PREDICCIÓN DE FALLAS PREMATURAS DE COMPONENTES EN UNA FLOTA DE CAMIONES MINEROS UTILIZANDO CIENCIA DE DATOS

Christian Andrés Martínez Morales
Código 9.015.395
Juan Camilo Perdomo Olarte
Código 9.014.270

Proyecto Aplicado para optar al título de
Magister en Ciencia de Datos

Directora Isabel Cristina García Arboleda

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, NOVIEMBRE 4 DE 2025

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

INTRODUCCIÓN

Una flota de camiones había estado experimentando fallas de manera prematura en algunos de sus componentes dentro de los cuales se encuentran suspensiones frontales, cilindros de dirección, ruedas, entre otros, afectando la disponibilidad, productividad y aumentando los costos de reparación. Los resultados de análisis de fallas realizados por la fábrica de estos camiones concluían que estos eventos no van relacionados a un problema de producto sino a una alta severidad de la operación, es decir, a condiciones en el uso de los camiones que atentan con el desempeño de una vida útil esperada. La fábrica centraba sus argumentos en un índice de severidad patentado cuya formulación es oculta y en el que la minera ha obtenido valores sobre los límites establecidos, por lo cual, la fábrica contemplaba este hecho como explicativo y suficiente para que se generen este tipo de fallas, sin embargo, para la minera estas conclusiones eran insuficientes y desconocía cuáles aspectos eran los que debía controlar. Al ser eventos que impactaban la seguridad, disponibilidad de los equipos y la producción, la minera buscaba alguna forma de predecir este tipo de fallas en función de variables a las cuales pueda realizarle seguimiento y control.

El proyecto planteó una solución para la problemática presentada bajo un enfoque en ciencia de datos a través del cual se buscó predecir las fallas de los componentes en función de variables a las que se le pudiera realizar seguimiento y control con el fin de que la minera pudiera anticipar las fallas, brindando así un insumo para la realización de planes de acción preventivos que apuntaran a la reducción de estos eventos.

La manera en cómo se abordó el problema inició con el análisis de la información disponible en una base de datos privada alimentada por los módulos de control de los camiones y donde se relacionan valores como carga acarreada, distancias recorridas, velocidades máximas, torque, cargas suspendidas máximas, microfallas por torsión y suspensión, entre otras; de igual manera, se analizó la información de fallas y reemplazo de componentes proveniente de los ingenieros de soporte que la fábrica de los camiones emplea para la mantención de la flota con el fin de definir aquellos componentes a los que se les realizarán modelos predictivos, posteriormente, se utilizaron diferentes técnicas supervisadas y no supervisadas tales como análisis descriptivo multivariado, análisis de conglomerados, XGBoost, regresión logística y perceptrones multicapa para la generación de modelos que fueron evaluados a través de métricas de rendimiento como la precisión y f1-score para seleccionar los modelos más adecuados de cada componente y modo de falla. Finalmente, se propuso un tablero de seguimiento y control para que la minera pudiera gestionar las fallas prematuras predichas.

Fig. 1 Metodología realizada en el proyecto. Fuente: Autores

El presente proyecto acotó su desarrollo, resultados y conclusiones sólo a la flota específica de camiones considerados y sólo para la operación en la mina seleccionada ya que cada marca de camión proporciona información y variables diferentes; de igual manera, el performance de estos equipos es dependiente de los perfiles de ruta, topología y operación específica de cada mina donde operan.

Los beneficios en la ejecución de este proyecto se centraron en obtener una operación enfocada a la seguridad, al cuidado del activo y en el mantenimiento predictivo abarcando varios grupos de interés donde se destacan el departamento de producción, mantenimiento, logística, e incluso, la fábrica de los camiones en cuestión, pues no contaban con un estudio similar que se hubiera reportado para estos equipos.

1. DEFINICIÓN DEL PROBLEMA

1.1. PLANTEAMIENTO DEL PROBLEMA

Una flota de camiones que opera en una mina de carbón en Colombia había estado experimentando fallas de manera prematura en algunos de sus componentes tales como suspensiones frontales, cilindros de dirección, ruedas, entre otros, afectando la disponibilidad, productividad y aumentando los costos de reparación. La fábrica de estos camiones provee mensualmente un indicador con el cual clasifica la severidad de la operación en baja, media o alta y determina unos valores límites dentro de los cuales se espera un desempeño de la vida útil de sus componentes de un 100%, así, a mayor nivel de severidad, la vida útil de los componentes se reducirá, mientras que a un menor nivel de severidad esta aumentará; además, debido a que el indicador es una patente de fábrica, su formulación y las variables de las que depende son desconocidas, por lo que un valor de severidad alto, podría explicar y advertir fallas prematuras pero de manera insuficiente.

La minera generalmente había mantenido un indicador de severidad medio/alto en su operación, por lo que la fábrica argumentaba que las fallas experimentadas no correspondían a un problema de producto sino más bien a un abuso de estos por la naturaleza severa de la operación. Bajo este contexto, la minera manifestó la necesidad de predecir este tipo de fallas en función de variables a las que pudiera realizarle seguimiento y control con el fin no solo de anticiparlas, sino también de realizar planes de acción preventivos que impacten a la disminución de estos eventos, sin embargo, se enfrentaba al desafío de manejar gran volumen de datos y múltiples variables, como registros de mantenimiento, datos de sensores y el indicador de severidad para entender mejor las causas subyacentes de estas fallas. Con el fin de tomar decisiones estratégicas basadas en los datos y utilizando técnicas de Ciencia de Datos, este análisis buscaba mejorar la disponibilidad y productividad de los camiones, así como reducir los costos de reparación asociados con las fallas prematuras, lo cual se alinea con las metas, los objetivos y las iniciativas de la minera.

1.2. FORMULACIÓN DEL PROBLEMA

Solucionar el problema planteado implicó responder principalmente el siguiente interrogante: ¿Cómo se pueden predecir las fallas prematuras de componentes de la flota de camiones mineros?

De manera sistemática, también implicó dar respuesta a los siguientes interrogantes: ¿Cuáles componentes y modos de falla pueden ser predichos?, ¿Cómo se desarrollarán los modelos predictivos de falla para los componentes y modos de falla identificados?, ¿Cómo se evaluarán los modelos desarrollados?, ¿Qué gráficas son adecuadas para el seguimiento y control de las fallas prematuras predichas?

2. OBJETIVOS DEL PROYECTO

2.1. OBJETIVO GENERAL

Predecir fallas prematuras de componentes de la flota de camiones mineros con el fin de anticiparlas y brindar un insumo con el que la minera pueda realizar planes de acción preventivos.

2.2. OBJETIVOS ESPECÍFICOS

2.2.1. Analizar los datos disponibles del desempeño de la flota e historial de fallos prematuros para identificar los componentes y modos de falla a los que se pueda predecir fallas prematuras.

2.2.2. Desarrollar modelos predictivos de falla para los componentes y modos de falla identificados a partir del análisis realizado.

2.2.3. Evaluar los modelos desarrollados a través de métricas de rendimiento para establecer límites de las variables predictoras a partir de los cuales se presentará una falla.

2.2.4. Desarrollar un prototipo de interfaz gráfica para el seguimiento y control de las fallas prematuras.

2.3. RESULTADOS ALCANZADOS

2.3.1. Listado de componentes y modos de falla a los que sea posible predecir fallas prematuras.

2.3.2. Modelos predictivos por componente y modo de falla de fallas prematuras.

2.3.3. Análisis y comparación de modelos desarrollados para cada componente y modo de falla.

2.3.4. Tablero de seguimiento y control de las fallas prematuras.

3. MARCO DE REFERENCIA

3.1. MARCO TEÓRICO

En esta sección se presentan los conceptos relacionados con el contexto del proyecto, así como de las técnicas de análisis y Ciencia de Datos utilizadas en la aplicación.

3.1.1. Ciclo de acarreo: Es el ciclo que realiza el camión y que determina su función en la mina, la cual es transportar el material de un punto a otro. Generalmente se divide en 5 etapas como lo son: Cargue, transporte, descarga, retorno y espera/demora [1].

La etapa de cargue, es aquella en la que se ingresa el material a transportar en la tolva del camión bien sea por una pala, cargador o mediante otro mecanismo como bandas transportadoras; la etapa de transporte es aquella en la que el camión se desplaza cargado hacia el punto de destino; la descarga es aquella etapa en el que el camión levanta la tolva para depositar el material en el punto destino que puede ser un botadero, retro llenado, trituradora, etc; la etapa de retorno es aquella en la que el camión vuelve al sitio del cargue para iniciar el ciclo nuevamente; finalmente, la espera/demora es aquella etapa donde se suman los tiempos muertos del ciclo ya sea por imprevistos o por la espera del turno para el cargue o descargue.

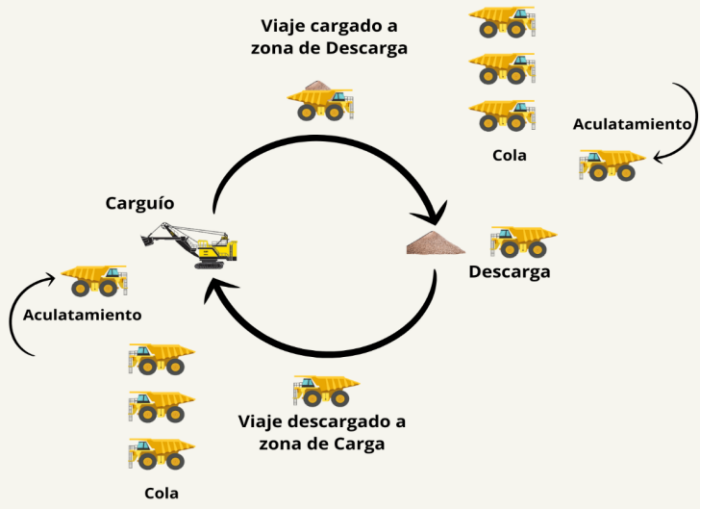


Fig. 2 Ciclo de acarreo de camiones mineros [2]

3.1.2. Esfuerzo de flexión: Es aquel que resulta al aplicar fuerzas perpendiculares al eje principal de un elemento que se encuentra apoyado en ciertos puntos. Este tipo de esfuerzo tiende a doblar el elemento sometiéndolo así a compresión en la parte cóncava y tensión en la parte convexa [3].

3.1.3. Esfuerzo de torsión: Es aquel que resulta al aplicar fuerzas a un elemento en direcciones contrarias y que tienden a girarlo [3].

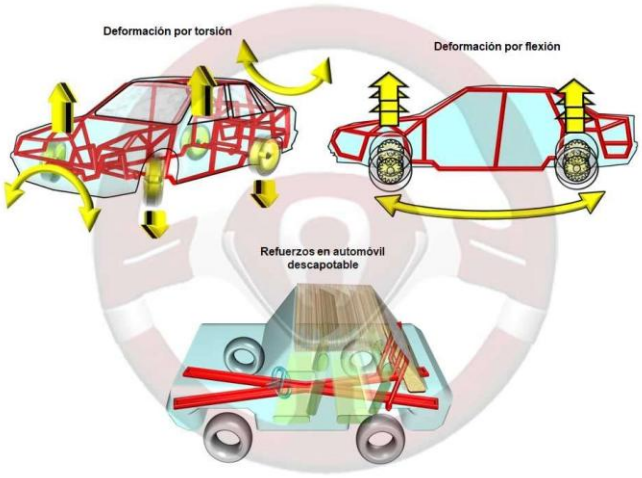


Fig. 3 Deformaciones por flexión y Torsión en el chasis de un vehículo [4]

3.1.4.Componentes mayores de un camión minero: Son aquellos que poseen grandes dimensiones tales como: Suspensiones frontales y traseras, ruedas delanteras, ruedas motorizadas, paquetes de frenos, chasis, nose cone, housing axle, cilindros y brazos de dirección, cilindros de levante, módulo de potencia, reservorios y tolva.

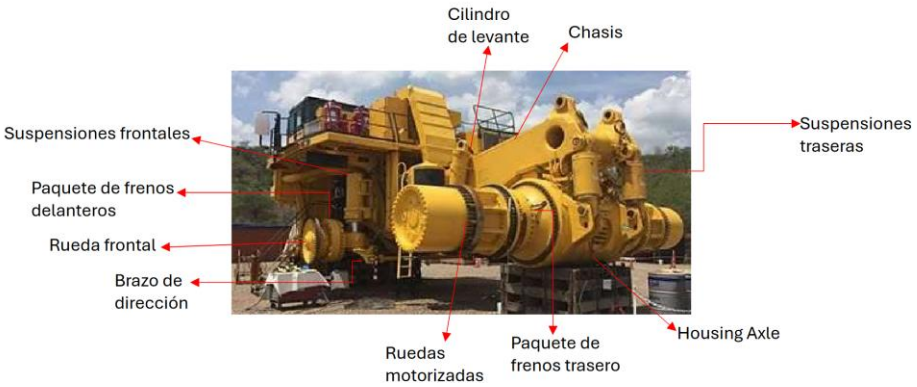


Fig. 4 Componentes mayores del camión minero [5]

3.1.5.Sistema de pesaje del camión minero: El sistema de pesaje del camión minero es aquel que calcula la carga acarreada por el equipo y otras variables de interés con el fin de realizar un transporte de material eficiente, de igual manera, almacena y provee información útil a partir de la cual se conoce el desempeño en producción del equipo a través de variables como velocidades máximas, picos de torsión experimentados por el chasis, picos de cargas máximas medidas, microfallas por torsión y flexión las cuales se utilizan para el cálculo del indicador de severidad.

El sistema, consta de sensores de presión en las suspensiones que envía una señal análoga al módulo quien a su vez recibe información de otros sensores para interpretar las señales y

almacenar información de interés donde cada registro almacenado corresponde a un ciclo de acarreo.

Las variables y unidades que proporciona el módulo se presentan en la siguiente tabla.

Tabla 1 Variables proporcionadas por el módulo de pesaje

Variables proporcionadas por el módulo de pesaje		
Variable	**Unidad**	**Tipo de variable**
Fecha y hora	dd/mm/yy hh:mm:ss	Cuantitativa
Pases de pala	Número entero	Cuantitativa
Banderas de error	Adimensional	Categórica
Carga remanente	Tons	Cuantitativa
Tiempo de ciclo	hh:mm:ss	Cuantitativa
Distancia recorrida con carga	km	Cuantitativa
Distancia recorrida sin carga	km	Cuantitativa
Velocidad máxima con carga	km/h	Cuantitativa
Velocidad máxima sin carga	km/h	Cuantitativa
Máximo torque	Tons*m	Cuantitativa
Máxima carga medida	Tons	Cuantitativa
TKPH de las ruedas frontales y traseras	Tons*km/h	Cuantitativa
Microfalla por torsión	Adimensional	Cuantitativa
Microfalla por carga suspendida	Adimensional	Cuantitativa
Peso en vacío	Tons	Cuantitativa
Carga bruta	Tons	Cuantitativa
Consumo de combustible por ciclo	Litros	Cuantitativa

3.1.6.Microfalla: Variable adimensional que relaciona el daño causado debido a los esfuerzos de torsión y carga suspendida del ciclo de acarreo ejecutado. Esta variable se relaciona con la regla de Palmgren-Miner de la sumatoria de la relación de ciclos $\sum \frac{n_i}{N_i}$ donde n_i representa el número

de ciclos que generaron un esfuerzo específico y N_i representa el número de ciclos a los que fallaría un componente a dicho nivel de esfuerzo específico. Cuando la sumatoria alcanza un valor de 1, se origina la falla por fatiga [6].

3.1.7. Análisis de Componentes Principales: Es una herramienta Estadística cuyo objetivo es reducir el número de variables. Las nuevas variables o componentes principales (los cuales son independientes de cada uno) son una transformación lineal de las variables iniciales u originales y una cantidad pequeña de componentes. Se eliminan los componentes que son menos explicativos o que presentan pérdidas mínimas de información [7].

3.1.8.Análisis de Correspondencia: El Análisis de Correspondencia es una técnica la cual, permite la representación de las categorías que poseen dos o más variables cualitativas en un espacio de reducidas dimensiones, de esta manera se agrupan las categorías dependiendo de las similitudes que tengan las variables representadas. El propósito es analizar las relaciones de dependencia entre variables categóricas, las cuales se muestran en tablas de contingencia [8].

3.1.9.Prueba Chi-cuadrado: La prueba chi-cuadrado hace parte de las técnicas estadísticas no paramétricas y se emplea con el objetivo de analizar la independencia entre dos variables categóricas. Esta prueba evalúa si son significativas las diferencias observadas entre las frecuencias esperadas de las categorías bajo una serie de supuestos estadísticos los cuales son: las frecuencias son absolutas y mutuamente excluyentes; otro supuesto es que los grupos de estudio son independientes lo que significa que no se aplica la prueba chi- cuadrado si los datos provienen de muestras pareadas o relacionadas; finalmente, las frecuencias esperadas son suficientes de tal manera que, al menos el 80% de los casos tienen valores esperados mayores o iguales a cinco y nunca menor a 1 [9].

La prueba Chi-Cuadrado se aplica al presente proyecto para analizar la relación de la falla con las variables categóricas y de esta manera determinar si algunas variables funcionan como variables predictoras. Se observó que en algunos casos no se cumplió el supuesto de frecuencias esperadas mayores o iguales a 5 y nunca menor a 1, sin embargo, al considerar los resultados arrojados, los modelos obtuvieron un rendimiento bueno.

3.1.10. Análisis no supervisado: El Análisis no supervisado es una técnica la cual usa algoritmos que logran identificar patrones, agrupaciones o estructuras en grupos de datos que no poseen categorías o etiquetas definidas previamente. Permite encontrar subgrupos naturales que muestran propiedades similares entre sí, y logra comprimir los datos a lo largo de los patrones encontrados para reducir la dimensionalidad del problema [10].

3.1.10.1. Análisis descriptivo multivariado (cluster): El análisis descriptivo multivariado (Análisis cluster) es una técnica de análisis exploratorio de los datos utilizada para resolver problemas de clasificación ordenando objetos en grupos o clusters por el grado de similitud

entre miembros del mismo grupo, este tipo de análisis permite descubrir asociaciones y estructuras en los datos que no son evidentes a primera vista. Se puede encontrar 2 tipos fundamentales de métodos de clasificación los cuales son jerárquicos y no jerárquicos; en los primeros, la clasificación resultante posee un número creciente de clases anidadas, mientras que en los segundos las clases no se presentan de forma anidada anidadas [11].

3.1.11. Análisis supervisado: El análisis supervisado en ciencia de datos permite el aprendizaje automático por medio de algoritmos que clasifican datos tomando como referencia un conjunto de datos de entrenamiento ya anteriormente etiquetado [12], explicado de otra manera, cada entrada de datos está relacionada con la salida correcta esperada por medio del aprendizaje automático del mapeo de entradas con salidas para lograr una predicción correcta. El proceso de entrenamiento conlleva la recopilación del conjunto de datos, la división del conjunto de datos, el entrenamiento del modelo y su evaluación permitiendo su uso para clasificar o predecir un valor numérico continuo. Dentro del análisis supervisado se encuentran varias técnicas como la regresión logística, la técnica de random Forest y la técnica de Support Vector Machine (SVM).

3.1.11.1. Regresión Logística: La regresión logística es una técnica de aprendizaje estadístico empleada para propósitos explicativos y predictivos. El propósito de su análisis consiste en la predicción de un valor, la predicción de probabilidad de ocurrencia de cierto evento, así como determinar qué variables pesan más para aumentar o disminuir el valor que se desea predecir [13]. Existen diferentes enfoques para el análisis de regresión logística basados en la variable dependiente, dentro de ellos se encuentran la regresión logística binaria (donde la variable dependiente sólo toma dos valores), regresión logística multinomial (cuando la variable dependiente presenta un número finito de posibles resultados) y regresión logística ordinal (cuando la salida se representa en rangos en lugar de valores) [14].

3.1.11.2. Random Forest: La técnica de Random Forest, la cual consiste en un método no paramétrico que funciona haciendo particiones sucesivas buscando valores de variables que maximizan la homogeneidad de las particiones resultantes. Funciona construyendo árboles de decisión múltiples, cada árbol es entrenado con una muestra aleatoria y en cada nodo es seleccionado de manera aleatoria un subconjunto de variables para determinar la división más adecuada. El Random Forest es un algoritmo del aprendizaje supervisado que tiene la capacidad de manejar grandes conjuntos de datos y también variables para proporcionar resultados robustos y precisos [15].

3.1.11.3. Support Vector Machine (SVM): Las máquinas de vectores soporte son una metodología para clasificación binaria que se utiliza para separar datos en dos clases mediante un hiperplano de separación. Este hiperplano maximiza el margen entre los datos más cercanos de cada clase que se llamarán vectores soportes. El propósito del Support Vector Machine es encontrar el hiperplano óptimo equidistante a las muestras que se encuentren más cercanas

de ambas clases. Existen ciertas variantes las cuales con SVM margen duro que funciona para datos linealmente separables y SVM margen blando que aplica cuando los datos no son perfectamente separables [16].

3.1.11.4. XGBoost: El algoritmo denominado XGBoost es considerado como el más avanzado modelo de los modelos que implican el uso de arboles de decisión. Es una técnica de aprendizaje supervizado, a modo de contexto la evolución ha sido de la siguiente manera: Arboles de decisión, seguido de Bagging, seguido de Random Forest, seguido de Boosting, seguido de Gradient Boosting, seguido de XGBoost. El algoritmo consiste en la secuencia de árboles de decisión donde los árboles aprenden del resultado de los árboles anteriores y corrigen los errores producidos por ellos hasta llegar a un límite donde no hay posibilidad de corregir más dicho error [17].

