Predicción temprana de averías en flotas de autobuses mediante machine learning para optimizar la gestión y rentabilidad a largo plazo.

Laureano Romero Velásquez

Trabajo de grado para optar al título de
Máster en Ciencia de Datos

Director:

Jose Armando
Ordoñez Cordoba

Tutor:

Jonnathan
Vargas

FACULTAD DE INGENIERÍA Y CIENCIAS APLICADAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI
2024

Resumen

El objetivo de este trabajo de grado es desarrollar modelos predictivos fundacionales para identificar posibles fallas en los buses del Consorcio Express, el principal operador privado de transporte en Colombia. La investigación se basa en una metodología rigurosa que abarca las siguientes etapas:

A) **Definición precisa del problema:** Definir claramente el problema específico que se busca resolver mediante técnicas de aprendizaje automático (machine learning).
B) **Establecimiento de objetivos y métricas de éxito:** Determinar los objetivos medibles y los criterios de éxito que guiarán el desarrollo y evaluación del modelo.
C) **Identificación de variables relevantes:** Formular hipótesis sobre las variables claves que podrían influir en el problema y alimentar el modelo.
D) **Evaluación de la disponibilidad de datos:** Analizar y cuantificar la accesibilidad y disponibilidad de los datos necesarios para entrenar y validar el modelo.
E) **Selección de métricas de evaluación:** Definir las métricas de evaluación apropiadas para medir el rendimiento y la precisión del modelo en relación con los objetivos establecidos.
F) **Colaboración interdepartamental:** Establecer reuniones periódicas con diferentes áreas de la empresa para identificar y acceder a fuentes de datos relevantes.
G) **Recopilación de datos:** Consumir la información identificada en la etapa anterior, garantizando la calidad y relevancia de los datos recopilados.
H) **Procesamiento y transformación de datos (ETL):** Utilizar procesos ETL (Extracción, Transformación y Carga) para limpiar, transformar, enriquecer y cargar los datos en un formato adecuado para el modelado.
I) **Creación de la tabla de modelado:** Construir una tabla estructurada y optimizada que contenga los datos preparados para el entrenamiento del modelo de machine learning.
J) **Desarrollo del modelo de predicción:** Aplicar técnicas de machine learning para construir un modelo capaz de predecir las fallas en los buses.
K) **Evaluación del rendimiento del modelo:** Evaluar el desempeño del modelo utilizando métricas como el área bajo curva, la Accuracy, la puntuación F1 y la matriz de confusión, para determinar su precisión y capacidad de generalización.
L) **Implementación del modelo:** Después de seleccionar el modelo óptimo, con base en su rendimiento del F1 Score este se almacena en la nube de Azure. Los parámetros óptimos del modelo son aplicados a los nuevos datos, a los cuales se les realizan algunos cálculos adicionales. Posteriormente, los resultados se guardan en archivos en formato .parquet, también alojados en la nube, para facilitar su consulta en Power BI. Esto permite el acceso desde dispositivos fijos y móviles a través del aplicativo alojado en **https://app.powerbi.com/**.

Los algoritmos, formas y lenguajes de programación de la ciencia de datos para afrontar la problemática y lograr plantear una posible solución son:

- Se utilizó esencialmente el lenguaje de programación llamado Python, como herramienta fundamental para todo el proceso de ETL.

- Se usaron principalmente las siguientes librerías:
    a. **Manipulación y Análisis de Datos**
    Pandas: Estructuras de datos y análisis eficiente.
    Numpy: Operaciones matemáticas y manejo de arreglos.
    b. Visualización de Datos
    Matplotlib & Seaborn: Gráficas personalizadas y análisis visual avanzado.
    c. Machine Learning
    Scikit-learn: Modelos como Random Forest, Logistic Regression, SVM, redes neuronales, imputación de valores (KNN).
    Skopt: Optimización Bayesiana para hiper parámetros y visualización de resultados.
    d. Bases de Datos
    SQLAlchemy: Conexión y manejo de bases de datos SQL.
    Pandasql: Consultas SQL sobre DataFrames.
    Pymongo: Conexión y operaciones con MongoDB.
    e. Almacenamiento en la Nube
    Azure Blob Storage: Acceso y procesamiento de datos desde Azure.
    Estructuras de datos y análisis eficiente.
    f. Utilidades
    Pickle & Joblib: Serialización de modelos y objetos.
    PIL: Manipulación de imágenes.
    Datetime: Manejo de fechas y tiempos.
    OS & Subprocess: Operaciones con el sistema operativo.

- Para el desarrollo de este trabajo de grado se usaron solamente modelos tradicionales tales como árboles de decisión, regresión lineal, redes neuronales poco profundas, SVM y árboles aleatorios.

El modelo escogido fue Decision Tree obtuvo un rendimiento sobresaliente en el conjunto de prueba, alcanzando un F1 Score de 0.856 con un punto de corte de 0.880.

FACULTAD DE INGENIERÍA
Maestría en Ciencia de Datos
Universidad ICESI

Tabla de contenido

1. Integrantes y directores del trabajo de grado.

Integrante:
Laureano Romero Velásquez

Director: Jose Armando Ordoñez Cordoba
Tutor: Jonnathan Vargas

2. Título del trabajo de grado.

Predicción temprana de averías en flotas de autobuses mediante machine learning para optimizar la gestión y rentabilidad a largo plazo.

3. Contexto y antecedentes.

El presente trabajo de grado se enmarca en el sistema de transporte masivo de pasajeros BRT (Bus Rapid Transit) en la ciudad de Bogotá, específicamente en el contexto de la operación de Transmilenio, administrado por la Empresa de Transporte del Tercer Milenio S.A. (TRANSMILENIO S.A.).

El sistema de transporte público de Bogotá, previo a la implementación de TransMilenio, se caracterizaba por su desorganización y fragmentación. Operado por empresas privadas, presentaba una heterogeneidad de vehículos en términos de tamaño y modelo, sin una regulación rigurosa en cuanto a tarifas, horarios y rutas. Además de los autobuses tradicionales, existían los llamados "colectivos", vehículos de menor capacidad que ofrecían rutas específicas a un precio más elevado.

Este sistema de transporte era un foco de problemas urbanos. Su falta de integración y planificación agravaba la contaminación y la congestión, deteriorando la calidad de vida. La inseguridad y la ineficiencia lo hacían poco atractivo para los usuarios, contribuyendo a un declive en los indicadores de calidad de vida urbana.

Actualmente[1] el SITP de la ciudad cuenta con 9.856 buses, de los cuales 910 corresponden a alimentadores, 2.089 a troncales y 6.857 a zonales.

Este sistema, fundamental para la movilidad de la capital colombiana, se caracteriza por su estructura de concesiones, donde diversos operadores se encargan de la gestión y operación de los buses en diferentes zonas de la ciudad.

El estudio se centrará en Consorcio Express S.A.S., el mayor operador de transporte BRT en Colombia, que desempeña un papel crucial en la Fase III de Transmilenio. Con una flota que supera los 2300 buses, con una participación del transporte de Bogotá en troncal del 25.3%, alimentadores del 12.7%, y zonal del 21.3%, Consorcio Express aporta una parte significativa de los vehículos que conforman el sistema de transporte público de Bogotá. La participación de esta empresa en el sistema se formaliza mediante el Contrato No. 008 del 2010, una concesión otorgada por TRANSMILENIO S.A. para la explotación preferencial y no exclusiva del servicio público de transporte de pasajeros en la Zona 4, que abarca las localidades de San Cristóbal y Usaquén, incluyendo tanto la operación troncal como alimentadora.

[1] La información de corte corresponde a enero 2024.

Consorcio Express, es una empresa líder en el sector del transporte público, que gestiona una extensa flota operativa de más de 2.300 autobuses. Esta flota se distribuye estratégicamente entre dos concesiones principales: Usaquén (Norte), que concentra el 66% de los vehículos, y San Cristóbal (Sur), con el 34% restante.

Para garantizar una operación eficiente y cubrir la amplia demanda de transporte en la ciudad, Consorcio Express cuenta con una infraestructura robusta compuesta por 14 centros de operación y 10 patios, desde los cuales se coordinan y ejecutan las actividades diarias de la flota.

La tipología de servicio ofrecida por Consorcio Express abarca cuatro modalidades principales, cada una con un enfoque específico para satisfacer las diversas necesidades de movilidad de los usuarios. El 13% de la flota se dedica al servicio troncal, que conecta los principales corredores de la ciudad. El servicio zonal, que cubre una amplia red de rutas en diferentes zonas de Bogotá, representa el 72% de la flota. El 9% de los autobuses operan en modalidad dual, combinando características de los servicios troncal y zonal, mientras que el 6% restante se destina al servicio de alimentación, que facilita el acceso de los usuarios a las estaciones y portales del sistema de transporte.

La distribución geográfica de los autobuses de Consorcio Express está diseñada para optimizar la cobertura y accesibilidad del servicio en toda la ciudad.

4. Planteamiento del problema y justificación

Como se ha destacado en la sección de antecedentes, para Consorcio Express resulta de vital importancia anticiparse y predecir cualquier tipo de eventos que afecte el buen funcionamiento de la flota de automóviles en este contexto nace la necesidad de conocer o estimar la probabilidad de sufrir una falla mecánica que pueda llevar a su inmovilización. Este tipo de eventos impacta directamente en la Evaluación Integral de Calidad del componente zonal-alimentación, conocida como EMIC, un indicador clave para la empresa, ya que determina el pago que recibe de TRANSMILENIO S.A. por los servicios prestados.

El EMIC es un indicador multidimensional que engloba seis elementos críticos para la calidad del servicio: seguridad vial, cumplimiento de los servicios programados, regularidad en la frecuencia de estos, distancia promedio recorrida por los vehículos, conductas operacionales del personal y, finalmente, la percepción de los usuarios a través de encuestas de satisfacción. Cada uno de estos elementos tiene un peso específico en el cálculo del EMIC, y su combinación determina el valor final del indicador.

A continuación, se presenta un análisis detallado de cada uno de los componentes del EMIC, incluyendo su forma de cálculo, los valores de referencia establecidos por TRANSMILENIO S.A. y su relevancia en la evaluación global de la calidad del servicio prestado por Consorcio Express. Este análisis permitirá comprender la importancia de anticipar y prevenir las fallas mecánicas, ya que su impacto negativo en el EMIC puede tener consecuencias financieras significativas para la empresa.

Formula No. 1 Cálculo EMIC, Mensual.

$$\boldsymbol{EMIC_M} = \underbrace{(15\% \times ISV_M)}_{\text{Seguridad Vial}} + \underbrace{(35\% \times ICS_M)}_{\text{Cumplimiento de Servicios}} + \underbrace{(15\% \times IRI_M)}_{\text{Regularidad del Servicio}} + \underbrace{(15\% \times DPV_M)}_{\text{Distancia Promedio entre Varados}} + \underbrace{(20\% \times ICO_M)}_{\text{Conductas Operacionales}} + \underbrace{(PSU_M)}_{\text{Encuesta de Satisfacción del Usuario}}$$

Fuente: Consorcio Express.

La naturaleza intrínseca de las empresas de transporte, especialmente Consorcio Express, implica una serie de desafíos que requieren soluciones metodológicas diversas. Uno de estos desafíos, y el que se abordará en este estudio, son las fallas mecánicas inherentes a la operación de la flota de autobuses. Estas fallas, aunque inevitables, presentan una naturaleza dual: por un lado, poseen un componente estocástico, sujeto a la aleatoriedad de eventos imprevisibles; por otro lado, también exhiben un componente determinístico, vinculado a factores como el desgaste natural de los componentes y las condiciones de operación.

La flota de autobuses de Consorcio Express, compuesta por una cantidad considerable de vehículos, se enfrenta a una exigencia operativa significativa. Los autobuses recorren diariamente una gran cantidad de kilómetros, sometidos a un uso intensivo que acelera el desgaste de sus piezas y sistemas. Además, factores externos e incontrolables, como las condiciones climáticas adversas y el estado de las vías, así como la intensidad del tráfico, pueden que contribuyen a aumentar una posible probabilidad de que ocurran fallas

mecánicas.

Para contrarrestar este problema, la empresa implementa un programa de mantenimiento preventivo basado en el kilometraje recorrido, recomendaciones de los fabricantes y la antigüedad de cada autobús. Sin embargo, esta medida, aunque fundamental, no puede garantizar la eliminación total de fallas inesperadas. La complejidad de los sistemas mecánicos y eléctricos de los autobuses, junto con la variabilidad de las condiciones de la operación, hacen que siempre exista un riesgo latente de que surjan problemas que afecten el funcionamiento general de los vehículos.

El impacto de las fallas mecánicas trasciende lo meramente técnico, generando consecuencias económicas significativas para la empresa. Un autobús fuera de servicio representa una pérdida directa de ingresos, ya que no puede cumplir con su función de transportar pasajeros. Además, la reducción en la frecuencia de las rutas puede ocasionar una insatisfacción entre los usuarios, lo que a su vez afecta la percepción de la calidad del servicio y puede llevar a una disminución en la demanda.

Pero el impacto más crítico de las fallas mecánicas es, sin duda, el que se refiere a la seguridad. Un vehículo que presenta problemas técnicos puede poner en riesgo la vida de los pasajeros y de otros usuarios de la vía. Un accidente de tránsito causado por una falla mecánica no solo tendría consecuencias trágicas para las personas involucradas, sino que también dañaría gravemente la reputación de la empresa y su relación con la comunidad.

El alcance de este proyecto, en su fase inicial durante el primer y segundo semestre del año 2024, se centrará en el desarrollo de un modelo de predicción anual de averías en la flota de autobuses. Para ello, se emplearán modelos tradicionales de machine learning, sometidos a un proceso iterativo de ajuste y recalibración de hiper parámetros con el objetivo de alcanzar un modelo óptimo. Si bien este modelo inicial puede no ser el más preciso o sofisticado.

