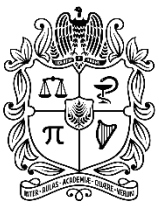


UNIVERSIDAD
NACIONAL
DE COLOMBIA

EVALUACIÓN DE ESTRATEGIAS DE AGRUPAMIENTO NO SUPERVISADAS EN LA DETERMINACIÓN DE PATRONES ASOCIADOS A FALLAS DE SISTEMAS TÉRMICOS EN TRACTOCAMIONES GRANELEROS

Andrés Mauricio Zapata Rincón

Universidad Nacional de Colombia

Facultad de Minas, Departamento de Ciencias de la Computación y la Decisión

Medellín, Colombia

2022

EVALUACIÓN DE ESTRATEGIAS DE AGRUPAMIENTO NO SUPERVISADAS EN LA DETERMINACIÓN DE PATRONES ASOCIADOS A FALLAS DE SISTEMAS TÉRMICOS EN TRACTOCAMIONES GRANELEROS

Andrés Mauricio Zapata Rincón

Tesis presentada como requisito parcial para optar al título de:

Magister en Ingeniería - Analítica

Director:

Ph.D., Alejandro Restrepo Martínez

Codirector:

Ph.D., John Willian Branch Bedoya

Línea de Investigación:

Mantenimiento Predictivo – Análisis de Datos

Grupos de Investigación:

Grupo de Promoción e Investigación en Mecánica Aplicada GPIMA

Grupo de Investigación y Desarrollo en Inteligencia Artificial GIDIA

Universidad Nacional de Colombia

Facultad de Minas, Departamento de Ciencias de la Computación y la Decisión

Medellín, Colombia

2022

Dedicatoria

A mis amados padres

Declaración de obra original

Yo declaro lo siguiente:

He leído el Acuerdo 035 de 2003 del Consejo Académico de la Universidad Nacional. «Reglamento sobre propiedad intelectual» y la Normatividad Nacional relacionada al respeto de los derechos de autor. Esta disertación representa mi trabajo original, excepto donde he reconocido las ideas, las palabras, o materiales de otros autores.

Cuando se han presentado ideas o palabras de otros autores en esta disertación, he realizado su respectivo reconocimiento aplicando correctamente los esquemas de citas y referencias bibliográficas en el estilo requerido.

He obtenido el permiso del autor o editor para incluir cualquier material con derechos de autor (por ejemplo, tablas, figuras, instrumentos de encuesta o grandes porciones de texto).

Por último, he sometido esta disertación a la herramienta de integridad académica, definida por la universidad.

Andrés Mauricio Zapata Rincón

28/02/2022

Fecha

Agradecimientos

A mis padres, Angela Rincón y Mario Zapata quienes me han formado desde el ejemplo, responsabilidad, disciplina y amor.

A mi Director Alejandro Restrepo por sus enseñanzas y el conocimiento brindado para poder potenciar mi camino en el desarrollo del aprendizaje siendo un ejemplo para seguir.

A mi Co director John Willian Branch por su apoyo, acompañamiento y confianza.

A Miguel Andrés Dávila, David Escobar y Jonathan González por su apoyo en el momento de estructurar el proyecto a desarrollar con los conocimientos adquiridos durante el proceso de la maestría.

A Diego Eusse por su apoyo en la implementación de las metodologías de esta tesis.

A la Universidad Nacional de Colombia, por la formación y competencias adquiridas para mi vida profesional con los principios de “Trabajo y Rectitud”.

Y a todas las personas que me expresaron su apoyo durante el tiempo de formulación y desarrollo de esta tesis.

Resumen

EVALUACIÓN DE ESTRATEGIAS DE AGRUPAMIENTO NO SUPERVISADAS EN LA DETERMINACIÓN DE PATRONES ASOCIADOS A FALLAS DE SISTEMAS TÉRMICOS EN TRACTOCAMIONES GRANELEROS.

Esta tesis tiene como finalidad evaluar estrategias de agrupamiento no supervisadas para datos asociados a tractocamiones graneleros, en la detección de patrones de falla en sistemas térmicos. El estudio de estas técnicas es importante en el ámbito del mantenimiento predictivo basado en datos con la implementación de algoritmos de aprendizaje de máquinas que permitan planificar adecuadamente cronogramas de mantenimiento en empresas de transporte de carga. Para el desarrollo de la tesis, se usa como fuente de información los dispositivos de telemetría de los tractocamiones graneleros de una empresa colombiana de transporte de carga que reportan datos en tiempo real de la medición de variables como velocidad, temperaturas, estado de operación del vehículo, entre otras para el año 2020. También se usa el histórico de ingresos a taller de la flota de 116 tractocamiones donde se analizan los ingresos a taller para la intervención de sistemas térmicos. Estos datos son el insumo para la evaluación de las estrategias de agrupamiento propuestas en este trabajo. Los resultados parten desde la obtención de los datos, preparación de estos y análisis descriptivos para implementar técnicas de reducción de dimensionalidad en la información y posteriormente evaluar el comportamiento de algoritmos de agrupamiento para la detección de patrones de falla que se relacionen a daños en sistemas térmicos de los vehículos. Con el desarrollo de este trabajo se encuentra un potencial para el ahorro en costos correctivos de la flota en taller que apunte a una adecuada gestión de la flota en modelos de pago por uso, apalancando la disponibilidad de los vehículos en las operaciones de transporte.

Palabras clave: Mantenimiento Predictivo, Aprendizaje de Máquinas, Sistemas térmicos, Agrupamiento no supervisado, Detección de patrones.

Abstract

EVALUATION OF UNSUPERVISED GROUPING STRATEGIES IN THE DETERMINATION OF PATTERNS ASSOCIATED WITH FAILURES OF THERMAL SYSTEMS IN BULK TRACTORS.

The purpose of this thesis is to evaluate unsupervised clustering strategies for data associated with bulk carrier trucks, in the detection of failure patterns in thermal systems. The study of these techniques is important in the field of data-based predictive maintenance with the implementation of machine learning algorithms that allow proper planning of maintenance schedules in freight transport companies. For the development of the thesis, the telemetry devices of the bulk tractor trucks of a Colombian cargo transport company are used as a source of information, which report data in real time of the measurement of variables such as speed, temperatures, state of operation of the vehicle, among others for the year 2020. The history of workshop entries of the fleet of 116 tractor-trailers is also used, where workshop entries for the intervention of thermal systems are analyzed. These data are the input for the evaluation of the grouping strategies proposed in this work. The results start from obtaining the data, preparing them and descriptive analysis to implement dimensionality reduction techniques in the information and subsequently evaluate the behavior of grouping algorithms for the detection of failure patterns that are related to damage in thermal systems. With the development of this work, there is a potential for savings in corrective costs of the fleet in the workshop that points to an adequate management of the fleet in pay-per-use models, leveraging the availability of vehicles in transport operations.

Keywords: Predictive Maintenance, Machine Learning, Thermal Systems, Unsupervised Clustering, Pattern Detection.

Contenido

Lista de figuras

Pág.

Lista de tablas

Pág.

Lista de abreviaturas

Abreviaturas

Abreviatura	Término
CPU	Central Processing Unit
DBSCAN	Density-based spatial clustering of applications with noise
IA	Inteligencia Artificial
KMEANS	K-Means clustering
PCA	Principal component analysis
SC	Spectral clustering
SOM	Self-organizing maps
t-SNE	T-distributed Stochastic Neighbor Embedding
UMAP	Uniform Manifold Approximation and Projection for Dimension Reduction

1. INTRODUCCIÓN

En este capítulo se realiza una descripción del mantenimiento predictivo con su importancia en entornos industriales, se realiza una descripción del aprendizaje de máquinas y sus bondades en el momento de apoyar procesos donde no se tienen datos etiquetados que indiquen la falla de un sistema mecánico. Posteriormente se enuncian las preguntas de investigación, hipótesis y objetivos de esta tesis.

1.1 Motivación

Los enfoques de mantenimiento predictivo se han aplicado ampliamente en las industrias para manejar el estado mecánico de los equipos industriales. Debido a la transformación digital, es posible recolectar cantidades masivas de datos de condiciones operativas para realizar una detección y diagnóstico de fallas automatizado con el objetivo de minimizar el tiempo de inactividad y aumentar la tasa de utilización de los componentes y aumentar su vida útil restante. El mantenimiento predictivo se basa en la monitorización continua de la máquina. Utiliza herramientas de predicción para medir cuándo son necesarias tales acciones de mantenimiento, por lo que las intervenciones mecánicas se pueden programar de manera planificada, tratando de minimizar la afectación de la operación (Çınar et al., 2020).

La recopilación automatizada de datos de sensores en el vehículo permite el desarrollo de técnicas de inteligencia artificial (IA) para procesos de diagnóstico y pronóstico de sistemas vehiculares con el objetivo de evaluar mejor el estado de los componentes mecánicos, predecir fallas y evaluar la vida residual de los sistemas de vehículos terrestres. Los sensores acumulan y transmiten continuamente información sobre las temperaturas de los subsistemas. Incluso si las temperaturas no son tan altas como para causar un sobrecalentamiento del motor, es posible detectar síntomas inusuales entre la misma flota de vehículos por grupos de control. Los sensores acumulan la información sobre las

tendencias en el comportamiento de las variables para permitir la predicción del tiempo de revisión de los componentes (Ranasinghe et al., 2020; Murakami et al., 2002).

Las técnicas no supervisadas permiten estimar el tiempo de falla y optimizar la programación del mantenimiento. Es importante detectar todos los estados de la máquina con precisión. Si no se detecta un estado en particular, el resultado puede ser catastrófico (Amruthnath, 2019). Existe el reto de la aplicación de estas técnicas y el aprovechamiento de sus beneficios ante la falta de expertos en las técnicas analíticas avanzadas en mantenimiento predictivo.

1.2 Justificación

Los datos en la compañía de estudio para esta investigación presentan una oportunidad desde soluciones analíticas para predecir comportamientos de falla y con el objetivo de tomar decisiones sobre políticas de mantenimiento.

Para el año 2020 los costos totales de mantenimiento de flota ascendieron a $1.298.000.000 donde se presenta interés en estudiar el comportamiento de los sistemas térmicos que en intervenciones correctivas de los paquetes de enfriamiento representaron $59.000.000 como se muestra en la Figura 1-1.

Figura 1- 1: Costos de mantenimiento flota tractocamiones año 2020.

La compañía opera bajo un modelo *Renting* o pago por uso, donde el objetivo es maximizar la disponibilidad de los equipos ya que la facturación depende de esta misma y con

políticas de mantenimiento apoyadas desde el análisis de los datos existe oportunidad para intervenir este fenómeno, en el año 2020 el tiempo en taller asociado a las intervenciones correctivas por sistemas térmicos fueron 10.282 horas y con una facturación dejada de percibir de $1.674.000.000

1.3 Metodología para el desarrollo de la tesis

En el desarrollo de la tesis se aborda el análisis de los daños en sistemas térmicos donde en el capítulo 2 se estudia el estado del arte para técnicas no supervisadas en mantenimiento predictivo de flotas desde una perspectiva enfocada a las técnicas de aprendizaje de máquinas aplicadas. Se realiza una revisión a las técnicas de agrupamiento para detectar estados de falla donde las estrategias se basan en la selección de métodos de reducción de dimensionalidad ante el reto de contar con grandes cantidades de datos multivariados y la necesidad de eliminar características que limiten el rendimiento en los algoritmos de aprendizaje de máquinas.

Posteriormente se analizan las técnicas no supervisadas empleadas en agrupamiento con el análisis respectivo también de las variaciones que poseen para abordar problemas con grandes cantidades de datos.

En el capítulo 3 se presenta la estrategia de tratamiento de los datos, describiéndose los datos desde las fuentes de información como lo son los reportes de los dispositivos de telemetría en la flota y la relación de ordenes de servicio en talleres para reparaciones en sistemas térmicos en el año 2020. Se realiza una introducción a los pasos en el análisis de datos que los describen desde el momento de la recopilación hasta el almacenamiento. Posteriormente, se describen y se analizan los datos de esta tesis para una comprensión general en el problema que se desarrolla.

En el capítulo 4 se presentan las técnicas de reducción de dimensionalidad abordadas junto con el desarrollo de las técnicas de agrupamiento no supervisadas propuestas y que se describen en el capítulo 5.

En el capítulo 6 se presentan los ejercicios de clasificación de estados de falla con las técnicas con la aplicación de las técnicas que mejor se ajustan a la solución de las

preguntas de investigación de esta tesis. Finalmente, los resultados obtenidos con las conclusiones sobre la evaluación de estrategias no supervisadas en la determinación de patrones asociados a fallas de sistemas térmicos en tractocamiones graneleros.

Finalmente, en el capítulo 7 se presentan las conclusiones y recomendaciones que se presentan en el desarrollo de esta tesis.

1.4 Preguntas de investigación

Ya que la detección temprana de fallas permite una correcta implementación de políticas de mantenimiento, las cuales aportan a la disminución de intervenciones correctivas y minimización de periodos de inactividad de los vehículos, que afectan los periodos de facturación de la compañía al trabajar bajo un modelo de Renting o pago por uso, ésta tesis evaluará el comportamiento de estrategias de agrupamiento no supervisadas en determinación de patrones asociados a fallos de sistemas térmicos en tractocamiones graneleros.

Las preguntas de investigación en la tesis son:

1. ¿Es posible detectar y diagnosticar tipos de falla usando métodos de agrupamiento?
2. ¿Cuáles métodos permiten detectar patrones de falla en sistemas térmicos para flotas de tractocamiones graneleros?

1.5 Hipótesis

Los algoritmos no supervisados como DBSCAN, SOM, K-MEANS y SC pueden detectar grupos relacionados con estado normal y de falla para sistemas térmicos de tractocamiones graneleros.

1.6 Objetivos

GENERAL

Evaluar el comportamiento de estrategias de agrupamiento no supervisadas en la determinación de patrones asociados a fallas de sistemas térmicos en tractocamiones graneleros.

ESPECÍFICOS

1. Establecer criterios para la preparación de los datos obtenidos desde los dispositivos de telemetría y entradas a taller, a partir de reportes de falla por sistemas térmicos.

2. Implementar las técnicas de reducción de dimensionalidad como PCA, t-SNE y UMAP en los datos analizados.

3. Diseñar una estrategia de selección de los métodos de agrupamiento con algoritmos no supervisados como K-MEANS, DBSCAN, SOM y SC, con base en el mejor rendimiento en tiempo y desempeño.

4. Evaluar los métodos propuestos en la identificación de patrones de fallas de sistemas térmicos para tractocamiones graneleros.

2. ESTADO DEL ARTE PARA TÉCNICAS NO SUPERVISADAS EN MANTENIMIENTO PREDICTIVO DE FLOTAS

En este capítulo se presenta la revisión del estado del arte en torno al aprendizaje de máquinas y la aplicación en el mantenimiento predictivo de flotas, donde en la sección 2.1 se especifica la importancia del mantenimiento predictivo con aprendizaje de máquinas desde un enfoque basado en disponibilidad del servicio y la integración de tecnologías como los sensores de telemetría en los tractocamiones. Se describen las estrategias utilizadas en la detección de patrones de fallas con estas tecnologías y las técnicas no supervisadas.

En la sección 2.2 se especifican los conceptos generales en el aprendizaje de máquinas supervisado y no supervisado con el fin de identificar la necesidad de desarrollar esta tesis con técnicas no supervisadas y ante la falta de etiquetas que determinan condiciones de falla en los sistemas térmicos. Se realiza un revisión de las técnicas aplicadas en mantenimiento con sus hallazgos más relevantes, logros y retos.

En la sección 2.3 se revisa de una manera general el estado del arte en los años 2011 a 2021 en la gestión de flotas con técnicas no supervisadas, se revisan los estudios que se implementaron entre el año 2015 a 2020 específicamente y los estudios enfocados en técnicas de agrupamiento entre el año 2016 a 2021 con los aportes fundamentales.

Finalmente, en la sección 2.4 se revisan las estrategias de reducción de dimensionalidad y agrupamiento desde su descripción y las implementadas en esta tesis. Se describen las métricas utilizadas para la selección y las etapas requeridas para la evaluación de estrategias. Esta sección se complementa con una revisión de las variaciones de los

algoritmos para trabajar en *Big Data* y la definición de la computación distribuida para abordar los problemas con grandes conjuntos de datos.

2.1 Mantenimiento predictivo desde las soluciones basadas en datos de telemetría.

El mantenimiento predictivo se está volviendo cada vez más importante, especialmente desde que el enfoque cambia del producto a la operación basada en servicios. Requiere, entre otras cosas, poder ofrecer a los clientes garantías de tiempo de actividad. Un beneficio es el ahorro en costos de mantenimientos correctivos, apalancados de correctos planes de mantenimiento general. Es difícil diagnosticar fallas por adelantado en la industria de vehículos debido a la disponibilidad limitada de sensores y algunos de los esfuerzos de diseño. Sin embargo, con el gran desarrollo en la industria automotriz, parece factible hoy analizar los datos del sensor junto con técnicas de aprendizaje de máquinas para la predicción de fallas (Prytz et al., 2013).

Los enfoques de mantenimiento predictivo se han aplicado ampliamente en las industrias para manejar el estado mecánico de los equipos industriales. Debido a la transformación digital, es posible recolectar cantidades masivas de datos de condiciones operativas para realizar una detección y diagnóstico de fallas automatizado con el objetivo de minimizar el tiempo de inactividad y aumentar la tasa de utilización de los componentes y aumentar su vida útil restante. El mantenimiento predictivo se basa en la monitorización continua de la máquina. Utiliza herramientas de predicción para medir cuándo son necesarias tales acciones de mantenimiento, por lo que las intervenciones mecánicas se pueden programar de manera planificada, tratando de minimizar la afectación de la operación (Çınar et al., 2020).

La recopilación automatizada de datos de sensores en el vehículo permite el desarrollo de técnicas de inteligencia artificial (IA) para procesos de diagnóstico y pronóstico de sistemas vehiculares para evaluar mejor el estado de los componentes mecánicos, predecir fallas y evaluar la vida residual de los sistemas de vehículos terrestres. Los sensores acumulan y transmiten continuamente información sobre las temperaturas de los subsistemas. Incluso si las temperaturas no son tan altas como para causar un sobrecalentamiento del motor,

es posible detectar síntomas inusuales entre la misma flota de vehículos por grupos de control. Los sensores acumulan la información sobre las tendencias en el comportamiento de las variables para permitir la predicción del tiempo de revisión de los componentes (Ranasinghe et al., 2020; Murakami et al., 2002).

Los desafíos en el mantenimiento predictivo se basan en la predicción de la confiablidad, implementación de políticas de mantenimiento basadas en datos y la identificación de los sistemas a predecir, sus principales aplicaciones se basan en la administración de flotas por medio de análisis de los sistemas mecánicos como se muestra en la Figura 2-1.

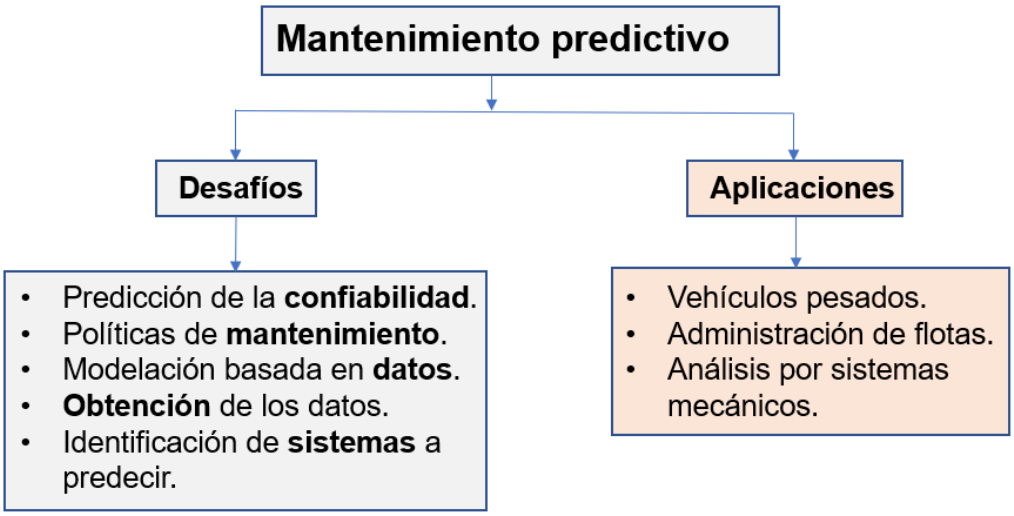


Figura 2- 1: Desafíos y aplicaciones del mantenimiento predictivo.

Como definen (Ibrahim, et. al., 2020), en la metodología propuesta para la detección de fallas con datos de monitoreo satelital y luego del proceso de preparación de los datos se debe realizar el proceso de reducción de dimensionalidad y posteriormente implementar un algoritmo de agrupamiento. En su publicación proponen la reducción de dimensionalidad con el algoritmo t-SNE y la implementación de KMEANS para el agrupamiento donde se encuentra una separación de datos en dos grupos que permiten identificar condiciones de falla. Para el diagnostico de fallas, las técnicas de agrupamiento son una herramienta principal para detectar condiciones anormales y ser soporte para el personal experimentado (Pacella & Papadia, 2020).

2.2 Aprendizaje de máquinas

El aprendizaje de máquinas comprende dos enfoques básicos: aprendizaje supervisado y aprendizaje no supervisado. El aprendizaje supervisado usa datos etiquetados para ayudar a predecir los resultados, mientras que el aprendizaje no supervisado no posee datos etiquetados. En el aprendizaje supervisado el algoritmo aprende del conjunto de datos de entrenamiento haciendo predicciones iterativas sobre los datos y ajustando la respuesta correcta, este método requiere una intervención humana inicial para etiquetar los datos de la manera adecuada. Los modelos de aprendizaje no supervisado funcionan por sí mismos para descubrir la estructura inherente de los datos sin etiquetar, son ideales para la detección de anomalías (Delua, 2021).

Como describen (Perr-Sauer, et. al., 2020), al elegir conjuntos de características independientes de una fuente de datos, se pueden realizar agrupamientos con resultados eficientes eligiendo las estrategias adecuadas. En su experimentación con sus conjuntos de datos de vehículos medianos y pesados en la identificación de condiciones anormales, encuentran resultados óptimos en agrupamiento con la implementación de técnicas de reducción de dimensionalidad PCA, t-SNE y con tres técnicas de agrupamiento (KMEANS, AGGLOMERATIVE y DBSCAN). Para obtener resultados satisfactorios, recomiendan la implementación de agrupamiento basado en características.

En la necesidad de encontrar patrones relevantes para las acciones de mantenimiento en conjuntos de datos que no poseen etiquetas es relevante el uso de algoritmos no supervisados.

2.3 Técnicas no supervisadas aplicadas en la gestión de flotas

En la revisión del estado del arte se presenta de manera general en la Figura 2-2, la línea temporal de los años 2011 al 2021 para casos de aplicación en gestión de flotas para el mantenimiento correctivo y la implementación de métodos de agrupamiento no supervisados para la detección de fallas. Se presentan también los estudios más relevantes en la gestión de flotas desde el año 2015 al 2020 detallados en la Figura 2-3 y

los estudios para aplicaciones de técnicas de agrupamiento desde el año 2016 al 2021 descritos en la Figura 2-4. El análisis de líneas de tiempo para la literatura revisada en la gestión de flotas y estrategias de agrupamiento para la clasificación de fallas se realiza teniendo en cuenta los avances de los últimos cinco años.

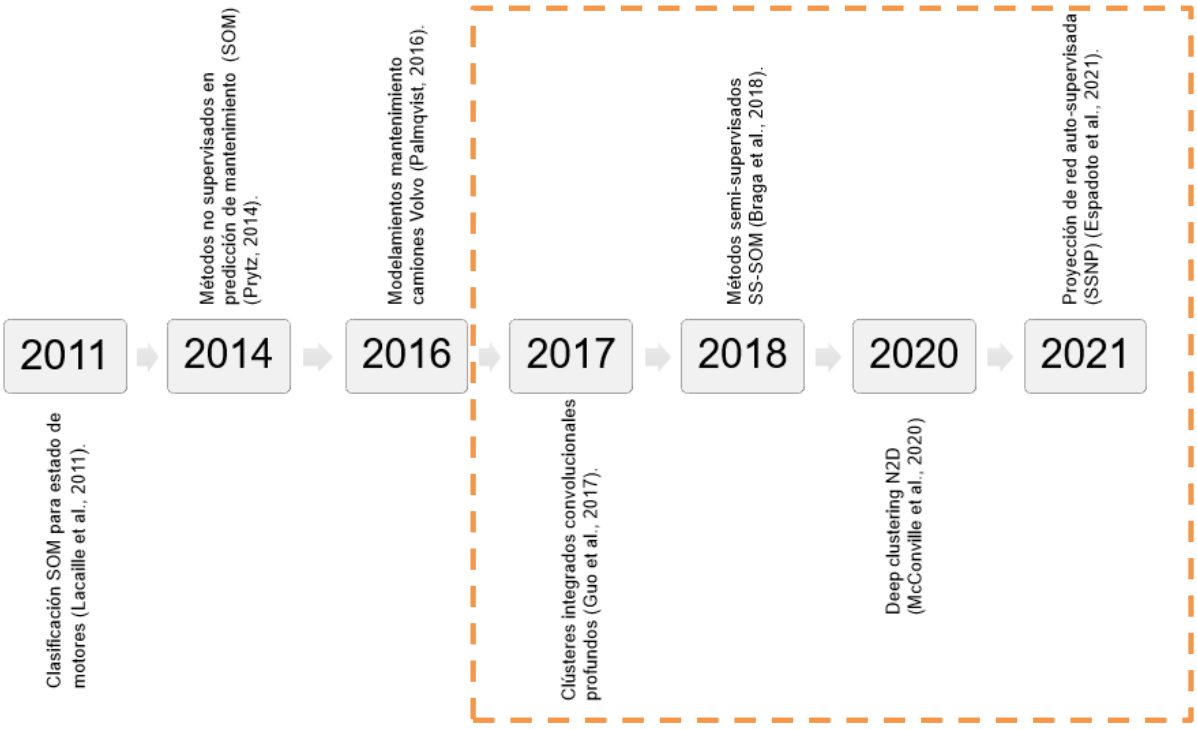


Figura 2- 2: Estado del arte para metodologías aplicadas en flotas y técnicas de agrupamiento.

Se encuentra que entre el año 2011 y 2016 se detallan las implementaciones de técnicas no supervisadas en casos de aplicación para clasificar estados con falla en motores con su predicción de mantenimiento y a partir del año 2017 se encuentran soluciones avanzadas para la clasificación de estados de falla que mejoran los análisis de agrupamiento para los problemas de reconocimiento de patrones tal como se presenta en la Figura 2-2. Los aportes fundamentales de estos trabajos consisten en presentar la evolución en la minería de datos dedicada al mantenimiento de flotas, aplicando modelos predictivos alimentados de los datos que se recopilan de los sensores en los vehículos,

donde la calidad de la información es determinante para el agrupamiento y la determinación de patrones de falla en la flota.

Entre el año 2015 al 2020 se presentan diferentes aplicaciones a la gestión de flotas en torno al mantenimiento predictivo, implementando técnicas que abordan el análisis de grandes conjuntos de datos y la optimización del mantenimiento predictivo con técnicas de agrupamiento y aprendizaje profundo como se detalla en la Figura 2-3.

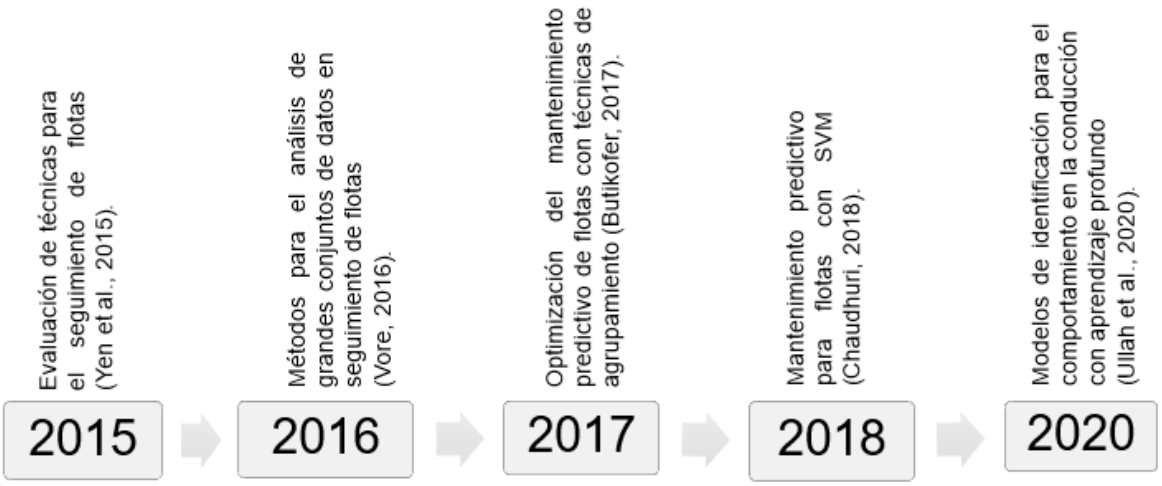


Figura 2- 3: Estado del arte para metodologías aplicadas en flotas desde el año 2015 al 2020.

Los aportes fundamentales se basan en el entendimiento global sobre la recolección de la información entregada por los sensores de los vehículos, procesamiento y análisis de grandes conjuntos de datos para el despliegue de analítica avanzada que permita la toma decisiones en tiempo real para la programación de mantenimiento a través de técnicas de agrupamiento, permitiendo obtener ahorro en costos de reparaciones, eficiencias logísticas y preservación de los vehículos.

Entre el año 2016 al 2021 se presentan los avances de las estrategias de agrupamiento implementadas que se aplican y trabajan en paralelo desde el análisis de grandes conjuntos de datos con base en *Big Data* como se describe en la Figura 2-4.

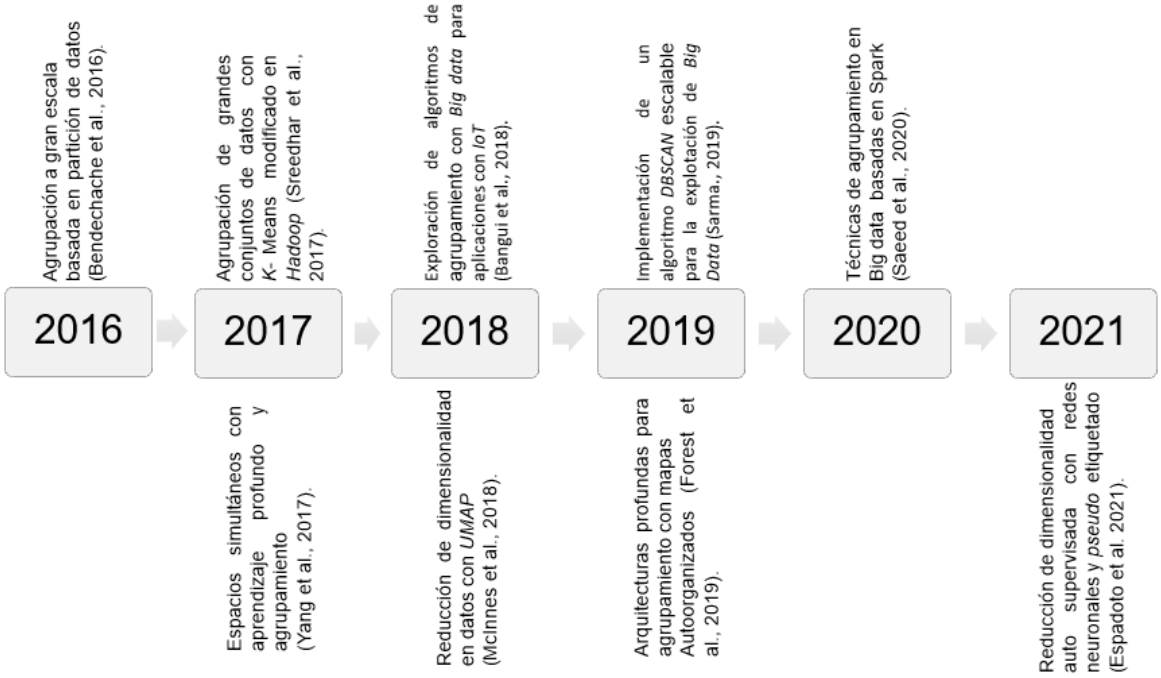


Figura 2- 4: Estado del arte para metodologías aplicadas en técnicas de agrupamiento desde el año 2016 al 2021.