3.1.11.5. Perceptrón Multicapa: El Perceptrón Multicapa es un modelo utilizado para la resolución de problemas de clasificación de regresión, al probar ser un aproximador universal de funciones. Se caracteriza por su organización en capas de celdas disjuntas que vita conexiones auto recurrentes ya que ninguna salida neuronal crea una entrada para las neuronas de la misma capa o capas anteriores [18].

3.1.12. Métricas de evaluación: Son medidas cuantitativas usadas para la evaluación del desempeño de los modelos de aprendizaje automático permitiendo comparar modelos entre sí [19]. Dentro de estas métricas se encuentran:

3.1.12.1. Exactitud o Accuracy: Es la proporción de instancias correctamente clasificadas y el total de observaciones, su rango es de 0 a 1 indicando que entre mayor sea su valor, mejor será el rendimiento del modelo, sin embargo, la precisión puede no ser apropiada para medir el rendimiento de modelos en escenarios de desbalance.

$$Exactitud = \frac{VP + VN}{VP + VN + FP + FN}$$

3.1.12.2. Sensibilidad o Recall: Mide la capacidad del modelo de predecir un valor positivo VP sobre el total de positivos del dataset. Su rango va de 0 a 1 y tomará relevancia en escenarios donde sea importante no perder casos positivos; un valor alto significa entonces que el modelo predice correctamente los verdaderos positivos, pero sin considerar la proporción de falsos positivos.

$$Sensibilidad = \frac{VP}{VP + FN}$$

3.1.12.3. Precisión: Mide la capacidad del modelo de predecir un valor positivo sobre el total de predicciones positivas del modelo, es decir, de cierta forma mide la calidad de los positivos del modelo.

$$Precision = \frac{VP}{VP + FP}$$

Su rango va de 0 a 1 y cobrará relevancia en escenarios donde sea importante reducir los falsos positivos; un modelo que tenga un valor alto de precisión predecirá menos falsos positivos, pero sin considerar la proporción de falsos negativos.

3.1.12.4. F1 – Score: Es la media armónica entre la precisión y la sensibilidad con el fin buscar un equilibrio entre estas, será útil en escenarios de desbalance con el fin de mitigar las consecuencias de obtener Falsos positivos o Falsos Negativos.

$$F_1 = 2\frac{Precision * Sensibilidad}{Precision + Sensibilidad}$$

Las métricas de evaluación para el desarrollo de este proyecto se eligieron considerando las consecuencias de los Falsos Positivos y Falsos Negativos, pues predecir una falla de manera errónea (Falso Positivo) aumentará los costos de operación de los equipos, causando indisponibilidad y tiempos muertos; por el contrario, no predecir una falla (Falso Negativo) puede asemejarse a la situación problema descrita en este proyecto, afectación a la disponibilidad, productividad y aumentando los costos de reparación debido a cambios necesarios no planeados.

3.2. ANTECEDENTES

Es preciso investigar casos de estudio los cuales compartan similitud con el objetivo del proyecto a ejecutar, esto para conocer qué se ha explorado, qué métodos, heurísticas o paso a paso se han empleado para cumplir con el objetivo plantado, qué modelos fueron útiles a la hora de buscar la solución a un problema determinado y cuáles fueron los resultados. También podremos conocer en qué medida el proyecto aplicado es innovador o en que sectores ha tenido implementación. Se presentan a continuación 5 casos aplicados que sirven de referencia y comprenden objetivos similares a la predicción de fallas en componentes.

3.2.1. Propuesta de modelo de predicción de fallos para componente crítico de camión minero, utilizando machine learning: El estudio se enfoca en el uso de un modelo de Machine Learning para lograr predecir fallos en motores diésel de vehículos mineros, empleando herramientas como la regresión logística, SVM y Random Forest. Este caso refleja datos no lineales lo cual es común en indicadores de severidad, por medio de múltiples pruebas se concluyó que la regresión logística resultó ser la adecuada para predicciones basadas en datos categóricos, logrando una precisión del 77.5%, mientras que con el Random Forest se alcanzó un RMSE de 76.91 [20]. Las diferencias con el proyecto a presentar radican principalmente en algunas técnicas adicionales tales como el cálculo con base Weibul y también algunos conceptos como

la creación de datos aleatorios. Ahora bien, en cuanto a similitudes podemos encontrar un cuadro comparativo de modelos, ya que de hecho uno de los objetivos es justamente la evaluación de las diferentes técnicas a realizar. También se puede considerar como una similitud las medidas de calidad que el proyecto de Lastarroa y Karem presentan, ya que se pretede establecer unos límites de confianza para sugerir, por ejemplo, un cambio de componentes.

Esta propuesta resulta provechosa para el proyecto que se pretende realizar ya que relaciona varias técnicas en ciencia de datos con el contexto al que pertenece este proyecto, por lo que se cuenta con información que permita contrastar los resultados teniendo en consideración que la diferencia radica en los componentes que se abordan.

3.2.2. Métodos de pronóstico de fallas en motores diésel de camiones mineros en base a indicadores de degradación probabilísticos: El estudio propone una metodología para lograr estimar el tiempo de falla en motores empleando información en tiempo real por medio del uso de métodos probabilísticos y técnicas no supervisadas. En la sección titulada “Discusión” se menciona que se logró desarrollar un modelo con un margen de error promedio de 20 días para predicciones de 40 días y un error de máximo 40 días para predecir 120 días. El caso muestra un enfoque en indicadores específicos para modelar la degradación. Como resultados, se logró una predicción precisa con un margen reducido de error en las simulaciones permitiendo establecer zonas de riesgo de fallar. Se empleó un modelo basado en las cadenas de Markov para lograr una proyección probabilística [21]. En cuanto a diferencias el antecedente muestra el uso de DBSCAN que para el caso a desarrollar no se aplica. Ahora bien en cuanto a similitudes nuevamente se evidencia el propósito de usar las técnicas en el contexto minero, también la exploración de reducciones de dimensionalidad que se ven reflejadas en el presente proyecto por medio del PCA.

Resulta útil tener en cuenta este estudio para el desarrollo del proyecto ya que brinda otra herramienta estadística que podría ser utilizada, de forma similar al antecedente expuesto anteriormente, la diferencia radica en el alcance de los componentes que se desea abordar.

3.2.3. Framework para la detección anticipada de fallas de equipos mineros mediante el uso de Machine Learning: En este estudio se propone la creación de un framework para lograr la detección anticipada de fallas, que puede ser una guía para la estructuración del modelo predictivo que busca realizar el proyecto. El framework probó ser útil al momento de reducir costos de mantenimiento y tiempos inactivos, este framework logró utilizar un conjunto de 140.000 registros históricos durante un periodo de 4 años con un error inferior a 7 días, lo cual resultó un plazo adecuado para planear los mantenimientos. Este caso refuerza la importancia de integrar datos en los modelos de predicción y su impacto económico [22]. En cuanto a diferencias importantes a resaltar, claramente se evidencia que los equipo a analizar son distintos, ya que en este caso no estamos hablando de camiones mineros, sin embargo, la idea del diseño de un framework sirve como iniciativa.

Este estudio es considerado porque, además de brindar información del uso de las herramientas en ciencia de datos en el contexto común y la predicción de fallas, involucra un enfoque de reducción de costos y tiempos inactivos que podría ser considerados para la extensión del proyecto y estudios posteriores diferenciándose por el hecho de que el proyecto propuesto tiene un enfoque de control en la operación del equipo minero según se definió en el alcance.

3.2.4. Procesamiento de datos en el pronóstico de fallos de rodamientos para el mantenimiento predictivo: Este estudio tiene el propósito de mejorar el rendimiento de los modelos de clasificación y regresión mediante un preprocesamiento de datos. Se menciona que se obtuvo un 96% de Accuracy en algoritmo SVM, mientras que con los atos en crudo solo llegaba a un 26% como máximo, incrementó en un 74.4% el tiempo para la planificación del mantenimiento y también se pudo determinar que con el método propuesto se genera un ahorro de 5,356,000 de dólares en costos por mantenimiento. Este proyecto difiere en varios aspectos del proyecto de predicción de fallas en camiones mineros ya que la temática es distina (ya que está enfocada a rodamientos), que utiliza modelos como K-Means, redes bayesianas y Fuzz Logic y también que se enfoca más en la preparación de los datos o almenos, ese es uno e sus principales enfoques, sin embargo se asemeja en el objetivo de predecir defectos y poder tener cierta anticipación para la toma de decisiones. Esta, de hecho, fue la razón por la cual seconsideró valioso dicho estudio para el proyecto de predicción de fallas en amiones mineros, y e la misma manera, comparar algoritmos que sí se proponen para el proyectopara probar y explorar como árboles de dcisón y regresión logística [23]

3.2.5. Pronóstica de fallas en redes de distribución de agua potable haciendo uso de herramientas Machine Learning: En este estudio se propone predecir el estado de las redes de distribución de agua potable, mediante el pronóstico del estado estructural haciendo uso de machine learning. Se logró la construcción de un mapa de distribución espacial del deterioro de redes, indicando con el, zonas de mayor vulnerabilidad. En cuanto a diferencias con el proyecto de predicción de fallas en camions mineros identificamos en primera instancia que la temática es distinta ya que está enfocado a otro sector (sistemas de distribución de agua) pero también tiene varias similiudes pese a ello; por ejemplo el uso de los modelos XGBoost y perceptrón multicapa, también, se diseña algo muy similar a una interfaz gráfica, con la cual se mapea el sistema de distribución de agua, en el proyecto de pedicción de fallas en vehíclos mineros también se propone el diseño de una interfaz. Este estudio fue considerado debido a que muestra el uso de modelos e interés para el proyecto con camiones mineros tales como XGBoost y perceptrones, también aporta e el sentido de que el tema central es predecir el estado de un objeto [24].

4. ANÁLISIS DE LOS DATOS DE DESEMPEÑO DE LA FLOTA E HISTORIAL DE FALLOS PREMATUROS

4.1. CREACION DE BASE DE DATOS

La data proporcionada se componía originalmente de varios archivos dentro de los cuales se encuentran archivos de Access, archivos de Excel, tablas de archivos pdf y archivos .csv. Para el desarrollo del proyecto, se consolidó toda esta información utilizando el gestor de base de datos de Oracle donde se encuentran 5 tablas cuya descripción se presenta a continuación.

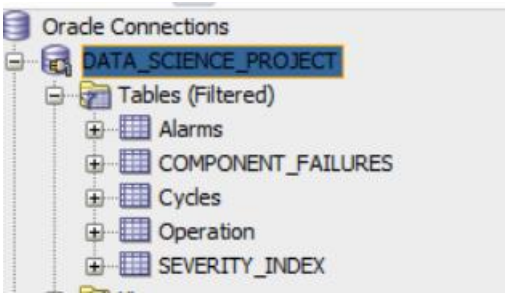


Fig. 5 Tablas generadas en Oracle. Fuente: Autores

Tabla 2 Descripción de las fuentes de datos

Nombre de Tabla	Descripción	Dimensión
Alarms	Contiene información de las alarmas de Sprung y Torque por camión junto con su severidad. Se presume que, a mayor cantidad de alarmas y mayor severidad, habrá más fallas prematuras por camión.	107.663 x 8
Component_Failures	Contiene el registro histórico de falla de componentes con información de modos de falla, horas trabajadas, camión donde estuvo instalado, así como fechas de instalación y remoción.	557 x 16
Cycles	Contiene información de la operación de los camiones almacenadas en variables como distancias recorridas, velocidades máximas, esfuerzos de torsión,	1.022.328 x 50

	esfuerzos de flexión y microfallas.	
Operation	Contiene información de la operación de los camiones donde se relacionan las cargas realizadas y los lugares donde se realizaron dichas cargas	338.920 x 6
Severity_Index	Contiene información del índice de severidad por camión mes a mes. Se presume que, a mayor índice de severidad, mayor número de fallas prematuras	959 x 6

4.2. ANALISIS EXPLORATORIO DE DATOS POR DATASET

Se realizó un análisis exploratorio de datos para cada dataset con el fin de verificar tipos de datos, datos erróneos, atípicos y realizar la limpieza de datos.

4.2.1. Component Failures

Este es el dataset de hechos, es decir, la información contenida en los demás datasets se relaciona con la información contenida en este. Este dataset no contiene valores nulos.

	TRUCK_ID	COMPONENT	LOCATION	TYPE	SMR_INSTALLATION	SMR_REMOVED	DATE_INSTALLATION	DATE_REMOVE	CUM_SMR	MODIFICATION	HOURS_WORKED	TOTAL_HOURS	REMOVAL_FAILURE	FAILURE_MODE	FAILED	INSTALLED
1	C109	ALTERNATOR	NA	OVH	21605	35295	31-AUG-22	11-DEC-24	0	ESTANDAR	13690	13690	UNEXPECTED	Coupling Mi...	Yes	No
2	C109	ALTERNATOR	NA	OVH	35295	36514	11-DEC-24	28-FEB-25	0	ESTANDAR	1219	1219	NA	NA	No	Yes
3	C109	BLOWER	LH	ORIGINAL	0	36514	16-NOV-18	28-FEB-25	0	ESTANDAR	36514	36514	NA	NA	No	Yes
4	C109	BLOWER	RH	ORIGINAL	0	36514	16-NOV-18	28-FEB-25	0	ESTANDAR	36514	36514	NA	NA	No	Yes
5	C109	ENGINE	NA	ORIGINAL	0	21605	16-NOV-18	31-AUG-22	0	ESTANDAR	21605	21605	OVH	NA	No	No

Fig. 6 Dataset Component Failures. Fuente: Autores

Los registros por camión de componentes que han fallado y los que no se presentan a continuación.

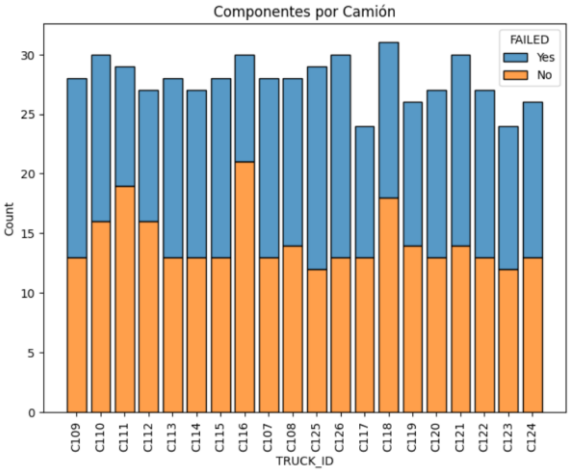


Fig. 7 Registros por Camión. Fuente: Autores

Los componentes contemplados en el Proyecto y el número de registros en el dataset se presentan a continuación.

Tabla 3 Cantidad de Componentes en Component Failures

Componente	Cantidad
Suspensiones Frontales	102
Cilindros de levante de tolva	61
Suspensiones traseras	60
Cilindros de dirección	165
Ruedas frontales	99
Ruedas motorizadas o traseras	70

Las fallas de los componentes enunciados en la tabla anterior se presentan en la Fig. 8.

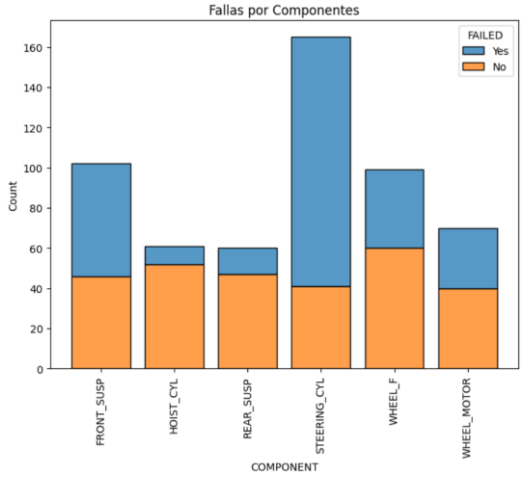


Fig. 8 Fallas por Componente. Fuente: Autores

Se observa que los componentes que más fallas han presentado son los cilindros de dirección, seguidos de suspensiones frontales, mientras que los que menos fallan son cilindros de levante (Hoist_Cyl) y suspensiones traseras (Rear_Susp).

Los componentes por camión en el dataset se presentan en la Fig. 9.

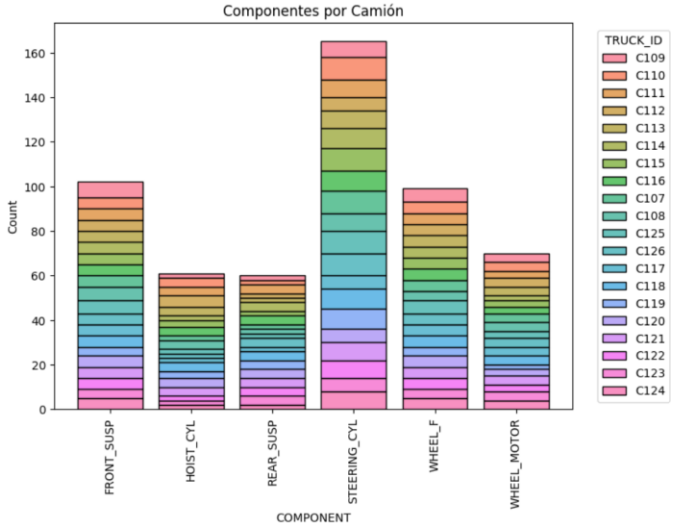


Fig. 9 Componentes por Camión. Fuente: Autores

En la Fig. 10 Componentes que han fallado por Camión, se presentan exclusivamente los componentes que han fallado por camión.

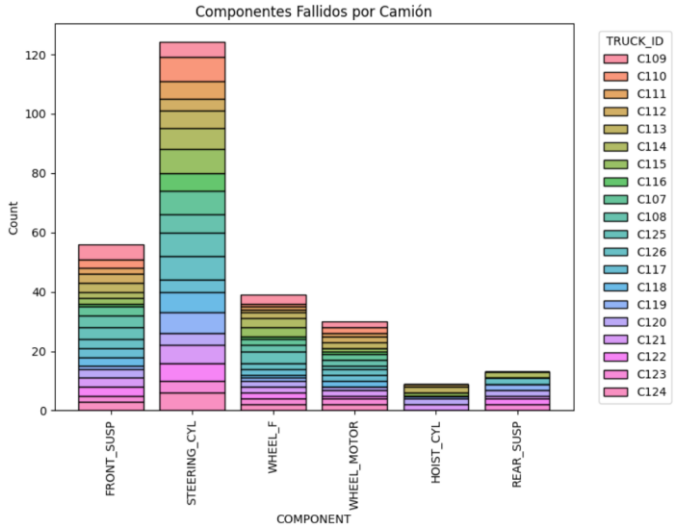


Fig. 10 Componentes que han fallado por Camión. Fuente: Autores

Se observa que los componentes que más han fallado de manera prematura son las suspensiones frontales junto con los cilindros de dirección, seguidos de las ruedas frontales y las ruedas motorizadas, mientras que los que menos han presentado fallas prematuras son los cilindros de levante y suspensiones traseras.

De igual manera se observa que las fallas por camión se comportan de manera similar, exceptuando el C109 que tiene un mayor número de fallas en suspensiones frontales respecto al resto de camiones. En las ruedas frontales (Wheel_F) se destaca el camión C125 con mayor número de fallas.

Cada componente, ha sufrido modificaciones buscando extender la vida útil, por lo que la relación de cada modificación por componente en el dataset se presenta en la Fig. 11.

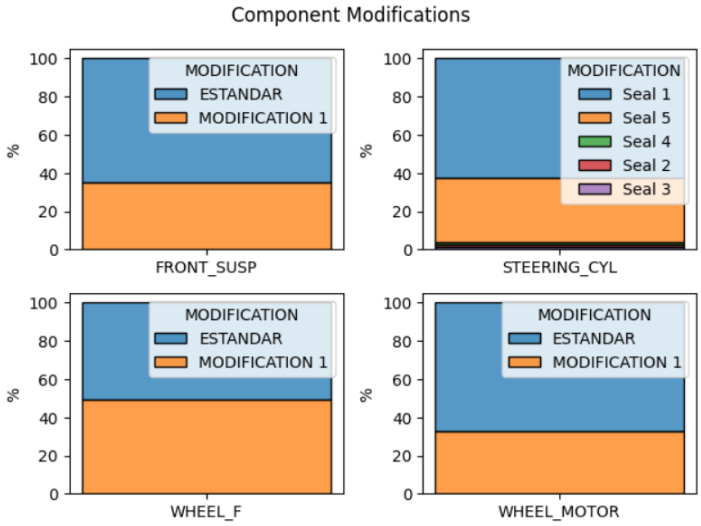


Fig. 11 Modificaciones por Componentes. Fuente: Autores

La relación de fallos por modificación por componentes se presenta a continuación.

