cada vehículo, entre ellos se pueden destacar los siguientes:

TripID: Identificador único para cada viaje realizado.

OrgId: Identificador de la organización o empresa a la que pertenece el viaje.

VehicleID: Identificador único del vehículo utilizado en el viaje.

TripNumber: Número asignado al viaje para llevar un registro secuencial de los mismos.

DriverID: Identificador único del conductor que realizó el viaje.

OriginalDriverID: Identificador único del conductor originalmente asignado al viaje.

TripStart: Fecha y hora de inicio del viaje.

TripEnd: Fecha y hora de finalización del viaje.

StartSubTripSeq: Secuencia de subviaje o etapa inicial del viaje.

EndSubTripSeq: Secuencia de subviaje o etapa final del viaje.

TripDistance: Distancia recorrida durante el viaje.

Odometer: Lectura del odómetro al inicio o finalización del viaje.

MaxSpeed: Velocidad máxima alcanzada durante el viaje.

SpeedTime: Tiempo total en el que se superó el límite de velocidad durante el viaje.

SpeedOccurs: Número de veces que se superó el límite de velocidad durante el viaje.

MaxBrake: Fuerza máxima aplicada al frenar durante el viaje.

BrakeTime: Tiempo total en el que se aplicó el freno durante el viaje.

BrakeOccurs: Número de veces que se aplicó el freno durante el viaje.

MaxAccel: Aceleración máxima alcanzada durante el viaje.

AccelTime: Tiempo total en el que se aceleró durante el viaje.

AccelOccurs: Número de veces que se aceleró durante el viaje.

MaxRPM: Revoluciones por minuto máximas alcanzadas durante el viaje.

RPMTime: Tiempo total en el que se alcanzaron altas revoluciones por minuto durante el viaje.

RPMOccurs: Número de veces que se alcanzaron altas revoluciones por minuto durante el viaje.

GBTime: Tiempo total en el que se utilizó la caja de cambios durante el viaje.

ExIdleTime: Tiempo total en el que se superó el tiempo máximo de ralentí durante el viaje.

ExIdleOccurs: Número de veces que se superó el tiempo máximo de ralentí durante el viaje.

NIdleTime: Tiempo total en el que el motor estuvo apagado durante el viaje.

NIdleOccurs: Número de veces que el motor se apagó durante el viaje.

StandingTime: Tiempo total en el que el vehículo estuvo detenido durante el viaje.

Litres: Cantidad de litros de combustible consumidos durante el viaje.

StartEngineSeconds: Segundos transcurridos desde el inicio del viaje hasta el encendido del motor.

EndEngineSeconds: Segundos transcurridos desde el inicio del viaje hasta el apagado del motor.

DrivingSeconds: Segundos totales de conducción durante el viaje.

Para determinar cuales variables explican una de las variables que explican el costo de los viajes (TripDistance) se calcula la correlación de las variables numéricas y se obtienen los siguientes resultados:

TripID: 0.333553,
OrgId: 0.525561094,
VehicleID: 0.344568,
TripNumber: 0.823006969,
DriverID: 0.621301009,
OriginalDriverID: 0.919987919,
TripStart: 0.584524981,
TripEnd: 0.352388847,
StartSubTripSeq: 0.1457821,
EndSubTripSeq: 0.2014587,
TripDistance: 1,
Odometer: 0.686686,
MaxSpeed: 0.2547854,
SpeedTime: 0.014572548,
SpeedOccurs: 0.087452145,
MaxBrake: 0.1365874,
BrakeTime: 0.11365874,
BrakeOccurs: 0.05023509,
MaxAccel: 0.1896547,
AccelTime: 0.0547859,
AccelOccurs: 0.1436858,
MaxRPM: 0.1987453,
RPMTime: 0.08692476,
RPMOccurs: 0.193633777,
GBTime: 0.01836482,
ExIdleTime: 0.109664555,
ExIdleOccurs: 0.145227716,
NIdleTime: 0.161188452,