En la segunda fase, prevista para el siguiente semestre, se buscará refinar y mejorar el modelo de predicción e implementarlo en la nube de AZURE, para luego ser consultado usando POWERBI. Para ello, se enriquecerá el modelo con una mayor cantidad de variables dependientes, ampliando así su capacidad de análisis y comprensión de los factores que influyen en las averías. Además, se ajustarán los parámetros y se realizarán las modificaciones necesarias para que el modelo pueda generar pronósticos con una granularidad temporal más precisa, específicamente cada 15 días. Esto permitirá a la empresa anticiparse de manera más efectiva a las posibles fallas y tomar medidas preventivas con mayor antelación, optimizando aún más la gestión de la flota y minimizando los costos asociados a las averías.

Ahora bien, dentro del contexto de la justificación del siguiente trabajo se plantea las siguientes preguntas:

¿Cuáles son los principales factores, patrones o señales que pueden contribuir a que un autobús quede fuera de servicio?

¿Qué tipo de datos históricos se necesitan para entrenar el modelo de machine learning para la operación de la flota de autobuses?

5. Objetivos del proyecto

5.1. Objetivo general

Implementar un sistema de predicción de averías en la flota de autobuses de consorcio express mediante el uso de machine learning, con el fin de anticipar y mitigar los costos asociados a los vehículos fuera de servicio, optimizando así la gestión de la flota y mejorando su rentabilidad en el largo plazo.

5.2. Objetivos específicos

- **5.2.1.** Diseñar y desarrollar un modelo de machine learning ajustable que anticipe las averías en la flota de autobuses, utilizando un conjunto de datos históricos de mantenimiento, operación y otras variables relevantes que puedan influir en la probabilidad de fallo.
- **5.2.2.** Llevar a cabo un proceso de selección y análisis de variables, identificando aquellas que tengan un mayor poder predictivo y contribuyan significativamente al rendimiento del modelo de machine learning.
- **5.2.3.** Realizar un análisis exploratorio de datos (EDA) para obtener una comprensión de los patrones, tendencias y relaciones presentes en los datos. Este análisis permitirá refinar la hipótesis inicial, ajustar el planteamiento del problema y garantizar que el modelo se adapte a las características específicas de los datos.
- **5.2.4.** Implementar metodologías de limpieza y depuración de datos, abordando problemas como valores faltantes, outliers y errores de registro. Este proceso garantizará la calidad y consistencia de los datos, optimizando así el rendimiento y la precisión del modelo de machine learning.
- **5.2.5.** Examinar el mejor modelo y realizar los cambios respectivos en los hiperparametros de tal manera que se obtenga un modelo óptimo.

6. Marco teórico

6.1. Dominio del problema

Las fallas mecánicas en una flota de autobuses, como se detalla a lo largo de este estudio, se desencadenan una serie de repercusiones que se extienden desde lo operativo hasta lo financiero, afectando la eficiencia, la calidad del servicio y en última instancia, la rentabilidad de la empresa. Dada la magnitud de estas consecuencias, resulta imperativo abordar esta problemática de manera proactiva y con herramientas que permitan anticiparse a los eventos. En este contexto, el aprendizaje automático (machine learning) emerge como una herramienta poderosa para predecir y prevenir fallas mecánicas, para esto se basó en eventos que resultaron a partir de la inmovilización de los vehículos, los cuales impactan negativamente el indicador EMIC. En la etapa inicial de selección de datos, se identificaron 16 posibles fuentes de información correspondientes al año 2023, incluyendo variables como la frecuencia de mantenimientos, el kilometraje total recorrido, la cantidad de rutas operadas, las órdenes de trabajo asociadas al

mantenimiento preventivo, las características técnicas de los vehículos, las variaciones en los intervalos de mantenimiento entre lotes de buses, el número de pasajeros transportados por vehículo, la cantidad de paradas realizadas por ruta y tipo de bus, la edad promedio de los conductores, el número de conductores asignados a cada vehículo, la cantidad de rutas realizadas por cada autobús, la cuantificación de la dificultad de las rutas en términos de inclinación o tipo de terreno, el consumo de combustible por kilómetro y los patrones de conducción (velocidad, aceleración, frenado). Además, se consideraron los datos telemétricos disponibles en algunos vehículos.

Sin embargo, no todas estas fuentes de datos pudieron ser utilizadas debido a diversas limitaciones. Algunas presentaban restricciones de acceso impuestas por la compañía, mientras que otras carecían de la completitud necesaria para su inclusión en el análisis. Adicionalmente, ciertas variables no estaban calculadas previamente, lo que dificulta su obtención y procesamiento. A pesar de estos desafíos, se logró recolectar información relevante de las siguientes fuentes:

6.1.1. Variables Seleccionadas

Características propias de los vehículos: Esta tabla de datos contiene información detallada sobre cada autobús, incluyendo su modelo, marca, placa, antigüedad y otras especificaciones técnicas.
Historial de mantenimiento preventivo: Se obtuvo un registro completo de los mantenimientos preventivos realizados en cada vehículo, junto con las fechas correspondientes.
Cantidad de rutas operadas: Se recopiló información sobre el número de rutas asignadas a cada autobús durante el período de estudio.
Intervalos de mantenimiento: Se calcularon los intervalos en días entre los mantenimientos preventivos de cada vehículo.
Cantidad de rutas por vehículo: Se registró el número de rutas completadas por cada autobús.
Tiempo registro conductor: Se anexó información de la cantidad de horas transcurridas, desde que el conductor empieza a manejar el vehículo hasta el momento que deja de conducir. También se registró la cantidad de conductores diferentes que conducen cada uno de los buses.
Esta información, aunque limitada en comparación con la lista inicial de fuentes potenciales, proporcionó una base sólida para el desarrollo del modelo de predicción de averías.

Un antecedente crucial para considerar en este análisis es la composición actual de la flota de Consorcio Express, la cual presenta una proporción significativa de vehículos antiguos que han superado su vida útil esperada. Esta situación, si bien ha sido temporalmente autorizada por las autoridades de transporte hasta el año 2025, plantea desafíos importantes en términos de eficiencia, costos de mantenimiento y, sobre todo, impacto ambiental.

La antigüedad de los vehículos no solo implica un posible mayor riesgo de fallas mecánicas y un aumento en los costos de mantenimiento, sino que también conlleva un mayor consumo de combustible y emisiones de gases contaminantes. Esto último es especialmente relevante en el contexto actual de creciente preocupación por el cambio climático y la necesidad de adoptar tecnologías más limpias y sostenibles en el sector del transporte.

La prórroga otorgada por las autoridades hasta 2025 representa una oportunidad para que Consorcio Express planifique y ejecute una transición gradual hacia una flota más moderna y ecológica. La incorporación de buses eléctricos o alimentados por gas natural no solo reduciría el impacto ambiental de la empresa, sino que también podría generar ahorros a largo plazo en costos de combustible y mantenimiento, mejorando así la rentabilidad y la competitividad de la empresa en el mercado.

Sin embargo, esta transición también plantea retos importantes, como la inversión inicial en nuevos vehículos, la adaptación de la infraestructura de carga y mantenimiento, y la capacitación del personal técnico y de conducción. Es fundamental que Consorcio Express aborde estos desafíos de manera estratégica y proactiva, aprovechando al máximo el tiempo disponible para garantizar una transición exitosa hacia una flota más moderna, eficiente y sostenible.

La ocurrencia de una falla mecánica en un vehículo de la flota de Consorcio Express desencadena una serie de consecuencias que se propagan en cascada, afectando diversos aspectos de la operación y, en última instancia, la rentabilidad de la empresa. Esta reacción en cadena se inicia con la inmovilización del vehículo averiado, lo que obliga a realizar ajustes en la planificación operativa para cubrir la ruta afectada. Estos cambios implican la reasignación de recursos, como conductores y otros autobuses, lo que puede generar retrasos en la programación y una disminución en la eficiencia general del servicio.

La alteración de la programación no solo afecta a la empresa, sino también a los usuarios del sistema de transporte. Los pasajeros pueden experimentar retrasos en sus desplazamientos, lo que genera una percepción negativa de la calidad del servicio y puede disuadirlos de utilizar el transporte público en el futuro. Esta insatisfacción del cliente puede traducirse en una disminución de la demanda y, por ende, en una reducción de los ingresos de la empresa.

Además de los costos directos asociados a la reparación del vehículo averiado, como repuestos y mano de obra, la empresa también debe asumir los gastos de grúa y el tiempo perdido en la gestión de la incidencia. Estos costos adicionales, sumados a la pérdida de ingresos por la disminución de la demanda, agravan el impacto económico de la falla mecánica.

La relación causa-efecto entre las fallas mecánicas y sus múltiples consecuencias es especialmente perjudicial para Consorcio Express debido a su impacto directo en el EMIC. Este indicador, utilizado por TRANSMILENIO para remunerar a la empresa, se ve afectado negativamente por las fallas mecánicas, ya que estas aumentan la frecuencia de mantenimientos y reparaciones, lo que se traduce en un menor rendimiento económico para Consorcio Express.

7. Estado del Arte

Si bien la ciencia de datos y el aprendizaje automático (machine learning) han ganado un protagonismo indiscutible en diversos sectores empresariales y académicos, su adopción e implementación en el ámbito del transporte público aún se encuentra en una etapa temprana de desarrollo. A pesar de que estas tecnologías ofrecen un enorme potencial

para optimizar la gestión de flotas, predecir fallas mecánicas y mejorar la eficiencia operativa, su aplicación en el sector del transporte ha sido relativamente limitada hasta el momento.

Consorcio Express, como empresa líder en el transporte público, no ha sido ajena a esta realidad. Hasta la fecha, no se han llevado a cabo proyectos internos que exploren el potencial del machine learning para abordar los desafíos específicos de la operación de una flota de autobuses. Esta falta de experiencia previa ha hecho necesario recurrir a fuentes de información externas para sentar las bases de este proyecto.

7.2. Trabajos seleccionados

La literatura académica y profesional sobre la aplicación del machine learning en el sector del transporte, aunque aún en desarrollo, ofrece un punto de partida valioso para comprender las posibilidades y limitaciones de estas tecnologías. A continuación, se presenta un resumen de los principales hallazgos y enfoques encontrados en la literatura consultada, que servirán como marco de referencia para el desarrollo de este proyecto, se apoya en diferentes documentos los cuales se encuentran anexados en la bibliografía, entre los cuales cabe destacar:

El estudio “Aprendizaje automático para el mantenimiento predictivo en los buses del transporte público” realizado por (Tipantuña, Arroba, 2019) en donde proponen un modelo predictivo de fallas mecánicas para las unidades de transporte de una empresa, aprovechando el poder del aprendizaje automático (machine learning) y los datos históricos disponibles. Este modelo buscó anticipar de manera proactiva las posibles fallas en los vehículos, utilizando algoritmos y técnicas de análisis de datos para identificar patrones y tendencias ocultas en los registros de mantenimiento y operación de la flota.

También se utilizó como antecedente el trabajo “diseño del plan de mantenimiento preventivo de los vehículos en operación para la empresa servitransguamal s.a.s” realizado por (Linares, 2023) en donde se abordaron temas en cuanto a la optimización de los recursos disponibles de las organizaciones, alineándose con los objetivos del Plan Estratégico de Seguridad Vial (PESV). A través de un enfoque proactivo y sistemático, se buscó identificar y abordar de manera anticipada las posibles fallas mecánicas en la flota de vehículos, contribuyendo así de manera significativa a la reducción de accidentes de tránsito y garantizando la seguridad tanto de los conductores como de los pasajeros.

El trabajo de grado “Análisis de datos para la optimización de la gestión de flotas vehiculares: Impacto en los costos operativos y rendimiento empresarial” toma como antecedente la investigación de (Ortiz, 2023) en donde se enfoca en desarrollar un análisis exhaustivo de datos históricos y actuales relacionados con los costos y gastos de la flota vehicular de una empresa de transporte ubicada en Medellín, donde realiza un modelo predictivo asociado a identificar los costos y gastos futuros asociados a la operación de dicha flota.

Por último, también se toma como referencia el trabajo llamado “Optimización del tráfico en el transporte público” (Pérez, 2019) en donde se aborda como objetivo primordial analizar en profundidad la ocupación de los autobuses a lo largo de sus recorridos. A través de un minucioso examen de los datos recopilados, se busca no solo cuantificar la ocupación en cada punto del trayecto, sino también evaluar la fiabilidad y eficiencia del

sistema de transporte en su conjunto.

A diferencia de las propuestas anteriores, esta implementación se destaca por el uso de un conjunto más amplio de algoritmos de entrenamiento, la optimización bayesiana y la puesta en producción del modelo mediante la generación de archivos .pkl. Además, se incorpora una herramienta de seguimiento del modelo basado en Power BI.

Tabla No. 1 Resumen de los criterios de comparación entre los artículos seleccionados y el proyecto de grado.

	Proyecto de grado propio	Aprendizaje automático para el mantenimiento predictivo en los buses del transporte público	diseño del plan de mantenimiento preventivo de los vehículos en operación para la empresa servitransguamal s.a.s	Análisis de datos para la optimización de la gestión de flotas vehiculares: Impacto en los costos operativos y rendimiento empresarial	Optimización del tráfico de usuarios en el transporte público
Fecha de publicación	2024	2019	2023	2023	2023
Autor	(Propio)	(Tipantuña, Arroba)	(Linares)	(Ortiz)	(Pérez)
país	Colombia	Ecuador	Colombia	Colombia	España
Tecnica de ML utilizada	RandomForest	RandomForest	Ninguno	Gradient Boosted Trees	clústeres K-Means
Aplica al sistema de transporte de BRT	Si	Si	No	No	Si
Aplica a pronosticas	Si	Si	Si	No	No

Fuente: Elaboración propia.