Desde los aportes en las estrategias de agrupamiento se obtiene el aporte del método UMAP, presentando eficiencias en tiempos de computación contra otros métodos de reducción de dimensionalidad como t-SNE y mejor desagregación de los datos en espacios de dimensiones menores. Se proponen métodos como la reducción de dimensionalidad con aprendizaje profundo en conjunto con técnicas de agrupamiento como KMEANS. Exploración y visualización de grandes conjuntos de datos con el algoritmo SOM y su eficiencia computacional.

Para los aportes en *Big data* se encuentran estrategias en el agrupamiento para grandes cantidades de datos en conjunto con plataformas de computación paralela como *Hadoop* y *Spark* reduciendo tiempos de computación, teniendo como línea base la integración entre *Big Data*, el internet de las cosas (*IoT*) y los métodos de agrupamiento.

La instalación de dispositivos de telemetría para el seguimiento del comportamiento de variables mecánicas y de conducción de vehículos, acompañada de políticas de

intervención en la operación pueden generar cambios positivos que apuntan al adecuado funcionamiento de las máquinas y sus componentes mecánicos (Yen et al., 2015).

El análisis de datos con dispositivos telemáticos es fuente primaria para el monitoreo de variables, donde las metodologías y técnicas deben trabajarse en paralelo con *Big Data* (Vore, 2016). Estas estrategias aportan bases fundamentales para la toma de decisiones basadas en datos, aprovechando la gran cantidad de información que se recolecta desde los dispositivos satelitales para mejorar la disponibilidad de equipos y la facturación que generan.

A través de la información de los dispositivos de telemetría se pueden obtener distintas mediciones de variables en los vehículos como velocidad del vehículo, posición del pedal de freno, aceleración, entre otros. La integración de estos datos de manera automática permite el desarrollo de técnicas de inteligencia artificial para sistemas vehiculares (Ullah et al., 2020; Ranasinghe et al., 2020). Con la información que se obtiene, la inteligencia artificial puede apoyar los procesos que requieren de expertos en el fenómeno de interés.

2.4 Estrategias de agrupamiento no supervisadas en la identificación de fallas

Las técnicas de agrupamiento permiten identificar las variables relevantes a los mecanismos de falla, utilizando los resultados del proceso en segmentaciones para el diagnóstico de motores en funcionamiento (Butikofer, 2017). En la identificación de fallas, el agrupamiento detecta valores atípicos en los datos, encuentra estos grupos y muestra los límites que determinan si un punto de datos es un valor atípico o no.

El agrupamiento es una tarea compleja, ya que el resultado se ve afectado por una serie de factores como la adquisición y representación de datos, el preprocesamiento como la reducción de la dimensionalidad con técnicas como PCA, t-SNE y UMAP en conjunto con la experimentación con algoritmos de agrupamiento como KMEANS, DBSCAN, SOM y SPECTRAL CLUSTERING (Yang et al., 2017).

Finalmente, para valorar el desempeño del algoritmo de agrupamiento se requiere el uso de métricas, donde algunas de ellas son *Silhouette score* y *Variation of information*. La métrica de *Silhouette score* es una medida de la compactación de cada grupo y la limpieza de la separación entre los mismos y la métrica *Variation of information* es una medida de la información mutua entre dos agrupaciones, la medición sólo es significativa con respecto a una agrupación de línea de base y es sólo el orden de la métrica (no necesariamente la magnitud) lo que se puede interpretar (Perr-Sauer et al., 2020). En la Figura 2-5 se presentan las etapas para la detección de patrones de falla con la implementación de algoritmos de agrupamiento.

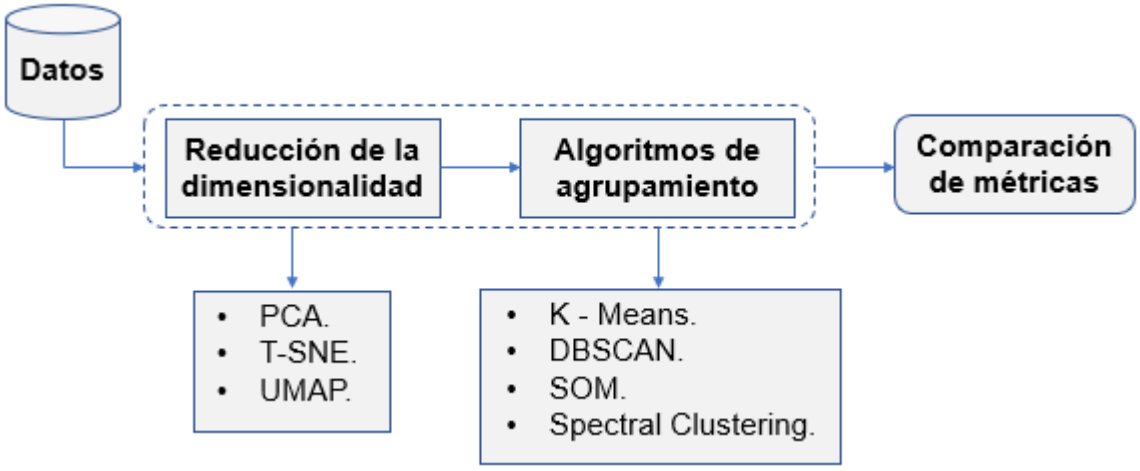


Figura 2- 5: Etapas para la detección de patrones de falla (Modificado de Perr-Sauer et al., 2020).

2.4.1 Reducción de la dimensionalidad

PCA es un método bien conocido para reducir la dispersión de las mediciones de los sensores y su dimensionalidad. El PCA da como resultado la transformación de las variables originales en una pequeña cantidad de características (componentes principales). El PCA reemplaza las medidas con un número menor de puntos que son una combinación lineal de los datos originales y considera estos nuevos puntos como variables escalares (Pacella et al., 2020).

Sin embargo, PCA es un método lineal y no funciona bien en los casos en que las relaciones no son lineales. Afortunadamente, existen múltiples métodos de aprendizaje

alternativos no lineales, y pueden clasificarse por su enfoque en los casos en los que las relaciones no son lineales y encontrar una estructura local o global (Allaoui et al., 2020).
El algoritmo de reducción de dimensionalidad t-SNE ayuda a identificar puntos de datos similares, la información de agrupamiento se mantiene y las estructuras locales de las variables independientes se agrupan por separado. Las estructuras locales se conservan cuando se reducen las dimensiones mediante t-SNE pero no se conservan las propiedades globales de los datos (Chekkala, 2020).

UMAP es un método propuesto, el cual presenta un mejor desempeño para preservar la estructura local y global de los datos. Comparado con t-SNE tiene amplias ventajas, entre ellas la capacidad para abordar bien grandes conjuntos de datos, mientras t-SNE típicamente posee dificultades con esto (Allaoui et al., 2020).

2.4.2 Técnicas de agrupamiento no supervisadas

Se introdujo una variante del algoritmo original de K-MEANS, denominada K-MEANS ++. Esta variante produce una mejor clasificación junto con una reducción del parámetro SSW y, por lo tanto, la compacidad de los grupos en comparación con los grupos iniciales. Cuando se compara con el algoritmo original de K-MEANS, el K-MEANS ++ muestra una mejor precisión de clasificación y una convergencia más rápida. El algoritmo de K-MEANS requiere la información preliminar sobre el número de grupos y, por tanto, el número de centroides alrededor de los cuales agregar los puntos más cercanos. Esta característica hace que el algoritmo K-MEANS sea particularmente atractivo en problemas de clasificación no supervisados (Pacella et al., 2020).

“El algoritmo DBSCAN es un método de agrupamiento basado en densidad, la característica más importante es la habilidad para detectar formas arbitrarias y ruido. Define Heidari et. al. (2019)”. Revisando el artículo, dentro de sus variaciones se encuentran las siguientes mejoras al algoritmo:

- OPTICS: Enfrenta una de las debilidades del DBSCAN la cual es el problema de detectar grupos significativos en datos con densidad variada. Es un algoritmo para

encontrar los grupos basados en densidad para datos espaciales creando una ordenación aumentada de los puntos de datos.

- AUTOEPSDBSCAN: Propone un algoritmo mejorado que automáticamente selecciona los parámetros de entrada. Puede detectar grupos con densidad variada, con diferentes formas y tamaños a partir de una gran cantidad de datos que contienen ruido y valores atípicos, requiere solo una entrada y ofrece una mejor salida que el algoritmo DBSCAN.

- VMDBSCAN: Una mejora del algoritmo DBSCAN que detecta grupos de diferentes formas y tamaños que difieren en densidad local. Este algoritmo primero encuentra los núcleos de cada grupo en los grupos generados después de aplicar DBSCAN.

- MR-DBSCAN: Es la implementación de DBSCAN distribuido con *Map/Reduce* sobre la plataforma HADOOP. Se centra en equilibrar la carga en conjuntos de datos a gran escala y velocidad eficiente para grandes conjuntos de datos. Consta de tres niveles los cuales son la partición de datos, agrupación local y fusión global.

Sin embargo, el paradigma en *Big data* declara que el algoritmo DBSCAN no es muy eficiente en el análisis de grandes volúmenes de datos mientras se ejecute en una sola máquina.

Adicionalmente, existe otra familia específica de algoritmos de agrupación, denominados mapas auto organizados, *Self Organizing Maps* (SOM) que realizan agrupaciones y visualización simultáneas mediante la proyección de datos de alta dimensionalidad en un mapa de baja dimensión (normalmente bidimensional con fines de visualización) que tiene una topología de cuadrícula. La cuadrícula está compuesta por unidades, también llamadas neuronas o células. Cada unidad de mapa está asociada con un vector prototipo del espacio de datos original (también llamado vector de código). Los algoritmos de mapas autoorganizados imponen una restricción topológica en el mapa, de modo que las unidades vecinas en el mapa corresponden a vectores prototipo que están cerca en el espacio original de alta dimensión (Forest et al., 2019).

Los algoritmos de agrupamiento espectral son herramientas eficientes en el procesamiento de señales para agrupación de conjuntos de datos recolectados por sistemas multisensores para el diagnóstico de fallas. SPECTRAL CLUSTERING (SC) es un enfoque eficaz. El método ha demostrado ser robusto con respecto a la geometría de los grupos, el ruido y los valores atípicos. El método es particularmente adecuado para la agrupación en un entorno de alta dimensión (Pacella et al., 2020).

Debido a que la detección temprana y precisa de fallas en las máquinas son cruciales para minimizar el tiempo de inactividad, las técnicas de monitoreo de procesos que han sido más efectivas en la práctica se basan en el análisis de datos históricos de procesos, en donde el agrupamiento espectral por núcleo, *Kernel Spectral Clustering* (KSC) tiene un buen desempeño para procesar datos históricos del sensor, distinguiendo en ellos condiciones de funcionamiento normales y situaciones anormales (Langone, 2013).

El SPECTRAL CLUSTERING se adapta a la variedad de los datos para manejar datos incompletos para la agrupación, siendo capaz de agrupar conjuntos de datos incompletos de una manera más sólida en comparación con los enfoques existentes (Løkse et al., 2017).

Otra estrategia aplicada a los agrupamientos, son las redes neuronales artificiales, donde las redes neuronales artificiales profundas han demostrado ser capaces de aprender representaciones para mejorar el agrupamiento de datos. En varias aplicaciones se cuenta con rendimientos prometedores, los algoritmos de agrupación profunda existentes no aprovechan bien las redes neuronales artificiales convolucionales o no preservan considerablemente la estructura local de distribución de generación de datos en el espacio de características aprendidas. Se proponen algoritmos integrados de agrupación convolucionales profundos, donde se desarrollan estructuras de codificadores automáticos convolucionales para aprender las funciones integradas en todo el proceso de agrupamiento (Guo et al., 2017).

2.4.3 Algoritmos de agrupamiento en grandes cantidades de datos

El agrupamiento es una técnica de minería de datos esencial como un método de análisis en *Big Data*. El principio de esta técnica es crear grupos o subconjuntos que contengan los objetos con características similares. Permite el análisis simple encontrando estructuras en los datos y clasificando cada dato acorde a su naturaleza (Bangui et al., 2018).

Por el reciente crecimiento de la cantidad de información, los métodos tradicionales de agrupamiento son altamente requeridos. Las Investigaciones proponen diseños para métodos de agrupamiento que aumentan los beneficios en plataformas para *Big Data*, tal como Apache Spark, la cuál es diseñada para una rápida distribución masiva en procesamiento de datos (Saeed et al., 2020; Espadoto et al., 2021).

En *Big data* se presentan grandes beneficios para el procesamiento y almacenamiento de grandes cantidades de información. Los actuales algoritmos de agrupamiento requieren soluciones escalables para gestionar grandes conjuntos de datos. Se presentan avances como *K-means Hadoop MapReduce* (KM-HMR), basados en el concepto de *Map Reduce* y la implementación del método estándar de KMEANS (Sreedhar et al., 2017).

Con respecto al método DBSCAN y como mencionan (Sarma et al., 2019), proponen un método basado en micro agrupamiento para conjuntos de grandes datos y mejorar los tiempos de procesamiento para este algoritmo y basados en el concepto de la técnica DBSCAN, esta técnica es nombrada µDBSCAN, la cual identifica los puntos centrales incluso sin computar las consultas entre los puntos vecinos del espacio, permitiendo ser un instrumento para la reducción de tiempos de procesamiento del algoritmo.

Según (Bendechache et al., 2016) presenta una técnica distribuida de agrupamiento para tratar eficientemente la generación de resultados locales y la generación de modelos globales por agregación capaz de analizar los conjuntos de datos con técnicas de agrupamiento como KMEANS y DBSCAN. La fase de agregación es diseñada tal que los agrupamientos finales son compactos y con resultados de alta calidad mientras que el

tiempo de procesamiento promedio es eficiente y también la utilización de memoria en la infraestructura computacional.

2.4.4 Computación distribuida

Las técnicas de agrupamiento son altamente atractivas para la extracción e identificación de patrones de conjuntos de datos. Sin embargo, su aplicación a grandes conjuntos de datos presenta retos por: la alta dimensionalidad de datos, la heterogeneidad y la complejidad de los algoritmos. En este contexto, el agrupamiento distribuido brinda una alternativa a los retos en Big Data en términos de volumen, variedad y velocidad. (Bendechache et al., 2016).

Con el rápido desarrollo de la era de la información, grandes cantidades datos son producidos diariamente por lo que una simple computadora no puede procesar estos volúmenes de información, nuevas tecnologías son requeridas para almacenar y extraer información de estos volúmenes de datos en el marco de Big data. La computación en la nube es una tecnología potente que habilita el acceso a redes o recursos computacionales compartidos. Puede ser definida como un sistema distribuido y paralelo producto de computadores interconectados (Heidari, 2019).

La ventaja básica de una arquitectura de memoria compartida es doble: primero, hay un costo bajo asociado con la comunicación entre procesos. Es decir, los procesadores pueden usar la memoria compartida físicamente para pasar información a otros procesadores. En segundo lugar, los sistemas de memoria compartida evitan la necesidad de crear réplicas redundantes de la comunicación entre procesos de datos (Kucukyilmaz, 2014).

Finalmente, el problema que se trabajará en esta tesis se basa en la detección de patrones de falla para sistemas térmicos en tractocamiones, y dado a que no se cuenta con etiquetas de clase o grupos, se decide abordar este problema desde el estudio e implementación de algoritmos de agrupamiento no supervisados.

2.5 Conclusiones

- En este capítulo se hace una revisión del estado del arte donde se encuentra la importancia de la medición de variables mecánicas y de operación en las flotas a través de dispositivos de telemetría, con la recolección de esta información se genera oportunidad en el desarrollo de técnicas de aprendizaje de máquinas que permitan estudiar la detección de fallas con técnicas no supervisadas. La detección anticipada de fallas apalanca un desarrollo en la programación de mantenimientos, garantizando tiempo de actividad en las flotas, beneficios en la disminución de intervenciones correctivas y aumento en la vida útil de los equipos.

- Ante la ausencia de datos que se puedan obtener con las etiquetas que indiquen si se presenta un estado de falla, las técnicas de agrupamiento no supervisadas juegan un papel fundamental en esta condición para la detección de anomalías.

- Para el desarrollo de las técnicas de agrupamiento en la detección de fallas, es recomendado el proceso de reducción de dimensionalidad previamente, teniendo en cuenta que en la recolección de información se presentan conjuntos de datos con gran cantidad de variables, por lo que es necesario capturar la información con mayor varianza explicada en conjuntos de menores dimensiones. Esto permitirá tener mejores resultados con los algoritmos de agrupamiento y mejorando su eficiencia computacional.

- Ante la gran cantidad de información que se puede obtener con los dispositivos de telemetría es necesario conocer las técnicas de *Big data* empleadas en conjunto con plataformas que permitan tener una computación eficiente por estos volúmenes de información.

- Para los estudios de estas técnicas es fundamental contar con las bases de datos para los estudios de flotas, explorando las técnicas en diferentes arquitecturas de computación, teniendo en cuenta la gran cantidad de datos que se puedan obtener y en ocasiones no ser eficiente desarrollar las técnicas en máquinas locales, requiriendo el uso de servicios de computación en la nube.

3. ANÁLISIS EXPLORATORIO DE DATOS PARA TRACTOCAMIONES GRANELEROS

En el desarrollo de esta tesis se realiza un análisis exploratorio de los datos para comprender los comportamientos de las variables a estudiar, desde la recolección de la información, el preprocesamiento y almacenamiento con el fin de dimensionar el problema que se tiene en términos de volumen y variedad de la información que permita preparar toda la información para la posterior aplicación de las técnicas de reducción de dimensionalidad y agrupamiento.

Se describen en este capítulo las fuentes de datos para el desarrollo de la tesis, las cuales son especificadas en la sección 3.1 Fuentes de datos, describiendo las fuentes de información desde dispositivos de telemetría e ingreso de la flota a taller durante el año 2020. Posteriormente, en la sección 3.2 Etapas en la preparación de los datos, se enuncian los pasos generales para tener en cuenta desde la recolección de la información hasta el almacenamiento de esta.

En la sección 3.3 Análisis descriptivos de los datos, se realiza un exploratorio sobre los datos que se obtienen de los dispositivos de telemetría y características como la resolución o frecuencia de la información recolectada. En la sección 3.4 Distribuciones estadísticas de las variables de telemetría, se revisa por medio de una prueba de escritorio en la fuente de información, el comportamiento estadístico de las variables obtenidas por los dispositivos de telemetría.

Para la sección 3.5 Análisis diferencia de medias para variables de temperatura del aceite y del refrigerante del motor, se analiza el comportamiento de estas variables en un caso específico antes de ingresar a taller el tractocamión de la prueba de escritorio definida y luego de salir de taller. Finalmente, en la sección 3.6 Estrategias para la preparación de los datos de tractocamiones graneleros, se describe la dificultad con los datos al no tener la misma temporalidad la recolección de la información de las variables, problema que se aborda desde el método de la interpolación de datos con su validación en los comportamientos estadísticos resultantes.

Con el análisis estadístico de las variables se genera un comprensión de manera general sobre los datos de esta tesis, se presenta una oportunidad con los hallazgos en las distribuciones de las variables para utilizar como futura herramienta de modelación en el comportamiento de las variables.

Al final del capítulo se realiza una descripción de las estrategias desarrolladas para poder trabajar con grandes datos en esta tesis desde las limitaciones de infraestructura computacional y su solución desde la experimentación por muestras de datos escalables.

3.1 Fuentes de datos

En el desarrollo de esta tesis se implementa la extracción y análisis de los datos de los sensores satelitales de la flota de 116 tractocamiones graneleros, donde se recopila información de los distintos valores de las variables mecánicas y de operación de los vehículos para el año 2020. También se incluye dentro de las fuentes de datos los reportes de ingresos a taller en el año 2020.

3.1.1 Datos recopilados de los sensores satelitales

Los vehículos cuentan con dispositivos de telemetría monitoreando en tiempo real variables asociadas a la operación de estos, tales como temperatura, velocidad, consumo de combustible, voltaje, carga del motor, entre otras.

Para los datos a procesar, se cuenta con 17.3 GB de información correspondientes a todos los reportes de la telemetría en el año 2020, para 18 variables en la flota de 116 tractocamiones graneleros. La descarga se realiza a través de la API (*Application Programming Interfaces*) del proveedor de la plataforma satelital y el software Python ya que directamente desde la plataforma de telemetría no es posible generar la descarga masiva de la información.

En el proceso de obtención de los datos, el algoritmo realiza 33 iteraciones de descarga para cada una de las 18 variables en pasos de tiempo de 10 días y garantiza la disponibilidad de toda la información para el desarrollo de esta tesis. Para las 18 variables

se cuenta con 16 variables continuas y 2 variables discretas que indican si la variable a medir se encuentra activa o no como se presenta en la Figura 3-1.

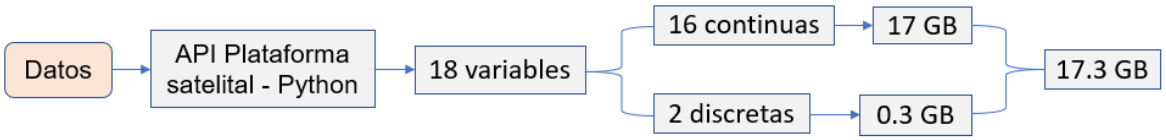


Figura 3- 1: Proceso de extracción de los datos.

En la Tabla 3-1, se presenta la descripción de las variables obtenidas en el proceso de descarga masiva de los datos, en seis columnas. La primera representa el número de la variable entre uno a 18, la segunda indica el nombre de la variable, donde se obtienen variables que en general representan mediciones de temperaturas, niveles de combustible, velocidades, entre otras. La tercera columna indica el peso en *Gigabytes* del conjunto de datos por variable. La cuarta columna indica si la variable es continua (valores numéricos) o variable discreta (Datos categóricos, 0-1). En la quinta columna se detallan las unidades de medición de cada variable y por último, la columna seis indica una descripción general de las variables.

Tabla 3- 1. Descripción de variables obtenidas desde los dispositivos de telemetría.

Número variable	Variable	Tamaño (GB)	Categoría de la variable	Unidades	Descripción
1	Combustible de viaje utilizado	0.0793	Continua	[Litros]	Litros consumidos de combustible entre cada una de las mediciones del sensor.
2	Combustible de viaje utilizado en ralentí	0.0837	Continua	[Litros]	Litros consumidos de combustible cuando el vehículo se encuentra en ralentí entre cada una de las mediciones del sensor.
3	Combustible total utilizado (desde la instalación del dispositivo de telemática)	7.61	Continua	[Litros]	Cantidad acumulada de litros en el consumo de combustible en el vehículo en función del tiempo.

Tabla 3-1: (Continuación)

Número variable	Variable	Tamaño (GB)	Categoría de la variable	Unidades	Descripción
4	Combustible total utilizado en ralentí (desde la instalación del dispositivo de telemática)	0.17	Continua	[Litros]	Cantidad acumulada de litros en el consumo de combustible en el vehículo cuando se encuentra en ralentí en función del tiempo.
5	*Ignition*	0.147	Discreta	[0-1] Apagado/Encendido	Estados del motor cuando se encuentre encendido o apagado.
6	Nivel de combustible	0.195	Continua	[Litros]	Nivel de combustible en litros del vehículo.
7	Nivel de DEF	0.292	Continua	[Litros]	Nivel de urea en litros.
8	Nivel de refrigerante	0.277	Continua	[%]	Nivel del refrigerante en porcentaje para el vehículo.
9	Odómetro	0.466	Continua	[Metros]	Distancia total recorrida en metros para el vehículo en función del tiempo.
10	Voltaje dispositivo de telemetría	0.504	Continua	[Voltios]	Voltaje del dispositivo de telemetría en el vehículo.
11	Temperatura de aceite del motor	0.609	Continua	[°C]	Temperatura del aceite del motor en función del tiempo.
12	Temperatura del refrigerante del motor	1.17	Continua	[°C]	Temperatura del refrigerante del motor en función del tiempo.
13	Temperatura exterior	0.126	Continua	[°C]	Temperatura exterior en función del tiempo.
14	Tensión de arranque	0.788	Continua	[Voltios]	Voltaje al iniciar el vehículo durante los primeros 10 segundos.

Tabla 3-1: (Continuación)

Número variable	Variable	Tamaño (GB)	Categoría de la variable	Unidades	Descripción
15	Tiempo de funcionamiento del motor	0.349	Continua	[Segundos]	Tiempo de uso del motor.
16	Vehículo activo (ralentí o en movimiento)	0.186	Discreta	[0-1] Movimiento/Ralentí	Estado del vehículo, indicando si se encuentra encendido y moviéndose o encendido y detenido.
17	Velocidad del motor	3.4	Continua	[RPM]	Velocidad del motor en función del tiempo.
18	Velocidad en carretera del motor	0.841	Continua	[Km/h]	Velocidad del vehículo en función del tiempo.

El tamaño total de los archivos corresponde a 17.3 Gb de información en archivos con formato CSV (*comma-separated values*).

3.1.2 Datos recopilados de órdenes de servicio en taller

Se cuenta con los documentos de los registros de las ordenes de servicio de ingresos a taller de los vehículos en el año 2020, detallando por cada orden de servicio el tiempo de atención, relación de repuestos suministrados para cada sistema del vehículo intervenido junto con los costos asociados como facturación no percibida, costo de reparaciones, tiempos en taller y cantidad de ordenes atendidas que se detallan en la Figura 3-2.

Figura 3- 2: Costos por órdenes de servicio de mantenimiento año 2020.

En la Figura 3-3 Se detallan los subsistemas reportados correspondientes al 80% de los costos de mantenimientos correctivos en el año 2020 en los componentes de los sistemas térmicos de los tractocamiones graneleros.

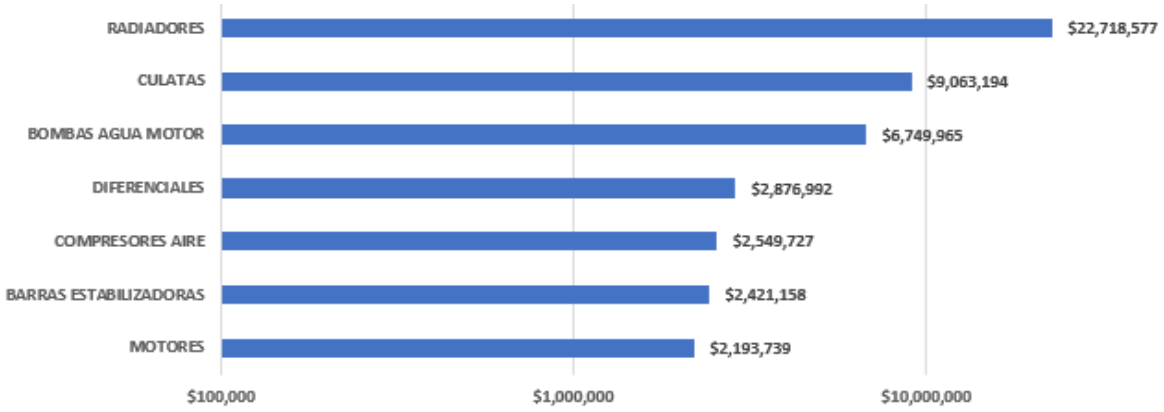


Figura 3- 3: Costos por subsistemas intervenidos año 2020.

Durante el año 2020 se presentaron en total 2170 órdenes de servicio donde se registraron 173 que incluyeron intervenciones para los sistemas térmicos, correspondientes al 8% de las órdenes de servicio totales.

En la Figura 3-4 se presenta el histograma para los costos de intervención en sistemas térmicos. En el eje “x” se encuentra el costo por cada intervención en taller asociada a

sistemas térmicos y en el eje “y” se encuentra la frecuencia o la cantidad de ordenes dentro del conjunto de las 173 facturas asociadas.

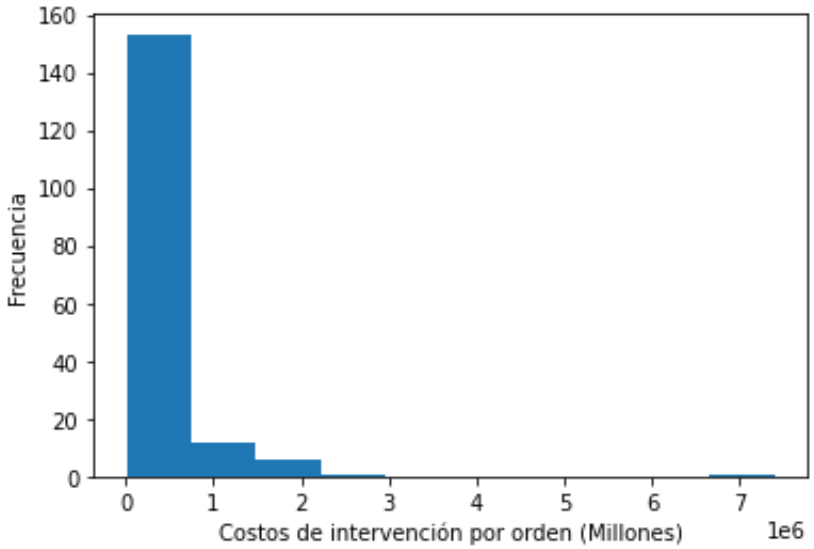


Figura 3- 4: Histograma del costo de intervención en sistemas térmicos por orden de servicio.

Se presenta en la Tabla 3-2 la estadística descriptiva para los costos anteriores.

Tabla 3- 2. Estadística descriptiva costos de intervención sistemas térmicos.

Conteo		173
Media	$	3,430,000
Desviación estándar	$	687,000
Mínimo	$	86,200
25%	$	104,460
50%	$	156,690
75%	$	249,080
Máximo	$	7,402,368

Con el análisis anterior se concluye que de manera descriptiva el 75% de las ordenes de intervención en sistemas térmicos alcanzan un costo de $249 mil pesos colombianos. Donde se alcanza un valor máximo de $7 millones de pesos asociado a una de las intervenciones más costosas que se presenta.

3.2 Etapas en la preparación de los datos

En la Figura 3-5, se resumen las seis etapas en el procesamiento de los datos como lo definen (PeerXP Team, 2017).

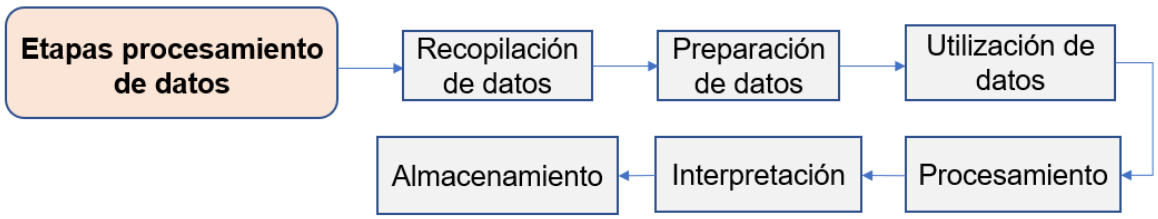


Figura 3- 5: Etapas para el procesamiento de datos.

3.2.1 Recopilación de datos

La recopilación de datos es el primer paso de las etapas. Es crucial la calidad de los datos recolectados ya que esto tiene un gran impacto en los resultados finales donde se toman decisiones con base en los hallazgos que proporcionan los datos procesados.