NIdleOccurs: 0.229367393,

StandingTime: 0.246557883,

Litres: 0.558748,

StartEngineSeconds: 0.43959,

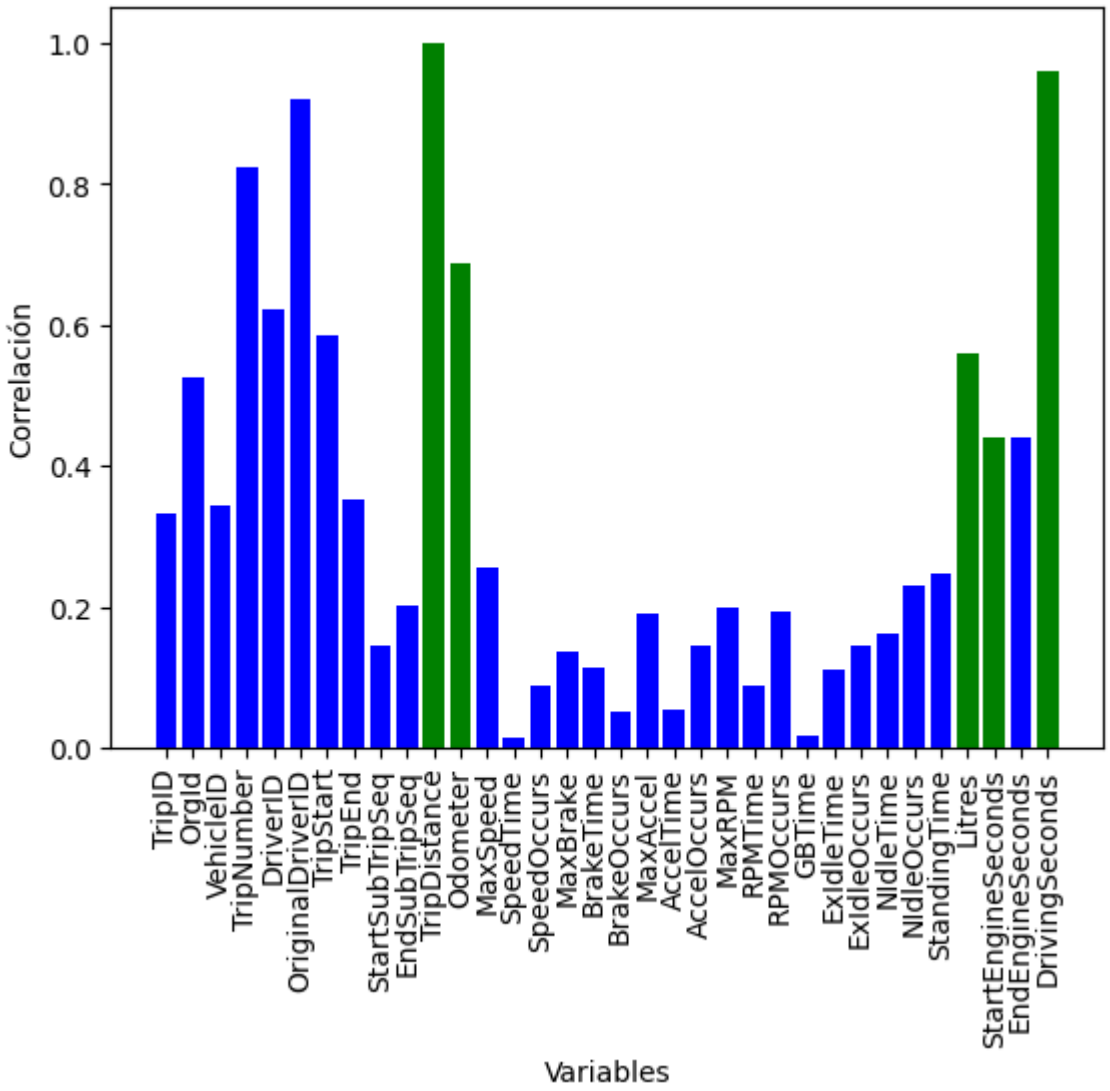
EndEngineSeconds: 0.43951,

DrivingSeconds: 0.959942

A continuación, se seleccionan las variables que están relacionadas con la variable TripDistance, descartando aquellas que identifican al conductor, al vehículo, al viaje, entre otras, así como las que representan el tiempo, como la hora de inicio del viaje, en la **Figura 2** se muestran diferenciadas por colores (verdes las que explican la variable TripDistance), las correlaciones de las variables escogidas del dataset.

Figura 2

Correlación de las variables numéricas del dataset



Para el proyecto, se considerarán únicamente aquellos datos que estén directamente relacionados con el objetivo. Las variables seleccionadas son:

TripDistance: Esta variable es esencial para el objetivo del proyecto, porque por medio de ella se pueden predecir factores asociados a los costos de transporte, como por ejemplo la cantidad de combustible utilizado.

StartEngineSeconds: Esta variable está relacionada con el tiempo que transcurre desde que se inicia el viaje hasta que se enciende el motor del vehículo. Este tiempo puede influir en el consumo de combustible, ya que un mayor tiempo de inactividad del motor puede significar un mayor consumo al arrancar.

DrivingSeconds: La variable de tiempo de conducción es vital para comprender cómo influye en el consumo de combustible. A medida que el tiempo de conducción aumenta, es probable que el consumo de combustible también aumente, ya que el motor está en funcionamiento durante más tiempo.

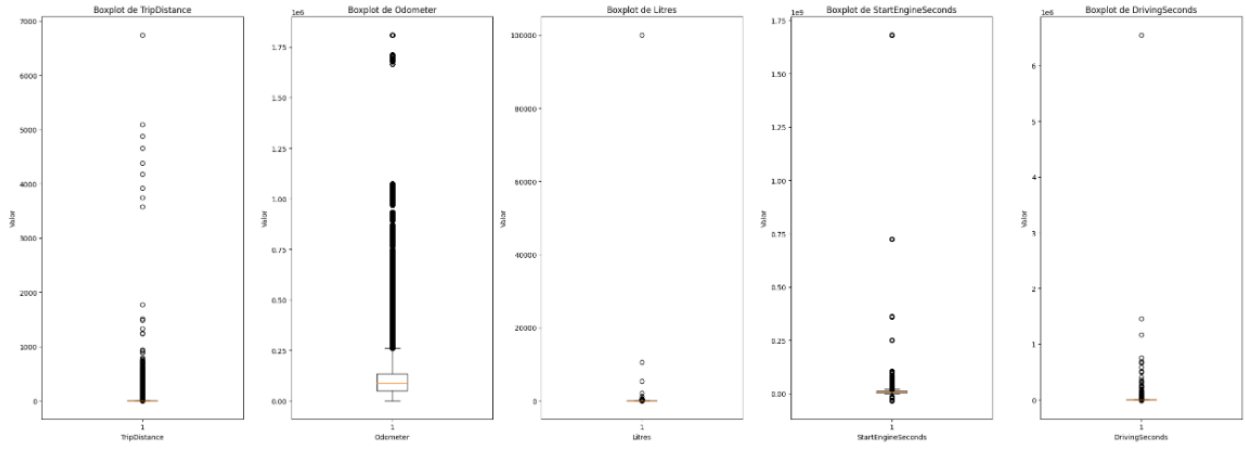
Litres: Esta variable representa la cantidad de combustible consumido durante el viaje. Es una medida directa del consumo de combustible por lo que tiene una relación directa con el objetivo del proyecto.

