UNIVERSIDADE D
COIMBRA

Roney Camargo Malaguti

DETERMINATION OF THE OPERATION CONDITION OF LUBRICATING OILS IN DIESEL ENGINES

PhD thesis in Mechanical Engineering, Advanced Production Systems, supervised by Professor Cristóvão Silva and Professor Nuno Lourenço, and presented to the Department of Mechanical Engineering of the Faculty of Sciences and Technology of the University of Coimbra.

December de 2022

Determination of the Operation Condition of Lubricating Oils in Diesel Engines

Roney Camargo Malaguti

PhD thesis in Mechanical Engineering, Advanced Production Systems, supervised by Professor Cristóvão Silva and Professor Nuno Lourenço, and presented to the Department of Mechanical Engineering of the Faculty of Sciences and Technology of the University of Coimbra.

December, 2022

UNIVERSIDADE D
COIMBRA

This work is partially funded by national funds through the FCT - Foundation for Science and Technology, I.P., within the scope of the project CISUC - UID/CEC/00326/2020 and by European Social Fund, through the Regional Operational Program Centro 2020 and by national funds through FCT - Fundação para a Ciência e a Tecnologia, under the project UIDB/00285/2020.

Life is a game that must be played in hard mode, because nobody evolves playing in easy mode.

Roney Malaguti

Abstract

Diagnosis and fault prediction processes based on real data are important tools in the Condition Based Maintenance (CBM) approach for diesel vehicle fleets. Companies are equipping their vehicle fleets with a large number of sensors, which allows the collection of large amounts of data about the current condition of each asset. This allows companies to invest efforts in the development of methods to accurately identify the state of wear and tear of a piece of equipment or system, making CBM more effective and reliable.

For this type of equipment condition analysis there are several techniques currently available, depending on the type of data to be collected, cost, and the need or not to use external diagnostic interfaces to analyze the condition of the vehicle. Intelligent lubricant oil analysis is one of the possible techniques for determining the condition of equipment and is therefore an important tool for fleet managers to determine the condition of heavy diesel vehicles. Because it is a relatively low-cost technique that allows the manager a clear view of the equipment's condition from real-time data collection, in this work we will present an intelligent analysis of lubricant data from 5 different vehicles, evaluating whether the variables collected make it possible to determine the condition of the lubricants and the vehicles they are used on. To this end, we initially present a study on the issues that guide the development of this work, through a literature review that presents not only pertinent information about lubricants, but also the necessary information about real-time data capture interfaces and systems based on machine learning. After analyzing the literature, presenting the hardware for real-time data collection, we confirm the importance of the study of lubricants through the analysis of common faults of heavy diesel vehicles.

From the knowledge acquired by reading and analyzing the literature, we begin the analysis and organization of the collected data. We start by analyzing the missing values of the variables, the need for data balancing, Principal Component Analysis (PCA) and correlation analysis on top of the collected data. This step, is called exploratory data analysis, is important for the development of the automatic condition determination system, as well as for evaluating the need for the inclusion of new variables designed from raw data for a better determination of this operating condition.

Using a solid and organized database, we developed the first supervised machine learning model based on the Random Forest (RF) classifier to determine the operating condition of lubricating oil in diesel engines. The presented results show that the selected variables have the potential to determine the running state, and that they are strongly related to the condition of the lubricant. One of the variables designed is kinematic viscosity, which is shown to play a relevant role in characterizing the condition of the vehicle. This model presented Recall results of 95.90%, an Accuracy of 97.50%, and an F1 score of 96.70%.

Finally, despite having results worthy of recognition for the initial model created from the RF algorithm, we still performed an in depth study using a diverse range of algorithms and different sets of

features as input data. This analysis was conducted using cross-validation, and considered the following algorithms: Logistic Regression, Perceptron, Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier. With this validation it is possible to demonstrate that the proposed approach is able to successfully identify the operating conditions of lubricating oil, with the predictive model Gradient Boosting Classifier with 7 variables, obtaining a Recall result of 97.90%, an Accuracy of 98.80% and an F1 score of 97.90%, even better than the model initially chosen. The cross validation also allows to identify the best combination of variables for the model.

Keywords: lubricating oil data, diesel-powered vehicle fleets, Condition Based Maintenance (CBM), machine learning algorithms

Resumo

Os processos de diagnóstico e previsão de falhas baseados em dados reais são ferramentas importantes na abordagem de Manutenção Baseada na Condição (CBM) para frotas de veículos a diesel. As empresas estão a equipar as suas frotas de veículos com um grande número de sensores, o que permite a recolha de grandes quantidades de dados sobre o estado actual de cada activo. Isto permite às empresas investir esforços no desenvolvimento de métodos para identificar com precisão o estado de desgaste de um equipamento ou sistema, tornando a CBM mais eficaz e fiável.

Para este tipo de análise do estado do equipamento existem várias técnicas actualmente disponíveis, dependendo do tipo de dados a recolher, do custo e da necessidade ou não de utilizar interfaces externas de diagnóstico para analisar o estado do veículo. A análise inteligente do óleo lubrificante é uma das técnicas possíveis para determinar o estado do equipamento e é, portanto, uma ferramenta importante para os gestores de frotas para determinar o estado dos veículos pesados a diesel. Como se trata de uma técnica de custo relativamente baixo que permite ao gestor uma visão clara do estado do equipamento a partir da recolha de dados em tempo real, neste trabalho apresentaremos uma análise inteligente dos dados de lubrificantes de 5 veículos diferentes, avaliando se as variáveis recolhidas permitem determinar o estado dos lubrificantes e dos veículos em que são utilizados. Para tal, apresentamos inicialmente um estudo sobre as questões que orientam o desenvolvimento deste trabalho, através de uma revisão bibliográfica que apresenta não só a informação pertinente sobre lubrificantes, mas também a informação necessária sobre interfaces e sistemas de recolha de dados em tempo real baseados na aprendizagem de máquinas. Após análise da literatura, apresentamos o hardware para recolha de dados em tempo real e confirmamos a importância do estudo de lubrificantes através da análise de falhas comuns de veículos pesados a diesel.

A partir dos conhecimentos adquiridos através da leitura e análise da literatura, iniciamos a análise e organização dos dados recolhidos. Começamos por analisar os valores em falta das variáveis, a necessidade de equilíbrio dos dados, a Análise de Componentes Principais (PCA) e a análise de correlação, para além dos dados recolhidos. Esta etapa, denominada análise exploratória de dados, é importante para o desenvolvimento do sistema de determinação automática da condição, bem como para avaliar a necessidade de inclusão de novas variáveis concebidas a partir de dados em bruto para uma melhor determinação desta condição operacional.

Utilizando uma base de dados sólida e organizada, desenvolvemos o primeiro modelo de aprendizagem supervisionada de máquinas baseado no classificador Random Forest (RF) para determinar o estado de funcionamento do óleo lubrificante em motores diesel. Os resultados apresentados mostram que as variáveis seleccionadas têm o potencial de determinar o estado de funcionamento e que estão fortemente relacionadas com o estado do lubrificante. Uma das variáveis concebidas é a viscosidade cinemática, que se mostra ter um papel relevante na caracterização do estado

do veículo. Este modelo apresentou resultados de Recall de 95.90%, uma Precisão de 97.50%, e uma pontuação de F1 de 96.70%.

Finalmente, apesar de termos resultados dignos de reconhecimento para o modelo inicial criado a partir do algoritmo RF, ainda realizámos um estudo aprofundado utilizando uma gama diversificada de algoritmos e diferentes conjuntos de variáveis como dados de entrada. Esta análise foi conduzida utilizando a validação cruzada, e considerou os seguintes algoritmos: Regressão Logística, Perceptron, Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier. Com esta validação é possível demonstrar que a abordagem proposta é capaz de identificar com sucesso as condições de funcionamento do óleo lubrificante, com o modelo preditivo Gradient Boosting Classifier com 7 variáveis, obtendo um resultado Recall de 97.90%, uma Precisão de 98.80% e uma pontuação F1 de 97.90%, ainda melhor do que o modelo inicialmente escolhido. A validação cruzada permite também identificar a melhor combinação de variáveis para o modelo.

Palavras Chave: Dados de Óleo Lubrificante, Frotas de Veículos a Diesel, Manutenção Baseada na Condição (CBM), Algoritmos de Aprendizagem de Máquinas

Acknowledgements/Agradecimentos[1]

Diante de todo o esforço e comprometimento com este trabalho chega a altura de agradecer a todos os que contribuíram para este momento que finaliza com a elaboração e apresentação deste documento. Não foram poucos os problemas, nem tão pouco dúvidas e desafios que foram sendo ultrapassados. Por isso, a ajuda de todos foi de fundamental importancia para que este sonho se tornasse realidade. Começo por agradecer aos meus orientadores: professor Nuno Lourenço e Cristovão Silva por todo o empenho e auxilio nesta caminhada, sem o apoio e conhecimento dos senhores este caminho seria muito mais complexo. Agradeço ainda o incentivo do professor Penousal Machado de iniciar um percurso na investigação, convite esse que contribuiu bastante para o meu desenvolvimento profissional. Agradeço também a todas as pessoas da Stratio Automotive que contribuiram direta e indiretamente, em especial ao Ricardo Margalho e ao Rui Sales por acreditarem e aceitarem contribuir com todo apoio não apenas financeiro mas de tempo e conhecimento, para além de toda a liberdade que me concederam na definição do que viria a ser a Tese. Por fim, agradeço-lhes todo o apoio nas múltiplas questões e linhas de investigação que foram surgindo. Obrigado aos meus avós Roberto Malaguti e Raymunda Marinho Malaguti, aos meus pais Roberto Marinho Malaguti e Maria do Socorro Silva Camargo pelo apoio incondicional, em especial um muito obrigado para a minha esposa Caroline Mota e ao meu filho Bernardo Mota Malaguti por nunca deixarem de acreditar nessa jornada e aceitarem tantas horas de ausencia e trabalhos nos finais de semana. Sem vocês dois, muito do que está neste documento não teria sido possível. Obrigado por me teres apoiado nos bons e maus momentos, por celebrarem as minhas vitórias e por não me deixares desistir diante das derrotas e frustrações. Não podendo esquecer um agradecimento ao meus amigos que aqui não consigo nomear todos por falta de espaço, por todos os momentos de descontração e minis. A todos vós, um agradecimento por estarem presentes nos momentos que mais precisei de força para não desistir de tudo. Por fim, agradeço à Fundação para a Ciência e Tecnologia (FCT), pelo financiamento no âmbito do projecto CISUC - UID/CEC/00326/2020 e pelo Fundo Social Europeu, através do Programa Operacional Regional Centro 2020 e por fundos nacionais no âmbito do projecto UIDB/00285/2020. Um obrigado a todos os outros que certamente me estou a esquecer mas que direta ou indiretamente contribuíram para o sucesso desta viagem.

Roney Malaguti

Coimbra, 9 de Dezembro de 2020

[1] For personal reasons, the acknowledges are written in Portuguese. Apologies to the non-Portuguese speakers.

Table of Contents

List of Figures

List of Tables

List of Acronyms

Condition Based Maintenance (CBM).
Principal component analysis (PCA)
Random Forest (RF)
Machine Learning (ML)
European Union (EU)
Artificial Intelligence (AI)
Random Forest Classifier (RFC)
Reliability Centred Maintenance (RCM)
Failure Mode and Effect Analysis (FMEA)
Predictive Maintenance (PdM)
Predictive Testing and Inspection (PT&I)
Potential failure (PF)
Top Dead Center (TDC)
Bottom Dead Center (BDC)
Revolutions Per Minute (RPM)
Lubricant Condition Monitoring (LCM)
Total Base Number (TBN)
Total Acid Number (TAN)
Oil Drain Interval (ODI)
Society of Automotive Engineers (SAE)
American Petroleum Institute (API)
Oil Stress Factor (OSF)
Fourier Transform Infrared Spectroscopy (FTIR)
International Organization for Standardization (ISO)
Zinc Dialkyldithiophosphate (ZDDP)
American Society for Testing and Materials (ASTM)
Equipment for Potentially Explosive Atmospheres (ATEX)
Field Scanning Electron Microscopy (FESEM)
Internet of Things (IoT)
On Board Diagnostic (OBD)
Parameter IDentification (PID)
Diagnostic Trouble Code (DTC)
Electronic Control Units (ECU)

Controller Area Network (CAN)
Malfunction Indicator Lamp (MIL)
K-Nearest Neighbors (K-NN)
Expert Systems (ES)
Human Experts (HE)
Knowledge Based Systems (KBS)
Knowledge Engineer (KE)
Intelligent Systems (IS)
Coding Expert (CE)
Before Big Data (BBD)
After Big Data (ABD)
Radio Frequency IDentification (RFID)
Data Mining (DM)
Knowledge Discovery in Databases (KDD)
International Telecommunication Union (ITU)
True Positive (TP)
False Negative (FN)
False Positive (FP)
True Negative (TN)
Receiver Operating Characteristic (ROC)
Area Under Curve (AUC)
Short Message Service (SMS)
Global Positioning System (GPS)
Exploratory Data Analysis (EDA)
Kolmogorov-Smirnov test (K-S test)
Principal Components (PC's)
Logistic Regression (LR)
Perceptron (PER)
Decision Tree Classifier (CART)
Gradient Boosting Classifier (GBM)

Chapter 1 - Introduction

Nowadays the transport sector, whether for passengers or goods, is of vital importance for society, customers safety, quality of service and level of emissions of pollutants. According to European Union (2021) the European Union (EU) transport policies are contributing to the dynamism of the sector's economy, through the development of a modern infrastructure network that makes travel faster and safer, while promoting digital and sustainable solutions. Thus, as society becomes increasingly mobile, EU policies must support the transport sector to address the major challenges it faces:

- Congestion which affects road and air traffic;
- Sustainability: transport still relies on oil for most of its energy and lubrication needs, which is environmentally and economically unsustainable;
- Air quality: the EU must reduce transport emissions by 60% below 1990 levels by 2050 and further reduce vehicle pollution;
- Infrastructure: the quality of infrastructure is not uniform throughout the EU;
- Competition: the European transport sector faces increasing competition from rapidly developing transport markets in other regions of the world.

According to data from BPstat (2021) the land transport sector in Portugal is composed by 20.898 companies, employing more than 108.200 people in 2020, representing a business volume of 7.168M€. Even though this financial volume was affected by the COVID-19 Pandemic, the data shows that this sector is extremely important for the Portuguese economy, and it is essential that Portugal aligns with the agenda presented by the EU for a sustainable and innovative transport sector. According to Nunes et al. (2019) it is necessary to increase transport efficiency in the use of resources and energy, contributing to the reduction of the environmental footprint of society, thus resulting in benefits to ecosystems and biodiversity, decarbonization of the economy and minimization of waste and pollution.

Following this line of development and following the agenda for decarbonization of the transport sector determined by the EU, it is necessary for companies to invest not only in the acquisition of new vehicles with renewable energy sources considered green, but also in technologies to maintain and help the current Portuguese fleet to improve its operating conditions and reduce the level of pollutant emissions. According to ACAP (2021), in the end of 2020, the heavy duty vehicle fleet in Portugal was composed of 149300 vehicles, with an average age for heavy passenger vehicles of 15.1 years and for heavy goods vehicles of 14.9 years. Taking into account these numbers regarding the vehicles age, a more reliable and sustainable maintenance is necessary to avoid failures in this old fleet and reduce the emission of pollutants and waste.

The identification and modelling of vehicle failure signatures to predict them before they occur has been subject of great interest for the engineers involved in the maintenance area, mainly due to the incomparable advantages, such as reduced downtime of the vehicle, lower maintenance cost and

increased security, when comparing with corrective maintenance. Vehicles in a fleet are critical for the production and billing of companies in the transport segment, whether for goods or people, requiring constant monitoring of their performance, since an unplanned interruption can have high-cost implications for the company. There are several maintenance techniques and concepts that can be applied to predict the occurrence of an equipment failure. In recent years, monitoring of health conditions and prognosis for lubricating oil has become a significant topic for people and groups in academia and industry. More and more efforts have been put into the development and research of fault diagnosis and prognosis systems based on lubricant tests. The purpose of most research is, by monitoring the oil degradation process, to provide early warning of machine failure and also to extend the operating life of the oil to increase machine availability, avoid unnecessary oil change costs, waste or environmental pollution (A. Kumar & Ghosh, 2016).

1.1 Problem Statement and Objectives

As maintenance solutions in the transportation field become more complex, with the insertion of new systems and increasing amounts of information, this sector needs new methodologies to become more efficient and reduce costs. According to Prytz (2014), new fleet management and maintenance decision support systems are gradually expanded with new features to improve reliability and planning. For the development of predictive diagnosis tools and new intelligent vehicle maintenance methods, extensive experimentation and lengthy modelling are required during the development of decision support systems (Keartland & Van Zyl, 2020). Thus, the use of Artificial Intelligence (AI) mechanisms, which can be adapted according to the needs of the organizations and the equipment present in the fleet, are fundamental. In this way, the main objective of this work is to develop a system based on data from lubricating oils of diesel engines of a fleet of passenger transportation vehicles, capable of helping the manager in the decision-making process of asset maintenance, avoiding high costs with the serious failure of vital organs of the equipment.

To achieve the general objective of this study, it is necessary to divide it into several development stages and to elaborate specific objectives:

1. Contextualize the digital transformation in maintenance and the categories of technologies needed to develop a fault prediction system;
2. Delve into the study of Machine Learning (ML) to build a model and its operationalization in organizations;
3. Understand the scenario and the failures with the highest number of occurrences, as a cut-out of the passenger transportation vehicles maintenance sector, besides the main applications of ML in this segment;
4. Conduct case studies with the goal of developing applications of ML to improve failure prediction in automotive lubrication systems in heavy-duty passenger vehicles;

5. Develop a data-driven failure prediction system capable of being implemented in a real environment;
6. Identify the context and the construction of the case study models to propose a failure prediction system that will be implemented in a case study environment;

1.1.1 Originality and Research Contribution

By using the model developed in this work fleet managers will be able to apply new maintenance techniques based on actual data and not just estimates. On the other hand, and not less important, the academic contribution of this decision making model is based on the detailed study of the parameters that indicate and determine recurrent failures in automotive systems, as well as on the elaboration of technical procedures for solving potential failures, which can be used for the theoretical enrichment of professionals in the maintenance field of organizations. In terms of originality, this research stands out in aspects related to the development of a computational model based on ML techniques, applied to fleet maintenance, capable of being operated by the entire organization and containing data and information relevant to the total operation of the assets that generate profits for the company. Another fundamental point of originality of this specific study is the determination of performance indicators that, in combination with practice and theory, generate a competitive differential in the strategic management of the organization's business where the developed model is inserted. The work presented in this document incorporates solutions and methods developed in an academic environment, through the development of intelligent decision-making models and results obtained in the real world business environment with the collection of data from lubricants used in the operation of fleets. The link between the academic and business environment gives this system a high degree of relevance, since many studies developed within this scope only focus on theoretical analysis without any real practical application. The prototype described and developed in this project is focused on automotive vehicle fleet companies, which provide services in the most diverse areas such as: long and short distance collective passenger transportation; long and short distance freight transportation and may eventually be suitable for other similar service companies. In summary, the transport sector, whether freight or passenger transport, is a network classified with a high degree of uncertain failures and needs to provide a service with high reliability. Thus, with the creation of a model capable of determining the operating condition of lubricants, which is a crucial vehicle system, and furthermore determining the general condition of the asset itself, transportation companies will have a key tool for improving maintenance management and reduce unnecessary costs.

1.2 Methodology

Scientific Methodology addresses the main rules of scientific production, providing the techniques, instruments and objectives for better performance and quality of scientific work. According

to Muchiri et al. (2018), in the last decade, several works on lubricant analysis have been developed, in many different business segments. It is important to mention that many of these studies are purely theoretical, focusing on the development of academic models based on the analysis of operational data from a wide variety of organizational sources, with the common goal of increasing equipment reliability and availability.

Since this project as a substantial component of research, planning is necessary and the researcher needs to be very clear about his research goal, how it is posed, how it is problematized, what are the hypotheses he is raising to solve the problem, what theoretical elements he can count on, what instrumental resources he has to carry out the research and what stages he intends to go through. Thus, it is important to begin this project by formulating the Research Questions (RQ), as identified in Figure 1:

- **RQ1:** Is it possible to improve the performance and determination of the degradation index of lubricating oils in diesel engines?
- **RQ2:** Is it possible to use real-time data to predict failures in automotive lubrication systems using ML methods and algorithms?
- **RQ3:** If we apply supervised learning models how will we acquire the labels of the lubrication system?
- **RQ4:** Will the system be able to be implemented in an enterprise environment with real data?

According to Dogan & Birant (2021), a research question is the statement of a specific inquiry that the researcher wishes to answer to address the research problem. The research question or questions guide the types of data to be collected and the type of study to be conducted. After determining the RQ's, we can determine the steps for the development of this project until its implementation and through the analysis of Figure 1, we can identify the following steps:

1. Bibliographic review of books and scientific papers that make up the state of the art in ML focused on the automotive sector and based on lubricant oil data;
2. Study of the main scientific papers that describe Maintenance, Diesel Engines, Lubricants, Big Data and ML techniques;
3. Documentation and archiving of all the tools used, all data generated and results obtained for future use.

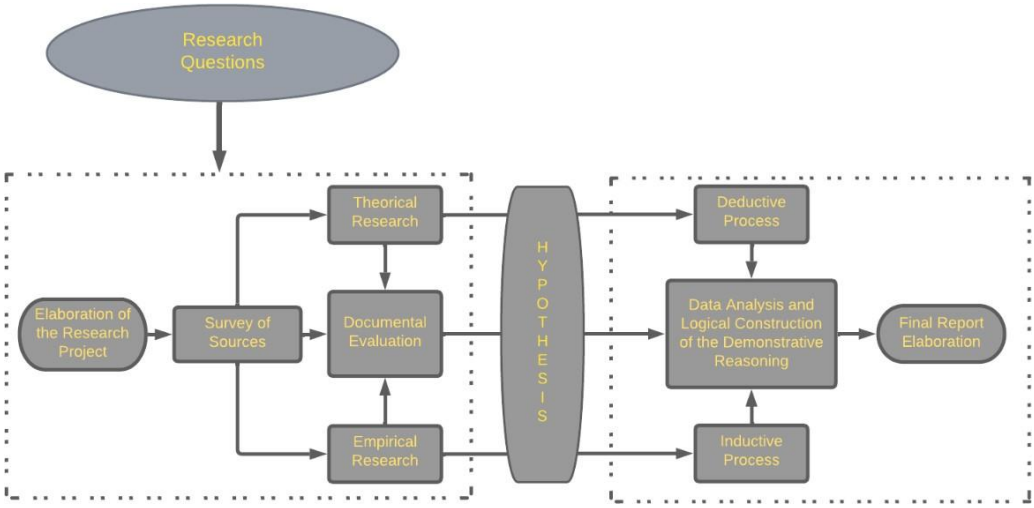

Figure 1 - Work Methodology Structure

According to Mitchell (1997), formulating hypotheses without having carefully reviewed the literature can lead to errors such as suggesting hypotheses about something that has been very well proven or something that has no theoretical or practical value. In short, the quality of hypotheses is positively related to a thorough literature review, in order to avoid wasting time on fruitless research without a viable contribution to either the academic framework or a practical approach. Therefore, based on the detailed reading and analysis of the available materials related to the determination of the condition of lubricants and diesel engines, we can elaborate our hypothesis:

- Machine Learning models allow automated prediction of lubricant and equipment condition, thus reducing the downtime and maintenance costs associated with diesel vehicles.

In this way the hypothesis fulfils the objectives and definitions of the hypothesis as a guideline for the execution of the research. Therefore, the terms used in the hypothesis must clarify, as precisely as possible, what they mean in the context of the research to be carried out. In this way we can proceed to the second block determined in Figure 1 and following the steps below.

1. Writing and publishing scientific articles in conferences and journals related to the theme of this project;
2. Dissemination and presentation to the local and international academic community of all the results obtained at the end of each stage developed in this work.

With a clear vision of the path to be taken, to further analyze the materials and documents collected as the theoretical basis of this development, a bibliographic mapping analysis, was performed in order to determine patterns in the use of the main lubricant analysis tools, in the use of equipment maintenance data, in reliability engineering concepts, in ML and data mining techniques used, in

addition to the authors' study trends. As for the database needed for the experiments, it was provided by Stratio Automotive (Automotive, 2021).

1.3 Contributions and Publications

As result of the work presented in this dissertation, the following contributions can be highlighted:

- **Wear and Tear: A Data Driven Analysis of the Operating Condition of Lubricant Oils - (Appendix A):** In this paper we present an intelligent data analysis from 5 different vehicles to evaluate whether the variables collected make it possible to determine the operating condition of lubricants. The results presented show that the selected variables have the potential to determine the operating condition, and that they are highly related with the lubricant condition. We also evaluate the inclusion of new variables engineered from raw data for a better determination of the operating condition. One of such variables is the kinematic viscosity which we show to have a relevant role in characterizing the lubricant condition. Moreover, 3 of the 4 variables that explaining 90% of the variance in the original data resulted from our feature engineering.

This document was presented in APMS 2021 - Advances in Production Management Systems. Artificial Intelligence for Sustainable and Resilient Production Systems (Malaguti et al., 2021b)

- **A Well Lubricated Machine: A Data Driven Model for Lubricant Oil Conditions - (Appendix B):** In this paper we present a supervised ML framework based on the Random Forest Classifier (RFC) to determine the operating condition of lubricant oil in diesel engines based on data from 5 different vehicles. We describe the how practitioners should collect and process data, and which features can be engineered to help describe the state of the lubrication system. This data will then be used by a RF model to determine the operational condition of the lubricating oil. The results presented show that the proposed approach is able to successfully identify the oil operating conditions, with the predictive model obtaining a Recall of 97.9%, a Precision of 99.5% and a F1-score of 98.7%. In addition, we evaluate the importance is the inclusion of new engineered features projected from raw data for better determination of the operating condition.

This document was presented in EPIA 2021 - Conference on Artificial Intelligence (Malaguti et al., 2021a).

- **Um modelo de aprendizagem supervisionado para a determinação das condições de operação de óleos lubrificantes - (Appendix C):** In this document we present a summary of the two previous documents presented at EPIA and APMS with a less technical structure in the area of artificial intelligence and ML algorithms, submitted in

the Portuguese language and targeted at operators and maintenance managers in Portugal. With this document we prioritized that the collaborators of the maintenance segment of transport companies could be aware of the new technologies used in the sector as a way to assist in the decision-making process when repairing an asset.

This document was presented in 16.º Congresso Nacional de Manutenção e 8.º Encontro de Manutenção dos Países de Língua Oficial Portuguesa.

- **A Supervised Machine Learning Model for Determining Lubricant Oil Operating Conditions – (Appendix D):** This paper presents an intelligent system for assessing the condition of lubricating oil in automotive diesel engines. To this end we analyze the use of raw data obtained from the sensors installed in the car and evaluate in conjunction with the insertion of engineered features designed the best way to determine the operating state of the oils. The results presented in this analysis show that to explain 90% of the variation in the original data only the variables Kinematic Viscosity, Dynamic Viscosity, Engine Oil Temperature and OSF_v3 are needed. After evaluating the quality of the variables, we conducted an experimental study to analyze the performance of various ML algorithms, taking into account the number of features as input data. The results show that the proposed system has the ability to identify the operating conditions of lubricating oil using 7 variables as input to a model based on Gradient Boosting, obtaining a Recall result of 93%, Precision of 96% and F1-score of 94%. We conducted a set of additional studies to understand how different subsets of variables affected the performance of the models, and the results show that the best combination includes information regarding the engine speed, coolant and oil temperature, oil pressure, the oil stress factor (OSF_v3), kinematic viscosity, and the dynamic viscosity).

  This document was published in Expert Systems, Online ISSN: 1468-0394, Clarivate Analytics Impact Factor JCR 2020 2.587 (Malaguti et al., 2022).

1.4 Thesis Framework and Structure

In terms of practical relevance, this study will enable entrepreneurs in the transportation sector conditions for their organizations to have a system capable of identifying the operating condition of lubricants and the equipment itself, thus assisting in vehicle fleet maintenance, minimizing the economic impacts caused by the catastrophic failure of important vehicle element. According to Carvalho et al. (2019), most problems occur when critical factors are neglected and the development of a ML model, can minimize the high costs generated in this area. The academic contribution of this work lies in the fact that this study reviews the parameters that indicate asset failures based on real lubricant data. Although this system was developed for implementation in a company in the passenger transportation sector based on real data, it is still part of the doctoral thesis of the author of this

document. In this way it is necessary to indicate the development structure of this project as a whole, culminating in the presentation of the final thesis document and the results achieved.

According to Figure 2, the project is divided into 3 segments: **Initial Development/Planning, Practical Development and PhD Thesis,** with each segment having its specific tasks and milestones that support and validate the final document.