8. Marco metodológico

En este proyecto, se ha optado por una metodología híbrida que combina los principios ágiles de Scrum con las etapas estructuradas de CRISP-DM (Cross Industry Standard Process for Data Mining), buscando aprovechar las fortalezas de ambos enfoques para lograr un desarrollo eficiente y efectivo del modelo de predicción de averías.

De la metodología Scrum, ampliamente utilizada en el desarrollo de software, se adoptan elementos como la iteración rápida, la planificación en sprints y las reuniones diarias (Daily). Estas prácticas fomentan la adaptabilidad, la comunicación constante y la mejora continua, aspectos cruciales en un proyecto de ciencia de datos donde los requisitos y las condiciones pueden cambiar rápidamente.

Por otro lado, la metodología CRISP-DM proporciona un marco sólido y probado para guiar el proceso de minería de datos, desde la comprensión del negocio y los datos hasta el despliegue del modelo. Sus etapas bien definidas (conocimiento del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue) garantizan un enfoque sistemático y riguroso en cada fase del proyecto.

La metodología propuesta, aunque flexible en cuanto a los plazos específicos, sigue un orden lógico y estructurado de tareas. Se divide en cinco fases principales:

Entendimiento del negocio y planteamiento de objetivos
En esta etapa inicial, se llevan a cabo reuniones exhaustivas con las partes interesadas de la empresa para comprender a fondo los datos disponibles, el funcionamiento de la flota y los desafíos específicos que se enfrentan. Esta comprensión profunda del contexto empresarial permite definir objetivos claros, medibles y alineados con las necesidades de la organización.

Identificación y evaluación de fuentes de datos
En esta fase, se revisan y evalúan las fuentes de datos disponibles, confirmando la relevancia de algunas y descartando otras que no cumplen con los criterios de accesibilidad, completitud o calidad. Este proceso es fundamental para asegurar la solidez de los datos que alimentarán el modelo de predicción.

Recopilación, preparación y modelado de datos
Esta etapa, dividida en sprints ágiles, abarca la recopilación final de los datos seleccionados, su limpieza y validación, la creación de nuevas variables relevantes, la construcción de la tabla de datos (minable) y el entrenamiento inicial del modelo de machine learning. La iteración y la retroalimentación constante en esta fase permiten ajustar el modelo y mejorar su rendimiento.

Implementación del modelo en Power BI:
Una vez que el modelo ha sido entrenado y validado, se procede a su implementación en un reporte interactivo de Power BI, accesible desde la nube. Esto facilita la visualización y análisis de los resultados, así como la toma de decisiones basadas en datos. Esto se realizará a nivel empresarial y si estará documentada en este trabajo de grado.

Documentación del proceso:
La fase final consiste en documentar detalladamente todo el proceso, desde la definición del problema hasta la implementación del modelo. Esta documentación servirá como guía para futuras mejoras y replicaciones del proyecto, así como para compartir el conocimiento adquirido con la comunidad científica y profesional.

A continuación, se presenta un resumen gráfico del marco metodológico:

Gráfica No. 1 Resumen gráfico del marco teórico.

Fuente: Elaboración propia.

9. Metodología

Con el propósito de cumplir con los objetivos planteados en el trabajo de investigación, se propusieron los siguientes pasos:

9.1. Planteamiento y entendimiento del problema.
- **9.1.1.** Recopilación de datos relevantes relacionados con el problema.
- **9.1.2.** Identificación de las métricas de evaluación adecuadas para medir el rendimiento del modelo.

9.2. Recopilación y Preparación de Datos.
- **9.2.1.** Realización de reuniones preliminares para identificar fuentes de datos.
- **9.2.2.** Manejo de datos.
  - **9.2.2.1.** Importación de librerías.
  - **9.2.2.2.** Importación fuente de datos.
- **9.2.3.** Descripción de variables y ETL.
  - **9.2.3.1.** ETL Mantenimiento.
  - **9.2.3.2.** ETL Registro conductores.
  - **9.2.3.3.** ETL Kilómetros troncales.
  - **9.2.3.4.** ETL 01-01-2023, características de buses
  - **9.2.3.5.** ETL BASE DPV 2023.xlsx, tabla de fallas.
- **9.2.4.** Unificación fuentes de datos - ETL- Construcción tabla minable

9.3. Entendimiento de los datos.
- **9.3.1.** Análisis exploratorio.
  - **9.3.1.1.** Análisis gráfico de los datos
  - **9.3.1.2.** Ajuste y validación de tabla minable - imputación.
  - **9.3.1.3.** Manejo de valores atípicos.
  - **9.3.1.4.** Análisis de asociación cualitativo.
  - **9.3.1.5.** Dumización de variables cualitativas.
  - **9.3.1.6.** Estandarización de variables numéricas y verificación de matriz de correlación.
  - **9.3.1.7.** Selección de variables.

9.4. Modelado.
9.4.1. Modelos con variables numéricas.
9.4.1.1. Regresión Logística.
9.4.1.2. Árbol de decisión.
9.4.1.3. SVM.
9.4.1.4. Redes neuronales.
9.4.1.5. Árboles aleatorios.
9.4.2. Elección del algoritmo de machine learning más adecuado.
9.5. Implementación del modelo.
9.6. Conclusiones y documentación.

Este análisis permitirá no solo evidenciar la rigurosidad del trabajo, sino también facilitar la replicabilidad y la transferencia de conocimiento a otros contextos similares. Además, se espera que esta descripción detallada sirva como guía para futuros investigadores y profesionales interesados en aplicar técnicas de machine learning para resolver problemas complejos en el ámbito del transporte público.

A través de esta presentación sistemática y organizada, se busca demostrar cómo cada etapa del proceso contribuyó de manera significativa al logro de los objetivos planteados, así como identificar oportunidades de mejora y lecciones aprendidas que puedan ser aplicadas en futuros proyectos.

9.1. Planteamiento y entendimiento del problema

9.1.1. Recopilación de datos relevantes relacionados con el problema.

La investigación de información que alimente el modelo tal como se mencionó en el marco teórico se realizó una lluvia de ideas para tener un panorama de las posibles variables que pueden intervenir en la variable objetivo a continuación se anexa la tabla con el listado de variables numéricas que se van a manejar:

Tabla No. 2 Variables numéricas inicialmente consideradas para hacer parte del modelo.

Variables	Implementada
suma de mantenimientos realizados, durante año 2023	X
Suma total de los kilometros recorridos por el bus durante el año 2023	X
Maxima cantidad de rutas que realiza en la operación	X
Cantidad de ordenes de trabajo asociadas al mantenimiento preventivo	
Promedio de días de intervalo entre mantenimientos	X
suma de pasajeros que trasporto durante el 2023	
Suma de paraderos hechos durante el año 2023.	
Edad promedio de los operdores que conducen el bus, durante el año 2023, Cantidad de conductores por bus.	
Suma total de rutas hechas por el bus durante el año 2023	X
Totalidad suma de rutas realizadas que incluyen pendientes o caminos agrestes o de dificil manejo	
Cantidad de combustible consumido por bus x kilometro	
Sensores y telemetría: Datos recopilados en tiempo real o periódicamente durante la operación del vehículo, como temperatura del	
2.1.10 Comportamiento de conducción como velocidad, aceleración, frenado	
Cosumo de gasolina Medias de Consumo de aceites. Lubricantes	
Cantidad de Conductores por bus	X
Cantidad de horas laboradas por bus	X

Fuente: Elaboración propia.

La consolidación de la información para este proyecto se vio obstaculizada por una serie de desafíos inherentes a la estructura de datos de la empresa. La dispersión, la volumetría de las fuentes de datos, que se encontraban en diferentes ubicaciones y formatos, dificulta la unificación y centralización de la información relevante. Además, la compañía se encuentra en un proceso de transición hacia la implementación de una política de datos más robusta, lo que implica que la infraestructura de datos aún no está completamente consolidada y optimizada para el análisis.

Esta situación, sumada a la variabilidad en la calidad y disponibilidad de los datos, hizo que la selección de las variables para el modelo de predicción fuera un proceso cuidadoso y riguroso. Se tuvieron que evaluar múltiples fuentes de información, descartando aquellas que presentaban problemas de accesibilidad, inconsistencias o falta de datos completos. A pesar de estos obstáculos, se logró identificar un conjunto de variables clave, detalladas en la tabla número 2 y marcadas con una "X" en la tabla anterior, que cumplieron con los criterios de relevancia, calidad y disponibilidad necesarios para el desarrollo del modelo.

9.1.2. Identificación de las métricas de evaluación adecuadas para medir el rendimiento del modelo

La identificación de métricas de evaluación adecuadas es crucial para comprender qué tan bien está funcionando un modelo de machine learning en la tarea para la que fue diseñado. Estas métricas proporcionan una medida cuantitativa del rendimiento del modelo en relación con los datos de entrada y las predicciones que genera.

Precisión (Accuracy): Es la proporción de predicciones correctas realizadas por el modelo sobre el total de predicciones realizadas. Se calcula dividiendo el número de predicciones correctas entre el número total de predicciones. Es una métrica adecuada para problemas de clasificación binaria o multiclase, pero puede ser engañosa cuando hay clases desbalanceadas.

F1-Score: Es la media armónica entre precisión y recall. Proporciona un equilibrio entre ambas métricas y es especialmente útil cuando hay un desequilibrio entre las clases de predicción.

Matriz de Confusión: Es una tabla que muestra la cantidad de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos producidos por un clasificador. Es útil para comprender en detalle cómo está clasificando el modelo las diferentes clases y para calcular otras métricas como precisión y recall.

Especificidad: Este estadístico mide la proporción de verdaderos negativos entre todos los casos que realmente son negativos. En otras palabras, indica qué tan bien el modelo puede identificar correctamente los casos negativos.

Recall: Este indicador también llamado como sensibilidad o tasa de verdaderos positivos, mide la proporción de verdaderos positivos entre todos los casos que realmente son positivos. Indica qué tan bien el modelo puede identificar correctamente los casos positivos.

En este estudio más adelante se va a notar que la especificidad es baja, sin embargo, la sensibilidad es alta, identificando correctamente los vehículos que se varan, pero fallando en los carros que no presentan problemas mecánicos, esto puede pasar por un evidente desbalance de las clases de la variable respuesta, como también no tener un umbral de corte óptimo, o que las características del modelo no sea las más óptimas. Dado que no se tiene la suficiente cantidad de registros (buses) para replicar los datos y la dificultad de la recolección de la información se hace necesario ajustar el umbral de decisión el cual va a ser el punto donde la curva ROC esté más cercana a la coordenada (0,1).

9.2. Recopilación y preparación de datos.

9.2.1. Reuniones preliminares para identificar las fuentes de datos.

En aras de obtener una comprensión holística de los datos y procesos involucrados en la operación de la flota de autobuses, se llevó a cabo una serie de reuniones estratégicas, tanto virtuales como presenciales, con diferentes equipos clave dentro de la organización. Estas reuniones permitieron establecer un diálogo fluido y constructivo entre el investigador y los expertos en cada área, facilitando la identificación y evaluación de las

fuentes de datos disponibles, así como la clarificación de conceptos y terminología específica del negocio.

Se realizaron diferentes reuniones virtuales como presenciales con una variedad de equipos con el fin de aterrizar y comprender diferentes conceptos de los datos y el negocio a continuación se enumeraron las áreas con las cuales se realizó este tipo de reuniones.

Mantenimiento: Se interactuó con el equipo de mantenimiento para comprender los procesos de inspección, reparación y registro de fallas mecánicas, así como para identificar las fuentes de datos generadas en estas actividades.
Centro de información: Se colaboró con el centro de información para explorar las bases de datos existentes, evaluar la calidad y disponibilidad de los datos, y discutir posibles restricciones de acceso o limitaciones técnicas.
Planeación: Se trabajó juntamente con el equipo de planeación para entender la dinámica de la operación de la flota, la asignación de rutas, la programación de mantenimientos y otros factores relevantes para la predicción de averías.

9.2.2. Manejo de Datos

Las tablas que se manejan son las siguientes:

“BASE DPV 2023.xlsx”: Es la fuente de datos donde están identificado el evento de falla que provoca la inmovilización, esta tabla originalmente viene conformada 28 columnas y muestra el detalle por día del año 2023, vehículo, placa del motivo de la inmovilización.

“Km Ejecutado por Vehículo 2023_Zonal.xlsx”: La información contiene 3 columnas y muestra la suma de kilómetros hechos durante el año 2023 para los buses zonales, esta información está representada en como información diaria y por código interno el bus.

“**PMCEXP2023.xlsx”**: La tabla de 36 columnas hace referencia a los mantenimientos preventivos realizados por los vehículos, por día durante el año 2023.

“01-01-2023.csv”: El archivo .csv contiene 53 columnas la cual hace referencia a información propia de las características de cada uno de los buses de tal manera se puede identificar el detalle de cada bus tal, como modelo, marca, placa, chasis, numero de motor, como tal información sensible para la compañía, por este motivo se precede a no mostrar o en tal caso a enmascarar la información.

“Kilometros_Troncal.xlsx”: este archivo en Excel de 3 columnas muestra la cantidad de kilómetros realizados durante el año 2023 para los buses que recorren las troncales.

“Km_x_Bus_2023.xlsx”: Este documento de Excel contiene la información acumulada de zonal para los vehículos por kilómetros contiene información de 7 variables.

“Tiempo_Conductor_Log.parquet”: Esta tabla contiene la información del conductor desde que se registra para operar el bus hasta su finalización, contiene aproximadamente 14 variables.