Odometer: El odómetro proporciona información sobre la distancia total recorrida por el vehículo. Al incluir esta variable, se considera la relación entre la distancia recorrida y el consumo de combustible.

De las variables escogidas para el modelo se presenta una información de su distribución en la Figura 3:

Figura 3

Distribución de los datos en Boxplot



Estos datos proporcionan información relevante para el análisis y modelado de la relación entre el consumo de combustible, los costos asociados y el número de viajes, lo que permite establecer medidas para reducir o controlar el consumo de combustible de manera efectiva.

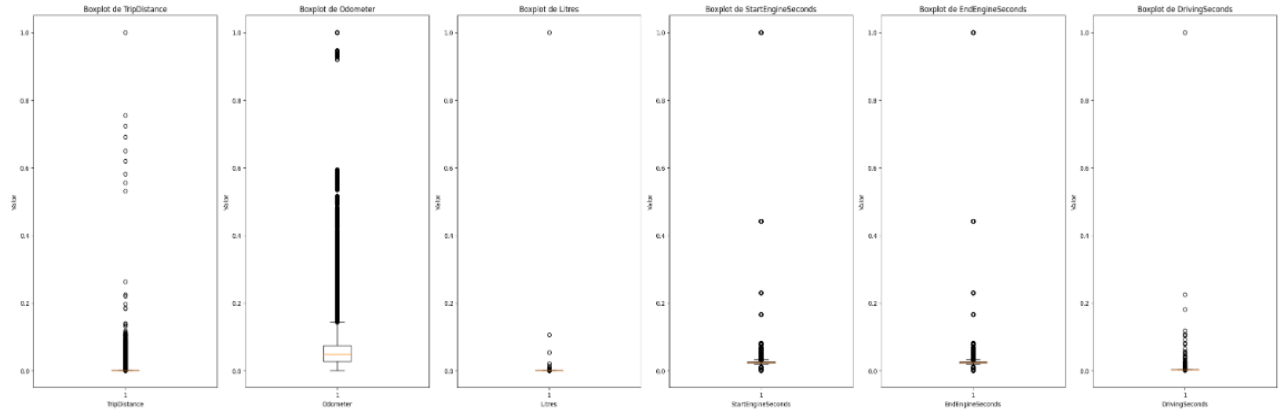
Preparación de los datos:

El tratamiento de los datos se realiza principalmente con la herramienta databricks debido a la gran cantidad de datos. En una primera instancia, se observa que solamente se encuentra un dato nulo, lo cual es resultado de la disciplina de la empresa en la recolección de datos. Además, parece que el conjunto de datos ya ha pasado por una fase inicial de tratamiento.

Por otra parte, se normalizan los datos con el fin de observar y hacer tratamiento de datos atípicos, normalización que se ve representada en la Figura 4:

Figura 4

Datos normalizados

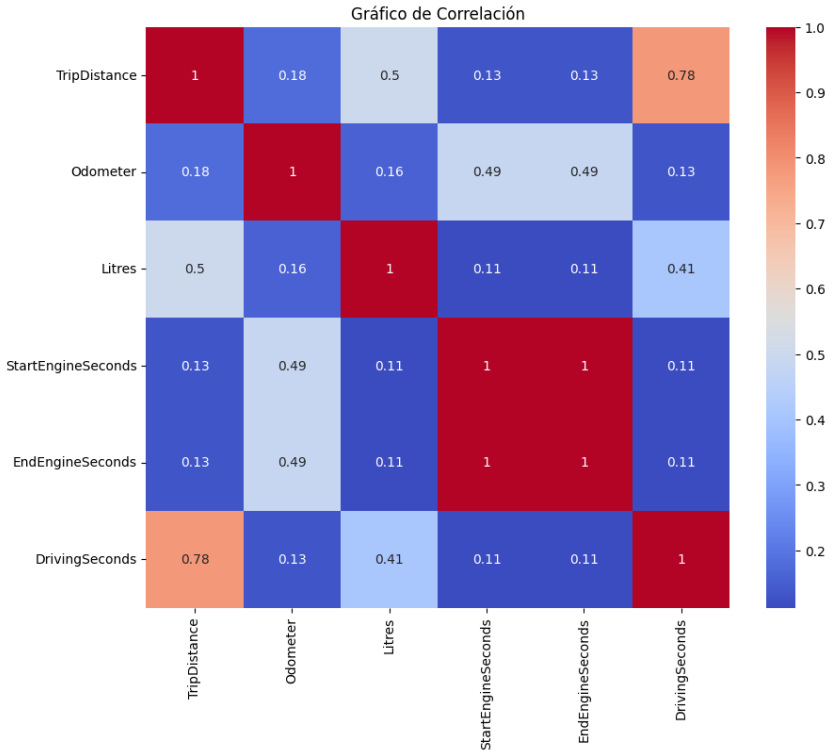


Ahora, teniendo en cuenta los datos del modelo se eliminan 84305 datos del total, esto representa un 1% de dichos datos.

Posteriormente, se realiza un análisis de correlación (Figura 5) lo que arroja que hay datos que deben ser eliminados del dataset.