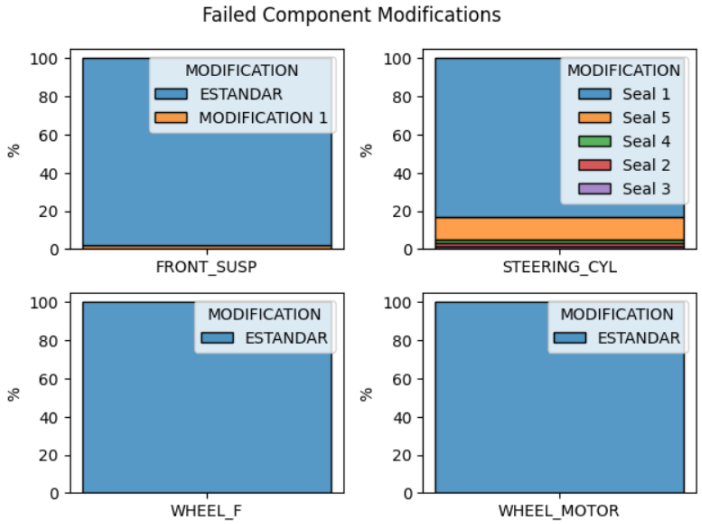


Fig. 12 Fallas por modificación por Componente. Fuente: Autores

Se observa que los componentes sin modificación (estándar) son los que presentan un mayor número de fallas, por lo que se podría pensar que las modificaciones realizadas han sido efectivas ya que el conteo de fallas es menor.

El desempeño en horas alcanzadas por modificación por componente se resume en los siguientes diagramas de cajas y violines.

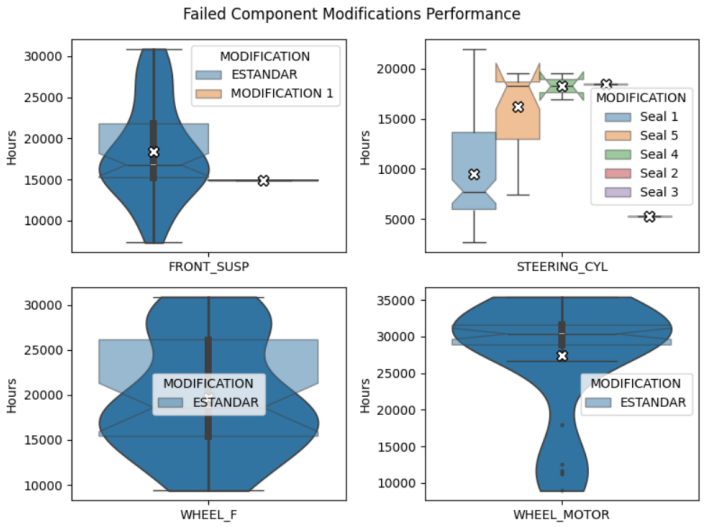


Fig. 13 Desempeño por Modificación por Componente. Fuente: Autores

Se observa que hay muy pocos datos de la modificación 1 en las suspensiones frontales para compararlo con la modificación estándar. Respecto a los cilindros de dirección, se observa que hay una diferencia significativa en el desempeño por cada modificación donde el mejor desempeño lo

alcanzó la modificación 4 y 2, sin embargo, hay muy pocos datos en comparación con la modificación 1 y 2. Finalmente, se observa el desempeño de las ruedas frontales y las ruedas motorizadas con la modificación estándar ya que con la modificación 1, ningún componente ha fallado.

El balance del dataset para la columna FAILED, la cual menciona si el componente falló o no de manera prematura, se presenta a continuación.

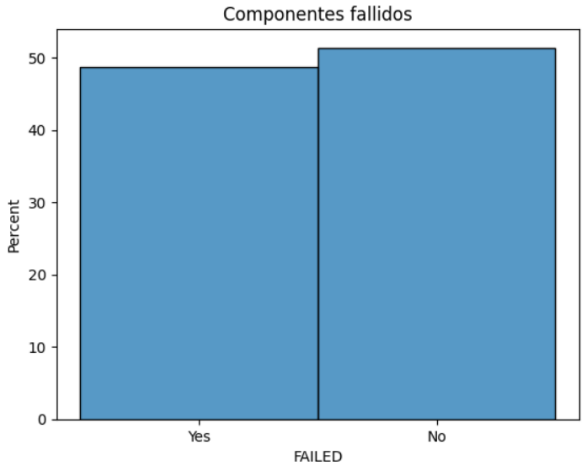


Fig. 14 Balance del dataset Component Failures. Fuente: Autores

Se observa que a nivel general el dataset se encuentra balanceado, sin embargo, es necesario tener en cuenta el balanceo por componente y por modificación a la hora de desarrollar los modelos.

4.2.2. Cycles

En este dataset se contiene la información histórica por camión y por fecha de variables con las cuales se pretende realizar la predicción de fallas prematuras. Naturalmente, es un dataset de alta dimensión, sin embargo, previamente se seleccionaron las variables que se presume podrían ser predictoras de fallas prematuras basado en la opinión de un experto, la nueva dimensión del dataset es de 1.022.328 x 18.

Las variables y valores nulos del dataset se presentan a continuación.

```
Truck_ID                    0
Date_Cycle                  0
Status Flag            963103
L-Haul Distance             0
E-Haul Distance             0
L-Max Speed                 0
E MaxSpeed                  0
Max Pos T                   0
Max Neg T                   0
Max Sprung                  0
LF TKPH                     0
RF TKPH                     0
R TKPH                      0
MicroFail Torque            0
MicroFail SprungWeight      0
Maneuvering T_Avg           0
FinalZone T_Avg             0
Gross Payload               0
```

Fig. 15 Valores nulos del dataset Cycles. Fuente: Autores

Se observan valores nulos en la columna Status Flag, esta columna contiene información de alertas durante los ciclos que podrían indicar imprecisiones en la data obtenida, por lo que funcionará como un filtro para descartar aquellos registros erróneos, además, un valor nulo significa que el registro no tuvo ningún tipo de alerta y por ende no sería necesaria una imputación.

Luego de hacer una limpieza inicial utilizando la información de la columna Status Flag, la dimensión del dataset es de 980.427 x 18.

Tabla 4 Resumen Cycles

	Date	L-Haul Distance	E-Haul Distance	L-Max Speed	E MaxSpeed	Max Pos T	Max Neg T	Max Sprung	LF TKPH	RF TKPH	R TKPH	MicroFail Torque	MicroFail SprungWeight	Maneuvering T_Avg	FinalZone T_Avg	Gross Payload
count	980427	980427	980427	980427	980427	980427	980427	980427	980427	980427	980427	980427	980427	980427	980427	980427
mean	41:41,6	4,35	4,72	35,77	42,22	90,91	93,63	632,91	853,2	850,11	748,59	1,18	1,06	-2,34	-1,52	302,7
min	26/10/2018	0,12	0	6	0	0	11,5	220	4	5	4	0	0	-92,67	-84,54	17,8
25%	24/07/2020	3,31	3,33	31,6	39,2	77,52	80,05	588,9	671	668	580	0,62	0,51	-12,14	-9,34	293,4
50%	31/03/2022	4,34	4,45	36	43,3	89,75	92,4	623,8	885	880	787	0,98	0,85	-2,51	-1,51	303,3
75%	30/09/2023	5,42	5,65	40,6	47,6	102,96	105,83	666,1	1054	1049	936	1,5	1,35	7,29	5,9	313,6
max	28/02/2025	66,8	81,03	102,2	67	417,24	403,73	1392,4	2277	2284	2399	45,57	33,19	202,57	120,27	401,9
std		1,73	2,88	7,28	7,23	19,63	20,18	64,54	274,88	274,63	250,19	0,82	0,87	14,82	11,48	16,29

Los registros por camión en el dataset se presentan a continuación.

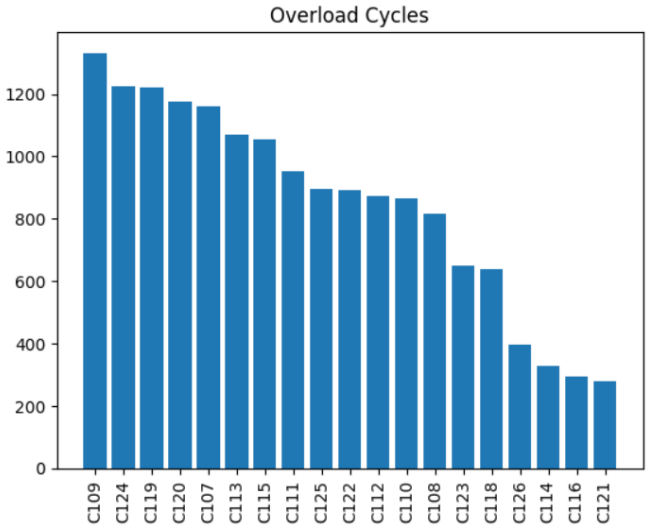


Fig. 16 Registros por Camión en Cycles. Fuente: Autores

A continuación, se presentan diagramas de caja para las variables numéricas del dataset.

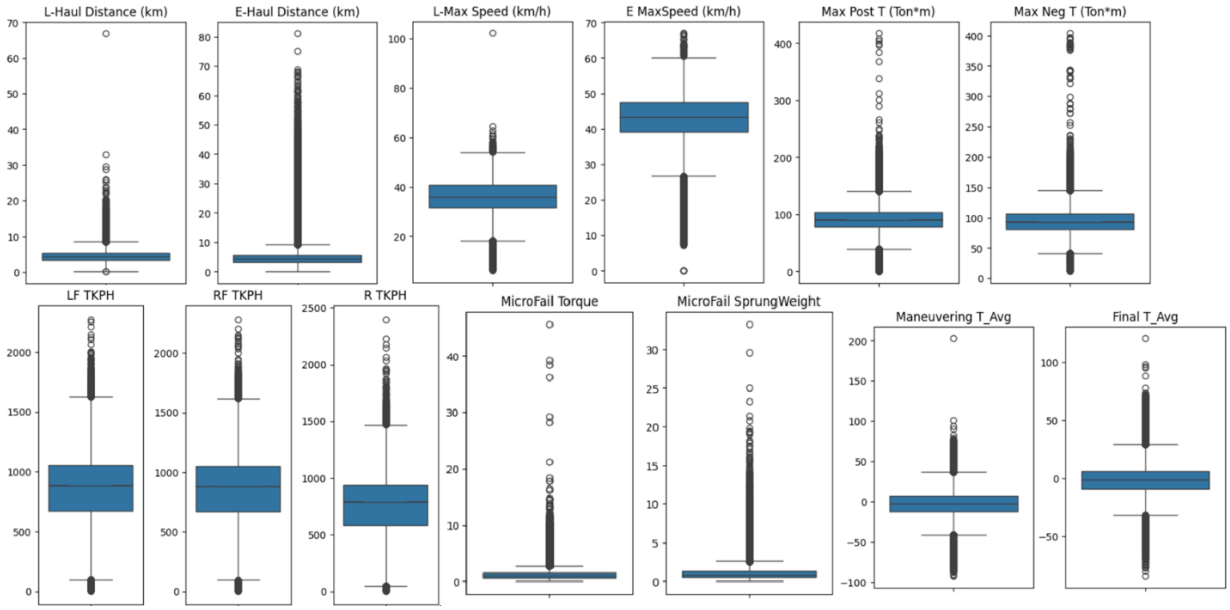


Fig. 17 Diagramas de caja dataset Cycles. Fuente: Autores

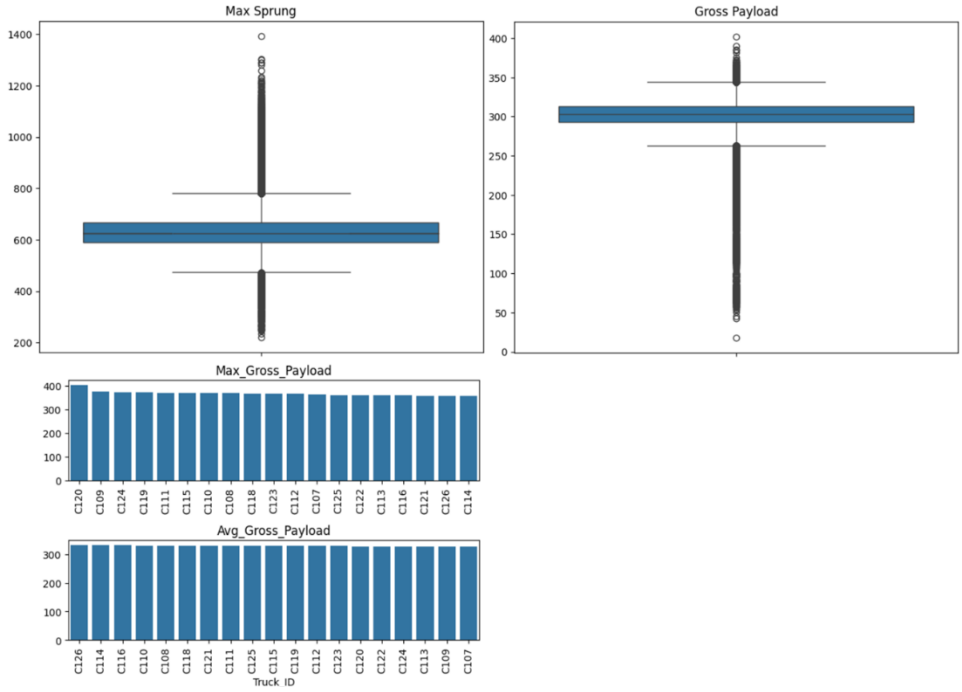


Fig. 18 Diagramas de caja dataset Cycles 2. Fuente: Autores

Se observa un comportamiento compacto en la mayoría de las variables con puntos muy extremos, lo que podría ser indicio de valores erróneos y/o atípicos.

4.3. LIMPIEZA DE DATOS ATIPICOS POR DATASET

Como se mencionó anteriormente, los dos datasets que se utilizarán para la generación de modelos son Component_Failures y Cycles. Respecto al primer dataset no hay presencia de datos atípicos ya que todos estos registros corresponden a componentes que fallaron y de los que se lleva registro descartando la presencia de valores erróneos; respecto al segundo, sí se detectaron valores erróneos por lo que a este dataset sí se le realizó una limpieza.

Considerando que el dataset tiene altas dimensiones, que tal caso puede generar ruido y que algunos atributos pueden ser irrelevantes para la detección de anormalidades [25] se eligió realizar un análisis univariado por cada atributo.

La técnica seleccionada fue Modified Z-Scores [26] la cual consiste en valores Z-scores que no se calculan con la media ni la desviación estándar por ser estimadores no robustos, sino con la mediana de la desviación absoluta de la mediana (Median Absolute Deviation about the Median, MAD) ya que se considera un estimador robusto cuyo punto de ruptura (brakedown point) es de aproximadamente el 50%, lo que implicaría que el estimador puede tolerar un 50% de valores errados sin arrojar un resultado incorrecto, además, ha sido probada satisfactoriamente en distribuciones pseudo-normales. Se considerará un valor como Outlier si el valor absoluto del Z-score modificado es mayor a 3.5.

Los límites para la determinación de datos atípicos por atributo se presentan en la siguiente tabla.

Tabla 5 Definición de datos atípicos dataset Cycles

Atributo	Limites según Modified Z-Scores (Superior – Inferior)	Opinión de Expertos
L-Haul Distance	9.96 – 0	Ok
E-Haul Distance	10.55 – 0	Ok
L-Max Speed	59.67 – 11.93	Ok
E MaxSpeed	66.03 – 20.37	60 – 20.37
Max Pos T	155.79 – 23.42	Ok
Max Neg T	159.22 – 25.34	Ok
Max Sprung	821.14 – 423.66	1400 – 423.66
Maneuvering T_Avg	47.90 – -52.88	Ok
FinalZone T_Avg	38.04 – -40.94	Ok
Gross Payload	359.54 – 247.46	Ok
LF TKPH	1868.10 – 0	Ok
RF TKPH	1857.92 – 0	Ok
R TKPH	1701.65 – 0	Ok
MicroFail Torque	3.10 – 0	Ok
MicroFail SprungWeight	2.88 – 0	Ok

Luego de calcular los límites con la técnica seleccionada, se consultó con los expertos con el fin de contrastar la información. En la mayoría de los atributos estuvieron de acuerdo, sin embargo, los límites de velocidad máxima fueron modificados ya que los camiones están configurados para no superar los 60km/h y, de igual manera, los límites de Sprung Weight fueron modificados ya que sí es posible que se registren valores por encima de 821 Toneladas.

Una vez realizada la limpieza de datos, el dataset se redujo a un 84% pasando de tener 1022328 de filas a tener 863772.

4.4. ANALISIS MULTIVARIADO DE DATOS

Como el dataset Component_Failures contiene el intervalo de fechas en las que los componentes estuvieron instalados, se combinó la información con el dataset Cycles el cual contiene la información operativa de los equipos durante dicho intervalo de fechas. Teniendo en cuenta que cada registro del dataset Cycles corresponde a un ciclo de acarreo ejecutado por un camión, a cada registro de Component_Failures le corresponderán múltiples registros de Cycles.

Si se observa, por ejemplo, el primer registro del Component_Failures mostrado en la Fig. 19:

	TRUCK_ID	COMPONENT	LOCATION	TYPE	SMR_INSTALLATION	SMR_REMOVED	DATE_INSTALLATION	DATE_REMOVE
7	C109	FRONT_SUSP	LH	ORIGINAL	0	13624	16-NOV-18	14-APR-21
8	C109	FRONT_SUSP	LH	NEW	13624	24458	14-APR-21	10-MAR-23
9	C109	FRONT_SUSP	LH	OVH	24458	36514	10-MAR-23	28-FEB-25

Fig. 19 Registro de ejemplo dataset Component_Failures. Fuente: Autores

Este registro corresponde a una suspensión frontal que estuvo instalada desde el 16/11/2018 hasta el 14/04/2021 en el camión C109. Para obtener información de la operación del camión durante ese intervalo de fechas, en el dataset Cycles será necesario filtrar la información por el camión C109 desde el 16/11/2018 hasta el 14/04/2021, lo que da como resultado 19393 registros o ciclos.

Lo anterior genera la necesidad de resumir esos 19393 registros para poderlos relacionar con el registro del dataset Component_Failures en una misma línea. Para ello, se decidió resumir los atributos de Cycles utilizando el valor máximo, la mediana y la sumatoria con el fin de añadirlos a los registros de Component_Failures, esto da resultado a un nuevo dataset llamado Combined_Component_Failures cuya dimensión es de 490 x 61 mostrado a continuación.

TRUCK_ID	COMPONENT	LOCATION	TYPE	SMR_INSTALLATION	SMR_REMOVED	DATE_INSTALLATION	DATE_REMOVE	CUM_SMR	MODIFICATION	...	Median_Gross Payload	Max_Gross Payload	Sum_Gross Payload
C109	FRONT_SUSP	LH	ORIGINAL	0.0	13624.0	2018-11-16	2021-04-14	0.0	ESTANDAR	...	297.200012	359.500000	4.861412e+06
C109	FRONT_SUSP	LH	NEW	13624.0	24458.0	2021-04-14	2023-03-10	0.0	ESTANDAR	...	303.500000	357.700012	3.997429e+06
C109	FRONT_SUSP	LH	OVH	24458.0	36514.0	2023-03-10	2025-02-28	0.0	MODIFICATION 1	...	312.100006	356.299988	4.508596e+06
C109	FRONT_SUSP	RH	ORIGINAL	0.0	15181.0	2018-11-16	2021-07-28	0.0	ESTANDAR	...	299.399994	359.500000	5.457539e+06
C109	FRONT_SUSP	RH	REINSTALLED	15181.0	18169.0	2021-07-28	2022-02-03	13358.0	ESTANDAR	...	300.399994	357.700012	1.093760e+06

Fig. 20 Dataset Combined_Component_Failures. Fuente: Autores

Posteriormente, al filtrar el dataset anterior por cada componente considerado en el proyecto (suspensiones frontales, cilindros de levante, suspensiones traseras, cilindros de dirección, ruedas frontales y ruedas motorizadas), se determinó cuáles de estos 61 atributos tenía incidencia en la falla bajo los siguientes criterios:

- Para atributos cuantitativos: Se utilizaron diagramas de cajas muescados en los que en el eje X se relaciona si el componente falló o no y en el eje Y el atributo cuantitativo. Si las muescas (que representan el intervalo de confianza del 95% para la mediana) se traslapan, indicará que no hay diferencia significativa de la mediana del atributo cuantitativo evaluado entre los componentes que fallaron y los que no, por lo que no se consideraría dicho atributo como predictor de la falla; si las muescas no se traslapan, habrá una diferencia significativa entre la mediana del atributo de los componentes que fallaron y los que no, por lo que sí se considerará dicho atributo como predictor de la falla.

  En la Fig. 21, por ejemplo, se evalúa si la mediana, la máxima y la sumatoria de la máxima velocidad sin carga vs la falla de suspensiones frontales.

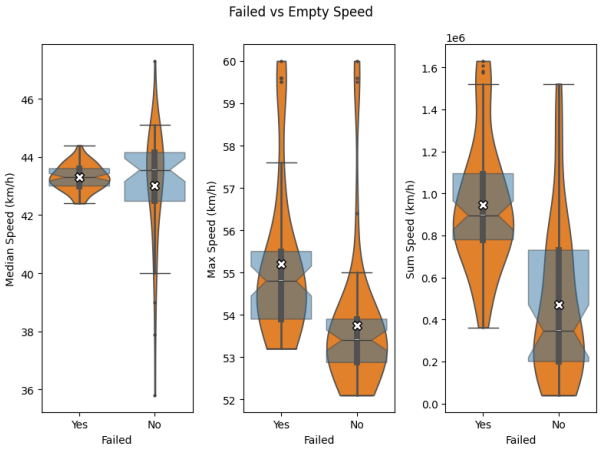


Fig. 21 Análisis Bivariado para determinar variables predictoras. Fuente: Autores

En este caso se determina que la Mediana de la velocidad máxima (primera gráfica de izquierda a derecha) no será considerada como variable predictora de la falla de suspensiones ya que las muescas se traslapan, mientras que los valores máximos y la sumatoria de este atributo sí lo serán, ya que las muescas de sus cajas no se traslapan indicando que hay una diferencia significativa en los valores de las suspensiones que sí fallaron y las que no.