3.2.2 Preparación de datos

La preparación de datos, a menudo llamada «preprocesamiento», es la etapa en la que los datos se construyen en *datasets* de una o más fuentes de datos para ser explorados y procesados. Analizar datos que no son cuidadosamente preparados pueden producir resultados erróneos que son altamente dependientes de la calidad de la información con que se preparan.

3.2.3 Utilización de datos

Es la tarea donde los datos verificados son codificados en una forma que pueda interpretar una máquina de procesamiento o aplicación. Este proceso requiere de velocidad y precisión. Los datos necesitan seguir una sintaxis formal ya que se requiere una gran potencia de procesamiento para desglosar los datos.

.

3.2.4 Procesamiento

El procesamiento es cuando los datos se someten a diversos medios y técnicas que utilizan algoritmos de aprendizaje automático e inteligencia artificial para generar una salida o interpretación sobre los datos. El proceso puede estar compuesto por múltiples hilos de ejecución que ejecutan instrucciones simultáneamente, según el tipo de datos.

3.2.5 Salida/Interpretación de datos

La salida y la interpretación es la etapa donde la información procesada ahora se transmite y se muestra al usuario. La salida se presenta a los usuarios en varios formatos de informe, como informes gráficos, audio, video o visores de documentos. La salida debe interpretarse para que pueda proporcionar información significativa que guiará las decisiones futuras de la empresa.

3.2.6 Almacenamiento de datos

El almacenamiento es la última etapa en el ciclo de procesamiento de datos, donde los datos y los metadatos (información sobre los datos) se guardan para uso futuro. La importancia de este ciclo es que permite acceso y recuperación rápida de la información procesada.

3.3 Análisis descriptivo de los datos

Se realiza una exploración sobre el conjunto de datos para cada variable, con el objetivo de identificar la resolución o la cantidad de registros que se puede tener por día en un tractocamión. Se toma como referencia el tractocamión identificado con caso número uno con la información recopilada de los sensores satelitales en el año 2020. Dentro de las fechas de muestreo se seleccionan aleatoriamente dos días laborales correspondientes a las fechas de 12 de febrero y 22 de septiembre de 2020, se seleccionan estas dos fechas ya que corresponden a meses en los que no se tuvo afectación en la operación de transporte efecto de la pandemia Covid-19.

En la tabla 3-3 se relacionan la cantidad de registros para cada día, seleccionados en los meses de febrero y septiembre para el año 2020. En la primer columna se especifica el

nombre de la variable que se revisa en cuanto a la cantidad de datos que se capturan por cada día, en la segunda columna se indica la cantidad de datos capturados por los dispositivos de telemetría el día 12 de febrero de 2020 al igual que en la tercera columna se indica la cantidad de datos obtenidos el día 22 de septiembre de 2020.

Tabla 3- 3. Número de registros por variable tractocamión caso uno.

	Mes-Día	
Variable	**Febrero 12**	**Septiembre 22**
Combustible de viaje utilizado	9	10
Combustible de viaje utilizado en ralentí	9	10
Combustible total utilizado (desde la instalación del dispositivo de telemática)	766	751
Combustible total utilizado en ralentí (desde la instalación del dispositivo de telemática)	15	15
Ignition	20	21
Nivel de combustible	15	19
Nivel de DEF	47	51
Nivel de refrigerante	42	0
Odómetro	56	73
Voltaje dispositivo de telemetría	48	43
Temperatura de aceite del motor	62	93
Temperatura del refrigerante del motor	71	277
Temperatura exterior	10	15
Tensión de arranque	140	78
Tiempo de funcionamiento del motor	35	51
Vehículo activo (ralentí o en movimiento)	20	21
Velocidad del motor	272	552
Velocidad en carretera del motor	140	78

Del anterior análisis de la cantidad de datos que se capturan por cada día y por cada variable, se encuentra que la captura de la información desde los dispositivos de telemetría no presenta una temporalidad definida, por lo que en cada instante de tiempo no se tiene la información de las 18 variables de la tesis y a la hora de consolidar estos datos se tendrá información faltante en el *dataset*, siendo una de las principales dificultades en la tesis, por lo que se aborda con técnicas de interpolación por variable para poder tener toda la información requerida en los posteriores experimentos con los algoritmos no supervisados.

Para el análisis estadístico que se presenta en la siguiente sección, se selecciona uno de los tractocamiones de la flota, el caso número dos nombrándolo el vehículo de prueba de escritorio, siendo el equipo con más registros de falla según la etiqueta de *Fault Engine Protection Torque Derate* la cual indica que el vehículo debe ser atendido prontamente en taller y el computador del motor limita la potencia de este, con los datos de los meses de julio y agosto del año 2020. En la Tabla 3-4 se detalla la cantidad de registros de falla para los siete casos de tractocamiones con más reportes.

Tabla 3- 4. Cantidad de reportes de falla por caso – número de tractocamión, año 2020.

Número de caso	2	3	4	5	6	7	8
Cantidad de reportes de falla	37488	36379	30314	27597	18960	18768	17264

La descripción estadística de los datos se realiza por medio del *software* Python, donde con este *software* se determina para cada variable si se cumplen pruebas de normalidad o determinar cuál es la distribución estadística que mejor se ajusta al comportamiento de los datos de cada variable en un conjunto de 28 funciones de distribución como se relaciona en la Figura 3-6.

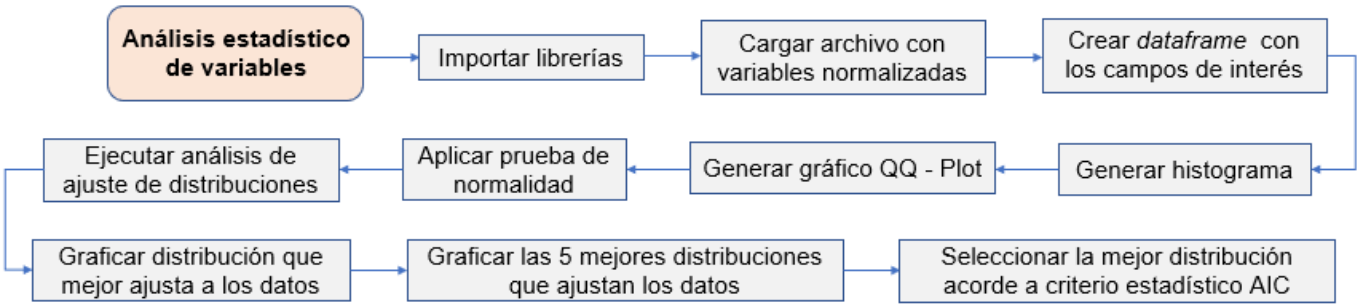


Figura 3- 6: Pasos para análisis estadístico por variable.

En la sección 3.4 se realiza una experimentación con el tractocamión con caso número 2 con el fin de determinar cómo se distribuyen estadísticamente las variables según los datos obtenidos por los dispositivos de telemetría que en trabajos futuros puede permitir modelar datos según los comportamientos estadísticos encontrados.

3.4 Distribuciones estadísticas de las variables de telemetría

En esta sección se presentan los análisis correspondientes sobre el tractocamión caso 2 e identificar cuáles son las distribuciones de probabilidad que mejor se ajustan al comportamiento de los datos, esto con el fin de poder modelar datos para trabajos futuros que faciliten la implementación de algoritmos de aprendizaje de máquinas.

Se relaciona a continuación, el análisis estadístico para dos variables del conjunto de datos donde se presenta el gráfico del histograma, el gráfico *QQ-Plot*, la prueba de hipótesis de normalidad de los datos y las cinco distribuciones que mejor se ajustan al comportamiento de los datos acorde al criterio estadístico AIC con valor menor.

Para determinar la prueba de hipótesis de normalidad se establecen las siguientes hipótesis:

$$H_0 = \ Los\ datos\ poseen\ una\ distribución\ normal$$
$$H_1 = \ Los\ datos\ no\ poseen\ una\ distribución\ normal$$

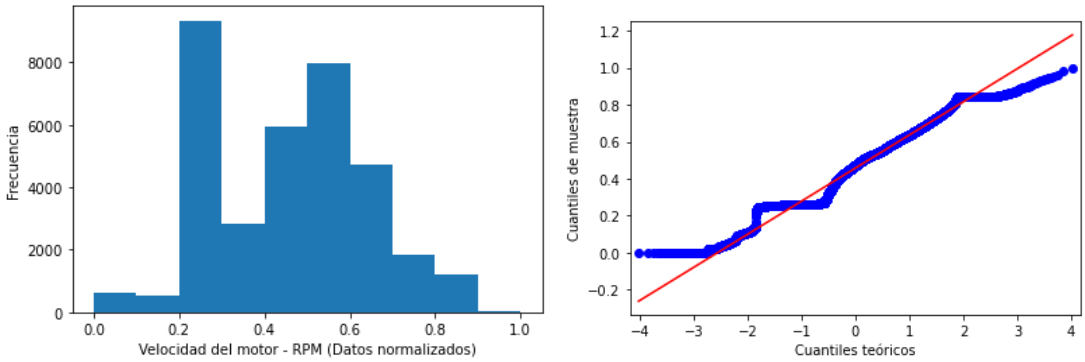
Donde:

H_o: $Hipótesis\ nula$ y H_1: $Hipótesis\ alterna$

Si valor p < 0.05 se rechaza la hipótesis nula y los datos no poseen una distribución normal.

3.4.1 Análisis variable velocidad del motor

La cantidad de registros de la prueba de escritorio para la variable velocidad del motor (RPM) corresponde a 34995 registros. En la Figura 3-7 se relacionan el histograma de la velocidad (a) del motor y el gráfico *QQ-Plot* (b).



(a) (b)

Figura 3- 7: Histograma variable velocidad del motor (a) y gráfico *QQ-Plot* (b).

Con la prueba de normalidad de *Shapiro-Wilk* se obtiene en el estadístico el valor de 0.965 y el valor p de 0.000, por lo que se rechaza la hipótesis nula y la variable velocidad del motor no se ajusta a una distribución normal.

Dado que los datos no poseen una distribución normal, se realiza la prueba para determinar cuál es la distribución que mejor se ajusta a los datos. Se realizan 28 pruebas de distribuciones estadísticas donde se detallan las cinco más relevantes y que mejor ajustan de acuerdo con el criterio AIC, seleccionándolas por su valor más negativo como se detalla en la tabla En la Tabla 3-5.

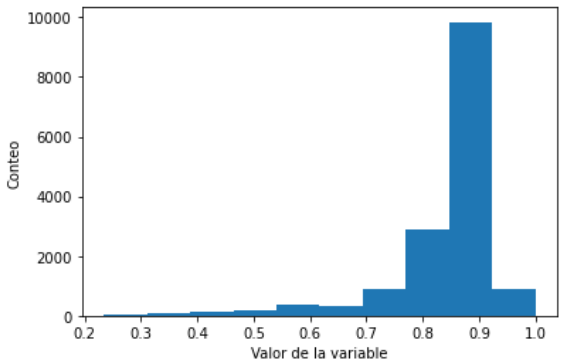
Tabla 3- 5. Ajuste de distribuciones estadísticas variable velocidad del motor.

	Distribución	Log_verosimilitud	AIC	BIC	Cantidad parámetros	Parámetros
1	gennorm	1155.8	-22305.7	-22280.3	3	'beta': 3.02, 'loc': 0.45, 'scale': 0.29
2	tukeylambda	10975.9	-21945.8	-21920.4	3	'lam': 0.26, 'loc': 0.45, 'scale': 0.14
3	genextreme	10849.1	-21692.3	-21666.9	3	'c': 0.27, 'loc': 0.39, 'scale': 0.17
4	powernorm	10655.2	-21304.4	-21279.0	3	'c': 0.32, 'loc': 0.30, 'scale': 0.12
5	pearson3	10590.62	-21175.2	-21149.8	3	'skew': 0.13, 'loc': 0.45, 'scale': 0.17

Bajo el criterio AIC la distribución que mejor se ajusta a los datos es la gennorm con parámetros {'beta': 3.02, 'loc': 0.45, 'scale': 0.29}.

3.4.2 Análisis variable temperatura de aceite del motor

La cantidad de registros de la prueba de escritorio para la variable temperatura del aceite del motor corresponde a 15612 registros. En la Figura 3-8 se relacionan el histograma de la temperatura del aceite del motor.



(a)

Figura 3- 8: Histograma de la temperatura del aceite del motor.

Con la prueba de normalidad de Shapiro-Wilk se obtiene en el estadístico el valor de 0.708 y el valor p de 0.000, por lo que se rechaza la hipótesis nula y la variable temperatura del aceite del motor no se ajusta a una distribución normal.

Dado que los datos no poseen una distribución normal, se realiza la prueba para determinar cuál es la distribución que mejor se ajusta a los datos. Se realizan 28 pruebas de distribuciones estadísticas donde se detallan las cinco más relevantes y que mejor ajustan de acuerdo con el criterio AIC, seleccionándolas por su valor más negativo como se detalla en la tabla en la Tabla 3-6.

Tabla 3- 6. Ajuste de distribuciones temperatura del aceite del motor.

	Distribución	Log_verosimilitud	AIC	BIC	Cantidad parámetros	Parámetros
1	dgamma	103002.4	-205998.8	-205975.8	3	'a':0.12, 'loc':0.88
2	dweibull	46279.4	-92552.3	-92529.8	3	'c':0.76, 'loc':0.88
3	gennorm	27801	-55596.4	-55573.5	3	'beta':0.28, 'loc':0.88
4	johnsonsu	23370	-46733.6	-46702.9	4	'a': 0.66, 'b':0.52
5	crystalball	21185	-42362.6	-42332	4	'beta':0.71, 'm':3.41

Bajo el criterio AIC la distribución que mejor se ajusta a los datos es la dgamma con parámetros {'a':0.12, 'loc':0.88, ‘scale’:0.08} como se detalla en la Figura 3-9.

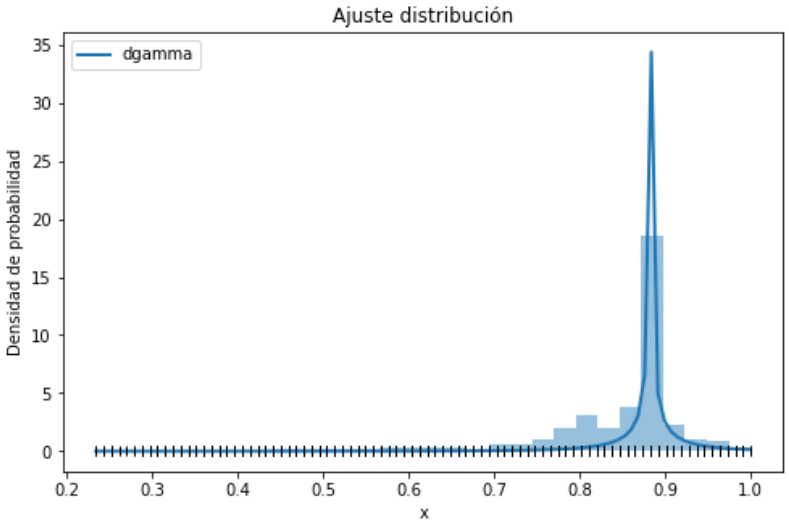


Figura 3-9: Distribución dgamma variable temperatura del aceite del motor.

Luego del análisis por variable, se obtienen las distribuciones estadísticas que mejor se ajustan a los datos, para trabajos futuros donde se realicen experimentos y toma de datos sin contar con grandes conjuntos de información es posible modelar el comportamiento de estas variables para generar datos sintéticos.

En la Tabla 3-7 se detalla el resultado del análisis sobre las funciones que mejor se ajustan a cada variable de la prueba de escritorio. En la primera columna se encuentra el nombre de la variable, en la segunda columna la cantidad de registros que fueron modelados con las 28 funciones de distribución estadística, en la tercera columna la distribución que mejor

se ajusta al comportamiento de los datos de cada variable acorde al criterio estadístico AIC indicado en la cuarta columna, en la quinta columna se encuentra la cantidad de parámetros de la distribución que mejor se ajusta y los parámetros con sus valores se encuentran en la sexta columna.

Según el análisis de las funciones de distribución de probabilidad por variable, de manera general se encuentra que para las variables de medición de combustible predomina la distribución pearson3, las variables categóricas *Ignition* y Vehículo activo la función dweibull es la que mejor modela estos comportamientos y para las variables de temperatura la función que mejor modela los datos es la dgamma. Estas funciones de distribución son herramientas para trabajos futuros en la modelación de datos e implementación de algoritmos de aprendizaje de máquinas.

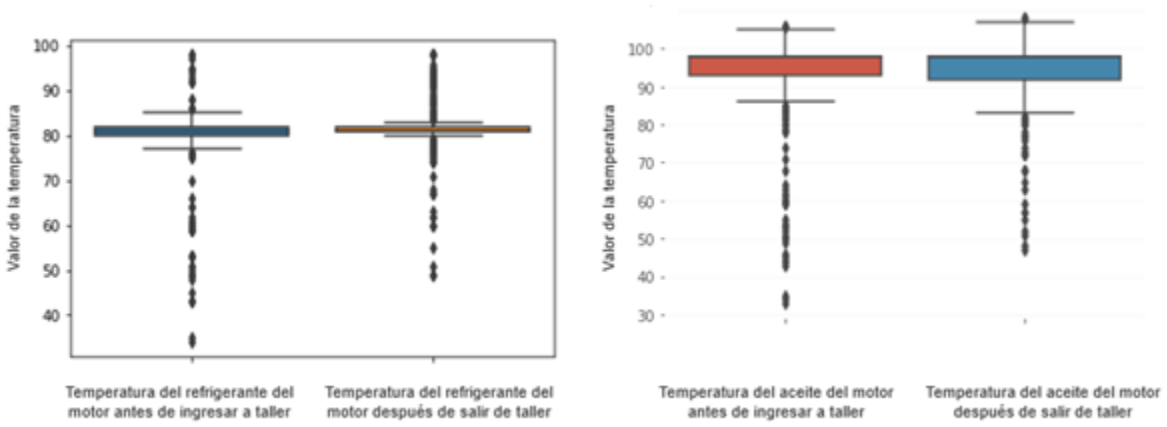
Tabla 3- 7. Ajuste de distribuciones todas las variables. Tractocamión prueba de escritorio.

	Variable	Cantidad de registros analizados	Distribución	AIC	Cantidad parámetros	Parámetros
1	Combustible de viaje utilizado	549	**pearson3**	-1912.77	3	skew': 2.89, 'loc': 0.08, 'scale': 0.12
2	Combustible de viaje utilizado en ralentí	549	**pearson3**	-1750.64	3	skew': 2.14, 'loc': 0.05, 'scale': 0.06
3	Combustible total utilizado (desde la instalación del dispositivo de telemática)	66000	**dgamma**	-2277080.00	3	a': 0.28, 'loc': 1.27e-26, 'scale': 6045.02
4	Combustible total utilizado en ralentí (desde la instalación del dispositivo de telemática)	572	**pearson3**	3913.10	3	skew': 2.17, 'loc': 8.99, 'scale': 9.79
5	Ignition	1100	**dweibull**	-41448.56	3	c': 426744117.70, 'loc': 0.49, 'scale': 0.50
6	Nivel de combustible	1201	**kappa4**	-1212.51	4	h': 1.02, 'k': 1.72, 'loc': -0.002, 'scale': 1.73
7	Nivel de DEF	14062	**pearson3**	-28777.84	3	skew': -2.20, 'loc': 0.838, 'scale': 0.18
8	Nivel de refrigerante	14022	**cauchy**	-1208821.00	2	loc': 1.0, 'scale': 5.42e-20
9	Odómetro	4131	**tukeylambda**	-10094.62	3	lam': 1.02, 'loc': 0.853, 'scale': 0.15
10	Voltaje dispositivo de telemetría	2899	**genextreme**	-10695.41	3	c': 0.59, 'loc': 0.92, 'scale': 0.04
11	Temperatura de aceite del motor	15612	**dgamma**	-205998.80	3	'a':0.12, 'loc':0.88, 'scale': 0.04
12	Temperatura del refrigerante del motor	25920	**dgamma**	-281980.92	3	a': 0.53, 'loc': 0.78, 'scale': 0.05
13	Temperatura exterior	901	**genextreme**	-1097.01	3	c': 0.22, 'loc': 0.58, 'scale': 0.12
14	Tensión de arranque	4513	**skewnorm**	-7327.27	3	a': -11.75, 'loc': 0.97, 'scale': 0.20
15	Tiempo de funcionamiento del motor	14003	**kappa4**	-33881.31	4	h': 1.00, 'k': 1.05, 'loc': 0.69, 'scale': 0.31
16	Vehículo activo (ralentí o en movimiento)	1101	**dweibull**	-38753.31	3	c': 122344589.29, 'loc': 0.49, 'scale': 0.50
17	Velocidad del motor	34995	**gennorm**	-22305.70	3	'beta': 3.02, 'loc': 0.45, 'scale': 0.29
18	Velocidad en carretera del motor	111713	**kappa4**	-63847.82	4	h': 1.07, 'k': 1.01, 'loc': -0.07, 'scale': 1.09

3.5 Análisis diferencia de medias para variables de temperatura del aceite y del refrigerante del motor

A continuación, se describe el análisis de medias comparando los datos un día antes de ingresar a taller y un día después de salir de taller para las variables temperatura del refrigerante del motor y temperatura del aceite del motor. Teniendo como referencia el día 26/08/2020, fecha en que el tractocamión ingresa a taller.

En la Figura 3-10 se presenta el análisis de medias para la temperatura del refrigerante del motor (a) y temperatura del aceite del motor (b).



(a) (b)

Figura 3- 10: Comparativo de medias temperatura del refrigerante del motor (a) y temperatura del aceite del motor (b) tractocamión caso número 2.

Para determinar la prueba de hipótesis para la diferencia de medias se define:

$$H_0 = \ No\ hay\ diferencia\ de\ medias\ entre\ los\ grupos$$
$$H_1 = \ Si\ hay\ diferencia\ de\ medias\ entre\ los\ grupo$$

Donde:

H_o: $Hipótesis\ nula$ y H_1: $Hipótesis\ alterna$

Si valor p < 0.05 se rechaza la hipótesis nula y los datos presentan diferencia de medias.

Para el análisis de medias de la temperatura del refrigerante del motor entre los grupos de datos antes y luego de salir de taller, el resultado del valor p con la prueba t-student es $1.24x10^{-7}$, por lo que se rechaza la hipótesis nula y se concluye que se presenta diferencia de medias entre los datos antes de taller y luego de salir de taller.

Para el análisis de medias de la temperatura del aceite del motor, el resultado del valor p con la prueba t-student es 0.007, por lo que se rechaza la hipótesis nula y se concluye que se presenta diferencia de medias entre los datos antes de taller y luego de salir de taller. Los anteriores resultados para el análisis de este tractocamión indican que se presentan cambios estadísticos en el comportamiento de estas variables antes de ser intervenido el vehículo y luego de salir del taller.

3.6 Estrategias para la preparación de los datos de tractocamiones graneleros

Como se indica en la sección 3.3 Análisis descriptivo de los datos. Una de las dificultades en esta tesis corresponde a que los datos de los sensores en los tractocamiones graneleros del estudio no capturan la información de las 18 variables en el mismo instante de tiempo, las frecuencias son diferentes por cada variable. En el momento de consolidar la información de todas las variables, el *dataset* generado cuenta con registros incompletos por variables.

Es por esto, que para cada experimentación realizada por tractocamión o consolidando varios vehículos a la vez, se propone realizar una interpolación por cada variable y luego de tener los datos interpolados incorporar este resultado nuevamente en el *dataset* que se consolide. Los resultados del método ayudan a completar estos registros que faltan en cada instante de tiempo de los datos para realizar los análisis descriptivos y preparación de los datos antes de implementar los algoritmos de reducción de dimensionalidad y agrupamiento posteriormente.

De manera específica, en la Tabla 3-8 se presentan los resultados de interpolación en el caso del tractocamión caso dos de la prueba de escritorio con los datos de todo el año 2020. Donde por cada variable se específica la cantidad de datos que se interpolan para la construcción del *dataset* con todas las variables de la tesis, donde se obtiene un archivo con un tamaño de 1.100.427 registros de 18 variables y un peso en memoria de 398 *Megabytes*.

Tabla 3- 8. Ajuste de distribuciones todas las variables. Tractocamión prueba de escritorio.

Variable	Datos antes de interpolar	Datos interpolados	Total datos con interpolación
Combustible total utilizado	289,463	810,964	1,100,427
Combustible total utilizado en ralentí	3,174	1,097,253	1,100,427
Combustible de viaje utilizado	2,981	1,097,446	1,100,427
Combustible de viaje utilizado en ralentí	2,977	1,097,450	1,100,427
Nivel combustible	6,348	1,094,079	1,100,427
Nivel DEF	47,925	1,052,502	1,100,427
Nivel Refrigerante	44,812	1,055,615	1,100,427
Odometro	27,995	1,072,432	1,100,427
Voltaje dispositivo de telemetría	17,211	1,083,216	1,100,427
Temperatura del aceite del motor	55,839	1,044,588	1,100,427
Temperatura del refrigerante del motor	92,049	1,008,378	1,100,427
Temperatura exterior	4,606	1,095,821	1,100,427
Tensión de arranque	29,988	1,070,439	1,100,427
Tiempo funcionamiento del motor	47,126	1,053,301	1,100,427
Vehículo activo	5,968	1,094,459	1,100,427
Velocidad del motor	153,133	947,294	1,100,427
Ignition	5,967	1,094,460	1,100,427
Velocidad en carretera del motor	514,599	585,828	1,100,427

En la Figura 3-11 se presentan los gráficos de la temperatura del refrigerante del motor y temperatura exterior con los datos antes y luego de la interpolación de los datos. Siendo (a) la temperatura del refrigerante del motor antes de la interpolación y (b) el grafico con los datos interpolados. Como también en (c) se encuentra la temperatura exterior antes de interpolar y en (d) el gráfico con los datos interpolados. Se encuentra que los patrones en la información son interpolados adecuadamente de acuerdo con el comportamiento de los

datos. Las secciones donde no se presentan datos corresponden a periodos de inactividad del equipo a raíz de la afectación en el transporte por el Covid-19.

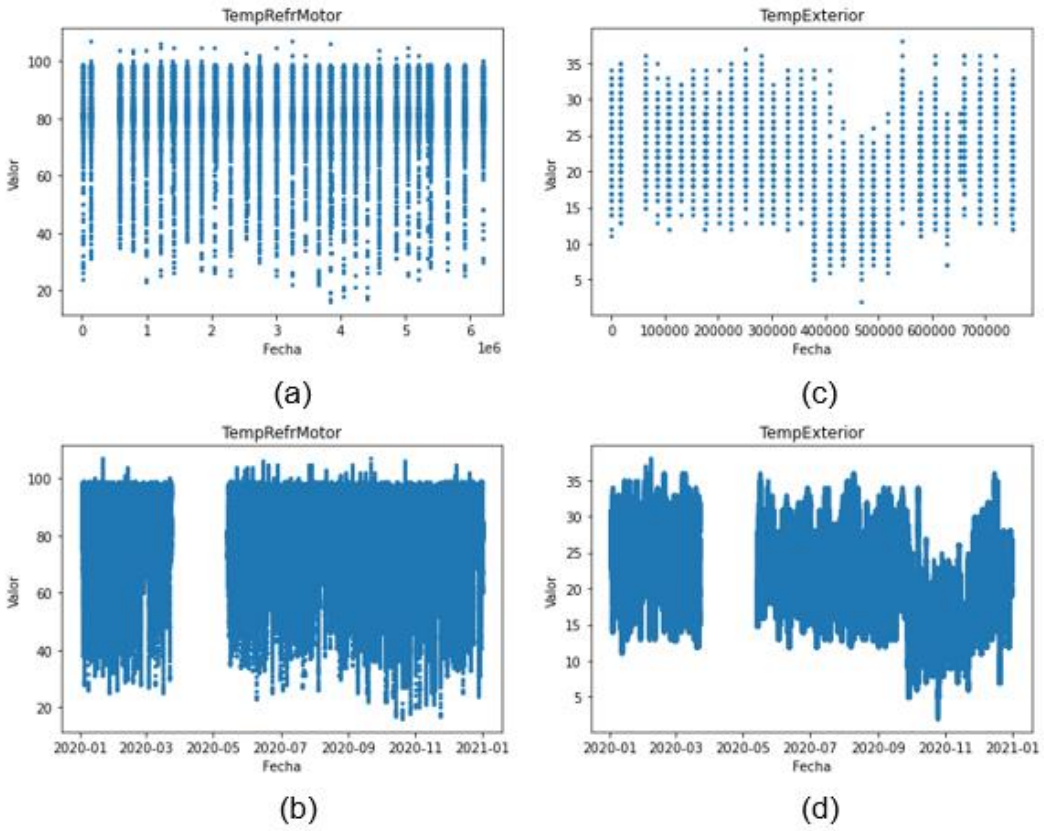
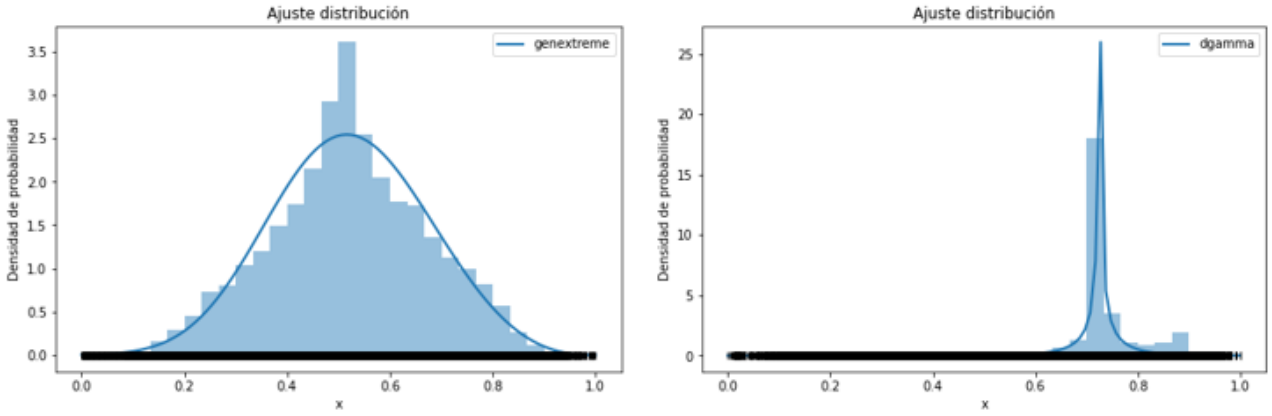


Figura 3- 11: Interpolación de variables temperatura refrigerante del motor. Temperatura del refrigerante del motor antes de la interpolación (a), datos interpolados temperatura del refrigerante del motor(b) e Interpolación temperatura del aceite del motor, temperatura exterior antes de interpolar (c) y datos interpolados temperatura exterior (d). Tractocamión caso número dos.

En la Figura 3-12 se realiza la validación sobre las distribuciones estadísticas para las variables temperatura exterior (a) y temperatura del refrigerante del motor (b). Se encuentra que las distribuciones estadísticas de estas variables son iguales antes de haber interpolado los datos, como se describe anteriormente en la sección 3.4.2.

Se concluye que el método de interpolación se ajusta al comportamiento de las variables al presentar la misma distribución estadística antes y después de interpolar los datos.



(a) (b)

Figura 3- 12: Distribuciones estadísticas variables temperatura refrigerante del motor (a) y temperatura exterior (b) tractocamión caso número dos después de interpolar.

En la Tabla 3-9 se presentan las distribuciones estadísticas de las variables anteriormente analizadas junto con sus parámetros de comportamiento.

Tabla 3- 9. Distribución estadística temperatura exterior y temperatura del refrigerante del motor después de interpolación. Tractocamión prueba de escritorio.