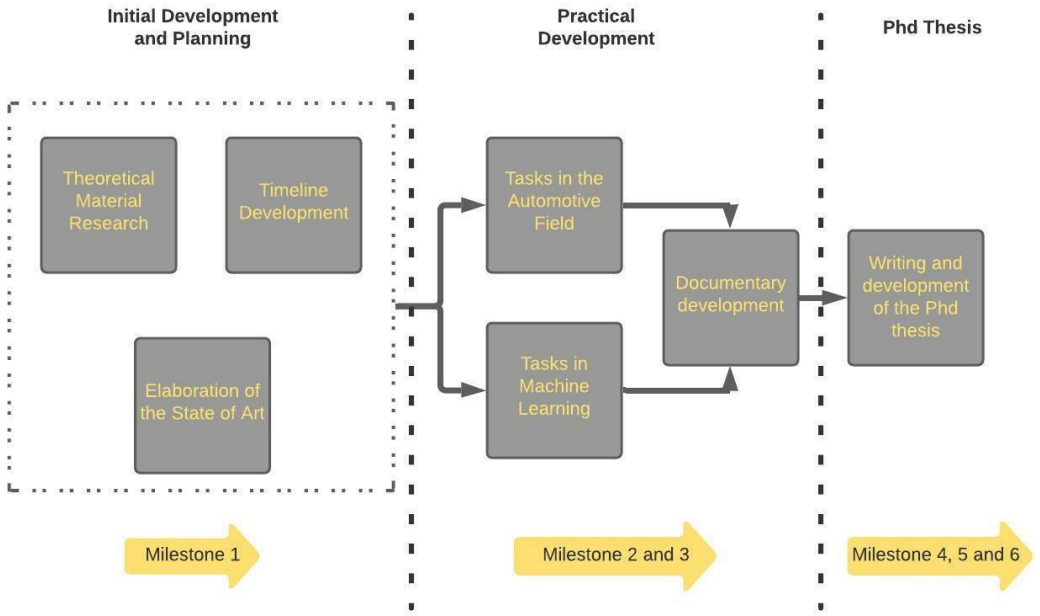

Figure 2 - Thesis Framework

As mentioned, to achieve the objective of this project, the structure shown in Figure 2 was elaborated. This work was subdivided into five main tasks, described hereafter.

1. **Initial Development/Planning**

   1.1. Development Schedule - Development of the work schedule with delivery schedules;

   1.2. Research of Theoretical Material - Research of practical theoretical reference material for the basis of the project;

   1.3. State of the art analysis - The need to introduce some pertinent issues for the development of this project makes the state of the art an important and crucial piece for its proper development of the model based on real data for implementation in a business environment;

   1.4. Milestone 1 - Submitting the doctoral project for evaluation - 1st year;

After the first stage of this project was developed the remaining tasks were segmented into stages and development areas, this type of structuring gave rise to the model needed to answer the project's general problem, the stages will be described below:

2. **Pratical Development – Automotive Field**

2.1. Study of the variable's behaviour - All variables measured in a vehicle, have limits and working areas determined by the manufacturer. This step consists of collecting system variables and analyzing their operational behaviour. (i.e., Engine System - Oil temperature, coolant temperature, engine speed, torque, etc.);

2.2. System modelling - Creation of a tool for analysis and comparison between variables - The first tool will be elaborated with the objective of determining comparison weights between variables of the same system;

2.3. Simulation of the system - Elaboration of the correspondence analysis of each variable, in this step several statistical and probabilistic methods were used, and the method with the lowest associated error was used to determine the capacity of the variables to answer the identified problem;

2.4. Milestone 2 - Article "Wear and Tear: A Data Driven Analysis of the Operating Condition of Lubricant Oils";

3. Pratical Development – Machine Learning

3.1. Study of programming software and function search - This subtask will focus on the research of software, intelligent methods and programs developed in the field of predictive maintenance;

3.2. Data cleaning and management platform development - In this subtask all databases will be collected and will be prepared in a standard format, which will start the extraction of knowledge and information from the data;

3.3. Development of testing and simulation algorithms - In this step a simulation code will be developed, based on the evolution of the behaviour of each variable, its combinations, its levels of influence on the systems (weights in the system) and the forecasts established. The tool will allow the user to analyze the reduction, increase and/or change in the behaviour of a variable or a group of variables and verify what level of influence they have on the vehicle's behaviour;

3.4. Testing and tuning - The development of the case study will be carried out from a passenger transportation company and the data will be collected from the various sources of information available in the organization. In this scenario, the application of the computational model for the analysis and evaluation of the several variables contained in the maintenance management, such as: failure, failure by nature, time, number of occurrences, among others, can be used in this study and system tuning;

3.5. Milestone 3 – Article "A Well Lubricated Machine: A Data Driven Model for Lubricant Oil Conditions";

4. Pratical Development – Documentary development

4.1. Organization and elaboration of the theoretical descriptive material of the project;

5. Final Writing and Conclusions

5.1. Milestone 4 – Article "Um modelo supervisionado de aprendizagem de máquinas para determinar as condições de funcionamento do óleo lubrificante";

5.2. Milestone 5 – Article "A Supervised Machine Learning Model for Determining Lubricant Oil Operating Conditions";

5.3. Write the PhD thesis;

5.4. Milestone 6 - Delivery and defence of a descriptive document of the project to obtain the title of Doctor in Mechanical Engineering with specialization in propulsion systems by the University of Coimbra -FCTUC

Following this design structure developed by the author, this document is structured as follows:

- Chapter 2, presents the related work and the literature review: In view of this continuation of the growing research in the field of maintenance, with a focus on diesel engine lubricating oils, the author of this document initially evaluated a documental database, from the year 2000 to the year 2020, which followed the same assumptions of the present study. This research and evaluation determined a consolidated theoretical basis, with the evolution of maintenance methodologies and techniques based on determining the condition of equipment through lubricant analysis.
- Chapter 3, describes the overview of the approach and the system by developing the interconnection of the concepts mentioned in the literature review chapter. The company involved in the development of this project, the selected fleet, and a brief summary of the author's work environment are also covered in this chapter;
- Chapter 4, details the process of data acquisition and the labels needed to confirm the results: After this documentary analysis, it was possible to develop a suitable predictive model for determining the condition of lubricants and equipment with a focus on implementing it in a real business environment. However, before mentioning the development of the model, it was necessary to develop several analyses according to the selected variables in order to determine their potential for the model's implementation;
- Chapter 5, describes the experimental scenarios used in our study, explaining the procedure carried out to discover patterns, detect anomalies and test the hypothesis, which allows a better understanding of the variables and the relationships between them through the use

of statistical techniques. In other words, it details the investigation of the data set, summarizing its main characteristics;

- Chapter 6, Finally, after confirming the ability of the selected variables to identify the operating conditions of the lubricant and/or the equipment, the general objective of this study, which is the Determination of the degradation index of lubricating oils in diesel engines, can be achieved by performing a sequence of tests to determine the best predictive model for implementation in a fleet;
- Chapter 7, brings together the main conclusions and possible future works.

Chapter 2 - Background and Related Work

The aim of this Chapter is to introduce the background of our work. Figure 3 presents a roadmap for the structure of the Chapter. We start by presenting concepts related to the field of maintenance. Then, we introduce the concepts related to Diesel engines, which is the equipment analyzed in this work, followed by a discussion of the importance of lubricants. Finally, and since we are interested in developing an intelligent system to the prediction of lubricant condition, we introduce some concepts regarding Data Analysis and Machine Learning.

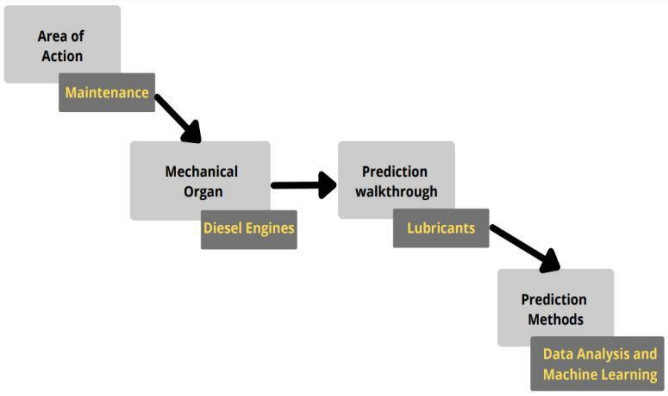

Figure 3 - Related Work Flowchart

2.1 Maintenance – Concepts and Definitions

According to Cabral (2013) maintenance tasks are done to keep equipment in proper working condition. To achieve this goal several approaches can be followed: planned and unplanned corrective maintenance; preventive maintenance; predictive maintenance; detective maintenance; and maintenance engineering (Figure 4).

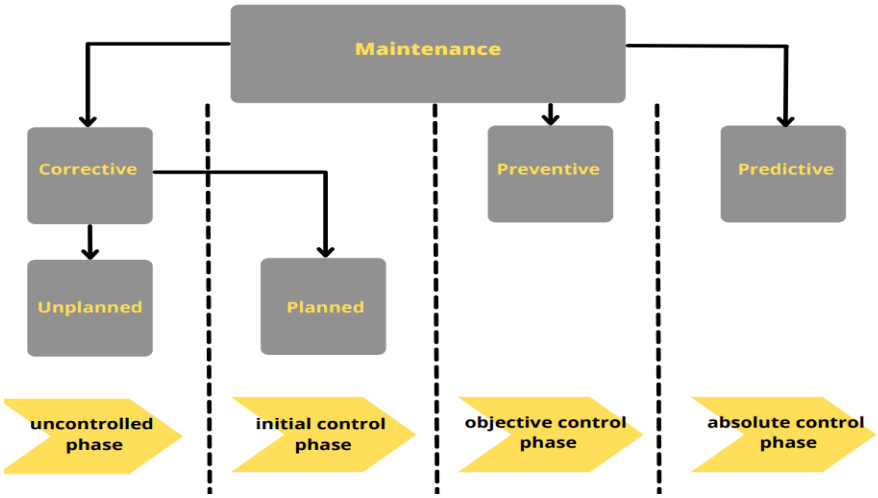

Figure 4 - Maintenance Types

Planned corrective maintenance refers to repair actions that are needed but can be deferred to a later date, which can be due to a limited budget, time, or staff. The unplanned corrective maintenance is performed at the time that the equipment does not perform its functions as expected. Preventive maintenance is based on performing its tasks according to a certain schedule, which aims at reducing failures, reducing costs and improving the equipment's performance. Predictive maintenance is performed when the equipment needs it. It implies identifying potential failures, allowing a better intervention planning to the manager. According to Dunn (2015), there are different concepts about the definition of preventive maintenance and predictive maintenance. Some experts in the field, as well as organizations that provide specialized services, define predictive maintenance as a specific subset of preventive maintenance techniques, with the goal of avoiding failures and downtime in the service of critical equipment. According to the general objective of this work, the concepts of predictive maintenance and preventive maintenance will be determined from Reliability Centred Maintenance (RCM) methodology (Whitaker et al., 2018). Thus, preventive maintenance is determined as a set of routine activities that aim for a particular component of equipment to be replaced or overhauled at a specific, predefined interval, regardless of its condition and predictive maintenance, is a set of inspection or testing activities aimed at determining the presence of failure warning conditions in a gradual process of condition points of the overall component or equipment. From the determination of this degradation index, a corrective maintenance action is then scheduled for replacement, repair or overhaul of the item before it suffers a breakdown in service.

For example, we can consider the replacement of oil in an engine every 10000 km as a preventive maintenance activity. On the other hand, the determination of the oil degradation index according to the type of use or oil viscosity index, among other factors, that determine the exact moment of failure, can be classified as predictive maintenance.

2.2 Reliability Engineering

According to Ndustries (2012), the goal of reliability engineering is based on the identification of failures in modules of certain systems that are critical to the operation of the organization, preventing these failures from occurring at the operational level. Based on reliability engineering, systems considered critical must take into consideration the functional safety approach that concerns the operation in a safe way for the operators and the informational safety that deals with the protection and integrity of the generated data. Ndustries (2012) also states that for the design process of a system, the relationship between safety and failure determination must take into consideration the need for implementation of automatic control mechanisms. According to this statement, a transportation maintenance management system, given the high risk in traffic, should adopt safety mechanisms with automatic control to avoid accidents.

2.2.1 Reliability Centered Maintenance (RCM)

According to Cabral (2013), the maintenance process and structure in an organization must include and ensure that all technical and organizational tasks operate within the degrees of reliability acceptable to the company. In this way it can be stated that maintenance services and repairs that follow specific technical premises reduce the chances of breakdowns and consequently unexpected expenses. On the other hand, one must take into consideration that the equipment in an organization, depending on their functions, have different needs and importance. As such, one needs to specify distinct maintenance policies and guidelines for distinct groups of equipment. According to Afefy (2010), RCM, defines a methodology that ensures components, systems or processes, maintain their tasks with risk control and environmental integrity. According to Cabral (2013), the RCM methodology is the preservation of the functions of a given system, identifying failures, with the help of the Failure Mode and Effect Analysis (FMEA), classifying and prioritizing the failures according to their respective consequences. In this sense Butdee & Kullawong (2015) states that RCM describes the types of failures according to their consequences, process safety and environmental impacts, focusing on the hidden failures that are unnoticeable to the operator or maintenance professional. Another important point of RCM is the use of probabilistic and statistical models referring to the failure modes and effects tools. Previously, the organizations adopted practices that did not prioritize the operational context in the elaboration of the maintenance plan, but through the application of RCM, the tasks performed in the equipment became the focus in the maintenance analysis. This change in the modus operandi context provided the creation of a proper format of maintenance analysis focused on the insertion of equipment and its components in the organizational context. According to Afefy (2010), RCM provides a structured method of selecting maintenance tasks in any process and organization and one of the main goals in this maintenance methodology is cost minimization, focusing on the main functions of the system and avoiding unnecessary maintenance actions. Thus, if an organization already has a structured maintenance program, the insertion of the RCM methodology will eliminate inefficient preventive maintenance activities.

2.2.1.1 Attributes of RCM

RCM can be seen as an organizational maintenance strategy that is implemented to optimize the maintenance program of a company, matching individual assets with the techniques most likely to deliver cost-effective outcomes. This results in a greater availability, reliability and reduction of operational costs. Additionally, it guarantees an adequate inventory planning for the minimization of industrial maintenance costs. Some attributes that define RCM are:

- Preservation of system function;
- Identification of failure modes and effects that enable the loss of priority functions;
- Prioritization of applicable and effective maintenance tasks.

This type of tool has been applied in the industry with a focus on ensuring that the equipment performed its functions properly, with the standards for which it was designed, in addition to guiding a policy of improvement in the maintenance adopted under a technical and economical point of view. Ndustries (2012) states that the need for a group of experienced professionals and the availability of a volume of reliable data that indicates the equipment failures, is of fundamental importance for the elaboration and implementation of this maintenance methodology. When implementing this methodology, the working groups are focused on increasing equipment reliability rates, concentrating their focus on equipment that is critical to the organization as a whole. RCM was built with the intention of maintaining the balance between cost and best preventive maintenance plan. According to Bloom (2005), RCM treats maintenance through a detailed and rigorous study of each equipment's reliability. In this process, the identification of each equipment's reliability rates and its relation to the process as a whole requires a maintenance team with a high degree of technical and theoretical capabilities for the elaboration of reliability studies. One way of characterizing RCM is the great interaction between the maintenance departments and other specialties of the organization, which together are the main developers of improvements in the company's equipment. We can enumerate several positive points of RCM:

1. Environmental protection and increased safety, which are determined from the information generated by identifying the possible risks of equipment failure;
2. Optimization of Operational Performance based on the collection of information, which are necessary techniques to determine the best maintenance practices, focusing on the guarantee of greater equipment availability. This greater availability of equipment can also be seen as a shorter downtime or reduction of repair time;
3. Greater Maintenance efficiency, that through all the volume of information collected, managers can develop better maintenance techniques guaranteeing the return on all investments in the maintenance sector. According to Bloom (2005), the RCM applied in a correct way can reduce 40 to 70% of the number of routine jobs, and 10 to 30% of emergency jobs;
4. Increase useful life of the equipment, which in fact is a result of the adoption of the best maintenance techniques that guarantee, to each component of the equipment, the necessary maintenance to fulfill its functions adequately to the organization;
5. Improved data storage, which is provided through the records generated by RCM, which can be used by both maintenance and operation, inspection and projects. These databases provide fundamental information for: the identification of the skills required to the maintenance operators and in the decision of which is the best stock policy;
6. Optimization of the Work Team, that through the analysis of solutions to daily problems becomes more motivated with routine tasks. The promotion of the integration of

multifunctional teams for the analysis of solutions to problems increases the level of commitment of employees;

7. Social Impact, which, according to the correct implementation of RCM that aims at minimizing the probability of failures through adequate procedures of control of the effects and consequences of failures, makes the organization use the natural resources necessary for activities in a rational way, without waste or possible accidents with an impact on the environment.

Thus according to Afefy (2010), RCM's basic principle is that the inherent reliability of equipment lies in the quality of design and construction, and although maintenance ensures this reliability, it does not increase it. In this sense, this increase becomes possible through equipment redesign or modification.

2.2.1.2 Condition Based Maintenance

A study conducted in 2008 show that 30% of maintenance operations in Europe were not planned, that is, the interventions were merely corrective, carried out after the equipment failure (Salgueiro et al., 2015). Scheduled maintenance, on the other hand, is more effective and is the most common maintenance strategy adopted by large companies. However, this type of maintenance is based on information regarding the expected lifetime of a component or machine, which means that the active-target can be discarded while it still is in good condition or suffer from premature breakage due to flaws in the initial construction project. Scheduled maintenance does not determine a way to measure wear during the use of a given vehicle, for example. It is estimated that 18% to 30% of all maintenance costs could be avoided by implementing a more reliable maintenance methodology or by determining vehicle wear during normal operation (Salgueiro et al., 2015). Condition based preventive maintenance, also known as Predictive Maintenance or Condition Based Monitoring Maintenance, commonly known in English as Condition Based Maintenance (CBM), Predictive Maintenance (PdM) or Predictive Testing and Inspection (PT&I), had its main development in the last 30 years, due to the progress in research, software and devices for collecting, processing and interpreting data, as well as the introduction of computers in the day-to-day of maintenance. It is important to mention that RCM is a process used to determine the maintenance requirements of any physical asset in its operating context and a subset of this process is CBM. CBM is a maintenance methodology based on evidence of need, and RCM provides the rules of evidence. Unlike periodic preventive maintenance, which performs active services even though there are no apparent defects, condition-based preventive maintenance management only performs interventions on equipment after real defects are found and the evolution of their deterioration is assessed. This type of maintenance consists of the inspection and monitoring of physical parameters such as vibrations, temperatures, pressure, lubricant conditions and machinery operation, to determine which combination of these offers the best indication of the wear and tear of the equipment. Since all assets deteriorate with use, monitoring the evolution of a component of equipment over time, assessing

its tendency to deteriorate and choosing the best time to apply corrective actions before failure is extremely important for failure prevention. This can avoid the high costs of unscheduled maintenance, not only in terms of materials and labor, but also in lost profits due to downtime.

In this way, CBM is a more reliable maintenance methodology and can be exemplified through the Potential Failure (PF) curve (Figure 5), which is an essential analytical tool for a maintenance plan that is based on reliability. Even with a wide variety of analysis methods, the main information that a developer of a predictive maintenance system needs to send to the manager, to avoid the possible consequences of the failure, is the proportion of a malfunction in relation to the PF interval.

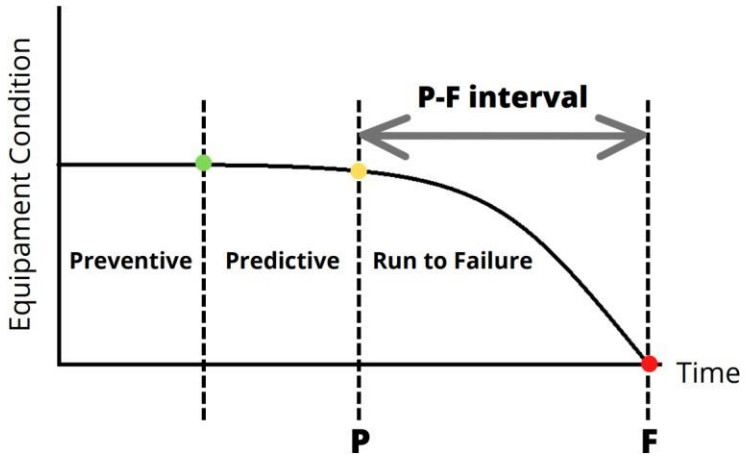

Figure 5 – Potential Failure (PF) Curve with the x-axis representing the time in service of an asset and the y-axis representing the performance of the asset in this time period – Adapted from (Prajapati & Bechtel, 2012)

According to Figure 5, it is possible to determine that the PF interval or period illustrates the failure behavior of a component or equipment, measured in hours, Km, cycles or manoeuvres. This interval is considered from the point P, which is the point at which the first sign of the existence of a potential break is detected and the point F at which the equipment has suffered a real functional failure.

In CBM most of the inspection technologies follow the same steps for data acquisition. Since the general object of this study are failures in lubrication systems in diesel engines, we will focus only on this segment. The main steps one has to follow for data acquisition are the following (Bousdekis et al., 2021):

1. The collected data are recorded and compared with historical data, reference data provided by standards and experience of technical personnel, drawings and manuals of manufacturers, among others, in order to check for the existence of defects and the respective evolution trend;
2. If a defect is found with a high rate of degradation, a service order is issued and the repair carried out before the failure occurs;

3. If the parameter is in normal conditions, or with a low or stable rate of degradation, in a next inspection it undergoes a new measurement for the purpose of checking its state of deterioration.

The results found also lead to a reassessment of the inspection program. As deterioration trends are assessed, the frequency of inspections may be subject to change, and may increase or decrease.

2.3 Diesel Engines

According to Pucher (2010), on February 27, 1892, the engineer Rudolf Diesel filed a patent with the Imperial Patent Office in Berlin for a ‘‘new rational heat engine’’. On February 23, 1893, he was granted the patent DRP 67207 for the ‘‘Working Method and Design for Combustion Engines’’ dated February 28, 1892. Diesel engine, is a type of internal combustion engine in which air is compressed to a high enough temperature to ignite diesel fuel injected into the cylinder, where combustion and expansion drives a piston. It converts the chemical energy stored in the fuel into mechanical energy, which can be used to power a wide range of small vehicles, cargo trucks, buses, large tractors, locomotives, and marine vessels (HUGHES RV, 1969).

2.3.1 Engine Characteristics

All the information needed to keep the vehicle running smoothly for as long as possible is described in the maintenance manual, including information regarding the ideal lubricating oil, which ensures the best engine performance. To develop a new vehicle model, designers and engineers work together on the project, discussing its design, mechanics, and aerodynamics. In addition to the visible aspects, the mechanical parts of the engine are also designed and tested, including specifying the lubricating oil. Thus, it is necessary to present the characteristics of the engine to which the lubricating oil is being applied:

- Engine Description: Most engine manufacturers use the name of the engine to determine some important information about it. Thus, from this name we can, in some cases, identify the technology used in it, the series and structure, among other data. However, even though we have direct access to this information, if it is necessary to have more details about the engine, the name (Ex: OM906hLA) and the brand (Mercedes Benz) are precise indicators for researching its technical file.
- Engine Cylinder Number: The cylinder of an engine is the place where a piston moves. Its name comes from its cylindrical shape. In internal combustion engines, it is in the cylinder that the fuel deflagration takes place, which is the origin of the mechanical force that enables the vehicle to move. The number of cylinders can vary from a single one to 12 or 16, depending on the project's specifications.

- Piston Diameter and Stroke: The piston in an engine is a cylindrical part usually made of aluminum or aluminum alloy, which moves inside the cylinder of internal combustion engines. To calculate the engine displacement, it is necessary to use the maximum useful volume inside the cylinder. For this volume calculation, the distance travelled inside the cylinder by the piston, called Stroke, and the cylinder diameter are taken into account.
- Engine Displacement: In the specific case of internal combustion engines, the displacement is the volume swept by a piston inside a cylinder between the Top Dead Center (TDC) and the Bottom Dead Center (BDC), for one round trip. For example, car engine with a displacement of 2 liters draws in and expels two liters of fuel for each crankshaft revolution. When the crankshaft makes one revolution, all the pistons have made one round trip. In two turns, two liters of fuel are drawn in and another two liters of exhaust gas are exhaled, or four liters in total. We can calculate the displacement in liters by the equation 1:

$$\mathrm{Displ}_{(\mathrm{liter})} = \frac{(0.785\ x\ (CD)^2\ \ x\ SL\ x\ CN\ )}{1000} \qquad (1)$$

Where: CD is a cylinder diameter, SL is a stroke length and CN is a number of cylinders.

- Engine Power: The power of an engine can be defined as the useful energy generated per unit of time. If torque is the energy generated in an explosion, then power is proportional to torque multiplied by speed. The units of power are energy/unit of time, i.e. Joule/s = Watt. In cars the standard unit is kW but in our market it is more usual to use Horse Power (HP) where 1 kW equals 1.36 HP. So an engine that delivers 100 kW at 6000 RPM is equivalent to saying that it has a power of 136 HP at 6000 rpm.
- Maximum Engine Speed: The unit Revolutions Per Minute (RPM), is widely used to characterize various types of engines and refers to the rotational speed of the crankshaft. In general, for rotating machines of any kind (electrical, hydraulic, mechanical or thermal), generator or motor, the term RPM is used to refer to the angular speed of the main axis of the machine (input, if it is a generator; output, if it is a motor). In this case, we have angular frequency and angular speed, each with its own conceptual basis, depending on the type of physical-mathematical analysis to be made. The speed range at which the motor will operate is determined during the motor design phase, and the maximum speed determines the speed limit to which the motor is dimensioned.
- Crankcase storage capacity: The crankcase surrounds the lower part of the engine, housing the crankshaft and protecting the moving parts of the engine from foreign objects, in addition to other different functions depending on the type of engine in which it is applied. On its bottom surface it has a screw lid through which the entire contents

can be emptied. The oil stored here is circulated by the oil pump, which spreads it in the moving parts of the engine. After lubricating the cylinders and crankshaft bearings, the oil returns to the crankcase and is dispersed again in a successive cycle.

2.3.2 Spark Ignition Engine VS Diesel Engine

One of the main differences between a diesel engine and a spark-ignition engine is the way ignition is accomplished. In diesel engines the ignition system does not exist and the explosion is dependent on the temperature of the air in the cylinder, that needs to be high enough to spontaneously ignite the fuel, which is added at the end of the compression stroke. This process is different from the spark ignition engine, which in the cylinder already contains an air-fuel mixture and a spark determines the ignition, not necessarily the pressure and temperature (Viskup, 2019).

Characteristcs/Engines	Diesel	Spark Ignition
Typical size range	1 kW to 80 MW	1 kW to 6.5 MW
Efficiency	30% to 48%	28% to 42%
Compression ratio	14:1 to 25:1	8:1 to 12:1

Table 1- Diesel Vs Spark Ignition Engine

By analyzing Table 1, it is possible to see that the compression ratios in diesel engines (14:1 to 25:1) are higher than in spark ignition engines (8:1 to 12:1). This is to achieve ignition, the air in the cylinder is compressed much more than in a spark-ignition engine and it is this compression that raises the temperature. To withstand the higher pressure, diesel engine components must be stronger than in a spark-ignition engine, thus making them heavier and more expensive. However, the technologies employed in diesel engines translate into higher efficiency, which can reach 50% energy conversion efficiency, significantly higher than in spark ignition engines. It is important to mention that with the high compression ratio diesel engines have a higher combustion temperature and therefore need a highly efficient cooling system.

To get a clear idea of the differences between diesel and spark ignition engines, as well as the reason why heavy vehicle fleets mainly use this type of engine, we can summarize the advantages and disadvantages of diesel engines in a few important points (Viskup, 2019):

1. Consumption: It is one of the biggest advantages of diesel engine vehicles. They can get up to 25 kilometers per liter in small vehicles, a number that is unthinkable in spark ignition vehicles of the same size;
2. Engine Price: in this aspect we have the first disadvantage of the diesel engine in comparison with spark ignition. Since its manufacture involves more advanced technologies, its costs are higher;
3. Maintenance: Although it appeared before the spark ignition engine, today the diesel engine has a more advanced technology. For this reason, its maintenance also has a higher cost;

4. Durability: another reason that makes the diesel engine more expensive is its robustness. With a well-designed maintenance plan and the proper care, these engines have a longer operating time than spark ignition engines;
5. High speed: for a high-performance vehicle at high speeds, the gasoline engine is ideal, since diesel engines are dimensioned for strength, not speed.

2.4 Lubricants

In this section we describe some important points about lubricants, detailing some concepts of condition-based maintenance with a focus on lubricants and present their physical properties and possible source of contamination. Lubricant Condition Monitoring (LCM) is considered an important condition monitoring technique, due to the extensive information derived from lubricant testing. According to J. M. Wakiru et al. (2019), LCM is not only used as an early warning system on machines, but also for the diagnosis and prognosis of CBM failures, increasing lubricant replacement intervals and consequently increasing equipment availability and causing a reduction in maintenance costs (Raposo et al., 2019). Many studies present the needs for an approach based on real data for the development of a condition-based maintenance policy. J. M. Wakiru et al. (2019) presents a detailed approach to recent research trends and the development of LCM based approaches applied to support maintenance decisions and applications in equipment diagnosis and prognosis. This study reviews and classifies LCM tests and parameters into several categories, which include physicochemical, elemental, contamination, and additive analyses. J. M. Wakiru et al. (2019) also reports approaches to analyze LCM derived data to support maintenance decisions, classifying them into four categories: statistical, model-based, artificial intelligence, and hybrid approaches. It can be concluded that this strategy of modelling or building intelligent models based on real data can accurately identify a state of wear and tear in a piece of equipment or system, making condition-based maintenance methodology more effective and reliable, and it is better to track the evolving health of a machine system than to just detect failures.

Another important point is that the study of oils can determine the rate of equipment degradation. Thus, lubricant analysis is an important tool for the development of a predictive maintenance methodology based on real data or CBM. Up to this point in the document, it can be seen that the use of lubricant analysis to determine vehicle condition is considered a CBM strategy and has been applied in several cases over the years and has shown reliable results for better maintenance management. As such, it is necessary to review concepts about lubricants, their properties, their classification and their contaminants to create an intelligent model to predict failures in diesel engines.

Lubricants are fundamental compounds for the perfect functioning of engines and equipment, reducing the friction of moving components, reducing wear and tear, and helping to cool the equipment

so that it can work and maintain the perfect temperature for its operation. The knowledge about them, their characteristics, applications and analysis, allows:

- General notion of the state of conservation of the equipment and the lubricant itself;
- Contributes to the creation of a reliable failure history of the equipment;
- Decrease in corrective maintenance costs and, therefore, increase in the company's profits;
- Implementation of a CBM methodology.