- Para atributos categóricos: Se utilizaron pruebas de chi-cuadrado para determinar la dependencia entre atributos categóricos y la falla, para ello se realizaron inicialmente tablas de contingencia. En el siguiente ejemplo se observan las tablas de contingencia de la falla de suspensiones traseras con los atributos Location y Type.

LOCATION	LH	RH
FAILED		
No	24	23
Yes	6	7

TYPE	NEW	ORIGINAL	OVH
FAILED			
No	4	27	16
Yes	0	13	0

```
FAILED and Location: Independent
FAILED and Type: Dependent
```

Fig. 22 Pruebas Chi-cuadrado de independencia de variables. Fuente: Autores

En este caso la prueba arroja que la falla de suspensiones traseras es independiente de la ubicación donde esté instalada (si es derecha o izquierda) y dependiente de si es nueva, original o reparada.

Las variables predictoras para cada componente se podrán encontrar en el Anexo A, las cuales fueron obtenidas utilizando la metodología anteriormente explicada. Posteriormente se generó un dataset por componente el cual contuviera exclusivamente las variables predictoras del componente correspondiente.

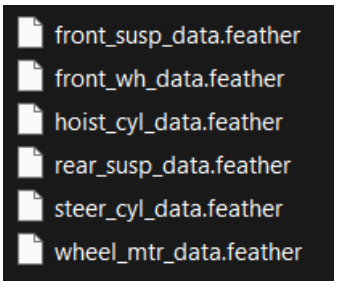


Fig. 23 Datasets generados por componente. Fuente: Autores

4.4.1. Análisis de Componentes Principales

Para cada uno de los datasets creados, se realizó un análisis de correlación y de factor de inflación de la varianza de las variables predictoras con el fin de evitar los errores que la multicolinealidad pudiera generar. En la Fig. 24 se observa la matriz de correlación del dataset de suspensiones frontales, donde se aprecia correlaciones fuertes entre variables predictoras.

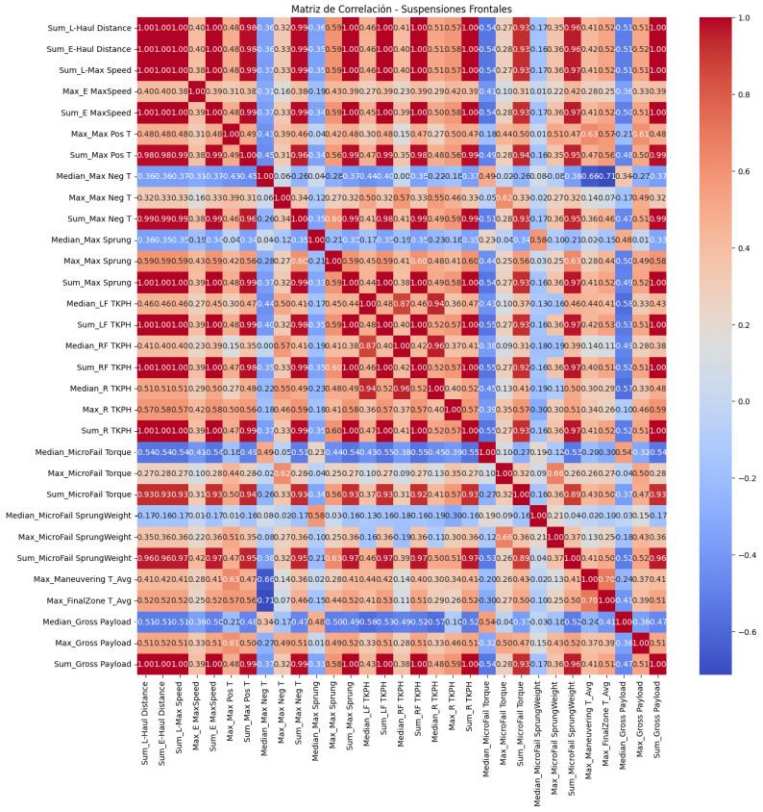


Fig. 24 Matriz de Correlación del dataset suspensiones frontales. Fuente: Autores

En cuanto al factor de inflación de la varianza (VIF), se muestran los 5 mayores valores en la Tabla 6

Tabla 6 VIF Suspensiones Frontales

Variable	VIF
Sum_Max Sprung	31987
Sum_RF TKPH	30725
Sum_Gross Payload	28895
Sum_LF TKPH	27844
Sum_R TKPH	22267

De manera similar, la mayoría de datasets por componente presentaban correlaciones fuertes entre las variables predictoras de falla, las matrices pueden observarse en el Anexo B.

Para solventar esta condición, se realizó un análisis de componentes principales (posterior al escalamiento de las variables) donde el número de componentes seleccionado fue aquel cuya varianza explicada acumulada superara o igualara el 95% de la varianza total. En el siguiente ejemplo, se observa que, para las suspensiones frontales, si se seleccionan las primeras 10 componentes principales, se explicará alrededor del 95% de la varianza.

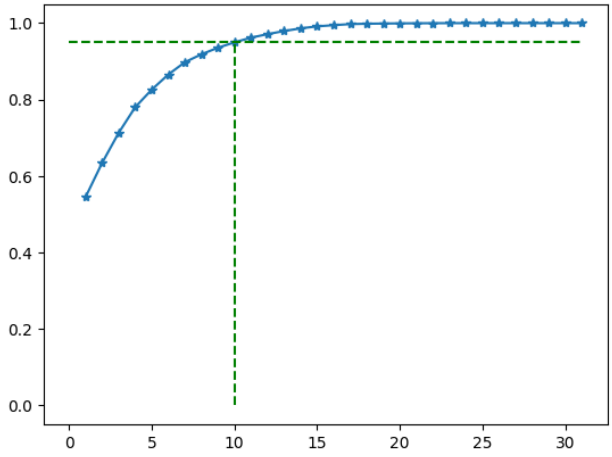


Fig. 25 PCA para Suspensiones Frontales. Fuente: Autores

Realizado lo anterior, se ejecutó una matriz de correlación junto con un cálculo del VIF para descartar multicolinealidad; en cada una de las componentes el valor VIF obtenido fue igual a 1.00

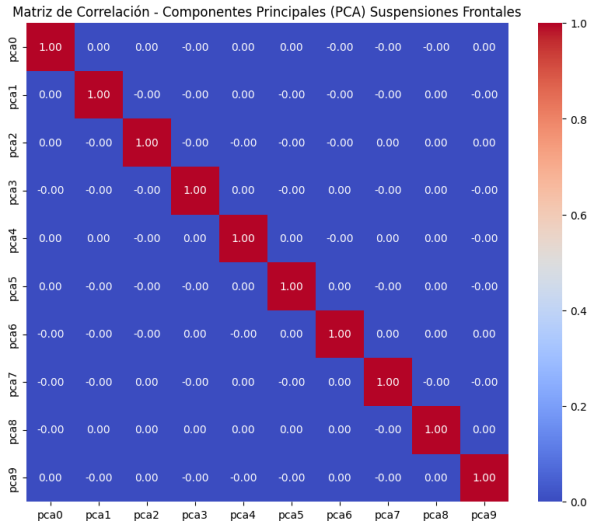


Fig. 26 Matriz de correlación en el dataset de Suspensiones Frontales. Fuente: Autores

De esta manera se procedió con todos los datasets por componente.

A sabiendas de que generar modelos utilizando componentes principales puede afectar la interpretabilidad de estos, se solicitó a los ingenieros de soporte de fábrica las variables que, a su criterio, pudieran ser removidas de los datasets por componente con el fin de eliminar las correlaciones fuertes sin necesidad de realizar el análisis de componentes principales. El resultado de este ejercicio se puede observar en el Anexo C.

A partir de la información provista por los ingenieros, se crearon otros datasets con el fin de contrastar el desempeño de los modelos generados sobre los datos cuando se aplica el análisis de componentes principales vs cuando no se aplica.

4.4.2. Análisis Cluster

Posteriormente, se exploró la opción de realizar un análisis cluster no jerárquico utilizando el criterio de maximización del índice de Silhouette para la determinación del número de conglomerados. Para algunos componentes como las suspensiones frontales, el valor del índice de Silhouette arrojó que agrupar en más de un conglomerado no sería provechoso, ya que su valor máximo se alcanza cuando hay uno solo.

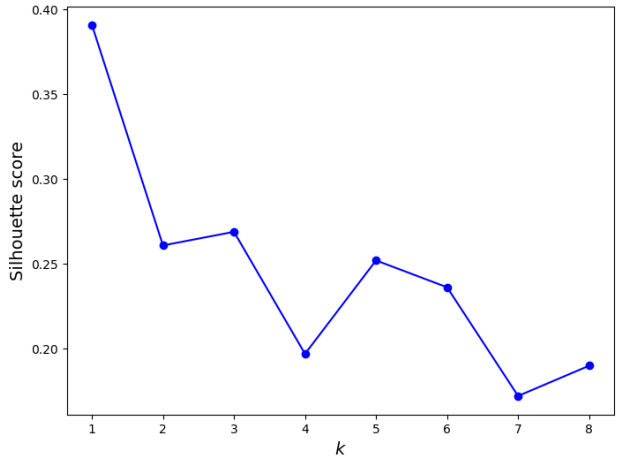


Fig. 27 Optimización del indice de Silhouette dataset Suspensiones Frontales. Fuente: Autores

Sin embargo, para otros componentes como cilindros de levante, sí resultaba provechoso realizar el análisis ya que cuando hay 2 conglomerados, se obtiene el mayor valor de Silhouette.

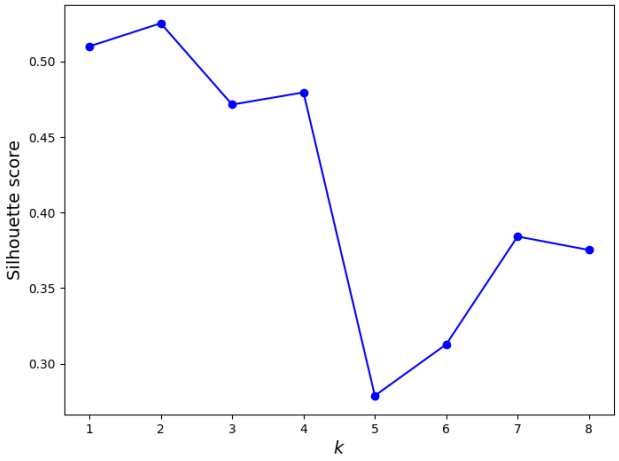


Fig. 28 Optimización del indice de Silhouette dataset Cilindros de Levante. Fuente: Autores

Al realizar el análisis cluster, se observa que los cilindros que pertenecen al grupo 1 no se relacionan con la falla, mientras que aquellos que pertenecen al grupo 0 sí han fallado.

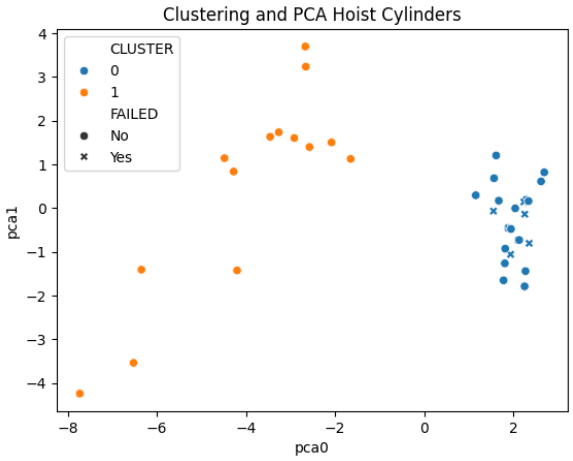


Fig. 29 Análisis Cluster Cilindros de Levante. Fuente: Autores

El resto de los resultados de análisis cluster se presenta en el Anexo D.

4.4.3. Análisis de Correspondencia

Finalmente, para explorar las relaciones entre atributos categóricos, se realizó un análisis de correspondencia múltiple. Para componentes como las suspensiones frontales, por ejemplo, se observa que el tipo OVH (reparadas) con modificación 1 se relacionan con la no falla; mientras que el tipo Original sin modificación (Estandar), se relaciona con la falla.

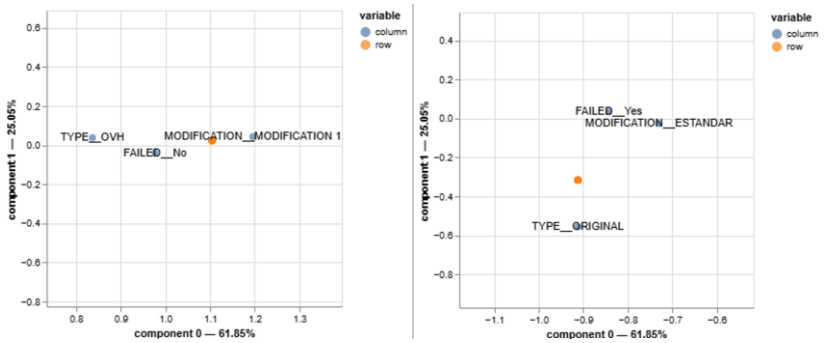


Fig. 30 Análisis de Correspondencia en Suspensiones Frontales. Fuente: Autores

Los resultados de todos los componentes se detallan en el Anexo D.

4.5. LISTADO DE COMPONENTES CONSIDERADOS

Se consideraron los componentes mayores (conocidos así por sus dimensiones y funcionamiento) que habían fallado prematuramente en un mínimo de 7 ocasiones, se consideró el valor de 7 ya que correspondía al 30% del total de camiones de la flota.

1. Suspensiones frontales
2. Front Wheels
3. Cilindros de levante

4. Suspensiones traseras
5. Cilindros de dirección
6. Wheel motors

5. DESARROLLO DE MODELOS PREDICTIVOS PARA LOS COMPONENTES

Se desarrollaron modelos de clasificación supervisados utilizando 4 técnicas por componente las cuales fueron Random Forest, XGBoost, Perceptrones Multicapa y Regresión Logística. Estas técnicas fueron elegidas considerando que no todos los datasets presentan un comportamiento marcado entre los componentes que fallaron y los que no, es decir, no todos los datasets se pudieron separar en más de 1 cluster, por lo que se descartaron técnicas como el Neareast K-Neighbors (que depende directamente de la distancia de las observaciones) y el SVM (que depende de la separación de clases en un plano).

Considerando la versatilidad de los árboles de decisión, quienes no dependen de la separabilidad ni escalabilidad de los datos, se eligieron los modelos Random Forest y el modelo XGBoost, el cual, según la literatura, generalmente ofrece un mejor rendimiento que el primero; además, con esta técnica se aprovechó la ventaja de conocer la importancia que cada atributo aporta a la predicción. Por otro lado, se consideraron los perceptrones multicapa teniendo en cuenta su versatilidad y gran aplicabilidad independientemente de la complejidad del dataset. Finalmente, como uno de los entregables del proyecto fue proponer límites para las variables predictoras, se consideró la regresión logística para buscar conseguir el objetivo a partir del análisis de los odds.

En cuanto al esquema de validación utilizado para la sintonización de hiperparámetros, se utilizó una búsqueda por grilla con validación cruzada realizando una división de los datasets del 64% para entrenamiento, 16% para validación y 20% para prueba. En el XGBoost la división fue del 60% entrenamiento, 20% validación y 20% de prueba ya que la programación de esta técnica permite seleccionar el subset donde se validará la sintonización.

Respecto al preprocesamiento de los datos, se utilizó la transformación en componentes principales realizadas sobre los datasets presentadas en el capítulo anterior, así como los datasets creados a partir de la experiencia de los ingenieros de fábrica, con el fin de contrastar el desempeño de los modelos generados cuando se aplica el análisis de componentes principales vs cuando no se aplica (ver Análisis de Componentes Principales); además se transformaron los atributos categóricos a través de la técnica One-hot encoding, en la cual, a pesar de aumentar la dimensionalidad, siempre se obtuvo un número de registros mayor al número de atributos.

De los atributos categóricos considerados, fue necesario excluir MODIFICATION debido a que presentaba un desbalance muy marcado en la mayoría de los componentes tal como lo muestra la Fig. 12, lo cual afectaba el rendimiento de los modelos.

De manera general, la grilla que se utilizó para la sintonización de hiperparámetros en técnica de Random Forest se presenta a continuación.

Tabla 7 Grilla para sintonización de hiperparámetros Random Forest

Parámetros	Rango
n_estimators	De 2 a 102 en intervalos de 20
criterion	gini, entropy, log_loss

max_depth	None y de 5 a 15 en intervalos de 2
min_samples_split	De 2 a 12 en intervalos de 2
min_samples_leaf	De 1 a 5
max_features	sqrt, log2 y None
max_leaf_nodes	None
min_impurity_decrease	0
bootstrap	True
oob_score	True
warm_star	True, False
class_weight	None
cc_alpha	De 0.0 a 0.03 en intervalos de 0.01
max_samples	None
monotonic_cst	None
Cv	De 2 a 7

La grilla que se utilizó para la sintonización de hiperparámetros en técnica del XGBoost se presenta a continuación.

Tabla 8 Grilla para sintonización de hiperparámetros XGBoost

Parámetros	Rango
n_estimators	De 1 a 20
max_depth	De 1 a 20
learning_rate	0.1, 1, 0.1
booster	gbtree, linear
gamma	De 0.1 a 0.5 en intervalos de 0.1
tree_method	auto, exact, approx
grow_policy	depthwise, lossguide
scale_pos_weight	None
Cv	5

La grilla que se utilizó para la sintonización de hiperparámetros en técnica del Perceptrón Multiple se presenta a continuación.

Tabla 9 Grilla para sintonización de hiperparámetros Perceptrón Múltiple

Parámetros	Rango
hidden_layer_sizes	Una capa de 1 a 3 neuronas y de 1 a 3 capas con 1 a 3 neuronas cada una
activation	logistic, tanh, relu
solver	lbfgs, adam, sgd

alpha	de 0.0001 a 0.5 en intervalos de 0.10
batch_size	auto
learning_rate	constant, invscaling, adaptive
learning_rate_init	De 0.001 a 0.12 en intervalos de 0.02
power_t	De 0.25 a 0.5 en intervalos de 0.75
shuffle	True
momentum	De 0.1 a 0.9 en intervalos de 0.3
nesterovs_momentum	True, False
early_stopping	True, False
validation_fraction	0.25
beta_1	0.5 y 0.9
beta_1	0.5 y 0.9
epsilon	1e-8
n_iter_no_change	10

La grilla que se utilizó para la sintonización de hiperparámetros en técnica de la regresión logística se presenta a continuación.

Tabla 10 Grilla para sintonización de hiperparámetros Perceptrón Múltiple

Parámetros	Rango
penalty	l2, none
dual	Falsee
C	De 0.1 a 1 en intervalos de 0.1
fit_intercept	True, False
intercept_scaling	0.2, 0.5, 0.7, 1 y 1.5
class_weight	None
solver	lbfgs, newton_cg, newton-cholesky, sag, saga
multi_class	auto

El rango de estos parámetros fue definido de acuerdo con la literatura encontrada y realizando pruebas previas sobre cada uno de los datasets.

En resumen, el rendimiento de los modelos generados sobre el subset de entrenamiento (80% de los datasets) se presenta a continuación.

Tabla 11 Resultados de entrenamiento de los modelos

		Random Forest						XGBoost						Multiple Perceptron						Logistic Regression					
Data set		PCA			Expert			PCA			Expert			PCA			PCA			PCA			Exprt_scaled		
Preprocesamiento		* Scaled * PCA *One hot			*One hot			* Scaled * PCA *One hot			*One hot			* Scaled * PCA *One hot * Solver LBFGS			* Scaled * PCA *One hot * Solver SGD or ADAM			* Scaled * PCA *One hot			*Scaled *One hot		
Componente	n	A	F1 (+)	F1 (-)	A	F1 (+)	F1 (-)	A	F1 (+)	F1 (-)	A	F1 (+)	F1 (-)	A	F1 (+)	F1 (-)	A	F1 (+)	F1 (-)	A	F1 (+)	F1 (-)	A	F1 (+)	F1 (-)
Susp front	76	.97	.98	.97	.97	.98	.97	.93	.94	.93	.95	.95	.94	.96	.96	.96	.93	.94	.93	.91	.91	.9	.89	.9	.89
Susp traseras	48				.98	.96	.99				.98	.96	.99	.98	.96	.99	.77	0	.87				.83	.73	.88
Cil levante	48	.98	.93	.99	.98	.93	.99	.92	.78	.95	.96	.88	.97	.85	0	.92	.85	0	.92	.69	.44	.78	.94	.82	.96
Wheel motors	52	1	1	1	.98	.98	.98	.96	.96	.96	.94	.94	.95	.92	.92	.93	.83	.81	.84	1	1	1	.56	0	.72
Cil direcc	90	1	1	1	1	1	1	.97	.97	.95	1	1	1	.97	.97	.95	.96	.96	.94	.97	.97	.95	.92	.94	.9
Front Wheels	76	.95	.94	.96	.96	.95	.97	.92	.91	.93	.91	.89	.92	.92	.91	.93	.96	.95	.97	.91	.89	.92	.92	.91	.93

En la Tabla 11, se presenta el rendimiento sobre el subset de entrenamiento de cada uno de los modelos realizados utilizando métricas como el accuracy (A), F1 score para la predicción positiva (F1 +) y negativa (F1 -). Se observa que en algunos componentes el rendimiento sobre el dataset que contiene variables transformadas (columna PCA) es el mismo sobre el que contiene variables seleccionadas por los ingenieros de fábrica (columna Exprt), mientras que en otros el mejor rendimiento varía entre un u otro dataset.