Variable	Cantidad de registros analizados	Distribución	Log_verosimilitud	AIC	BIC	Cantidad parámetros	Parámetros
Temperatura exterior	1,100,427	**genextreme**	523939.20	-1047872.48	-1047836.75	3	'c': 0.28, 'loc': 0.46, 'scale': 0.15
Temperatura del refrigerante del motor	1,100,427	**dgamma**	4345047.50	-8690089.02	-8690053.29	3	'a': 0.20, 'loc': 0.72, 'scale': 0.07

El método de interpolación funciona adecuadamente para la modelación de los datos, donde se encuentra que la distribución obtenida en la prueba de escritorio se comporta de la misma manera cuando se realiza en el conjunto de datos de todo el año 2020 para las variables analizadas. La variable temperatura exterior cuenta con 4.606 datos antes de interpolar y con 1.100.427 datos luego de interpolar, preservando el comportamiento estadístico de una distribución genextreme como se describe en los resultados anteriores con la prueba de escritorio. Adicional, la variable temperatura del refrigerante del motor cuenta con 92.049 datos antes de interpolar y con 1.100.427 datos después de interpolar,

donde se preserva el comportamiento de una distribución dgamma presentada también en la experimentación con el tractocamión de la prueba de escritorio.
Con la validación de los resultados de interpolación se procede como estrategia en la formación de los *datasets* con todas las variables para los casos en que se analicen tractocamiones individuales o en conjunto.

Para el desarrollo de la tesis se cuenta con grandes cantidades de datos, como se observa en las anteriores experimentaciones, en el momento de interpolar los datos completos del año 2020 y solo para un tractocamión, en promedio se encuentra un conjunto de datos con 1 millón de registros para las mediciones de 18 variables y un peso en memoria de 400 *Megabytes* por archivo generado. El recurso local para el desarrollo de este trabajo comprende una máquina Intel Core i5 2.5 GHz con memoria RAM de 12 *Gigabytes*.

La estrategia desarrollada para trabajar en grandes conjuntos de datos consistió en realizar experimentaciones de los algoritmos con muestras pequeñas de toda la información disponible, como lo es el caso de la prueba de escritorio con la información de dos meses de un tractocamión. Con esta muestra pequeña de la información se realiza la experimentación con los algoritmos, con el fin de conocer su comportamiento y desempeño en tiempos de computación.

Con el procesamiento de las técnicas en la máquina local se comienzan a encontrar dificultades con conjuntos de datos completos del año para cada placa que se iba a analizar (1 millón de datos x 18 variables), donde el equipo de cómputo alcanzaba tiempos de procesamiento de 12 a 24 horas dependiendo la complejidad del algoritmo en evaluación, como también se presentaron casos donde la cantidad de operaciones matemáticas que computaban los algoritmos colapsaba la infraestructura computacional disponible limitando el avance en la experimentación.

Ante la necesidad de continuar experimentando con más información se explora el procesamiento distribuido en máquinas como servicio en la nube. Para la continuidad de la experimentación con estos conjuntos de datos se contratan máquinas en la nube de uno de los proveedores de estos servicios.

3.7 Conclusiones

- En este capítulo se realiza un revisión de las fuentes de datos que se cuentan en la tesis, las cuales corresponden a los datos recopilados de los sensores satelitales en la flota de tractocamiones graneleros, dispositivos que capturan en tiempo real información sobre variables como temperaturas, velocidad, consumo de combustible, niveles de líquidos, entre otras que serán el insumo fundamental en la implementación de técnicas no supervisadas en la detección de patrones de fallos. Como también la información asociada a ingresos a taller durante el año 2020 para intervenciones de la flota en sistemas térmicos que permitirán validar la eficiencia de los patrones encontrados de las técnicas de agrupamiento.

- Se realiza una descripción de las etapas en la preparación de los datos, para obtener una guía general que especifica los pasos a seguir desde la recolección de la información, preparación, utilización de los datos, procesamiento, interpretación y posterior almacenamiento.

- En el análisis descriptivo de la información se revisa el comportamiento estadístico de las variables, donde se realiza la prueba de normalidad en los datos y determinar si las variables poseen este comportamiento. Se evidencia que las variables de la tesis no cuentan con comportamientos bajo una distribución normal, por lo que se analizan cuáles son las funciones de distribución de probabilidad que mejor se ajustan a los datos junto con los parámetros de estas funciones. En la implementación de estas técnicas no se pueden realizar supuestos de normalidad en los datos antes de realizar las respectivas pruebas, por lo que con base al análisis descriptivo de los datos de los tractocamiones graneleros se requiere normalizar los datos antes de implementar las técnicas no supervisadas en esta tesis.

- De acuerdo con la revisión de la cantidad de información recolectada para cada variable en un día, se identifica que para las variables de la tesis, los datos no son recolectados en el mismo instante de tiempo, por lo que al unir los conjuntos de datos se encuentran momentos en el tiempo donde algunas variables no tendrá

información asociada. La estrategia implementada consistió en realizar interpolación de los datos para capturar los comportamientos de las variables y asignar datos en los momentos de ausencia de información con este método descrito y que pudo conservar el comportamiento estadístico en cada uno de los casos interpolados.

4. REDUCCIÓN DE LA DIMENSIONALIDAD

En este capítulo se presentan los métodos de reducción de dimensionalidad propuestos en esta tesis. En la sección 4.1 se describe el algoritmo PCA (*Principal component analysis*) junto con el criterio de selección de componentes principales que explican el 85% de la varianza de la información en el tractocamión con número de caso dos con el que se implementa la prueba de escritorio.

En la sección 4.2 se describe el algoritmo t-SNE (*T-distributed Stochastic Neighbor Embedding*) junto con las experimentaciones realizadas en la variación de hiperparámetros con el objetivo de mejorar el desempeño del algoritmo basándose en las recomendaciones encontradas por (Wattenberg, et. al., 2016)

Finalmente, en la sección 4.3 se describen las características principales del método UMAP (*Uniform Manifold Approximation and Projection for Dimension Reduction*) con la estrategia de selección de hiperparámetros que se describe en la documentación del algoritmo para mejorar el desempeño del algoritmo.

Con la revisión de estas técnicas se persigue encontrar a través de la experimentación con los datos de los tractocamiones que técnica es la más adecuada para reducir la dimensionalidad y posteriormente evaluarla con las técnicas de agrupamiento no supervisadas en la detección de patrones de falla asociado a sistemas térmicos.

4.1 PCA (*Principal component analysis*)

Como definen (Pacella & Papadia, 2020), el método de reducción de dimensionalidad PCA (*Principal component analysis*) es capaz de reducir la dispersión de las medidas de múltiples sensores y su dimensionalidad. PCA da como resultado la transformación de variables originales en un número pequeño de características o componentes principales.

Como definen Pacella & Papadia (2020), se encuentra que la operación de PCA consiste en la computación de la matriz de covarianza del conjunto de datos y su descomposición.

Los valores propios resultantes son organizados en orden decreciente, donde cada valor propio es relacionado a la fracción de varianza explicada por los componentes principales. Correspondiente a los valores propios ortogonales que describen una base del espacio cuyas direcciones se refieren a la máxima variabilidad de los datos.

La descripción matemática de (Pacella & Papadia, 2020), considera el caso de la cantidad de variables nombradas como P, con M muestras o datos tomados. La muestra de datos es almacenada en una matriz designada como $\mathbf{X} \in \mathbb{R}^{P\,x\,M}$ y direccionada con índices: j,i relacionados a filas y columnas de $\mathbf{X}$ respectivamente.

Donde $\bar{\mathbf{x}} = \frac{1}{M} \sum_{i=1}^{M} \boldsymbol{x_i}$ corresponde al vector promedio de los datos y $\boldsymbol{x_i^c} = \boldsymbol{x_i} - \bar{\boldsymbol{x}}$ corresponde al vector centrado obtenido de $\boldsymbol{x_j}$ por sustracción del vector promedio. El conjunto de datos completo puede ser representado por la matriz $\boldsymbol{X_c} \in \mathbb{R}^{P\,x\,M}$.

El objetivo de PCA es resolver el problema de aproximar la matriz de datos $\boldsymbol{X_c}$ con otra matriz $\widehat{\boldsymbol{X}}^{\boldsymbol{c}}$ la cual posea un rango inferior y donde el objetivo sea minimizar la distancia entre $\boldsymbol{X_c}$ y $\widehat{\boldsymbol{X}}^{\boldsymbol{c}}$.

Ahora, $\boldsymbol{N}$ es un límite superior dado para el rango de la matriz $\widehat{\boldsymbol{X}}^{\boldsymbol{c}}$ ($\boldsymbol{N} < \boldsymbol{P}$), por consiguiente, se denota por $\widehat{\boldsymbol{U}}$ la matriz formada por las primeras $\boldsymbol{N}$ columnas de $\boldsymbol{U}$, lo cual corresponde a los primeros $\boldsymbol{N}$ valores únicos más grandes de $\boldsymbol{X^c}$, un vector de datos de muestra de $\boldsymbol{P}$ puntos $\boldsymbol{x_i}$ (i = 1,…,M) es proyectado al espacio de características como $\widehat{\boldsymbol{U}}^{\boldsymbol{T}} = (\boldsymbol{x_i} - \bar{\boldsymbol{x}})$. Este es el vector de $\boldsymbol{N}$ coordenadas $\boldsymbol{t_i} = (\boldsymbol{t_{i1}}, \ldots, \boldsymbol{t_i})$ el cual representa los valores de los componentes principales del método PCA.

PCA como proceso de reducción de dimensionalidad permite obtener un conjunto de datos de menor tamaño, lo que permite disminuir los tiempos de procesamiento en los algoritmos de aprendizaje de máquinas, entre ellos los algoritmos no supervisados para agrupamiento.

El parámetro base del algoritmo PCA es la cantidad de componentes que explican la mayor varianza de los datos en el resultado de la reducción de la dimensionalidad, (Pedregosa et. al., 2011)

En la elección de los componentes principales se realiza experimentación con la prueba de escritorio correspondiente al tractocamión con caso número dos del conjunto de datos, correspondiendo a 1.100.427 registros y las 18 variables de la tesis con el fin de identificar la cantidad de componentes principales para elegir como parámetro en la ejecución del algoritmo como se detalla en la Figura 4-1.

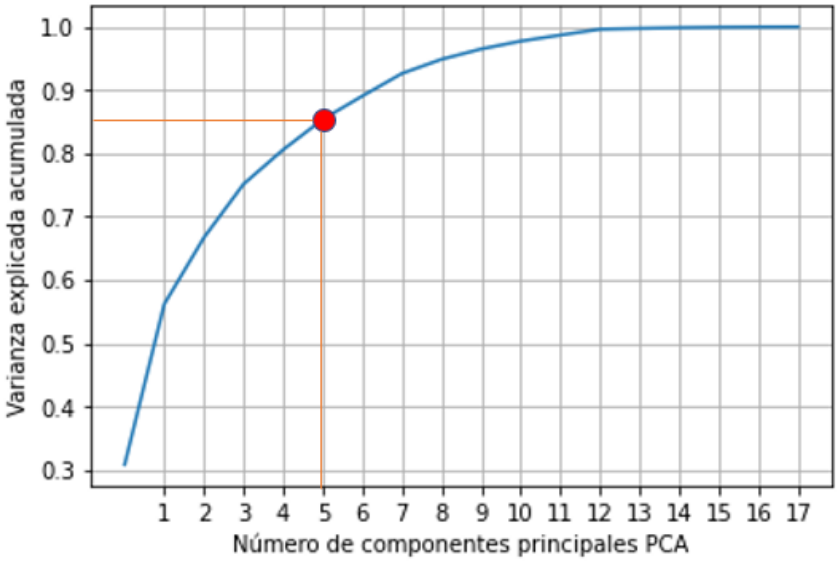


Figura 4- 1: Selección de componentes principales PCA tractocamión caso número dos, varianza explicada acumulada.

Con base en el resultado anterior y para la ejecución de los algoritmos de agrupamiento se establece como parámetro elegir cinco componentes principales, según la experimentación estos cinco componentes principales de PCA explican el 85% de la varianza de la información, valor que se toma como umbral en esta tesis para la implantación posterior de los algoritmos de agrupamiento.

4.2 t-SNE (*t-Distributed Stochastic Neighbor Embedding*)

Como mencionan (Van der Maaten & Hinton, 2008), t-SNE es una herramienta que permite visualizar datos de alta dimensión en un espacio de baja dimensión, antes de implementar el t-SNE se recomienda usar otro método de reducción de dimensionalidad como el PCA para reducir el número de dimensiones a una cantidad razonable si la cantidad de características del conjunto de datos es grande permitiendo disminuir el ruido en la información y acelerando el tiempo de computación del algoritmo t-SNE.

La descripción matemática de (Van der Maaten & Hinton, 2008), es explicada en el sentido que t-SNE emplea una distribución t-Student con un grado de libertad (distribución Cauchy) como una distribución en un mapa de baja dimensión. Con esta distribución, las probabilidades conjuntas $\boldsymbol{q_{ij}}$ son definidas como:

$$\boldsymbol{q_{ij}} = \frac{(1 + \| \boldsymbol{y_i} - \boldsymbol{y_j} \|^2)^{-1}}{\sum_{k \neq l}(1 + \| \boldsymbol{y_k} - \boldsymbol{y_l} \|^2)^{-1}}$$

Se usa una distribución t-Student con un único grado de libertad porque $(1 + \| \boldsymbol{y_i} - \boldsymbol{y_j} \|^2)^{-1}$ tiene la particular propiedad de aproximarse a una ley del inverso del cuadrado para grandes distancias entre pares de datos $\| \boldsymbol{y_i} - \boldsymbol{y_j} \|$ en el mapa de bajas dimensiones. Esto hace que la representación del mapa de probabilidades conjuntas sea casi invariante a cambios en la escala del mapa para puntos que están alejados entre ellos. Esto también significa que grandes grupos de puntos que están muy separados interactúan de la misma manera que los puntos individuales, por lo que la optimización opera de la misma manera en todas las escalas excepto en las más detalladas.

El hiperparámetro base del algoritmo es la cantidad de componentes, donde por defecto es dos, siendo esto las dimensiones resultantes del proceso de reducción de dimensionalidad. En el desarrollo de esta tesis se realiza experimentación con otros dos hiperparámetros los cuales son perplejidad y la cantidad de iteraciones del algoritmo con el fin de optimizar los resultados del proceso.

Como lo describen también (Van der Maaten & Hinton, 2008), la perplejidad es relacionada al número de vecinos más cercanos que son usados en el proceso de generación de clases en el conjunto de datos. Los grandes conjuntos de datos usualmente requieren un valor alto de perplejidad. Los diferentes valores de la perplejidad pueden entregar diferentes resultados al proceso de reducción de dimensionalidad.

La cantidad de iteraciones también es un factor de importancia para la convergencia del algoritmo. No existe una definición sobre la cantidad fija de iteraciones que indiquen la estabilidad en el resultado. Diferentes conjuntos de datos pueden requerir diferentes números de iteraciones para obtener convergencia y datos desagregados. Se recomienda realizar experimentación en las iteraciones hasta encontrar una convergencia esperada en la reducción de dimensionalidad (Wattenberg, et al., 2016)

Con las anteriores definiciones de los hiperparámetros se realiza experimentación sobre el tractocamión de la prueba de escritorio descrita anteriormente correspondiente al tractocamión caso número dos, donde previamente se realizó el método PCA para 1.100.427 registros y 18 variables la elección de cinco componentes principales y posteriormente se realiza el método t-SNE realizando cinco pruebas donde se mantiene constante la cantidad de 5.000 iteraciones y variando la perplejidad con valores de 2, 5, 30, 50 y 100 como se puede visualizar en la Figura 4-2 respectivamente.

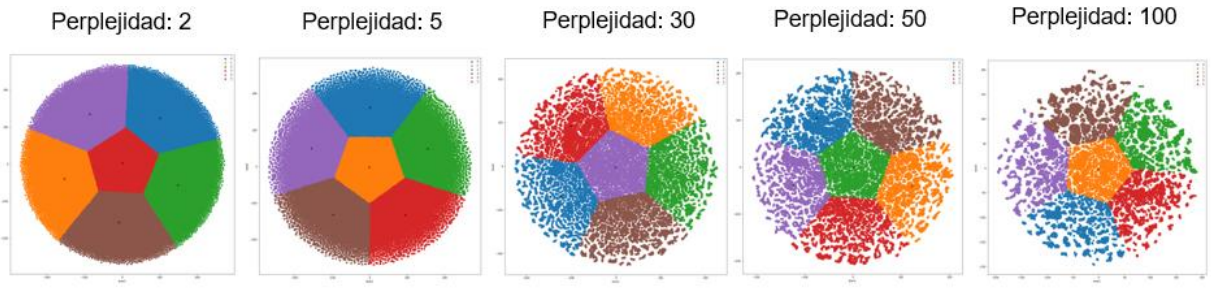


Figura 4- 2: Experimentación t-SNE con cantidad de iteraciones constantes y variación de la perplejidad tractocamión caso número dos.

Con respecto a los anteriores resultados, se presenta que variando el hiperparámetro de perplejidad se encuentra una desagregación de los datos con el valor de perplejidad 100

que no es tan marcada en todo el espacio dimensional resultante pero visualmente mejor con respecto a la experimentación con valores menores de perplejidad. El método t-SNE no presenta resultados satisfactorios ante la desagregación de los datos de tractocamiones graneleros aun realizándose variaciones en el hiperparámetro de perplejidad para posteriormente aplicar técnicas de agrupamiento.

Adicional, se realiza la experimentación con valor constante de 30 para la perplejidad y variando la cantidad de iteraciones para el algoritmo en 250, 500, 1000, 2000 y 5000 como se puede visualizar en la figura 4-3 respectivamente.

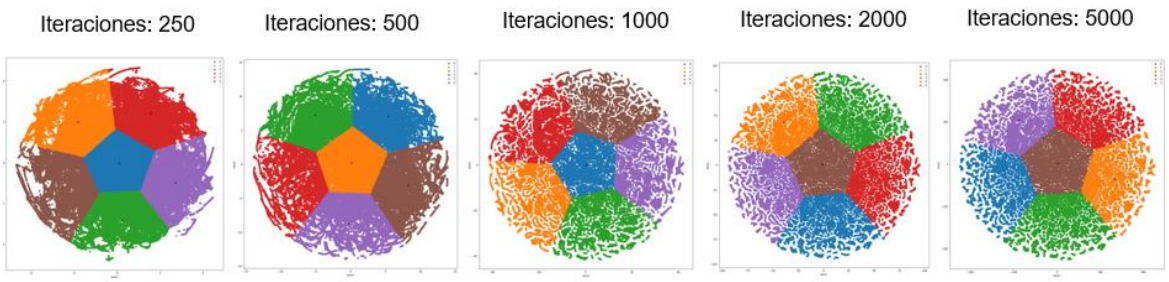


Figura 4- 3: Experimentación t-SNE con valor constante en perplejidad y variación de la cantidad de iteraciones tractocamión caso número dos.

Con el resultado de las dos anteriores experimentaciones en la variación de los hiperparámetros perplejidad y cantidad de iteraciones no se encuentra una desagregación de los datos para el caso de los tractocamiones graneleros y aplicado en el tractocamión caso número dos precisamente que permita que los algoritmos de agrupamiento realicen una separación de datos marcada en esta técnica de reducción de dimensionalidad.

4.3 UMAP (*Uniform Manifold Approximation and Projection)*

Como describen (McInnes et al., 2018), el algoritmo de reducción de dimensionalidad UMAP (*Uniform Manifold Approximation and Projection*) es competitivo con t-SNE en calidad de visualización y preserva mejor la estructura global de los datos con una mejora en tiempo de procesamiento. UMAP no tiene restricciones computacionales para construir

representaciones topológicas de datos con altas dimensiones, lo cual lo hace viable en la elección de técnicas de reducción de dimensionalidad para aprendizaje de máquinas.

En las definiciones de (Allaoui, et. al., 2020), UMAP representa los puntos de datos en un grafo altamente dimensional ponderado, con pesos que representan la probabilidad de que dos puntos se encuentren conectados. UMAP utiliza una distribución de probabilidad exponencial para calcular la similitud entre puntos de alta dimensión con el siguiente fundamento matemático:

$$\boldsymbol{p_{ij}} = \exp\left(-\frac{d(x_i, x_j) - \rho_i}{\sigma_i}\right)$$

Donde $d(x_i, x_j)$ es la distancias entre los i-esimos y j-esimos puntos de datos y ρ es la distancia entre los i-esimos puntos de datos y su primer vecino más cercano. En los casos que los pesos del grafo entre los nodos i y j no son iguales a los pesos entre los nodos j e i, UMAP utiliza una simetrización de la probabilidad altamente dimensional:

$$\boldsymbol{p_{ij}} = \boldsymbol{p_{i|j}} + \boldsymbol{p_{j|i}} - \boldsymbol{p_{i|j}}\,\boldsymbol{p_{j|i}}$$

El grafo construido es un gráfico de probabilidad y UMAP necesita especificar el número K de vecinos más cercanos:

$$K = 2\sum_i p_{ij}$$

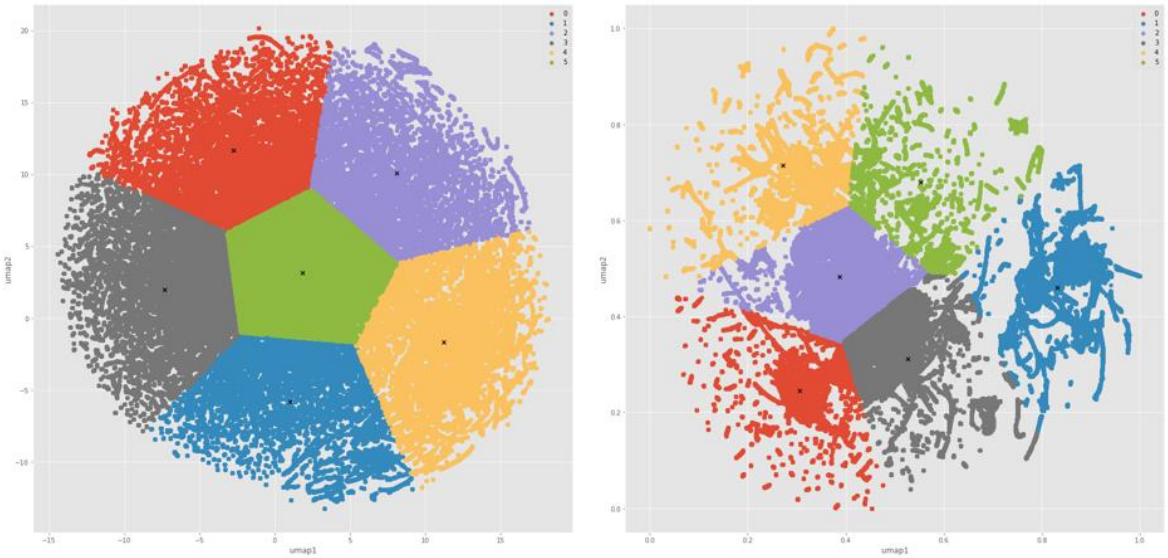
Una vez, el grafo altamente dimensional es construido, UMAP construye y optimiza la salida de un espacio de baja dimensión análogo tan similar sea posible.

En la documentación del algoritmo se menciona que la variación de los hiperparámetros puede tener un impacto significativo en los resultados de la separación de las clases en la reducción de dimensionalidad, en las experimentaciones que se realizan en la documentación se encuentra que los principales parámetros objetivos son la cantidad de vecinos en el conjunto de datos, este parámetro controla como UMAP realiza el balance la estructura local y global en los datos.

Los valores bajos en el hiperparámetro de la cantidad de vecinos dan como resultado una concentración sobre las estructuras locales de los datos, mientras que valores altos obligan a UMAP a buscar vecindarios más grandes de cada punto perdiendo estructura de los detalles obteniendo resultados más amplios. Adicional, se presenta también el hiperparámetro de distancia mínima que controla que tan estrictamente se le permite a UMAP agrupar puntos. El hiperparámetro define la distancia mínima entre los puntos que estarán en la representación de baja dimensión. Los valores bajos generan separaciones de los datos más desagregadas y los valores altos evitan que UMAP empaquete los puntos y se centra en estructuras más amplias en la reducción de dimensionalidad (McInnes et al., 2018),

Para efectos de la experimentación de esta tesis, con base en el anterior análisis de hiperparámetros y la cantidad de información que se tiene, se seleccionan como valores de los hiperparámetros el valor de 100 para la cantidad de vecinos en la agrupación de clases y como distancia mínima un valor de 0.0 para lograr encontrar pequeños componentes conectados, grupos y cadenas en los datos, enfatizando estas características en la reducción de la dimensionalidad sin pérdida de detalle en la estructura topológica de los datos, como se define en (McInnes et al., 2018).

En la Figura 4-4 se presenta el resultado de la experimentación con los datos del año 2020 para el tractocamión prueba de escritorio. Donde (a) representa el resultado del método UMAP, sin la variación de hiperparámetros mencionada anteriormente, los hiperparámetros con los que opera el algoritmo son los estándar del método que corresponden a una distancia mínima con valor de 0.1 y la cantidad de vecinos con un valor de 15. Se encuentra que los datos no son desagregados en esta experimentación. Al comparar con (b), donde se asigna el valor de 0.0 para la distancia mínima y un valor de 100 para la cantidad de vecinos, se encuentra que los datos son desagregados, encontrando resultados prometedores para esta estrategia y próximas experimentaciones en conjunto con los algoritmos de agrupamiento.



(a) (b)

Figura 4- 4: Experimentación resultados UMAP datos tractocamión prueba de escritorio con variación de hiperparámetros. Resultado sin ajuste de hiperparámetros (a). Resultado con ajuste de hiperparámetros (b).

Se concluye de este capítulo que para la reducción de dimensionalidad con el método PCA se toma la cantidad de cinco componentes que explican el 85% de la varianza de la información. Con este resultado posteriormente, se pueden aplicar las técnicas de agrupamiento para evaluar la formación de grupos y que tan separados pueden estar unos de otros.

El método t-SNE con los hiperparámetros ajustados de perplejidad con valor de 100 y 5.000 iteraciones no encuentra una desagregación marcada en los datos donde se identifiquen grupos separados unos de otros, sin embargo se realiza la experimentación de este resultado con los algoritmos de agrupamiento para evaluar resultados de los grupos que se generen.

Los resultados de UMAP con la variación de hiperparámetros para la distancia mínima con un valor de 0.0 y la cantidad de vecinos con valor de 100, indican resultados prometedores en la desagregación de los datos, factor importante para la implementación de los algoritmos de agrupamiento.

Se espera que la implementación de estos métodos de reducción de dimensionalidad con los hiperparámetros sugeridos en la literatura permitan alivianar la cantidad de información que se tiene con los datos de la flota de tractocamiones sin perder información relevante y que genere una mejora en el rendimiento de los algoritmos de agrupamiento no supervisados en la determinación de grupos que puedan identificar estados de falla en los sistemas térmicos.

Con la revisión de las técnicas de reducción de dimensionalidad se busca encontrar una técnica adecuada donde se puedan desagregar los datos en un espacio de baja dimensión para continuar con la estrategia para detectar patrones de falla en conjunto con los algoritmos de agrupamiento.

4.4 Conclusiones

- En este capítulo se presentaron los métodos de reducción de dimensionalidad implementados en la tesis, correspondientes a PCA, t-SNE y UMAP. Donde se exploraron los hiperparámetros con base en las estrategias que se encuentran en la revisión de la literatura para optimizar sus resultados en la reducción de dimensionalidad para el conjunto de datos de tractocamiones graneleros.

- En la revisión de la técnica PCA se encuentra que en la identificación de la varianza acumulada por componentes principales y conforme a la experimentación realizada, la cantidad de componentes principales a elegir con este método corresponde a los cinco primeros componentes principales encontrados, explicando el 85% de la información. Con este resultado se implementa en el capítulo 5 las técnicas de agrupamiento bajo esta metodología.

- Para la exploración de los hiperparámetros con la técnica t-SNE se encuentra que la literatura recomienda realizar experimentación con los hiperparámetros perplejidad y cantidad de iteraciones, donde se seleccionan para las experimentaciones desarrolladas en la tesis, un valor de perplejidad de 100 y 5000 iteraciones, sin embargo para los datos de los tractocamiones graneleros no se encuentra desagregación de los datos en la reducción de dimensionalidad. Esto es

debido a que la experimentación cuenta con una cantidad de datos correspondiente a 1.100.427 registros de las 18 variables de la tesis y los datos se encuentran esparcidos ampliamente en la región de dos dimensiones que entrega como resultado el método t-SNE.

- En la experimentación con la técnica UMAP se exploran los hiperparámetros cantidad de vecinos con un valor de 100 en la agrupación de clases y como distancia mínima entre datos un valor de 0.0 para lograr encontrar pequeños componentes conectados en los datos, esto con base en la experimentación de los resultados con estos hiperparámetros propuestos en la literatura y que para el caso de experimentación se encuentran patrones en la desagregación de los datos con esta técnica de reducción de dimensionalidad apuntando a resultados prometedores.

- Se hace necesario tener estrategias cualitativas y respaldas en experimentación para determinar cuáles son los hiperparámetros a asignar con las técnicas de reducción de dimensionalidad, con estas estrategias se pueden mejorar los resultados en la desagregación de los datos para posteriormente implementar los algoritmos de agrupamiento en los datos de la flota de tractocamiones graneleros.

5. EVALUACIÓN ESTRATEGIAS DE AGRUPAMIENTO NO SUPERVISADAS EN LA DETECCIÓN DE FALLAS

En el desarrollo de este capítulo se presentan las técnicas de agrupamiento donde se pueden clasificar en dos categorías, la primera hace referencia a la aplicación en conjuntos de datos multivariados donde se aplican las técnicas KMEANS, DBSCAN y SOM. En la segunda categoría se encuentran los conjuntos de datos definidos por una temporalidad o series de tiempo donde se aplica la técnica de agrupamiento no supervisada *Spectral Clustering.*

Se presentan los resultados de los métodos de agrupamiento propuestos para la experimentación en un conjunto de datos aleatorios de 1.062.899 registros con 18 variables para el tractocamión con caso número 10. El objetivo del capítulo es encontrar cual es la técnica que mejor se puede desempeñar en el agrupamiento de datos para determinar estados de falla en los sistemas térmicos de los tractocamiones.

5.1 Agrupamientos en conjuntos de datos multivariados

Para el conjunto de datos de la tesis se presentan la unión de las variables medidas por los sensores satelitales, analizar estas variables conjuntamente apoyan el análisis de la relación y entendimiento de las mismas variables con el efecto que tienen en el comportamiento de los sistemas térmicos. Se presentan los resultados obtenidos con los métodos de reducción de dimensionalidad y los algoritmos de agrupamiento en el conjunto de datos multivariado.

5.1.1 KMEANS

Se detallan los resultados obtenidos con los métodos de reducción de dimensionalidad, PCA, t-SNE y UMAP para el algoritmo de KMEANS con los datos del tractocamión con caso número diez en un conjunto de 1.062.899 datos de 18 variables de la tesis.

Para la selección de la cantidad de grupos a calcular con el algoritmo KMEANS se determina a través del método del codo que nos sugiere la cantidad optima de agrupaciones de acuerdo con la distancia media entre cada uno de los puntos a los centroides que conforman cada grupo.

En la Figura 5-1 se detalla el resultado del método del codo, se encuentra que luego de seleccionar la cantidad de seis grupos la variación en la distancia media de los puntos a los centroides no tiene diferencias significativas, por lo que se selecciona el valor de seis para la cantidad de agrupamientos a parametrizar en el algoritmo KMEANS.