The automotive industries are intensely passionate about increasing engine efficiency to meet fuel economy, emissions and performance standards. In this way, many studies have been carried out to develop reliable approaches and models to understand the lubrication mechanisms and calculate energy losses (Delprete & Razavykia, 2018). So it is necessary to understand the central object that will be studied in this document which is the development of a system capable of determining the degradation index of lubricants and assist fleet managers in the predictive maintenance of vehicles.

2.4.1 Physical Properties of Lubricants

Lubricants are selected according to their ability to form a sliding film that protects moving parts, resisting constant attempts by heat and oxygen to change their properties, resisting shocks and mechanical loads without changing their characteristics, in addition to removing heat from components equipment. The main characteristics of lubricants are Density, Viscosity, Total Base Number (TBN), Total Acid Number (TAN), Dielectric Constant, Density and Flash Point (Diaby et al., 2013).

- Density is the ratio between mass and volume of oil at a given temperature. It is important to mention that this characteristic of the lubricant varies with temperature in an inverse manner, i.e. the higher the temperature, the lower the density. The density value of a fluid is used in viscosity calculations. This is therefore the main reason for performing oil density analysis, as without this information, it is not possible to determine the viscosity;
- Viscosity is the property of fluids corresponding to the microscopic transport of the amount of motion by molecular diffusion. That is, the higher the viscosity, the lower the speed at which the fluid moves. It is defined by Newton's law of viscosity (equation 2):

$$\tau = \mu \left(\frac{du}{dy}\right) \quad (2)$$

Where: μ= Dynamic Viscosity, τ = Shear stress and $\frac{du}{dy}$= Rate of shear deformation.

Viscosity is one of the main physical characteristics of lubricating oil and it is important to explain its variation in relation to temperature. In recent years, there has been a demand for lubricants for high performance engines, especially in the aerospace and automotive industries, which has led to

the development of different types of synthetic lubricants, with modifiers and additives, which change properties and make lubricants capable of operating at high temperatures without decomposition and with a low risk of combustion. Several studies show that during its normal operation in internal combustion engines, the oil viscosity decreases with increasing temperature and over time it loses its viscous properties. In this way, it can be determined that this reduction in viscosity depends mainly on the technical condition of the engine, the amount of oil in the lubricating oil system, the ignition mode, its operation and the operating time of the oil in the system. The engine is more likely to fail if there is inefficient lubrication or the lubricant viscosity is inadequate for the system to operate (Młynarczyk & Sikora, 2014).

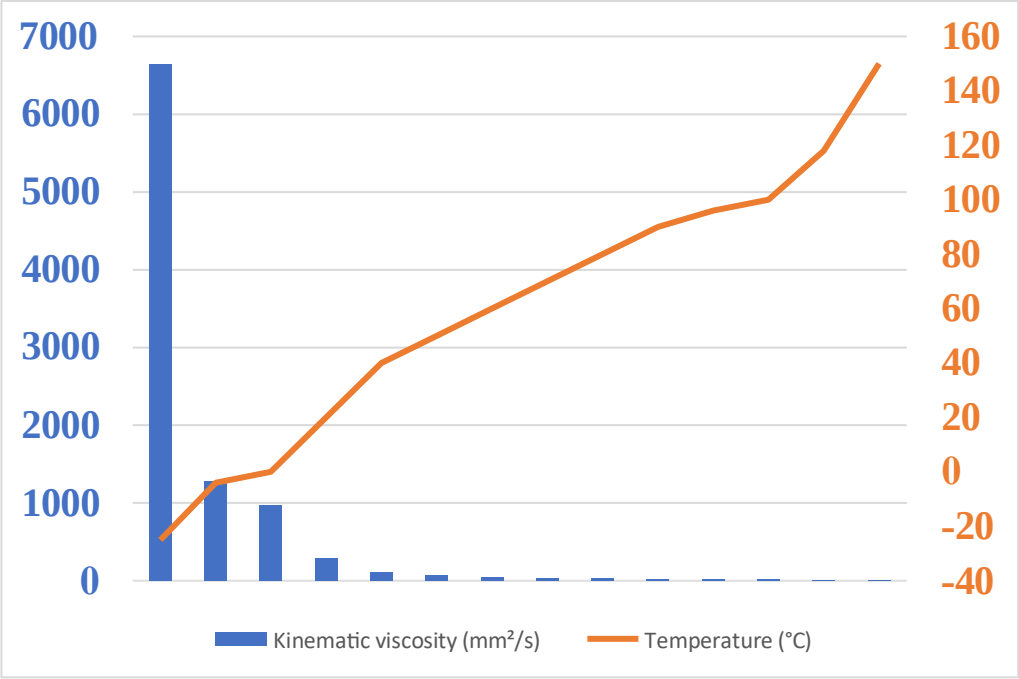

Figure 6 - Temperature Vs Kinematic Viscosity [Tauzia et al, 2012]

Looking a Figure 6 (Tauzia et al., 2019), the oil viscosity generally increases when the oil temperature is lower than the nominal one, which in turn increases friction. This effect is known to have some important disadvantages, but the detailed description of these effects is rarely presented in the literature. One can also cite the work developed by Sejkorová et al. (2017), having the objective of presenting a comparison between analyses of two samples of motor oils worn in an IVECO CITY bus. Sejkorová et al. (2017) refers that the lubricants should serve integrally in various operating conditions, which are extreme and/or in adverse climatic conditions, maintaining their capacity without determining risks to the engine;

- TBN and TAN: TBN determines the effectiveness and control of acids that arise during the combustion process. The higher the TBN, the greater the effectiveness in eliminating the contaminants that cause wear and tear and reducing the corrosive effects of acids over

a prolonged period. The TAN represents the mass necessary to neutralize one gram of oil, being the measure of all substances contained in the oil that react with potassium hydroxide. The most common constituents of such acid products are organic acids, metal soaps, oxidation products, nitrites and other compounds, which can be present as additives and which react with potassium hydroxide (Zadorozhnaya et al., 2016). The study by Zadorozhnaya et al. (2016) informs that any oil whose pH is between 4.0 and 11.0 will have a TBN, equivalent to the amount of acid needed to bring the pH up to 4.0 and a TAN needed to bring the pH to the upper limit of 11.0;

- Dielectric Constant: We live surrounded by a multitude of materials and some of them have properties that favor or hinder the passage of electricity. Thus, it is necessary to determine a constant that classifies the material according to the ease with which electric charges move through its structure. Pawashe et al. (2017) presents an analysis carried out in MATLAB with reference to the variation of the lubricating oil dielectric constant during the period of use, comparing the measurements obtained from this constant in oils of the same type and brand. The authors conclude that this variation of values in the dielectric constant allows to determine the degree of change in the oil condition. The dielectric change is directly related to oil degradation and contamination level and allows the user to obtain optimized intervals between oil changes and to detect greater mechanical wear as well as the loss of lubricating properties of the oil. In summary, the variation in the rate of iron, water or other compound content in lubricants causes a variation in the value of the lubricants dielectric constant. According to Pawashe et al. (2017), choosing the right time for oil change can avoid the risk of damage to the engine and also reduce the cost of using oil during the period of operation. The study presents the range of various parameters according to the dielectric constant analysis.

    1. The TAN must be between 0.8 and 1.76 mg KOH / g.
    2. The maximum iron content in the oil can be up to 69 μg / g of oil.
    3. The minimum oil density can reach 780 kg / m3.

Also according to Pawashe et al. (2017), if the measured values violate any of the 3 conditions above, the lubricant does not present the ideal working conditions and must be changed;

- Flash point is the lowest temperature (ºC) at which a liquid gives off vapor or gas in an amount sufficient to form a flammable mixture. That is, an oil sample develops enough vapors under specified conditions for the air-vapor mixture above the sample to ignite for the first time without burning further afterwards. Flash point alone is not a sufficient quality characteristic of oil, nor does it allow conclusions about the suitability of the oil, but by evaluating the data from this characteristic with others such as viscosity, working

pressure and temperature, we can determine the optimum working conditions of the system (Zadorozhnaya et al., 2016);

2.4.2 Lubricant Classification

To compare the degradation indexes of different lubricants it is important to classify them according to their properties. The best-known classification of engine oils is the one proposed by the Society of Automotive Engineers (SAE). This classification is based on the viscosity of lubricants, not considering factors of quality or performance. The first standard of lubricants, J300, was developed in 1911, and although this standard has been revised and updated many times, it is still used worldwide today for engine oil applications (Dv- & Dv-, 2014). Table 2, presents the classification of lubricants according to their viscosity.

SAE Viscosity Grade	Viscosity (cP) at Temp., (°C) Maximum		Viscosity (cSt) at 100 °C		Viscosity after shear (cP) at 150 °C
	Start	Pumping	Min	Max	
0W	3250 at -30	60000 at -40	3.80	-	-
5W	3500 at -25	60000 at -35	3.80	-	-
10W	3500 at -20	60000 at -30	4.10	-	-
15W	3500 at -15	60000 at -25	5.60	-	-
20W	4500 at -10	60000 at -20	5.60	-	-
25W	6000 at -5	60000 at -15	9.30	-	-
20			5.60	< 9.30	2.60
30			9.30	< 12.50	2.90
40			12.50	< 16.30	2.90
40			12.50	< 16.30	3.70
50			16.30	< 21.90	3.70
60			21.90	< 26.10	3.70

Table 2 - Society of Automotive Engineers (SAE) Classification According to Viscosity

Through the SAE classification and the analysis of Table 2, it is possible to identify that the thinner the lubricant is, the lower its grade number. In the SAE viscosity grade, the number refers to the viscosity at a given temperature, and the letter "W" following the SAE grade comes from the word "Winter" and indicates that the oil is suitable for use in cold weather locations, but within the limits of one of the grades for high temperature locations. Oils that carry the W designation must have an adequate viscosity value when measured at the low temperatures established by SAE. The grades that do not include the "W" define oil grades for use in higher temperature locations, i.e., warm climates. Although SAE is the most widely used classification for lubricants, there are other important lubrication service classifications that are based on characteristics other than viscosity. Developed by the American Petroleum Institute (API), this classification is based on the service characteristics, i.e., it is based on

how severely the oil can work. It is the classification of the technology used in the development of the oil and is formed by two letters, being category 'A' the oldest and 'N' the one that offers the most recent technological resources (SA = oil developed in 1920 and SN = oil developed in 2011) (Dv- & Dv-, 2014);

2.4.3 Operational Properties of Lubricants

Diesel engine oils licensed and classify by the SAE and API are formulated to provide the best engine protection properties and to provide the longest life possible. However, even the highest quality oil can be challenged by a number of factors that exacerbate its own degradation. Not only are the physical characteristics of the engine oil critical, but also whether the right grade of engine oil performance is selected to counteract the stresses and exposures that can shorten an oil's life. It is therefore necessary to present two lubricant characteristics that are generally determined by the manufacturers and that are not directly related to the physical and chemical characteristics, but are rather related to the oil's operating condition and to tests performed by the lubricant manufacturers.

- Oil Drain Interval (ODI): Vehicle manufacturers generally set two oil change intervals, one being measured in kilometers (km interval), while the other is measured in months (time interval). Both are important, and the reason why vehicle and lubricant manufacturers choose to use these two measurements simultaneously is simple:
    1. Per mileage achieved, the more you drive, the more you use the engine and the oil. The more the oil is used, the more its reserves of additives are depleted and the faster the long-chain hydrocarbon molecules separate;
    2. Per time achieved, it is important to mention that lubricants work on the basis of additives and these additives lose their shelf life and sometimes it happens that the vehicle is left idle for a long time, allowing harmful sediments to accumulate. In other words, even when not in operation, some properties of the oil can change, deposits form, oxidation occurs, its viscosity changes, even its pH value changes. Because they are caused by the accumulation of contaminants, the process continues even when the engine is turned off.
- Oil Stress Factor (OSF): The OSF is a method for quantifying the stress under which engine oil is placed and allows some prediction of oil degradation as a function of engine conditions. There are a few ways to calculate this factor and depending on the type of solution one wants to develop, certain factors can be substituted or suppressed. For our work, 2 different ways will be used to determine the OSF:
    1. $\boldsymbol{Oil_z}$ ($^{\boldsymbol{kW.h}}/_{\boldsymbol{liter^2}}$): Oil stress factor determined by method 1 described in equation 3 (Prasad & Lakshminarayanan, 2012):

$$Oil_z = \left(\frac{1}{V_{max}}\right)\ x\ \left(\frac{N_{Vmax}}{N_{Pemax}}\right)\ x\ \left(\frac{P_e}{V_h}\right)\ x\ \left(\frac{ODI}{O_{vol}}\right) \quad (3)$$

Where: V_{max} = Vehicle top speed (km/h); N_{Vmax} = Engine top speed (rpm); N_{Pemax} = Engine speed at maximum engine power (rpm); P_e = Maximum engine power (kW); V_h = Engine displacement (Litres); ODI = Oil drain interval (km); O_{vol} = Oil volume including top ups (Litres).

2. **OSF_v3** ($^{W}/_{liter^2\ x\ 10^4}$)**:** Oil stress factor determined by method 2 described in equation 4 (Lee et al., 2005):

$$OSF\_v3 = \left(\frac{P}{V_d}\right)\ x\ \left(\frac{C_e}{V_s}\right) \quad (4)$$

Where: P = Power output (W); V_d = Volume displaced by cylinder (m^3); C_e = Number of engine combustion cycles; V_s = Volume of oil in sump (m^3).

2.4.4 Lubricant Contamination

According to Felipe Lima Bronté et al. (2015), vehicles engines are more powerful nowadays and more technologically advanced, making then more sensitive to contaminants in their lubrication systems. This implies the use of modern techniques to monitor machine and vehicle components. W. Wang (2007) refer that Shell in its advisory report, cited a series of failure statistics in which, for diesel engines, approximately 70% of service failures are due to lubricant contamination and 30% are due to wear-related problems. With such a relevant percentage of failures related to contaminants, it becomes necessary to identify which are the most common ones, in order to integrate their effects into an intelligent forecasting model. We can consider the following contaminants:

- **Metallic Compounds:** In internal combustion engines, some mechanical components that participate in the movement process, such as cylinder liner, piston, injectors and valves, are usually worn due to friction. One of the lubricant functions consists in removing harmful substances, such as metal chips and other hard materials, which are produced by incomplete combustion, abrasion and wear, on the surface of the mentioned mechanical parts, thus aiming to ensure a smooth, stable and reliable work process for the engines (Hoang & Pham, 2019). When talking about metallic compounds, we must first make a brief introduction of lubricant additives, which have the purpose of adding important improvements such as anti-wear, antioxidant, anti-corrosion, defoamer, modify viscosity, emulsify, lower the pour point, among others. To formulate these additives, various chemical substances and metallic compounds are added to the base lubricant so that it performance is improved. Table 3 presents the metallic compounds

that are identified as contaminants in the lubricants in relation to the type of wear (Aucélio et al., 2007).

Metallic Compound	Presence in additives	Possible wear
Iron	The iron element is not added to lubricants as an additive	It occurs due to the wear of cylinders, pistons, gears, rings, axles, oil pump, crankshaft and support points.
Copper	The copper element is not added to lubricants as an additive	It occurs due to the wear of valve guides, piston rings and support points.
Nickel	Nickel-containing organometallic additives are added in small amounts to the lubricating oil as an anti-wear additive	Occurs due to block wear with included cylinders, connecting rods, camshaft and intake and discharge valves.
Lead	Organic lead complexes are usually added to lubricating oil as an extreme pressure additive	It occurs due to possible wear of the support points and bearings.
Zinc	The zinc metal is added to the lubricating oil as a multifunctional additive, performing the functions of antioxidant, corrosion inhibitor, anti-wear, detergent and extreme pressure.	It occurs due to the wear of galvanized systems.

Table 3 - Contaminants in the Lubricants

From Table 3, is possible to understand that some metallic compounds are present in some additives. If the presence of these metallic compounds is above a certain concentration it means that the lubricant is contaminated. On the other hand, there are metallic compounds that are not present in additives and their presence, independently of their concentration, indicates wear of the mechanical components;

- **Fuel:** Contamination can happen when the fuel is not completely burned in the combustion chamber, so it flows into the piston skirt sump and mix with the engine oil. According to this type of anomaly, one can have several effects. The washing of the liner walls, the piston skirt and the segments, which implies that the fuel removes the oil, leaving the area without lubrication and the liner walls polished, meaning that the metallic surfaces are in direct contact with each other, causing premature wear of the mechanical parts. When polished, it is more difficult for the coating to keep the lubricant in the area, which can lead to failures, such as blocking the piston and the coating itself. Dilution of the oil, which causes the lubricant to lose its viscosity and other properties, meaning that the films formed are weaker and less able to withstand high loads that can occur at certain points, such as the areas of the rod and crankshaft bearings. A third effect of the fuel that passes through the crankcase is related to biofuel. Currently, both diesel and gasoline include biofuels in their formula (biodiesel in the first case and bioethanol in the second). As the fuel is subject to high temperatures in the sump, part of it evaporates, which means that, in the case of diesel, the portion of the biofuel becomes concentrated. This makes biodiesel less fluid and more viscous than diesel, which causes the lubricant to thicken and create greater resistance for the contact of the metal parts of the engine.
  Given the exposed effects, it can be determined that contamination through fuel dilution can result in the degradation of the engine oil in several different ways, either by having a lower viscosity and causing greater friction between the metal parts of the engine, or by increasing viscosity (in the case of biofuels), causing greater resistance to movement in the metal parts of the engine, which can lead to increased engine wear and / or unexpected failure (J. Wakiru et al., 2017);

- **Soot:** During the combustion process of the diesel engine, soot particles are produced that can be absorbed by the engine lubricant during the combustion cycles. This percentage of soot particles that contaminate the lubricant is directly linked to significant changes in lubricant performance and, consequently, greater engine wear. The main wear mechanism related to soot is abrasion, but at higher levels of soot content in the lubricant, contact deprivation can occur, which can further increase wear and corrosion of metal parts. High concentrations of soot can increase the local acid level around the piston, where high temperatures and volatile gases can coexist (Green & Lewis, 2008). Studies identify that the valve train is an area of the engine where wear is generally more likely to occur through soot lubricant contamination, because it components require a continuous supply of oil during operation. Being located close to the top of the engine, the valve train tends to operate with inadequate lubrication, particularly during a cold

start, when the oil pressure will initially be insufficient to pump the oil to the top of the engine and with increasing opening of the EGR (Exhaust Gas Recirculation) valve, which implies an increase in soot inside the engine (Green & Lewis, 2008). Motamen Salehi et al. (2017) studies the effect of oil contamination and its degradation in relation to the friction and wear of the engine oil pump, verifying that the consumption of additives by soot in the engine oil during the aging process has a significant effect on component wear. Ibrahim et al. (2013) presents in his study a review of soot measurement techniques in lubricants in terms of principle and measurement process and soot characteristics, thus informing specific advantages and disadvantages of each method. According to Ibrahim et al. (2013) it is necessary to develop methods that are capable of facing more complicated measurement challenges today, with a faster response time to generate results with comparatively high levels of accuracy and repeatability.

- **Water or Coolant:** This type of contamination causes the lubricant to thicken, decreasing its viscosity and not allowing it to have a proper flow. This can lead to boundary conditions in parts of the engine that require a certain fluidity for adequate protection. According to J. Zhu et al. (2013), contamination from coolant, also creates an acidic environment within the oil, resulting in corrosion in the system, especially on copper surfaces. It is also important to mention that with coolant contaminating the lubricant, the filters become clogged sooner, which can cause reduced flow and eventually, once the bypass pressure is reached, a condition where the oil is no longer performing its filtering role. Thus, particles are formed that would normally have been filtered out might clogs the system, disrupting the lubricating film and resulting in surface damage to components. According to Bordatchev et al. (2014), antifreeze also mixes with the oil to form small globules called oil balls. Although very small, typically 5 to 40 microns in size, they can cause major problems. These bubbles are abrasive and create surface erosion, which can produce all kinds of fatigue and lead to lubrication failures in areas with a reduced tolerance such as engine cylinders.

Some of the main functions of lubricating oils are to provide low bearing friction, transfer heat, and protect components from corrosion. However as mentioned earlier, contaminants affect the oil's ability to meet these requirements, and the most common forms of oil contamination include: Water or coolant contamination, fuel, soot deposits, and debris such as wear particles from engine components. It is important to mention that each of these types of contamination can cause a different type of problem leading to engine lubricant degradation.

2.5 Vehicle Maintenance and Lubricant Analysis

To setup a reliable database, with a group of documents to support this study, several searches of articles and academic documents were carried out, in the main online databases: B-on - Online Knowledge Library[2], Web of Science[3], ScienceDirect[4] and Google Scholar[5]. This research has allow us to determine important issues for the development of a system for determining the degradation index of lubricating oils in diesel vehicles. Thus, the following is a presentation of some of the issues: The tests of lubricants, data acquisition in vehicles and intelligent systems.

2.5.1 Lubricant Oil Testing

From the analysis of the selected documents, it is possible to understand that oil analysis can be performed using various laboratory tests by means of a sample or through mathematical analysis of real-time data of the lubricants. In these various tests, it is possible to detect whether the lubricant is contaminated with water, fuel, and to check the wear of components by analysing metal particles. This information allows us to take actions aimed at correcting the origin of the contaminations and anticipate failures arising from wear and tear. According to Taylor et al. (2005), it is important to mention that tests for lubricant oil analysis can vary based on the source component and environmental conditions, but should almost always include tests for viscosity, elemental analysis (spectrometry), moisture levels, particle counts, Fourier Transform Infrared Spectroscopy (FTIR) and acid number. Other tests that rely on source equipment include analytical ferrography, ferrous density, demulsibility, and base number testing. Table 4, summarises how tests are commonly used in each of the three main categories of oil analysis.

Oil Analysis Category	Tests
Fluid Properties	Viscosity, Acid/Base Number, FTIR
Contamination	Particle Counting, Stain Test
Wear Debris	Ferrography, FTIR, Spectrometry, Stain test

Table 4 - Oil Analysis and Tests

Looking at Table 4, we can see that the tests are normally divided into 3 categories. Each category determines which results are obtained from the performed test. The first category is "Fluid Properties" and the tests that are presented measure physical and chemical properties of the lubricant, such as: Viscosity; TAN; TBN; coloration; particle shape and size. The second category measures the level of contaminants in the lubricant, which can be coolant, water, fuel, soot and ash. Finally, the last category represents the wear of the components that the lubricant has interaction with and is it collected

[2] https://www.b-on.pt/en/
[3] https://clarivate.com/webofsciencegroup/solutions/web-of-science/
[4] https://www.sciencedirect.com/
[5] https://scholar.google.com.br/

from the measurement of metal debris/particles present in the lubricant. The referred tests are briefly described below:

1. Several methods are used to measure viscosity, which is reported in terms of kinematic or absolute viscosity. While most industrial lubricants classify viscosity in terms of International Organization for Standardization (ISO) standardized viscosity grades (ISO 3448), this does not imply that all lubricants with an ISO VG 320, for example, are exactly 320 centistokes (cSt). According to the ISO standard, each lubricant is considered to have a particular viscosity grade as long as it falls within 10 percent of the viscosity midpoint (typically that of the ISO VG number) (Lee et al., 2005), (Kumbár & Sabaliauskas, 2013);
2. Acid number and base number tests are similar but are used to interpret different lubricant and contaminant-related questions. In an oil analysis test, the acid number is the concentration of acid in the oil, while the base number is the reserve of alkalinity in the oil. Results are expressed in terms of the volume of potassium hydroxide in milligrams required to neutralize the acids in one gram of oil. Acid number testing is performed on non-crankcase oils, while base number testing is for over-based crankcase oils (Idros et al., 2012), (Guan et al., 2011);
3. FTIR is a quick and sophisticated method for determining several oil parameters including contamination from fuel, water, glycol and soot oil degradation products like oxides, nitrates and sulfates, as well as the presence of additives such as Zinc Dialkyldithiophosphate (ZDDP) and phenols. The FTIR instrument recognizes each of these characteristics by monitoring the shift in infrared absorbance at specific or a range of wavenumbers. Many of the observed parameters may not be conclusive, so, often these results are coupled with other tests and used more as supporting evidence (Sejkorová et al., 2017), (Z. Wang et al., 2018);
4. Particle counting measures the size and quantity of particles in the oil. Many techniques can be used to assess this data, which are reported on the ISO 4406:99. This standard designates three numbers separated by a forward slash providing a range number that correlates to the particle counts of particles greater than 4, 6 and 14 microns (Felipe Lima Bronté et al., 2015);
5. Stain test is one of the oldest techniques used to identify excessive engine soot, evaluate the dispersion of a lubricant, and detect the presence of glycol, diesel fuel, and other contaminants in diesel engine crankcase lubricants. To do this, it is necessary to stop the warm engine, remove the dipstick and drip 2 drops onto a filter paper and place the paper with the drop on a flat surface, so that the oil drop can spread horizontally, for 5-10 minutes. After this period, the diameters of three areas of the stain are measured to

identify the colorations and uniformity of oil spreading (Raadnui, 2005), (Wei et al., 2021);

6. Ferrography was developed in the 1970s as a response to the limitations of particle size detection existing in spectrographic methods. The technique measures concentration of both large and small ferrous wear debris in an oil sample, performed by passing the sample through a treated glass tube that is positioned over a high gradient magnetic field (S. Wang et al., 2017), (A. Kumar & Ghosh, 2016).
7. Spectrometry is a technique for detecting and quantifying the presence of elements in a material based on the fact that each element has a unique atomic structure. When subjected to the addition of energy, each element emits light of specific wavelengths, or colors. Since no two elements have the same pattern of spectral lines, the elements can be differentiated according to the intensity of the light emitted in proportion to the amount of it presence in the material studied (Guan et al., 2011), (Silveira et al., 2010).

It is important to mention that some tests may obtain results in different categories, but all must follow standardized methods such as those provided by the American Society for Testing and Materials (ASTM). According to Mujahid & Dickert (2012), in 1994, a concise technical report on the analysis of oxidized motor oil was published, explaining the classic ASTM methods. In this work, different analytical tools were described to measure engine oil contaminants, such as, insoluble pentane and toluene, soot, fuel, water, ethylene glycol, and metal ions from engine wear. The condition of the engine oil was examined by measuring the viscosity, acid and base numbers, pentane and toluene insolubles, fuel dilution, water, and ethylene glycol in the oil. All the methods mentioned have advantages and disadvantages as can be seen in Table 5.

Test	Strengths	Weaknesses
Viscosity	Viscosity is the single most important property of a lubricant. The lubricant's viscosity is what allows it to form the protective layer required for separation of moving surfaces	Measure Viscosity alone cannot be used to determine the health of a lubricant, as different parameters can cause changes in viscosity outside of the normal aging process.
FTIR	This test method is relatively quick to perform and is capable of simultaneously detecting multiple parameters, including antioxidants, water, soot, fuel, glycol, oil oxidation, and certain additives. Adding to the power of this qualitative measurement, it is possible to indicate the amount of the specific material found in the sample.	The IR spectrum is very complicated and interpretation depends on a lot of experience. This is why the equipment that performs this type of test is complex and costly.

Acid number and base number tests	This test has a high accuracy following the standards required in the industry and has a high tolerance for contaminated samples.	To perform the tests requires expensive equipment making the cost per test relatively high. To perform the tests requires the use of solvents and a qualified operator. Therefore, this test is usually performed in a laboratory.
Particle Counting	In clean systems a change in particle count can be a sensitive early warning sign of an active wear pattern. Used as a screening tool for ferrous density or complete analytical ferrography, particle counting is an effective test, which can be run onsite with minimal cost per sample.	Particle counting is not limited to wear debris and cannot differentiate between wear debris and non-metallic particles, including ingested contaminants
Ferrography	In the hands of a skilled analyst, ferrography is able to detect and analyze active machine wear and can often provide a definitive "root cause analysis" based on the morphology of the wear particles Used in conjunction with forensic-type tests such as heat treatment and chemical microscopy, ferrography can often pinpoint the root cause of a specific wear problem	Due to the method of sample preparation, ferrography is biased, though not limited to ferrous particles. This can be circumvented to some degree by preparing a filtergram instead of a ferrogram. The test is non-quantitative and its effectiveness depends critically on the knowledge and experience of the analyst. Because of the skills required of this analyst and the time the analysis usually takes, the test can be quite expensive compared to other test methods
Spectrometry	Inexpensive. No additional testing is required. No interferences from soot or water.	Spectroscopy is perhaps the most important and useful test in used-oil analysis, but is does have limitations. A key drawback is the size limit of the particles it can vaporize. It does not detect particles beyond the five- to eight-micron range. While this limit does not affect the detection of most wear situations, there are times when it could be a problem

Stain Test	The spot test has the advantages of speed, simplicity of equipment, and a high degree of sensitivity.	This test is performed by comparing the intensity of the coloration of the dots obtained on the filter paper with the color of a standard. Therefore, a skilled technical operator is required to interpret the obtained results

Table 5 - Strengths and Weaknesses of Lubricant Tests

By looking at Table 5, one can see that a variety of methods are available, from rapid field tests to expensive laboratory methods. In a laboratory environment, methods are selected based on the highest repeatability and accuracy that can be obtained with adequate throughput. In the field, a reliable result must be obtained quickly so that corrective or preventive maintenance action can be taken before any major equipment failure. The best method is the one that meets the application requirements with a reliable result.