Se observa también que el mejor modelo para predecir las fallas en las suspensiones frontales fue el Random Forest sobre cualquiera de los dos datasets; en las suspensiones traseras fueron las técnicas Random Forest, XGBoost y el Perceptrón múltiple con el solver LBFGS alcanzando el mismo rendimiento; en los cilindros de levante fue el Random Forest sobre los dos datasets; en los Wheel motors fue tanto el Random Forest y la Regresión logística sobre el dataset PCA; en los cilindros de dirección el Random Forest sobre los dos datasets y el XGBoost sobre el dataset Exprt; en los Fronts Wheels el Random Forest sobre el dataset Exprt junto con el Perceptrón Múltiple con el solver Adam o SGD.

El detalle del tratamiento por componente y los hiperparámetros del mejor clasificador por técnica se presentan en el Anexo E.

6. EVALUACIÓN DE MODELOS PREDICTIVOS PARA LOS COMPONENTES

Una vez reentrenados los modelos con los parámetros de los mejores clasificadores y utilizando el 80% de los dataset, se evaluaron los mismos con el 20% de prueba, subset que no había sido utilizada para el entrenamiento ni la validación. En la siguiente tabla se presenta el rendimiento de los modelos en la fase de entrenamiento y prueba.

Tabla 12 Evaluación de Modelos Generados

			Random Forest						XGBoost						Multiple Perceptron						Logistic Regression					
Data set			PCA			Expert			PCA			Expert			PCA			PCA			PCA			Exprt_scaled		
Preprocesamiento			* Scaled * PCA *One hot			*One hot			* Scaled * PCA *One hot			*One hot			* Scaled * PCA *One hot * Solver LBFGS			* Scaled * PCA *One hot * Solver SGD or ADAM			* Scaled * PCA *One hot			*Scaled *One hot		
Component	Split	n	A	F1 (+)	F1 (-)	A	F1 (+)	F1 (-)	A	F1 (+)	F1 (-)	A	F1 (+)	F1 (-)	A	F1 (+)	F1 (-)	A	F1 (+)	F1 (-)	A	F1 (+)	F1 (-)	A	F1 (+)	F1 (-)
Susp Front	Train	76	.97	.98	.97	.97	.98	.97	.93	.94	.93	.95	.95	.94	.96	.96	.96	.93	.94	.93	.91	.91	.9	.89	.9	.89
	Test	19	.84	.88	.77	.79	.82	.75	.84	.88	.77	.79	.82	.75	.74	.8	.62	.79	.83	.71	.74	.78	.67	.74	.78	.67
Susp traseras	Train	48				.98	.96	.99				.98	.96	.99	.98	.96	.99	.77	0	.87				.83	.73	.88
	Test	12				1	1	1				1	1	1	1	1	1	.83	0	.91				.92	.8	.95
Cil levante	Train	48	.98	.93	.99	.98	.93	.99	.92	.78	.95	.96	.88	.97	.85	0	.92	.85	0	.92	.69	.44	.78	.94	.82	.96
	Test	13	1	1	1	1	1	1	.85	.67	.9	.85	.67	.9	.85	0	.92	.85	0	.92	.69	.33	.8	.92	.8	.95
Wheel motors	Train	52	1	1	1	.98	.98	.98	.96	.96	.96	.94	.94	.95	.92	.92	.93	.83	.81	.84	1	1	1	.56	0	.72
	Test	14	.86	.88	.83	.93	.93	.92	.86	.88	.83	.93	.93	.92	.93	.93	.92	.86	.86	.86	.79	.82	.73	.5	.46	.53
Cil direcc	Train	90	1	1	1	1	1	1	.97	.97	.95	1	1	1	.97	.97	.95	.96	.96	.94	.97	.97	.95	.92	.94	.9
	Test	23	.96	.97	.94	.87	.9	.82	.83	.87	.75	.87	.9	.82	.78	.85	.62	.83	.87	.75	.83	.87	.75	.78	.84	.67
Front Wheels	Train	76	.95	.94	.96	.96	.95	.97	.92	.91	.93	.91	.89	.92	.92	.91	.93	.96	.95	.97	.91	.89	.92	.92	.91	.93
	Test	19	.84	.8	.87	.84	.8	.87	.84	.8	.87	.84	.82	.86	.84	.8	.87	.84	.8	.87	.84	.82	.86	.84	.82	.86

Se observa que el mejor modelo para la predicción de las fallas de las suspensiones frontales es el Random Forest y el XGBoost aplicado sobre el dataset PCA, sin embargo, al juzgar por la diferencia entre el rendimiento obtenido en el entrenamiento y el de prueba, la menor diferencia se obtiene con el XGBoost, por lo que se seleccionó como el mejor modelo para este componente.

En cuanto a las suspensiones traseras, el Random Fores, XGBoost y Perceptón Multiple con solver lbfgs tienen un muy buen rendimiento con poco sobreajuste, juzgando por el menor tiempo de sintonización y entrenamiento, el mejor modelo para este componente fue el XGBoost.

Respecto a los cilindros de levante, el Random Forest fue el mejor modelo, alcanzando una predicción muy buena con muy poco sobreajuste en ambos datasets.

Para el Wheel motor, los mejores modelos fueron el Random Forest y XGBoost aplicados sobre el dataset exprt junto con el perceptrón múltiple con solver lbfgs, juzgando por las diferencias entre

el rendimiento en la fase de entrenamiento y la de prueba, el mejor modelo fue el perceptrón multicapa.

Para los cilindros de dirección, el mejor modelo fue el Random Forest aplicado sobre el dataset PCA.

Finalmente, para las ruedas frontales los mejores modelos fueron el XGBoost aplicado sobre el dataset exprt y la regresión logística aplicada sobre ambos datasets. Debido a que presentan diferencias entre el rendimiento en la fase de entrenamiento y prueba similares, se seleccionó como mejor modelo el XGBoost aplicado sobre el dataset exprt para evitar perder explicación al utilizar transformaciones con la regresión logística aplicada sobre el dataset PCA.

Debido a que la Regresión Logística no obtuvo el mejor rendimiento en ninguno de los componentes, se determinó no ahondar en la validación de supuestos ni análisis de coeficientes, sino que más bien se realiza el análisis de la importancia de los atributos en las técnicas que presentaron mejor rendimiento por componente (para el wheel motor cuyo mejor modelo fue el perceptrón multicapa se presenta el XGBoost cuyo resultado fue el mismo).

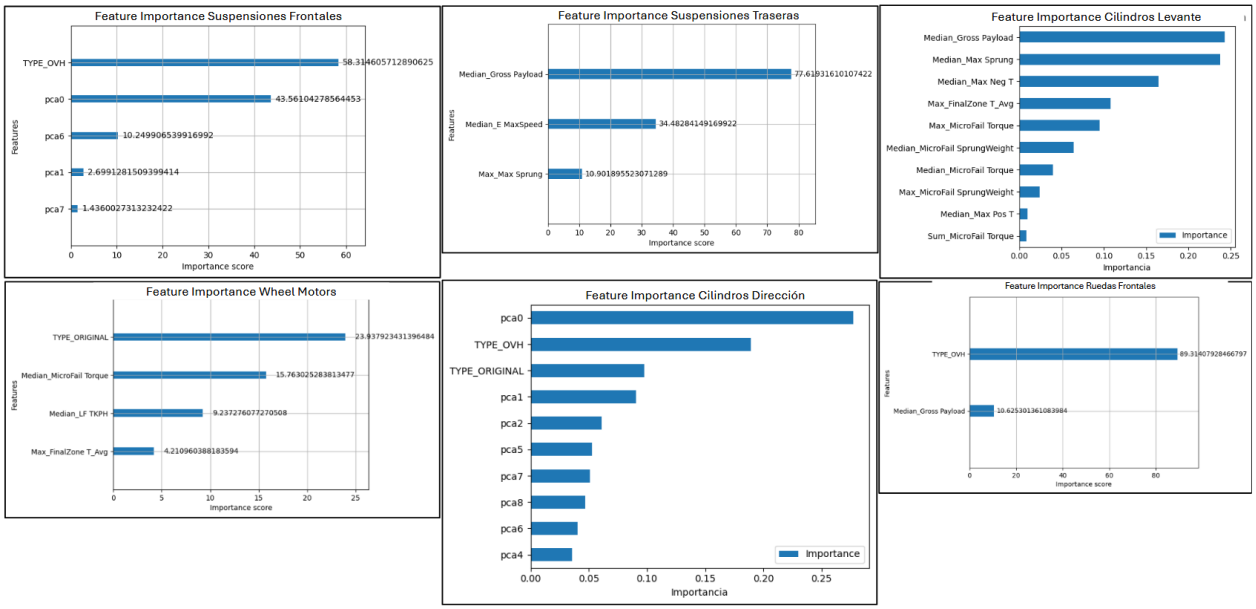


Fig. 31 Importancia de atributos en predicción de fallas por componente. Fuente: Autores

En general, se observa que los atributos Median Gross Payload (carga acarreada por ciclo) y Type_OVH (el cual informa si el componente fue reparado en overhaul) son los más comunes en los componentes. Los pesos de la componente principal 0 de las suspensiones frontales (que representa el 54% de la varianza) y de la componente principal 0 de los cilindros de dirección (que representa un 38% de la varianza), se presentan a continuación

Tabla 13 Coeficientes Componente 0 Suspensiones Frontales

Feature	Loading (ϕ)	ϕ^2

Sum_LF TKPH	0,23831528	5,68%
Sum_R TKPH	0,23804686	5,67%
Sum_L-Max Speed	0,23794802	5,66%
Sum_E-Haul Distance	0,23785919	5,66%
Sum_RF TKPH	0,23785388	5,66%
Sum_L-Haul Distance	0,23768232	5,65%
Sum_E MaxSpeed	0,23753278	5,64%
Sum_Max Sprung	0,23714987	5,62%
Sum_Gross Payload	0,23659445	5,60%
Sum_Max Pos T	0,23540413	5,54%
Sum_Max Neg T	0,23376773	5,46%
Sum_MicroFail SprungWeight	0,23121674	5,35%
Sum_MicroFail Torque	0,22000775	4,84%

Tabla 14 Coeficientes Componente 0 Cilindros Dirección

Feature	**Loading (ϕ)**	**ϕ^2**
Median_R TKPH	0,43103169	18,58%
Median_LF TKPH	0,41890659	17,55%
Median_RF TKPH	0,40529772	16,43%
Median_L-Max Speed	0,39096476	15,29%
Median_Gross Payload	-0,29793078	8,88%
Median_MicroFail Torque	-0,27355028	7,48%
Median_E MaxSpeed	0,25720825	6,62%
Median_Max Neg T	-0,18153046	3,30%
Max_FinalZone T_Avg	0,16215554	2,63%
Max_E MaxSpeed	0,10118335	1,02%
Median_Max Sprung	-0,09831802	0,97%
Sum_MicroFail Torque	-0,09570074	0,92%
Max_MicroFail Torque	-0,05958034	0,35%

Se observa que los atributos relacionados a TKPH son los más relevantes en ambos casos.

En general, se observa un buen rendimiento de los modelos para la predicción de fallas en cada uno de los componentes considerados, por lo que se afirma que los componentes a los que se pueden predecir fallas son los siguientes.

Tabla 15 Listado de Componentes y Fallas a Predecir

Componente	Falla a predecir	Tecnica	Preprocesamiento	Accuracy	F1 (+)	F1 (-)	AUC
Suspensiones Frontales	Fuga de aceite por sellos	XGBoost	Escalamiento, PCA, One hot	0,84	0,88	0,77	0,81
Suspensiones traseras	Fuga de aceite por sellos	XGBoost	One hot	1	1	1	1
Cilindros de levante	Fuga de aceite por sellos	Random Forest	One hot	1	1	1	1
Wheel Motors	Fuga de aceite por sellos	Perceptrón multicapa	Escalamiento, PCA, One hot	0,93	0,93	0,92	0,92
Cilindros de direccion	Fuga de aceite por sellos	Random Forest	Escalamiento, PCA, One hot	0,96	0,97	0,94	0,97
Ruedas frontales	Fuga de aceite por sellos	XGBoost	One hot	0,84	0,82	0,86	0,88

7. PROTOTIPO DE INTERFAZ DE SEGUIMIENTO PARA LA PREDICCION DE FALLAS

Una vez desarrollados los modelos, se propuso crear una tabla por cada componente al que se le realizaría seguimiento en la base de datos, es decir, una tabla de suspensiones frontales, traseras, cilindros de levante, cilindros de dirección, ruedas frontales y traseras. En dichas tablas se propuso consignar cada componente instalado en los camiones (y por ende operativos, sin falla) con su respectivo SN e información que posibilite su identificación.

Como la predicción de falla se realizó a partir de la data descargadas de los módulos de los camiones (las cuales se realizan mensualmente por la minera), se propuso agregar una fila por cada mes que el componente estuvo operativo, de esta manera se crearía una tendencia mes a mes de las variables predictoras y de la estimación de la probabilidad de falla. La estructura de cada tabla propuesta se presenta en la Tabla 16

Tabla 16 Tabla propuesta para seguimiento de predicción de fallas

Nombre Columna	Descripción	Tipo de Variable
ID	Id del registro	Cuantitativa
SN	Serial del componente	Categórica
LIFE	Nro de veces que el componente ha sido reutilizado	Cuantitativa
TRUCK	Camión donde está instalado el componente	Categórica
LOC	Localización (Izquierda o derecha)	Categórica
DATE_INST	Fecha de instalación	Cuantitativa
DATE_CURRENT	Fecha en la que se hace la predicción	Cuantitativa
V1	Variable predictora 1	
V2	Variable predictora 2	
...	...	
Vn	Variable predictora n	
FAILED_	Si el componente falló o no	Categórica
FAILED_PROB	Probabilidad estimada de falla	Cuantitativa
FAILED_PREDIC	Predicción de la falla	Booleana

En la Fig. 32 se presenta el ejemplo de la tabla para los Suspensiones Traseras.

ID	SN	LIFE	TRUCK	LOC	DATE...	DATE_CU...	MEDIAN_E_MAXSPEED	MAX_MAX_SPRUNG	MEDIAN_GROSS_PAYLOAD	TYPE_	FAILED_	FAILED_PROB	FAILED_PREDIC
1	345	1	C108	RH	03-NOV-18	28-FEB-25	43.5	1090.8	304.3	ORIGINAL	No	0.48944867	0
2	345	1	C108	RH	03-NOV-18	31-JAN-25	43.5	1090.8	304.1	ORIGINAL	No	0.48944867	0
3	345	1	C108	RH	03-NOV-18	31-DEC-24	43.5	1090.8	304.0	ORIGINAL	No	0.48944867	0
4	345	1	C108	RH	03-NOV-18	30-NOV-24	43.5	1090.8	303.7	ORIGINAL	No	0.48944867	0
5	345	1	C108	RH	03-NOV-18	31-OCT-24	43.5	1090.8	303.6	ORIGINAL	No	0.48944867	0
6	345	1	C108	RH	03-NOV-18	30-NOV-24	43.5	1090.8	303.7	ORIGINAL	No	0.48944867	0
7	345	1	C108	RH	03-NOV-18	31-AUG-24	43.3	1090.8	303.4	ORIGINAL	No	0.86049217	1
8	345	1	C108	RH	03-NOV-18	31-JUL-24	43.3	1090.8	303.3	ORIGINAL	No	0.86049217	1
9	345	1	C108	RH	03-NOV-18	30-JUN-24	43.3	1090.8	303.2	ORIGINAL	No	0.98389614	1

Fig. 32 Tabla de seguimiento de predicción de fallas Suspensiones Traseras. Fuente: Autores

El flujo propuesto consistió en que mes a mes, los ingenieros de soporte de fábrica introdujeran los SN de los componentes instalados junto con los campos LIFE, TRUCK, LOC, DATE_INST,

DATE_CURRENT y FAILED_; posteriormente, una vez almacenadas las descargas de los módulos en la tabla Cycles, un script de Python se encargaba de seleccionar las variables predictoras para cada componente, realizar su preprocesamiento, estimar la probabilidad de falla y completar los demás campos de la tabla. Una vez predicha las fallas, la información podría observarse en un dashboard elaborado en Power BI donde se destacan aquellos componentes cuya predicción fue positiva.

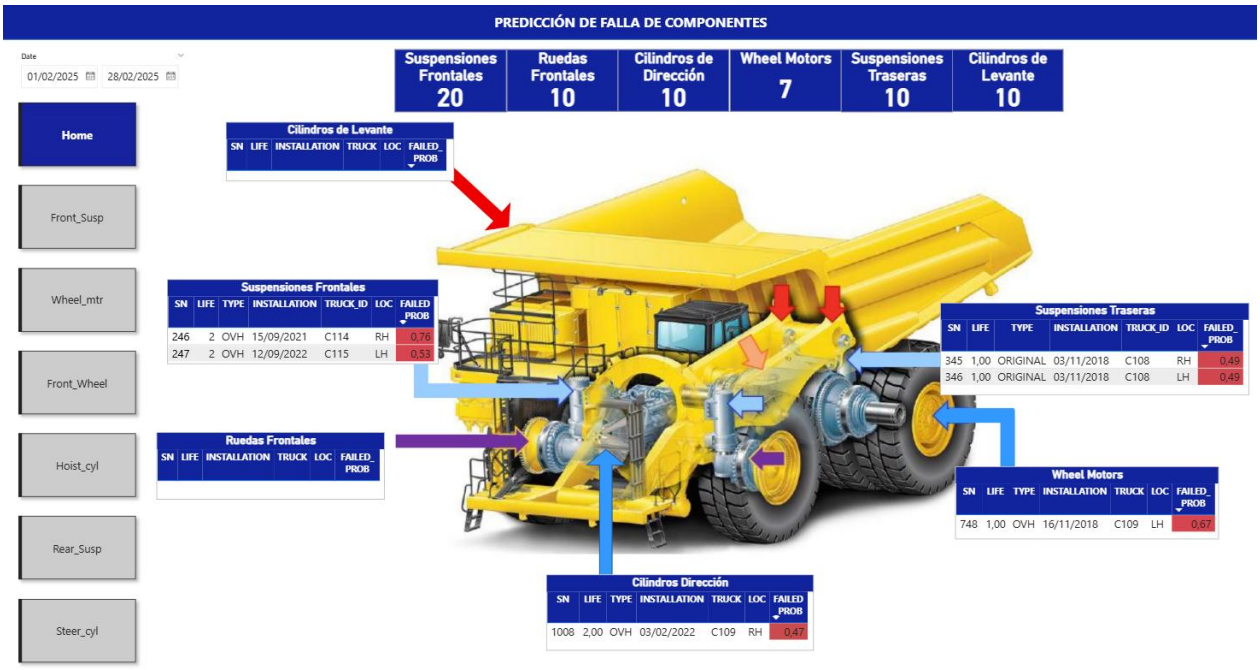


Fig. 33 Interfaz de Seguimiento para la predicción de fallas. Fuente: Autores

En la interfaz se añadió una página por componente donde se presentan las tendencias de las variables predictoras más importantes y/o cuyo coeficiente de variación fue superior al 20%. Al gráfico de tendencia de cada variable, se le agregó el intervalo de confianza del 95% a partir del cual se podría presentar la falla, esto con el fin de identificar tendencias peligrosas que puedan ocasionar fallas prematuras.

Fig. 34 Interfaz de Seguimiento para Suspensiones Frontales. Fuente: Autores

Con la información provista por la interfaz, se podrá entonces listar componentes para ser reemplazados de manera preventiva bajo el enfoque de su condición y vida útil la cual no es medida sólo por las horas de operación, sino por el uso u operación que realmente ha prestado.

En resumen, el flujo de trabajo propuesto se presenta en la Fig. 35

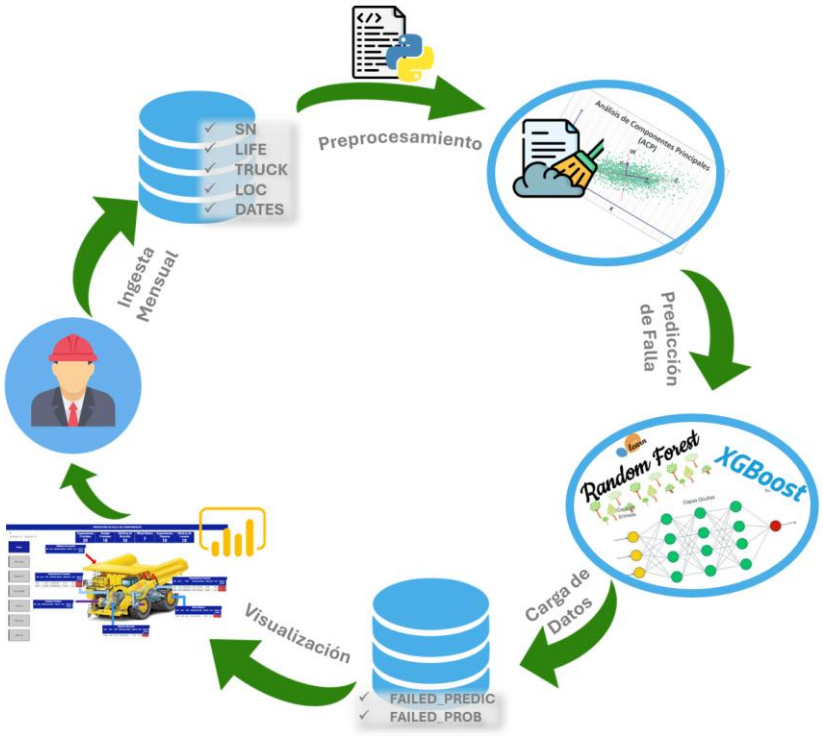


Fig. 35 Flujo para la Predicción de Fallas. Fuente: Autores

8. CONCLUSIONES Y TRABAJOS FUTUROS

En este capítulo se listan las conclusiones basadas en el desarrollo de cada uno de los objetivos del proyecto aplicado, así como futuros trabajos que podrían ejecutarse para complementar la aplicación y beneficios.