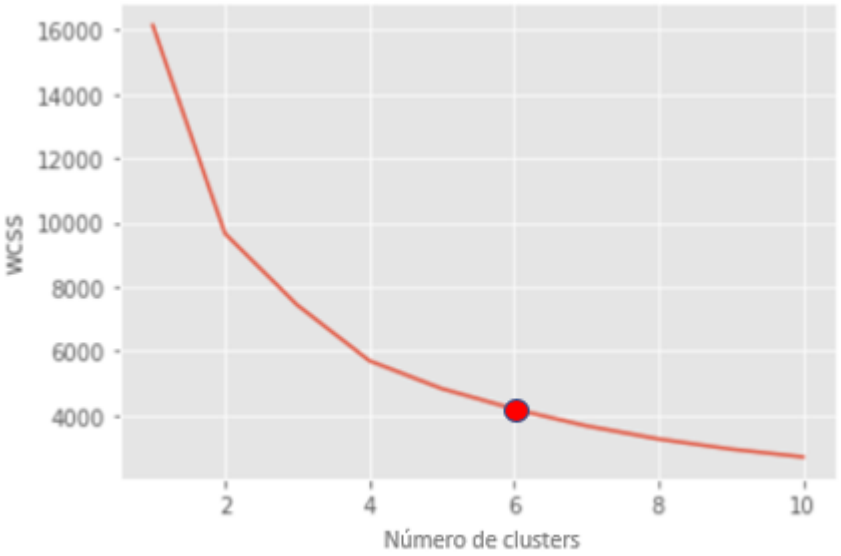


Figura 5- 1: Método del codo para la determinación de grupos para el algoritmo KMEANS.

5.1.1.1 PCA - KMEANS

En la reducción de dimensionalidad con PCA, el 85% de la varianza de los datos es explicada por cinco componentes del PCA como se detalla en la Figura 5-2, donde el 85% de la varianza en la información es explicada por cinco componentes principales del PCA, por lo tanto, antes de la aplicación del algoritmo KMEANS se realiza la reducción de dimensionalidad con esta cantidad de componentes principales.

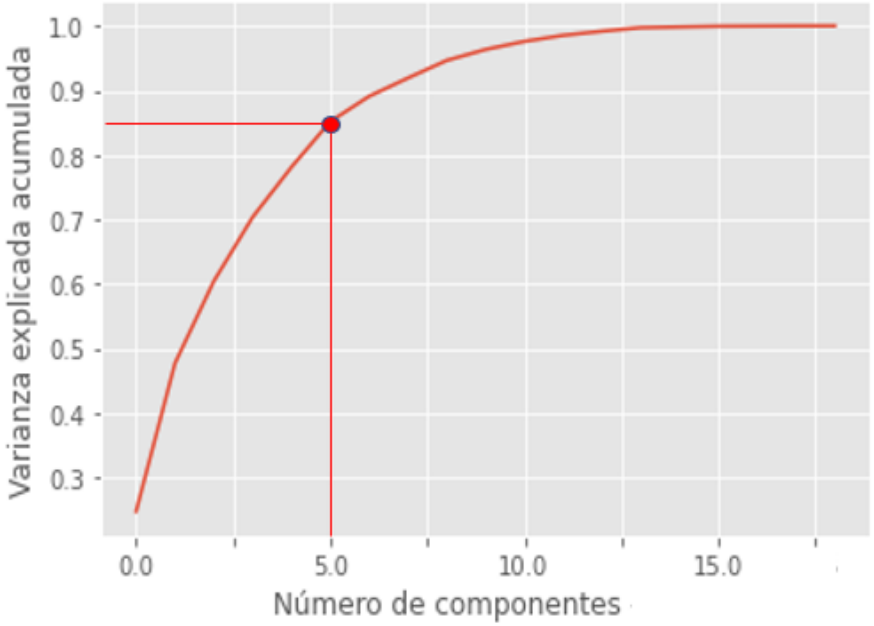
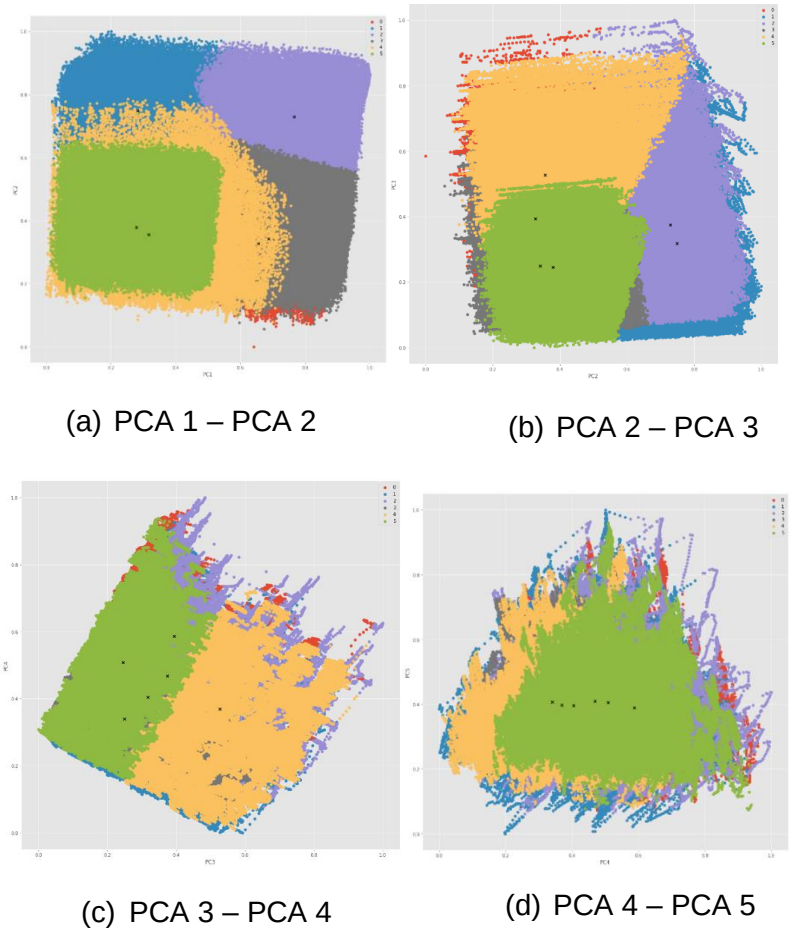


Figura 5- 2: Varianza explicada por componentes PCA.

En la Figura 5-3 se presenta el análisis de PCA por componentes junto con la técnica KMEANS con 6 grupos. En los resultados obtenidos se identifican grupos encima de otros, por lo que no se encuentran resultados prometedores para esta técnica con los datos de los tractocamiones graneleros.



(a) PCA 1 – PCA 2

(b) PCA 2 – PCA 3

(c) PCA 3 – PCA 4

(d) PCA 4 – PCA 5

Figura 5- 3: Comparativo PCA-KMEANS por pares de componentes principales. PCA1-PCA2 (a), PCA2-PCA3 (b), PCA3-PCA4 (c) y PCA4-PCA5 (d). Tractocamión caso número 10.

Se encuentra que con la estrategia de agrupamiento PCA-KMEANS no se cuenta con un agrupamiento de clases marcado con seis grupos resultado del algoritmo KMEANS. Los datos no son desagregados adecuadamente con este método en la proyección de los datos a un espacio de menores dimensiones, indica que la relación que se tiene entre las variables no cumple linealidad entre ellas, el algoritmo de KMEANS como técnica de agrupamiento no tiene un buen desempeño al generar los grupos en los datos de los tractocamiones graneleros.

5.1.1.2 t-SNE – KMEANS

Como se describe en el capítulo 4.2, para la selección de los hiperparámetros a elegir con el método t-SNE se asigna la perplejidad con valor de 100 y 5000 iteraciones para la ejecución del algoritmo KMEANS.

En la Figura 5-4 se relaciona el resultado del algoritmo de reducción t-SNE ajustado en hiperparámetros junto con seis grupos generados con el método KMEANS. Donde se encuentra que el algoritmo t-SNE no genera una proyección en dos dimensiones donde los datos se encuentren desagregados, se generan seis grupos con la técnica KMEANS donde tampoco se encuentra un resultado adecuado en términos de la visualización de la estructura global de los datos (distancia entre grupos) y las estructuras locales (agrupamiento de puntos de datos vecinos) como lo define (ICHI.PRO, 2020).

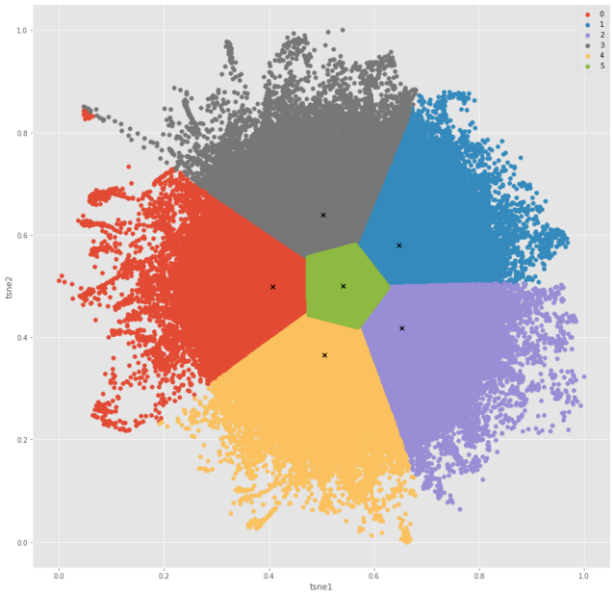


Figura 5- 4: t-SNE ajustado en hiperparámetros con 6 grupos para KMEANS. Tractocamión caso número 10.

La técnica t-SNE-KMEANS no presenta potencial para la detección de patrones de falla en los sistemas térmicos desde el análisis de la proyección de los datos.

5.1.1.3 UMAP - KMEANS

Como se describe en el capítulo 4.3, para la selección de los hiperparámetros a elegir con el método UMAP se asigna la cantidad de vecinos con valor de 100 y la distancia mínima de 0.0 para lograr encontrar pequeños componentes conectados, como se describe anteriormente desde los conceptos básicos del algoritmo y posteriormente la ejecución del algoritmo KMEANS con seis grupos.

En la Figura 5-5 se detalla el resultado de la reducción de dimensionalidad con UMAP y la aplicación de KMEANS con dos grupos en el tractocamión caso número diez con el fin de generar dos clases y posicionar los puntos correspondientes a zonas de atención en taller, donde los puntos rojos dentro del grupo naranja generado por KMEANS corresponden a un día antes del tractocamión ser intervenido por sistemas térmicos y los puntos verdes en el grupo azul también generado por KMEANS corresponden a los datos de un día después de las fechas que el tractocamión salía de la intervención en taller. Este conjunto de datos consta de 1.062.899 datos de las 18 variables de la tesis.

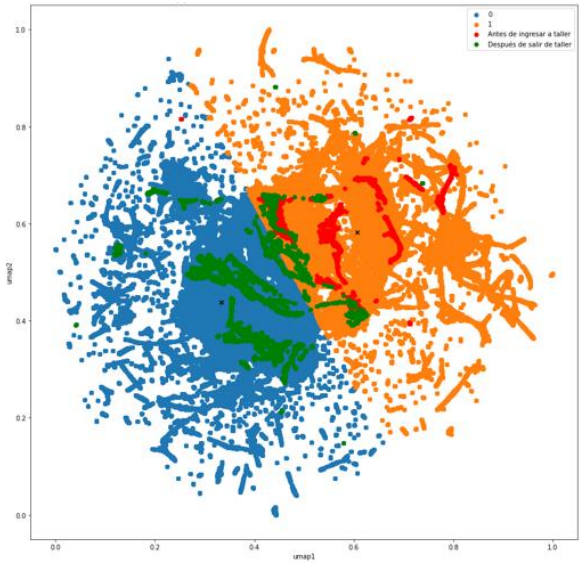
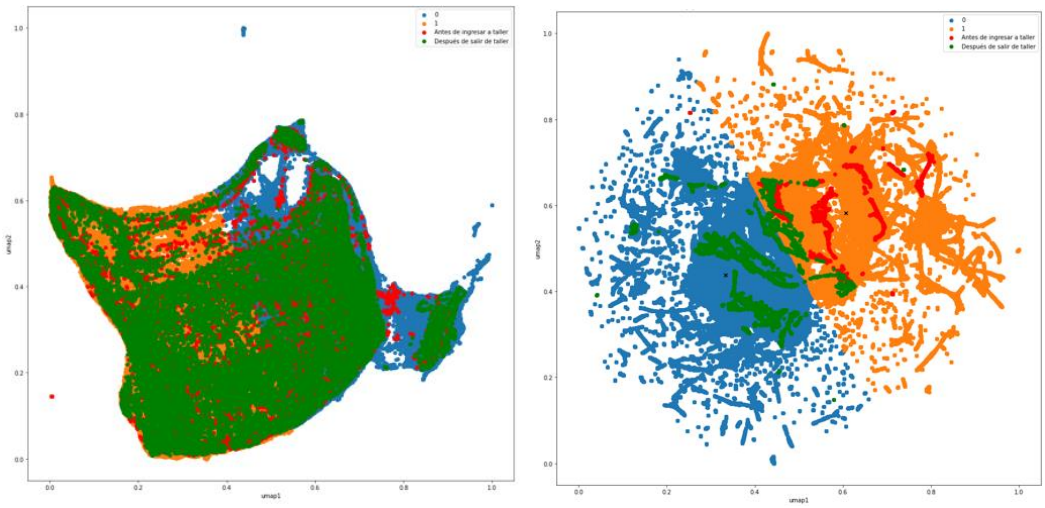


Figura 5- 5: UMAP con hiperparámetros ajustados con 2 grupos de KMEANS para el tractocamión caso número 10.

Con las experimentaciones realizadas se encuentra que una estrategia para la selección de tractocamiones, consiste en seleccionar los vehículos con menor cantidad de ingresos a taller para intervenciones en sistemas térmicos como es el caso número 10 con respecto a vehículos que tengan mayor cantidad de ingresos, ya que como hipótesis, estos tractocamiones pueden tener daños en otros sistemas mecánicos que enmascaran los daños térmicos y el agrupamiento de datos no es específico para la detección de estos patrones de falla como se presenta en la Figura 5-6, comparando el tractocamión con caso número 10 con pocas intervenciones contra el tractocamión caso número 11 con mayor cantidad de intervenciones en taller en el año 2020.



(a) (b)

Figura 5- 6: UMAP con hiperparámetros ajustados con 2 grupos de KMEANS tractocamiones caso número 11 (a) y caso número 10 (b).

En el tractocamión caso número 11 (a) se encuentra que al posicionar los datos antes de ingreso a taller (puntos rojos) y después de salir de taller (puntos verdes) se ubican en las dos regiones donde se generan los grupos con KMEANS, no es posible detectar los patrones de falla con los datos de este tractocamión que presenta más intervenciones de taller. En el tractocamión caso número 10, los datos de taller se posicionan según su estado en cada uno de los dos grupos generados, por lo que la estrategia indica realizar así la selección de los vehículos a estudiar para el fenómeno de interés.

Se concluye que en la evaluación de esta técnica, se dividen mejor las categorías de agrupamiento con respecto a los métodos de PCA-KMEANS y t-SNE-KMEANS y con la validación de las zonas donde el vehículo está a punto de ingresar a taller y un día después de la intervención. El análisis de los resultados del tractocamión con caso número nueve, indica que se puede realizar experimentación para evaluar clasificadores en la determinación de patrones de falla.

5.1.2 DBSCAN

El algoritmo DBSCAN es una técnica de agrupamiento basada en densidad, sus dos parámetros principales son *Epsilon*, que representa la distancia máxima que puede tener cualquier punto con el punto central en el agrupamiento. El segundo parámetro corresponde a la cantidad de puntos mínimos que se debe tener en el vecindario para generar los grupos en el conjunto de datos.

En el momento de seleccionar los valores para estos parámetros se presentan dificultades con su elección ya que pueden variar los resultados en el agrupamiento. Sin embargo, como se menciona en (Sefidian, 2020), una estrategia para seleccionar *Epsilon* se basa en calcular la distancia promedio entre cada punto y sus k vecinos más cercanos, en el gráfico se organizan los puntos contra las distancias definidas y en el punto donde se presenta la máxima curvatura se encuentra el valor óptimo para *Epsilon*.

Finalmente, como estrategia para seleccionar la cantidad mínima de puntos se calcula con la formula *MinPts* = 2**Dim*, donde *Dim* corresponde a la dimensión del conjunto de datos, como estrategia también se puede seleccionar como la cantidad mínima de puntos el valor que representa la cantidad de dimensiones.

Para el conjunto de datos de experimentación se selecciona el tractocamión con caso número diez con los registros en los que el vehículo presenta ingresos a taller en el año 2020, el conjunto de datos cuenta con 13.032 datos de las 18 variables de la tesis.

Una estrategia futura en la implementación de esta técnica con una de sus variaciones es con la técnica OPTICS, capaz de ejecutar el agrupamiento detectando automáticamente

los grupos a generar como se describe en la sección 2.4.3. Por restricciones computacionales actuales en la elaboración de esta tesis, no se realiza esta técnica pero en trabajos futuros se puede explorar su implementación.

5.1.2.1 PCA – DBSCAN

Como se define en la sección anterior, en la selección de los hiperparámetros implementados con la técnica DBSCAN, el valor de la cantidad mínima de puntos es cinco, teniendo en cuenta que esta es la cantidad de características o de componentes que tiene el resultado de la reducción de dimensionalidad con PCA. Para la selección de *Epsilon* se relaciona en la Figura 5-7 el grafico de distancias para seleccionar el valor adecuado. Del resultado donde las distancias no cambian significativamente se encuentra que el valor óptimo para *Epsilon* es de 0.05

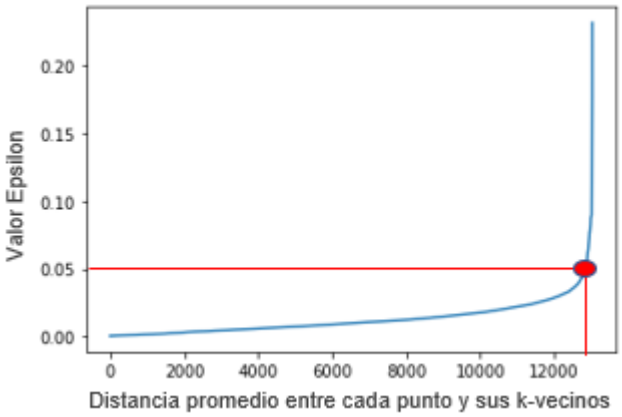
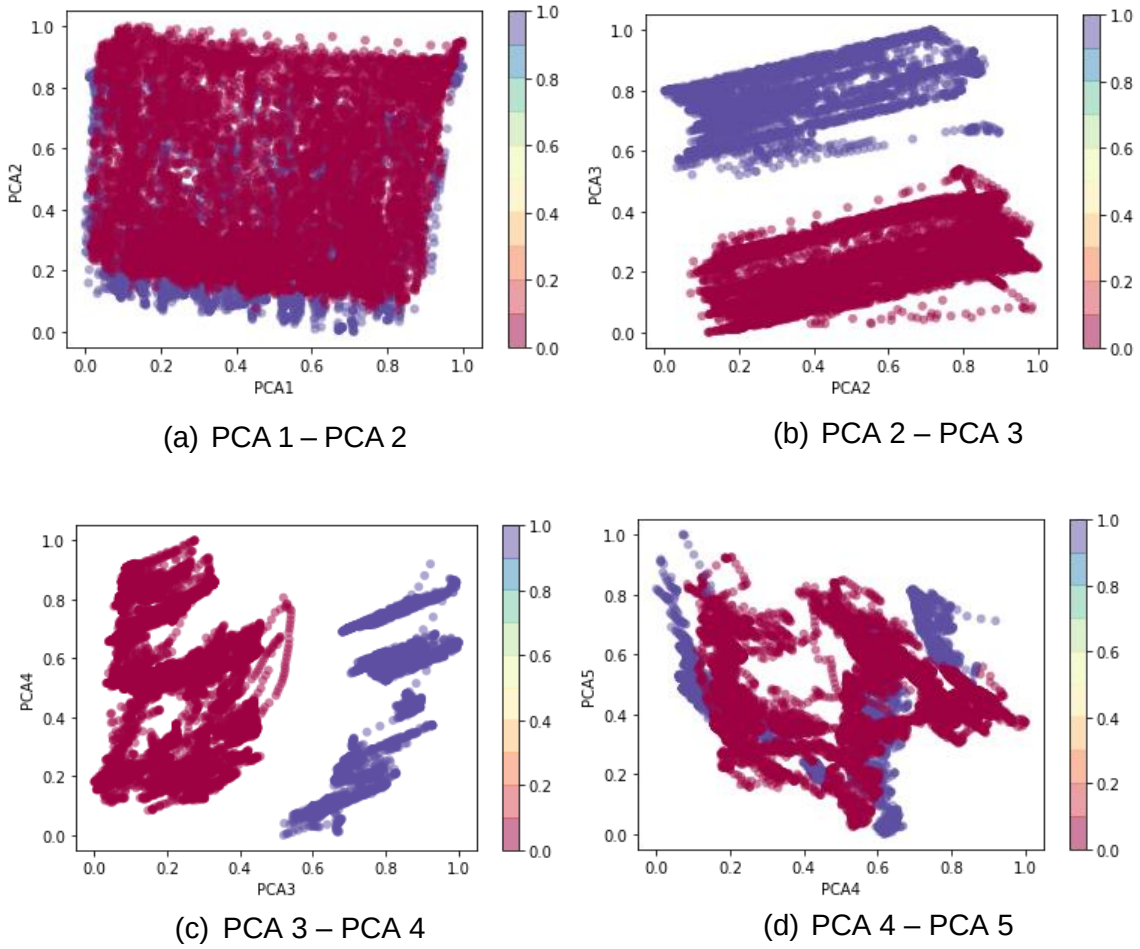


Figura 5- 7: Distancias vecinos más cercanos y selección de *Epsilon* PCA-DBSCAN tractocamión caso número 10.

Con la selección de los anteriores hiperparámetros, se presenta en la Figura 5-8, el análisis por pares de componentes principales en el resultado de agrupamiento con DBSCAN.



(a) PCA 1 – PCA 2

(b) PCA 2 – PCA 3

(c) PCA 3 – PCA 4

(d) PCA 4 – PCA 5

Figura 5- 8: Agrupamiento con PCA-DBSCAN por pares de componentes principales. PCA1-PCA2(a), PCA2-PCA3(b), PCA3-PCA4(c) y PCA4-PCA5(d). Tractocamión caso número 10.

Se encuentra que en el análisis por pares de componentes principales, los datos no son agrupados en (a) y (d), mezclándose los dos grupos generados en esta experimentación. Mientras que en (b) y (d) los datos son separados y los dos grupos resultado de DBSCAN son separados entre sí. En el análisis de estos comparativos por componentes principales,

no se observa bien el ejercicio de agrupamiento con DBSCAN implementando la reducción de dimensionalidad de PCA con el análisis por pares de componentes en general.

5.1.2.2 t-SNE – DBSCAN

Para realizar la selección de los hiperparámetros en el método DBSCAN, el valor de la cantidad mínima de puntos es cuatro, teniendo en cuenta que el algoritmo t-SNE realiza la reducción de dimensionalidad a dos componentes, como se describe en la sección 5.1.2.

Para la selección de *Epsilon* se relaciona en la Figura 5-9 el grafico de distancias para seleccionar el valor adecuado. Se encuentra que el valor óptimo para *Epsilon* es de 0.003

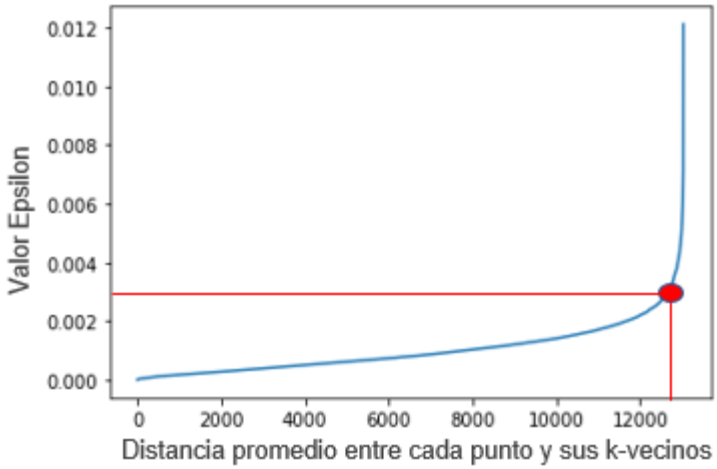
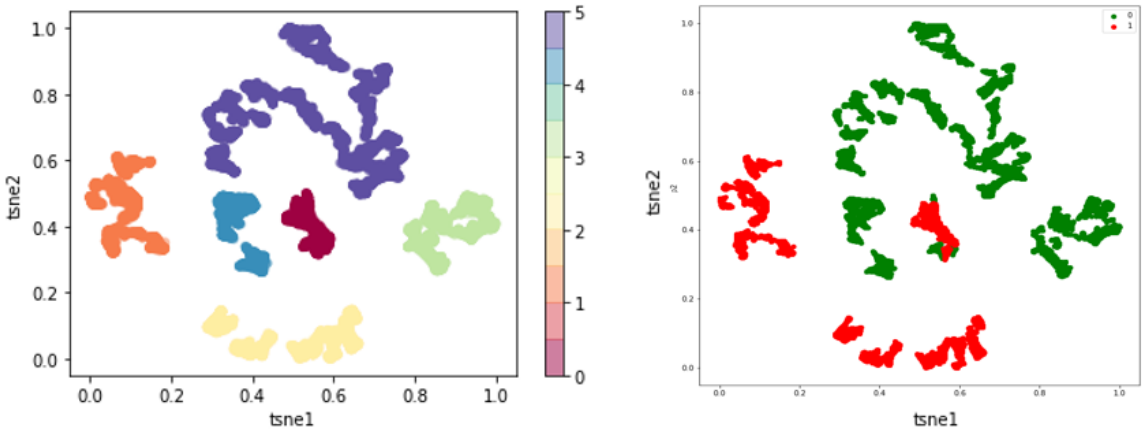


Figura 5- 9: Distancias vecinos más cercanos selección de *Epsilon* t-SNE y DBSCAN. Tractocamión caso número 10.

Con la selección de los anteriores hiperparámetros, se presenta en la Figura 5-10 (a), el resultado de la implementación del algoritmo de reducción de dimensionalidad t-SNE con los hiperparámetros ajustados y la aplicación del algoritmo de DBSCAN con *Epsilon*=0.003 y *MinPts*=4. En la Figura 5-10 (b) se presenta el resultado de este mismo agrupamiento con DBSCAN, donde los puntos en color rojo representan los datos antes de ingresar a taller el tractocamión y los puntos en color verde representan los datos luego de salir de taller.



(a) (b)

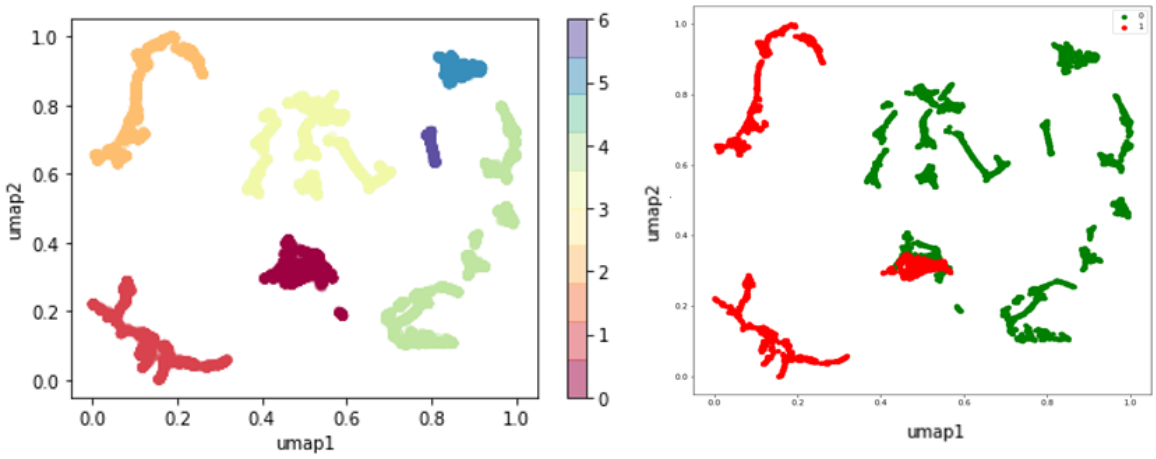
Figura 5- 10: Agrupamiento con t-SNE y DBSCAN. Resultado algoritmo con hiperparámetros ajustados(a) y datos con estados de taller (b). Tractocamión caso número 10.

Se encuentra que los datos para este caso de experimentación son desagregados marcadamente, el algoritmo DBSCAN detecta seis grupos separados entre sí como se visualiza en (a), en la comparación con (b), los datos antes de ingresar a taller se ubican en los grupos de color naranja, amarillo y vinotinto, en este último grupo también se encuentran algunos datos del tractocamión cuando sale de taller. Finalmente, los datos del tractocamión después de salir de taller se ubican en los grupos de color vinotinto, azul purpura y verde.

La técnica t-SNE con el algoritmo de DBSCAN para esta experimentación es potencial para la detección de patrones de falla con posibles afectaciones en un ejercicio de clasificación posterior por la intersección de datos de taller en el grupo color vinotinto.

5.1.2.3 UMAP – DBSCAN

En la Figura 5-11 se relaciona el resultado de la implementación del algoritmo de reducción de dimensionalidad UMAP con los hiperparámetros ajustados y la aplicación del algoritmo de DBSCAN con *Epsilon*=0.00075 y *MinPts*=4.



(a) (b)

Figura 5- 11: UMAP-DBSCAN. Resultado algoritmo con hiperparámetros ajustados(a) y datos con estados de taller (b). Tractocamión caso número 10.

Los datos son desagregados también con la implementación del algoritmo UMAP, DBSCAN detecta siete grupos separados entre sí como se visualiza en (a), comparando con (b), los datos antes de ingresar a taller se ubican en los grupos de color naranja, rojo y vinotinto, en este último grupo también se encuentran algunos datos del tractocamión cuando sale de taller. Finalmente, los datos del tractocamión después de salir de taller se ubican en los grupos de color vinotinto, amarillo, purpura, azul y verde.

La técnica UMAP con el algoritmo DBSCAN para esta experimentación también es potencial para la detección de patrones de falla y con posibles afectaciones en un ejercicio de clasificación posterior por la intersección de datos de taller en el grupo color vinotinto que también se presenta en este caso.

5.1.3 MAPAS AUTOORGANIZADOS (*Self Organizing Maps*)

Como lo menciona (Forest et al., 2019), los mapas autoorganizados (SOM) realizan agrupaciones y visualización de manera simultánea con la proyección de datos de alta dimensionalidad en un mapa de baja dimensión con una tipología de cuadrícula. Dicha cuadricula se compone de neuronas o células donde cada neurona está asociada con un vector prototipo del espacio de datos original o vector de código.

Se realiza la implementación del algoritmo SOM tomando como fuente de información los datos de los diez tractocamiones que presentaron mayor cantidad de ingresos a taller durante el año 2020 para intervenciones o reparaciones en los sistemas térmicos, donde los tractocamiones ingresaron 26 veces a taller entre todos. El algoritmo SOM es capaz de separar y generar visualización de los resultados por cada una de las variables, en esta experimentación se analizan 10.416.307 datos de las 18 variables de la tesis con un tiempo de ejecución de 20 horas del algoritmo

Cuando se implementa el algoritmo por cada millón de datos el tiempo promedio de ejecución que se obtiene es de 30 minutos, lo que permite analizar poblaciones de una manera más efectiva y sin uso de computación distribuida en la nube.

En la Figura 5-12 se relaciona la visualización de este método de agrupamiento para los datos de los diez tractocamiones con mayor cantidad de ingresos a taller en el año 2020, representando el agrupamiento para cada una de las 18 variables de la tesis.

Los gráficos resultado correspondientes a las variables son:

(a): Combustible total utilizado, (b): Combustible total utilizado en ralentí, (c): Combustible de viaje utilizado en ralentí, (d): Combustible de viaje utilizado, (e): Nivel combustible, (f): Nivel DEF, (g): Nivel refrigerante, (h): Odómetro, (i): Voltaje del dispositivo de telemetría, (j): Temperatura del aceite del motor, (k): Temperatura del refrigerante del motor, (l): Temperatura exterior, (m): Tensión de arranque, (n): Tiempo de funcionamiento del motor, (o): Vehículo activo, (p): Velocidad del motor, (q): *Ignition*, (r): Velocidad en carretera del motor.