2.5.2 Degradation Analysis Methods

In J. M. Wakiru et al. (2019), the condition of motor oil was examined by appearance and smell, measuring viscosity, water, and fuel content in oxidized oil. Samples of aged motor oil were taken from automobiles and then subjected to classical laboratory analysis using the tests described in Table 5. According to Mujahid & Dickert (2012), a detailed, technical analysis of motor oil was published in 1968, focusing on the selection and use of lubricants. The complete profile of engine oil in terms of its appearance and smell, water presence due to oxidation, kinematic viscosity, fuel dilution, insoluble matter, and ash content followed by some spectroscopic analyses were discussed. Since engine oil degradation depends mainly on contamination and oxidative stress, the analysis of the above-mentioned parameters is important for accessing the condition of the lubricant. But according to X. Zhu et al. (2017) a detailed analysis following all regulations and using large laboratory equipment, takes time and can be costly. Thus, with the advancement of miniaturization of components, which enabled the creation of sensors with better capabilities (see Table 6) and with the technological advances in the areas of ML and Data Analysis (see Table 7) the analysis of lubricant oils may have its response time and costs reduced.

Reference	Summary
(Clark & Fajardo, 2012)	The physical properties of diesel engine oil were quantified through direct, real-time measurements using an onboard sensor. The sensor measures the lubricant temperature, density, dynamic viscosity, and dielectric constant.

(C. Zhang, 2003)	A micro-acoustic wave sensor based on the thickness shear mode (TSM) quartz resonator has been investigated for engine oil quality monitoring through the viscosity measurement.
(Agoston et al., 2005)	The investigated parameter is the viscosity of the lubricating oil, which can be efficiently measured using microacoustic sensors.
(Raadnui & Kleesuwan, 2005)(Raadnui & Kleesuwan, 2005)(Raadnui & Kleesuwan, 2005)	The development of a low cost condition monitoring sensor for used oil is described. of the principle of the system consist in measuring the relative variation of the dielectric constant of lubricant caused by contaminants such as water, fuel dilution, water, wear debris.

Table 6 - Example of "Sensor" Articles for Lubricant Analysis

In Table 6 we just present a few examples of sensors used to analyse lubricant oils. Other examples could be presented, since there are many sensors, either for the determination of the degradation index or for the determination of contaminants, described in the literature and available in the market. Most of them use the principles of the tests mentioned in Table 5.

Some sensors measure the oil's dielectric constant, which changes as the oil degrades or becomes contaminated. A substance's dielectric constant reflects its ability to keep an electric field from forming in it. Others measure optical characteristics and compare them to model conditions to assess the oil's quality (a technique called Fourier transform infrared spectroscopy). There also are sensors that use magnetic fields to detect and classify metallic particles in the oil (a sign of wear). Finally some o them use x-ray emissions to detect the presence of foreign elements.

Oil sensors need to be placed on or near the asset that is being monitored. For this reason, oil analysis sensors are not suited to monitoring assets that are:

- Inaccessible (such as underground pumps);
- Remote or widely spaced (such as offshore wind turbines);
- Located in hard-to-reach places;
- Located in hazardous environments, such as areas referred to by the European standard for Equipment for Potentially Explosive Atmospheres (ATEX), concerning the danger of explosive atmospheres;
- Located in harsh conditions, such as hot strip steel mills where extreme temperatures can damage the sensors and the resultant flow of data.

Another method of determining the degradation of lubricating oils that has been made possible by technological advances, are methods that use ML tools to analyze large amounts of data and quickly and effectively determine the degradation rate. According to Keartland & Van Zyl (2020), data-

driven prognostic models require large amounts of data to provide adequate solutions. Furthermore, the use of this type of method for determining the degradation index of lubricants is only becoming possible due to the increasing availability of real-time information about the condition of machines, due to low-cost sensors and Internet connections. Just as the development of sensors becomes a trend in selected articles in Table 6, we can see a new branch forming strongly on the development of ML systems to determine the condition or degradation of lubricating oils, as can be seen in Table 7.

References	Summary
(W. Wang & Zhang, 2005)(W. Wang & Zhang, 2005)(W. Wang & Zhang, 2005)	This paper reports on a study using the available oil monitoring information, such as the data obtained using the Spectrometric Oil Analysis Programme (SOAP), to predict the residual life of a set of aircraft engines.
(Rodrigues et al., 2020)	This paper describes a model to automatically classify the oil condition, using Artificial Neural Networks and Principal Component Analysis. The study was done using data obtained from two passenger bus companies in a country of Southern Europe.
(Raposo et al., 2019)	This paper presents a case study and a model to predict maintenance interventions based on condition monitoring of diesel engine oil in urban buses by accompanying the evolution of its degradation
(J. M. Wakiru et al., 2019)	This paper systematically reviews recent research trends and development of LCM based approaches applied for maintenance decision support, and specifically, applications in equipment diagnosis and prognosis.

Table 7 - Example of "ML" articles for lubricant analysis

According to the analysis of Table 7, this ramification can be confirmed with the work of Felipe Lima Bronté et al. (2015), which mentions that lubricating oil analysis, which estimates the life cycle of components and increases machine availability, brings immense benefits for efficient vehicle fleet maintenance planning. Moreover, this type of analysis consists of taking oil samples at certain periods of engine operation, as well as after operation, and evaluating them to determine both the condition of the engine and the condition of the lubricant and whether it maintains its properties. Felipe Lima Bronté et al. (2015) states that oil analysis can be done in the laboratory or online (during operation). In the tests sent to the laboratory, the evaluation occurs by analyzing the number of particles found in its volume, the size of these particles, their shape and composition.

2.5.3 Discussion of the Literature Review

Through the document analysis performed, it can be concluded that in addition to anticipating risks, oil analysis reduces maintenance costs, avoids unnecessary downtime, controls wear and tear, and increases the useful life of equipment. Additionally, it eliminates the need for intrusive inspections of the equipment, which is often unnecessary and costly compared to taking a lubricating oil sample. This is because oil analysis performs a diagnostic of the physical and chemical condition of the lubricants and the equipment itself. If the company does not perform predictive maintenance with oil analysis, it is more prone to failures, and when these occur, corrective maintenance will be required, which is much more costly and affects productivity that could be easily avoided with predictive maintenance. In other words, despite the initial investment with predictive maintenance, it ends up generating a greater cost-benefit by making surgical corrections to avoid larger problems.

Therefore, lubricant analysis is one of the tools that allows the anticipation of catastrophic equipment failures as such over the years we have seen advances in the determination of the lubricant degradation index and flow. These advances may be related to the use of classical techniques and tools as presented in Table 4 and Table 5, but they are directed towards using new maintenance methodologies based on real-time data analysis, either using online sensors or through the development of mathematical models and/or intelligent ML systems.

It can be concluded from this analysis, that lubricant analysis, is currently divided into various techniques and methods for determining the state or rate of lubricant degradation. However, it should be noted that the selection of a procedure appropriate to the user's needs as a means of effective maintenance planning, should not only take into consideration the choice of a technique, be it the most reliable or the most reactive, but it should take into account the costs associated with each technique, the response time and the ability of the maintenance team to interpret the results. Therefore, to proceed with a technique or method to determine the lubricant degradation index and use this information as equipment condition, some factors should be considered namely: The point of oil sample collection; The cost associated with the analysis; The types of parameters to be evaluated; The interpretation of the results. Taking into account these factors for an oil analysis test to be reliable it should cover three categories of analysis: Fluid Properties, Contaminants, and Wear Residues. Therefore, it is important to mention the work of J. Zhu et al. (2012), which provides a comprehensive review of existing solutions for monitoring lubricating oil conditions and characteristics, along with the classification and evaluation of each technique. In this review, the techniques are analyzed and classified into four categories: electrical (magnetic), physical, chemical, and optical. The characteristic of each solution and its detection technique is evaluated with a set of properties crucial for oil health monitoring, diagnosis and prognosis.

It is still important to mention that the techniques can be combined to add value to the lubricant analysis by utilizing the strengths of each technique and reducing their weaknesses. This link between

techniques and methods generates more reliable information about the condition of the equipment, but an expert is always needed to interpret the results of each test. With this information, we can mention the study conducted by A. Kumar & Ghosh (2016), which makes a combination of analysis techniques, performing particle separation by means of ferrography with Ferrogram (Model: T2FM), image analysis by bichromatic microscope (Model:) MET-233) and Field Scanning Electron Microscopy (FESEM) (Model: ZEISS SUPRA 55 FESEM), and finally the study of the variation of properties in relation to the time of use were analyzed with Viscometer (Model: SVM 3000) and FTIR spectroscopy (Model: Perkin Elmer spectrum 2000). Sheng et al. (2012) develops his study based on the union of ferrographic analysis and spectrum analysis, thus demonstrating that ferrographic analysis can effectively monitor the wear condition of the engine, while spectrum analysis can provide additional information about the sources of abnormal wear.

The combination of several techniques, allows for a proper characterization of the degradation index. However, when combining techniques it is necessary to take into account the associated costs, with each technique influencing an increase in costs in analyzing each sample without an expected return of information. In this way, some companies are opting for simpler tests with reduced costs as can be seen through the analysis of the items in Table 7, such as of the use of ML techniques and intelligent models, among others that maintain a strong relationship with the use of data to determine the degradation index.

Finally, based on the literature review and, considering that this is an academic project with business connections, we decided to use a combination of techniques at a lower cost and with potential for advancement and generalization to other fleets and vehicles. In this way, we have chosen to use stain tests to acquire comparison information in conjunction with data analysis and ML techniques.

2.6 Vehicle Information Acquisition Methods

To reduce the problems associated with vehicle fleets, it is important to develop a predictive maintenance system based on data acquired in real time to determine the degradation index of lubricants. To this end, it is necessary to capture a wide range of variables, either from the behavior of the vehicles in operation or the lubricant itself and to perform a detailed analysis of the operation of the lubrication system and the behavior of the variables. In this section we describe the platform we will use for this purpose, presenting some concepts pertinent to this type of information acquisition.

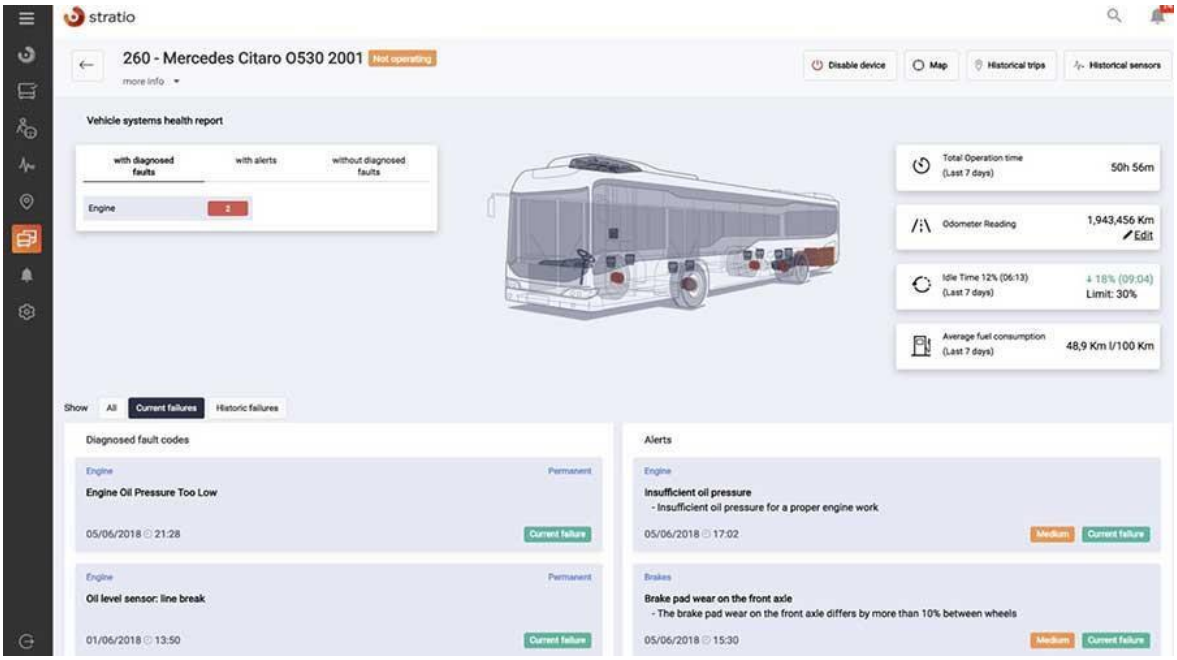

Figure 7 - Sending Data in Real Time on Stratio's Platform – Source: (Automotive, 2021)

Figure 7, presents an example of a dashboard of a platform, which can acquire real-time information from all systems of the vehicle, present information of potential failures and if necessary, the user can have access to more details of the behavior of the variables of each system. However even with the use of this type of platform it is necessary for the user to be an expert to determine the importance of this information and analyze it reliably, or to have someone to process the results that can be passed on to management. This type of platform often does not directly present results in an automated way that is easy for management to interpret. Thus, before starting the development of a system capable of providing results that are easy to interpret and in a dynamic way to determine the lubricant operating condition based on the information collected from the vehicles, it is necessary to present fundamental concepts for this type of acquisition and how these data relate to this project. These concepts are: Vehicle Communication Protocol, interface On Board Diagnostic (OBD), Parameter IDentification (PID), Diagnostic Trouble Code (DTC).

2.6.1 Vehicle Communication Protocol

According to Neumann et al. (2017) the increasing regulatory requirements and different technological advances determined the continuous development of automotive electronic systems. Improvements in the areas of safety, comfort, compatibility with environmental laws, legal requirements, and stricter guidelines have driven the integration of electronic systems with sensors connected to processing units, enabling connection to computers and communication by radio frequency signal. Based on these requirements, the individual systems of the vehicles had to be grouped in an integrated system, with the information being exchanged through the Controller Area Network (CAN) protocol between the Electronic Control Units (ECU). The standardization of the systems and

subsystems are an essential condition for the tuning between the modules. Because of this regulatory process, there was a reduction in development time, an increase in system reliability, and a reduction in redundant components, which were successively integrated to improve the applicability of information from other systems. The implementation of the integrated systems requires that all ECU's have a standard communication interface. This way the interaction with the subsystems will be defined according to standards, and all the variables involved will be efficiently controlled and analyzed, regardless of the manufacturer. According to Bozdal et al. (2020) large technological advances have also influenced the creation of devices that follow a trend of decentralization of applications, where smaller modules of a system such as sensing, processing and actuation modules can work separately. The technologies incorporated in vehicles have increased, with the electronic control modules available no longer being items of luxury automobiles, thus leading the industry to develop more efficient and versatile communication protocols.

SAE describes a classification of network communication in relation to application requirements:

- **A:** This is used for non-critical applications such as lamp and stereo control in a vehicle;
- **B:** Network used for important applications, but not essential to the proper functioning and safety of the vehicle. The A and B ratings are applied to the electronic body;
- **C:** Used in applications of maximum security, for distributed systems with information display in real time.

Taking into account this classification it is possible to see that the choice for a distributed solution presents a number of benefits compared to a centralized solution, such as: fewer wires in the system, making wiring simpler and cheaper; a flexible system; easy system maintenance; and the fact that each unit can be developed and tested individually without affecting the operation of the system as a whole. Currently, there are many industrial solutions available for data communication between devices. However, when choosing one, we need to pay attention to the transmission rate of each protocol, so that there is no data loss between units.

Protocol	Transmission Rate
SAE J1708/J1587	9600 bps
SAE J1939	33.3 kbps / 83.33 kbps
CAN 2.0/ISO 11898	10 kbps - 1 Mbps
k-line/ISO 14230	10.4 kbps

Table 8 – Example of Automotive Communication Protocols

Table 8, presents the differences in the transmission rate of each protocol and this directly influences the type of application and the amount of information made available (Neumann et al., 2017). Furthermore, companies usually use several communication protocols at the same time in the vehicle structure depending on the need of each system. According to Neumann et al. (2017) we can exemplify the network architecture of a truck and bus body that uses the SAE J1939 standard, with a transmission

rate of 250kbps for priority systems, and the ISO14230 for diagnostic functions. According to Axelsson et al. (2004), we can present another example of this kind of use of communication protocols from the architecture of a Volvo FH truck, which is composed of three functional groups: powertrain, passive safety and entertainment. The first group contains the main control and diagnostic modules. The second group, performs the functions of controlling the active safety systems (airbags and immobilizer) and climate control. Finally, the third group concentrates the connectivity functions.

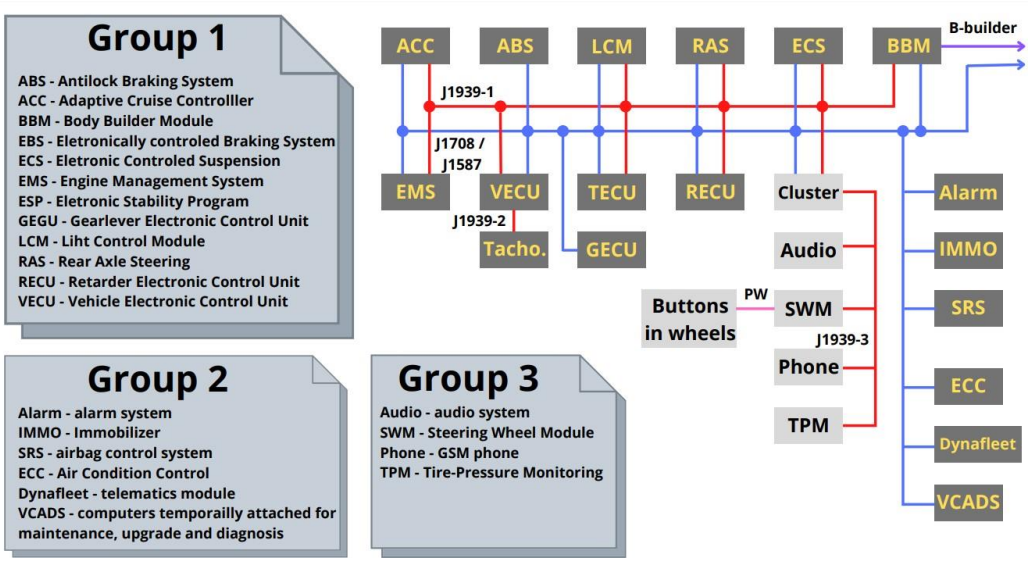

Figure 8 – Volvo Truck Network Architecture (Source: (Axelsson et al., 2004)*)*

Looking at Figure 8, we can see that there are three J1939 communication lines (J1939-1, J1939-2, J1939-3), one J1708/J1587 communication line and one PW communication line. This distribution of the ECUs, categorizes as a priority the basic vehicle operation and safety systems in networks with higher capacity and data transmission speed (250 kbps), leaving the systems of lesser importance to the J1708 network, with transmission rate of 9.6 kbps. This distribution makes it possible to: expand the network to structures outside the chassis; a lower volume of data traffic on the network; greater security in case of failure for the main systems and thus ensure a higher quality of transmission.

2.6.2 On Board Diagnostic (OBD)

According to T. Wang et al. (2021) during the 1970s and early 1980s, automobile manufacturers began to use electronic means to control vehicle functions. Following the advancement in these types of control devices and technologies soon manufacturers realized the potential of these devices to perform diagnostics, but they struggled with development of a tool that had such capabilities. Manufacturers started the developments of these diagnostic tools to meet mainly the standards for emissions of pollutants gases, by creating an electronic system which is incorporated into passenger cars, commercial and heavy vehicles with internal combustion engines to control and diagnose faults and failures, called On Board Diagnostic (OBD). T. Wang et al. (2021) states that with the evolution of technology and of the OBD system itself, associated with the dissemination and acceptance of the

interface by users and maintenance operators, the system has become more sophisticated and robust, able to offer greater control of the vehicle's operating variables, not only focused on emissions regulations, as it had been initially conceived, but also started to incorporate the control of all vehicle systems and diagnose potential failures. T. Wang et al. (2021) also indicates that in 1996 an evolution of the OBD was presented in the automotive field, called OBDII, based on a solid structure of standards developed by the SAE in accordance with the ISO. The evolution of the OBD system into OBDII was characterized by a new approach to detecting failures or poor performance of a particular vehicle component or system. Another objective covered by the OBDII technology is the minimization of the time between the occurrence of the malfunction in the operating variables of the vehicle, detection of the root problem of this malfunction and its final repair. With this, this system has become a fundamental part in the identification of anomalies, failures and errors in the components and subsystems of the vehicle, all of which are connected to the ECU. According to Pranjoto et al. (2018) the OBDII interface specifies the basis in data communication between vehicles and their control unit, as well as determines the pattern of the electrical signals and furthermore the physical pattern of the connector and location of the signals in it. The interface is determined as a 16-pin connector located inside the vehicle and should by standard be located within 2 meters of the steering wheel and have direct access without the need to use tools.

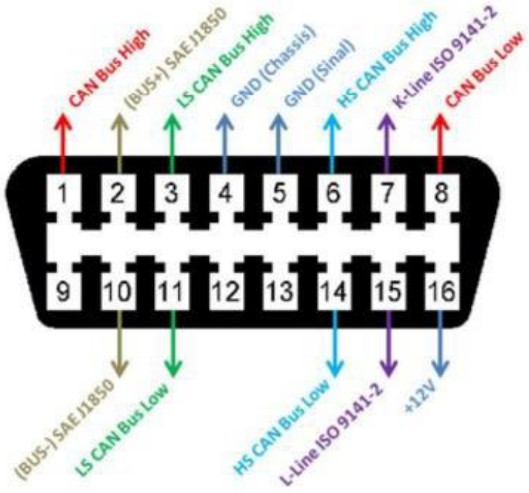

Figure 9 - On Board Diagnostic (OBDII) Pinout – Adapted from (Abbott-mccune & Shay, 2016)

Pranjoto et al. (2018) reports the physical specifications of the OBD-II connector (see Figure 9), its mechanical characteristics and its pins according to the standardization of SAEJ1962 standard. There are ten different operating modes of OBD-II, dependent on the SAEJ1962 that is used. The OBDII data engine can be visualized according to Table 9.

Mode (Hexadecimal)	Description
01	Requests current data from a given PID
02	Shows vehicle sensor data at the time of a diagnosed malfunction (Freeze frame data)

03	Requests diagnostic error codes
04	Clear error codes stored in the ECU
05	Request oxygen sensor test results
06	Request on-board test results from monitored systems
07	Request error codes detected in the current or previous drive cycle (impacting emissions)
08	Request on-board system, test or component control
09	Request vehicle information
0A	Request error codes impacting emissions

Table 9 - On Board Diagnostic (OBDII) Modes

It is important to mention that vehicle manufacturers are not obligated to support all the service modes expressed in Table 9. Each manufacturer can define additional services, for example: service 22 as defined by SAE J2190 for Ford/GM, service 21 for Toyota.

2.6.3 Parameter IDentification (PID)

According to Pranjoto et al. (2018), when it comes to PID's, the SAE J1979 standard defines a large amount of possible parameters, however not all vehicles will support all of the given ones and there may be custom parameters by manufacturers that are not defined by SAE. These parameters allow access to data such as vehicle speed, engine speed, fluid temperature, among other variables of the vehicle.

PID (Hexa)	Bytes	Description	Min Value	Max Value	Unit
5	1	Coolant Temp.	-40	215	°C
0C	2	Engine speed	0	16383.75	Rpm
0D	1	Vehicle speed	0	255	Km/h
11	1	throttle position	0	100	%

Table 10 - Example of Parameter Identification (PID's) According to SAE J1979

Table 10 shows examples of the PID's in the OBD-II standard as defined by SAE J1979. Pranjoto et al. (2018) states that the expected response for each PID is provided, along with information on how to translate the response into meaningful data.

2.6.4 Diagnostic Trouble Code (DTC)

According to Vigneswaran et al. (2021), early versions of OBD would simply illuminate a malfunction indicator light on the instrument panel called Malfunction Indicator Lamp (MIL) if a problem was detected, but would not provide any information about the nature of the problem. Modern OBD implementations use a standardized digital communication port to provide real-time data, in addition to a standardized series of diagnostic trouble codes, or DTC's, that allow the user to quickly identify and correct malfunctions within the vehicle. A fault in a component or subsystem that has a direct impact on the perfect operation of the vehicle causes a DTC fault code to be written to the ECU, which contains the description and location of the problem that was identified. This can be seen in Figure 10.

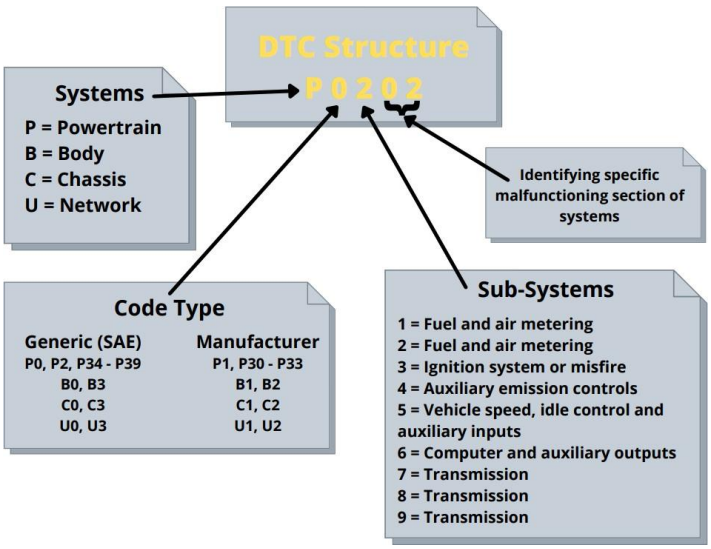

Figure 10 - Standard Diagnostic Trouble Code (DTC) Reading – Adapted from (Vrachkov & Todorov, 2018)

In this way, analyzing Figure 10 and according to Vigneswaran et al. (2021) the DTC's are composed of a letter, which indicates the system where the fault was diagnosed, followed by four numerical values according to the SAE J2012 standard, which indicate the type of code, the subsystem affected, the section and the type of malfunction or fault. In this way a car that has a DTC recorded in the ECU, according to Figure 10, "P0202", represents the diagnosis of a failure of the "Engine or transmission" system, which has a PID defined by ISO/SAE (Generic), in the subsystem or component determined by the manufacturer as being 2 and failure number 02 of this subsystem.

2.7 Intelligent Systems

Intelligent systems can be considered as implementations of the branch of science that study the set of paradigms that intend to justify how intelligent behavior can emerge from artificial implementations, in computers. An intelligent system incorporates intelligence into applications being handled by machines to support complex activities. Complex systems should not be confused with intelligent systems. For example, a manipulator robot that applies weld spots on a vehicle bodywork, despite performing a complex sequence of movements, having acute real-time operation and safety requirements, is not considered intelligent. This robot merely repeats a previously stored sequence of movements. This system lacks the ability to adapt to its environment. One of the characteristics of intelligent systems is precisely the ability to learn and to adapt itself to an unfamiliar environment or a new situation. According to Charissis et al. (2021), intelligent systems can be considered to be technological platforms that perform several functions that approximate the rational capacity of the human being to solve problems. The use of these systems is on the rise because large companies need to automate their processes. It aims to meet specific demands of the industry, promoting improved

performance, greater competitiveness, new business and thickening of the value chain from the introduction of new products and processes.

Since the main goal of this work is to research and develop a system capable of identifying the operating conditions of lubricating oils in diesel engines, we will introduce topics related to intelligent systems using Big Data and Data Mining concepts, as tools for analyzing large volumes of data, in addition to Machine Learning (ML) techniques and expert systems.

2.7.1 Big Data

When we talk about extracting knowledge from data it is necessary to specify how this information will be stored, since it should be available in a timely manner for further analysis for the acquisition of knowledge whenever necessary. The term Big Data (Mell & Grance, 2011) refers to a large volume of data that is stored in an unstructured way and needs to be analyzed in real time, which briefly represents the inability of traditional architectures to store, handle and analyze large amounts of information. The characteristics of the big data concept imply a new architecture known as the "Vs" of Big Data as presented in Figure 11.

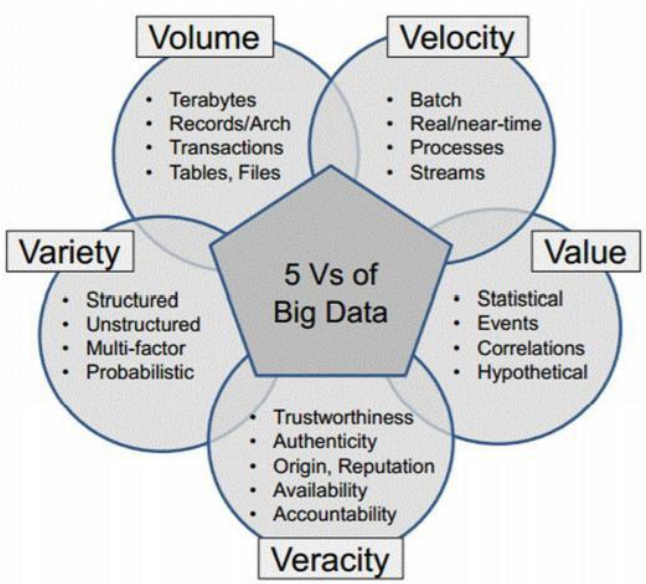

Figure 11 - "V's" of Big Data – Source: (Ricardo, 2014)

According to Figure 11, the "Vs" of Big Data are: Volume (size of the data set); Variety (data from multiple repositories, domains or types); Velocity (rate of data flow); Variability (Consistency in the data set) and Value (Data in itself is of no use, they need to be converted in something valuable to extract knowledge). According to Alvine et al. (2018), industries have recently shown great interest in the potential of Big Data, with many organizations announcing major plans and investments in research and applications. A key point in Big Data is in its relationship with the Internet of Things (IoT), in which its data sources come from sensors, which capture various types of information on various

devices/machines, such as vehicles in a fleet. According to M. Chen et al. (2014), by 2030 sensors will reach the one trillion mark and will be a fundamental part of the big data concept.