8.1. CONCLUSIONES

El desarrollo de este proyecto aplicado se centró en responder a la problemática de la predicción de fallas prematuras debido a fugas hidráulicas de 6 componentes en una flota de camiones mineros tales como Suspensiones Frontales, Ruedas Frontales, Cilindros de dirección, Suspensiones Traseras, Ruedas Motorizadas y Cilindros de Levante. Utilizando diferentes técnicas de análisis y Ciencia de Datos fue posible alcanzar los objetivos propuestos obteniendo precisiones superiores al 84% en la predicción, donde el XGBoost fue la técnica que mejor rendimiento presentó seguida por el Random Forest, mientras que otras técnicas utilizadas, como la Regresión Logística, presentaron un rendimiento menor en términos de precisión con valores muy cercanos a los alcanzados por el desarrollo realizado por Karem Lastarria en 2024 en la predicción de fallas en motores diésel de quipos mineros (78%) [20].

En cuanto a las variables en las que la minera debiera enfocarse para realizar seguimiento y control en busca de evitar este tipo de fallas, se encontró que el valor acumulado, máximo y medio de atributos como las microfallas por torsión y flexión, cargas suspendidas, valores de torsión y carga acarreada son fundamentales para la vida de estos componentes. En aras de proporcionar valores críticos a dichos atributos, se halló el intervalo de confianza del 95% a partir del cual se puede presentar la falla y se graficaron tomando como ejemplo el desarrollo realizado por Martín Valderrama en el 2022 en su proyecto de pronóstico de fallas en motores diésel basado en indicadores de degradación probabilísticos [21].

Finalmente, con los elementos creados para el desarrollo del proyecto, dentro de los que se incluyen la base de datos SQL, los scripts del preprocesamiento de datos y predicción de fallas y la interfaz de seguimiento, se propuso un flujo de trabajo como lo hicieron Pablo Samillan y Eveling Castro en Perú [22] con la diferencia de que en lugar de Redes Neuronales Recurrentes se utilizan otras técnicas y preprocesamientos.

La aplicación de este proyecto sirve a la minera como un insumo para tener el criterio necesario a la hora de decidir qué componentes deben ser reemplazados, previniendo así tiempos muertos indeseados y las ineficiencias que las fallas imprevistas generan en toda la operación.

8.2. TRABAJOS FUTUROS

Debido a la disponibilidad de los datos, la predicción de falla se centró en los 6 componentes mencionados con un solo modo de falla, fugas hidráulicas, a medida que aparezcan más modos de falla, conviene expandir el proyecto con el fin de desarrollar modelos para la predicción de nuevos modos de falla; de igual manera, a medida que se recopile información de la falla de otros

componentes como grietas en componentes estructurales, por ejemplo, conviene expandir el proyecto para desarrollar nuevos modelos que abarquen otros componentes.

En cuanto al comportamiento de las tendencias, sería conveniente realizar un pronóstico que permita estimar la fecha en la cual los componentes podrían fallar, tal como lo propuso Martín Valderrama en la falla de motores diésel [21], esto permitiría predecir con cierto tiempo de antelación, lo que traería mayores beneficios a otros grupos interesados como Planeación, Repuestos y Comercio exterior.

Finalmente, otra propuesta de trabajo futuro sería expandir este proyecto a nivel regional por medio de la fábrica de los camiones, pues estos equipos son muy utilizados en mineras de Perú y Chile, por lo que los beneficios adquiridos a través de este proyecto podrían ser replicados para más clientes e incluso ofrecido como un servicio que la fábrica de camiones podría ofrecer a nivel mundial, esto a su vez permitiría un desarrollo de modelos con mayor alcance, pues al tener mucha más información, los limitantes descritos en este proyecto (como modos de falla y componentes considerados) pueden aminorarse debido a una mayor disponibilidad de información.

ANEXOS

Anexo A: Variables Predictoras por Componente según Análisis Multivariado

1. **Suspensiones frontales**

- Sum_L-Haul Distance
- Sum_E-Haul Distance
- Sum_L-Max Speed
- Max_E MaxSpeed
- Sum_E MaxSpeed
- Max_Max Pos T
- Sum_Max Pos T
- Median_Max Neg T
- Max_Max Neg T
- 'Sum_Max Neg T
- Median_Max Sprung
- Max_Max Sprung
- Sum_Max Sprung
- Median_LF TKPH
- Sum_LF TKPH
- Median_RF TKPH
- Sum_RF TKPH
- Median_R TKPH
- Max_R TKPH
- Sum_R TKPH
- Median_MicroFail Torque
- Max_MicroFail Torque
- Sum_MicroFail Torque
- Median_MicroFail SprungWeight
- Max_MicroFail SprungWeight
- Sum_MicroFail SprungWeight
- Max_Maneuvering T_Avg
- Max_FinalZone T_Avg
- Median_Gross Payload
- Max_Gross Payload
- Sum_Gross Payload
- TYPE
- MODIFICATION
- FAILED

2. **Cilindros de Levante**

- Median_Max Pos T
- Max_Max Pos T
- Sum_Max Pos T
- Median_Max Neg T
- Max_Max Neg T
- Sum_Max Neg T
- Median_Max Sprung
- Median_MicroFail Torque
- Max_MicroFail Torque
- Sum_MicroFail Torque
- Median_MicroFail SprungWeight
- Max_MicroFail SprungWeight
- Sum_MicroFailSprungWeight
- Max_FinalZone T_Avg
- Median_Gross Payload
- Max_Gross Payload
- Sum_Gross Payload
- FAILED

3. **Suspensiones traseras**

- Median_E MaxSpeed
- Max_Max Sprung
- Median_Gross Payload
- TYPE
- FAILED

4. **Cilindros de dirección**

- Median_L-Max Speed
- Median_E MaxSpeed
- Max_E MaxSpeed
- Median_Max Neg T
- Median_Max Sprung
- Median_LF TKPH
- Median_RF TKPH
- Median_R TKPH
- Median_MicroFail Torque
- Max_MicroFail Torque
- Sum_MicroFail Torque
- Max_FinalZone T_Avg
- Median_Gross Payload
- TYPE
- MODIFICATION
- FAILED

5. **Ruedas frontales**

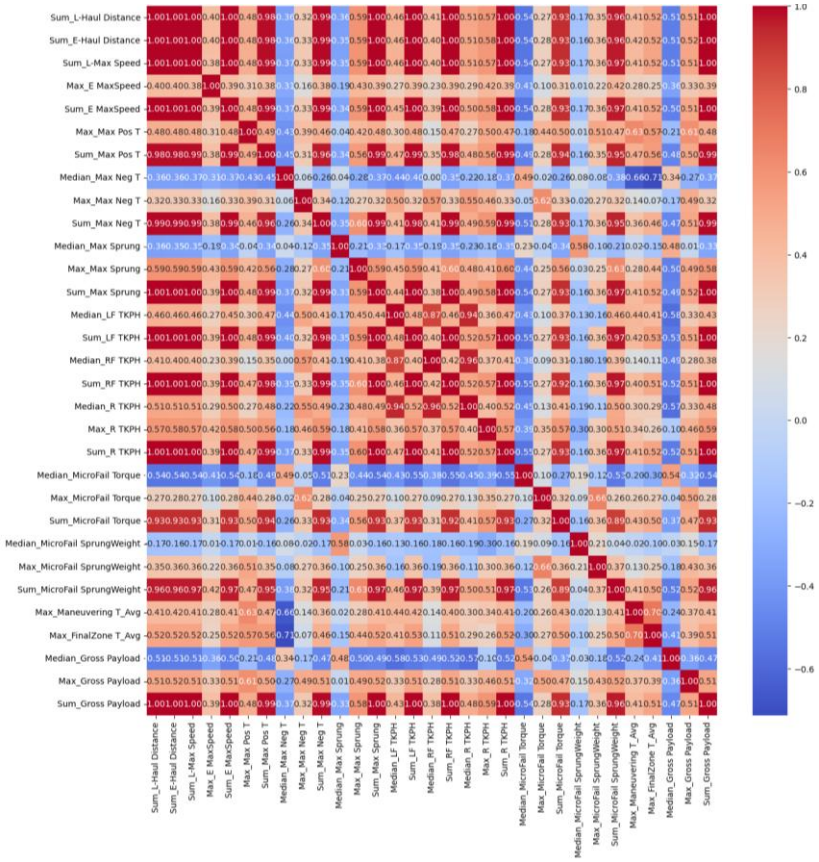
- Median_L-Haul Distance
- Sum_L-Haul Distance
- Median_E-Haul Distance
- Sum_E-Haul Distance
- Median_L-Max Speed
- Sum_L-Max Speed
- Max_E Max Speed
- Sum_E Max Speed
- Sum_Max Pos T
- Median_Max Neg T
- Sum_Max Neg T
- Median_Max Sprung
- Max_Max Sprung
- Sum_Max Sprung
- Median_LF TKPH
- Sum_LF TKPH
- Median_RF TKPH
- Sum_RF TKPH
- Median_R TKPH
- Sum_R TKPH
- Median_MicroFail Torque
- Sum_MicroFail Torque
- Sum_MicroFail Sprung Weight
- Max_FinalZoneT_Avg
- Median_Gross Payload
- Max_Gross Payload
- Sum_Gross Payload
- TYPE
- MODIFICATION
- FAILED

6. **Ruedas motorizadas**

- Median_L-Haul Distance
- Max_L-Haul Distance
- Sum_L-Haul Distance
- Median_E-Haul Distance
- Sum_E-Haul Distance
- Median_L-Max Speed
- Max_L-Max Speed
- Sum_L-Max Speed
- Max_E MaxSpeed
- Sum_E MaxSpeed
- Sum_Max Pos T
- Median_Max Neg T
- Sum_Max Neg T
- Max_Max Sprung
- Sum_Max Sprung
- Median_LF TKPH
- Max_LF TKPH
- Sum_LF TKPH
- Median_RF TKPH
- Max_RF TKPH
- Sum_RF TKPH
- Median_R TKPH
- Max_R TKPH
- Sum_R TKPH
- Median_MicroFail Torque
- Sum_MicroFail Torque
- Sum_MicroFail SprungWeight
- Max_Maneuvering T_Avg
- Max_FinalZone T_Avg
- Median_Gross Payload
- Sum_GrossPayload
- TYPE
- MODIFICATION
- FAILED

Anexo B: Matrices de Correlación por Dataset de Componentes

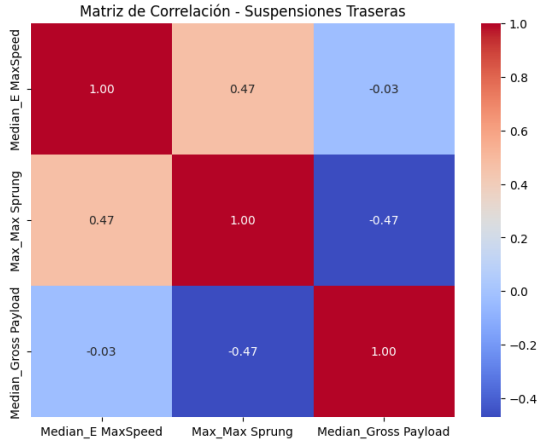
Suspensiones Frontales



TOP 5 VIF

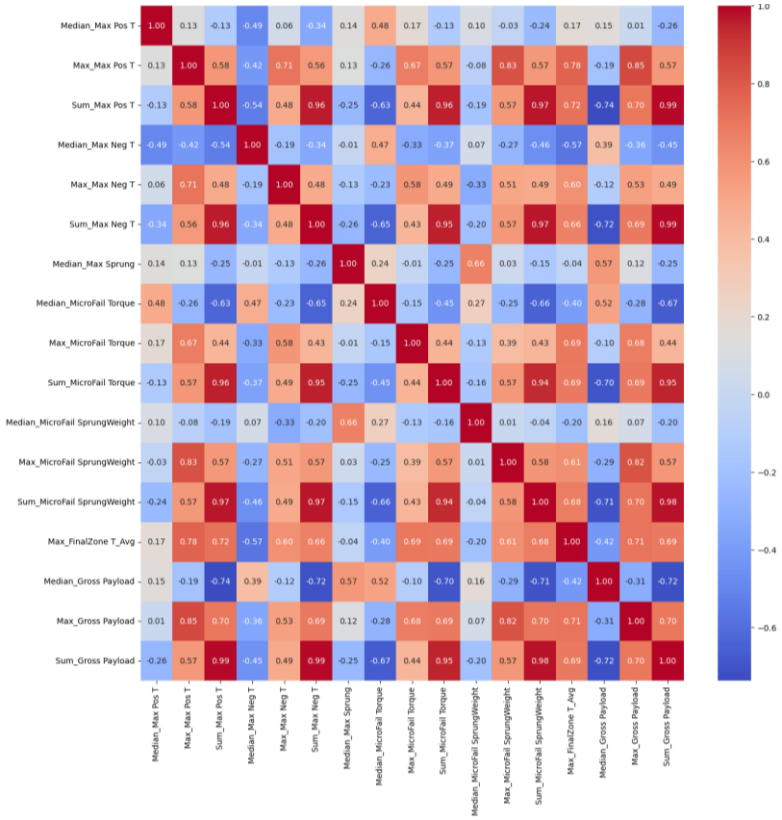
Variable	VIF
Sum_Max Sprung	31987
Sum_RF TKPH	30725
Sum_Gross Payload	28895
Sum_LF TKPH	27844
Sum_R TKPH	22267

Suspensiones Traseras



Variable	VIF
Max_Max Sprung	1.75
Median_E MaxSpeed	1.36
Median_Gross Payload	1.36

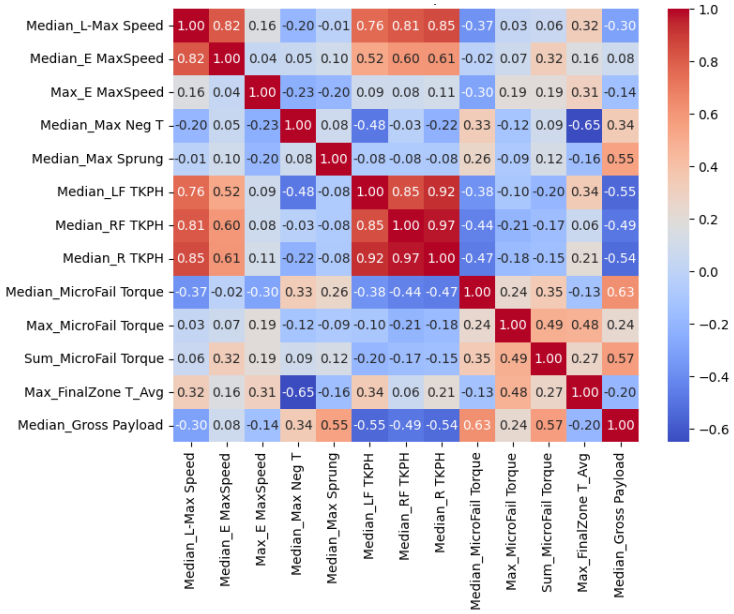
Cilindros de Levante



TOP 5 VIF

Variable	VIF
Sum_Gross Payload	2.40e+06
Sum_Max Neg T	7.82e+03
Sum_Max Pos T	3.08e+03
Sum_MicroFail Torque	2.42e+03
Sum_MicroFail SprungWeight	3.38e+02

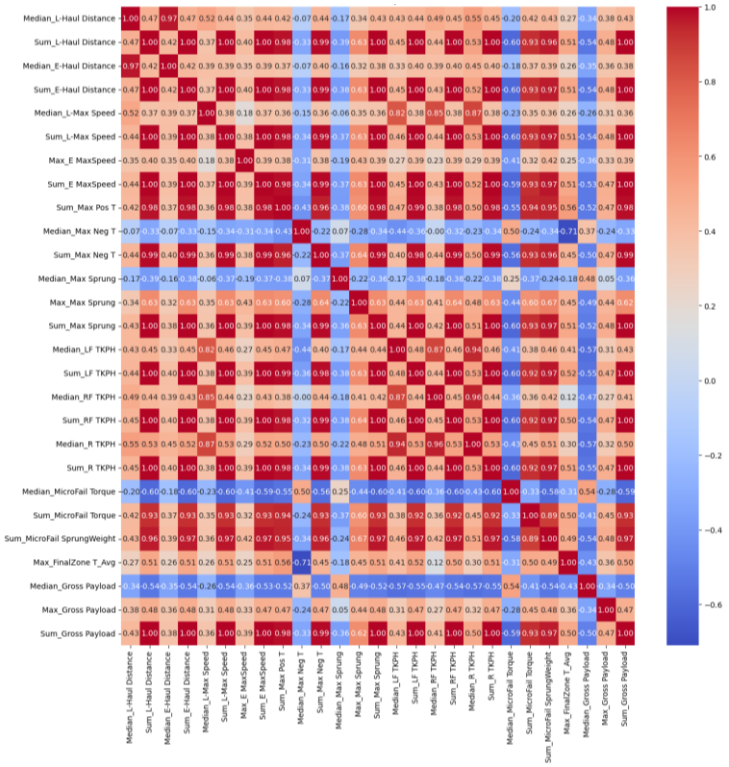
Cilindros de dirección



TOP 5 VIF

Variable	VIF
Median_R TKPH	66.87
Median_RF TKPH	60.66
Median_LF TKPH	38.11
Median_Max Neg T	12.58
Median_L-Max Speed	12.30

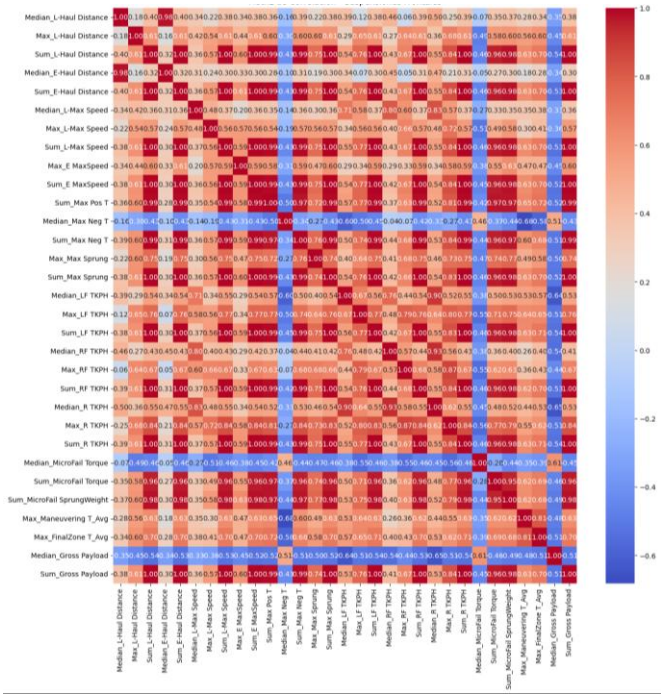
Front Wheels



TOP 5 VIF

Variable	VIF
Sum_R TKPH	29447
Sum_LF TKPH	28596
Sum_RF TKPH	23970
Sum_Gross Payload	23316
Sum_Max Sprung	20315

Wheel motors



TOP 5 VIF

Variable	VIF
Sum_R TKPH	60081
Sum_LF TKPH	55385
Sum_RF TKPH	43862
Sum_L-Haul Distance	41054
Sum_E-Haul Distance	38522

Anexo C: Variables Predictoras por Componente según Expertos

1. **Suspensiones frontales**

- Sum_Microfail Torque
- Max_E MaxSpeed
- Max_Max Pos T
- Median_Max Neg T
- Max_Max Neg T
- Median_Max Sprung
- Max_Max Sprung
- Median_LF TKPH
- Max_R TKPH
- Median_Microfail Torque
- Max_Microfail Torque
- Median_Microfail SprungWeight
- Max_Microfail SprungWeight
- Max_Maneuvering T_Avg
- Max_FinalZone T_Avg
- Median_Gross Payload
- Max_Gross Payload

2. **Cilindros de Levante**

- Median_Max Pos T
- Max_Max Pos T
- Sum_Microfail Torque
- Median_Max Neg T
- Max_Max Neg T
- Median_Max Sprung
- Median_Microfail Torque
- Max_Microfail Torque
- Median_Microfail Sprungweight
- Max_Microfail Sprung Weight
- Max_FinalZone T_Avg
- Median_Gross Payload

3. **Suspensiones traseras**

- Median_E MaxSpeed
- Max_Max Sprung
- Median_Gross Payload
- TYPE
- FAILED

4. **Cilindros de dirección**

- Median_L-Max Speed
- Max_E MaxSpeed
- Median_Max Neg T
- Median_Max Sprung
- Median_LF TKPH
- Median_Microfail Torque
- Max_Microfail Torque
- Sum_Microfail Torque
- Max_FianlZone T_Avg
- Median_Gross Payload