9.2.2.1. Importación de librerías

Tal como se mencionó en la parte de antecedentes se implementará el programa Python versión 3.5 con la interfaz de Júpiter notebook 6.5.4, en la mayoría de las ejecuciones se utilizará alguno de los equipos locales los cuales ambos tienen Windows 11, Para Python se usó algunas librerías utilizadas en la ciencia de datos las cuales se asocian principalmente al manejo, transformación de datos, diagramación de gráficos, modelamiento de datos e implementación del modelo en la nube. Por lo tanto, se relacionan las siguientes librerías:

Azure-storage-blob: Es una biblioteca de programación que permite interactuar con el servicio de almacenamiento de blobs de Azure, permitiendo almacenar grandes cantidades de datos no estructurados de forma segura y escalable.
Bayesian-optimization: es una poderosa herramienta que se utiliza en este proyecto de grado para encontrar los mejores valores de los hiper parámetros de un modelo. En términos más simples.
Datetime: Se utilizó para el manejo de fechas y tiempos.
IPython: Este entorno interactivo de desarrollo proporciona una interfaz intuitiva y potente para la ejecución de código Python, la visualización de resultados y la experimentación con diferentes enfoques de análisis.
Matplotlib: Esta librería proporciona una amplia gama de herramientas para la creación de gráficos y visualizaciones de alta calidad. Su flexibilidad y capacidad de personalización permiten representar los datos de manera clara y efectiva, facilitando la identificación de patrones y tendencias relevantes para el análisis.
Numpy: Proporciona soporte para arreglos y matrices multidimensionales, junto con una colección de funciones matemáticas para operar con estos arreglos de manera eficiente.
Pandas: Esta librería se usa para la manipulación y análisis de datos en Python. Ofrece estructuras de datos flexibles y eficientes, como DataFrames y Series, que permiten realizar operaciones de filtrado, transformación, agregación y visualización de datos de manera intuitiva y concisa.
Pandasql: Esta librería permite la integración de consultas SQL en el flujo de trabajo de Pandas, facilitando la interacción con los datos y la aplicación de operaciones familiares a aquellos usuarios acostumbrados a trabajar con bases de datos relacionales.
Pickle & Joblib: El paquete sirvió para la Serialización de modelos y objetos.
PIL (Pillow): Esta librería ofrece un amplio conjunto de funciones para el procesamiento de imágenes, como la lectura, escritura, manipulación y visualización de imágenes en diferentes formatos.
Pymongo: La librería permitió la conexión y operaciones con MongoDB.
OS & Subprocess: Se implementa para mejorar las operaciones con el sistema operativo.
Subprocess: Esta librería permite la ejecución de comandos y procesos externos desde Python, lo que puede ser útil para automatizar tareas e integrar herramientas externas en el flujo de trabajo.
Scikit-learn: Esta librería es una de las más utilizadas en machine learning proporciona una amplia gama de algoritmos y herramientas para tareas de clasificación, regresión, clustering y reducción de dimensionalidad, entre otras.
Seaborn: Basada en Matplotlib, esta librería ofrece una interfaz de alto nivel para crear gráficos estadísticos atractivos e informativos.
Skopt: sirve para la implementación de la optimización Bayesiana para encontrar los mejores hiper parámetros y hacer una visualización de resultados.
SQLAlchemy: este paquete tiene la funcionalidad de realizar la conexión y manejo de bases de datos SQL.

9.2.2.2. Importación fuentes de datos

Para la adquisición de los datos contenidos en las tablas seleccionadas, se emplearon las funciones read_csv, read_parquet y read_excel de la biblioteca Pandas de Python. Estas funciones permiten cargar los archivos CSV, Parquet y Excel desde una ubicación local, asegurando la correcta interpretación de los caracteres acentuados propios del idioma español mediante la especificación del encoding "latin-1". En el caso de los archivos Excel con múltiples hojas, se indicó explícitamente la hoja a leer para garantizar la precisión de la extracción de datos.

Con miras a futuras mejoras en el proceso de ingesta de datos, se contempla la implementación de una conexión directa con Microsoft Azure Blob Storage.

9.2.3. ETL y Descripción de variables

ETL Registro de conductores:

- Usar expresiones regulares para estandarizar los códigos de los buses.
- Ajustar las variables de fecha para regular las horas.
- Extracción de la parte numérica identificadora del bus.
- Resumir la información por bus para tener las horas y la cantidad de conductores que manejan cada vehículo.

Tabla No. 3 Dataframe resumen_horas_2023.

	No. Interno	Total_Horas_Transcurridas	cantidad_conductores_maneja
0	D1000	3497.525000	208
1	D1001	3906.254444	209
2	D1002	4842.707222	219
3	D1003	4139.165000	216
4	D1004	4298.793333	210
...	...	...	...
2411	Z17-4054	2821.746667	289
2412	Z17-4055	4283.121389	208
2413	Z17-4060	3000.311667	299
2414	Z70-7035	5.023333	4
2415	Z90-7102	16.467778	14

Fuente: Elaboración propia.

ETL Mantenimiento:

- Cambio de formato de fechas a la forma: format='%d.%m.%Y'.
- Cálculo de la variable “días_de_diferencia”, la cual contabiliza la cantidad de días los días que pasa desde el último mantenimiento preventivo, la obtención de esta nueva variable tiene el objetivo de identificar alguna anormalidad de los tiempos de mantenimiento.
- Renombrar la nueva variable creada.
- Resumir los datos para obtener cantidad de mantenimientos por placa.
- Cálculo de la variable cantidad de grupos la cual hace referencia a la cantidad promedio de paquetes de mantenimientos que se le hace a un bus durante un año.
- Estandarización de tipos de variables para que el dataframe los lea bien.
- Fusionar en una sola tabla las dos medidas calculadas “días_de_diferencia”, “cantidad de mantenimientos” y “la cantidad de mantenimientos por placa”.

Tabla No. 4 Dataframe depurado de mantenimientos.

	Placa	dias de diferencia	cantidad_mantenimientos	cantidad_grupos
0	0	55.666667	7	2.333333
1	1	68.250000	6	2.000000
2	2	50.666667	7	2.333333
3	3	50.666667	7	2.333333
4	4	45.500000	7	2.333333

Fuente: Elaboración propi

ETL Kilómetros troncales:

- Homologación de variables eliminando algunos caracteres utilizando la sentencia `str.replace`.
- Reemplazar el nombre de las variables usando el comando `rename`.
- Agrupación de los datos para ver la cantidad de rutas por cada una de las placas de los carros, utilizando el comando `groupby`. Para ver la variable Ruta
- Se concatena la información de kilómetros tanto de zonales como troncales.
- Descripción y conocimiento de las variables.
- Verificación del nombre de las variables utilizando la sentencia `values`.

ETL 01-01-2023, características de buses:

- Identificación de las variables, con la sentencia `values`.
- Análisis de valores únicos de las tablas `nunique`.
- Eliminación de vehículos que no entran al estudio, dejando solamente las categorías "alimentación", "complementario", "especial", "troncal", "urbano".
- Dejar solo las variables necesarias optimizando la cantidad de registros.
- Visualización de filas y columnas.

ETL BASE DPV 2023.xlsx, tabla de fallas:

- Visualización de la cantidad de variables que contiene la tabla de fallas.
- Análisis de valores únicos de las tablas nunique.
- Dejar solo los registros con afectaciones que afectan el indicador el EMIC.
- Visualización de la cantidad de filas y columnas, con la sentencia shape.
- Selección de las variables necesarias.
- Creación del resumen de veces con las cuales se vara un bus.

Tabla No. 5 Dataframe depurado de fallas.

	Placa	conteo_varados
0	GUZ220	1
1	GUZ221	2
2	GUZ223	1

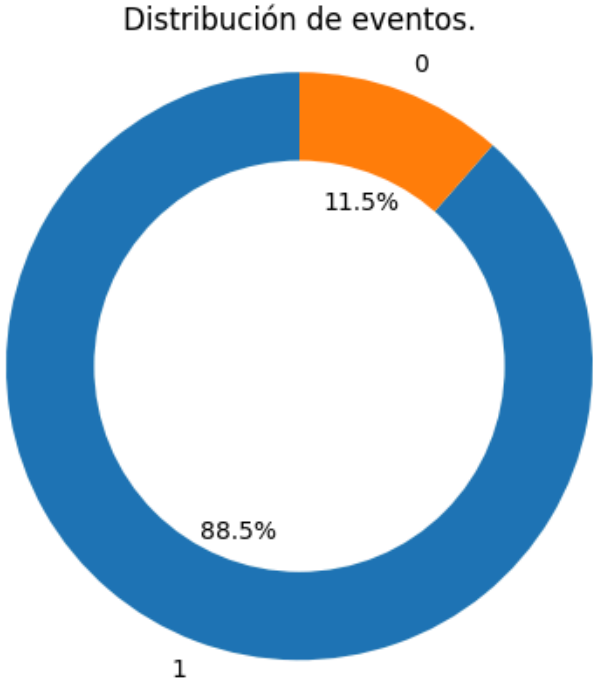
Fuente: Elaboración propia.

- Unificación de fuentes de datos -ETL- Construcción tabla minable, la cual tiene inicialmente 2.337 registros (buses) y 34 variables.

- Anexo de las tablas de registro de manejo de conductores, varados, kilómetros y mantenimientos usando ***"left join"*** y una llave del número interno de vehículo a la tabla de características de buses.

- Definición de la variable llamado evento la cual es una dicotómica donde 1 (uno) es la presencia del evento o sea presenta falla durante el año y 0 (cero) no ha presentado fallas. Esta variable dicotómica servirá como eje central del modelo de predicción, permitiendo identificar los factores que influyen en la probabilidad de ocurrencia de una falla mecánica y, en última instancia, anticipar y prevenir dichas fallas para optimizar la gestión de la flota.
- Gráfica del porcentaje fallas que se presentan por los buses en un año.
- Eliminación de fechas vacías.
- Eliminación de las variables "fecha", "Placa", "No. Interno", “Tipo Adquisición”, “Tipo Acceso Discapacitado”, “Espacio Discapacitados.”

Gráfica No. 2 Resumen gráfico de fallas.



Fuente: Elaboración propia.

- Verificación de la dimensionalidad de las categorías, cantidad de valores únicos y tipos de variables.
- Renombre de variables, cálculo de años de vida útil, eliminación de variables no relevantes.

9.3. Entendimiento de los datos.

9.3.1. Análisis exploratorio.

Se hace uso de estadísticas descriptivas para analizar el comportamiento de las variables numéricas de lo cual podemos decir lo siguiente.

- **Modelo**: El año promedio del modelo es 2013.
- **Cilindraje**: El cilindraje promedio es 6393.93.
- **Capacidad de Pasajeros**: La capacidad promedio de pasajeros es 76.39.
- **Años de Vida Útil**: El promedio de años de vida útil es 28.15. Sin embargo, el valor máximo de 2012 años es un outlier extremo y probablemente un error de

registro, ya que es poco realista que un autobús tenga una vida útil tan larga.

- **Valor**: El valor promedio es $ 196.809.000 millones de pesos.

Tabla No. 6 Estadísticas numéricas de tabla minable.

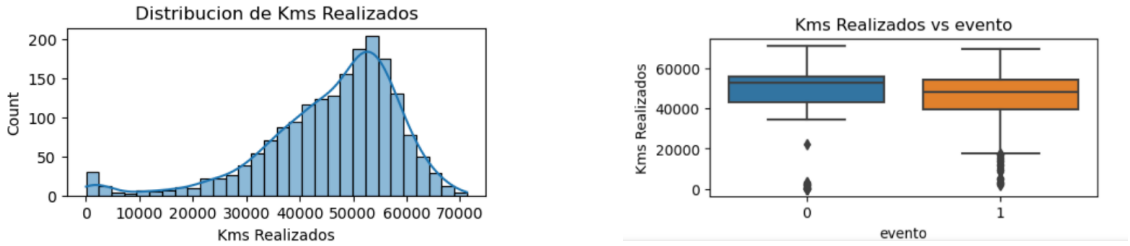
	Modelo	Cilindraje	Capacidad Pasajeros	Años Vida Útil	\
count	2930.000000	2930.000000	2930.000000	2832.000000	
mean	2013.852560	6393.934812	76.397270	28.153249	
std	4.605365	2544.177757	55.634315	175.570356	
min	2003.000000	1598.000000	2.000000	12.000000	
25%	2012.000000	4570.000000	50.000000	12.000000	
50%	2014.000000	5193.000000	55.000000	12.000000	
75%	2015.000000	7201.000000	80.000000	14.000000	
max	2023.000000	12130.000000	250.000000	2012.000000	

	Valor
count	1.628000e+03
mean	1.968090e+08
std	3.059210e+08
min	0.000000e+00
25%	0.000000e+00
50%	0.000000e+00
75%	4.790451e+08
max	8.311561e+08

Fuente: Elaboración propia.

9.3.1.1. Análisis gráfico de los datos.

Gráfica No. 3 Resumen gráfico de kilómetros realizados.

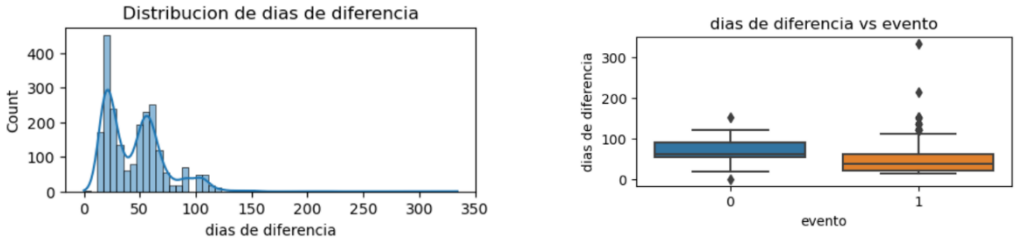


Fuente: Elaboración propia.

Para el periodo de entrenamiento de los datos, se ha identificado que la mayoría de los buses registran un kilometraje entre 50.000 y 60.000 unidades.

Es particularmente intrigante observar que algunos vehículos presentan fallas en sus primeros kilómetros recorridos, lo cual resulta inusual y añade un aspecto aún más interesante a este estudio.