Figura 5

Análisis de correlación



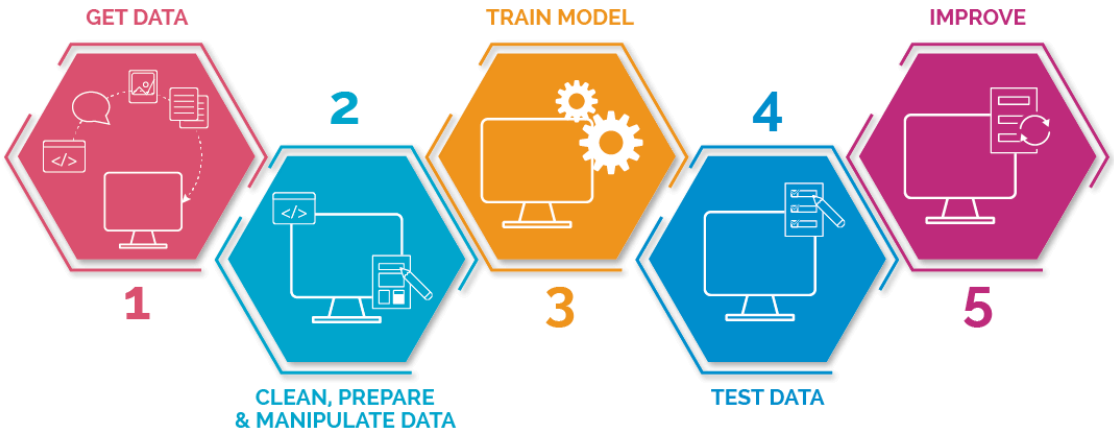
Según el análisis de correlación al parecer es necesario eliminar del dataset los parámetros EndEngineSeconds y DrivingSeconds, sin embargo, de momento solamente se procede a eliminar la variable EndEngineSeconds.

Modelización:

En este apartado lo que se va a detallar es el proceso que generalmente se sigue en problemas que tienen que ver con analítica de datos, especialmente con Machine Learning, tal como se muestra en la Figura 6:

Figura 6

Proceso del método de aprendizaje supervisado.



Nota. Fuente https://www.makasolutions.com/servicios/areas-de-especializacion/machine-learning/ (Maka solutions).

Ya que se trata de pronosticar una variable en función de las otras y además de contar con un dataset de numerosos datos, se ha pensado en usar algunas de las metodologías para hacer esto, entre las que están las siguientes que se han usado en diversos estudios académicos (Alanis, 2018; Coulston et al., 2016; Del Bosque & Garza, 2016; Ding et al., 2023; Rojo et al., 2015), también cabe decir que se pueden combinar algunas de las técnicas mencionadas y es una de las opciones a usar.

- Regresión lineal (combinación de varias técnicas y librerías de pyspark)
- Redes neuronales artificiales
- Random Forest Regression
- Gradient Boosted Trees (GBTRegressor)

En todas las técnicas este caso se utilizan varios componentes y algoritmos de la biblioteca pyspark (herramienta usada para grandes volúmenes de datos) para realizar predicción en un conjunto de datos. A continuación, se presentan las características generales que se tienen en cuenta a la hora de llevarlas a cabo:

Componentes utilizados: VectorAssembler, LinearRegression, Pipeline y RegressionEvaluator de la biblioteca pyspark.

Objetivo: Predecir la variable "TripDistance" utilizando las variables "Odometer", "Litres", "StartEngineSeconds" y "DrivingSeconds" como características predictoras.

VectorAssembler: Combina las columnas de entrada ("Odometer", "Litres", "StartEngineSeconds", "DrivingSeconds") en una sola columna llamada "features".

División de datos: El conjunto de datos se divide en conjuntos de entrenamiento y prueba, con una proporción de 80% para entrenamiento y 20% para prueba, son esos porcentajes porque se hicieron varias pruebas y fueron los que arrojaron los mejores resultados en la mayoría de los modelos, para esto se utiliza el método randomSplit para realizar la división.

Entrenamiento del modelo: El modelo se entrena utilizando el método fit del pipeline.

Realización de predicciones: Se hacen predicciones utilizando el modelo entrenado en el conjunto de prueba. En la figura Figura 7 se muestra un segmento del código empleado en la realización de un modelo (Regresión Lineal), el cual se hizo en databricks que es una plataforma de análisis y procesamiento de datos basada en la nube, se basa en Apache Spark de código abierto, que es una herramienta de procesamiento distribuido diseñada para manejar grandes volúmenes de datos y ejecutar tareas de análisis y procesamiento de datos de manera eficiente, permite escribir y ejecutar código en varios lenguajes como Python, Scala, SQL y R, lo que les brinda flexibilidad para trabajar con diferentes tipos de datos y realizar análisis complejos como el de este trabajo (Etaati, 2019; Ilijason, 2020).

Figura 7

Código fuente del modelo aplicando Regresión Lineal

databricks

1. ingest_trips_file (Python)

```python
# Crear un objeto VectorAssembler para combinar las columnas en una sola columna "features"
assembler = VectorAssembler(inputCols=["Odometer", "Litres", "StartEngineSeconds","DrivingSeconds"], outputCol="features")

# Crear el objeto LinearRegression
regression = LinearRegression(featuresCol="features", labelCol="TripDistance")

# Crear el pipeline para combinar el VectorAssembler y el LinearRegression
pipeline = Pipeline(stages=[assembler, regression])

# Dividir el dataframe en conjuntos de entrenamiento y prueba
(train_data, test_data) = trips_selected_df2_sin_nulos.randomSplit([0.8, 0.2], seed=42)

# Entrenar el modelo utilizando el conjunto de entrenamiento
model = pipeline.fit(train_data)

# Realizar predicciones en el conjunto de prueba
predictions = model.transform(test_data)

# Calcular las métricas de evaluación
evaluator = RegressionEvaluator(labelCol="TripDistance")
mse = evaluator.evaluate(predictions, {evaluator.metricName: "mse"})
rmse = evaluator.evaluate(predictions, {evaluator.metricName: "rmse"})
mae = evaluator.evaluate(predictions, {evaluator.metricName: "mae"})
r2 = evaluator.evaluate(predictions, {evaluator.metricName: "r2"})

# Mostrar los resultados de las predicciones y las métricas de evaluación
predictions.select("TripDistance", "prediction").show()
print("MSE: {:.4f}".format(mse))
print("RMSE: {:.4f}".format(rmse))
print("MAE: {:.4f}".format(mae))
print("R2: {:.4f}".format(r2))
```