5. **Ruedas frontales**

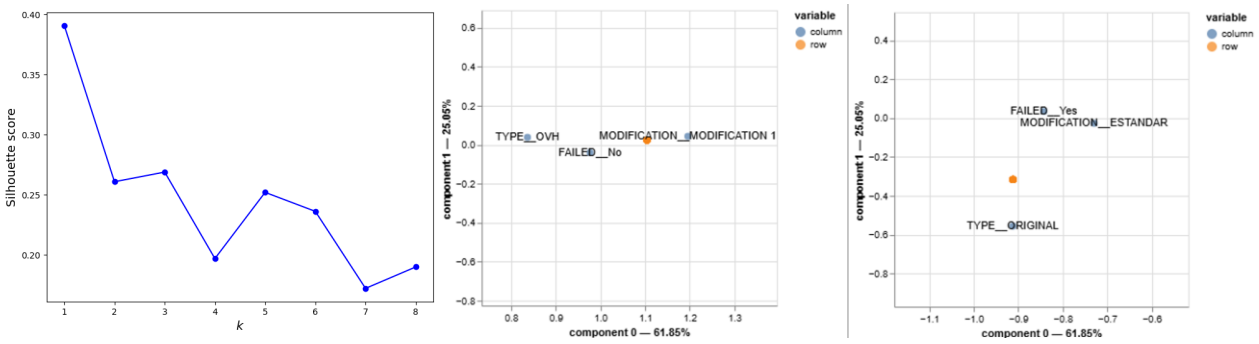
- Median_L-Haul Distance
- Sum_MicroFail Torque
- Median_L-Max Speed
- Max_E MaxSpeed
- Median_Max Neg T
- Median_Max Sprung
- Max_Max Sprung
- Median_Microfail Torque
- Max_FinalZone T_Avg
- Median_Gross Payload
- Max_Gross Payload

6. Ruedas motorizadas

- Median_L-Haul Distance
- Max_L-Haul Distance
- Sum_Microfail Torque
- Median_L-Max Speed
- Max_L-Max Speed
- Max_E MaxSpeed
- Median_Max Neg T
- Max_Max Sprung
- Median_LF TKPH
- Max_LF TKPH
- Median_RF TKPH
- Max_RF TKPH
- Median_Microfail Torque
- Max_Maneuvering T_Avg
- Max_FinalXonte T_Avg
- Median_Gross Payload

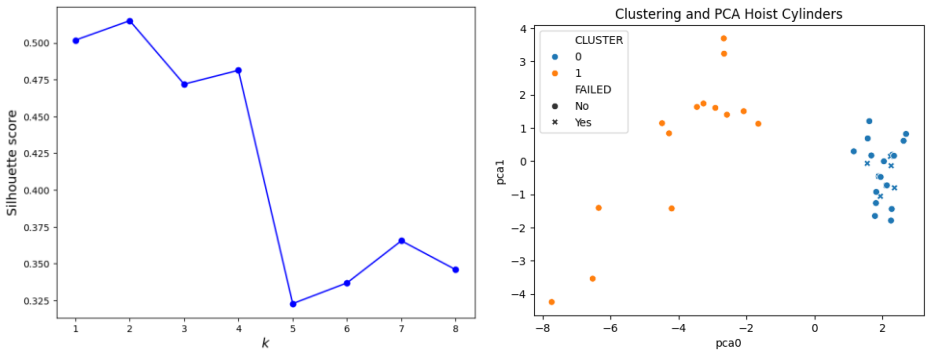
Anexo D: Análisis de Conglomerados y Correspondencia por Componente

Suspensiones Frontales



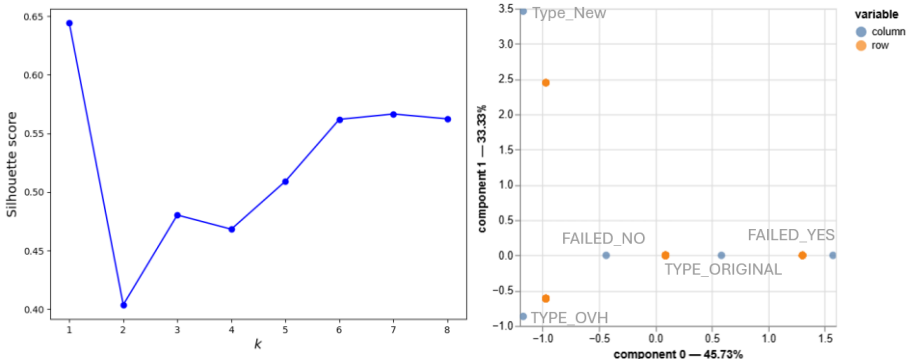
Debido a que no hay mejoría en el índice de Silhouette, no se realizó análisis. El análisis de correspondencia mostrado en la FIG se observa que el tipo OVH (reparadas) con modificación 1 se relacionan con la no falla; mientras que el tipo Original sin modificación (Estandar), se relaciona con la falla.

Cilindros de levante



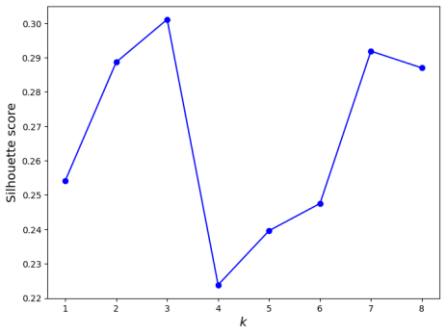
Al realizar el análisis cluster, se observa que los cilindros que pertenecen al grupo 1 no se relacionan con la falla, mientras que aquellos que pertenecen al grupo 0 sí han fallado. Este dataset no contiene variables categóricas, por lo que no fue necesario realizar un análisis de correspondencia.

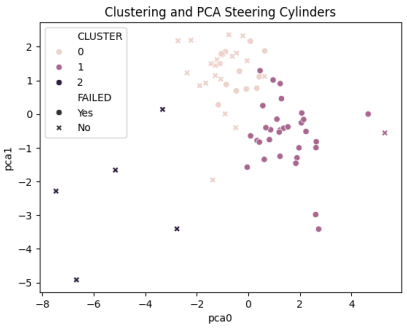
Suspensiones traseras

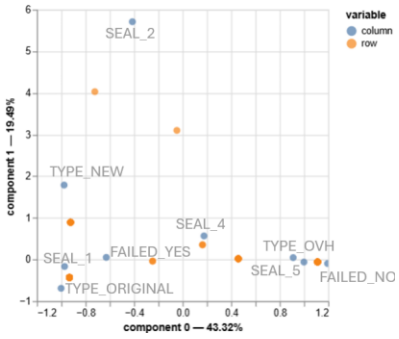


Debido a que no hay mejoría en el índice de Silhouette, no se realizó análisis. Respecto al análisis de correspondencia, se observa que la falla se relaciona más con las suspensiones originales (las que se instalaron por primera vez), mientras que la no falla con las suspensiones que han sido reparadas (OVH) y con las suspensiones nuevas.

Cilindros de dirección

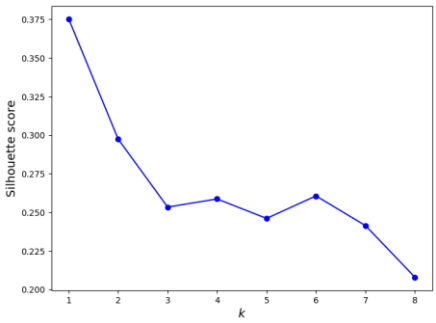


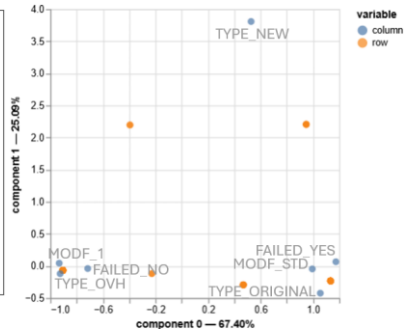




El análisis de conglomerado muestra 3 grupos en los que el grupo 0 presenta componentes fallidos y no fallidos, mientras que el grupo 1 presenta componentes no fallidos y el 3 componentes que sí fallaron. Del análisis de correspondencia se observa que la no falla se relaciona más con los cilindros cuya modificación consiste en el uso del sello número 5 y que han sido reparados, mientras que la falla se relaciona más con el resto de tipos de sellos en componentes originales y nuevos.

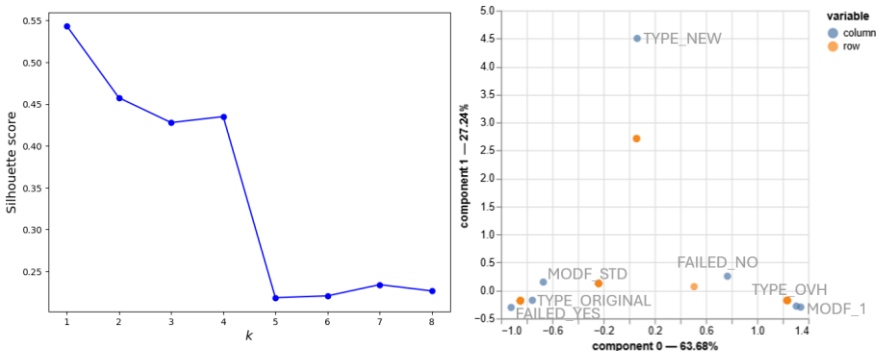
Front Wheels





Debido a que no hay mejoría en el índice de Silhouette, no se realizó análisis. Respecto al análisis de correspondencia, se observa que la falla se relaciona más con los front wheels que no tienen modificación (std), originales y nuevos, mientras que la no falla se relaciona con aquellos que tienen la modificación tipo 1 y que han sido reparados.

Wheel Motors



Debido a que no hay mejoría en el índice de Silhouette, no se realizó análisis. Respecto al análisis de correspondencia, se observa que la falla se relaciona más con los Wheel Motors originales y que no tienen modificación, mientras que la no falla se relaciona más con los reparados y que cuentan con la modificación 1.

Anexo E: Hiperparámetros del mejor clasificador por Componente

1. Suspensiones Frontales

1.1. Random Forest

El mejor estimador en el dataset PCA obtuvo los siguientes hiperparámetros

Parámetros	Rango
n_estimators	22
criterion	entropy
max_depth	5
min_samples_split	4
min_samples_leaf	1
max_features	Sqrt
max_leaf_nodes	None
min_impurity_decrease	0
Bootstrap	True
oob_score	True
warm_star	True
class_weight	None
cc_alpha	0.0
max_samples	None
monotonic_cst	None
Cv	5
Tiempo de Sintonización	18:32

En la validación (16% de los datos) se obtuvieron los siguientes resultados.

Accuracy	Precision	Recall	F1
0.86	0.90	0.83	0.85

El mejor estimador en el dataset exprt obtuvo los siguientes hiperparámetros

Parámetros	Rango
n_estimators	62
criterion	Gini
max_depth	5
min_samples_split	4
min_samples_leaf	1
max_features	Sqrt
max_leaf_nodes	None
min_impurity_decrease	0
Bootstrap	True
oob_score	True
warm_star	True
class_weight	None
cc_alpha	0.0
max_samples	None
monotonic_cst	None
Cv	5

Tiempo de Sintonización	21:59

En la validación (16% de los datos) se obtuvieron los siguientes resultados.

Accuracy	Precision	Recall	F1
0.86	0.89	0.84	0.85

1.2. XGBoost

El mejor estimador en el dataset PCA obtuvo los siguientes hiperparámetros

Parámetros	Rango
n_estimators	5
max_depth	2
learning_rate	0.3
booster	gbtree
gamma	0
tree_method	auto
grow_policy	depth_wise
cv	5
Tiempo de Sintonización	09:08

En el dataset de validación (20% de los datos) se obtuvieron los siguientes resultados.

	Precision	Recall	F1-Score	Muestras
Fail No	0.89	0.67	0.76	12
Fail Yes	0.60	0.86	0.71	7
	Accuracy		0.74	19

El mejor estimador en el dataset exprt obtuvo los siguientes hiperparámetros

Parámetros	Rango
n_estimators	14
max_depth	2
learning_rate	0.3
booster	gbtree
gamma	0
tree_method	auto
grow_policy	depth_wise
cv	5
Tiempo de Sintonización	07:18

En el dataset de validación (20% de los datos) se obtuvieron los siguientes resultados.

	Precision	Recall	F1-Score	Muestras
Fail No	0.89	0.67	0.76	12
Fail Yes	0.60	0.86	0.71	7
	Accuracy		0.74	19

1.3. Perceptrón Multicapa

El mejor estimador con el solver lbfgs obtuvo los siguientes parámetros.

Parámetros	Rango
Hidden_layer_sizes	(1,)
Activation	Tanh
Solver	lbfgs
Alpha	0.10010000000000001
Tiempo de Sintonización	00:10

En la validación (16% de los datos) se obtuvieron los siguientes resultados.

Accuracy	Precision	Recall	F1
0.86	0.86	0.92	0.88

El mejor estimador con el solver SGD o ADAM obtuvo los siguientes parámetros.

Parámetros	Rango
Hidden_layer_sizes	(2, 2)
Activation	logistic
Solver	adam
Alpha	0.10010000000000001
Batch_size	auto
Learning_rate	constant
Learning_rate_init	0.101
Power_t	0.25
Shuffle	True
Momentum	0.1
Nesterovs_momentum	True
Early_stopping	True
Validation_fraction	0.25
Beta_1	0.5
Beta_2	0.5
Epsilon	1e-8
N_iter_no_change	10

Tiempo de Sintonización	172:38

En la validación (16% de los datos) se obtuvieron los siguientes resultados.

Accuracy	Precision	Recall	F1
0.58	0.39	0.50	0.42

1.4. Regresión Logística

El mejor estimador en el dataset PCA obtuvo los siguientes hiperparámetros

Parámetros	Rango
Penalty	L2
Dual	False
C	0.1
Fit_intercept	True
Intercept_scaling	0.2
Solver	Lbfgs
Multi_class	auto
Tiempo de Sintonización	00:32

En la validación (16% de los datos) se obtuvieron los siguientes resultados.

Accuracy	Precision	Recall	F1
0.87	0.90	0.87	0.88

El mejor estimador en el dataset exprt obtuvo los siguientes hiperparámetros

Parámetros	Rango
Penalty	L2
Dual	False
C	0.1
Fit_intercept	True
Intercept_scaling	0.2
Solver	Lbfgs
Multi_class	auto
Tiempo de Sintonización	00:32

En la validación (16% de los datos) se obtuvieron los siguientes resultados.

Accuracy	Precision	Recall	F1
0.84	0.87	0.83	0.84

2. Suspensiones Traseras

Recordando que el dataset de este componente no es de alta dimensión y no presentó multicolinealidad, no se realizó la transformación en componentes principales; además, como el dataset presenta un desbalance en la clase de falla bastante importante (ver Fig. 8), se utilizó el parámetro scale_pos_weight en el XGBoost calculado como el cociente del número de componentes sin falla entre el número de componentes con falla; en la regresión logística se utilizó el parámetro class_wieght para balancear los datos. Los resultados se presentan a continuación.

2.1. Random Forest

El mejor estimador obtuvo los siguientes hiperparámetros

Parámetros	Rango
n_estimators	82
Criterion	gini
max_depth	5
min_samples_split	2
min_samples_leaf	1
max_features	Sqrt
max_leaf_nodes	None
min_impurity_decrease	0
Bootstrap	True
oob_score	True
warm_star	True
class_weight	Balanced
cc_alpha	0.0
max_samples	None
monotonic_cst	None
Cv	5
Tiempo de Sintonización	18:02

En la validación (16% de los datos) se obtuvieron los siguientes resultados.

Accuracy	Precision	Recall	F1
0.83	0.61	0.85	0.68

2.2. XGBoost

El mejor estimador obtuvo los siguientes hiperparámetros

Parámetros	Rango
n_estimators	16
max_depth	4
learning_rate	0.5
booster	gbtree
gamma	0
tree_method	auto
grow_policy	depth_wise
scale_pos_weight	4.14
cv	5

Tiempo de Sintonización	07:34

En el dataset de validación (20% de los datos) se obtuvieron los siguientes resultados.

	Precision	**Recall**	**F1-Score**	**Muestras**
Fail No	1	0.88	0.93	8
Fail Yes	0.80	1	0.89	4
	Accuracy		0.92	12

2.3. Perceptrón Multicapa

En la grilla, se aumentó el número de capas y de neuronas máximo por capa de 3 a 6. El mejor estimador con el solver lbfgs obtuvo los siguientes parámetros.

Parámetros	**Rango**
Hidden_layer_sizes	(5,3)
Activation	Logistic
Solver	lbfgs
Alpha	0.0001
Tiempo de Sintonización	00:36

En la validación (16% de los datos) se obtuvieron los siguientes resultados.

Accuracy	**Precision**	**Recall**	**F1**
0.8	0.48	0.59	0.5

El mejor estimador con el solver SGD o ADAM obtuvo los siguientes parámetros.

Parámetros	**Rango**
Hidden_layer_sizes	(1,)
Activation	Tanh
Solver	adam
Alpha	0.10010000000000001
Batch_size	auto
Learning_rate	constant
Learning_rate_init	0.021
Power_t	0.25
Shuffle	True
Momentum	0.1
Nesterovs_momentum	True
Early_stopping	True
Validation_fraction	0.25

Beta_1	0.9
Beta_2	0.9
Epsilon	1e-8
N_iter_no_change	10
Tiempo de Sintonización	165:05

En la validación (16% de los datos) se obtuvieron los siguientes resultados.

Accuracy	Precision	Recall	F1
0.67	0.09	0.23	0.11

2.4. Regresión Logística

El mejor estimador obtuvo los siguientes hiperparámetros

Parámetros	Rango
Penalty	L2
Dual	False
C	0.3
Fit_intercept	True
Intercept_scaling	0.2
Solver	Lbfgs
Multi_class	auto
Tiempo de Sintonización	00:29

En la validación (16% de los datos) se obtuvieron los siguientes resultados.

Accuracy	Precision	Recall	F1
0.78	0.56	0.92	0.68

3. Cilindros de levante

Se emplearon los parámetros scale_pos_weight y class_wieght para lidiar con el desbalance de los datos. Los resultados se presentan a continuación.

3.1. Random Forest

El mejor estimador en el dataset PCA obtuvo los siguientes hiperparámetros

Parámetros	Rango
n_estimators	42
criterion	entropy
max_depth	7
min_samples_split	2

min_samples_leaf	1
max_features	Sqrt
max_leaf_nodes	None
min_impurity_decrease	0
Bootstrap	True
oob_score	True
warm_star	True
class_weight	Balanced
cc_alpha	0.0
max_samples	None
monotonic_cst	None
Cv	5
Tiempo de Sintonización	18:22

En la validación (16% de los datos) se obtuvieron los siguientes resultados.

Accuracy	Precision	Recall	F1
0.81	0.16	0.14	0.14

El mejor estimador en el dataset exprt obtuvo los siguientes hiperparámetros

Parámetros	Rango
n_estimators	42
criterion	Gini
max_depth	5
min_samples_split	2
min_samples_leaf	1
max_features	None
max_leaf_nodes	None
min_impurity_decrease	0
Bootstrap	True
oob_score	True
warm_star	True
class_weight	Balanced
cc_alpha	0.0
max_samples	None
monotonic_cst	None
Cv	5
Tiempo de Sintonización	21:19

En la validación (16% de los datos) se obtuvieron los siguientes resultados.

Accuracy	Precision	Recall	F1
0.81	0.06	0.05	0.05

3.2. XGBoost

El mejor estimador en el dataset PCA obtuvo los siguientes hiperparámetros

Parámetros	Rango

n_estimators	12
max_depth	1
learning_rate	0.6
booster	Gbtree
gamma	0.1
tree_method	auto
grow_policy	depth_wise
Scale_pos_weight	6.2
cv	5
Tiempo de Sintonización	14:41

En el dataset de validación (20% de los datos) se obtuvieron los siguientes resultados.

	Precision	**Recall**	**F1-Score**	**Muestras**
Fail No	0.90	0.90	0.90	10
Fail Yes	0.50	0.50	0.50	2
	Accuracy		0.83	12

El mejor estimador en el dataset exprt obtuvo los siguientes hiperparámetros

Parámetros	**Rango**
n_estimators	17
max_depth	1
learning_rate	0.7
booster	gbtree
gamma	0
tree_method	auto
grow_policy	depth_wise
Scale_pos_weight	6.2
cv	5
Tiempo de Sintonización	05:41

En el dataset de validación (20% de los datos) se obtuvieron los siguientes resultados.

	Precision	**Recall**	**F1-Score**	**Muestras**
Fail No	0.83	1	0.91	10
Fail Yes	0	0	0	2
	Accuracy		0.83	12

3.3. Perceptrón Multicapa

El mejor estimador con el solver lbfgs obtuvo los siguientes parámetros.

Parámetros	Rango
Hidden_layer_sizes	(1,1)
Activation	Logistic
Solver	lbfgs
Alpha	0.0001
Tiempo de Sintonización	00:10

En la validación (16% de los datos) se obtuvieron los siguientes resultados.

Accuracy	Precision	Recall	F1
0.81	0.01	0.02	0.01

El mejor estimador con el solver SGD o ADAM obtuvo los siguientes parámetros.

Parámetros	Rango
Hidden_layer_sizes	(1, 1)
Activation	logistic
Solver	Sgd
Alpha	0.3001
Batch_size	auto
Learning_rate	Invscaling
Learning_rate_init	0.101
Power_t	0.25
Shuffle	True
Momentum	0.7
Nesterovs_momentum	True
Early_stopping	True
Validation_fraction	0.25
Beta_1	0.5
Beta_2	0.5
Epsilon	1e-8
N_iter_no_change	10
Tiempo de Sintonización	171:23

En la validación (16% de los datos) se obtuvieron los siguientes resultados.