Gráfica No. 4 Resumen gráfico de días de diferencia.



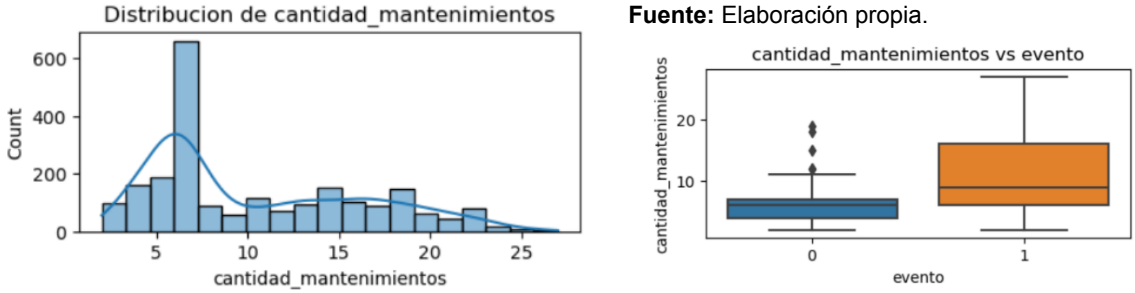
Fuente: Elaboración propia.

**Días de diferencia: Se refiere a la cantidad de días transcurridos entre la realización de dos mantenimientos preventivos consecutivos.*

La representación gráfica sugiere la presencia de dos grupos distintos en el mantenimiento de los vehículos: uno comprendido entre 0 y 40 días, y otro entre 50 y 70 días. Además, se observa que aquellos vehículos con un mayor número de días de mantenimiento probablemente presentan fallas mecánicas.

Sin embargo, si excluimos los vehículos con más de 100 días de mantenimiento, encontramos que la mediana de los días de mantenimiento de los vehículos con fallas es menor que la del grupo de vehículos sin fallas. Este hallazgo resulta inusual, ya que se esperaría que un mantenimiento preventivo más frecuente estuviera asociado con una menor incidencia de fallas mecánicas.

Gráfica No. 5 Resumen gráfico de Cantidad de mantenimientos.

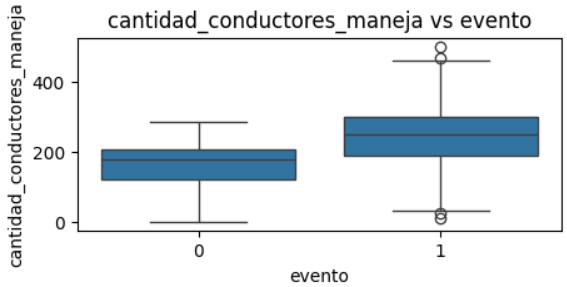


Fuente: Elaboración propia.

Fuente: Elaboración propia.

En este análisis, se destaca la práctica del área de mantenimiento de llevar a cabo un mantenimiento preventivo más exhaustivo en los vehículos que podrían estar mostrando signos de fallos o que han experimentado problemas previos, lo cual proporciona una perspectiva reveladora sobre los costos asociados con los vehículos que requieren atención especial.

Gráfica No. 6 Resumen gráfico de Cantidad de conductores que manejan.



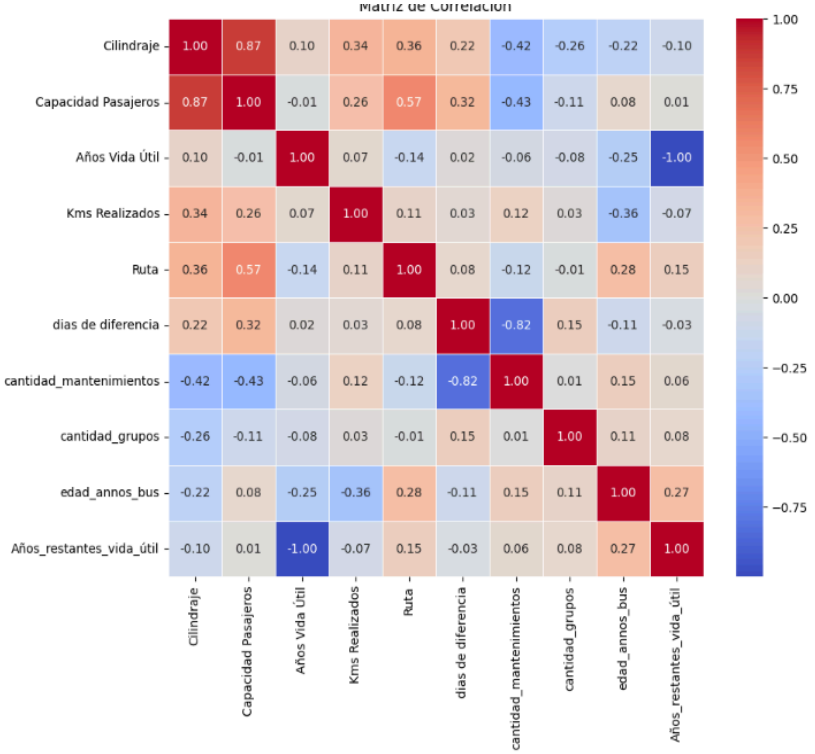
Fuente: Elaboración propia.

En la anterior grafica se observa que la ocurrencia de los varados parece estar asociada con una mayor dispersión en la cantidad de conductores que manejan por bus. Los valores atípicos en el caso del evento sugieren que, en algunas ocasiones, cuando ocurre

el varado, hay un aumento notable en la cantidad de conductores involucrados.

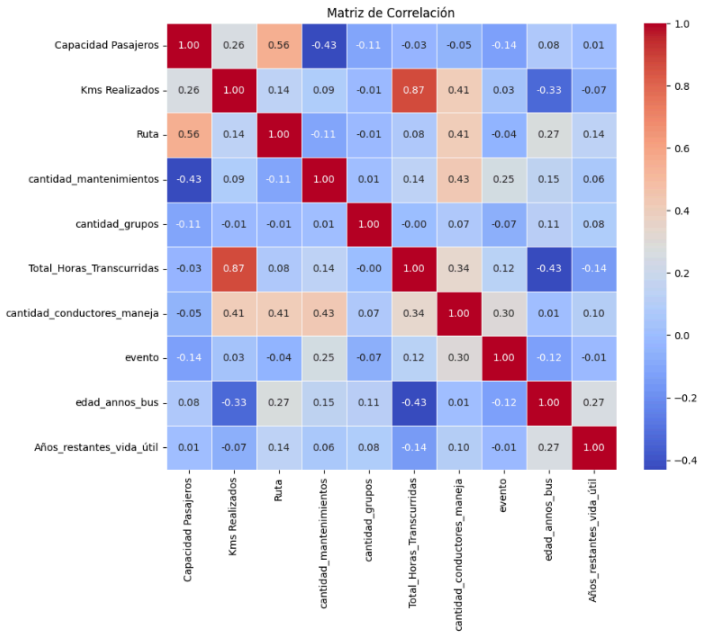
Después de realizar un análisis de correlación de descartar aquellas variables con un coeficiente de Pearson mayor a .80, lo cual evitaría problemas de multicolinealidad dentro del modelo, quedando así:

Gráfica No. 7 Resumen gráfico correlación antes de multicolinealidad.



Fuente: Elaboración propia.

Gráfica No. 8 Resumen gráfico correlación depurado para evitar multicolinealidad.



Fuente: Elaboración propia.

9.3.1.2. Ajuste y validación de tabla minable-Imputación.

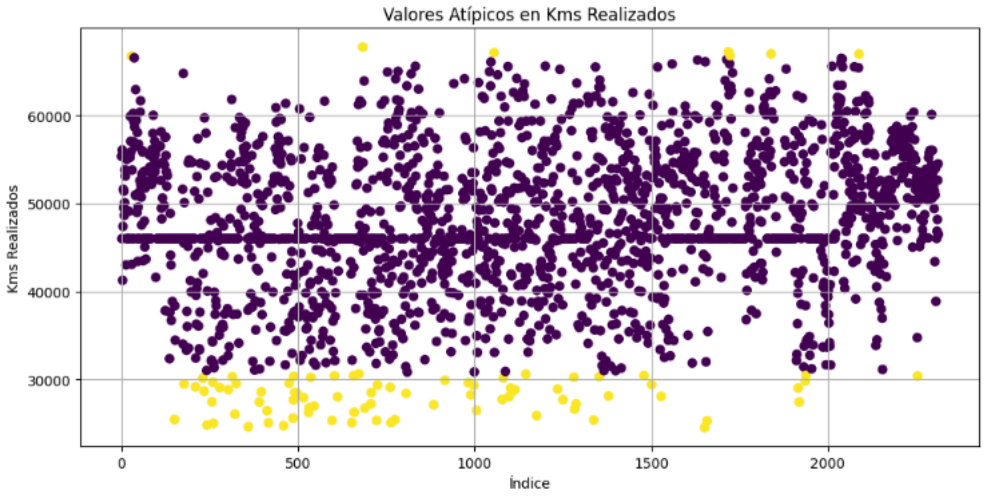
Se encuentra un número óptimo de vecinos a considerar al imputar valores faltantes con KNN. Al utilizar validación cruzada y una métrica de evaluación relevante, se puede estar más seguro de que el valor seleccionado para n_neighbors dará buenos resultados en el modelo de predicción, el cual dio como mejor hiperparametro: 5.

- Imputación de valores vacíos usando KNN.
- Predicción del valor a imputar.
- Uso de validación cruzada para encontrar los mejores parámetros para imputar
- Aplicación de un pipeline para ajustar la imputación.
- Declaración y entrenamiento del modelo para realizar la imputación.

9.3.1.3. Manejo de valores atípicos

- Para las variables numéricas del modelo, se crea una función en Python para detectar y cambiar los valores atípicos, por la mediana de los registros no atípicos.
    - Se calcula el MAD (desviación absoluta mediana), los límites que se estén por fuera se consideran como atípicos.
- Se realizan gráficas con colores, para facilitar la identificación de los atípicos y hacer cambios respectivos.
- El análisis de los kilómetros realizados se muestra de forma representativa en este documento. El análisis completo de todas las variables está disponible en el notebook adjunto.

Gráfica No. 9 Resumen gráfico de identificación de valores atípicos para kilómetros realizados.



Fuente: Elaboración propia.

9.3.1.4. Estandarización de variables

Se utiliza este procedimiento para mejorar el rendimiento y la interpretabilidad de los modelos, especialmente para aquellos que son sensibles a la escala de medida para las variables independientes, lo cual puede ocasionar un sesgo no voluntario de los resultados. Al tener una escala más pequeña el descenso del gradiente se optimiza convergiendo más rápido.

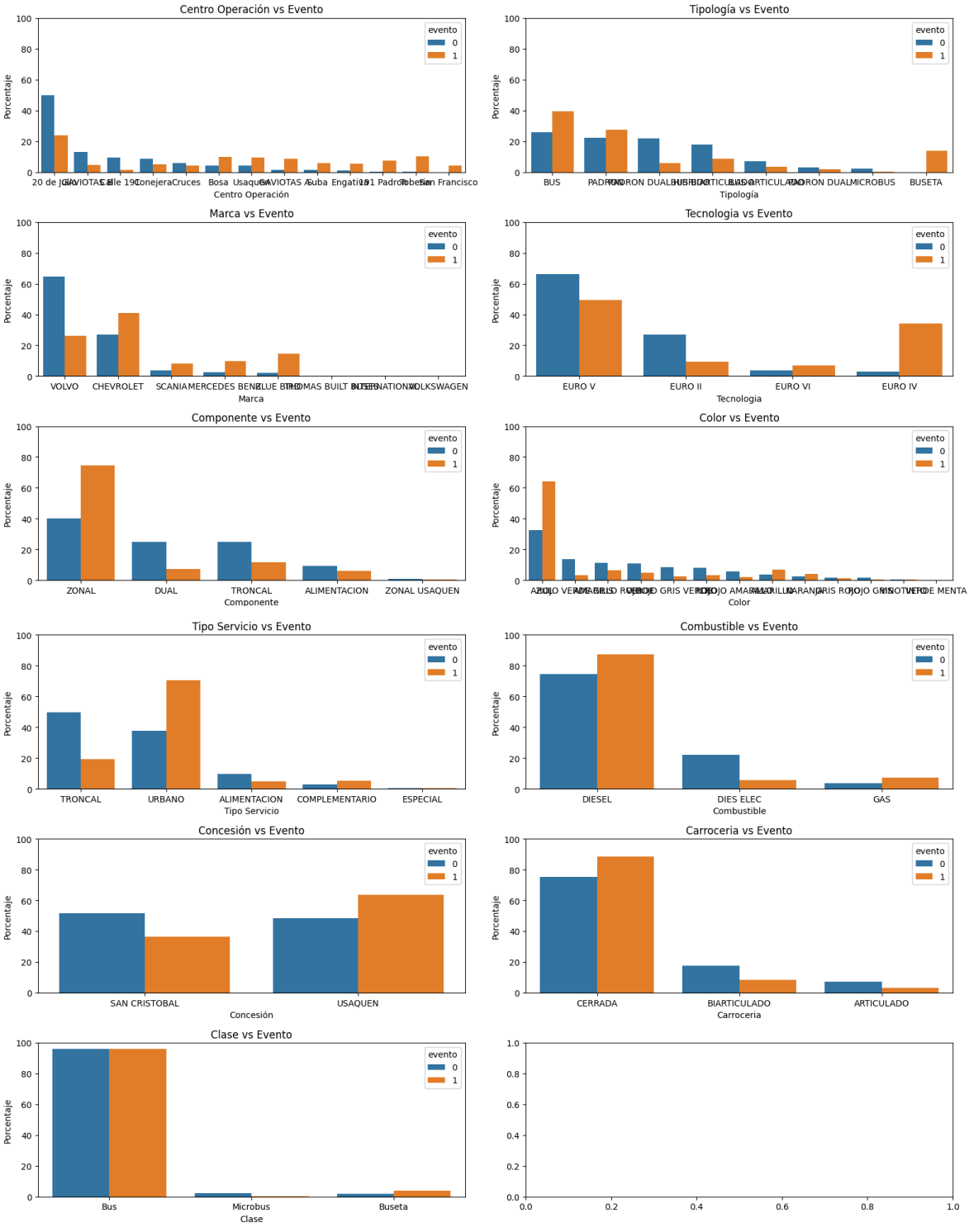
La estandarización que se hace es restando la media y dividiendo por la desviación estándar, lo que da como resultados variables con media 0 y desviación estándar 1.