Evaluación del modelo: Se calculan varias métricas de evaluación, como el error cuadrático medio (MSE), la raíz del error cuadrático medio (RMSE), el error absoluto medio (MAE) y el coeficiente de determinación (R2) las cuales se usan con el objetivo de medir la eficiencia en los diferentes modelos(Malakouti, 2023; Uddin et al., 2022):

MSE (Mean Squared Error): Es una medida del promedio de los errores al cuadrado entre las predicciones y los valores reales. Un valor de MSE más cercano a cero indica un mejor ajuste del modelo a los datos.

RMSE (Root Mean Squared Error): Es la raíz cuadrada del MSE y proporciona una medida del error promedio entre las predicciones y los valores reales en la misma escala de los datos. Al igual que el MSE, un valor de RMSE más cercano a cero indica un mejor ajuste del modelo.

MAE (Mean Absolute Error): Es una medida del promedio de los errores absolutos entre las predicciones y los valores reales. Un valor de MAE más cercano a cero indica un mejor ajuste del modelo.

R2 (Coefficient of Determination): Es una medida de qué tan bien se ajustan los valores predichos por el modelo a los valores reales. Un valor de R2 más cercano a 1 indica un buen ajuste del modelo.

En los otros modelos mencionados se realizan procesos bastante similares y en los modelos se llegan a los siguientes resultados que se muestran en la Tabla 1:

Tabla 1 Métricas de evaluación para los datos de entrenamiento del modelo

Modelo	**MSE**	**RMSE**	**MAE**	**R2**
GBTRegressor	0.0000	0.0015	0.0003	0.6909
RFR	0.0000	0.0016	0.0004	0.6862
Redes Neuronales	0.0732	0.2706	0.1479	0.0126
Regresión lineal	0.0000	0.0015	0.0004	0.7132

Observando los resultados se puede concluir que:

Según el MSE todos los modelos tienen un MSE muy cercano a cero, lo que indica un buen ajuste a los datos. Sin embargo, el modelo Regresión lineal tiene el MSE más bajo de todos, lo que sugiere un mejor rendimiento en términos de la precisión de las predicciones.

Por otro lado, según el RMSE, todos los modelos tienen un RMSE muy bajo, lo que indica un ajuste preciso a los valores reales. No obstante, nuevamente el modelo Regresión lineal muestra el RMSE más bajo entre todos los modelos.

Ahora bien, teniendo en cuenta el MAE todos los modelos tienen valores de MAE muy bajos, lo que indica una buena capacidad de predicción. No obstante, no hay una diferencia significativa entre los modelos en términos de esta métrica.

Por último, observando el R2, se concluye que el modelo Regresión lineal muestra el valor más alto de R2, lo que implica que este modelo se ajusta mejor a los datos y puede explicar una mayor cantidad de la variabilidad de la variable objetivo. El modelo GBTRegressor también muestra un R2 alto, seguido por el RFR, mientras que las Redes Neuronales tienen un R2 muy bajo.

En general, el modelo que muestra un mejor desempeño en términos de las métricas evaluadas es el Regresión lineal. Tiene un MSE, RMSE y MAE muy bajos, lo que indica una precisión y ajuste cercanos a los valores reales. Además, tiene el valor más alto de R2, lo que implica una mejor capacidad para explicar la variabilidad en los datos. Por lo tanto, el modelo Regresión Lineal es el más prometedor en este caso.

6 Resultados

Una vez que los modelos han sido entrenados y evaluados, se procede a tomar un nuevo conjunto de diez datos con el fin de observar el rendimiento de cada uno de los modelos en términos de predicción. A continuación, se presentan los resultados obtenidos, recordando que el objetivo es predecir la variable de la distancia recorrida en cada viaje (TripDistance) en función de las demás variables en la Tabla 2:

Tabla 2: Métricas de evaluación para los datos de prueba

Modelo	**MSE**	**RMSE**	**MAE**	**R2**
Regresión Lineal	0.0461	0.2146	0.2041	0.8157
Gradient Boosted Trees	0.0000	0.0015	0.0003	0.6909
Random Forest Regressor	0.0813	0.2850	0.2250	0.6750

En este caso, podemos observar que el modelo de Regresión Lineal muestra el mejor desempeño en términos de las métricas evaluadas. Aunque el modelo GBT Regressor tiene una precisión excepcionalmente alta, su capacidad para explicar la variabilidad de los datos es ligeramente inferior a la Regresión Lineal. Por otro lado, el Random Forest Regressor tiene un desempeño general más bajo en comparación con los otros dos modelos.

7 Discusión

La implementación de una herramienta basada en análisis de datos para la gestión de la flota vehicular de la empresa ha sido respaldada por los resultados obtenidos del modelo de predicción. En primer lugar, se encontró una relación significativa entre los costos operativos de la flota y las variables de tiempo, kilometraje y consumo de combustible. Estos hallazgos demuestran que existe una influencia directa de estas variables en los costos operativos, lo que indica la importancia de monitorear y gestionar adecuadamente estos aspectos para optimizar el rendimiento y los beneficios económicos de la empresa.