A large number of researches indicate that for the success of big data a strong connection with IoT and its effective integration with cloud computing is required (M. Chen et al., 2014). Another term widely used is analytics, which is related to the discovery, interpretation and communication of significant patterns in data. This analysis has added value in a wide range of areas, because that can be applied to the study of organizational data with the ability to forecast, improve, aid in decision making, risk analysis, among other applications (Chiang et al., 2018). According to Chiaburu (2016), in the organizational sphere of maintenance the implementation of this concept with the aid of mathematical models can demonstrate predictive task planning. Chiaburu (2016) further states, that scholars in the realm of smart data have experienced two eras of the use of analytics, these eras are referred to as: Before Big Data (BBD) and After Big Data (ABD). According to H. Chen et al. (2018), there is a timeline according to smart data, called: BI&A 1.0, the era of Business Intelligence (BI), which deals with understanding business phenomena by analyzing production processes, sales, and customer interactions, granting the manager greater decision-making power H. Chen et al. (2018). BI&A 2.0, the web era, is driven by companies such as Google, Yahoo, Amazon and eBay, which from the analysis of customer purchase profile collect a large volume of data to design or redesign products (H. Chen et al., 2018). BI&A 3.0, the mobile and sensor era, is based on the large volume of data generated by a large set of sensors and is driven by technologies such as Radio Frequency IDentification (RFID) and IoT (H. Chen et al., 2018). Considering this timeline, it is possible to indicate the direction of new technologies that drive the use of smart concepts, such as: smart cities, smart cars, among others (H. Chen et al., 2018). Already according to Dunn (2015), "big data" is the determination for a large volume of data that are arranged in diverse systems, with complex formats and modes of corelations. Dunn, (2015) states that the importance of a large volume of data for maintenance management comes from their storage and analysis to allow greater information gathering. With advances in IoT and other information technology systems, industry and maintenance services have the ability to store and analyze a more complete picture of asset integrity based on a more complete set of data drawn from a variety of sources.

2.7.2 Data Mining

The rapid increase in volume and diversity of data creates challenges in the long term storage and analysis. According to Shafi et al. (2018), Data Mining (DM), allows the user or developer to explore data to extract real value from it. According to Dogan & Birant (2021), the growth of data generates the need for the development and implementation of new tools and techniques capable of extracting knowledge in an automatic and intelligent way. All the knowledge extracted is of great value for managers in making critical decisions for the organization. However, the extraction of knowledge

from this large volume of data coming from various devices is not a task easily performed using conventional database management systems.

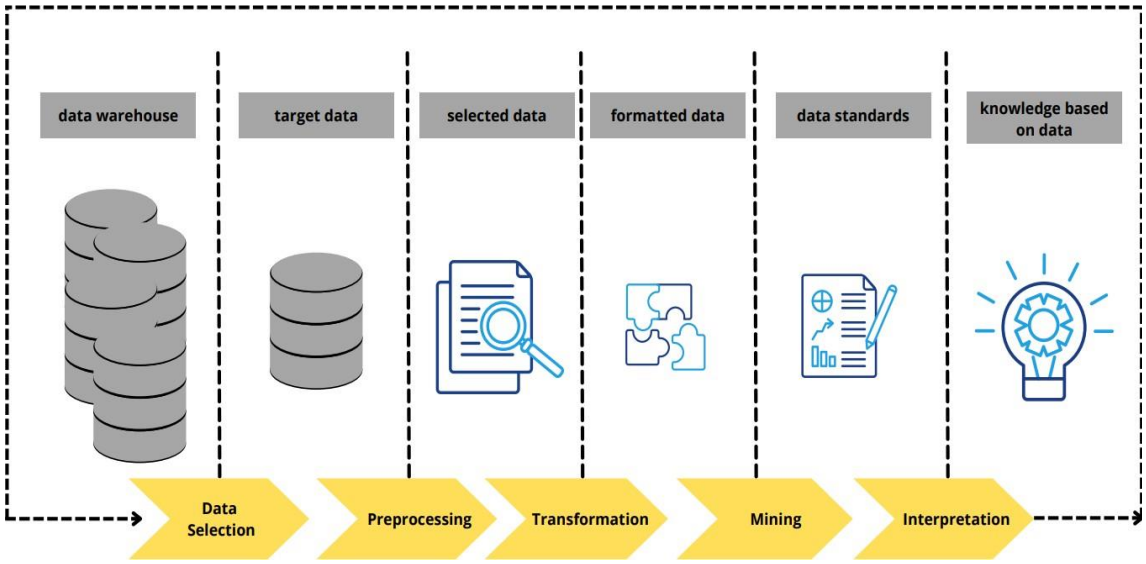

Figure 12 - Data Mining Flow proposed for two class problem divided by processes and milestones for knowledge extraction – Adapted from (Shadroo & Rahmani, 2018)

Figure 12 presents the proper flow for knowledge extraction. Shafi et al. (2018) identify the important technologies and tools for the data mining process. Shafi et al. (2018) states that cloud databases, artificial intelligence engines, statistics, pattern recognition, knowledge-based systems, information retrieval, high performance computing and data visualization are some of the tools needed for the procedure of knowledge extraction from data. Konar et al. (2015) explains that DM is a methodology that was born from the interaction and exchange of information from three major areas: classical statistics, artificial intelligence and ML. Still according to Konar et al. (2015) Data Mining can be considered part of a larger process called Knowledge Discovery in Databases (KDD), which allows the retrieval of initially unknown and potentially useful knowledge from a database. Dogan & Birant (2021) emphasizes that data mining is the non-trivial process of identifying valid patterns in data until then unknown, but potentially useful. According to Systems et al. (2021), in DM, the information from a database is processed in order to ensure the identification of factors and trends that are important in recognizing patterns in business activities. This type of procedure is used by managers to help them make decisions about strategic operational and business changes to obtain competitive advantages in the market. According to Dogan & Birant (2021), DM technology is structured from a set of tools, which through the use of ML algorithms or statistical models, are able to process a large volume of data, to extract knowledge in the form of hypotheses and rules. Thus, DM can be seen as a method that seeks a logical or mathematical description of the data, which can sometimes be complex in nature. Shadroo & Rahmani (2018) reports that hypothesis formulation is the basis for traditional data analysis, and that this is accomplished through DM techniques. Another important point is that relationships between data are not assumed to be known a priori, and so a variety of techniques are used to explore

and bring to light complex relationships in a large data set by exploring interrelationships. According to Shadroo & Rahmani (2018) and Mikut & Reischl (2011), Table 11 contains information about commercial data mining software, as well as some of their features:

Name	Manufacturer	Functions	Featured
Intelligent Miner	IBM	Algorithms for association rules, classification, regression, sequential patterns and clustering.	Integrated with IBM DB2 DBMS great scalability of algorithms
MineSet	Silicon Graphics Inc.	Algorithms for association rules, classification and statistical analysis.	A robust set of advanced visualization tools
Clementine	Integral Solutions Ltd.	Algorithms for induction rules, neural networks, classification and visualization tools	Object-oriented interface
DBMiner	DBMiner Technology Inc.	Algorithms for association rules, classification and clustering.	Data Mining using OLAP
Genamics Expression	Genanics Developer	Sequential analysis algorithms	Protein and DNA sequence analysis

Table 11 - Example of Commercial Data Mining Software

Mikut & Reischl (2011), refers that data mining technology is based on several techniques from different areas, like:

- ML algorithms, which are used to extract knowledge from a database, these algorithms are able to find examples of high-level rules, which are understandable to humans (Mikut & Reischl, 2011);
- Data Warehousing, a structure for managing data repositories from various sources, with the objective of having detail of part or all of a business. The final product obtained from a Data Warehousing project is the grouping of data from various sources of an organization for it efficient management (Mikut & Reischl, 2011);
- Statistical Methods, used to study the collection, organization and interpretation of numerical data, especially for the analysis of population characteristics, through inferences from sample groups. The statistical techniques have fundamental importance in the Data Mining process, because most of the methods used had their origin within the Statistics area (Mikut & Reischl, 2011);
- Data Visualization techniques and tools, which are indispensable tools in the Data Mining process. These techniques improve the understanding of the results obtained and the communication between users during the execution of the knowledge extraction steps. Data Visualization stimulates human perception and intelligence, aiming to increase the capacity of understanding and association of new patterns related to data (Mikut & Reischl, 2011).

When analysing these two studies, Shadroo & Rahmani (2018) and Mikut & Reischl (2011), it is perceptible that commercial software adapts well with the techniques needed for the development of DM. Thus, the techniques used in DM go beyond a simple analysis, generating new information that

in the future may be part of the common knowledge of an organization and be applied in Information Systems to support decision making. For example, the use of visualization tools (trees, rules, graphs), combined with other techniques can greatly improve the process of understanding and recognizing patterns in DM. Dogan & Birant (2021) states that the basic function of DM is to find knowledge from the large volumes of information stored in the databases of organizations, to allow agility in decision making. In this way, DM is an indispensable technology in current organizations, capable of selecting relevant data, learning from it, extracting deductions, generating information with hypothesis, correlating apparently unrelated facts, making predictions, revealing the important attributes, generating scenarios, reporting and discovering interesting knowledge to the company's managers, and assisting in a fast and automatic way in decision making. DM can create correlations, for example, between people and products or services, and thus predict costumers' habits or behaviours. DM can be used for prediction, identification, classification and optimization, allowing the analysis of association rules; classification and prediction; analysis of sequential patterns; cluster analysis or analysis of exceptions. According to Konar et al. (2015), it is necessary to highlight that each data mining technique adapts better to certain problems than to others, which determines the inexistence of a standard technique or a universal method. Each problem has its own characteristics and the success of a task is directly linked to the developer's experience and intuition. Furthermore, despite being a powerful and profitable tool, it poses challenges with respect to privacy protection.

2.7.3 Machine Learning

According to Jayabharathi & Ilango (2021), in 1959, Arthur Samuel defined the term ML, as being the field of study that allows computers to have a certain level of learning without necessarily having to have prior programming. For Mitchell (1997), the term ML is defined as a set of techniques that allow improvement in the performance of a given complex task, using previous experiences and information. ML can be divided into 3 fundamental areas, depending on the way we use the data to learn: supervised, unsupervised and reinforcement learning (see Figure 13). In supervised learning we have the data organized in a structured way, where for each input we have the desired output. In unsupervised learning, we have no prior information and the goal is to establish the best data structure autonomously, based on a dynamic interaction. Reinforcement learning, is used when the objective is that systems learn from acquired experiences. In these cases, when humans program the algorithm, they define which result is expected without indicating the best way to achieve it. Thus, the machine is responsible for figuring out how to achieve its goal. The algorithm runs a series of experiments in which it gets errors and successes, being rewarded for what went right and being penalised for the actions that led to failure.

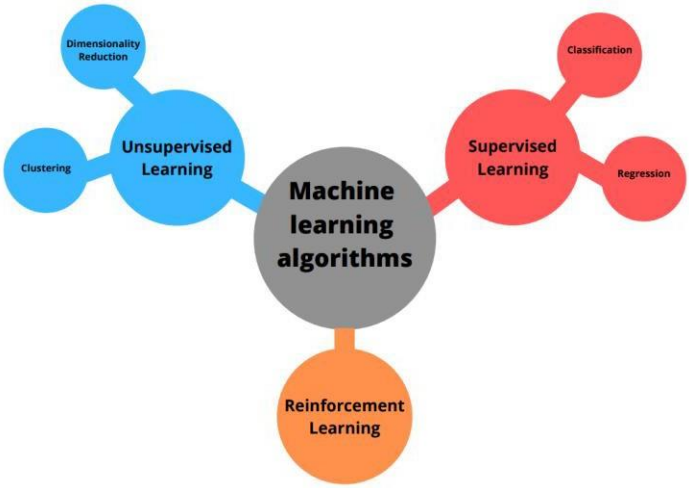

Figure 13 - Machine Learning Algorithm Models – Adapted from (Mitchell, 1997)

It is important to mention that all groups of algorithms, whether supervised or unsupervised or reinforcement, must have an ultimate goal, which is proposed by the developer at the beginning of the intelligent system development. We can exemplify this goal as being the prediction of a failure in a vehicle or the operating condition of a certain vehicular system. Figure 13 shows that supervised learning can be subdivided into two types: classification and regression. Mitchell (1997) refers that there are algorithms that are able to perform time series analysis, with the goal of learning a function, mapping initial facts for determination and/or extrapolation to a real value. Several algorithms exist for this type of problem: linear regression, regression trees and artificial neural networks are given as some examples. Mitchell (1997) finally informs about algorithms that have the goal of learning a function based on the description of an initial input: Bayesian networks, logistic regression and K-Nearest Neighbors (K-NN), are some given examples.

According to Prytz (2014), ML can be broadly defined as computational methods using experience to improve performance or to make accurate predictions. In this case experience refers to information previously available to the learning system, which usually takes the form of electronic data collected and made available for analysis. This data can be in the form of digitized human training sets or other types of information obtained via interaction with the environment. Thus, it is important to mention Expert Systems developed to simplify and automate the work of the technician responsible for a given equipment, allowing to analyze and report its operation to track a robust and diverse amount of information.

2.7.3.1 Expert Systems

According to N. Kumar & Parrek (2013), Expert Systems (ES) were proposed in the late 1980s and early 1990s, stemming from systems that were based on a knowledge structure determined by an individual, who was considered as the expert. The knowledge extracted from the experts according to N. Kumar & Parrek (2013), was used by inference mechanisms or rules, answering questions without

any interaction with other systems. Over the years the concept of expert systems has evolved. For example:

- According to Feigenbaum (1979), ES are computer programs that used knowledge and a rules model to solve complex problems that required human expertise and judgment;
- According to Beynon-Davies (1991), ES are computer systems that made use of human knowledge in a particular scope to perform tasks with characteristics similar to a human expert.

Taking into account these two definitions, one can see that they share some similarities. They both agree that ES are systems programmed with computational tools and that are developed from the knowledge of one or more Human Experts (HE) to develop tasks or answer complex problems. With the dissemination of distributed platforms and administrative management models, N. Kumar & Parrek (2013) affirms that the ES's have evolved for complex problem solving with the ability to learn by not being restricted only to small-scale applications and consequently incorporated into the so-called Knowledge Based Systems (KBS) that are based on the implementation of standardized processes of an organization with the help of the incorporated knowledge of a human specialist. In this way, KBS have their concept defined as being systems with problem solving capabilities using specific knowledge according to the scope of the application, generally inserted in the management and decision making process of a system or organization. These systems collect data and manipulate various aspects of knowledge, with the fundamental requirement being information about the skill, experience and heuristics used by the expert in problem solving. With a high degree of complexity, the development of these systems requires a deep interaction between the Knowledge Engineer (KE) that will model and/or develop the system (N. Kumar & Parrek, 2013). According to Rezende (2003), KBS can be identified as ES at the time they are developed, whose functioning is determined in an isolated manner from other systems, with the objective of being implemented in applications in which the knowledge to be manipulated focuses on a specific scope and is supported by a high degree of specialization and heuristic knowledge.

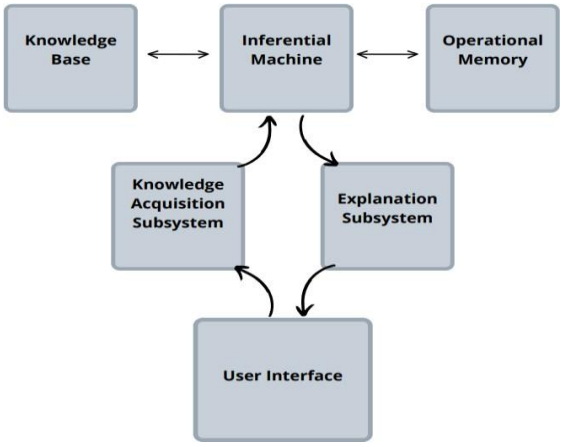

Figure 14 - Expert System Structure - Adapted from (Rezende, 2003)

Figure 14 summarizes the characteristics of ES in the context of Intelligent Systems (IS). ES must have their knowledge stored and organized in a way that makes it easy for a non-expert user to query solutions. This type of system is extremely useful for training and teaching, since a large part of an expert's knowledge is stored and organized to solve problems as efficiently as possible. The database of all this knowledge according to Charissis et al. (2021) is formed by a set of rules and procedures that the HE uses to solve problems, which are modelled in the system by the Coding Expert (CE), who implements it in a practical way according to the chosen representation, and the knowledge base provides the functional characteristics of the system. According to Alvine et al. (2018), the operational memory, which functions as a short-term or volatile memory, has the function of storing facts determining the problem during solution processing. The data regarding the facts can be obtained from sensors, keyboard responses, databases, or other programs. The inference machine or rule engine, according to Alvine et al. (2018), works by comparing the data stored in operational memory with the knowledge stored in the knowledge base. The inference machine is the component of ES's that processes reasoning and logical planning. Its processing works as follows Charissis et al. (2021): At the point where the knowledge base is structured from rules, the inference machine determines which conditional of the rule satisfies the facts held in operational memory and adds the conclusion of this rule to operational memory. There are two ways of implementing inference: Forward Chaining, which begins with evidence to reach a conclusion; and Backward Chaining, which begins with a conclusion and searches for evidence to prove it. The knowledge acquisition subsystem has the function of introducing or removing knowledge from the knowledge base (Charissis et al., 2021). The explanation subsystem is used to explain to the user the line of reasoning that the SE used to determine the resolution of the problem. This characteristic is of great importance because it allows the user to access additional information in the system, besides enabling the system for educational purposes (Charissis et al., 2021). The user interface establishes a means of communication between the user and the system (Charissis et al., 2021).

KBS's possess the ability to: Question the user, using an easy language with the focus on obtaining the information he needs; Develop a line of reasoning from the information obtained and the knowledge extracted to determine appropriate solutions; and finally, explain logical reasoning at the moment of questioning the user on how he reached his conclusions. For this explanation the system must be able to memorize the rules performed in the processing and present them in an understandable way to the user. According to Rezende (2003), the fundamental differences between a conventional system and a KBS can be highlighted considering seven points (see Figure 15): 1: Organization, 2: How it incorporates Knowledge, 3: Execution Techniques, 4: Form of Control, 5: Modification process, 6: Processed Information, 7: Outputs.

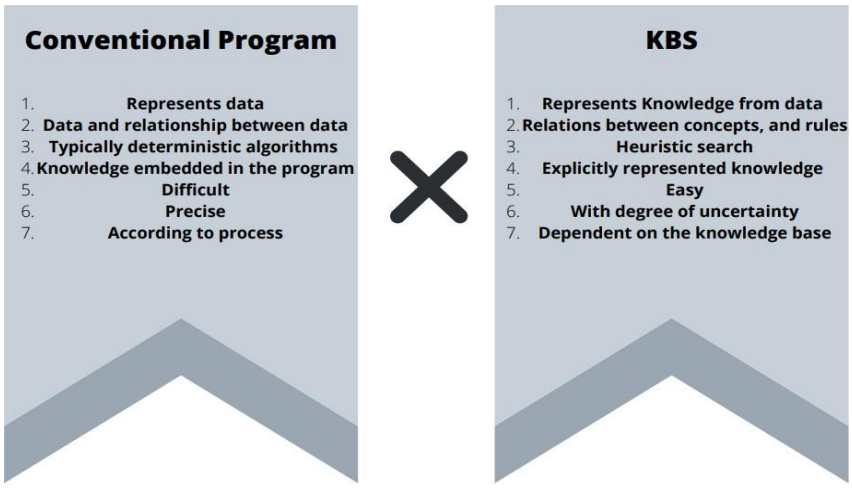

Figure 15 - Conventional Systems VS Knowledge Based Systems KBS's – Adapted from (Beynon-Davies, 1991)

Figure 15, allows to conclude that the most notorious difference between a conventional systems and a KBSs is in the aggregate power of knowledge and in the system's capacity to learn and to be modified as it is being used.

2.7.3.2 Model Evaluation Metrics

After explaining some important concepts in the area of intelligent model development for solving real problems, it is necessary to specify how we can evaluate the results of these models and systems. According to Keartland & Van Zyl (2020), in classification and failure prediction problems, it is possible to find numerous metrics to evaluate the model, however, it is of utmost importance to know which ones and when to use them. Before mentioning the evaluation metrics commonly used for the type of problem we are proposing to solve (prediction of the operating condition of diesel engine lubricating oils), it is necessary to explain a basic concept for the calculation of each metric. This concept is based on the confusion matrix, which shows the classification frequencies for each class of the system, allowing the extraction of metrics that help in it evaluation. This matrix is exemplified in Table 12:

		Predicted	
		Positive	**Negative**
Real	**Positive**	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Table 12 - Confusion Matrix shows the classification frequencies for each class

According to Hanafy & Ming (2021), a confusion matrix is a table that indicates the errors and hits of a given model, compared to the expected result (or labels/labels). Table 12, presents important concepts:

- True Positive (TP) - correct classification of the class Positive;

- False Negative (FN) - error in which the model predicted Negative class when the actual value was Positive class;
- False Positive (FP) - error where the model predicted class Positive when the actual value was class Negative;
- True Negative (TN) - correct classification of the Negative class.

It is important to mention that Negative class does not necessarly means failure and Positive class no failure. The determination of these classes depends on how the problem was structured. That is, in our problem we are looking for a prediction of the lubricant's operating condition, and in this case Positive class means that it is a failure and Negative class means that there is no failure and the lubricant is in operating condition. With the values obtained in this matrix it is possible calculate three evaluation metrics commonly used to evaluate intelligent systems.

$$Recall = \left(\frac{TP}{TP + FN}\right) \tag{4}$$

Where Recall score is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive).

$$Precision = \left(\frac{TP}{TP + FP}\right) \tag{5}$$

Where Precision score is the number of correct positive results divided by the number of positive results predicted by the classifier.

$$F1\ Score = 2\ x\ \left(\frac{Precision\ x\ Recall}{Precision + Recall}\right) \tag{6}$$

Where F1-score is the Harmonic Mean between precision and recall. This evaluation metric indicates how precise the classifier is (how many instances it classifies correctly), as well as how robust it is (it does not miss a significant number of instances). High precision but low recall, means that the model is extremely accurate, but it misses a large number of instances that are difficult to classify. The greater the F1 Score, the better is the performance of the model

By analyzing equations 4, 5 and 6, we can determine that precision can be used in a situation where FP are considered more detrimental than FN. For example, when classifying an action as an optimal time to change the lubricant in a given equipment, the model must be correct, even if it ends up classifying good times as bad times (a FN situation) in the process. In other words, the model must be precise in its classifications, because if we consider a good moment when in fact it is not, this leads to loss of operating time and potentially profits for the company. Recall can be used in a situation where FN are considered more damaging than FP. For example, the model should find all situations where the

lubricant is not in operating condition anyway, even if it classifies some situations that determine the operating condition as not being a condition (FP situation) in the process. In other words, the model must have high recall, because classifying the non-operating condition as being healthy for the equipment to operate with the lubricant oil can lead to catastrophic losses. F1-Score is a way to look at only 1 metric instead of two (precision and recall) in some situation. In other words, when you have a low F1-Score, it is an indication that either the precision or the recall is low.

Finally, and not less important we can use Receiver Operating Characteristic (ROC) curve, which is one of the popular metrics used in the industry to binary classification problems. This curve is exemplified in Figure 16.

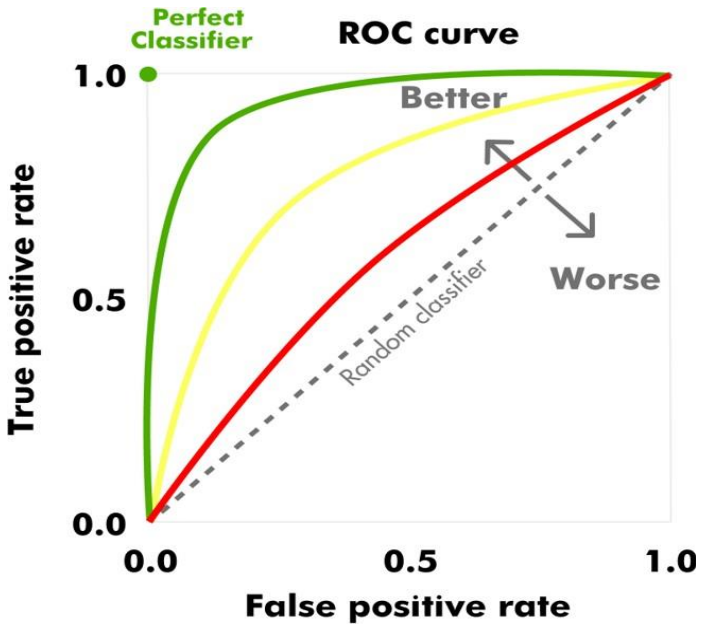

Figure 16 - Receiver Operating Characteristic (ROC) curve, is a chart that show the ability of a binary classifier system

ROC curves typically feature True Positive (TP) Rate on the Y axis, and False Positive (FP) Rate on the X axis. This means that the top left corner of the plot is the "ideal" point indicating a FP rate of zero, and a TP rate of one. The "steepness" of ROC curves is also important, since it is ideal to maximize the TP rate while minimizing the FP rate. Using this curve we can compute the Area Under Curve (AUC), that show the probability of the classifier to rank a randomly chosen positive example higher than a randomly chosen negative example. This value of the AUC varies between 0 and 1, it does mean that a larger area under the curve is usually better and is close to the value 1.

Chapter 3 - Fleet Maintenance Case Study

After discussing the background and concepts for the development of this work we will present an overview of the approach, and the interconnection of the concepts mentioned in the previous chapter. At this point, it is important to refer that the author of this work is part of a company (Stratio) team responsible for the development of a device to collect real-time information from vehicles to predict failures and reduce maintenance costs by creating predictive models on a proprietary platform. It can be said that by collecting repair data, DTC's stored in the vehicle control unit and real-time data on the fleet's operational profile, one can realize that a catastrophic failure does not occur immediately, but rather it occurs over a certain period of time with anomalous indications in variables and data that can be analysed to pre-emptively determine a potential minor failure that can be addressed early at a reduced cost. For example, an engine failure that may have a high maintenance cost for a fleet manager can be detected by analyzing lubricant variables and data, and consequently can be tackled by changing the lubricant and/or reducing oil change intervals. This type of minor intervention has a reduced cost and is less damaging to the vehicle than a catastrophic engine failure occurring due to lack of inefficient lubrication.

3.1 Stratio

Stratio specializes in Artificial Intelligence and Automotive Engineering, with a mission to advance ML in automotive (Automotive, 2021). Stratio envisions creating a company that pushed the boundaries of predictive intelligence transportation and mobility going beyond existing methodologies, and help automotive manufacturers to accelerate towards a zero-downtime future, enabling early fault detection and insights into vehicles health, systems condition and performance.

Figure 17 - Stratio Data Acquisition Hardware – Source (Automotive, 2021)

According to technological developments in the maintenance sector, one can highlight the author's presence in this scope of development, since he currently develops activities as an automotive engineer in the research department of Stratio Automotive. As can be seen in Figure 17, Stratio

developed a heavy-duty vehicle data acquisition device that connects to the vehicle's OBD port, to acquire real-time data on various aspects of the vehicles and operators using PID's and DTC's as a source of information. The information is sent wirelessly, via mobile network and is stored in a database. Then, it is processed analyzed, and presented to customers on a proprietary online platform. It is important to mention that the author's insertion in this company not only allows access to real-time information from the vehicles, through the use of the acquisition device, but also opens the opportunity to access the information that is not stored in the vehicle's control unit memory, stored in physical format at the clients Stratio servers. Even though this information is not used directly in the development of the prediction system, it serves as guide for the development of many of the fault prediction functionalities.

The Artificial Intelligence algorithms developed by Stratio analyze thousands of pieces of data extracted from sensors in real time, which help engineers identify important variables for the development of methods to predict vehicle failures and performance issues. The technology enables the empowerment and acceleration of root cause failure analysis with critical insight, saving time and resources; it allows manufacturers to add value to their cars by retrieving intelligence from the data, preventing recurring failures, reducing warranty costs, and benefit by proactively selling maintenance and repair services. Stratio technology works in three main stages, Acquisition, Analysis and Alerts:

- **Acquisition -** After Stratio Databox, i-e., the acquisition device is installed in the vehicle, fault codes and data from engine systems, batteries, transmission, among others, are collected and transmitted in real time to the server. The Stratio platform is also compatible with pre-existing data acquisition systems as long as these have the same density and granularity of the required parameters and the transmission of this data is performed in real time;
- **Analysis -** The received data is continuously analyzed by proprietary Machine Learning algorithms according to rules developed by automotive engineers. The information system constantly "learns" from the newly acquired data, adapting and refining the results obtained for each new malfunction detected. The malfunctions found during real-time diagnostics are also processed the moment they become active;
- **Alerts -** Notifications are configured as fault and anomaly alerts, displayed on the website, without the need for an external application. Additionally, the alerts are configured to be sent, by email and Short Message Service (SMS), according to the client's needs. The alerts are associated with the values registered in the sensors and systems associated with the presented failure, for a more precise analysis of the root cause of the problem.