Accuracy	Precision	Recall	F1
0.66	0.66	0.29	0.09

3.4. Regresión Logística

El mejor estimador en el dataset PCA obtuvo los siguientes hiperparámetros

Parámetros	Rango
Penalty	L2
Dual	False
C	0.2
Fit_intercept	True
Intercept_scaling	0.2
Solver	Lbfgs
Multi_class	auto
Class_weight	Balanced
Tiempo de Sintonización	00:25

En la validación (16% de los datos) se obtuvieron los siguientes resultados.

Accuracy	Precision	Recall	F1
0.57	0.2	0.75	0.31

El mejor estimador en el dataset exprt obtuvo los siguientes hiperparámetros

Parámetros	Rango
Penalty	None
Dual	False
C	0.1
Fit_intercept	True
Intercept_scaling	0.2
Solver	Newton-cg
Multi_class	auto
Tiempo de Sintonización	00:23

En la validación (16% de los datos) se obtuvieron los siguientes resultados.

Accuracy	Precision	Recall	F1
0.63	0.19	0.61	0.28

4. Wheel Motors

4.1. Random Forest

En la grilla se consideró el número de estimadores 100 o 200. El mejor estimador en el dataset PCA obtuvo los siguientes hiperparámetros

Parámetros	Rango
n_estimators	100

criterion	entropy
max_depth	5
min_samples_split	5
min_samples_leaf	1
max_features	Sqrt
max_leaf_nodes	None
min_impurity_decrease	0
Bootstrap	True
oob_score	True
warm_star	True
class_weight	None
cc_alpha	0.0
max_samples	None
monotonic_cst	None
Cv	5
Tiempo de Sintonización	00:30

En la validación (16% de los datos) se obtuvieron los siguientes resultados.

Accuracy	Precision	Recall	F1
0.86	0.83	0.89	0.85

El mejor estimador en el dataset exprt obtuvo los siguientes hiperparámetros

Parámetros	Rango
n_estimators	200
criterion	Gini
max_depth	5
min_samples_split	2
min_samples_leaf	2
max_features	Sqrt
max_leaf_nodes	None
min_impurity_decrease	0
Bootstrap	True
oob_score	True
warm_star	True
class_weight	None
cc_alpha	0.0
max_samples	None
monotonic_cst	None
Cv	5
Tiempo de Sintonización	00:30

En la validación (16% de los datos) se obtuvieron los siguientes resultados.

Accuracy	Precision	Recall	F1
0.85	0.84	0.84	0.83

4.2. XGBoost

El mejor estimador en el dataset PCA obtuvo los siguientes hiperparámetros

Parámetros	Rango
n_estimators	4
max_depth	2
learning_rate	0.7
booster	gbtree
gamma	0
tree_method	auto
grow_policy	depth_wise
cv	5
Tiempo de Sintonización	07:30

En el dataset de validación (20% de los datos) se obtuvieron los siguientes resultados.

	Precision	Recall	F1-Score	Muestras
Fail No	0.78	0.88	0.82	8
Fail Yes	0.75	0.60	0.67	5
	Accuracy		0.77	13

El mejor estimador en el dataset exprt obtuvo los siguientes hiperparámetros

Parámetros	Rango
n_estimators	6
max_depth	1
learning_rate	0.6
booster	gbtree
gamma	0
tree_method	auto
grow_policy	depth_wise
cv	5
Tiempo de Sintonización	06:22

En el dataset de validación (20% de los datos) se obtuvieron los siguientes resultados.

	Precision	Recall	F1-Score	Muestras
Fail No	0.88	0.88	0.88	8
Fail Yes	0.80	0.80	0.80	5
	Accuracy		0.85	13

4.3. Perceptrón Multicapa

El mejor estimador con el solver lbfgs obtuvo los siguientes parámetros.

Parámetros	Rango
Hidden_layer_sizes	(3,1)
Activation	Logistic
Solver	lbfgs
Alpha	0.01
Tiempo de Sintonización	00:18

En la validación (16% de los datos) se obtuvieron los siguientes resultados.

Accuracy	Precision	Recall	F1
0.77	0.66	0.73	0.67

El mejor estimador con el solver SGD o ADAM obtuvo los siguientes parámetros.

Parámetros	Rango
Hidden_layer_sizes	(2,)
Activation	Relu
Solver	adam
Alpha	0.0
Batch_size	auto
Learning_rate	constant
Learning_rate_init	0.101
Momentum	0.1
Nesterovs_momentum	True
Early_stopping	False
Beta_1	0.9
Beta_2	0.5
Epsilon	1e-8
N_iter_no_change	10
Tiempo de Sintonización	07:40

En la validación (16% de los datos) se obtuvieron los siguientes resultados.

Accuracy	Precision	Recall	F1
0.62	0.36	0.46	0.39

4.4. Regresión Logística

El mejor estimador en el dataset PCA obtuvo los siguientes hiperparámetros

Parámetros	Rango
Penalty	None
Dual	False
C	0.1
Fit_intercept	False
Intercept_scaling	0.2
Solver	Newton-cholesky
Multi_class	auto
Tiempo de Sintonización	00:34

En la validación (16% de los datos) se obtuvieron los siguientes resultados.

Accuracy	Precision	Recall	F1
0.8	0.75	0.86	0.79

El mejor estimador en el dataset exprt obtuvo los siguientes hiperparámetros

Parámetros	Rango
Penalty	None
Dual	False
C	0.1
Fit_intercept	False
Intercept_scaling	0.2
Solver	Newton-cholesky
Multi_class	auto
Tiempo de Sintonización	00:37

En la validación (16% de los datos) se obtuvieron los siguientes resultados.

Accuracy	Precision	Recall	F1
0.82	0.77	0.89	0.81

5. Cilindros de dirección

Se identificó que la clase FAILED estaba en desbalance con una frecuencia de falla de 59 y una frecuencia de no falla de 31. Para lidiar con el desbalance, se utilizó el parámetro scale_pos_weight en la técnica del XGBoost contemplando los valores 31/59 y 1 en la grilla de búsqueda con el fin de evaluar cuál de estas opciones mejoraba el rendimiento. Para la regresión logística se utilizó el parámetro class_weight = balanced.

5.1. Random Forest

En la grilla se utilizó el número de estimadores de 100 o 200, el mejor estimador en el dataset PCA obtuvo los siguientes hiperparámetros

Parámetros	Rango
n_estimators	100
criterion	entropy
max_depth	5
min_samples_split	2
min_samples_leaf	1
max_features	Sqrt
max_leaf_nodes	None
min_impurity_decrease	0
Bootstrap	True
oob_score	True
warm_star	True
class_weight	Balanced
cc_alpha	0.0
max_samples	None
monotonic_cst	None
Cv	5
Tiempo de Sintonización	00:30

En la validación (16% de los datos) se obtuvieron los siguientes resultados.

Accuracy	Precision	Recall	F1
0.92	0.95	0.93	0.94

El mejor estimador en el dataset exprt obtuvo los siguientes hiperparámetros

Parámetros	Rango
n_estimators	100
criterion	Gini
max_depth	5
min_samples_split	2
min_samples_leaf	1
max_features	Sqrt
max_leaf_nodes	None
min_impurity_decrease	0
Bootstrap	True
oob_score	True
warm_star	True
class_weight	Balanced
cc_alpha	0.0
max_samples	None
monotonic_cst	None
Cv	5
Tiempo de Sintonización	00:30

En la validación (16% de los datos) se obtuvieron los siguientes resultados.

Accuracy	Precision	Recall	F1
0.92	0.95	0.93	0.94

5.2. XGBoost

El mejor estimador en el dataset PCA obtuvo los siguientes hiperparámetros

Parámetros	Rango
n_estimators	100
max_depth	1
learning_rate	0.05
booster	gbtree
gamma	0
tree_method	auto
grow_policy	depth_wise
Scale_pos_wiegth	1
cv	5
Tiempo de Sintonización	00:11

En el dataset de validación (20% de los datos) se obtuvieron los siguientes resultados.

	Precision	Recall	F1-Score	Muestras
Fail No	0.69	0.90	0.78	10
Fail Yes	0.90	0.69	0.78	13
	Accuracy		0.78	23

El mejor estimador en el dataset exprt obtuvo los siguientes hiperparámetros

Parámetros	Rango
n_estimators	100
max_depth	20
learning_rate	0.1
booster	gbtree
gamma	0
tree_method	auto
grow_policy	depth_wise
cv	5
Scale_pos_weight	1
Tiempo de Sintonización	00:06

En el dataset de validación (20% de los datos) se obtuvieron los siguientes resultados.

	Precision	Recall	F1-Score	Muestras
Fail No	0.77	1	0.87	10
Fail Yes	1	0.77	0.87	13
	Accuracy		0.83	23

5.3. Perceptrón Multicapa

El mejor estimador con el solver lbfgs obtuvo los siguientes parámetros.

Parámetros	Rango
Hidden_layer_sizes	(1,3)
Activation	Tanh
Solver	lbfgs
Alpha	0.0001
Tiempo de Sintonización	00:40

En la validación (16% de los datos) se obtuvieron los siguientes resultados.

Accuracy	Precision	Recall	F1
0.84	0.86	0.85	0.85

El mejor estimador con el solver SGD o ADAM obtuvo los siguientes parámetros.

Parámetros	Rango
Hidden_layer_sizes	(1, 3)
Activation	Relu
Solver	adam
Alpha	0.0
Batch_size	auto
Learning_rate	constant
Learning_rate_init	0.1001
Momentum	0.1
Nesterovs_momentum	True
Early_stopping	False
Beta_1	0.9
Beta_2	0.9
Epsilon	1e-8
N_iter_no_change	10
Tiempo de Sintonización	31:31

En la validación (16% de los datos) se obtuvieron los siguientes resultados.

Accuracy	Precision	Recall	F1
0.59	0.51	0.56	0.51

5.4. Regresión Logística

El mejor estimador en el dataset PCA obtuvo los siguientes hiperparámetros

Parámetros	Rango
Penalty	None
Dual	False
C	0.1
Fit_intercept	True
Intercept_scaling	0.2
Solver	Lbfgs
Multi_class	auto
Class_weight	Balanced
Tiempo de Sintonización	00:23

En la validación (16% de los datos) se obtuvieron los siguientes resultados.

Accuracy	Precision	Recall	F1
0.9	0.97	0.87	0.92

El mejor estimador en el dataset exprt obtuvo los siguientes hiperparámetros

Parámetros	Rango
Penalty	L2
Dual	False
C	0.4
Fit_intercept	True
Intercept_scaling	0.2
Solver	Lbfgs
Multi_class	auto
Class_weight	Balanced
Tiempo de Sintonización	00:28

En la validación (16% de los datos) se obtuvieron los siguientes resultados.

Accuracy	Precision	Recall	F1
0.87	0.93	0.87	0.9

6. Ruedas frontales

De manera similar a algunos componentes se emplearon los parámetros scale_pos_weight y class_weight para tratar el desbalance de los datos.

6.1. Random Forest

En la grilla de búsqueda, se implementaron 100 o 200 estimadores, el mejor estimador en el dataset PCA obtuvo los siguientes hiperparámetros

Parámetros	Rango

n_estimators	200
criterion	Gini
max_depth	5
min_samples_split	2
min_samples_leaf	2
max_features	Sqrt
max_leaf_nodes	None
min_impurity_decrease	0
Bootstrap	True
oob_score	True
warm_star	True
class_weight	Balanced
cc_alpha	0.0
max_samples	None
monotonic_cst	None
Cv	5
Tiempo de Sintonización	00:30

En la validación (16% de los datos) se obtuvieron los siguientes resultados.

Accuracy	Precision	Recall	F1
0.9	0.84	0.97	0.88

El mejor estimador en el dataset exprt obtuvo los siguientes hiperparámetros

Parámetros	Rango
n_estimators	100
criterion	Gini
max_depth	5
min_samples_split	5
min_samples_leaf	1
max_features	Sqrt
max_leaf_nodes	None
min_impurity_decrease	0
Bootstrap	True
oob_score	True
warm_star	True
class_weight	Balanced
cc_alpha	0.0
max_samples	None
monotonic_cst	None
Cv	5
Tiempo de Sintonización	00:30

En la validación (16% de los datos) se obtuvieron los siguientes resultados.

Accuracy	Precision	Recall	F1
0.89	0.84	0.92	0.87

6.2. XGBoost

El mejor estimador en el dataset PCA obtuvo los siguientes hiperparámetros

Parámetros	Rango
n_estimators	8
max_depth	1
learning_rate	0.9
booster	gbtree
gamma	0
tree_method	auto
grow_policy	depth_wise
Scale_pos_weight	1
cv	5
Tiempo de Sintonización	09:07

En el dataset de validación (20% de los datos) se obtuvieron los siguientes resultados.

	Precision	Recall	F1-Score	Muestras
Fail No	0.93	1	0.97	14
Fail Yes	1	0.80	0.89	5
	Accuracy		0.95	19

El mejor estimador en el dataset exprt obtuvo los siguientes hiperparámetros

Parámetros	Rango
n_estimators	4
max_depth	1
learning_rate	0.3
booster	gbtree
gamma	0
tree_method	auto
grow_policy	depth_wise
Scale_pos_weight	1
cv	5
Tiempo de Sintonización	07:18

En el dataset de validación (20% de los datos) se obtuvieron los siguientes resultados.

	Precision	Recall	F1-Score	Muestras
Fail No	0.82	1	0.90	14
Fail Yes	1	0.40	0.57	5
	Accuracy		0.84	19

6.3. Perceptrón Multicapa

El mejor estimador con el solver lbfgs obtuvo los siguientes parámetros.

Parámetros	Rango
Hidden_layer_sizes	(1,2)
Activation	Tanh
Solver	Lbfgs
Alpha	5
Tiempo de Sintonización	00:38

En la validación (16% de los datos) se obtuvieron los siguientes resultados.

Accuracy	Precision	Recall	F1
0.75	0.47	0.56	0.5

El mejor estimador con el solver SGD o ADAM obtuvo los siguientes parámetros.

Parámetros	Rango
Hidden_layer_sizes	(2, 2)
Activation	Relu
Solver	Sgd
Alpha	0.0
Batch_size	auto
Learning_rate	constant
Learning_rate_init	0.1001
Momentum	0.4
Nesterovs_momentum	True
Early_stopping	False
Beta_1	0.5
Beta_2	0.5
Epsilon	1e-8
N_iter_no_change	10
Tiempo de Sintonización	08:31

En la validación (16% de los datos) se obtuvieron los siguientes resultados.

Accuracy	Precision	Recall	F1
0.67	0.37	0.46	0.39

6.4. Regresión Logística

El mejor estimador en el dataset PCA obtuvo los siguientes hiperparámetros

Parámetros	Rango
Penalty	L2
Dual	False
C	0.2
Fit_intercept	False
Intercept_scaling	0.2
Solver	Lbfgs
Multi_class	auto
Class_weight	Balanced
Tiempo de Sintonización	00:30

En la validación (16% de los datos) se obtuvieron los siguientes resultados.

Accuracy	Precision	Recall	F1
0.86	0.79	0.89	0.83

El mejor estimador en el dataset exprt obtuvo los siguientes hiperparámetros

Parámetros	Rango
Penalty	L2
Dual	False
C	0.3
Fit_intercept	False
Intercept_scaling	0.2
Solver	Lbfgs
Multi_class	auto
Class_weight	Balanced
Tiempo de Sintonización	00:36

En la validación (16% de los datos) se obtuvieron los siguientes resultados.

Accuracy	Precision	Recall	F1
0.84	0.77	0.85	0.79

REFERENCIAS BIBLIOGRÁFICAS

[1] C. Escalante, «Ciclos y Tiempos de Carguio y Acarreo de Mineral y Desmonte,» 11 enero 2018. [En línea]. Available: https://www.scribd.com/document/368888415/Ciclos-y-Tiempos-de-Carguio-y-Acarreo-de-Mineral-y-Desmonte. [Último acceso: 2024 noviembre 21].

[2] E. Tapia, «Sistema de despacho minero,» *Revista Inglomayor,* nº 22.

[3] M. Torres, «Estructuras,» Xunta de Galicia, 2014.

[4] Tecnología del Automovil, «Torsión y flexión de la carrocería,» Tecnología del Automovil, [En línea]. Available: https://www.tecnologia-automovil.com/torsion-y-flexion-de-la-carroceria/. [Último acceso: 21 noviembre 2024].

[5] Países minero, «Komatsu 930 E4 - El Armado de un Gigante!,» Países minero, [En línea]. Available: https://www.paisminero.co/component/tags/tag/komatsu. [Último acceso: 21 noviembre 2024].

[6] R. Budynas y J. Keith, «Esfuerzos variables y fluctuantes,» de *Diseño en ingeniería mecánica de Shigley*, 9na ed., Mexico D.F, Mc Graw Hill, 2012, pp. 307-308.

[7] Temas de Análisis Estadístico Multivariado, Temas de Análisis Estadístico Multivariado, Editorial Universidad de Costa Rica.

[8] M. J. Rodríguez y R. Mora, «Análisis de correspondencia,» *Publicaciones de la Universidad de Alicante,* pp. 43-56, 2001.

[9] M. McHugh, «The Chi-square test of independence,» Enero 2013. [En línea]. Available: https://www.researchgate.net/publication/253336860_The_Chi-square_test_of_independence.

[10] A. Calviño y J. Alonso, «ANALISIS NO SUPERVISADO,» de *INTRODUCCIÓN A LA CIENCIA DE DATOS CON R: PREPARACIÓN DE LOS DATOS Y ANÁLISIS NO SUPERVISADO*, Madrid, García Maroto Editores, 2022.

[11] J. Villardon, «Introducción al análisis de Cluster,» 2007. [En línea]. Available: https://scholar.google.com/citations?view_op=view_citation&hl=es&user=qeq7fbgAAAAJ&citation_for_view=qeq7fbgAAAAJ:LkGwnXOMwfcC.

[12] J. Velazco, MACHINE LEARNING Fundamentos, algoritmos y aplicaciones para, España: Editorial Diaz de Santos, 2020.

[13] H. Chitarroni, «La regresión Logística,» Instituto de Investigación en Ciencias Sociales, Diciembre 2002. [En línea]. Available: https://racimo.usal.edu.ar/83/1/Chitarroni17.pdf.

[14] Amazon, «¿Qué es la regresión logística?,» Amazon, [En línea]. Available: https://aws.amazon.com/es/what-is/logistic-regression/.

[15] F. Cánovas, F. Sarría y F. Gomariz, «MODIFICACIÓN DEL ALGORITMO RANDOM FOREST PARA SU EMPLEO EN,» Research Gate, [En línea]. Available: https://www.researchgate.net/profile/Fulgencio-Canovas-Garcia/publication/304825355_Modificacion_del_algoritmo_Random_Forest_para_su_empleo_en_clasificacion_de_imagenes_de_teledeteccion/links/577be67a08aec3b743366b69/Modificacion-del-algoritmo-Random-Fore.

[16] G. Valenzuela, «Aprendizaje Supervisado: Métodos, Propiedades y Aplicaciones,» Universidad de Málaga, [En línea]. Available: https://riuma.uma.es/xmlui/bitstream/handle/10630/25147/TFG_Aprendizaje_Supervisado_GVG.pdf?sequence=4&isAllowed=y.

[17] J. J. Espinosa-Zúñiga, Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito, vol. XXI, UNAM (Facultad de Ingeniería, México), 2020.

[18] J. M. Fernández Fernández, Las Redes Neuronales Artificiales, A Coruña, 2008.

[19] O. Rainio, J. Teuho y R. Klén, «Evaluation metrics and statistical tests for machine learning,» *Scientific Reports,* vol. 14, nº 6086, 2024.

[20] K. Lastarria, «Propuesta de modelo de predicción de fallos para componente crítico de camión minero, utilizando machine Learning,» Universidad Técnica Federico Santa María, 2024. [En línea]. Available: https://repositorio.usm.cl/handle/11673/57720. [Último acceso: 21 noviembre 2024].

[21] M. Valderrama, «Métodos de pronóstico de fallas en motores diésel de camiones mineros en base a indicadores de degradación probabilísticos,» Universidad de Chile, 2022. [En línea]. Available: https://repositorio.uchile.cl/handle/2250/192663. [Último acceso: 11 noviembre 2024].

[22] P. Samillan y E. Castro, «Framework para la Detección Anticipada de Fallas de Equipos Mediante el Uso de Machine Learning,» LACCEI, [En línea]. Available: https://laccei.org/LACCEI2020-VirtualEdition/work_in_progress/WP283.pdf. [Último acceso: 21 noviembre 2024].

[23] M. L. S. R. S. S. N. A. C. S. José Luis Molina Salgado, Preprocesamiento de datos en el pronóstico de fallos de rodamientos para el mantenimiento predictivo, vol. 28, 2024, p. 1811–1821.

[24] C. A. Orjuela Ortiz, Pronóstico de fallas en redes de distribución de agua potable haciendo uso de herramientas machine learning, Bogotá D.C.: Universidad de Los Andes Departamento de Ingeniería Civil y Ambiental, 2023.

[25] C. Aggarwal, Outlier Analysis, Springer International Publishing, 2017.

[26] B. Iglewicz y D. Hoaglin, How to detect and Handle Outliers, American Society for Quality Control, 1993.