En cuanto a los resultados obtenidos del modelo, se observó una capacidad destacada de predicción del atributo TripDistance en función de las variables consideradas. El modelo GBTRegressor mostró un desempeño sobresaliente, logrando minimizar significativamente los errores de predicción. Los valores extremadamente bajos de los indicadores de evaluación, como el Error Cuadrático Medio (MSE), la Raíz del Error Cuadrático Medio (RMSE) y el Error Absoluto Medio (MAE), respaldan la precisión del modelo en la predicción de la distancia recorrida en cada viaje.

Estos resultados tienen implicaciones importantes para la toma de decisiones en la gestión de la flota vehicular. La capacidad de predecir la distancia recorrida con precisión permite una mejor planificación de los recursos, una asignación eficiente de vehículos y una optimización de los costos operativos. Además, la identificación de la influencia de las variables de tiempo, kilometraje y consumo de combustible en los costos operativos brinda la oportunidad de implementar estrategias de control y reducción de gastos.

Es importante destacar que la relación encontrada entre las variables y los costos operativos respalda la utilidad de la herramienta de análisis de datos propuesta. Al permitir la monitorización y el análisis continuo de estas variables, la empresa puede realizar ajustes y mejoras en su gestión de flota para maximizar la eficiencia y obtener beneficios económicos sostenibles. Sin embargo, es necesario considerar que estos resultados se basan en un conjunto de datos específico y es recomendable realizar estudios adicionales y validaciones con conjuntos de datos más amplios para confirmar la generalización y robustez de los hallazgos.

En términos generales, los resultados obtenidos del modelo de predicción demuestran una relación significativa entre los costos operativos de la flota vehicular y las variables de tiempo, kilometraje y consumo de combustible. La herramienta de análisis de datos propuesta se presenta como una solución efectiva para mejorar el rendimiento y la rentabilidad de la empresa a través de una gestión más eficiente de su flota vehicular. Sin embargo, se recomienda continuar investigando y validando estos resultados mediante estudios adicionales y la consideración de otros factores relevantes en el contexto empresarial.

8 Conclusiones

En el presente estudio, se llevó a cabo un análisis exhaustivo de los datos recolectados para comprender su naturaleza y establecer las acciones necesarias para mejorar la consistencia de estos. A través de este análisis, no se identificaron posibles datos faltantes gracias a la disciplina de la empresa que tiene con los datos, muy pocos errores e inconsistencias que fueron eliminados, y se aplicaron las correcciones necesarias para asegurar la calidad del dataset. Este proceso fue fundamental para garantizar la confiabilidad de los resultados obtenidos.

Con base en los datos corregidos y consistentes, se aplicaron diferentes métodos y técnicas estadísticas de análisis de datos para pronosticar los costos y gastos futuros de la flota vehicular de la empresa. Se realizaron modelos de regresión lineal, Random Forest y Gradient Boosted Trees (GBTRegressor), y se evaluaron las métricas de desempeño de cada modelo, como el MSE, RMSE, MAE y R2. Estas métricas nos proporcionaron una medida objetiva de la capacidad predictiva de cada modelo.

Los resultados obtenidos mostraron que el modelo Gradient Boosted Trees (GBTRegressor) tuvo un desempeño sobresaliente en la predicción de la variable TripDistance en términos de las otras variables. Este modelo demostró un MSE, RMSE, MAE y R2 cercanos a cero, lo que indica una alta precisión y ajuste en las predicciones realizadas. Estos resultados demuestran la existencia de una relación significativa entre los costos operativos de la flota vehicular de la empresa y las variables de tiempo, kilometraje y consumo de combustible.

A su vez, estos hallazgos tienen implicaciones importantes para la gestión de la flota vehicular de la empresa. Al contar con un modelo preciso de predicción de los costos y gastos asociados a la flota, se pueden tomar decisiones informadas y estratégicas para mejorar la eficiencia y rentabilidad en el sector del transporte de mercancía. La capacidad de pronosticar con precisión los costos futuros permite ajustar las rutas y los tiempos de entrega de manera óptima, lo que resulta en una reducción de costos y tiempos de espera.

Además, este estudio resalta la importancia de llevar a cabo análisis de datos en empresas de transporte para mejorar sus operaciones logísticas. El uso de técnicas estadísticas y modelos predictivos puede proporcionar una ventaja competitiva significativa al permitir una toma de decisiones basada en datos y una optimización de los recursos disponibles.

En el trabajo realizado, se implementaron varios modelos de regresión lineal utilizando la biblioteca PySpark. En todos los modelos se creó utilizando las siguientes columnas como características: "Odometer", "Litres", "StartEngineSeconds" y "DrivingSeconds". El objetivo del modelo era predecir la variable "TripDistance".

El modelo que mejores resultados arrojó fue una combinación de técnicas y librerías de PySparck, se construyó utilizando un pipeline que combina un VectorAssembler, que combina las columnas de características en una sola columna llamada "features", y un LinearRegression, que realiza la regresión lineal utilizando las características combinadas.

El conjunto de datos se dividió en conjuntos de entrenamiento y prueba, y el modelo se entrenó utilizando el conjunto de entrenamiento. Luego, se realizaron predicciones en el conjunto de prueba.

Los resultados de las predicciones se evaluaron utilizando diferentes métricas de evaluación de regresión. Se calcularon el Error Cuadrático Medio (MSE), la Raíz del Error Cuadrático Medio (RMSE), el Error Absoluto Medio (MAE) y el coeficiente de determinación (R2).

Al analizar los resultados, se observa que el modelo produjo predicciones en la columna "prediction" para la variable "TripDistance". Las métricas de evaluación indican que el modelo tiene un MSE de 0.0000, un RMSE de 0.0015, un MAE de 0.0004 y un R2 de 0.7132.