9.3.1.5. Análisis descriptivo de las variables cualitativas

Se llevó a cabo un análisis descriptivo de las variables categóricas para comprender su distribución y frecuencia de categorías. Este análisis preliminar permitió identificar redundancias y oportunidades para reducir la dimensionalidad de los datos, con el fin de mejorar el desempeño del modelo.

Sin embargo, se anexa a continuación algunas variables como ilustración del ejercicio.

Gráfica No. 10 Resumen gráfico de análisis de frecuencia para variables cualitativas.



Fuente: Elaboración propia.

9.3.1.6. Análisis de asociación cualitativo

Utilizando la prueba Chi-cuadrado evalúa si la distribución de la variable dependiente varía significativamente entre las diferentes categorías de las variables independientes. En otras palabras, busca determinar si existe una asociación o dependencia entre las variables. Lo cual permite conocer si se encuentra una asociación significativa, el valor del estadístico Chi-cuadrado y su correspondiente valor p pueden proporcionar una idea de la fuerza de esta asociación. Un valor Chi-cuadrado más alto indica una asociación más fuerte, mientras que un valor p más bajo indica una mayor confianza en que la asociación no se debe al azar lo cual, puede revelar patrones interesantes en los datos, por último, puede ser una guía la selección de variables para modelos predictivos, ayudando a identificar las variables cualitativas independientes que son más relevantes para predecir la variable dependiente. Estas variables pueden luego ser incluidas en modelos para mejorar su rendimiento. Para lo cual se realizaron los siguientes pasos:

a) Validar hipótesis: En algunos casos, el análisis Chi-cuadrado se puede utilizar para probar hipótesis específicas sobre la relación entre las variables. Por ejemplo, se podría plantear la hipótesis de que no hay asociación entre el género de una persona y su preferencia por un determinado producto.

a) Cuantificación de la cantidad de categorías de las variables no numéricas.
b) Realización de pruebas Chi-cuadrado de asociación con la variable objetivo de tal manera que permita ver el nivel de correspondencia.
c) Uso de gráficos de asociación para verificar de manera visual la influencia de la variable target en cada una de las categorías de las variables independientes categóricas.
d) Elaboración de las tablas de contingencia en donde se cruzan cada una de las variables categóricas frente a la variable respuesta.

Tabla No. 7 Análisis de resultados, tablas de contingencia.

.

Tabla de frecuencia para Componente:

evento	0	1
Componente		
ALIMENTACION	0.181159	0.818841
DUAL	0.301370	0.698630
TRONCAL	0.213592	0.786408
ZONAL	0.064474	0.935526
ZONAL USAQUEN	0.166667	0.833333

Tabla de frecuencia para Tipo Servicio:

evento	0	1
Tipo Servicio		
ALIMENTACION	0.198413	0.801587
COMPLEMENTARIO	0.067308	0.932692
ESPECIAL	0.142857	0.857143
TRONCAL	0.250000	0.750000
URBANO	0.064201	0.935799

Tabla de frecuencia para Concesión:

evento	0	1
Concesión		
SAN CRISTOBAL	0.169231	0.830769
USAQUEN	0.088152	0.911848

Tabla de frecuencia para Clase:

evento	0	1
Clase		
Bus	0.116920	0.883080
Buseta	0.060241	0.939759
Microbus	0.545455	0.454545

Tabla de frecuencia para Marca:

evento	0	1
Marca		
BLUE BIRD	0.019231	0.980769
CHEVROLET	0.078366	0.921634
INTERNATIONAL	NaN	1.000000
MERCEDES BENZ	0.034653	0.965347
SCANIA	0.038961	0.961039
THOMAS BUILT BUSES	NaN	1.000000
VOLKSWAGEN	NaN	1.000000
VOLVO	0.244604	0.755396

Tabla de frecuencia para Combustible:

evento	0	1
Combustible		
DIES ELEC	0.329545	0.670455
DIESEL	0.099695	0.900305
GAS	0.053571	0.946429

Tabla de frecuencia para Centro Operación:

evento	0	1
Centro Operación		
191 Padron	0.006494	0.993506
20 de Julio	0.215686	0.784314
Bosa	0.050000	0.950000
Calle 191	0.462963	0.537037
Conejera	0.185484	0.814516
Cruces	0.147059	0.852941
Engativa	0.026786	0.973214
GAVIOTAS A	0.018293	0.981707
GAVIOTAS B	0.271318	0.728682
San Francisco	NaN	1.000000
Suba	0.032520	0.967480
Toberin	0.004695	0.995305
Usaquen	0.043478	0.956522

Tabla de frecuencia para Tecnologia:

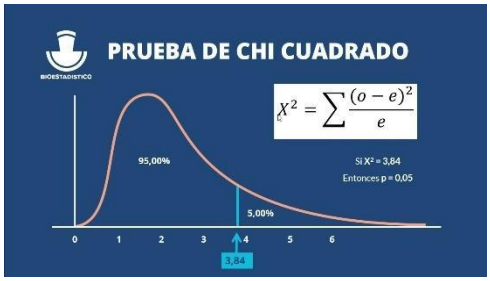
evento	0	1
Tecnologia		
EURO II	0.267925	0.732075
EURO IV	0.011380	0.988620
EURO V	0.149063	0.850937
EURO VI	0.053571	0.946429

Tabla de frecuencia para Tipo Adquisición:

evento	0	1
Tipo Adquisición		
NUEVO	0.099946	0.900054
USADO	0.210682	0.789318

Fuente: Elaboración propia.

Tabla No. 8 Análisis de resultados pruebas chi-cuadrado.



Variables más significativas:	P-valor
Centro Operación:	0,00
Tipología:	0,00
Marca:	0,00
Tecnologia:	0,00
Color:	0,00
Componente:	0,00
Tipo Servicio:	0,00
Combustible:	0,00
Tipo Adquisición:	7.874.862,12
Carroceria:	18.719.422,73
Concesión:	18.984.047.100,02
Clase:	114.326.879.012,62

Fuente: Elaboración propia.

e) Centro Operación: El COP (Centro de Operaciones) donde se encuentran los buses está fuertemente asociado con la probabilidad de falla. Algunos centros de operación pueden tener tasas de falla mucho más altas que otros, esto nos puede dar una hipótesis en cuanto a que la ubicación puede influir en la falla de buses.
f) Tipología: El tipo de bus está fuertemente asociado con la probabilidad de falla. Algunos tipos de automotores pueden ser más propensos a fallar que otros.
g) Marca, Tecnología y Componente: Aunque también son estadísticamente significativas, estas variables parecen tener una asociación más débil con la falla del bus en comparación con el centro de operación y la tipología.

9.3.1.7. Agrupaciones variables Cualitativas

Este procedimiento se realiza para reducir la dimensionalidad de las variables no numéricas, haciendo un análisis estadístico de frecuencia y se valida con una asociación empírica basada en el conocimiento del negocio.

a) Se agrupan los valores "MICROBUS" y "BUSETA" en una sola categoría llamada "BUSETA".
b) Los demás valores ("BUS", "PADRON", "BUS BIARTICULADO", "PADRON DUAL HIBRIDO", "BUS ARTICULADO" y "PADRON DUAL") se mantienen sin cambios.
c) Se agrupan los valores "ESPECIAL" y "COMPLEMENTARIO" en una sola categoría llamada "COMPLEMENTARIO".
d) Los demás valores ("URBANO", "TRONCAL" y "ALIMENTACION") se mantienen sin cambios.

9.3.1.8. Dumización de variables cualitativas.

En este paso se preparan las variables categóricas en el conjunto de datos train para su uso en modelos de aprendizaje automático mediante la aplicación de la técnica One-Hot Encoding. Esto facilitará el análisis y la modelización de los datos.

a) Se crea una función que permite agrupar las categorías con poca frecuencia anexándolas a similares u otras.
b) Uso de la sentencia `get_dummies` para aplicar la codificación one-hot a las variables tipo objeto.
c) Verificación de la codificación.

9.4. Modelado.

Se entrenaron los siguientes modelos de machine learning:

- *Regresión logística.*

- *Random Forest.*
- *Árbol de decisión.*
- *SVM.*
- *Redes neuronales.*

9.4.1. Selección de variables

El objetivo principal de este paso es realizar una selección de características en el conjunto de datos llamado ***train*** para mejorar el rendimiento de los modelos. La selección de características es el proceso de elegir un subconjunto de las variables más relevantes para el problema que se está modelando. En este caso, se utilizó la técnica Lasso ***(Least Absolute Shrinkage and Selection Operator)*** para identificar y seleccionar las características más importantes. Y esto se hace para evitar el sobreajuste, eliminación del ruido y mejorar la interpretabilidad de los modelos.

9.4.2. División del grupo de datos en conjuntos de entrenamiento y prueba

Partición del 60% en tablas de entrenamiento y 40% de los datos va a estar en el grupo de validación con una semilla aleatoria, para sus posteriores replicaciones igual a 42.

9.4.3. Establecimiento de los estadísticos de evaluación

La matriz de confusión es una tabla que se utiliza para evaluar el rendimiento de un modelo de clasificación. Resume los resultados de las predicciones del modelo en términos de las clases reales y las clases predichas.
Verdaderos Positivos (VP): El modelo predijo correctamente la clase positiva.
Verdaderos Negativos (VN): El modelo predijo correctamente la clase negativa.
Falsos Positivos (FP): El modelo predijo la clase positiva, pero la clase real era negativa.
Falsos Negativos (FN): El modelo predijo la clase negativa, pero la clase real era positiva.
Cuando se trabaja con clasificación binaria, donde las clases se denominan positivas y negativas, la exactitud (accuracy) puede ser analizada en mayor detalle a través de cuatro métricas fundamentales: verdaderos positivos (VP), verdaderos negativos (VN), falsos positivos (FP) y falsos negativos (FN). Estas métricas nos permiten comprender no solo el porcentaje general de aciertos del modelo, sino también su desempeño específico en la identificación de cada clase.

Fórmula No. 2 Cálculo Accuracy.

$$Accuracy = (VP + VN) / (VP + VN + FP + FN)$$

Fuente: Elaboración propia.

Otro estadístico que se utiliza en este trabajo de grado es el F1 que es una métrica de evaluación utilizada en problemas de clasificación binaria, especialmente cuando existe un desbalance entre las clases. Se calcula a partir de la matriz de confusión, que resume los resultados de las predicciones de un modelo en términos de verdaderos positivos (VP), falsos positivos (FP), verdaderos negativos (VN) y falsos negativos (FN). Este estadístico puede ser considerada como un punto medio entre la precisión y el Recall

Precisión: Mide la proporción de predicciones positivas que son realmente correctas.

Formula No. 3 Cálculo Precisión.

(VP / (VP + FP))

Fuente: Elaboración propia.

Recall: Mide la proporción de verdaderos positivos que el modelo identifica correctamente.

Formula No. 4 Cálculo Recall.

(VP / (VP + FN))

Fuente: Elaboración propia.

F1: Esta es la métrica que más se utilizará porque equilibra la precisión (qué tan bien el modelo evita falsos positivos) y el recall (qué tan bien el modelo encuentra todos los positivos verdaderos). Útil cuando las clases están desbalanceadas.

Formula No. 5 Cálculo F1.

*2 * (Precisión * Recall) / (Precisión + Recall)*

Fuente: Elaboración propia.

9.4.4. Modelo de regresión logística.

Este es un modelo de clasificación que utiliza una función sigmoidal (similar a una "***S***") para transformar una combinación lineal de variables predictoras en una probabilidad de pertenencia a una clase (0,1). Esta probabilidad se utiliza para tomar decisiones de clasificación, asignando una observación a la clase con mayor probabilidad. Entregando los siguientes resultados:

Tabla No. 9 Tabla estadísticos de evaluación.

Statistics	F1 Score		Precisión		Matriz de confusión								Especificidad		Recall	
Modelos	Train	Validation	Train	Validation	Train				Validation				Train	Validation	Train	Validation
							Pronóstico				Pronóstico					
							0	1			0	1				
Logistic Regression	0,943	0,950	0,902	0,911	Real	0	46	123	Real	0	20	76	0,272	0,208	0,989	0,992
						1	14	1219		1	7	832				

Fuente: Elaboración propia.

Tabla No. 10 Tabla resultado features más relevantes del modelo.

```
Características más significativas en el modelo:
                          Feature  Coefficient  ExpCoefficient
21              Tecnologia_EURO IV     2.347947       10.464068
22               Tecnologia_EURO V     1.494625        4.457665
6      Centro Operación_GAVIOTAS B    -1.486866        0.226080
9         Centro Operación_Toberin     1.451426        4.269200
2       Centro Operación_Calle 191    -1.378364        0.251991
```

Fuente: Elaboración propia.