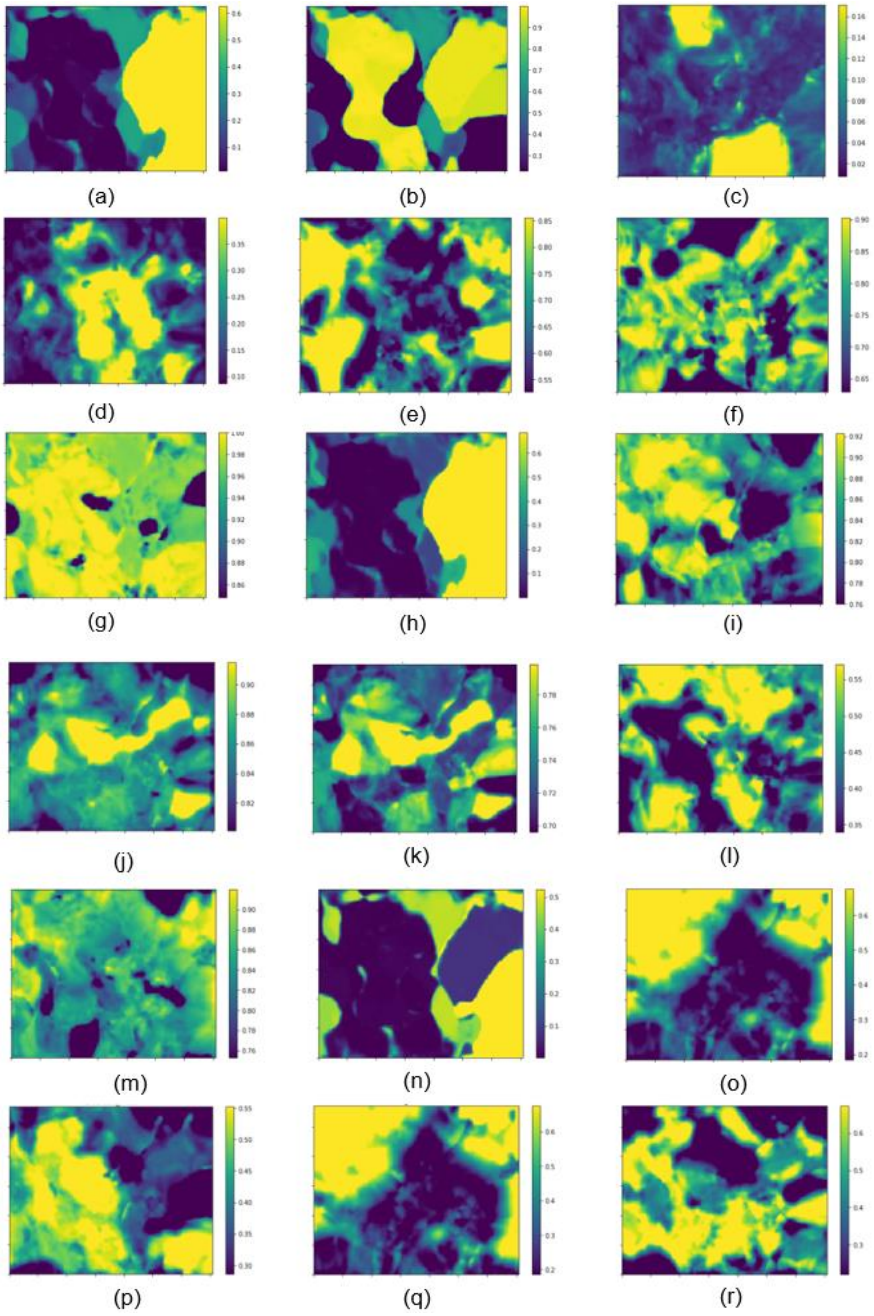


Figura 5- 12: Mapas auto organizados por variable para los diez tractocamiones con mayor cantidad de ingresos a taller en el año 2020. (a): Combustible total utilizado, (b): Combustible total utilizado en ralentí, (c): Combustible de viaje utilizado en ralentí, (d): Combustible de viaje utilizado, (e): Nivel combustible, (f): Nivel DEF, (g): Nivel refrigerante, (h): Odómetro, (i): Voltaje del dispositivo de telemetría, (j): Temperatura del aceite del motor, (k): Temperatura del refrigerante del motor, (l): Temperatura exterior, (m): Tensión

de arranque, (n): Tiempo de funcionamiento del motor, (o): Vehículo activo, (p): Velocidad del motor, (q): *Ignition*, (r): Velocidad en carretera del motor.

En el resultado de agrupamiento del algoritmo se identifican patrones o zonas demarcadas en tonalidad de azul y amarillo. En los patrones de color azul se encuentran agrupados los valores más bajos para cada variable y en los patrones de color amarillo se encuentran agrupados los datos con los valores más altos. Se busca encontrar si estos grupos generados pueden agrupar los datos de la variable que se puedan considerar particulares al comportamiento promedio y que puedan indicar los valores de la variable un estado de falla allí, con la comparación de los valores antes de ingresar y luego de salir de taller para los tractocamiones.

Como caso particular y de análisis, teniendo en cuenta la potencia en la visualización de resultados del algoritmo SOM se realiza una revisión aislada al tractocamión caso número dos de la prueba de escritorio para las variables de temperatura.

En la Figura 5-13 se presentan las cuadriculas o resultados del algoritmo SOM para las variables de temperatura del aceite del motor (a) y temperatura del refrigerante del motor (b) y se excluye del análisis la temperatura exterior al ser una variable externa y teniendo en cuenta que cuando el tractocamión se interviene en taller esta variable no cambia por su naturaleza.

Se encuentran patrones similares en el agrupamiento que determinan una relación entre las variables temperatura del aceite del motor (a) y temperatura del refrigerante del motor(b). Sin embargo, en (b) se encuentran zonas por colores más difusas, mientras que en (a) las zonas cuentan con tonalidades más fuertes (color amarillo más fuerte), lo que indica que se concentran datos con valores más altos en la temperatura del aceite del motor contra la temperatura del refrigerante del motor. El anterior análisis lo indica también los valores medios para estas temperaturas presentadas en la Tabla 5-1 y la Tabla 5-2, donde el valor medio de la temperatura del aceite del motor es mayor que la temperatura del refrigerante del motor.

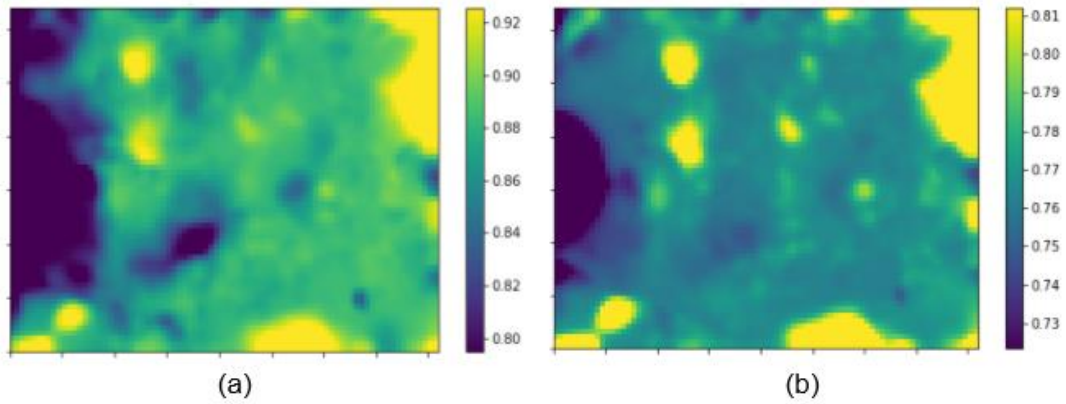


Figura 5- 13: Mapas auto organizados tractocamión caso número dos – Variables de temperatura. Temperatura del aceite del motor (a) y temperatura del refrigerante del motor(b).

En la Figura 5-14 se relacionan los histogramas de las variables de temperatura de aceite del motor (a) y del refrigerante del motor (b), donde se analiza desde los registros donde la placa estuvo tres días antes de entrar a taller y tres días después de la intervención en taller.

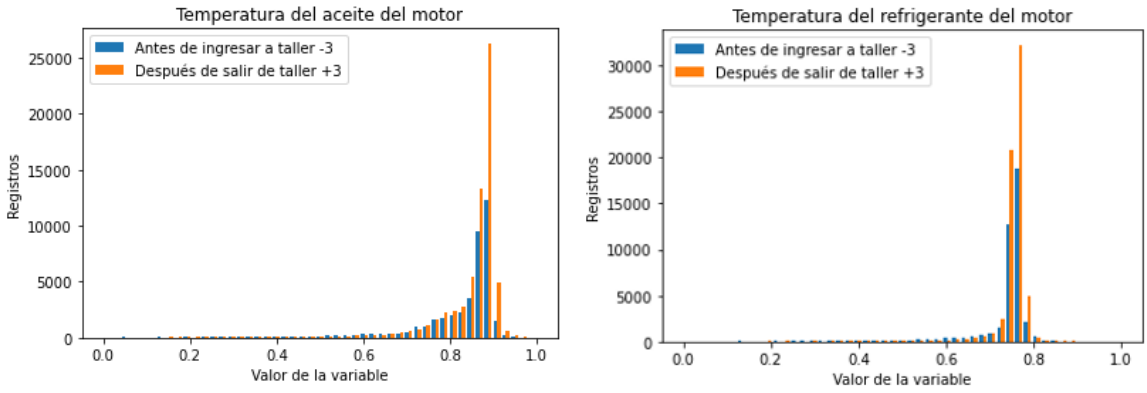


Figura 5- 14: Histogramas temperaturas de aceite del motor (a) y refrigerante del motor (b) antes y después de salir de taller.

En la Tabla 5-1 se presenta la estadística descriptiva de las variables tres días antes de las fechas en que ingresó a intervenciones en taller para los sistemas térmicos el tractocamión, como también se detalla en la tabla 5-2 la estadística descriptiva de las variables tres días después que el tractocamión fue intervenido en taller.

Tabla 5- 1. Análisis descriptivo de temperaturas antes de ingresar a taller tractocamión caso número dos.

Análisis tres días antes de ingresar a taller	Temperatura del aceite del motor	Temperatura del refrigerante del motor
Registros	41690	41690
Media	0.8139	0.7313
Desviación estándar	0.1250	0.0909
Mínimo	0.0120	0.0000
25%	0.7977	0.7440
50%	0.8675	0.7559
75%	0.8795	0.7662
Máximo	0.9518	0.8571

Tabla 5- 2. Análisis descriptivo de temperaturas después de salir de taller tractocamión caso número dos.

Análisis tres días después de salir de taller	Temperatura del aceite del motor	Temperatura del refrigerante del motor
Registros	65690	65690
Media	0.8437	0.7473
Desviación estándar	0.0930	0.0620
Mínimo	0.0482	0.0519
25%	0.8434	0.7504
50%	0.8766	0.7582
75%	0.8810	0.7662
Máximo	0.9759	0.9221

Del análisis descriptivo se encuentra que los valores de las medias para la temperatura del aceite del motor son menores antes de que el vehículo ingresara a taller en comparación a las mediciones cuando salió de taller (0.8139 < 0.8437) teniendo como referencia tres días antes del ingreso al taller y tres días después de salir del taller. Del histograma se puede observar que aproximadamente en el valor de 0.9 se presentan más mediciones en la temperatura del aceite del motor después que el tractocamión sale de la intervención en taller. A medida que el aceite se calienta a temperaturas de operación, se va volviendo menos delgado y puede proteger el motor mejor que un aceite a temperaturas más bajas. También puede absorber más calor como se menciona en (Car and Driver, 2019).

Por lo anterior y con base en el resultado del agrupamiento con SOM las zonas amarillas donde están agrupados los valores entre 0.90 y 0.92 para la temperatura del aceite del motor pueden corresponder a las temperaturas de operación del tractocamión y las zonas verde azuladas corresponden a temperaturas menores de operación.

En el análisis descriptivo de la temperatura del refrigerante del motor se encuentra que los valores son menores antes de que el vehículo ingresara a taller en comparación a las mediciones cuando salió de taller (0.7313 < 0.7473). Del histograma se puede observar que en el valor de 0.8 se presentan más mediciones, valor de la zona amarilla en el resultado de SOM (Después de salir de taller).

Con base en el resultado del agrupamiento con SOM las zonas amarillas donde están agrupados los valores entre 0.80 y 0.81 pueden corresponder a las temperaturas de operación del tractocamión para la temperatura del refrigerante del motor y las zonas verde azuladas corresponden a temperaturas menores de operación.

5.2 SPECTRAL CLUSTERING. Agrupamientos con datos temporales

Como definen (Castellanos & Rodríguez, 2011), “el agrupamiento en series de tiempo presenta como objetivo resaltar la estructura inherente de un conjunto de datos de serie de tiempo agrupando los datos en un numero de grupos homogéneos, de forma que la similitud entre datos de un grupo es máxima”.

En este capítulo se detallan los resultados obtenidos con el método *Spectral Clustering* para la variable temperatura del aceite del motor, aplicado al tractocamión caso número cuatro, el segundo vehículo con más ingresos a taller en el año 2020. Este tractocamión ingresa a taller el 25 de agosto de 2020, se analiza el comportamiento de la variable tres días antes de esta fecha y tres días después.

En la Figura 5-15 se presentan los resultados del agrupamiento, identificando la zona donde la placa ingresa a taller, como también los patrones que detecta el algoritmo antes y después de salir de taller para la variable temperatura del aceite del motor.

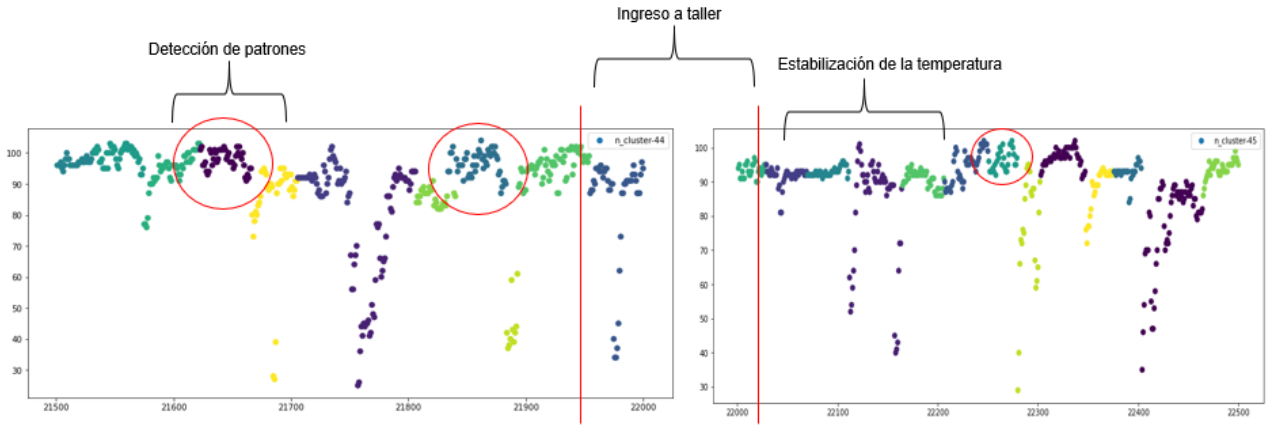


Figura 5- 15: *Spectral Clustering* para tractocamión caso número cuatro.

Del anterior resultado, se puede identificar el agrupamiento cuando se generan cambios estadísticos en la temporalidad de la variable, el algoritmo puede detectar estos cambios, identificando patrones atípicos en el comportamiento de los datos.

Sin embargo, no se encuentran patrones en el comportamiento de la variable antes de ingresar a taller y luego de salir de taller analizando las regresiones en estas fechas para los 13 eventos que se presentaron en el año 2020 donde el tractocamión tuvo que ser intervenido por sistemas térmicos y como se detalla en la Tabla 5-3.

Tabla 5- 3. Análisis descriptivo de temperaturas después de salir de taller tractocamión caso número dos.

		Antes de ingreso a taller			Luego de salida a taller			
Muestra	**Fecha**	**Ecuación**	**Pendiente**	**R2**	**Ecuación**	**Pendiente**	**R2**	**Cambio en la pendiente de la recta**
1	3-feb	-0.8321x + 1763	-0.8321	0.7647	1.8476x - 3767.8	1.8476	0.7887	Negativa a positiva
2	9-mar	0.0851x - 230.21	0.0851	0.2886	-0.0349x + 236.96	-0.0349	0.1044	Positiva a negativa
3	10-mar	2.3462x - 9316.2	2.3462	0.7582	-0.7363x + 3136.7	-0.7363	0.8064	Positiva a negativa
4	19-mar	0.0588x - 234.99	0.0588	0.0789	0.0438x - 158.99	0.0438	0.0859	Positiva a positiva
5	22-abr	0.0692x - 324.25	0.0692	0.3847	0.1165x - 637.91	0.1165	0.1966	Positiva a positiva
6	21-may	0.3742x - 3750.9	0.3742	0.2729	-0.07x + 829.24	-0.07	0.0711	Positiva a negativa
7	1-jun	0.0994x - 1051.2	0.0994	0.2243	1.1284x - 13054	1.1284	0.6545	Positiva a positiva
8	25-jun	0.1095x - 1362.8	0.1095	0.3337	-0.2431x + 3406.6	-0.2431	0.1003	Positiva a negativa
9	14-jul	0.1676x - 2693.6	0.1676	0.0941	1.115x - 18726	1.115	0.6352	Positiva a positiva
10	28-jul	-0.2515x + 4866.8	-0.2515	0.1206	0.1583x - 2946.4	0.1583	0.3791	Negativa a positiva
11	31-jul	-0.2219x + 4424.5	-0.2219	0.0985	1.5974x - 31401	1.5974	0.7955	Negativa a positiva
12	25-ago	-0.0275x + 690.13	-0.0275	0.0247	1.844x - 40087	1.844	0.8139	Negativa a positiva
13	21-sep	2.33x - 57149	2.33	0.5376	1.3036x - 32094	1.3036	0.7022	Positiva a positiva

En los 13 eventos, analizando el comportamiento de las rectas de regresión antes y luego de taller no se encuentra un patrón que indique el cambio en la pendiente de la recta de regresión cuando es intervenida la placa. El 31% de los eventos presentan datos con valores decrecientes para la temperatura del aceite del motor antes de ingresar a taller y luego de salir de taller valores crecientes en la variable. El 31% de los eventos, valores crecientes para la temperatura antes del ingreso al taller y valores decrecientes después de salir de taller. Finalmente, el 38% de los eventos presentan valores crecientes antes de ingresar a taller como también cuando el tractocamión sale de taller.

Por lo anterior, no se encuentran patrones para estos datos analizados en esta variable y que pueda ser explicado con la técnica *Spectral Clustering,* debido a que los agrupamientos realizados son asignados donde hay un cambio estadístico en la variable, pero de una manera global no se encuentra agrupamiento entre datos con distintos comportamientos.

En el siguiente capítulo se evalúa la implementación de la técnica UMAP-KMEANS para generar etiquetas en los datos de los tractocamiones graneleros y posteriormente entrenar modelos de clasificación para la detección de patrones de falla.

En la Tabla 5-4 se presenta la complejidad computacional de las técnicas no supervisadas implementadas en esta tesis.

Tabla 5- 4. Complejidad computacional para técnicas no supervisadas.

TÉCNICA	**ORDEN DEL ALGORITMO**
PCA	$O(nd \times min(n, d) + d^3)$
t-SNE	$O(n^2)$
UMAP	$O(n^{1.14})$
KMEANS	$O(nCd)$
DBSCAN	$O(nLog(n)d)$
SOM	$O(nC)$
SPECTRAL CLUSTERING	$O(n^2d + n^3 + nC^2)$

De la anterior tabla, se tienen las siguientes variables:

n: tamaño de los datos

C: número de grupos

d: dimensionalidad de los datos

Finalmente, se presentan los tiempos de cómputo de los experimentos realizados según la técnica empleada, detallados en la Tabla 5-5 donde para cada experimentación realizada se describe la estrategia de agrupamiento implementada, la cantidad de datos, el tiempo de procesamiento del algoritmo, el tipo de computación que permitió obtener los resultados, detallando si se implementó en la maquina local con 1 CPU, 12 *Gigabytes* en memoria RAM y un procesador Intel Core i5 con 2.5 Ghz o si se realizó con infraestructura computacional en la nube con 8 CPU's , 61 Gigabytes en memoria RAM y GPU con 16 *Gigabytes*.

Esta selección de infraestructuras se realiza ante la necesidad de la aplicación de las técnicas no supervisadas cuando se trabaja con grandes cantidades de información y la

maquina local no era capaz de realizar la computación por sus características propias y los requerimientos de cada algoritmo con base en su complejidad.

Tabla 5- 5. Tiempos de cómputo por técnicas no supervisadas.

Técnica	Cantidad de datos	Tiempo de procesamiento	Tipo de computación	Infraestructura
PCA(5)-KMEANS(6)	999754 filas x 5 columnas	7 segundos	Distribuida - Nube AWS	8 CPU´s, 61 GB, GPU 16 GB
t-SNE(2)-KMEANS(6)	999754 filas x 2 columnas	15 minutos	Distribuida - Nube AWS	8 CPU´s, 61 GB, GPU 16 GB
UMAP(2)-KMEANS(6)	1043297 filas x 2 columnas	1 hora	Distribuida - Nube AWS	8 CPU´s, 61 GB, GPU 16 GB
PCA(5)-DBSCAN	1314036 filas x 5 columnas	1.5 minutos	Secuencial - Local	1 CPU, 12 GB RAM, Intel Core i5 2.5 GHz
t-SNE(2)-DBSCAN	1314036 filas x 2 columnas	16 minutos	Secuencial - Local	1 CPU, 12 GB RAM, Intel Core i5 2.5 GHz
UMAP(2)-DBSCAN	1314036 filas x 2 columnas	1.25 horas	Distribuida - Nube AWS	8 CPU´s, 61 GB, GPU 16 GB
SOM	10416307 filas x 18 columnas	20 horas	Secuencial - Local	1 CPU, 12 GB RAM, Intel Core i5 2.5 GHz
SOM	892564 filas x 3 columnas	20 minutos	Secuencial - Local	1 CPU, 12 GB RAM, Intel Core i5 2.5 GHz
SPECTRAL CLUSTERING	500 datos temporales	0.4 segundos	Secuencial - Local	1 CPU, 12 GB RAM, Intel Core i5 2.5 GHz

5.3 Conclusiones

- Para la técnica DBSCAN, no se presenta un funcionamiento correcto con las técnicas PCA ya que los grupos generados se ubican unos con otros o superpuestos entre sí en el análisis por pares de componentes principales, los datos no son desagregados y la técnica DBSCAN basada en densidad no puede separar grupos con esta técnica de reducción de dimensionalidad. t-SNE genera una desagregación de los datos donde el algoritmo pueda detectar grupos separados unos de otros. El algoritmo genera la creación de grupos de datos lo que permite analizar cada grupo contra la información de ingresos y salidas de taller del tractocamión de la experimentación y posteriormente realizar modelos de clasificación en la detección de patrones de falla asociados a sistemas térmicos. La reducción de dimensionalidad con UMAP también genera resultados interesantes en la desagregación de los datos y la implementación de la técnica de agrupamiento con DBSCAN al detectar los grupos separados.

- Con la implementación del algoritmo SOM se encuentra potencial en la capacidad de procesar la información de varios tractocamiones a la vez y la proyección de los agrupamientos de datos en cada una de las variables para determinar en qué

grupos se pueden ubicar los datos de condiciones comunes de operación de las variables y en que grupos los datos que no son condiciones comunes de operación analizando en conjunto el comportamiento de las variables con los datos antes de ingresar y luego de salir de taller.

- Para la técnica SPECTRAL CLUSTERING, los agrupamientos obtenidos son asignados solamente en las secciones temporales donde hay cambios estadísticos en la información, pero de una manera global al analizar con los datos antes de ingresar y luego de salir de taller el tractocamión, no realiza agrupamiento entre datos que permitan detectar patrones en el comportamiento antes y luego de salir de taller. Aunque no se encontraron estos patrones en la experimentación es una herramienta adecuada para realizar agrupamiento en cambios temporales entre señales.

- Se concluye que luego de la evaluación de las técnicas de agrupamiento se encuentra potencial en la técnica UMAP-KMEANS al validar los datos de taller como se ubican en los grupos generados por el algoritmo, donde los datos antes de ingresar a taller se posicionan en el primer grupo y los datos luego de salir de taller el tractocamión se posicionan en el segundo grupo con un resultado satisfactorio donde el conjunto de datos de la experimentación se realiza con 1.062.899 datos respecto a 13.032 datos con que funciona la técnica DBSCAN, teniendo en cuenta que con la misma cantidad de información en experimentación, los datos en su baja desagregación no permiten generar grupos con esta última.

6. RESULTADOS

En este capítulo se presentan los resultados obtenidos para la evaluación de estrategias de agrupamiento no supervisadas, presentando la propuesta del método de reducción de dimensionalidad UMAP en conjunto con el algoritmo de agrupamiento KMEANS por los beneficios que se encontraron en la separación de grupos y detección de estados de falla en sistemas térmicos de la flota de tractocamiones graneleros.

En la sección 6.1 se implementan los algoritmos no supervisados UMAP y KMEANS para determinar dos etiquetas en el agrupamiento de datos que permitan entrenar clasificadores, evaluarlos y posteriormente realizar predicciones con los modelos obtenidos. Se presenta la experimentación realizada con los 10 tractocamiones que presentaron mayor cantidad de ingresos a taller durante el año 2020, se implementan los clasificadores *Logistic Regression, K-Neighbors, Decision Tree, Random Forest* y *MLPC* según las etiquetas del algoritmo KMEANS para entrenar los modelos de clasificación con la selección de tres tractocamiones de la muestra de estudio descrita anteriormente. Se realiza también el análisis de los resultados con la técnica SOM para los tres tractocamiones de estudio.

Con el agrupamiento generado se implementan pruebas con los clasificadores propuestos, evaluándolos con los datos antes de ingresar y luego de salir de taller para validar su desempeño.

En la sección 6.2 se propone un método supervisado de clasificación, donde se selecciona uno de los tractocamiones con menor cantidad de ingresos a taller bajo la hipótesis que con esta selección estratégica de tractocamiones se pueden implementar modelos de clasificación para la detección de fallas asociadas a sistemas térmicos.

Finalmente, en la sección 6.3 se presentan las conclusiones para los resultados del capítulo.

6.1 Agrupamiento con la estrategia UMAP-KMEANS con dos clases

Para el agrupamiento de los estados de falla se realiza la experimentación con los datos recolectados de las 18 variables de operación medidas para el año 2020 en los diez tractocamiones que más presentaron intervenciones en taller asociados a sistemas térmicos. Donde el orden se representa de manera descendente con base en la cantidad de ingresos a taller, siendo el primer tractocamión el vehículo con más ingresos a taller en el año y el décimo tractocamión el vehículo con menos ingresos a taller.

En la Figura 6-1 se presentan los resultados obtenidos de los diez tractocamiones implementado el algoritmo UMAP con la variación de hiperparámetros de distancia mínima con valor de 0.0 y cantidad de vecinos con valor de 100, en conjunto con la implementación del algoritmo KMEANS con 2 grupos.

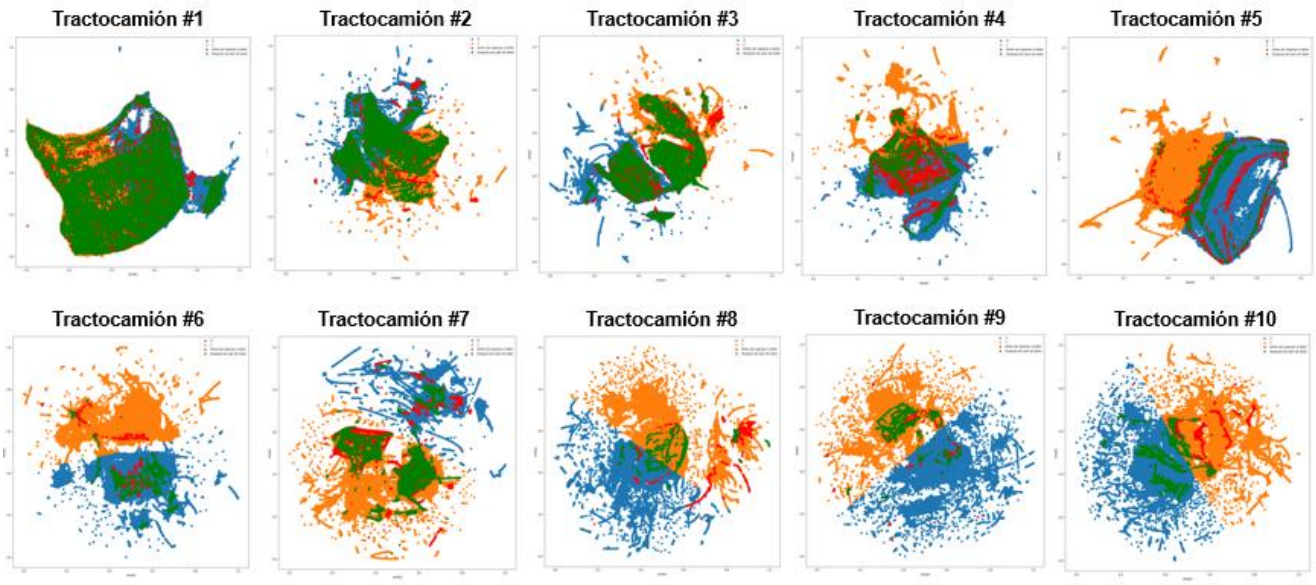


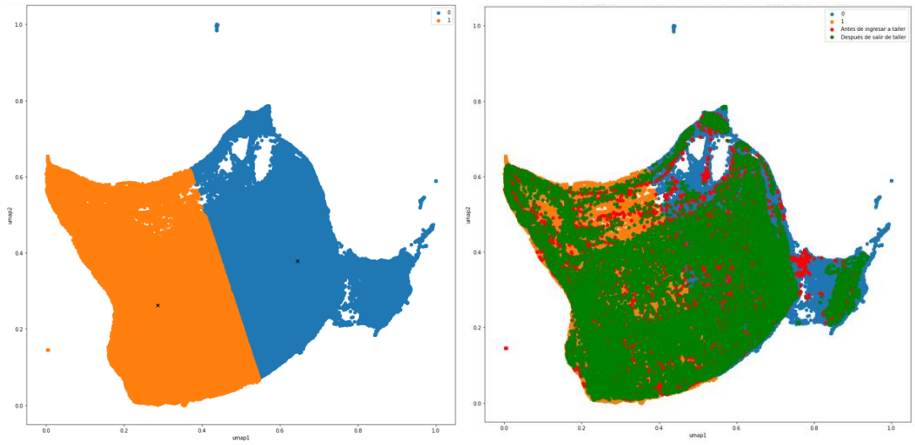
Figura 6- 1: UMAP-KMEANS con dos clases para los diez tractocamiones con mayor cantidad de ingresos a taller.

Se evidencia según esta muestra de tractocamiones que mientras mayor cantidad de ingresos a taller tenga cada tractocamión los datos antes de ingresar y luego de salir de taller se mezclan entre los grupos resultado de la técnica UMAP-KMEANS, lo que no

permite detectar estados de falla con la técnica como se visualiza en los tractocamiones #1 a #4.

Se relacionan los resultados del análisis de la estrategia de agrupamiento para tres tractocamiones de la muestra de estudio mencionada, correspondientes a los tractocamiones #1 (Cantidad mayor de ingresos a taller), #5 (Cantidad intermedia de ingresos a taller) y #10 (Cantidad menor de ingresos a taller.