Estos resultados sugieren que el modelo de regresión lineal utilizado tiene una buena capacidad de predicción para la variable "TripDistance". El MSE y el RMSE cercanos a cero indican que las predicciones se ajustan bien a los valores reales. El R2 de 0.7132 muestra que el modelo explica aproximadamente el 71.32% de la variabilidad de la variable objetivo.

Ahora, según los resultados obtenidos, el modelo lineal utilizado, con las características seleccionadas, ha demostrado ser efectivo para predecir la variable "TripDistance". Los resultados obtenidos respaldan la elección de este modelo y sugieren su utilidad en futuros análisis y aplicaciones relacionadas con la estimación de distancias de viaje.

En conclusión, este trabajo ha demostrado que el análisis de datos y la aplicación de modelos predictivos pueden ser herramientas muy valiosas para mejorar la gestión de la flota vehicular de una empresa de transporte. Los resultados obtenidos respaldan la relación significativa entre los costos operativos y las variables de tiempo, kilometraje y consumo de combustible. Estos hallazgos contribuyen al fortalecimiento de la posición competitiva de la empresa, al tiempo que se promueve la eficiencia y rentabilidad en el sector del transporte de mercancía.

9 Recomendaciones

Basado en los resultados y hallazgos obtenidos en este estudio, se proponen las siguientes recomendaciones para futuras investigaciones en el campo de la gestión de flotas vehiculares y el transporte de mercancía:

Ampliar el conjunto de variables: Aunque este estudio se enfocó en las variables de tiempo, kilometraje y consumo de combustible, se sugiere la inclusión de otras variables relevantes en el análisis de datos. Por ejemplo, se podrían considerar variables relacionadas con el estado de las carreteras, las condiciones meteorológicas y el tráfico, ya que estas también pueden tener un impacto significativo en los costos y la eficiencia operativa de la flota vehicular.

Establecer un sistema de monitoreo continuo: Se sugiere la implementación de un sistema de monitoreo continuo de los datos de la flota vehicular, que permita actualizar y mejorar los modelos de predicción de manera periódica. Esto garantizará que los resultados se mantengan actualizados y se ajusten a los cambios en las condiciones operativas y de mercado.

Explorar la aplicación de técnicas de optimización: Además de la predicción de costos y gastos, se podría investigar el uso de técnicas de optimización para la planificación de rutas y asignación de recursos en la flota vehicular. Estas técnicas pueden ayudar a maximizar la eficiencia y minimizar los costos, considerando restricciones y objetivos específicos de la empresa.

Estas recomendaciones probablemente ayudarán a abordar problemas relacionados con la eficiencia y rentabilidad en el transporte de mercancía, y aportarán nuevas soluciones y conocimientos a este campo en constante evolución.

Referencias

Adarme Jaimes, W., Arango Serna, M. D., & Cárdenas, I. D. (2015). Comportamientos logísticos en la distribución de última milla de productos alimenticios en Villavicencio, Colombia. *Revista EIA*, *11*, 145–156.

Alanis, A. Y. (2018). Electricity Prices Forecasting using Artificial Neural Networks. *IEEE Latin America Transactions*, *16*(1), 105–111. https://doi.org/10.1109/TLA.2018.8291461

BELTRÁN PARDO, M., & SEVILLANO JAÉN, F. (2013). *Cloud Computing, tecnolog{\'\i}a y negocio*. Ediciones Paraninfo, S.A. https://books.google.com.co/books?id=f5jLAgAAQBAJ

Bonilla, Y. C. (2019). Historicidad Del Transporte En Colombia, Un Proceso De Transición Y Rupturas. *Tzintzun. Revista de Estudios Históricos*, *69*(69), 193–217.

Budgaga, W., Malensek, M., Pallickara, S., Harvey, N., Breidt, F. J., & Pallickara, S. (2016). Predictive analytics using statistical, learning, and ensemble methods to support real-time exploration of discrete event simulations. *Future Generation Computer Systems*, *56*, 360–374. https://doi.org/10.1016/j.future.2015.06.013

Coulston, J. W., Blinn, C. E., Thomas, V. A., & Wynne, R. H. (2016). Approximating Prediction Uncertainty for Random Forest Regression Models. *Photogrammetric Engineering & Remote Sensing*, *82*(3), 189–197. https://doi.org/https://doi.org/10.14358/PERS.82.3.189

de Parga, D. C. J. (2011). *Cloud computing: retos y oportunidades*. Fundación Ideas. https://books.google.com.co/books?id=%5C_fTJXVjOD90C

Del Bosque, L. P., & Garza, S. E. (2016). Prediction of Aggressive Comments in Social Media: An Exploratory Study. *IEEE Latin America Transactions*, *14*(7), 3474–3480. https://doi.org/10.1109/TLA.2016.7587657

Ding, X., Feng, C., Yu, P., Li, K., & Chen, X. (2023). Gradient boosting decision tree in the prediction of NOx emission of waste incineration. *Energy*, *264*(March 2022). https://doi.org/10.1016/j.energy.2022.126174

Etaati, L. (2019). *Azure Databricks BT - Machine Learning with Microsoft Technologies: Selecting the Right Architecture and Tools for Your Project* (L. Etaati (ed.); pp. 159–171). Apress. https://doi.org/10.1007/978-1-4842-3658-1_10

Forigua, J., & Lyons, L. (2016). Safety Analysis of Transportation Chain for Dangerous Goods: A Case Study in Colombia. *Transportation Research Procedia*, *12*(June 2015), 842–850. https://doi.org/10.1016/j.trpro.2016.02.037