El modelo muestra un rendimiento similar en ambos conjuntos de datos, lo que sugiere una buena capacidad de generalización y un riesgo medio de sobreajuste. El F1 score y la precisión son altos, indicando un buen equilibrio entre identificar correctamente los casos positivos y minimizar los falsos positivos.
Las variables más significativas en el modelo, según el valor absoluto de sus coeficientes, son:
Tecnología_EURO IV y V: Un coeficiente positivo alto indica que los autobuses con esta tecnología tienen una mayor probabilidad de presentar fallas.
Centro Operación_Calle 191 y Gaviotas: Un coeficiente negativo alto sugiere que los autobuses que operan en este centro tienen una menor probabilidad de fallas.
probabilidad de fallas en comparación con otros tipos de autobuses.

Interpretación de Odds Ratios:

OR > 1: Un aumento de una unidad en la variable predictora se asocia con un aumento en las probabilidades de que ocurra el evento (falla mecánica).
OR<1: Un aumento de una unidad en la variable predictora se asocia con una disminución en las probabilidades de que ocurra el evento.
OR = 1: La variable predictora no tiene efecto sobre las probabilidades del evento.

Tecnologia_EURO IV: Los autobuses con esta tecnología tienen 10.4 veces más probabilidades de fallar que los que no la tienen.
Tecnologia_EURO V: Los autobuses con esta tecnología tienen 4.4 veces más probabilidades de fallar que los que no la tienen.
Centro Operación_Calle 191, Centro Operación_GAVIOTAS B, Centro Operación_Cruces: Los autobuses que operan en estos centros tienen una probabilidad significativamente menor de fallar.

9.4.5. Modelo de árboles de decisión.

Este modelo utiliza una estructura jerárquica de decisiones para predecir o clasificar datos. En cada nodo del árbol, se realiza una pregunta sobre una característica, y la respuesta determina la rama a seguir. El objetivo es dividir los datos en grupos cada vez más homogéneos con respecto a la variable objetivo.
Los criterios de homogeneidad se utilizan para medir qué tan "puro" o similar es un conjunto de datos en relación con la variable objetivo. Los criterios más comunes son:
Índice de Gini: Mide la impureza de un nodo como la probabilidad de que una muestra aleatoria del nodo sea clasificada incorrectamente si se le asigna la etiqueta de la clase más frecuente en ese nodo. Un índice de Gini de 0 indica un nodo perfectamente puro.
Entropía: Mide la cantidad de incertidumbre o desorden en un conjunto de datos. Un valor

de entropía de 0 indica un nodo perfectamente homogéneo.
Chi-cuadrado: Mide la independencia estadística entre una característica y la variable objetivo. Un valor alto de chi-cuadrado indica una fuerte asociación entre la característica y la variable objetivo, lo que sugiere que la característica es útil para dividir los datos en grupos más homogéneos.

Tabla No. 11 Tabla estadísticos de evaluación.

Statistics	F1 Score		Precisión		Matriz de confusión								Especificidad		Recall	
Modelos	Train	Validation	Train	Validation	Train				Validation				Train	Validation	Train	Validation
							Pronóstico				Pronóstico					
							0	1			0	1				
Decision Tree	0,950	0,949	0,911	0,909	Real	0	53	116	Real	0	17	79	0,314	0,177	0,993	0,993
						1	9	1224		1	6	833				

Fuente: Elaboración propia.

Tabla No. 12 Variables Importantes del modelo.

```
Análisis de Significancia de Variables:
                          Feature  Importance
29       Total_Horas_Transcurridas    0.393234
20                     Marca_VOLVO    0.222334
25                  Kms Realizados    0.186007
31                  edad_annos_bus    0.103854
5      Centro Operación_GAVIOTAS A    0.039800
30     cantidad_conductores_maneja    0.035014
1            Centro Operación_Bosa    0.012604
26                            Ruta    0.005683
28                 cantidad_grupos    0.001470
```

Fuente: Elaboración propia.

Conjunto de entrenamiento: El modelo muestra un buen rendimiento en el conjunto de entrenamiento, con un puntaje F1 de 0.950 y una precisión de 0.949. Esto indica que el modelo es capaz de predecir con alta precisión las fallas en los datos que ha visto durante el entrenamiento.
Conjunto de pruebas: El rendimiento en el conjunto de prueba es ligeramente inferior, con un puntaje F1 de 0.911 y una precisión de 0.909. Esto sugiere que el modelo generaliza razonablemente bien a nuevos datos, aunque hay un ligero grado de sobreajuste.
Total_Horas_Transcurridas: Esta es la variable más importante del modelo, lo que sugiere que el desgaste asociado a la cantidad de horas de uso del vehículo es un posible factor determinante de los daños en la flota.
Marca_VOLVO: La marca del autobús también juega un papel importante, indicando que ciertos modelos o marcas pueden influir positiva o negativamente en las fallas, dependiendo de la ubicación del nodo dentro del árbol de decisión.
Kilómetros realizados (Kms Realizados): Es la variable más importante, lo que sugiere que el desgaste acumulado por el uso es un factor determinante en la probabilidad de avería.
Edad del bus (edad_annos_bus): La edad del autobús es otro factor relevante, ya que a medida que los vehículos envejecen, aumentan las probabilidades de que presenten fallas.

Centro Operación_GAVIOTAS A: El centro de operación donde se encuentra el autobús también influye en la probabilidad de falla, lo que podría deberse a diferencias en las condiciones de operación, rutas o prácticas de mantenimiento.
cantidad_conductores_maneja: Una mayor cantidad de conductores por bus es un fuerte indicio de un mayor desgaste del vehículo, lo cual puede conducir a fallas.
El modelo de árbol de decisión proporciona información valiosa sobre los factores que influyen en las fallas de los autobuses, como el kilometraje, la marca, el centro de operación, la edad y el tipo de vehículo. Aunque el modelo muestra un buen rendimiento general, se debe tener en cuenta el desbalance de clases y la ligera disminución en el rendimiento en el conjunto de prueba.

9.4.6. Modelo de red neuronal

Este es un modelo computacional inspirado en el funcionamiento del cerebro humano. Aprende a reconocer patrones en los datos a través de capas de neuronas interconectadas que ajusta sus pesos durante el entrenamiento. Aunque son poderosas para tareas complejas, las redes neuronales tienen una "caja negra": es difícil interpretar cómo cada variable independiente influye en la predicción final. Esto se debe a la interacción compleja de múltiples capas y neuronas, lo que dificulta aislar el efecto individual de cada variable.
Matriz de Confusión para el conjunto de entrenamiento:

Tabla No. 13 Variables Importantes del modelo.

Statistics	F1 Score		Precisión		Matriz de confusión								Especificidad		Recall	
Modelos	Train	Validation	Train	Validation	Train				Validation				Train	Validation	Train	Validation
							Pronóstico				Pronóstico					
Neural Network	0,950	0,950	0,910	0,909	Real	0	52	117	Real	0	15	81	0,308	0,156	0,993	0,995
						1	9	1224		1	4	835				

Fuente: Elaboración propia.

El modelo de red neuronal (con 3 capas ocultas y 500 iteraciones), muestra un excelente rendimiento en el conjunto de entrenamiento, con un puntaje F1 de 0.950 y una precisión de 0.950. Esto indica que el modelo ha aprendido a clasificar las instancias de entrenamiento con alta precisión.
El modelo tiende a ser más conservador en sus predicciones, priorizando la reducción de falsos negativos (no predecir una falla cuando sí ocurre) sobre la reducción de falsos positivos.

9.4.7. Modelo SVM

Este es un algoritmo de aprendizaje automático supervisado utilizado para clasificación y regresión. En general su objetivo principal es encontrar el mejor hiperplano que separe las clases de datos con el margen más amplio posible. Esto lo hace maximizando la distancia entre los puntos de datos más cercanos de cada clase, conocidos como vectores de soporte. SVM puede manejar datos lineales y no lineales utilizando el mejor kernel, que transforma los datos a un espacio de mayor dimensión donde pueden ser separados linealmente.

Tabla No. 14 Variables Importantes del modelo.

Statistics	F1 Score		Precisión		Matriz de confusión								Especificidad		Recall	
Modelos	Train	Validation	Train	Validation	Train				Validation				Train	Validation	Train	Validation
							Pronóstico				Pronóstico					
							0	1			0	1				
SVM	0,952	0,955	0,912	0,916	Real	0	52	117	Real	0	19	77	0,308	0,198	0,995	0,998
						1	6	1227		1	2	837				

Fuente: Elaboración propia.

El modelo SVM muestra un excelente rendimiento en el conjunto de entrenamiento, con un puntaje F1 de 0.952 y una precisión de 0.912. Esto indica una alta capacidad para clasificar correctamente las instancias de entrenamiento.

9.4.8. Modelo Random Forest

se puede ejemplarizar a partir como donde cada "árbol" es un modelo simple que aprende a clasificar o predecir la falla. El Random Forest combina las predicciones de todos los árboles para tomar una decisión final más precisa y robusta. Es como si cada árbol votará, y la respuesta más popular ganara.

Tabla No. 15 Variables Importantes del modelo.

Statistics	F1 Score		Precisión		Matriz de confusión								Especificidad		Recall	
Modelos	Train	Validation	Train	Validation	Train				Validation				Train	Validation	Train	Validation
							Pronóstico				Pronóstico					
							0	1			0	1				
Random Forest	1,000	0,931	1,000	0,901	Real	0	169	0	Real	0	33	63	1,000	0,344	1,000	0,964
						1	0	1233		1	30	809				

Fuente: Elaboración propia.

Conjunto de entrenamiento: El modelo Random Forest presenta un rendimiento excepcional en el conjunto de entrenamiento, con un puntaje F1 de 1 y una precisión de 0.931. Esto indica una capacidad casi perfecta para clasificar correctamente las instancias de entrenamiento.
Conjunto de pruebas: El rendimiento en el conjunto de prueba, aunque ligeramente inferior, sigue siendo muy bueno, con un puntaje F1 de 1 y una precisión de 0.901. Esta pequeña disminución sugiere un mínimo sobreajuste, pero en general, el modelo demuestra una excelente capacidad de generalización a nuevos datos, lo que es fundamental para su aplicación en situaciones reales.

9.5. Ajuste de los modelos haciendo uso del punto de corte óptimo.

Durante la presentación de los resultados de los modelos fundacionales, se evidenció que los modelos tienen una alta sensibilidad (mayor a 0.98), pero una baja especificidad (menor a 0.32). Esto indica que el modelo es muy bueno identificando los vehículos con fallas; sin embargo, no es muy preciso para identificar los buses que no se van a varar.
Para dar solución a esta problemática, la literatura[2] provee varias alternativas, que van desde

[2] Data Mining: Practical Machine Learning Tools and Techniques" de Ian H. Witten, Eibe Frank y Mark A. Hall.

el balanceo de datos hasta la anexión de una mayor cantidad de registros o variables que puedan explicar mejor el fenómeno. Sin embargo, al tener pocas filas (2337), el balanceo no es una buena opción. La consecución de nuevas variables puede ser una buena alternativa, pero como se ha mencionado en este documento, puede ser compleja. Ante este panorama, se optó por encontrar el punto de corte óptimo, que es un equilibrio entre una sensibilidad muy alta y una especificidad relativamente baja. El uso de la curva ROC nos puede ayudar a encontrar este balance que necesita nuestro modelo.

Tabla No. 15 Estadísticas después de usar el punto de corte.

Statistics	Statistics	F1 Score		Precisión		Especificidad		Recall	
Modelos	Punto de Corte Óptimo	Train	Validation	Train	Validation	Train	Validation	Train	Validation
Logistic Regression	0.897	0,850	0,834	0,767	0,742	0,911	0,885	0,748	0,726
Decision Tree	0.880	0,859	0,856	0,781	0,772	0,905	0,906	0,764	0,757
Neural Network	0.910	0,832	0,834	0,745	0,743	0,953	0,938	0,717	0,721
SVM	0,878	0,809	0,818	0,717	0,722	0,941	0,938	0,686	0,697
Random Forest	0,916	0,943	0,856	0,907	0,771	1,000	0,833	0,894	0,764

Fuente: Elaboración propia.

Después de implementar el punto de corte óptimo, se evidencia que los estadísticos de especificidad y sensibilidad (recall) se estabilizan un poco mejor si se compara con los resultados anteriores. Como siguiente paso, se procederá a analizar el mejor F1 Score de los datos de validación.

9.6. Elección del algoritmo de machine learning más adecuado.

En consecuencia, a los resultados anteriores, se selecciona el Árbol de Decisión como el modelo óptimo, con un F1 Score de 0.856 en los datos de validación, resultado obtenido después de tratar un posible desbalance de las clases. A pesar de que empata con el modelo Random Forest, este último muestra un menor desempeño en los estadísticos de los datos de validación en comparación con el conjunto de datos de entrenamiento, lo cual sugiere un posible sobreajuste. Por este motivo, se procede a descartar este modelo como el mejor.

Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow.