3.2 Common Vehicle Failures

There are many types of failures in automotive systems. To create an intelligent vehicle maintenance model with high relevance for a fleet manager and with a reliable failure prediction response, it is first necessary to divide the vehicle structure into several systems and subsystems, to reduce the level of complexity. To catalogue the common failures of an automotive system and to proactively tackle the problems with the highest number of occurrences, which determine high maintenance costs for fleet managers, some authors such as Nahim et al. (2016) present a literature review on failures occurring more frequently in diesel engine vehicles. The characterization of these common failures is determined through the collection of repair data from various fleets, with the objective of preparing a database on diesel engine failures to help researchers develop accurate diagnostic and forecasting strategies. Another important point in determining and using this failure database is to assist in the development of simulators of the behaviour of the vehicles in the presence of failures. With this information and through the study of common faults determined by Nahim et al. (2016), we started an investigation about the most common failures in the fleet that will serve as a source of data for the system that will be developed. The results are summarised in Table 13

Occurrences of faults in diesel engine vehicles		
Class of defect	**Mean Occurrence (%)**	**System**
Fuel-injection equipment and fuel supply - (Collacott, 1977) and (Diesel, 1993)	26.05	Engine
Water leakages - (Collacott, 1977) and (Diesel, 1993)	15.20	Engine
Valves and seating - (Collacott, 1977) and (Diesel, 1993)	14.65	Body
Bearings - (Collacott, 1977)	3.50	Suspension and Steering
Piston assemblies - (Collacott, 1977)	3.30	Engine
Oil leakages and lubrication systems - (Collacott, 1977) and (Diesel, 1993)	5.75	Engine
Turbochargers - (Collacott, 1977) and (Diesel, 1993)	3.75	Engine
Gearing and drives - (Collacott, 1977)	1.95	Gearbox
Governor gear - (Collacott, 1977)	1.95	Gearbox
Fuel leakages - (Collacott, 1977)	1.75	Engine
Gas leakages - (Collacott, 1977) and (Diesel, 1993)	3.75	Engine
Breakages and fractures, other than mentioned - (Collacott, 1977)	1.25	Body
Miscellaneous - (Collacott, 1977) and (Diesel, 1993)	13.15	Body
Foundations - (Collacott, 1977)	0.45	Body
Crankshafts - (Collacott, 1977)	0.10	Engine
Electrical - (Diesel, 1993)	3.45	Electrical

Table 13 – Percentage of Common Breakdowns on Diesel Engine Vehicles

In concrete we present the average percentage of the most common faults in a vehicle from two references presented in Nahim et al. (2016) and add a column (System) that determines the group or system in which these faults appear. It is important to mention that this division into systems was

performed taking into account our domain expertise, as well as through document analysis of repair data obtained from Stratio customers. This type of screening is important to give engineers an idea about the systems and components that should be prioritized when creating intelligent predictive models. Through the analysis of Table 13 it is possible to see in parentheses in the column "Class of defect" the types of faults that are present in Collacott (1977) and Diesel (1993). They show us the commonalities and disagreements between the references collected by Nahim et al. (2016). 82.3% of referred faults are common to both references. 54.5% of these faults occurs in the system called "Engine". It is important to mention that the other system that has common faults in both references is the "Body" system and represents 24.7% of the faults. In spite of representing a significant amount of the total faults, the "Body" system will be excluded from our analysis because it does not have a great detail of information that can be collected in real time through sensors, since it is directly related to non- predictable situations such as traffic accidents, rear-view mirror breakage, and automotive aesthetics. To complement the analysis of the data presented in Nahim et al. (2016), we analysed Stratio database to identify the failures occurred on the fleet of one of its clients. We only looked for failures occurred in the engine system. The failure analysed were classified and their rate of occurrence determined. The results of this analysis are presented in Figure 18.

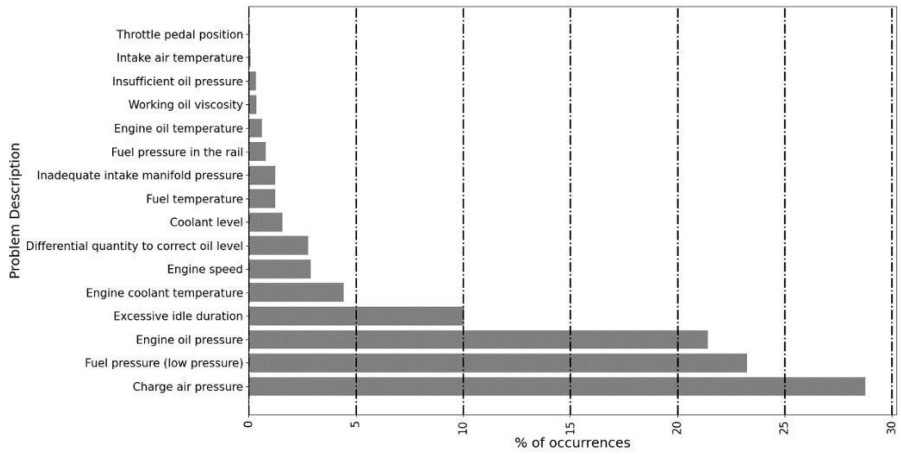

Figure 18 – Faults Histogram

By analyzing Figure 18, it is possible to determine the most common faults found in the repair data collected from the customer's engine system. Another important piece of information extracted from the analysis of Figure 18 is that there are several problems that are correlated among themselves and have a common element, which is the lubricant. In fact, it can be seen that lubrication-related problems have a high percentage of occurrence (25.53%):

1. Engine oil pressure – 21.42%;
2. Differential oil quantity to correct oil level – 2.78%;
3. Engine oil temperature – 0.63%;
4. Working oil viscosity – 0.37%;
5. Insufficient oil pressure – 0.33%.

All this data supports the choice made for this work, which consist in concentrating the analysis to the engine system and the lubrication problem.

To build the model intended for failure prediction, it is important to understand which variables will be available. To do so, we analyze the vehicle systems considering the parameters that can be acquired in real time. The parameters chosen for the model must be transversal to the fleet (independently of the individual age and technology used in each vehicle of the fleet), to have a failure prediction model relevant to the maintenance manager. The results of this analysis is presented in in Figure 19.

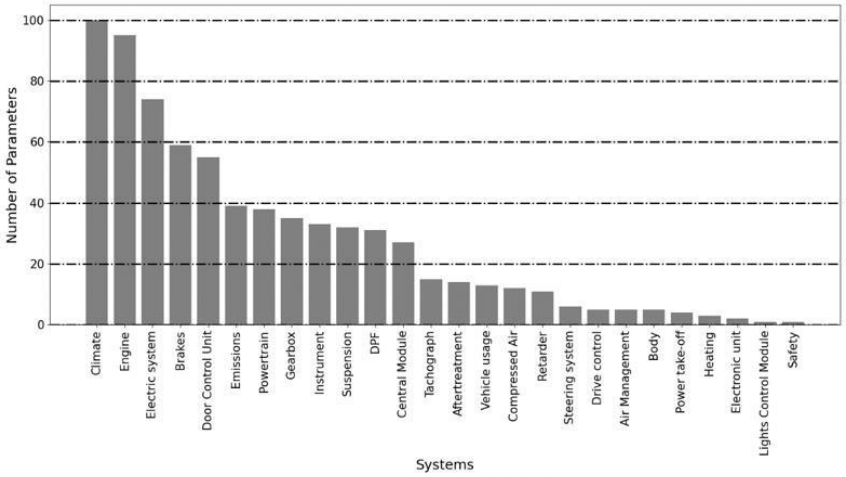

Figure 19 - Parameters Available by System

Figure 19 presents the parameters that can be collected from the communication bus of the fleet considered in this project, grouped by vehicle system. The most important conclusion we can draw from Figure 19, is that there is a considerable number of available parameters, related to the engine, the system we intend to study.

Finally, to confirm the importance of considering the engine system and, particularly, problems related to lubrication problems, we analyze the DTC's and alerts originated in real time, by the vehicle fleet to be considered in this project. The results of this analyzis are presented in Figure 20.

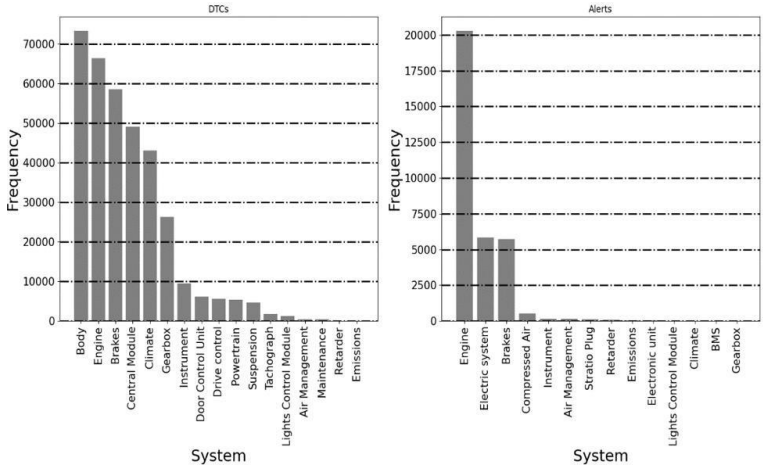

Figure 20 - DTC's and Alerts - Frequency Histogram

Figure 20 confirms that the engine is a critical system of the vehicle, presenting the largest number of alerts and having the second higher number of DTC's.

Considering all the information presented in Table 12 and Figure 18, Figure 19 and Figure 20 we can conclude that the decision to develop an intelligent system to predict faults in the engine system due to lubrication problems is adequate. In fact, it is not possible to identify an isolated engine problem to cope with the intended intelligent system. Nevertheless, it is possible to conclude that a large number of engine problems are related to lubrication anomalies.

With all the previous information in mind and considering the author practical experience, it was decided to study the problems involved with the lubrication system, since its structure is of fundamental importance to all vehicle systems, and we can have easy access to real-time testing and data with better quality and lower cost when compared to turbo failures. In other words, we can develop an intelligent model to predict failures in the lubrication system, which is relevant, considering it impact in the failures as a whole and from which we have a large number of parameters needed to create the intended tool to determine the operational condition of the asset. Furthermore, this option has a reduced cost in confirmation tests when compared to other identified problems.

3.3 Interconnection of Subjects

After analyzing common failures and having a clear vision of the system and the problem that we will tackle, we can establish a connection between the subjects presented in the literature review. We start with the CBM concept, which states that maintenance should only be performed when certain indicators show signs of decreasing performance or future failure. It is important to mention that checking these indicators on a vehicle can include non-invasive measurements, such as:

- **Vibration analysis:** Rotating equipment such as compressors, pumps, and motors all exhibit some degree of vibration. As they degrade or become misaligned, the amount of vibration increases. Vibration sensors can be used to detect when they becomes excessive;
- **Infrared:** Infrared cameras can be used to detect high temperature conditions in energized equipment;
- **Ultrasonic:** Can be used to detect deep subsurface defects, such as hull corrosion;
- **Acoustic:** Used to detect gas, liquid or vacuum leaks;
- **Lubricant Analysis:** measures the number, size, shape and/or presence of particles in a sample to determine equipment wear;
- **Electrical:** Motor current readings using clamp meters;
- **Operational Performance:** Sensors throughout the system allow to measure pressure, temperature, flow, among other parameters.

By checking these non-intrusive methods, we can see the presence of lubricant analysis which is present as a common element of several problems highlighted in our failure analysis. Recent developments in computer and transducer technologies, signal processing and real-time data acquisition have made it possible to implement CBM more effectively. We can cite the work developed by Cabral (2013), which states that information and monitoring technologies are still widely used in transport fleet management only to monitor driver behavior and whether the route is being executed according to plan. But through these monitoring and data collection techniques from the vehicle's built-in sensors in real time, it is possible to determine the driving condition safely and reliably, i.e., maximum speed, hard braking, engine speed, consumption as well as the operation and maintenance condition (failure prediction) (He et al., 2014). This condition of driving can be performed using Global Positioning System (GPS) transmitters and an on-board computer, which records the vehicle's route as well as the data necessary for efficient management of this type of service. Going beyond just monitoring operational information, the use of new IoT technologies, cloud data storage, big data, allows the manager the ability to manage maintenance, stock control, breakdown forecasting, determining equipment degradation rates, among other organizational capabilities. Another important point that can be determined from the author's insertion in the scope of maintenance and innovation of management methods is determined according to Oliveira et al. (2013), which states that the knowledge acquired through the literature deals with the management of maintenance in vehicle fleets concentrated in the survey of maintenance costs, repair estimates supported only by the monitoring of occurrences. Prytz (2014), states that, although there is the formation of a database of failures, maintenance occurrences and costs involved, there are cases in which scholars establish few standards that support the decision making of the manager. Thus, we can highlight some studies that present maintenance decision guidelines to improve CBM, using detailed analysis of used oil data as a source of information to determine whether the system fluid is healthy and suitable for other services or ready for a change, in addition to observing the condition of the equipment through these analyses (Muchiri et al., 2018). However, it should be considered that mechanical triboparts wear is a dynamic and gradual process that goes through different phases from normal to failure. Even with all the technological advances it is necessary to pay attention to the concepts already used in maintenance, such as the use of the PF curve, which shows how a typical wear process (S. Wang et al, 2017). Thus, it is apparent that predictive maintenance represents the best form of maintenance to be adopted by companies in the automotive segment and that the projections of the useful life of some vehicle systems, such as the lubrication system, can be better defined with the support of new data analysis tools. Oliveira et al (2013) reports that a good integrated maintenance system should have features and functionalities that ensure that vehicles are kept running as long as possible, avoid failures and reduce the use of trailers. Beauvir (2017), reports that analysis of the lubricating oil used for wear metals, contaminants and additive elements is a valuable diagnostic tool for scheduling preventive maintenance of engines and equipment in a vehicle fleet and fits into the CBM methodology. Following this path, the main objective of the

integrated maintenance systems is based on the decision support of the manager in several areas of the organization, such as: quality, environment, safety, stock and financial, which is totally related to the premises of the realization of the model proposed in this thesis, which aims to determine the working condition of lubricants and assets, in order to reduce unnecessary costs and increase safety in the provision of services of automotive companies, using the theoretical and practical foundations.

3.4 Fleet Characterization

Public transport in general and especially the urban buses for transporting passengers, represent, in addition to its indispensable role for a significant part of the population, an important alternative to individual transport. For this reason, in today's globalized economy, the survival of these companies depends on their ability to innovate and develop more reliable maintenance methods. In addition to providing a good service, companies must also develop methods for energy saving, environmental conservation, equipment renewal or replacement, reliability, maintainability, efficiency, and optimization of industrial processes in order to be prepared for the adversities that may arise.

Since the urban passenger transport sector is of extreme importance to society, we selected a fleet with a diverse set of vehicles from a transport company that has a diverse and enough set of data to allow the determination of the operating conditions of lubricating oils. We decide to collect data in a sample of the fleet, instead of working with all the buses that compose it. This option was made to simplify and reduce the costs of data collection. To avoid biases and guaranty the generalization of the results it was decided to have an heterogeneous set of buses in the sample.

The sample that will be used to collect data to develop and evaluate the proposed model, presented in the next chapters, consists of 5 vehicles (buses), considering different manufacturers, models, engines, used oil and manufacturing date as shown in Table 14.

Vehicle Identification	Brand	Model	Year of Fabrication	Engine	Engine Oil
21	Mercedes	Citaro O530	2002	OM906h LA	GALP ULtra S3 10W40
34	MAN	12.240 HOCL NL	2007	D0836L OH56	GALP ULtra LS 10W40
67	MAN	14.240 HOCL NL	2009	D0836L OH56	GALP ULtra LS 10W40
80	Temsa	Avenue LF12	2017	ISBe6.7	GALP LD Supra 15W40
82	Temsa	Avenue LF12	2017	ISBe6.7	GALP LD Supra 15W40

Table 14 - Selected Fleet for the Case Study

It is important to mention that, given the differences in the year of manufacture of the selected vehicles, we will have a varied wear structure, either in the lubrication system or in the engine, since the difference between the manufacturing year of the oldest and the newest vehicle is 19 years. Another extremely important point in the selection of these vehicles is the use of different types of oils for each

brand. This will assure that the proposed model will be not only generalist with regard to vehicles and engines, but also with regard to the viscosity of the oils, their inherent characteristics, and the technology employed in them.

3.5 System Architecture

According to Siebert et al. (2021) a machine learning model is a mathematical model or piece of software that an engineer or data scientist makes intelligent by training it with input data. As such, the quality of the model depends on the quality of the training data, so much so that, if we provide false information or unworked data, the trained model will give wrong answers. In this section, and to better understand the pipeline we follow to develop our intelligent system for the detection of the lubricant oil condition, we present an overview of its architecture and the steps it encompasses, Figure 21.

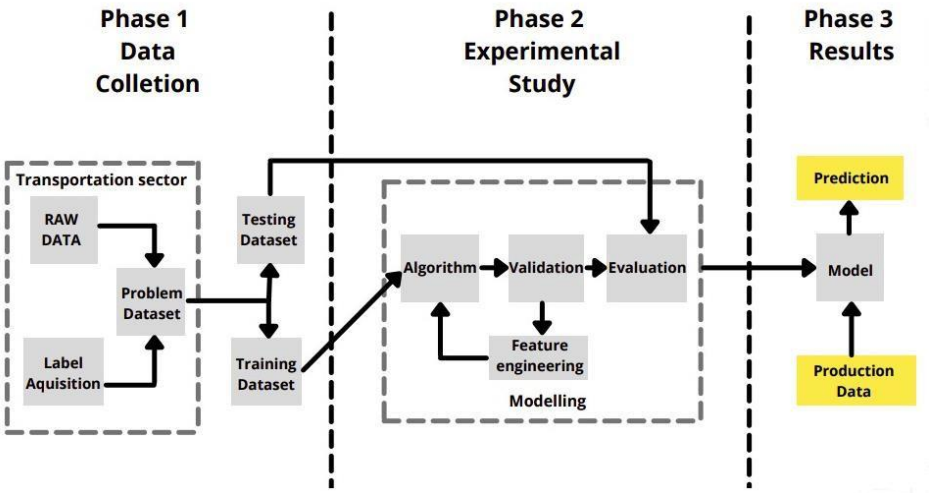

Figure 21 - Overview of the Developed System

Our proposal is composed by three main phases. The first phase is Data Collection; The second phase consists in Modelling; and the third phase the Deployment.

In the data collection step, we start by collecting the data, in real-time, using the proprietary device that is directly linked to the central system of the vehicles. However, it is important to be aware that the collected data cannot be used directly, as it needs to be cleaned, organised and filtered (Siebert et al., 2021). To solve this issue, a data preparation is required and so we split our original dataset into two separate datasets: Test dataset which consists of the data we will use to test our trained model and training dataset which is the data that will train the model so that we have the predictions for our problem.

The second phase is modelling, in this phase we initially select the algorithm we will use to answer our problem and with the selected algorithm and the training dataset we create the first functional model. At this point we validate the initial model with the validation dataset to get the initial results. If the model performance is not sufficient or adequate for the problem, we start the feature

engineering process to create new variables to improve the model performance. It is important to mention that this process is cyclical and is always redone until the model obtains reliable and satisfactory results considering the evaluation metrics.

The Deployment phase is where the implementation of the final training model occurs and is the last step within all the identified phases. The integration of our model into an uncontrolled environment, is an essential step of this phase. There is also a monitoring responsibility that must be undertaken to evaluate the performance of the model while in a production environment. This is to ensure that the model is working well enough and is fit for purpose, and if the results are not sufficient in the production environment, we return to the Modelling phase to retrain and update the system.

It is important to mention that the quality of the ML model is affected by several aspects: the type of task to be solved (i.e. classification, anomaly detection), the type of model (Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Gradient Boosting Classifier), the data used for construction, the way the data is separated for training and test, and with runtime complexity requirements or security constraints. Once the system is created, the experiments are managed (using hyper-parametric search, cross-validation, independent separation of training tests) so that evaluation of the developed model with respect to the results obtained is performed. It does this by measuring performance measures (such as precision or recall for classification tasks), performing sensitivity analysis, or testing against contradictory examples (Siebert et al., 2021).

Chapter 4 - Intelligent System for Oil Classification

With the arrival of artificial intelligence, many projects in lubricant analysis can be improved with the use of ML techniques. But learning the ramification of ML, understanding how it works and why these techniques should be used, especially knowing if the project in question supports the gains and losses of ML are important factors before any implementation. Because of that, in this section we detail the experimental study conducted to evaluate the possibility of using a data-driven approach to automatically determine the condition of a lubricating oil. We start by describing the process followed for data collection, the variables selected, the vehicle characteristics, the label collection and the specific protocol to perform the data collection. Then, we go through the exploratory analysis, where we evaluate the selected variables through analysis of missing values, operational behaviour of the variables, PCA analysis and correlation analysis lastly we present the predictive models and the evaluation metrics used. The study of the models was divided in two stages: the first consists in the application of a simple model, with the purpose of understanding if the selected selected variables allowed us to determine the oil condition. The second stage consists in a more detailed evaluation of several models, where a series of cross-validations were performed to obtain the best model and the best combination of variables. Finally, after selecting the best model and the best combination of variables, two additional studies were performed: the first consists in inserting unknown data in the test dataset and the second analysis consists in inserting data of only one temperature variable in the training and test datasets.

4.1 Data Collection

The data collection stage is very important, and the data collected have to come from a representative sample. In this way, relevant information about the oil and the vehicles were collected, namely:

- What type of oil was used;
- How long it has been in use in the vehicle;
- What are the vehicle's specific characteristics.

Each sample collected is labelled with information about the vehicle to which it refers and stored in a database. Note that the more information we have during collection, the better the diagnosis of the model and the accuracy of the response. In this way, we perform the data collection from January 1, 2020 to August 1, 2020, always bearing in mind that the developed model will be implemented in a real business environment, in a fleet composed by vehicles with different behaviours. Thus, we cannot simply collect as much data as possible without first understanding whether these variables are present in a considerable range of vehicles in the studied fleet, to allow the extrapolation of the results to other companies operating in the same segment.

4.1.1 Variables Description

For the data collection to be generally replicable, it is necessary to describe all the variables that will be used in this experiment. The variables are divided into 3 groups: Engine Characteristics (see Table 15), Oil Characteristics (see Table 16) and Real Time Data (see Table 17).

4.1.1.1 Engine Characteristics

According to Morgan & Liu (2009), the satisfactory operation of Diesel engines depends largely on the formation and maintenance of lubricating films between parts subject to friction and, at the same time, on the nature of these films. Several elements in the cycle of a Diesel engine act together in the lubrication. The engines are designed to run for thousands of kilometers and as such they are built with a complex lubrication system. Thus, it is necessary to collect the technical data of the engines to have a perception of the operation of the lubrication system for which it is intended.

Vehicle Id	Eng description	Eng Cyl Number	Piston Stroke Lenght	Cylinder Diameter	Engine Power (kW)	Max Speed Rpm	Engine Displ (liter)	Crank sump (liter)
21	OM906hLA	6	0.130	0.102	205	2500	6.370	25.0
34, 67	D0836LOH56	6	0.125	0.108	177	2650	6.871	27.5
80, 82	ISBe6.7	6	0.124	0.107	209	2600	6.700	15.1

Table 15 - Engine Characteristics

By analyzing Table 15, we can verify that all the information described in section 2.3.1 is available for each selected vehicle. This information will serve as the basis for the calculations needed for the OSF, which is a fundamental feature for the development of the proposed model.

4.1.1.2 Oil Characteristics

According to Plumley et al. (2018), for the satisfactory performance of the lubrication system, it is necessary to have a lubricating oil in defined quantities, with appropriate characteristics, specific finishing of the surfaces in contact, ideal clearances between the parts and the specific pressure of the contact surfaces. Thus, as mentioned in section 2.4.1, which describes the characteristics of lubricating oils and their importance, to develop an intelligent system to determine the operating condition of lubricating oils, we need to collect the characteristics of the oils that are used by the vehicles (Table 16).

Oil Id	Oil Name	SAE Grade	ODI (km)	Oil dens At 15C (g/mL)	Oil visc At -25C (cP)	Oil visc At 40C (cSt)	Oil visc At 100C (cSt)	Vehicle Id
Oil_1	GALP Ultra S3 10W40	10W40	30000	0.873	7000	90	14.6	21
Oil_2	GALP Ultra LS 10W40	10W40	20000	0.860	7000	102	15.2	34, 67
Oil_3	GALP LD Supra 15W40	15W40	15000	0.877	7000	110	14.5	80, 82

Table 16 - Oil Characteristics

By analyzing Table 16, it is possible to understand that we did not collect information about all the characteristics referred in section 2.4.1. Although TAN, TBN, dialetric constant, and flash point are important characteristics for identifying a lubricant, they were not used because they require more complex chemical tests and equipment and, they did not meet the needs of the system that was developed.

4.1.1.3 Real Time Data

To determine the operating condition of lubricating oils in diesel engines it is necessary to collect a large amount of data generated in real time from the operation of the vehicle. This data are used by the control units to determine the optimal operation of the vehicle and indicate system faults. In our model we will use this same data that passes through the vehicle's communications bus to generate all the information needed to build the model. The hardware developed by Stratio, can provide a range of more than 200 vehicle parameters in real time, but this amount of parameters is dependent on the technology employed in the vehicle, which can vary depending on manufacturer, model and/or year of manufacture. In other words, the technology employed in each vehicle results in a greater or lesser availability of information and a difference in the speed of data acquisition. Therefore, since our selected fleet has vehicles of different brands and models, as well as different manufacturing years, we selected the minimum portion of parameters which we knew would be available for all vehicles and would be possible to collect in an acceptable quantity. Furthermore, this list was selected to ensure that the application could be used in a real environment, and prepared to work in all vehicles of the fleet. Table 17 summarizes the collected and the calculated variables.

Variable identifiers	Variable Name	Unit Measure
Collected Variables		
PID100	Engine Oil Temperature	°C
PID101	Coolant Temperature	°C
PID108	Engine Oil Pressure	mbar
PID114	Vehicle speed	km/h
PID118	Engine speed	rpm
Calculated Variables		
visc_cin	Kinematic Viscosity	mm^2/s
visc_din	Dynamic Viscosity	Pa.s
Oil_z	Oil Stress Factor	$kW.h/liter^2$
OSF_v3	Oil Stress Factor	$W/liter^2 \times 10^4$

Table 17 - Variables Identifiers

When analyzing Table 17, we can verify that in the "Variable identifiers" column we have identifiers for each variable. This type of nomenclature will be used from now on to refer to each of the

variables according to the "Variable Name" column. This type of identifier is used to facilitate the collection and storage of data, either by the vehicle manufacturer or by Stratio which provided the data. It is also important to mention that all the parameters mentioned in Table 17 will be acquired with their respective units of measurement, grouped and normalized to remove outliers or data that could hinder the construction of the model.

4.1.2 Staining Test

According to Delgado Ortiz et al. (2014), the maintenance of internal combustion engines, can be performed using different lubricant analysis techniques to determine the condition of the oil and the system. For more than 50 years, a rapid method of analyzing the condition of the lubricant, called the oil slick test (Figure 22), has been used and improved to estimate the degradation of the lubricant and the presence of contaminants. With this test, through a visual inspection, it is possible to differentiate the contours that experts look for in interpreting the oil slick. According to Delgado Ortiz et al. (2014), the central zone (Figure 23) is characterized by its dark and uniform intensity. The intermediate or diffusion zone of the slick indicate the degree of dispersion of the carbon particles (Figure 24), and presence of contaminants (Figure 25 and Figure 26). With these parameters we can collect, information about the dispersion and detergency conditions of a lubricant by dropping an oil stain on a filter paper, that provide the model labels.

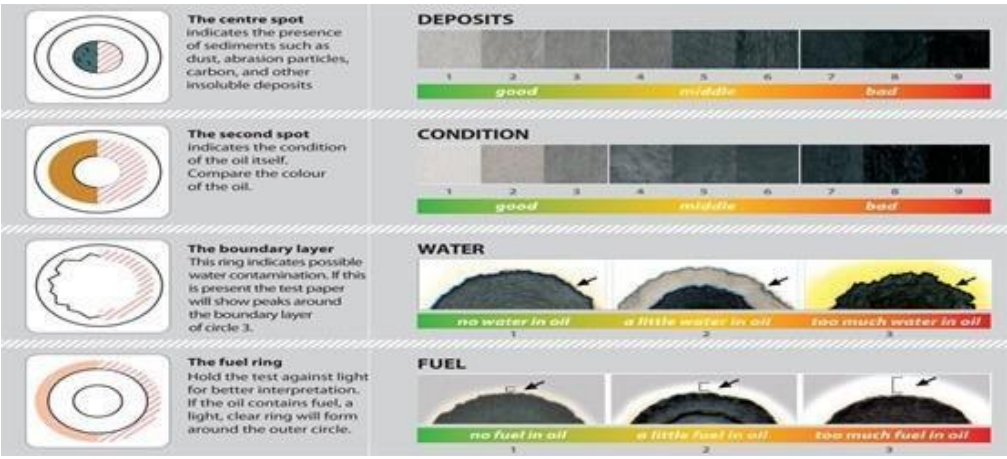

Figure 22 - Oil Stain Tests – Source: (MOTORcheckUP, 2022)

After dropping the oil on the test paper, an image with up to 4 circles based on the chromatographic effect appears. Depending on the age of the oil, the final result will take between 3 to 10 hours. It is best to leave the test result overnight and then compare it with the reference images on the Test Instructions Sheet. The interpretation of the oil slick can be summarized as follow:

1. **The Inner Circle:** shows whether the oil has been contaminated by small particles - for example, exhaust particles, dirt, dust, abrasions, or other substances (Figure 23);

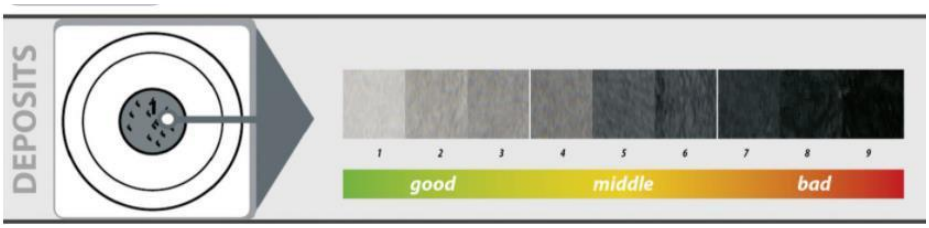

Figure 23 - Inner Circle Show us if Lubricant is Contaminated by Deposits – Source: (MOTORcheckUP, 2022)

2. **The second circle:** shows if the oil is still in good condition. Compare the coloration of the oil (Figure 24);

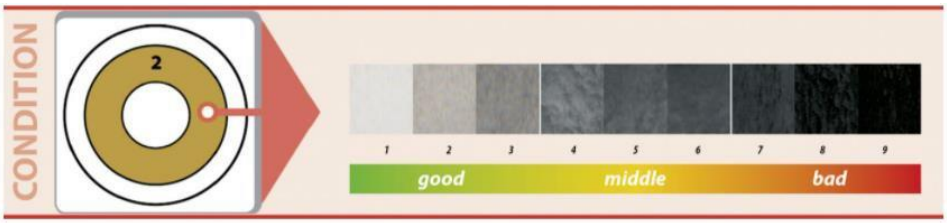

Figure 24 - Second Circle Show us the Condition of the Lubricant – Source: (MOTORcheckUP, 2022)

3. **The Jagged Circle:** It allows to understand if there is water or coolant are present in the oil. The water in the oil forms definite peaks in the outer areas, while the coolant forms a yellow ring around the jagged circle (Figure 25);

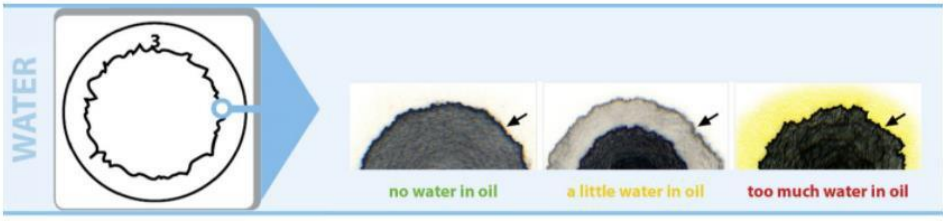

Figure 25 - Jagged Circle Show us if Lubricant is Contaminated by Water or Coolant – Source: (MOTORcheckUP, 2022)

4. **The Fuel Circle**: It allows to understand if there is fuel in the oil. The larger the transparent ring around the outside, the more fuel is in the oil (Figure 26).