Gallego, R. S. (2003). *Introducción al análisis de datos experimentales: tratamiento de datos en bioensayos*. Universitat Jaume I. Servei de Comunicació i Publicacions. https://books.google.com.co/books?id=NLUVJTK7EIoC

García Estrella, C. W., Barón Ramírez, E., & Sánchez Gárate, S. K. (2021). La inteligencia de negocios y la analítica de datos en los procesos empresariales. *Revista Científica de Sistemas e Informática*, *1*(2), 38–53. https://doi.org/10.51252/rcsi.v1i2.167

Ilijason, R. (2020). *Getting Started with Databricks BT - Beginning Apache Spark Using Azure*

Databricks: Unleashing Large Cluster Analytics in the Cloud (R. Ilijason (ed.); pp. 27–38). Apress. https://doi.org/10.1007/978-1-4842-5781-4_3

Kleijnen Jack P C. (1995). Verification and validation of simulation models. *European Journal of Operational Research*, *82*(mimic), 145–162. https://doi.org/10.1109/WSC.1998.744907

Malakouti, S. M. (2023). Utilizing time series data from 1961 to 2019 recorded around the world and machine learning to create a Global Temperature Change Prediction Model. *Case Studies in Chemical and Environmental Engineering*, *7*(February), 100312. https://doi.org/10.1016/j.cscee.2023.100312

Martínez-Jaramillo, J. E., Arango-Aramburo, S., Álvarez-Uribe, K. C., & Jaramillo-Álvarez, P. (2017). Assessing the impacts of transport policies through energy system simulation: The case of the Medellin Metropolitan Area, Colombia. *Energy Policy*, *101*(June 2016), 101–108. https://doi.org/10.1016/j.enpol.2016.11.026

Merino, A. P., D\'\iaz, M. Á. R., & Castellanos, R. S. M. (2009). *Análisis de datos I : en ciencias sociales y de la salud*. Editorial S{\'\i}ntesis, S.A. https://books.google.com.co/books?id=kea3cQAACAAJ

Molina, F., Pérez, S., & Rivera, S. (2017). *Formulación de Funciones de Costo de Incertidumbre en Pequeñas Centrales Hidroeléctricas dentro de una Microgrid*. *8*(March), 29–36. https://doi.org/10.21500/20275846.2683

Montoya-Torres, J. R., Moreno, S., Guerrero, W. J., & Mejía, G. (2021). Big Data Analytics and Intelligent Transportation Systems. *IFAC-PapersOnLine*, *54*(2), 216–220. https://doi.org/10.1016/j.ifacol.2021.06.025

Pasquinelli, M. (2019). How a Machine Learns and Fails: A Grammar of Error for Artificial Intelligence. *Spheres. Journal for Digital Cultures*, *5*(Spectres of AI), 1–17.

Phasinam, K., Singh, A. K., Singh, T., & Sharma, M. K. (2022). *Fundamental of Machine Learning*. Blue Rose Publishers. https://books.google.com.co/books?id=-XCJEAAAQBAJ

Quintero, J. B., Villanueva, D. M., Luis, F., & Montaya, G. (2018). Analítica de datos para sistemas de costos basados en actividades en la era de big data. *Revista Del Instituto Internacional de Costos, ISSN-e 2718-8507, Nº. Extra 1, 2018 (Ejemplar Dedicado a: Edición Especial XV Congreso de Costos), Págs. 64-82*, *1*, 64–82. https://dialnet.unirioja.es/servlet/articulo?codigo=7457929&info=resumen&idioma=SPA%0Ahttps://dialnet.unirioja.es/servlet/articulo?codigo=7457929

Rojo, J. D., Carvajal, L. F., & Velásquez, J. D. (2015). Streamflow prediction using a forecast combining system. *IEEE Latin America Transactions*, *13*(4), 1035–1040. https://doi.org/10.1109/TLA.2015.7106354

Sarri, P., Kaparias, I., Preston, J., & Simmonds, D. (2023). Using Land Use and Transportation Interaction (LUTI) models to determine land use effects from new vehicle transportation technologies; a regional scale of analysis. *Transport Policy*, *135*(March 2022), 91–111. https://doi.org/10.1016/j.tranpol.2023.03.012

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, *181*(2019), 526–534.

https://doi.org/10.1016/j.procs.2021.01.199

Shafique, U., & Qaiser, H. (2014). A Comparative Study of Data Mining Process Models (KDD , CRISP-DM and SEMMA). *International Journal of Innovation and Scientific Research*, *12*(1), 217–222. http://www.ijisr.issr-journals.org/

Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., & Lichtendahl, K. C. (2017). *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R* (1st ed.). Wiley Publishing.

Treviño, R., Rivera, F., & Garza, J. (2020). La analítica de datos como ventaja competitiva en las organizaciones. *VinculaTégica*, *6*(2), 1063–1074. http://www.web.facpya.uanl.mx/vinculategica/Vinculategica6_2/5_Treviño_Rivera_Garza.pdf

Uddin, M. G., Nash, S., Mahammad Diganta, M. T., Rahman, A., & Olbert, A. I. (2022). Robust machine learning algorithms for predicting coastal water quality index. *Journal of Environmental Management*, *321*(June), 115923. https://doi.org/10.1016/j.jenvman.2022.115923

Vidya, V. M., & Deepa, N. (2019). Big data analytics in intelligent transportation systems using hadoop. *International Journal of Recent Technology and Engineering*, *7*(6), 75–80.

Zadrozny, B. (2004). Learning and evaluating classifiers under sample selection bias. *Proceedings, Twenty-First International Conference on Machine Learning, ICML 2004*, 903–910. https://doi.org/10.1145/1015330.1015425