9.7. Implementación y puesta en producción del mejor modelo

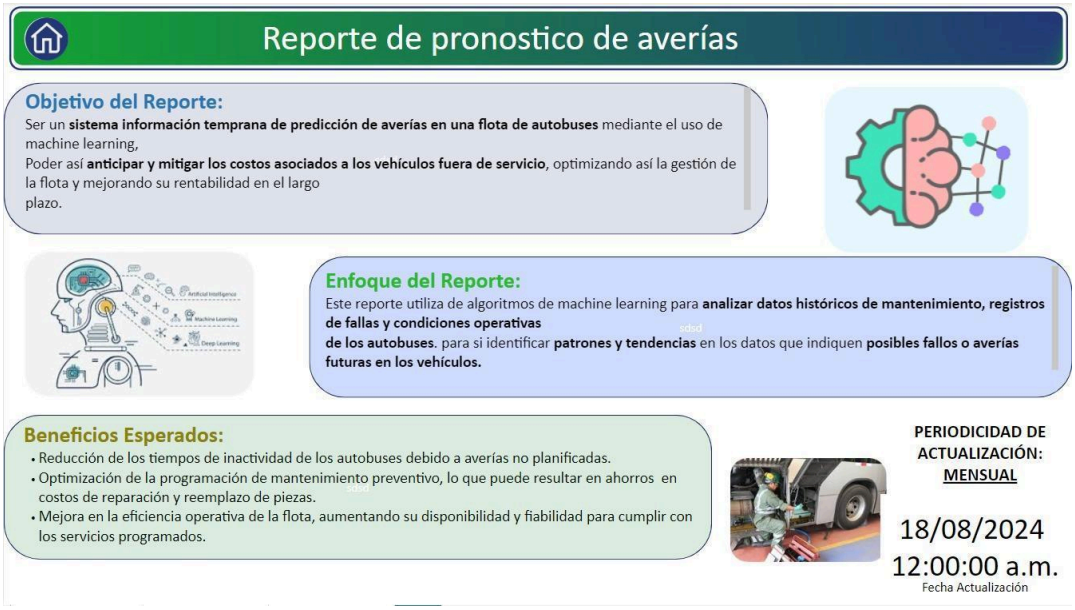
Se empleó Azure para almacenar los mejores modelos de machine learning en formato pickle (.pkl). Pickle es una técnica de serialización que convierte objetos complejos de Python en secuencias de bytes, facilitando su almacenamiento y transporte. Esta solución nos permite mantener la integridad de los modelos, garantizar una carga rápida y aprovechar las ventajas de la nube de Azure.

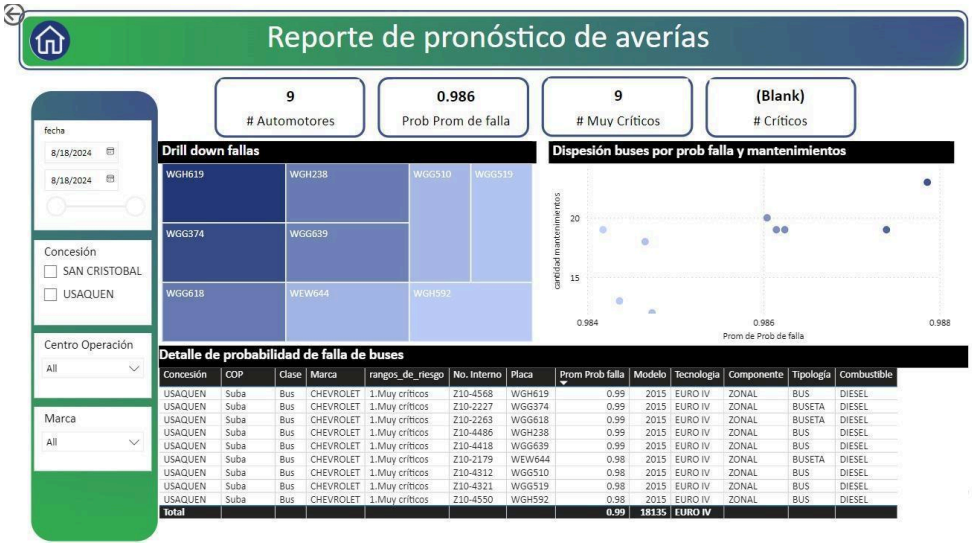
A continuación, se creó un notebook de entrenamiento adicional, idéntico al anterior en cuanto a las reglas de depuración y carga de datos. Sin embargo, este notebook se distingue por realizar consultas en un entorno productivo, empleando una ventana móvil de 370 días y prescindiendo de la fase de modelado. Este último notebook se utiliza para poner en producción el modelo y su posterior visualización en un reporte de BI.

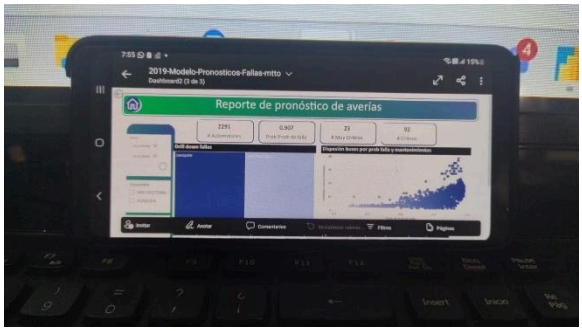
9.8. Construcción del modelo de BI para seguimiento de rendimiento del modelo.

Con el objetivo de evaluar de manera exhaustiva el desempeño de nuestro modelo, se añadió una nueva variable al código. Esta variable permite agrupar los vehículos en percentiles, facilitando así el análisis de los resultados a través de herramientas de inteligencia de negocios. Los reportes obtenidos ofrecen una visión detallada del comportamiento de cada segmento, lo que nos permite identificar oportunidades de mejora.

Gráfica No. 11 Resumen gráfico del tablero de control del modelo de pronóstico de fallas.







Fuente: Elaboración propia.

Como se evidencia en las gráficas presentadas anteriormente, el modelo es completamente funcional y está implementado para ser consultado regularmente por las áreas interesadas,

tanto desde dispositivos móviles como desde computadoras portátiles. El acceso al reporte se realiza a través de herramientas de Microsoft, como Power BI, lo que garantiza una experiencia dinámica y adaptada a las necesidades del proyecto. Esto permite cumplir con los objetivos planteados inicialmente en este proyecto de grado.

9.9. Conclusiones y recomendaciones.

Este estudio se destaca por ser pionero en la aplicación de modelos de machine learning para la predicción de averías en flotas de autobuses de Consorcio Express, abriendo un camino para futuros desarrollos en esta área. Además, se busca plantear el modelo mediante la incorporación de nuevas variables relevantes, lo que permitirá una predicción más precisa y confiable.

Uno de los principales retos enfrentados fue la identificación y selección de las variables más influyentes en la ocurrencia de averías. Esta búsqueda exhaustiva de datos y variables ha permitido reafirmar la importancia de factores tradicionales, como el kilometraje recorrido, en el mantenimiento preventivo. Sin embargo, también ha dado lugar a nuevas interrogantes, como la posible relación entre la frecuencia de mantenimientos preventivos y la probabilidad de avería. Esta hipótesis, surgida en esta primera fase de la investigación, plantea la necesidad de analizar si los autobuses que reciben más mantenimiento preventivo son también los que presentan un mayor número de fallas, lo que podría indicar problemas subyacentes en la calidad de los mantenimientos o en la selección de los vehículos que los reciben.

Este estudio no solo contribuye al desarrollo de modelos predictivos más efectivos, sino que también genera nuevas preguntas de investigación que podrían guiar futuros estudios y mejorar aún más la gestión de la flota de autobuses.
Modelos: Dieron resultados coherentes y no alejados con la realidad la cual es la finalidad teórica del machine learning, validando hipótesis que se tenían como ciertas y otras que se consideran como nuevas.
Predicción semanal: La granularidad semanal en la predicción de averías presenta desafíos considerables debido a la falta de información completa y a la presencia de valores faltantes en los datos. Además, se detectaron errores inherentes a la recopilación de datos, como referencias a vehículos inexistentes o datos incompletos para ciertas semanas. Estos problemas, que afectan la precisión del modelo a nivel semanal, se mitigan al resumir los resultados anualmente.
Por lo tanto, se propone mantener el modelo de predicción inicial, ejecutándose cada 15 días y utilizando una ventana móvil para agregar los resultados y obtener pronósticos anualizados. Esta estrategia permite aprovechar la información disponible y minimizar el impacto de los errores en los datos, al tiempo que se evalúa la eficacia del modelo en un contexto de predicción a corto plazo.
Acceso y confiabilidad de los datos: La fragmentación y falta de accesibilidad de los datos fuente dificulta la obtención de información confiable. Es crucial abordar esta problemática para garantizar la calidad y validez del modelo.
Mejora del modelo: Se recomienda mantener la estructura actual del modelo, pero enriquecerlo con variables adicionales, tanto cuantitativas como cualitativas, para mejorar su capacidad predictiva de la tasa de los falsos positivos y reducir el pequeño sobreajuste observado.
Actualidad: Se automatizó las consultas, conectándose a la nube de Microsoft Azure (Azure), e implementar el modelo mejorado, incorporando nuevas variables.

Foco en variables clave: Se recomienda centrar los esfuerzos de mantenimiento, disminuir la cantidad de conductores por flota y prevención en los autobuses con características asociadas a un mayor riesgo de falla, como la tecnología EURO IV, la tipología BUSETA y los centros de operación identificados. Dado que la variable "cantidad_mantenimientos" resultó significativa en el modelo y se asocia con un aumento en la probabilidad de fallas, se recomienda reevaluar la estrategia de mantenimiento preventivo.
Análisis de la calidad de los mantenimientos: Investigar si los mantenimientos preventivos se están realizando de manera adecuada y si están solucionando efectivamente los problemas potenciales.
Revisión de los intervalos de mantenimiento: Evaluar si los intervalos de mantenimiento actuales son óptimos para cada tipo de vehículo y tecnología. Es posible que algunos modelos requieren mantenimientos más frecuentes que otros.
Personalización del mantenimiento: Considerar la posibilidad de implementar un enfoque de mantenimiento preventivo más personalizado, basado en el historial de fallas y las características específicas de cada vehículo.
Capacitación del personal: Asegurar que el personal de mantenimiento esté debidamente capacitado y cuente con las herramientas y recursos necesarios para realizar un trabajo de calidad.
Uso de tecnología: Explorar el uso de tecnologías de monitoreo y diagnóstico que permitan detectar fallas potenciales antes de que se conviertan en problemas mayores.
Investigación adicional: Es importante investigar las razones detrás de la mayor probabilidad de falla en ciertas tecnologías, tipologías y centros de operación. Esto podría revelar problemas específicos que pueden ser abordados para mejorar la fiabilidad de la flota.
Considerar otras variables: Aunque el modelo actual identifica algunas variables importantes, es posible que existan otros factores relevantes que no se han incluido en el análisis. Se recomienda explorar la posibilidad de incorporar nuevas variables para mejorar la precisión del modelo.
Evaluación continua: El modelo debe ser monitoreado y reevaluado periódicamente para garantizar su relevancia y precisión a medida que cambian las condiciones de operación y la composición de la flota.
Este análisis proporciona una visión inicial de los factores que influyen en las fallas mecánicas de los autobuses. Sin embargo, es importante recordar que se trata de un modelo simplificado y que la realidad puede ser más compleja. Se recomienda utilizar este análisis como punto de partida para una investigación más profunda y una toma de decisiones informada en la gestión de la flota.
Estas conclusiones y recomendaciones ofrecen una hoja de ruta para futuras mejoras y optimizaciones del modelo de predicción de averías, garantizando su utilidad y precisión en la gestión de la flota de autobuses de Consorcio Express.

10. Bibliografía

Aggarwal, C. C. (2015). Data Mining: The Textbook. Springer. https://link.springer.com/book/10.1007/978-3-319-14142-8

Brownlee, J. (2019). Machine Learning Mastery With Python. Machine Learning Mastery. https://machinelearningmastery.com/machine-learning-with-python/

Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media. https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/

Daniel Peña Sánchez de Rivera. (1989). Estadística Modelos y métodos 2. Modelos lineales y series temporales.

de Ian H. Witten, Eibe Frank y Mark A. Hall (2016). Data Mining: Practical Machine Learning Tools and Techniques".

Icesi, apuntes de clases maestría Ciencia de datos.

McKinney, W. (2017). Python for Data Analysis. O'Reilly Media. Retrieved from https://www.oreilly.com/library/view/python-for-data/9781491957653/

Roberto Behar G. Mario Yepes. (1996). Estadística con un enfoque descriptivo.

SAS (2011). Applied Analytics Using SAS Enterprise Guide. Course Notes.

SAS (2012). Statistics 1: Introduction to ANOVA, Regression and Logistic Regression. Course Notes.

SAS (2012). Statistics 2: ANOVA and Regression. Course Notes.

Stanley I. Grossman (2001). Algebra lineal, quinta edición

Razonpublica.com.
https://razonpublica.com/los-nuevos-buses-de-transmilenio-no-tan-buenosni-tan-malos/

TransMilenio S.A. (2023). Datos Abiertos. https://datosabiertos-transmilenio.hub.arcgis.com/

TransMilenio S.A. (2023). Manual de Indicadores de Calidad del Servicio.
https://www.transmilenio.gov.co/

Zhou, Z. H. (2021). Machine Learning. Springer. Retrieved from https://www.springer.com/gp/book/9789811519827

11. Anexo

CARTA DE AVAL

Señores
UNIVERSIDAD ICESI
Maestría en Ciencia de Datos
Cali – Colombia

En mi calidad de Gerente General de la empresa CONSORCIO EXPRESS SAS, con NIT 900.365.740-3, ubicada en la ciudad de Bogotá - Colombia, doy mi aval para que el funcionario Laureano Romero presente ante la Universidad Icesi. en la Maestría en Ciencia de datos el anteproyecto denominado como "IMPLEMENTAR UN SISTEMA DE PREDICCIÓN DE AVERÍAS EN UNA FLOTA DE AUTOBUSES MEDIANTE EL USO DE MACHINE LEARNING, CON EL FIN DE ANTICIPAR Y MITIGAR LOS COSTOS ASOCIADOS A LOS VEHÍCULOS FUERA DE SERVICIO, OPTIMIZANDO ASÍ LA GESTIÓN DE LA FLOTA Y MEJORANDO SU RENTABILIDAD EN EL LARGO PLAZO", se implementará un proyecto para desarrollar un modelo que detecte de manera temprana las fallas que provocan anomalías y causan la salida de servicio de los buses. Por tanto, estoy presto a colaborar con la información requerida para este fin académico.

Este aval se da únicamente con fines académicos y estará sujeto a las condiciones de confidencialidad de la información que se maneja dentro del proyecto.

La presente se expide a solicitud del interesado, el día 12 de junio de 2024.

Atentamente,

Arturo Fernando Rojas Rojas
CC. 79.643.884 de Bogotá
Gerente General
Consorcio Express SAS