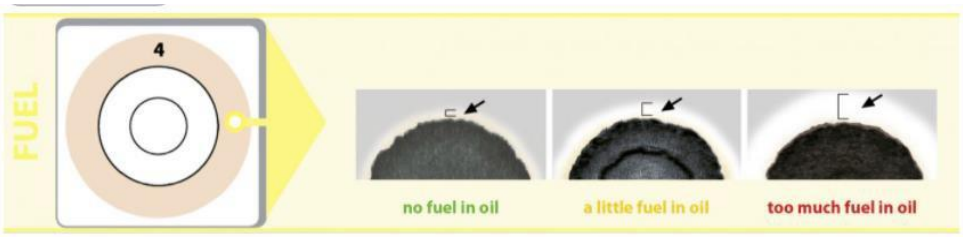

Figure 26 - Fuel Circle Show us if Lubricant is Contaminated by Fuel – Source: (MOTORcheckUP, 2022)

Although the spot test provides results on a scale of 0 to 9 for solid particle contamination and condition, and a scale of no presence, somewhat present and very present for coolant and fuel in the oil, we considered a condition scale with less granularity. In concrete, we define a binary scale, where "0" indicates that the oil is in good condition and "1" indicates that it is not good.

4.1.3 Data Collection Protocol

Oil analysis is a way for the industry to try to expand the oil's operating time. In addition to optimizing the application of the product in the machines, the reduction in the number of oil changes avoids waste, and makes costs with lubricants lower in the production line. This can be achieved through periodic evaluations of the integrity of the lubricants used by the equipment. Therefore, to develop our system, it is necessary to collect and evaluate the lubricants. The collection is considered the step prior to the lubricant analysis, which is performed following the procedure presented bellow. Note that the steps are divided into two segments (direct and indirect) and the materials and equipment used are determining factors in the oil evaluation and need to be handled with care.

The Direct Method corresponds to the steps needed to collect oil samples, which will be used to label the lubricant's operating condition. The procedure followed for this method was:

1. Start the car and wait 10 minutes for the oil to run through the entire engine and warm up to operating temperature;
2. Turn off the car and wait 3 minutes for the oil to return completely to the sump;
3. Take the oil sample wearing unused gloves to avoid contamination of the sample by particles in the environment;
4. Remove a drop of oil, through the dipstick, and place it in the center of the stain test;
5. Wait 3 to 10 hours for the oil drop to be absorbed and for the cotton to separate the contaminants;
6. Compare the test result with the results sheet;
7. Take a picture of the test and store it in a contamination free environment.

It is important to mention that this sequence of steps ensures that the particle concentrations reach a uniform distribution throughout the lubricant without contaminating the sample or losing valuable information. To store the results of the analysis digitally, we created a spreadsheet. This spreadsheet contains the details of the steps followed by the direct method. An example is shown in Table 18.

Sample collection Id	Vehicle Id	Date	Run engine	Stop engine	Time to analyze	Picture
Sp_01	21	11/01/2020	10:05	10:18	ok	ok
Sp_02	34	11/01/2020	10:03	10:17	ok	ok
Sp_03	67	11/01/2020	10:11	10:22	ok	ok
Sp_04	80	11/01/2020	10:17	10:28	ok	ok
Sp_05	82	11/01/2020	10:23	10:32	ok	ok

Table 18 - Direct Method to Collect Oils Samples

As it can be seen from Table 18, for each vehicle we have registred information about the vehicle, the engine start and shutdown time, as well as if the time for analysis was met and if a picture was taken to store the test in digital format for other possible analyses and search results. Another important point to mention is that all sample collections have a unique identifier (i.e., Sp_01)

The Indirect Method: corresponds to the data that will support the determination of the lubricant's working condition. In other words, complementary data obtained with the staining tests and that will be important elements to indicate anomalies during the development of this system, as well if the results are satisfactory at the end of the data analysis process. The additional information collected is detailed bellow:

1. External Temperature Measurement: Collecting the external temperature at the time of the collection of the lubricating oil sample is important because, temperature is a key factor in measuring viscosity. Viscosity is one of the most important characteristics of lubricants and it is necessary to understand if the external temperature influences this variable. Although we have the collection of data stipulated in the period between January 1, 2020 to August 1, 2020 with average temperatures between 8°C and 32°C in Portugal, the lubricants are not manufactured only to operate in this range of values. If we were to implement this system in a fleet that constantly operates in a location with high temperatures (above 35 degrees daily), or on the contrary in locations with low temperatures (below 10 degrees), we could analyze whether the constant high temperatures of the lubricant at rest in the vehicle's crankcase could influence premature wear. On the other hand, assuming a situation of constantly low temperatures, we could check whether prolonged and more frequent heating of the lubricant from a lower temperature to the operating temperature (between 83° and 93° Celsius) would also cause premature wear of the lubricant;

2. Measurement of the lubricant temperature at the time of collection: The measurement of this temperature is related to the differentiation between vehicles that were operating normally shortly before the sample was collected and vehicles that had been at rest for a long period of time. That is, in cases where the vehicle was operating normally, the lubricant had been at the ideal operating temperature period longer than the 10 minutes determined by the direct method. This type of situation corresponds, in principle, to an ideal lubricant viscosity when compared to the situation of vehicles that are at rest. The resting vehicles do not have the possibility to get to operating temperature earlier and work under optimal conditions gradually, running the lubricant through the whole system several times. This information might help justify some anomaly in the collection of the sample or even in the oil level, knowing that the lower the level of lubricant in the carter

the greater the possibility of finding sediments or contaminating materials in the sample when the sample was taken;

3. Check the vehicle mileage at the time of collection: The collection of this information is important to monitor both the operation of the vehicle (highest and lowest level of operation) and the lubricant's operating time. When the oil samples are taken at 15-day intervals, we should pay attention to whether there is an anomaly in the operation between samples. A week in which the vehicle worked for more hours than usual, or for more kilometers, may cause excessive wear during that period, either on the vehicle, because it operates for long distances in a short period, or on the oil because it operates close to its mileage limit. On the other hand, we can also calculate and determine that the degradation index found at the time the sample was taken corresponds to the mileage of the vehicle and the mileage of the oil;

4. Check the oil level at the time of collection: As mentioned earlier, to maintain the perfect operation of the vehicle and diesel engine, it is necessary to keep the lubrication system always operating at an adequate level and with ideal lubricant. If a situation arises in which the vehicle presents a low level of lubricant when the sample is taken, it indicates that the vehicle might have been operating with inefficient lubrication and that somewhere in the system there is oil leakage, whether due to leakage in the sealing rings, defects in the sealing surface, failures in the vacuum pumps, or even excessive oil pressure. Another important point to be mentioned with this measurement is that with less oil operating in the system, more compounds and/or contaminating materials will be dragged to the bottom of the sump, and thus the condition of the remaining oil we want to sample will be worse;

5. Oil refill information: In many fleets of heavy vehicles, a consumption of around 0.6-0.8 liters per 10000 km is common in relatively new vehicles, but this all depends on the condition of the engine, the pump, the oil filter or the rotation level reached of the engine. In addition, this consumption increases over time due to the natural wear of engine parts, and it can even differ from period to period if the vehicle exceed its normal capacity, such as, vehicles that are carrying a heavier load than normal and/or are being used under extreme conditions. In other words, if the engine is working hard, it consumes more oil. These numerous variables make this average consumption relatively complex to maintain in a diversified fleet such as the one selected. So, to understand if we are determining the degradation rate of the oil that actually operated throughout the interval between samples, we needed to collect the amount of oil that was inserted daily in the vehicles, to avoid performing tests on oils that were new due to the excess of insertion. Through this information we can determine whether the oil we are evaluating at the time of sample

collection, has high amounts of new oil, since the total amount consumed by the vehicle has already been replaced in its entirety over the course of its needs. Assuming that heavy duty diesel engines need 25 liters, we can have a situation where during the week 5 liters of lubricant were refilled due to the level maintenance needs. This value corresponds to refills on Sunday (2 liters), Wednesday (3 liters) and Saturday (1 liter). In the end these refills represent about 20% of the total oil in the vehicle, without major wear;

6. Verification of services and date of the last lubrication - When we talk about maintenance of vehicles in a fleet, it is normal to have a well-structured sequence of planned maintenance, such as: changing brake pads at an interval of 30 thousand kilometers, changing engine lubricants at an interval of 20 thousand kilometers and changing oil filters at an interval of 2 oil changes. Usually, these maintenances are determined by the manufacturers according to brand tests and are generic without any optimization for the real situation of the fleet. However, at this point, considering the oil change and the oil itself, we must keep in mind the information about when we have the new oil in perfect conditions to operate free of any contamination from wear and tear of the equipment and when we have the new oil filter installed, whose function is to store the largest amount of impurities in the system when the oil circulates between the crankcase and the engine, avoiding excessive loss of lubricant properties. This information, whether from oil change or filter change, is important to characterize the real condition of the lubricant and the equipment in general. This prevents us from obtaining positive results for the operating condition of a really bad lubricant and a worn out piece of equipment because the filter has stored the largest quantity of particles that would determine its real operating condition. This will also avoid the errors associated to the analysis of an oil that has just been changed, in an equipment that has already undergone excessive wear, but the lubricant has not had time to operate and circulate throughout the system to collect the impurities.

It is worth noting that the collection of this additional information will help to validate if the steps in the direct method were followed correctly, as well as help us verify that external factors, such as outside temperature, differentiation in operating style, and even the insertion of new oil into the lubrication system, will not alter the expected results. Another important point to mention is that all the information presented was carefully catalogued, accordingly to Table 19, Table *20* and Table *21* structure.

Sample collection Id	Ext_temp (°C)	Oil_temp_AT (°C)	Km_actual	Oil_Nvl_actual
Sp_01	10	68	882555	25
Sp_02	10	80	730018	25
Sp_03	10	55	694629	25
Sp_03	10	75	149369	25
Sp_03	10	86	148950	25

Table 19 - Oil Sample Collection

Last_OD_ID	Date	Type of Service	Components	Oil Id	Oil Quantity	Vehicle Id
LOD_01	10/02/2019	Oil - TIPO 1	Oil	Oil_01	25	21
LOD_02	10/06/2019	Oil - TIPO 2	Oil + Filters	Oil_01	25	21
LOD_03	10/10/2019	Oil - TIPO 1	Oil	Oil_01	25	21

Table 20 - Last Oil Drain: Type of Service and Dates

Oil_rep_ ID	Oil_Id	Vehicle _Id	Week	Sun.	Mon.	Tue.	Wed.	Thu.	Fri.	Sat.
Oilrep_01	Oil_01	21	1 de 2020	2	0	0	3	0	0	1
Oilrep_02	Oil_01	34	1 de 2020	0	2	0	0	0	3	0
Oilrep_03	Oil_01	67	1 de 2020	2	1	1	3	0	0	1
Oilrep_04	Oil_01	80	1 de 2020	0	0	2	0	0	1	0
Oilrep_05	Oil_01	82	1 de 2020	0	2	0	0	0	1	0

Table 21 - Oil Reposition Week Before the Test

Through the analysis of Table 19, Table 20 and Table 21, we can see that all data are easily accessible and well organized, following the same approach existing for the direct method and using a unique identifier for each set of information. In addition, storing this information in digital format with unique identifiers for each step structure (i.e., Sp_01, LOD_01, Oilrep_01) or element (vehicle: 21, oil: Oil_01) facilitated the corelation of this data and the development of the system as a whole, since from a structured database we can perform several tests without worrying about failures in the results that are not scaled and that could not be evaluated in depth if they occurred. Exemplifying this interconnection of information, we can see that vehicle 21 is a 2002 Mercedes Citation O530 (Table 14) with an OM606h LA engine (Table 15) and uses GALP Ultra S3 10W40 lubricant (Table 16). On January 11, 2020, the sample Sp_01 was collected with the engine start time at 10:05 and shutdown time at 10:18, meeting the required time for analysis and storage of the test in digital format (Table 18). During this collection the external temperature was 10º Celsius, the initial oil temperature was 68º Celsius, and its mileage was 882555, with the oil level at a maximum of 25 liters (Table 19). Checking other information, as of the date of the collection there was no oil or filter change (Table 20) and the

replacement for level maintenance had a total of 5 liters in the previous week (Table 21) and there was an anomaly in the operation of this vehicle this week. Finally, following this procedure of information collection and interconnection of data sets, in the end 65 oil samples were collected in the period of eight months, Table 22 summarizes the number of samples collected per vehicle.

Vehicle Id	Number of samples
21	8
34	13
67	12
80	16
82	16

Table 22 - Number of Samples Collected per Vehicle

By analyzing Table 22 we can see that there is a lower number of samples from vehicle 21 compared to the rest. This difference was due to an undiagnosed failure in this vehicle, which was not related to the lubrication system, but left it inoperable for 2 months. There is one more reason to the difference in the number of samples, and it is related to the fact that the collections were always made in determined periods (every 15 days), at a fixed time (morning) and in a real environment, where we had little control over the availability of the vehicles by the client. In other words, if the client needed to use the vehicle, or if the vehicle was not parked in the yard, we could not make the pickup, and thus lost that sample.

Chapter 5 - Exploratory Data Analysis

Before starting a machine learning project, it is important to ensure that the data is ready for the modelling work. Exploratory Data Analysis (EDA) ensures the readiness and ability of the data to be used in a machine learning system and allows to better understand the domain of the problem. Thus, we perform the following analysis: Missing Values Analysis, Variable Behaviour analysis, analyze the need for data balancing, PCA and Correlation analysis.

5.1 Missing Values Analysis

According to Y. Zhang et al. (2021), missing values appear in real world scenarios due to multiple reasons such as collection errors or the data acquisition format itself. These missing values might lead to the deterioration of the performance of analytical monitoring applications. As such, analysing the existence of missing values helps to address several concerns caused by incomplete data, since this lack of information can lead to misleading interpretations and possibly failures in the models.

Before beginning the analysis, it is important to mention some information about the dataset. The initial set of the variables was collected in real time (Table 17) from 5 selected vehicles (see Table 14) obtained through the device developed by Stratio and with a variable acquisition rate. This variable acquisition rate is a fundamental point in the development of this study and is related to two factors:

- The bus communication technology and speed which is dependent on vehicle manufacturer, model and year;
- The priority level that the variable has in the bus.

Since Stratio's data collection equipment gathers information through various communication protocols, and each protocol has its particularities in terms of communication speed, it is necessary to manage these variable acquisition rates internally. However, even performing this management through proprietary software, Stratio's equipment is unable to collect the same amount of information from all vehicles, because of limitations in the protocols implemented by the manufacturers. Therefore, the initial dataset is formed by data parcels from all selected vehicles, separated by selected variable, with an average acquisition rate of 1 Hz. It is still important to mention that the Stratio device sends all information to a cloud storage structure through a mobile network or wifi to reduce the costs. After the message is received on the server, it is broken into pieces of information (variable points) with their respective timestamps and identifiers, and it is stored in a searchable data structure.

Metrics/Parameters	Oil Temp	Coolant Temp	Oil Pressure	Vehicle speed	Engine speed
Count	183451	313893	3448095	3137167	3462106
Mean	72.38	72.77	3027.69	27.15	1135.39
Std	11.71	10.65	695.77	14.49	258.12

Min	4.19	4.00	0.00	0.00	0.00
Quartile 25 %	67.69	68	2560	16.53	1008
Quartile 50 %	75.5	76	3000	27.35	1130
Quartile 75 %	81	81	3520	37.08	1264
Max	126.70	127	5840	89.10	2676
Size	4542016	4542016	4542016	4542016	4542016
% of Null values	95.96	93.08	24.08	30.93	23.77

Table 23 - Metrics of the Data Before the Missing Value Reduction

Looking at the percentage of missing values (last line of Table 23) and through the descriptive metrics of the initial data set, it is possible to see that some of the selected parameters contained a large amount of missing data and should be addressed prior to their use in the study. In this analysis, it can be seen that the parameters oil temperature and coolant temperature have a high percentage of missing values compared to vehicle speed and engine speed. This is due to some data acquisition characteristics of the Stratio hardware (Automotive, 2021). The data collection by the Stratio system is governed by two central rules, value variation and time between acquisitions. These rules exist to avoid interferences with the vehicle's internal bus communication to prevent communication failures in safety systems (brakes and steering). As such, the system tries to keep the information request rate low. The parameters related to temperature (coolant and oil) have little variation of values and a slower variation when compared to the vehicle speed and engine speed parameters.

Taking into account the way the parameters are collected by the Stratio device, we can still mention the structure for grouping the data that explains this difference in the number of points for each parameter. As mentioned before, at the time of data acquisition, each parameter is acquired with its identifier and timestamp. By grouping these parameters in the same time series, we can see that the largest amount of acquired values are from the parameters with the largest range of variation and requiring the shortest acquisition time. However, the distribution of values per timestamp is not homogeneous, that is, there are periods in which we receive more values from one parameter than from another. For example, when the vehicle is turned on we receive more values of temperatures than of engine rotation and speed, since the vehicle is stopped in idle (without speed and rotation between 550 and 650 rpm) and the temperature is rising to stabilize at the operating temperature (see Table 24).

Timestamp	Oil Temp.	Coolant Temp.	Oil Pressure	Vehicle Speed	Engine Speed
02/01/2020 05:49:36	null	5.94	null	null	null
02/01/2020 05:49:37	null	null	5	null	0
02/01/2020 05:49:44	null	null	null	0.00	null
02/01/2020 05:49:49	6.09	null	null	null	null
02/01/2020 05:49:56	null	null	3495	null	582
02/01/2020 05:50:04	null	null	3468	null	614
02/01/2020 05:50:20	7.18	null	null	null	null
02/01/2020 05:50:21	null	null	3445	null	null
02/01/2020 05:50:29	null	7.69	3420	null	null

02/01/2020 05:50:37	8.21	null	null	null	null
02/01/2020 05:50:57	null	9.44	3395	null	null
02/01/2020 05:51:04	10.00	null	null	null	null
02/01/2020 05:51:06	null	null	null	null	590
02/01/2020 05:51:14	null	10.60	null	null	null
02/01/2020 05:51:21	11.10	null	null	null	null
02/01/2020 05:51:30	null	null	3375	null	null
02/01/2020 05:51:51	12.70	null	null	null	null
02/01/2020 05:51:58	null	13.10	null	null	null
02/01/2020 05:52:06	null	null	3398	null	null
02/01/2020 05:52:13	14.00	null	null	null	null
02/01/2020 05:52:32	null	14.60	null	null	null
02/01/2020 05:52:39	15.20	null	null	null	null
02/01/2020 05:52:41	null	null	3418	null	null
02/01/2020 05:52:59	null	15.90	null	null	null
02/01/2020 05:53:07	16.50	null	null	null	null
02/01/2020 05:53:08	null	null	3438	null	null
02/01/2020 05:53:33	null	17.10	null	null	null
02/01/2020 05:53:40	17.70	null	null	null	null

Table 24 - First Data Analysis to Identify Miss Values

It is important to mention that when fetching a piece of data from Stratio's database, it is necessary to enter the start and end date information for the period we are looking at. Since we are collecting samples from January 1, 2020 to August 1, 2020, we need to collect the real-time information from this same period. With the insertion of the start and end date information into our database, the result obtained is a dataframe sequence of points. That is, the database returns as rows the values of each selected variable indexed to the message reading timestamp. Table 24, shows is an example of the result of a request to Stratio's database. We can see the data formatting and the example of the vehicle's starting moment, where we have more temperature and pressure data than engine rotation and vehicle speed data. It is also important to mention that the message can contain more than one parameter (see timestamp line 02/01/2020 05:50:57, Coolant Temp. = 9.44 and Oil Pressure = 3395) or can contain only one parameter (see timestamp line 02/01/2020 05:49:49, Oil Temp. = 6.09).

Timestamp	Oil Temp.	Coolant Temp.	Oil Pressure	Vehicle Speed	Engine Speed
02/01/2020 06:11:57	null	null	Null	46.00	null
02/01/2020 06:12:00	null	null	4495	null	1360
02/01/2020 06:12:05	null	null	Null	43.00	null
02/01/2020 06:12:07	41.80	null	Null	null	null
02/01/2020 06:12:08	null	41.10	4345	null	null
02/01/2020 06:12:09	null	null	null	null	1120
02/01/2020 06:12:14	null	null	null	38.00	null
02/01/2020 06:12:18	null	null	4458	null	1452
02/01/2020 06:12:24	null	null	null	45.00	null
02/01/2020 06:12:29	43.10	null	null	null	null

02/01/2020 06:12:36	null	42.40	null	null	1360
02/01/2020 06:12:41	null	null	null	40.00	null
02/01/2020 06:12:44	null	null	4323	null	null
02/01/2020 06:12:45	null	null	null	null	1098
02/01/2020 06:12:51	44.20	null	null	null	null
02/01/2020 06:12:53	null	43.40	4425	null	1290
02/01/2020 06:13:01	null	null	4668	null	null
02/01/2020 06:13:02	null	null	null	null	1456
02/01/2020 06:13:06	null	null	null	23.00	null
02/01/2020 06:13:10	null	null	4358	null	1146
02/01/2020 06:13:15	null	null	null	33.50	null
02/01/2020 06:13:19	null	null	4293	null	1216
02/01/2020 06:13:26	null	null	null	49.50	null
02/01/2020 06:13:31	46.90	null	null	null	null
02/01/2020 06:13:37	null	46.90	null	null	null
02/01/2020 06:13:38	null	null	4458	null	1320

Table 25 - Second Data Analysis to Identify Miss Values

Looking at Table 25, we can identify another example of the heterogeneity of the distribution of values by timestamp. In this situation the vehicle is operating and there are more engine speed and vehicle speed values than temperature values. The same is true to the pressure values that are directly correlated with the engine speed. As data collection is more important when the vehicle is operating, the largest amount of values come from when the vehicle is on the move. Thus, to deal with this difference in the number of real values between the variables, we performed three additional steps to accommodate some specifics aspects related to our problem:

- **Step 1 - Split Dataset into Trips:** Which corresponds to the identification of the periods when the vehicle was running, i.e., making a trip. To do this, we will analyze the periods where the engine speed was above 550 RPM and the difference between timestamps is greater than 5 minutes, i.e., engine speed >= 550 AND timestamp difference > 5 minutes. The rotation information will indicate that the vehicle engine is running, because it operates in idle speed between 550 to 650 rpms and everything above these values indicates that, either the vehicle is moving, or the vehicle is stopped but connected and the accelerator pedal was actuated to increase the rotation and load the systems faster. Through the timestamp difference information, we will identify the beginning and end of the trip, since the acquisition systems does not work with the engine off. That is, if the server stops receiving data for more than 5 minutes it means that the vehicle has shut down and the trip has ended.

Timestamp	Oil Temp.	Coolant Temp.	Oil Pressure	Vehicle Speed	Engine Speed
02/01/2020 21:39:22	null	null	4153	null	1022
02/01/2020 21:39:27	null	null	null	12.50	null
02/01/2020 21:39:30	null	null	4290	null	1074

02/01/2020 21:39:43	null	null	4388	null	1180
02/01/2020 21:39:48	null	null	null	18.20	null
02/01/2020 21:40:01	null	null	3773	null	964
02/01/2020 21:40:08	null	null	null	11.00	null
02/01/2020 21:40:11	null	null	2108	null	null
02/01/2020 21:40:12	null	null	null	null	594
02/01/2020 21:40:16	null	null	null	3.30	null
02/01/2020 21:40:20	null	null	2150	null	null
02/01/2020 21:40:24	null	null	null	0.10	null
02/01/2020 21:40:37	null	null	1818	null	312
03/01/2020 05:47:19	null	14.10	null	null	null
03/01/2020 05:47:20	null	null	298	null	584
03/01/2020 05:47:27	null	null	null	0.00	null
03/01/2020 05:47:32	13.20	null	null	null	null
03/01/2020 05:47:33	null	12.80	3520	null	null
03/01/2020 05:47:42	null	null	null	null	614
03/01/2020 05:47:54	null	null	3495	null	null
03/01/2020 05:48:11	null	13.90	null	null	null
03/01/2020 05:48:12	null	null	null	null	592
03/01/2020 05:48:17	14.50	null	null	null	null

Table 26 - Third Data Analysis to Identify Miss Values

Through the analysis of Table 26 it is possible to see that there is a pause in the operation of this vehicle (between timestamp 02/01/2020 21:40:37 and 03/01/2020 05:47:19), but this comes sequentially when collecting the data from Stratio's database. That is, for a human analyst with knowledge about the data, this stop time indicates that a trip was finished at 21:40:37 on 02/01/2020 and a new trip started on 03/01/2020 at 05:47:19, but for an automated system it is considered a sequential data without any distinction between them. To be able to impute missing data, we need to reduce the possibility of error and for this we must perform the division in trips, so that the data that are inserted are not outside the range of operation of that particular period. We can exemplify this in the following way: When we take the data from the trip that started on 03/01/2020 at 05:47:19 we know that the first real value of the Oil Temp. was 13.20, of the Coolant Temp. was 14.10, of the Oil Pressure was 298, of the Vehicle Speed was 0 and of the Engine Speed was 584, that is, we know the exact behavior of the variable at the beginning of the trip, in the same way that if the trip ended on timestamp 03/01/2020 05:48:17 we would know the final values. This would give us the possibility to complete the missing values more accurately knowing for example that the Oil Temp. values went from 13.20 to 14.50, in an ascending manner, took about 40 seconds and had 5 missing points between the known initial value and the known final value;

- **Step 2 - Using value interpolation:** According to (F. Wang et al., 2021), interpolation is the process of using known data values to estimate unknown data values. One of the simplest methods is the linear interpolation, that only requires knowledge of two points and the constant rate of change between them. This step generates a smoother and more

realistic variation of values than using average values between acquired data, since a given value does not change to a much larger or smaller value immediately, there is always a gradual increase or a gradual decrease. For example, a temperature increases or decreases degree by degree (depending on the sampling level) and does not change immediately from 10°C to 30°C, if we chose to use the average, we would have all the values within this range at 20°C, which does not represent reality as well as using 19 values in this same range of increase (11°C, 12°C, 13°C, 14°C, ... up to 30°C).

Timestamp	Oil Temp.	Coolant Temp.	Oil Pressure	Vehicle Speed	Engine Speed
03/01/2020 05:47:19	null	14.10	null	null	null
03/01/2020 05:47:20	null	13.77	298	null	584
03/01/2020 05:47:27	null	13.45	1372	0.00	591
03/01/2020 05:47:32	13.20	13.12	2446	null	599
03/01/2020 05:47:33	13.41	12.80	3520	null	606
03/01/2020 05:47:42	13.63	13.16	3507	null	614
03/01/2020 05:47:54	13.85	13.53	3495	null	606
03/01/2020 05:48:11	14.06	13.90	null	null	599
03/01/2020 05:48:12	14.28	null	null	null	592
03/01/2020 05:48:17	14.50	null	null	null	null

Table 27 - Fourth Data Analysis to Identify Miss Values

By analyzing Table 27, it is possible to notice that some missing values were replaced by real values, when we use the linear interpolation method. This statement can be confirmed when we compare the rows between timestamps 03/01/2020 05:47:32 and 03/01/2020 05:48:17 of the variable oil temp. in Table 26 (before interpolation) and Table 27 (after interpolation). Note that the linear method ignores the index and treats the values as equally spaced, only needing to know two real values (initial and final value) and the number of points that need to be inserted. But when performing this type of method, we only replace the missing values by real values in the range of known values, i.e., when looking at the result of this method we can see that the values in the first row could not be filled, because the filling direction of the values is forward (in the case of values of the Oil Temp. from the value 13.20 towards the value 14.50) and there is no previous value that could have been used in the interpolation;

- **Step 3 - Filling missing values:** To fill the nulls values two other methods were used sequentially: backward fill and forward fill. Is important to mention that these two methods were used to replace the “null” values by a known real value, since we did not have a range as needed in the interpolation method, so that we could smoothly estimate the increase or decrease of the variable values.

Timestamp	Oil Temp.	Coolant Temp.	Oil Pressure	Vehicle Speed	Engine Speed
03/01/2020 05:47	13.20	14.10	298	0.00	584
03/01/2020 05:47	13.20	13.77	298	0.00	584
03/01/2020 05:47	13.20	13.45	1372	0.00	591

03/01/2020 05:47	13.20	13.12	2446	0.00	599
03/01/2020 05:47	13.41	12.80	3520	0.00	606
03/01/2020 05:47	13.63	13.16	3507	0.00	614
03/01/2020 05:47	13.85	13.53	3495	0.00	606
03/01/2020 05:48	14.06	13.90	3495	0.00	599
03/01/2020 05:48	14.28	13.90	3495	0.00	592
03/01/2020 05:48	14.50	13.90	3495	0.00	592

Table 28 - Fifth Data Analysis to Identify Miss Values

Backward fill is used to fill the missing values backwards, copying the last known value and forward fill will propagate the last valid observation forward. Through the analysis of Table 28, we can see that for Oil Temp. values, the backward fill method was used from timestamp 03/01/2020 05:47:32 backward, and for Oil Pressure from timestamp 03/01/2020 05:47:54 the forward fill method was used forward.