En la Figura 6-2 se presentan los resultados para el tractocamión #1 de la muestra de estudio, donde se grafica la posición de los datos un día antes de ingresar a taller y un día después de salir de taller en los dos agrupamientos de la técnica UMAP-KMEANS. La figura (a) representa el resultado de agrupamiento de la técnica y la figura (b) representa la superposición de los datos en los estados antes de ingresar a taller y luego de salir de la intervención en taller con la técnica de agrupamiento.



(a) (b)

Figura 6- 2: UMAP-KMEANS con dos grupos. Agrupamiento UMAP-KMEANS (a), datos de taller en resultado UMAP-KMEANS(b). Tractocamión #1.

Con el resultado obtenido en (a) para el agrupamiento, se verifica que los datos de los estados antes de ingresar a taller (color rojo) y luego de salir de taller (color verde) se posicionan en el espacio de dos dimensiones del método UMAP y se distribuyen en los dos grupos que entrega el método KMEANS como se visualiza en (b).

El resultado esperado es encontrar que en cada clase se posicione de una manera separada los datos antes de ingresar a taller y luego de ser intervenido el tractocamión, lo cual no se nota para esta experimentación, como hipótesis se puede plantear que al ser el vehículo con más ingresos a taller en el año 2020 se pueden presentar daños no solamente en sistemas térmicos sino también en otros sistemas que enmascaran el patrón de fallos.

Se realiza la validación de los resultados con los clasificadores *Logistic Regression, K-Neighbors, Decision Tree, Random Forest y MLPC*, con respecto a los datos antes de ingresar y luego de salir de taller del tractocamión #1. En la Figura 6-3 se presentan las curvas ROC para los modelos de clasificación. Donde (a) es la curva ROC para *Logistic Regression*, (b) para *K-Neighbors*, (c) para *Decision Tree*, (d) para *Random Forest* y (e) para *MLPC*.

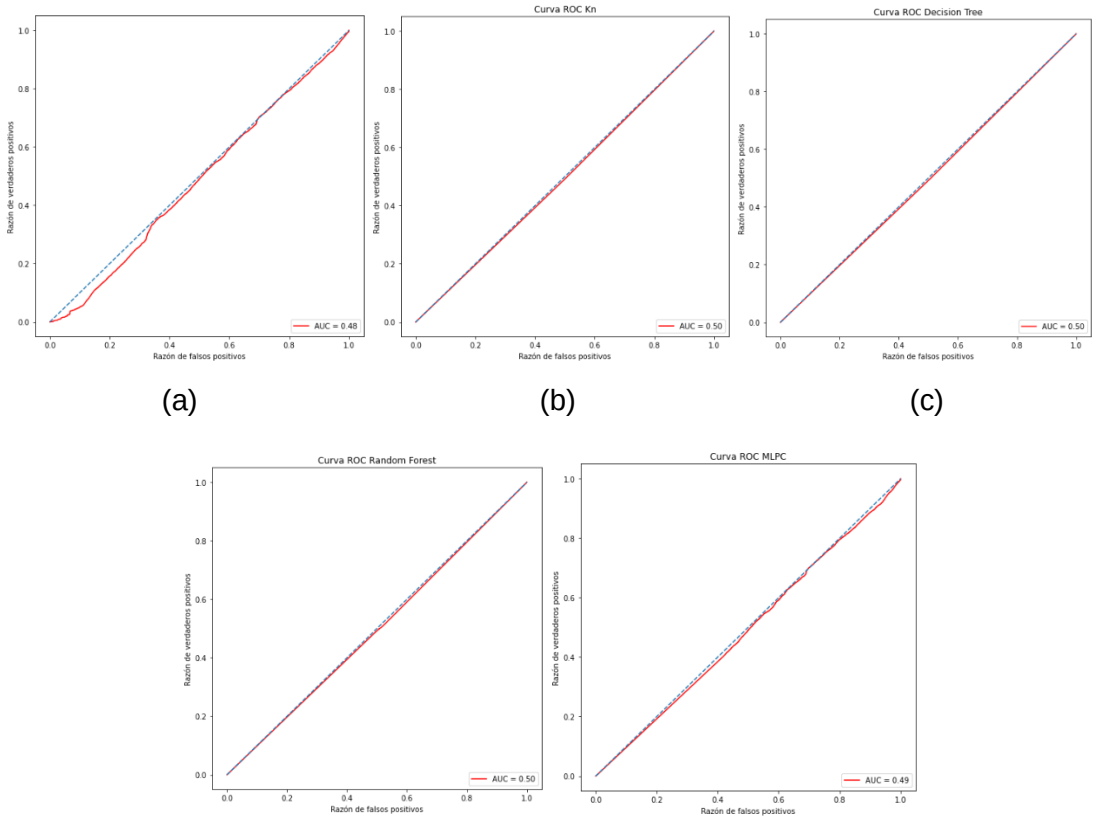


Figura 6- 3: Curvas ROC modelos de clasificación. *Logistic Regression* (a), *K-Neighbors* (b), *Decision Tree* (c), *Random Forest* (d) y *MLPC* (e). Tractocamión #1.

A continuación, en la Tabla 6-1 se detallan los valores del área bajo la curva ROC para cada clasificador. Las áreas bajo las curvas ROC con los clasificadores implementados son cercanas al valor de 0.5, lo que indica que los datos antes de ingresar y luego de salir de taller para el tractocamión #1 no se clasifican correctamente en los dos grupos resultado de la técnica UMAP-KMEANS en el tractocamión con más ingresos a taller durante el año 2020 bajo la hipótesis que se pueden presentar otros tipos de daños en este caso.

Tabla 6- 1. Área bajo la curva ROC por clasificador. Tractocamión #1.

Clasificador	Área bajo la curva ROC
Logistic Regression	0.48
K-Neighbors	0.50
Decision Tree	0.50
Random Forest	0.50
MLPC	0.49

Se presentan las matrices de confusión de cada uno de los clasificadores en la Figura 6-4 donde (a) es la matriz de confusión para *Logistic Regression*, (b) para *K-Neighbors*, (c) para *Decision Tree*, (d) para *Random Forest* y (e) para *MLPC* junto los resultados de las métricas en la Tabla 6-2.

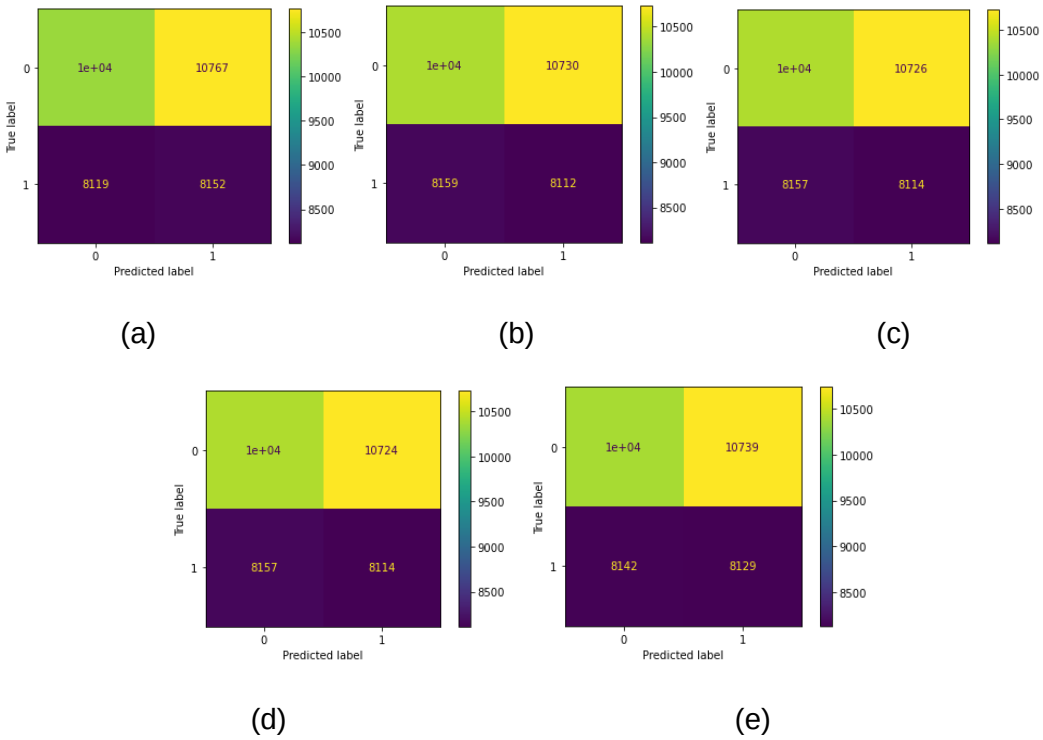


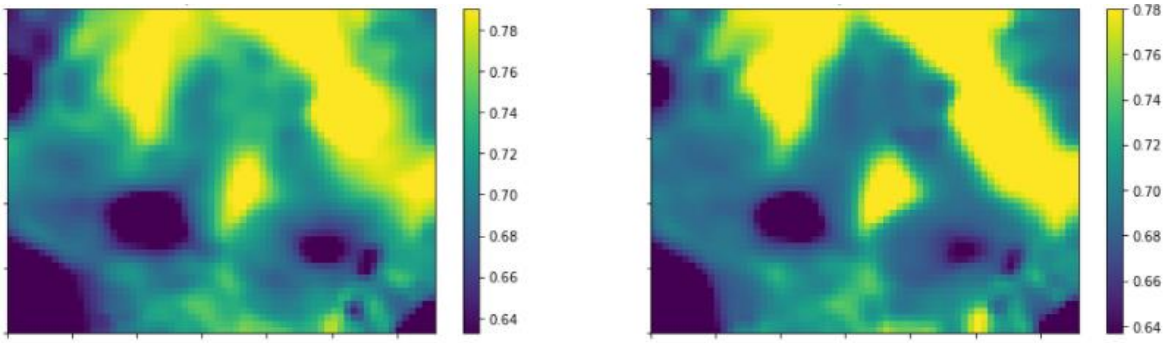
Figura 6- 4: Matrices de confusión modelos de clasificación. *Logistic Regression* (a), *K-Neighbors* (b), *Decision Tree* (c), *Random Forest* (d) y *MLPC* (e). Tractocamión #1.

Tabla 6- 2. Métricas de desempeño por clasificador tractocamión #1.

CLASIFICADOR	PRECISIÓN	EXACTITUD	ESPECIFICIDAD	SENSIBILIDAD
Logistic Regression	43.09%	49.01%	48.15%	50.10%
K-Neighbors	43.05%	48.95%	48.24%	49.86%
Decision Tree	43.07%	48.96%	48.25%	49.87%
Random Forest	43.07%	48.96%	48.25%	49.87%
MLPC	43.08%	48.98%	48.22%	49.96%

Los desempeños obtenidos en cada una de las métricas indican que los modelos propuestos no logran clasificar las clases correctamente al evaluar para los datos antes de ingresar y luego de salir de taller con los clasificadores entrenados con las etiquetas de la técnica UMAP-KMEANS. Las técnicas no supervisadas para la detección de fallas en sistemas térmicos no funcionan en el tractocamión #1 con mayor cantidad de intervenciones en sistemas térmicos.

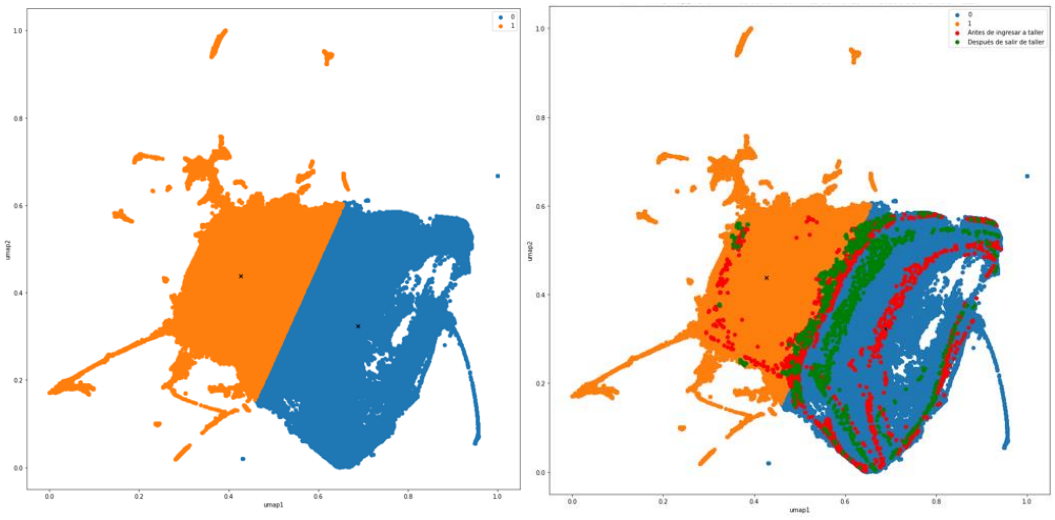
En la Figura 6-5 se presenta el resultado de la aplicación del algoritmo SOM para el tractocamión #1. Donde (a) representa el agrupamiento que se obtiene para la temperatura del aceite del motor y (b) la temperatura del refrigerante del motor. Para el análisis de estas temperaturas en este vehículo se encuentran patrones similares en el agrupamiento de los datos, lo cual para la clasificación de fallas puede ser una buena estrategia en este tractocamión con la mayor cantidad de ingresos a taller, debido a que con la técnica UMAP-KMEANS no se separan bien las clases. Esto indica que la estrategia de la selección de la placa debe contener una cantidad baja de ingresos a taller o analizar desde la implementación del algoritmo SOM.



(a) (b)

Figura 6- 5: SOM. Temperatura del aceite del motor(a) y Temperatura del refrigerante del motor(b). Tractocamión #1

La segunda experimentación corresponde al tractocamión #5 de la muestra analizada. En la Figura 6-6 se presentan los resultados de la técnica, donde la figura (a) indica el resultado de agrupamiento y la figura (b) la superposición de los datos de los estados de taller para el análisis de la separación de grupos según el patrón de falla de estudio.



(a) (b)

Figura 6- 6: UMAP-KMEANS con dos grupos. Agrupamiento UMAP-KMEANS (a), datos de taller en resultado UMAP-KMEANS(b). Tractocamión #5.

Como se visualiza en (b) los datos de los estados antes de ingresar a taller y luego de salir de taller se concentran mayoritariamente en la clase 0 del agrupamiento (color azul), por lo tanto la estrategia de agrupamiento no funciona de la manera esperada para estos datos seleccionados. Se preserva la hipótesis que el vehículo presenta también daños asociados en otros sistemas al pertenecer a esta población de vehículos con más ingresos a taller.

Se realiza la validación de los resultados con los clasificadores propuestos, con respecto a los datos antes de ingresar y luego de salir de taller del tractocamión #1. En la Figura 6-7 se presentan las curvas ROC para los modelos de clasificación. Donde (a) es la curva ROC para *Logistic Regression*, (b) para *K-Neighbors*, (c) para *Decision Tree*, (d) para *Random Forest* y (e) para *MLPC*.

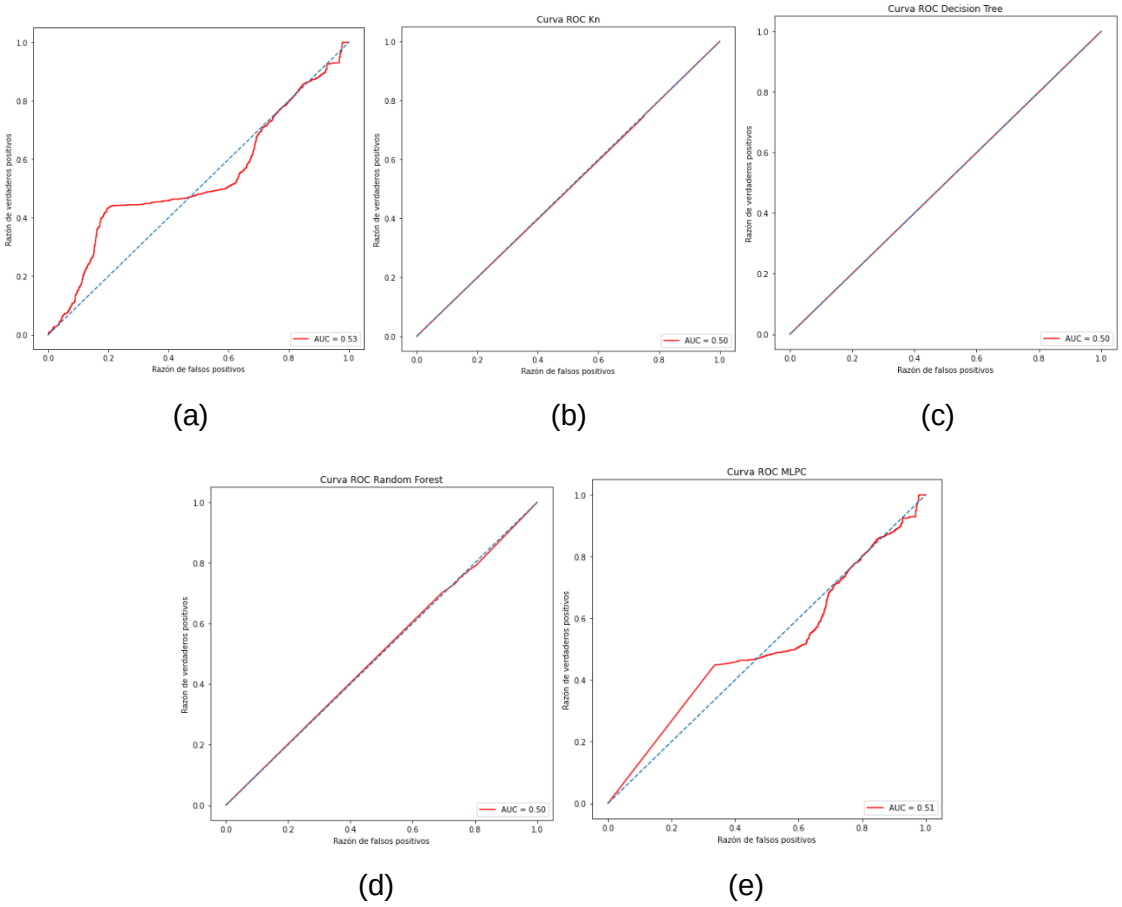


Figura 6- 7: Curvas ROC modelos de clasificación. *Logistic Regression* (a), *K-Neighbors* (b), *Decision Tree* (c), *Random Forest* (d) y *MLPC* (e). Tractocamión #5.

En la Tabla 6-3 se detallan los valores del área bajo la curva ROC para cada clasificador, con valores cercanos a 0.5, lo que indica que los datos antes de ingresar y luego de salir de taller para el tractocamión #5 no se clasifican correctamente en los dos grupos resultado de la técnica UMAP-KMEANS en el tractocamión con una cantidad intermedia de ingresos a taller en la muestra de análisis de este capítulo, bajo la hipótesis que se pueden presentar otros tipos de daños en este caso como se presentó anteriormente también con el tractocamión #1.

Tabla 6- 3. Área bajo la curva ROC por clasificador. Tractocamión #5.

Clasificador	Área bajo la curva ROC
Regression Logistic	0.53
K-Neighbors	0.5
Decision Tree	0.5
Random Forest	0.5
MLPC	0.51

Se presentan las matrices de confusión de cada uno de los clasificadores en la Figura 6-8 donde (a) es la matriz de confusión para *Logistic Regression*, (b) para *K-Neighbors*, (c) para *Decision Tree*, (d) para *Random Forest* y (e) para *MLPC* junto los resultados de las métricas en la Tabla 6-4.

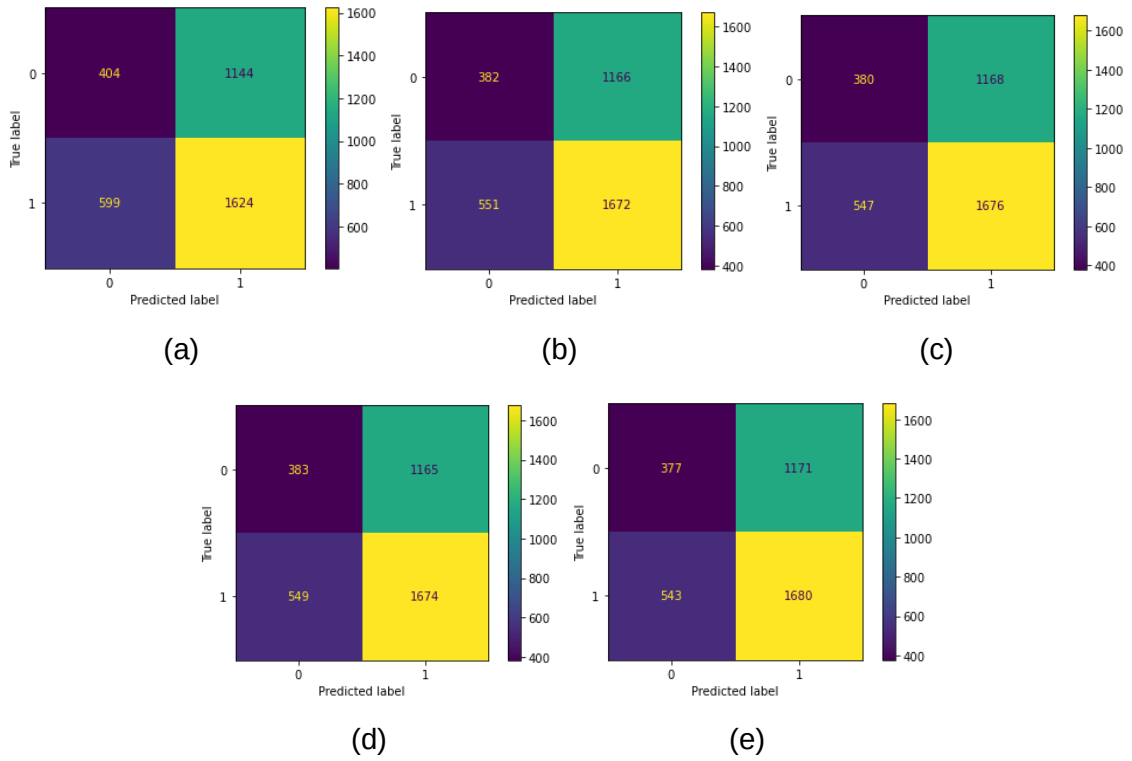


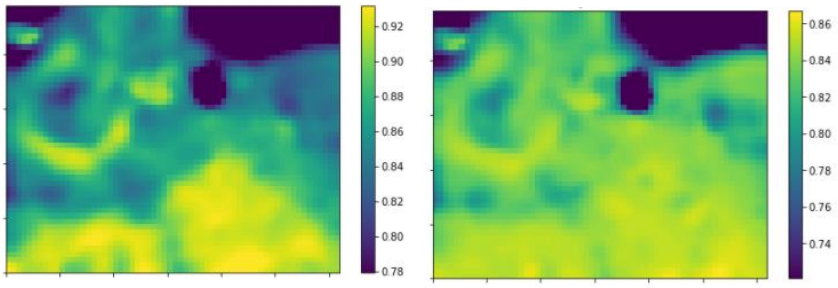
Figura 6- 8: Matrices de confusión modelos de clasificación. *Logistic Regression* (a), *K-Neighbors* (b), *Decision Tree* (c), *Random Forest* (d) y *MLPC* (e). Tractocamión #5.

Tabla 6- 4. Métricas de desempeño por clasificador tractocamión #5.

CLASIFICADOR	PRECISIÓN	EXACTITUD	ESPECIFICIDAD	SENSIBILIDAD
Logistic Regression	58.67%	53.78%	26.10%	73.05%
K-Neighbors	58.91%	54.47%	24.68%	75.21%
Decision Tree	58.93%	54.52%	24.55%	75.39%
Random Forest	58.96%	54.55%	24.74%	75.30%
MLPC	58.93%	54.55%	24.35%	75.57%

En el tractocamión #5 los resultados en la precisión para cada uno de los modelos propuestos son cercanos al 60%, se presenta una mejora en el rendimiento de esta métrica con respecto a los resultados obtenidos en el tractocamión #1, sin embargo los resultados se encuentran afectados aun por lo que se visualiza en la Figura 6-6 donde los datos antes de ingresar y luego de salir de taller se mezclan notablemente en los dos grupos resultado de la técnica UMAP-KMEANS y que es validado por el resultado de las métricas de los clasificadores. Para el tractocamión #5 la precisión obtenida con la aplicación de técnicas no supervisadas no funciona y también como hipótesis se pueden presentar otros tipos de daños que no permitan clasificar correctamente los asociados a sistemas térmicos.

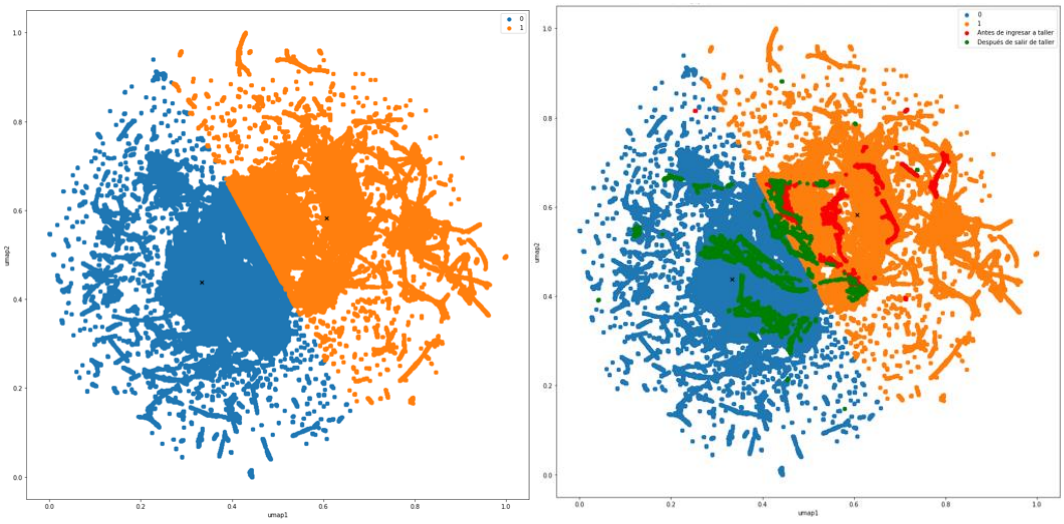
En la Figura 6-9 se implementa el algoritmo SOM para este tractocamión, donde también se observan patrones similares en el agrupamiento para las variables temperatura del aceite del motor (a) y temperatura del refrigerante del motor (b). Como en el caso del tractocamión #1, esta técnica puede explorarse en la determinación de fallas al no tener resultados satisfactorios con la técnica UMAP-KMEANS.



(a) (b)

Figura 6- 9: SOM. Temperatura del aceite del motor(a) y Temperatura del refrigerante del motor(b). Tractocamión #5.

Finalmente, la tercera experimentación corresponde a los datos obtenidos con el tractocamión #10 (menor cantidad de ingresos a taller) de la muestra de estudio. En la Figura 6-10 se presentan los resultados de la técnica indicado en la figura (a) y la superposición de los datos para los estados de intervención en taller indicado en la figura (b).



(a) (b)

Figura 6- 10: UMAP-KMEANS con dos grupos. Agrupamiento UMAP-KMEANS (a), datos de taller en resultado UMAP-KMEANS(b). Tractocamión #10.

En este caso particular, al experimentar la técnica en un tractocamión con poca cantidad de ingresos a taller, se hace posible identificar como los estados de las intervenciones se separan en los grupos que entrega UMAP-KMEANS con dos clases, bajo la hipótesis que dentro de la cantidad de intervenciones de taller presentadas no se enmascaran otros daños en el vehículo diferentes a sistemas térmicos.

Los datos antes de ingresar a taller (color rojo) se posicionan en el grupo con etiqueta 1 (color naranja) y mayoritariamente los datos después de salir de taller (color verde) se posicionan en el grupo con etiqueta 0 (color azul) como se visualiza en la figura (a).

Se realiza la validación de los resultados con los clasificadores, con respecto a los datos antes de ingresar y luego de salir de taller. En la Figura 6-11 se presentan las curvas ROC

para los modelos de clasificación. Donde (a) es la curva ROC para *Logistic Regression*, (b) para *K-Neighbors*, (c) para *Decision Tree*, (d) para *Random Forest* y (e) para *MLPC*.

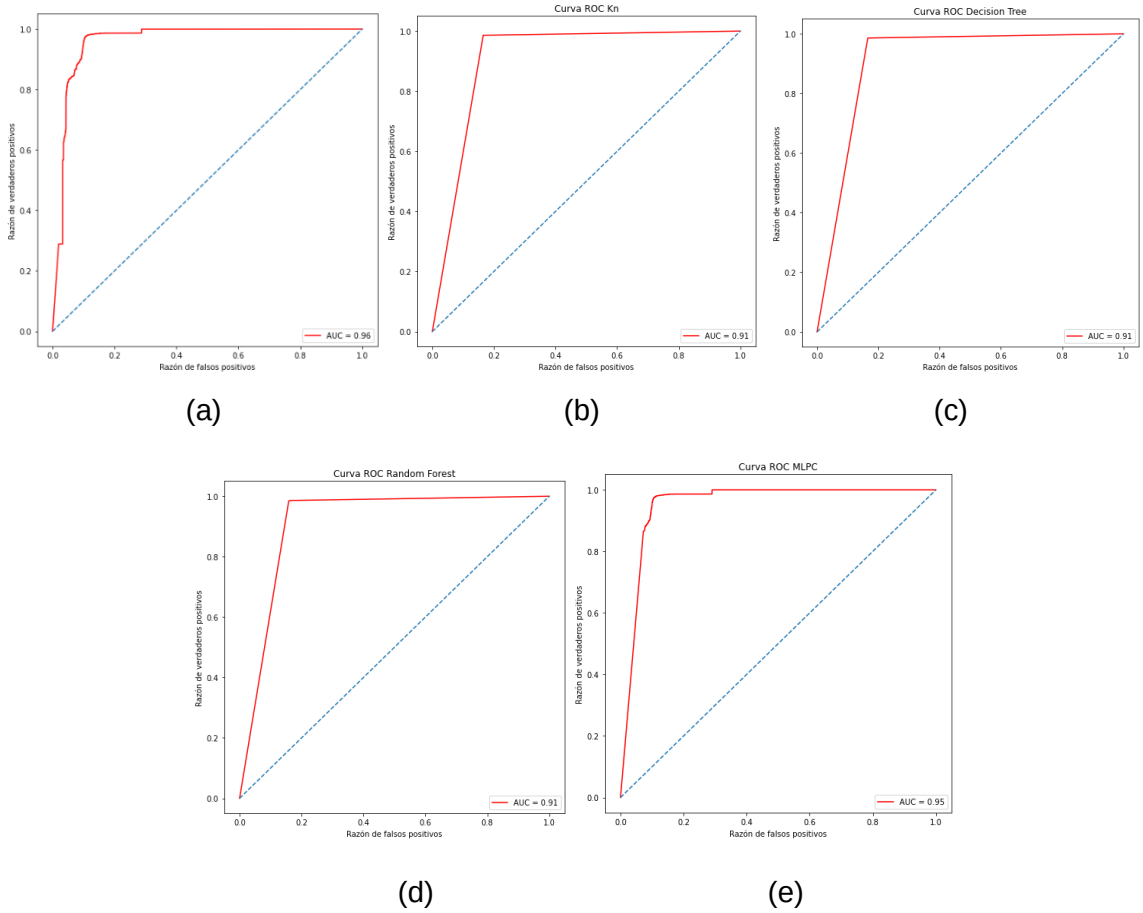


Figura 6- 11: Curvas ROC modelos de clasificación. *Logistic Regression* (a), *K-Neighbors* (b), *Decision Tree* (c), *Random Forest* (d) y *MLPC* (e). Tractocamión #10.

A continuación, en la Tabla 6-5 se detallan los valores del área bajo la curva ROC para cada clasificador. Las áreas bajo las curvas ROC con los clasificadores implementados son superiores al valor de 0.91 para cada uno, lo que indica que los datos antes de ingresar y luego de salir de taller para el tractocamión #10 se clasifican con un buen rendimiento en los modelos propuestos en los dos grupos resultado de la técnica UMAP-KMEANS.

Tabla 6- 5. Área bajo la curva ROC por clasificador.

Clasificador	Área bajo la curva ROC
Logistic Regression	0.96
K-Neighbors	0.91
Decision Tree	0.91
Random Forest	0.91
MLPC	0.95

Se presentan las matrices de confusión de cada uno de los clasificadores en la Figura 6-12 donde (a) es la matriz de confusión para *Logistic Regression*, (b) para *K-Neighbors*, (c) para *Decision Tree*, (d) para *Random Forest* y (e) para *MLPC* junto los resultados de las métricas en la Tabla 6-6.

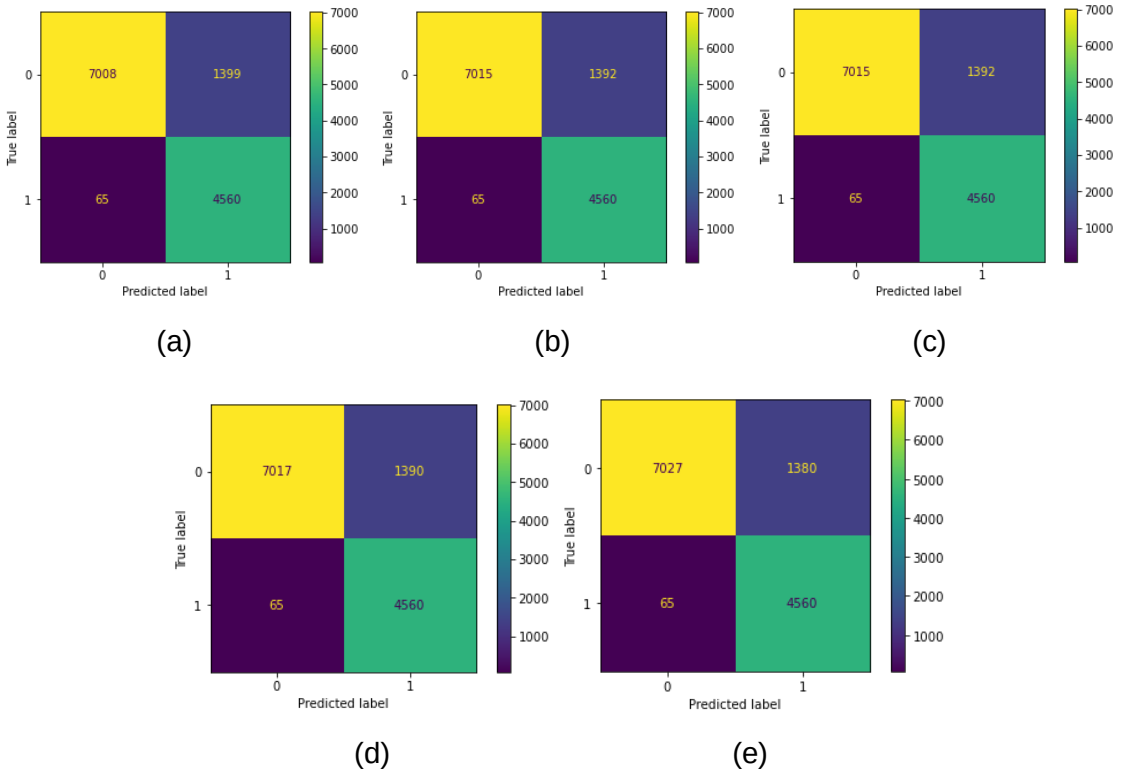


Figura 6- 12: Matrices de confusión modelos de clasificación. *Logistic Regression* (a), *K-Neighbors* (b), *Decision Tree* (c), *Random Forest* (d) y *MLPC* (e). Tractocamión #10.

Tabla 6- 6. Métricas de desempeño por clasificador. Tractocamión #10.

CLASIFICADOR	PRECISIÓN	EXACTITUD	ESPECIFICIDAD	SENSIBILIDAD
Logistic Regression	76.52%	88.77%	83.36%	98.59%
K-Neighbors	76.61%	88.82%	83.44%	98.59%
Decision Tree	76.61%	88.82%	83.44%	98.59%
Random Forest	76.64%	88.84%	83.47%	98.59%
MLPC	76.77%	88.91%	83.59%	98.59%

Se presenta una mejora en el resultado de la métrica de precisión con respecto a los tractocamiones #1 y #5, sin embargo también se puede notar que el rendimiento es afectado por los datos antes de ingresar y luego de salir de taller que se mezclan en los dos grupos de la técnica UMAP-KMEANS como se visualiza en la Figura 6-10, donde una parte de los datos luego de salir de taller del tractocamión se ubican en el grupo donde se encuentran los datos antes de ingresar a taller. Con esta experimentación se encuentra que bajo la hipótesis de menor cantidad de ingresos a taller para un tractocamión se pueden mejorar los resultados en la detección de fallas asociadas a sistemas térmicos donde no se presenten otros tipos de daños diferentes que enmascaren el análisis en cada caso de estudio.

En la Figura 6-13, con la implementación del algoritmo SOM, también se encuentran patrones similares en las variables temperatura del aceite del motor (a) y la temperatura del refrigerante del motor (b). La estrategia de selección de placas con baja cantidad de ingresos a taller presenta resultados prometedores puesto que la técnica UMAP-KMEANS y SOM presentan agrupamientos en los datos marcados para evaluar la identificación de patrones de falla.

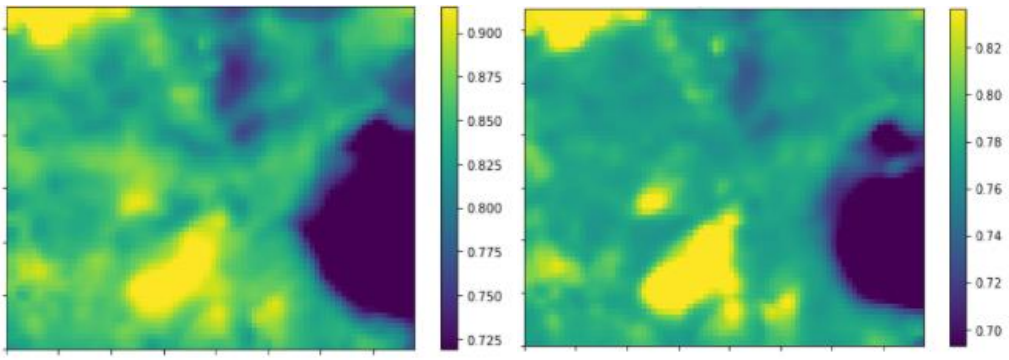


Figura 6- 13: SOM. Temperatura del aceite del motor(a) y Temperatura del refrigerante del motor(b). Tractocamión #10.

Con los resultados de esta técnica en este vehículo es posible realizar el entrenamiento de los clasificadores para detectar estados de falla según el comportamiento de los datos y teniendo en cuenta estos datos antes de ingresar a taller y luego de salir de taller para realizar el entrenamiento, validación y comparación de resultados con los modelos de clasificación propuestos, teniendo en cuenta que se puede encontrar un rendimiento afectado por los datos que se combinan en los grupos analizados.

6.2 Modelos de clasificación supervisados.

Como experimentación adicional en este capítulo, se evalúan los clasificadores propuestos con anterioridad al conjunto de datos del tractocamión #10, correspondientes a los registros antes y después de salir de taller en la fecha 20 de febrero de 2020.

La estrategia se basa en la selección de esta fecha particular, donde se toman 799 datos correspondientes a la información de un día antes de ingresar a taller el tractocamión, luego se seleccionan los 799 datos un día después de salir de taller y finalmente, se seleccionan aleatoriamente 799 datos que no corresponden a fechas donde el vehículo se encuentra en taller. Para un total de 2397 datos.

Con esta selección de los datos, se generan las etiquetas para tener un modelo supervisado, donde la etiqueta con valor igual a 0 corresponde a los datos después del tractocamión salir de taller (color verde), la etiqueta con valor de 1 corresponde a los datos antes de ingresar a taller (color rojo) y la etiqueta con valor de 2 corresponde a datos por fuera de los dos conjuntos anteriores, que llamaremos como “otros”. Lo anterior se visualiza en la Figura 6-14.

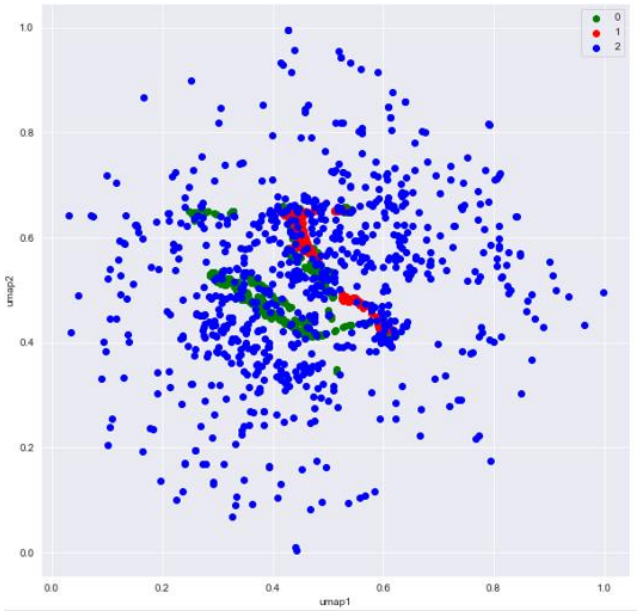
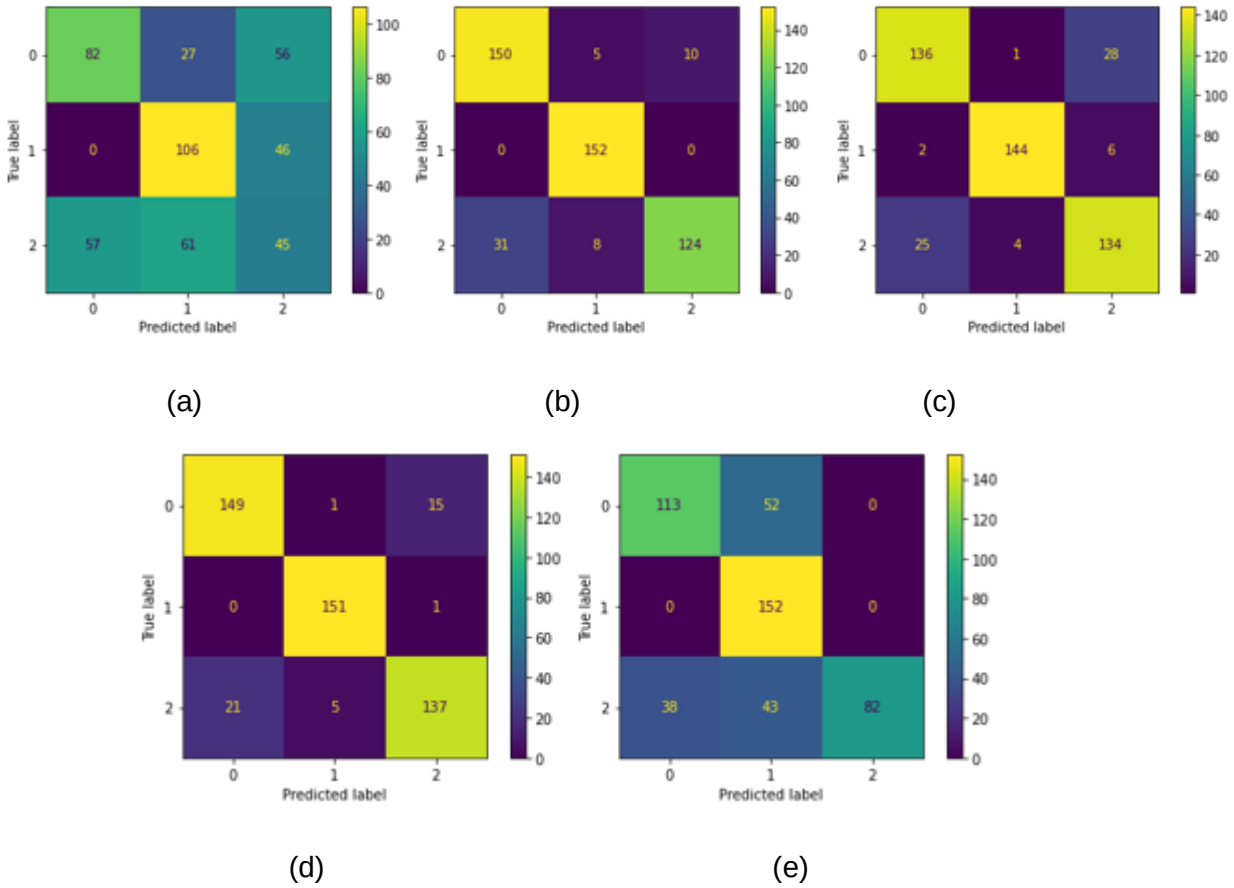


Figura 6- 14: Datos para modelos de clasificación supervisados. Tractocamión #10.

Los modelos de clasificación son entrenados con el 80% de los datos anteriormente descritos y se validan con el 20% de los datos restantes. En la Figura 6-15 se presentan las matrices de confusión de cada uno de los clasificadores, donde (a) es la matriz de confusión para *Logistic Regression*, (b) para *K-Neighbors*, (c) para *Decision Tree*, (d) para *Random Forest* y (e) para *MLPC* junto al resultado de la métrica de precisión en cada uno de los modelos evaluados como se describe en la Tabla 6-7.



(a) (b) (c)

(d) (e)

Figura 6- 15: Matrices de confusión modelos de clasificación supervisados. *Logistic Regression* (a), *K-Neighbors* (b), *Decision Tree* (c), *Random Forest* (d) y *MLPC* (e). Tractocamión #10.

Tabla 6- 7. Métrica de precisión por clasificador supervisado. Tractocamión #10.

CLASIFICADOR	PRECISION
Regression Logistic	48.54%
K-Neighbors	88.75%
Decision Tree	86.25%
Random Forest	**91.04%**
MLPC	72.29%

Los resultados son prometedores para los modelos de clasificación K-Neighbors y Random Forest, siendo este último el de mejor desempeño con la métrica de precisión (91%), se recomienda implementar este clasificador en la detección de estados de falla para sistemas

térmicos en tractocamiones graneleros cuando se implementen modelos supervisados y se cuenten con las etiquetas como se presentan en esta experimentación.

6.3 Conclusiones

- En este capítulo se realizó la experimentación con los diez tractocamiones que tuvieron más ingresos a taller en el año 2020 para intervenciones en sistemas térmicos, donde se detectó que a mayor cantidad de ingresos a taller el algoritmo de reducción de dimensionalidad UMAP en conjunto con el algoritmo de agrupamiento KMEANS, si bien es capaz de separar los grupos no es posible detectar los patrones de falla en los grupos al comparar con los datos que se asociaron a las entradas y salidas de taller ya que los datos se mezclan en los dos grupos como se evidenció en la experimentación del tractocamión #1.

- Como hipótesis se tiene que a mayor cantidad de ingresos a taller se pueden tener otros tipos de fallos que no se asocian a sistemas térmicos y enmascaran el patrón de fallo que queremos identificar, en este caso para sistemas térmicos. Con la experimentación con el tractocamión #10 de la muestra y siendo el que menor cantidad de ingresos a taller presentó, se encontró que los datos antes de ingresar a taller y luego de salir de taller se posicionan cada uno en los dos grupos generados por la técnica UMAP-KMEANS con una cantidad de los datos después de salir de taller que se posicionan en el grupo de los datos antes de ingresar a taller, lo que puede afectar el desempeño en la posterior implementación de un modelo de clasificación para la predicción de estados de falla en sistemas térmicos.

- En cada una de las experimentaciones realizadas se encuentran resultados prometedores con el algoritmo SOM, donde se hace posible evidenciar agrupamientos cuando se analizan las variables e identificar cuáles son los grupos donde se pueden posicionar los datos asociados a comportamientos anómalos de operación para la identificación de los umbrales en los que el sistema térmico puede empezar a presentar estado de falla.

- Debido a que para la detección de patrones de falla con técnicas no supervisadas no es sencillo aplicar las técnicas en cualquier placa sin realizar la estrategia de selección del tractocamión con pocos ingresos a taller que no enmascaren otro tipo de daños, se experimentan modelos de clasificación supervisados donde se realiza una selección de la información solo con datos antes y luego de salir de taller, se etiquetan estos estados para entrenar los modelos de clasificación donde el clasificador Random Forest presentó una precisión del 91% en la validación del modelo. Bajo esta técnica supervisada se puede proponer implementar este modelo para determinar si con los datos que presente un tractocamión en su comportamiento puede presentar estado de falla en sistemas térmicos.

7. CONCLUSIONES Y RECOMENDACIONES

Este trabajo evalúa la estrategia de técnicas no supervisadas para determinar patrones asociados a fallas de sistemas térmicos en tractocamiones graneleros con el objetivo de tener un insumo en la programación de mantenimiento predictivo de la flota que pueda garantizar disponibilidad y ampliar la vida útil de los equipos en modelos de pago por uso.

A continuación, se enuncian los resultados encontrados y el trabajo futuro que se propone.

- Con respecto a las preguntas de investigación de la tesis, es posible detectar y diagnosticar tipos de falla usando métodos de agrupamiento desde que se implemente la selección de tractocamiones con base en pocos ingresos a taller, donde no se enmascaren tipos de falla distintos a los sistemas térmicos. Para los tractocamiones con mayor cantidad de ingresos a taller las estrategias de agrupamiento no presentan resultados satisfactorios.

- Los métodos propuestos que permitieron detectar patrones de falla fueron las técnicas UMAP-KMEANS para conjuntos de datos con un promedio de 1 millón de registros, presentando desagregación de los datos en la reducción de dimensionalidad y agrupamientos separando los datos para la detección de patrones. El algoritmo SOM desde el análisis de agrupamiento por variables y con un alcance de hasta 10 millones de datos con la información de hasta 10 tractocamiones a la vez, permite visualizar los agrupamientos y analizar condiciones anormales en el comportamiento de las variables.

- Bajo la hipótesis de la tesis sobre la detección de grupos relacionados con estado normal y de falla para sistemas térmicos de tractocamiones graneleros, el algoritmo no supervisado DBSCAN presenta deficiencias al trabajar con conjuntos de datos de 1 millón de registros por la alta distribución espacial de la información donde no es posible encontrar agrupamientos separados entre los datos. Sin embargo, al trabajar con casos particulares como lo fue con solo los datos antes de ingresar y luego de salir de taller para un tractocamión presenta oportunidad en la agrupación

de estos estados para implementar posteriormente modelos de clasificación a partir de esta técnica.

- La técnica SOM permite detectar grupos anómalos en los datos por variable, donde a través del análisis de los umbrales donde en la cuadrilla de visualización se presentan cambios en el agrupamiento se puede proponer el análisis donde un tractocamión pueda estar próximo a fallar en los sistemas térmicos.

- El algoritmo KMEANS puede detectar estados de falla desde que se implemente la estrategia de selección de tractocamiones con pocos ingresos a taller y con un proceso previo de reducción de dimensionalidad con la técnica UMAP.

- Para la técnica SPECTRAL CLUSTERING los agrupamientos obtenidos no permitieron identificar patrones marcados en los análisis antes de ingresar y luego de salir de taller el tractocamión. Aunque de manera global no se encontraron estos patrones, la técnica permite realizar agrupamiento en cambios temporales entre señales.

- En el desarrollo de los objetivos de la tesis, se establecen los criterios para la preparación de los datos obtenidos desde los dispositivos de telemetría, donde es fundamental la sincronización temporal de los datos al conformar los conjuntos de información con el método de interpolación. Posteriormente a esto normalizar los datos teniendo en cuenta que en el análisis de distribuciones estadísticas por variable no se ajustan a distribuciones normales. Con la anterior preparación de los datos se hace posible la correcta implementación posterior de técnicas no supervisadas.

- En la implementación de las técnicas de reducción de dimensionalidad es necesario experimentar en la variación de hiperparametros donde los algoritmos puedan desagregar los datos en los espacios de baja dimensión, esto es fundamental para luego obtener mejores resultados con los algoritmos de agrupamiento.

- En la estrategia de selección de los métodos de agrupamiento no supervisados en cuanto a rendimiento y tiempo de desempeño, se realizan experimentaciones desde una cantidad pequeña de datos correspondientes a dos meses de información en el tractocamión con que se realiza la prueba de escritorio y ejecutando dichas experimentaciones en la maquina local, hasta encontrar que con los servicios de computación en la nube se podía trabajar con datos de un tractocamión con un millón de registros en promedio para las variables de la tesis.

- A partir de un millón de datos en información para los tractocamiones y con la metodología propuesta, los algoritmos KMEANS y DBSCAN no se podían procesar con la infraestructura contratada en la nube. Como trabajo futuro se puede investigar sobre la implementación de los algoritmos para Big data en conjunto con la computación distribuida de la nube.

- El algoritmo SOM presenta oportunidad de realizar el agrupamiento con 10 millones de registros para las 18 variables analizadas desde la maquina local con un tiempo de procesamiento de 25 horas, aprovechando los recursos computacionales locales con que se aborda esta tesis.

- La ejecución del algoritmo SPECTRAL CLUSTERING se pudo implementar en la maquina local con solo unos pocos segundos en el análisis de información por cada 500 datos temporales analizados.

- En la evaluación de los métodos propuestos en la identificación de patrones de falla se realiza la validación con la información referente a los estados de antes y luego de salir de taller para la comparación de que los datos se posicionen de manera separada en los grupos generados por las técnicas candidatas, donde con la técnica UMAP-KMEANS y los vehículos con pocos ingresos a taller se encontraron resultados prometedores para posteriores ejercicios con modelos de clasificación en la detección de fallas.

- Se propone abordar el fenómeno de la determinación de fallos en sistemas térmicos con modelos de clasificación supervisados, donde con la información de los estados de taller se puedan asignar etiquetas a los datos y poder entrenar los modelos. Esto con base en que los datos de la flota estudiada en conjunto no permiten detectar los patrones al contar con grandes conjuntos de información posiblemente afectados también por otros tipos de fallas.

Como trabajo futuro en la ampliación de esta tesis, se puede partir de la base de estudio sobre cuáles son los tipos de variables que se deben analizar para la determinación de patrones asociados a fallas de sistemas térmicos en tractocamiones graneleros, esto con el fin de apalancar de entrada los resultados que se puedan obtener para la reducción de dimensionalidad en la desagregación de los datos y posterior implementación de técnicas de agrupamiento para el fenómeno de estudio y con una correcta selección de hiperparametros en cada técnica como también evaluar las variaciones de los algoritmos que permitan trabajar con *Big data*, optimizando la utilización de los servicios en la nube y poder trabajar con mayor cantidad de datos para experimentación.

BIBLIOGRAFÍA

Allaoui, M., Kherfi, M. L., & Cheriet, A. (2020). Considerably improving clustering algorithms using UMAP dimensionality reduction technique: A comparative study. En *Lecture Notes in Computer Science* (pp. 317–325). Springer International Publishing.

Amruthnath, N., & Gupta, T. (2019). Fault diagnosis using clustering. What statistical test to use for hypothesis testing? *Machine Learning and Applications An International Journal*, *6*(1), 17–33. https://doi.org/10.5121/mlaij.2019.6102

Bangui, H., Ge, M., & Buhnova, B. (2018). Exploring big data clustering algorithms for internet of things applications. *Proceedings of the 3rd International Conference on Internet of Things, Big Data and Security*.

Bendechache, M., Kechadi, M.-T., & Le-Khac, N.-A. (2016). Efficient large scale clustering based on data partitioning. *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*.

Braga, P. H. M., & Bassani, H. F. (2018). A semi-supervised self-organizing map for clustering and classification. *2018 International Joint Conference on Neural Networks (IJCNN)*.

Butikofer Lagos, G. (2017). Optimización del mantenimiento preventivo de flotas en base a técnicas de clustering y aprendizaje supervisado.

Car and Driver. (2019, junio 15). *¿Cómo influye la temperatura del aceite en el motor?*. Recuperado de https://www.caranddriver.com/es/coches/planeta-motor/a60565/temperatura-aceite-como-influye-en-el-motor/

Castellanos, G. C., & Rodríguez, J. E. R. (2011). Agrupamiento de datos de series de tiempo. Estado del arte. Revista vínculos, 8(1), 210-231.

Chaudhuri, A. (2018). Predictive maintenance for industrial IoT of vehicle fleets using hierarchical modified fuzzy support vector machine. En *arXiv [cs.AI]*. http://arxiv.org/abs/1806.09612

Chekkala, V. L. (2020). Predictive Maintenance for Fault Diagnosis and Failure Prognosis in Hydraulic System (Doctoral dissertation, Dublin, National College of Ireland).

Çınar, Z. M., Abdussalam Nuhu, A., Zeeshan, Q., Korhan, O., Asmael, M., & Safaei, B. (2020). Machine learning in predictive maintenance towards sustainable smart manufacturing in Industry 4.0. *Sustainability*, *12*(19), 8211. https://doi.org/10.3390/su12198211

Delua, J. (12 de marzo de 2021). *Supervised vs. Unsupervised Learning: What´s the Difference?*. IBM. https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning

Espadoto, M., Hirata, N., & Telea, A. (2021). Self-supervised dimensionality reduction with neural networks and pseudo-labeling. *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*.

Forest, F., Lebbah, M., Azzag, H., & Lacaille, J. (2019). Deep architectures for joint clustering and visualization with self-organizing maps. En *Lecture Notes in Computer Science* (pp. 105–116). Springer International Publishing.

Guo, X., Liu, X., Zhu, E., & Yin, J. (2017). Deep Clustering with Convolutional Autoencoders. En *Neural Information Processing* (pp. 373–382). Springer International Publishing.

Heidari, S., Alborzi, M., Radfar, R., Afsharkazemi, M. A., & Rajabzadeh Ghatari, A. (2019). Big data clustering with varied density based on MapReduce. *Journal of Big Data*, *6*(1). https://doi.org/10.1186/s40537-019-0236-x

Ibrahim, S. K., Ahmed, A., Zeidan, M. A. E., & Ziedan, I. E. (2020). Machine learning techniques for satellite fault diagnosis. *Ain Shams Engineering Journal*, *11*(1), 45–56. https://doi.org/10.1016/j.asej.2019.08.006

Kucukyilmaz, T. (2014). Parallel K-means algorithm for shared memory multiprocessors. *Journal of computer and communications*, *02*(11), 15–23. https://doi.org/10.4236/jcc.2014.211002

Lacaille, J., & Come, E. (2011). Visual mining and statistics for a turbofan engine fleet. *2011 Aerospace Conference*.

Langone, R., Alzate, C., De Ketelaere, B., & Suykens, J. A. K. (2013). Kernel spectral clustering for predicting maintenance of industrial machines. *2013 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*.

Løkse, S., Bianchi, F. M., Salberg, A.-B., & Jenssen, R. (2017). Spectral clustering using PCKID - A probabilistic cluster kernel for incomplete data. En *arXiv [stat.ML]*. http://arxiv.org/abs/1702.07190

McConville, R., Santos-Rodriguez, R., Piechocki, R. J., & Craddock, I. (2021). N2D: (not too) deep clustering via clustering the local manifold of an autoencoded embedding. *2020 25th International Conference on Pattern Recognition (ICPR)*.

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. En *arXiv [stat.ML]*. http://arxiv.org/abs/1802.03426

Murakami, T. (2002). Development of Vehicle Health Monitoring System (VHMS / WebCARE) for Large-Sized Construction Machine. Construction, 48(150), 15–21.

Nowaczyk, S. S., Rognvaldsson, T., Byttner, S., Prytz, R., Nowaczyk, S. S., Rögnvaldsson, T., Thorsteinn, R., Byttner, S., Rognvaldsson, T., Byttner, S., Prytz, R., Nowaczyk, S. S., & Rögnvaldsson, T. (2013). Analysis of Truck Compressor Failures Based on Logged Vehicle Data. 9th International Conference on Data Mining, Las Vegas, Nevada, USA, July. https://www.researchgate.net/publication/256486984.

Pacella, M., & Papadia, G. (2020). Fault diagnosis by multisensor data: A data-driven approach based on spectral clustering and pairwise constraints. *Sensors (Basel, Switzerland)*, *20*(24), 7065. https://doi.org/10.3390/s20247065

Palmqvist, A. (2016). Exploratory data analysis of Volvo trucks repair history towards modelling a trucks lifetime maintenance needs (Dissertation). Retrieved from http://urn.kb.se/resolve?urn=urn:nbn:se:hh:diva-32244

PeerXP Team. (2017, October 17). *The 6 stages of data processing cycle - PeerXP team*. Medium. https://medium.com/@peerxp/the-6-stages-of-data-processing-cycle-3c2927c466ff

Perr-sauer, J., Duran, A., Phillips, C., Perr-sauer, J., Duran, A., & Phillips, C. (2020). Clustering Analysis of Commercial Vehicles Using Automatically Extracted Features from Time Series

Prytz, R. (2014). Machine learning methods for vehicle predictive maintenance using off-board and on-board data. In Thesis (Vol. 9, Issue 9). www.hh.se/hupData Clustering Analysis of Commercial Vehicles Using Automatically Extracted Features from Time Series Data. January.

Ranasinghe, K., Kapoor, R., Gardi, A., Sabatini, R., Wickramanayake, V., & Ludovici, D. (2020). Vehicular sensor network and data analytics for a health and usage management system. *Sensors (Basel, Switzerland)*, *20*(20), 5892. https://doi.org/10.3390/s20205892

Saeed, M. M., Al Aghbari, Z., & Alsharidah, M. (2020). Big data clustering techniques based on Spark: a literature review. *PeerJ. Computer Science*, *6*(e321), e321. https://doi.org/10.7717/peerj-cs.321

Sarma, A., Goyal, P., Kumari, S., Wani, A., Challa, J. S., Islam, S., & Goyal, N. (2019). MDBSCAN: An exact scalable DBSCAN algorithm for big data exploiting spatial locality. *2019 IEEE International Conference on Cluster Computing (CLUSTER)*.

Sefidian, Amir. (18 de diciembre de 2020). *How to determine Epsilon and MinPts parameters of DBSCAN clustering*. http://www.sefidian.com/2020/12/18/how-to-determine-epsilon-and-minpts-parameters-of-dbscan-clustering/

Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

Sreedhar, C., Kasiviswanath, N., & Chenna Reddy, P. (2017). Clustering large datasets using K-means modified inter and intra clustering (KM-I2C) in Hadoop. *Journal of Big Data*, *4*(1). https://doi.org/10.1186/s40537-017-0087-2

tSNE vs.UMAP: Estructura global. (2020, March 5). ICHI.PRO. https://ichi.pro/es/tsne-vs-umap-estructura-global-85213320100375

Ullah, S., & Kim, D.-H. (2020). Lightweight driver behavior identification model with sparse learning on in-vehicle CAN-BUS sensor data. *Sensors (Basel, Switzerland)*, *20*(18), 5030. https://doi.org/10.3390/s20185030

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of machine learning research, 9(11).

Vore, S. (2016). Methods to analyze large automotive fleet-tracking datasets with application to light-and medium-duty plug-in hybrid electric vehicle work trucks (Doctoral dissertation, Colorado State University).

Wattenberg, et al., "How to Use t-SNE Effectively", Distill, 2016. http://doi.org/10.23915/distill.00002

Yang, B., Fu, X., Sidiropoulos, N. D., & Hong, M. (2017). Towards K-means-friendly spaces: Simultaneous deep learning and clustering. 34th International Conference on Machine Learning, ICML 2017, 8, 5888–5901.

Yen, K. S., Ravani, B., & Lasky, T. A. (2015). DOE Fleet In-Vehicle Data Acquistion System (FIDAS) Technical Support and Testing (No. CA16-2516).