It is important to mention that the insertion of missing values with a low error margin is only possible, if before using the filling methods, we identify the trips for each vehicle. This step of splitting into trips is fundamental, since it prevents data from one trip being extrapolated to another that started long after the engine was shutdown. After inserting the missing values following the two determined steps, we can observe in Table 29 that the percentage of usable values in the model has increased significantly, making the grouping of parameters much more stabilized and without large discrepancies in the amount of data.

Metrics/Parameters	Oil Temp.	Coolant Temp.	Oil Pressure	Vehicle speed	Engine speed
Count	4337234	4339705	4442110	4519072	4534626
Mean	73.09	72.71	3027.69	26.90	1103.78
Std	9.22	9.20	695.77	14.95	266.48
Min	4.19	4.00	0.00	0.00	0.00
Quartile 25 %	68.71	68.28	2560	16.17	986.00
Quartile 50 %	75.04	74.61	3000	27.39	1112.00
Quartile 75 %	80.23	79.85	3520	37.21	1244
Max	126.7	127	5840	89.10	2676
Size	4542016	4542016	4542016	4542016	4542016
% of Null values	4.50	4.45	2.19	0.50	0.16

Table 29 - Metrics of the Data After the Missing Value Imputation

It is important to mention, that the value insertion model adopted did not significantly affect the initial statistics identified in Table 23 it only reduced the occurrence of missing values (% of null values in the Table 29). This shows that the inserted data are close to reality and does not negatively affect the results. Besides, we obtain a negligible percentage (less than 5%) of missing values that makes it possible to use this data set for the study.

5.2 Variables Behaviour Analysis

Visual analysis of data represents one of the most frequently used initial data analysis techniques in the development of intelligent fault prediction systems. By visually analyzing the data, and/or investigating the behaviour of the variables that will serve as the basis for model development, the developer can better understand the problem at hand and make an informed decision about the algorithms that will be needed. To perform this type of analysis, we chose to use density plots which allow us to get a quick visual understanding of the distribution of the values of the variables. According Scott, (2019) in this type of graph, the x-axis represents a data point of the variable as in a histogram and the y-axis is the probability density function for the kernel density estimate. However, we must be careful to specify that this is a probability density and not a probability. The difference is that the probability density is the probability per unit on the x-axis. To convert to an actual probability, we need to find the area under the curve for a specific interval on the x-axis. In other words, the y-axis values are the determination of the probability of a point occurring between two values x1 and x2, represented by the area under the curve between these two points. We generally use the y-axis values of a density plot as a value only for relative comparisons between different categories, and in this way this type of plot is useful to visualize three properties:

1. Skewness: describes the symmetry of a distribution. Density curves allow us to quickly see whether or not a graph is skewed to the left, to the right, or if it is centered around a certain value:

2. Mean and median: Depending on the skewness of a density curve, we can quickly tell if the mean or median is larger in a given distribution. In particular:

    a. If a density curve is left skewed, then the mean is lower than the median;

    b. If a density curve is skewed to the right, then the mean is greater than the median;

    c. If a density curve is no skew, then the mean is equal to the median.

3. Number of Peaks: Density curves also allow us to quickly see how many "peaks" there are in a given distribution. When the distributions has only one peak, it is describe as unimodal. However, some distributions can have two peaks which we call bimodal distributions and, in rare cases, we can also have multimodal distributions that have more than two peaks.

Thus, from this type of analysis, as shown in Figure 27 to Figure 41 we have conditions to identify trends of the data, determining similarity or not of the behavior of the variables among the selected vehicles, despite being of different brands, models and years of manufacture. It is important to mention that vehicles 30 and 80 are not being presented in this study because the data are similar to vehicles 67 and 82 respectively.

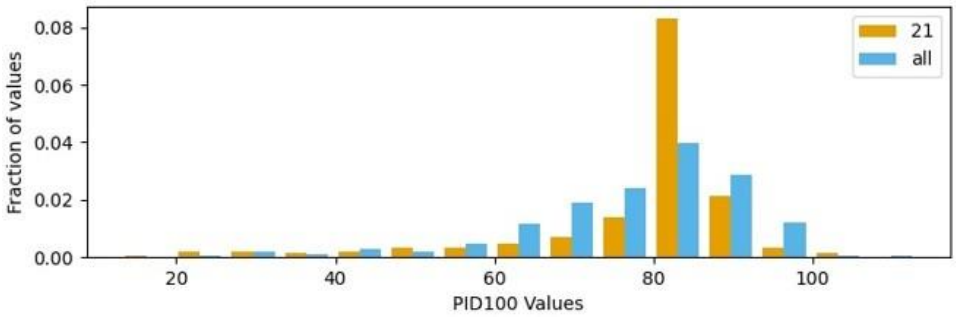

Figure 27 - Oil Temperature Data from Vehicle 21

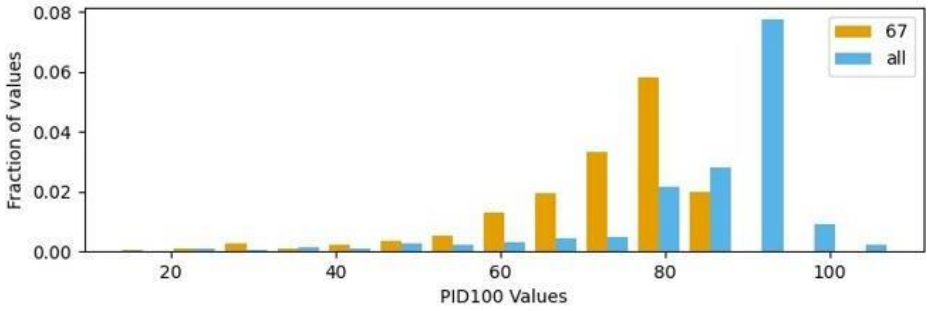

Figure 28 - Oil Temperature Data from Vehicle 67

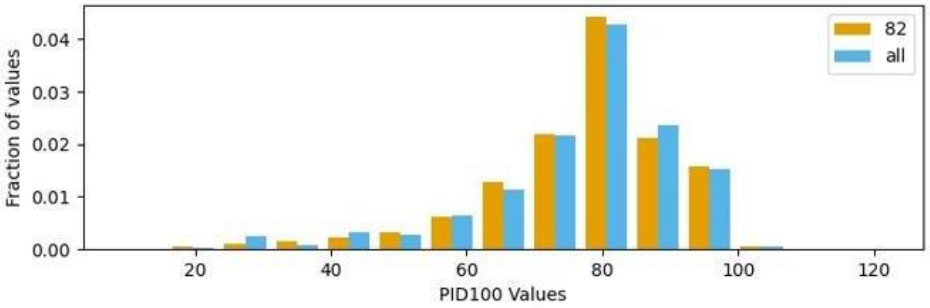

Figure 29 – Oil Temperature Data from Vehicle 82

Figure 27, Figure 28 and Figure 29 represent the behavior of Oil Temp. in vehicles 21, 67 and 82. Looking at data presented in the figures it is possible to see that the oil temperature has a similar range of values for all vehicles (0 to 120° C). However we have a variation in the percentage of values in each temperature segment. Figure 27 shows that vehicle 21 works more frequently in a zone close to 80 degrees, having a probability density of values almost double the sum of all the other ranges of values, while in vehicles 67 and 82 this distribution of values is smoother. By analyzing Figure 28, we can determine that vehicle 67 does not exceed 90°C, showing that this vehicle has a very strict temperature control and during data collection never had an episode of engine overheating. On the other hand, vehicles 21 and 82 have values above 100°C and despite not being high temperatures that imply a reduced viscosity (see Figure 6) it may indicate a potential failure of the cooling, temperature control and/or the lubrication system, since one of the functions of the lubricant is to cool the engine. When we analyze Figure 29 we see that vehicle 82 has a more homogeneous temperature variation compared to the other two vehicles (21 and 67), maintaining itself at the ideal operating temperature determined by

the manufacturer (83° to 93° C). Regarding vehicle 67, it has a higher percentage of values below this oil temperature, remaining cooler and with a higher viscosity. Note that this does not imply a failure or potential failure, but determines an increase in fuel consumption due to the need for more power to move the engine components.

We now move on to another analysis, following the same principles but using the coolant temperature data as shown in Figure 30, Figure 31 and Figure 32.

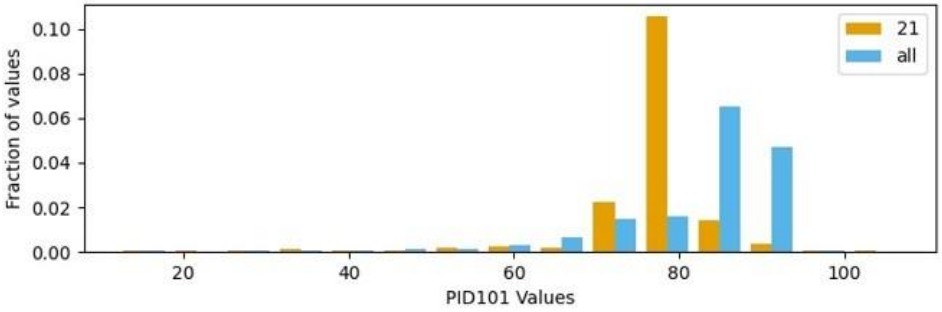

Figure 30 - Coolant Temperature Data from Vehicle 21

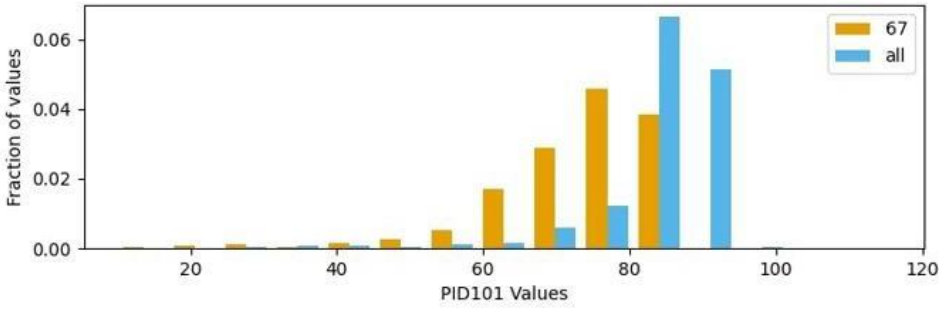

Figure 31 - Coolant Temperature Data from Vehicle 67

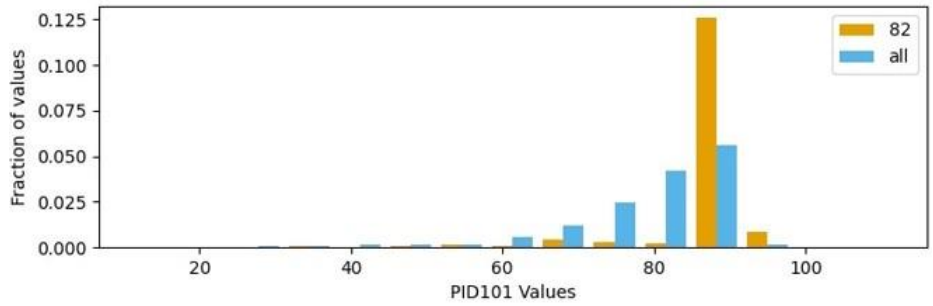

Figure 32 - Coolant Temperature Data from Vehicle 82

It is normal to have a similar behaviour between the oil temperature and the coolant temperature. This is due to the fact that these two variables should be kept at a difference of 5° to 8°C between them. By analyzing Figure 30, Figure 31 and Figure 32 we realize that this is true, i.e., the behaviour is similar between the two variables but the variation of values is much more rigid, having a percentage of values very close to each other. This rigidity and/or proximity of values is due to the fact that the cooling fan has a direct influence on this variable, and, in many cases, it does not have a smooth

rotation control, often staying at constant rotation percentages. Analyzing Figure 30 we notice that the vehicle 21 keeps the coolant temperature almost always a little below 80°C, which indicates that the fan is on all the time in a constant rotation, avoiding the variation of the coolant temperature. Figure 32 show that the same is true for vehicle 82, with the temperature staying around 85°C. When analyzing Figure 31 we notice that the fan is configured in a way to allow the temperature variation but still keep the coolant a little cold, out of the ideal operation zone.

Following our analysis, we move on to the next variable that is represented by oil pressure as shown in Figure 33, Figure 34 and Figure 35.

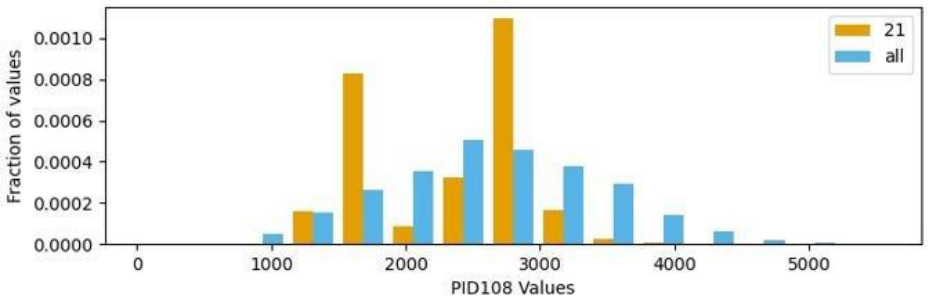

Figure 33 - Oil Pressure Data from Vehicle 21

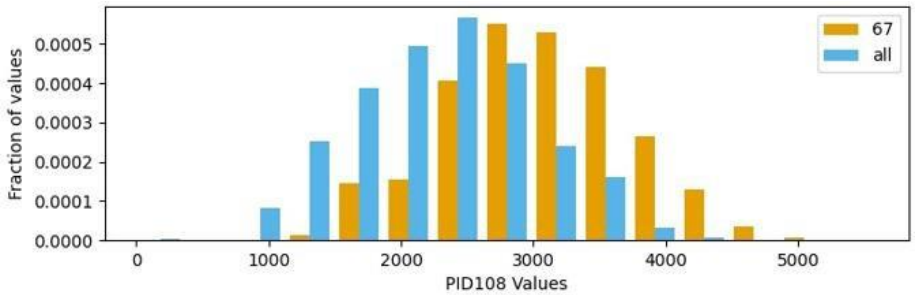

Figure 34 - Oil Pressure Data from Vehicle 67

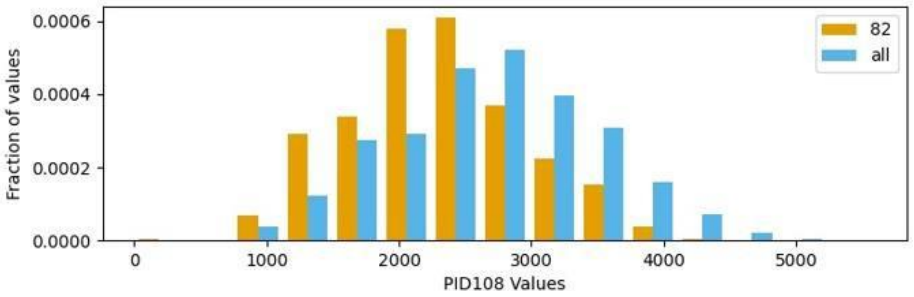

Figure 35 – Oil Pressure Data from Vehicle 82

Looking at Figure 33, Figure 34 and Figure 35 we see no significant differences in the ranges of values, with the working zone of the oil pressure being common to all vehicles. However, we can notice two peaks for vehicle 21 (Figure 33), which means, this data structure is bimodal, informing that this car has two frequent working zones. Considering that the ideal working zone for oil pressure is

between 2000 mbar and 3500 mbar, the frequency of values for vehicle 21 in the 1500 mbar zone may indicate a potential failure in the oil pump of this vehicle or pressure leaks.

The next analyzed parameter is the vehicle speed. However, it is important to mention that this variable is directly related to the type of operation, place of operation, and driver of the vehicle. Although speed is extremely important to calculate the OSF features, we cannot measure all the unknowns that influence this variable, since we have no control over the place where the vehicle will operate, how it will operate and by whom the vehicle will be driven. That is, the use of this variable is necessary for the development of the intelligent system, since the lubrication system and the engine work according to the driving style. However, their differences cannot indicate a potential failure like the other variables, since they are variables directly related to human behavior and not to the behavior of the machine itself.

In Figure 36 through Figure 38 we present the behavior of the variable vehicle speed.

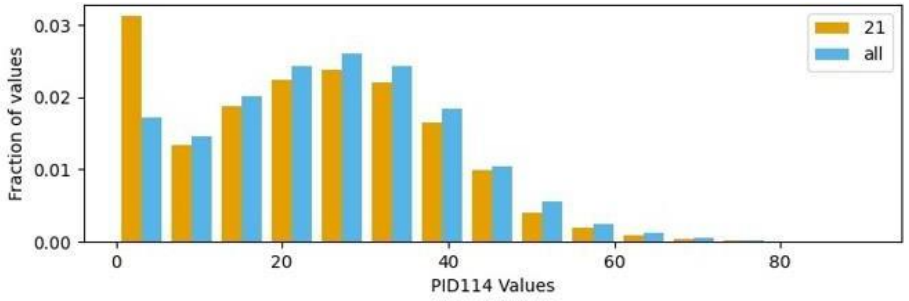

Figure 36 - Vehicle Speed Data from Vehicle 21

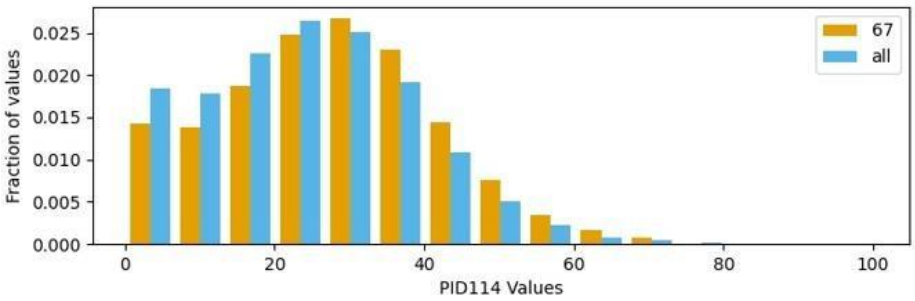

Figure 37 - Vehicle Speed Data from Vehicle 67

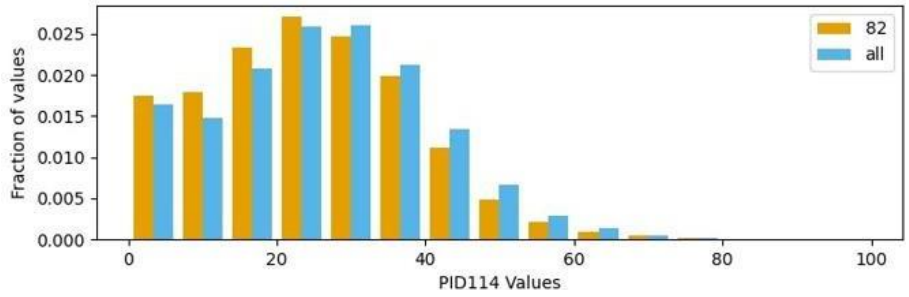

Figure 38 - Vehicle Speed Data from Vehicle 82

By analyzing the Figure 36, Figure 37 and Figure 38, we realize that all cars have the same range of values. Figure 36 shows that vehicle 21 operates more at low speeds than the other vehicles, which may indicate an operation in areas with speed control, such as near places with large flow of pedestrians.

Finally we will analyze the behavioural pattern of the engine speed in Figure 39, Figure 40 and Figure 41.

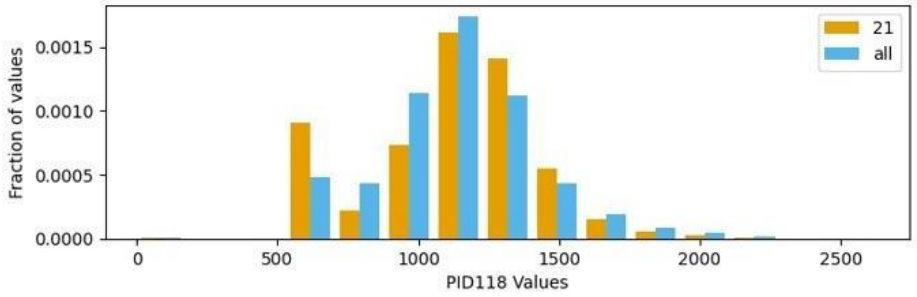

Figure 39 - Engine Speed Data from Vehicle 21

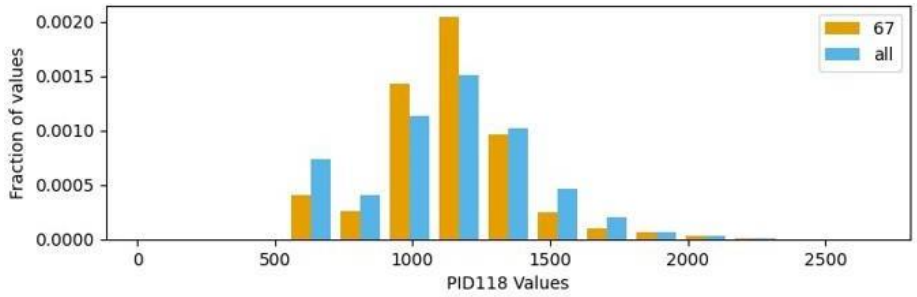

Figure 40 - Engine Speed Data from Vehicle 67

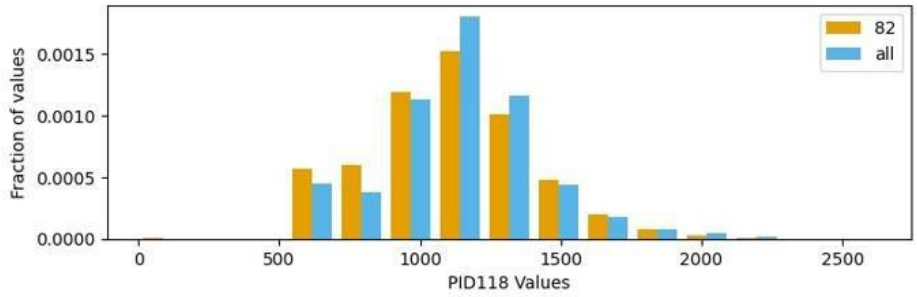

Figure 41 - Engine Speed Data from Vehicle 82

When analyzing Figure 39, Figure *40* and Figure *41*, we can see that the operation zone is the same in all vehicles, with vehicle 21 having some differences when compared to the other vehicles. This vehicle has a higher probability density in values near idle when compared to the other vehicles, which confirms the previous analysis that indicates that this vehicle usually operates at lower speeds, i.e., in stop and start zones, which can cause greater stress to the lubricant and the engine, and can also cause an anomalous heating pattern of the vehicle components (see Figure 18 and Figure 30). Another

important point to note is that all vehicles operate in a zone close to the Extra Economic Point, which is 1200 to 1500 RPM, on average, for heavy-duty vehicles. This point is important because it shows that drivers do not push their vehicles to the limit, i.e., they do not frequently reach the danger zone that goes from 2.200 to 2.500 RPM, which is the range that can cause irreversible damage to the engine, such as bending the valves in the cylinder head and even bend the connecting rods, because in this range the engine completely loses synchronism. In some types of electronic engines there are protection sensors that prevent this from happening by cutting off the fuel flow.

Thus, it can be concluded from this analysis of the behaviour of the variables that all variables have a certain level of asymmetry. Indicating that the data may not be normally distributed, however it indicates the range of values in which these parameters operate most of the time and confirms the limits and optimal values of operation. Another important point of this analysis is the direction we can go to potential problems, the types of operation, and the similarities and differences in the data for each vehicle, not to mention that we have already eliminated potential problems with the presence of outliers.

To confirm whether there was a strong or weak relationship between the distribution of values across months and vehicles, we applied the Kolmogorov-Smirnov test (K-S test). According to Lilliefors (1967), the application of the K-S test in comparing datasets provides a means of testing whether a specific dataset (sample) belongs to or resembles a dataset of some fully specified continuous distribution. It is important to mention that K-S test returns two values when applied to compare two samples and/or datasets:

- D value: quantifies a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution, or between the empirical distribution functions of two samples (Lilliefors, 1967);
- P-value: returns the possibility that we can reject the null hypothesis that the two samples were drawn from the same distribution which will hapen if the p-value is less than the significance level (Lilliefors, 1967).

According to this statement, and through the analysis of the graphs of Figure 42 to Figure 56 separated by variable, we can see in the x axis the values presented by each parameter and in the y axis the D values of the K-S test. As with the probabilistic density plot analysis, it is important to report that vehicles 30 and 81 are not being presented in this study because the data are similar to vehicles 67 and 82 (in the case of vehicle 30 similar to 67 and in the case of vehicle 81 similar to vehicle 82 - see Table 14).

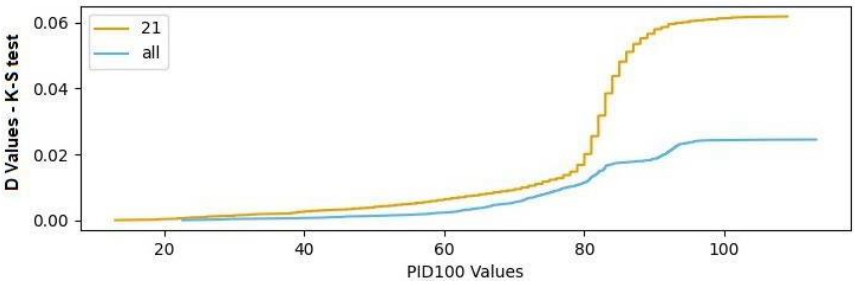

Figure 42 - Kolmogorov-Smirnov (K-S) Test of Oil Temperature Data from Vehicle 21

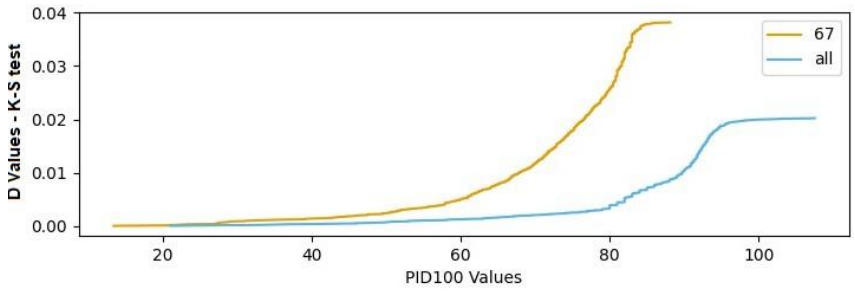

Figure 43 - Kolmogorov-Smirnov (K-S) Test of Oil Temperature Data from Vehicle 67

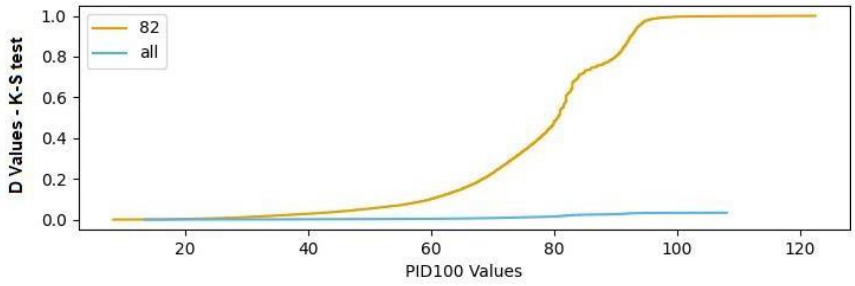

Figure 44 - Kolmogorov-Smirnov (K-S) Test of Oil Temperature Data from Vehicle 82

When we check the oil temperature, in Figure 42, Figure 43 and Figure 44 in vehicles 21, 67 and 82 respectively, we see a striking difference in relation to the D values in vehicle 82 (Figure 44), the maximum being approximately 0.97 in the zones close to 80°C, in comparison to 0.031 in vehicle 67 (Figure 43) and 0.039 in vehicle 21 (Figure 42). This kind of difference is directly linked to the acquisition mode of this parameter. This vehicle, despite having the same technical characteristics as vehicle 80, has a different data acquisition configuration and therefore the sample rate of this parameter in this vehicle is much higher than that of the other selected vehicles. This type of data acquisition configuration is determined internally by Stratio automotive in conjunction with the vehicle owner and cannot be changed as it is a business rule. Next, we will look at Figure 45, Figure 46 and Figure 47 which show the D values obtained from the coolant temperature data.

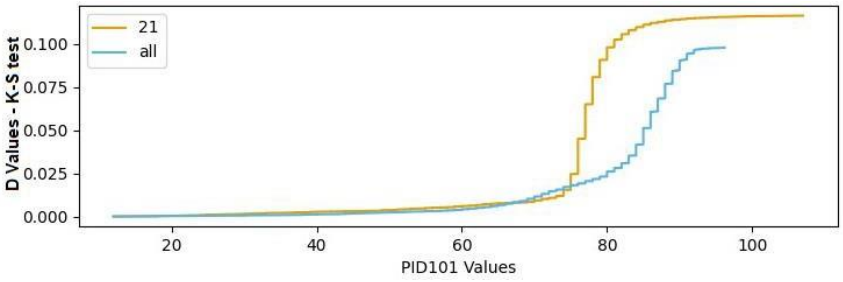

Figure 45 - Kolmogorov-Smirnov (K-S) Test of Coolant Temperature Data from Vehicle 21

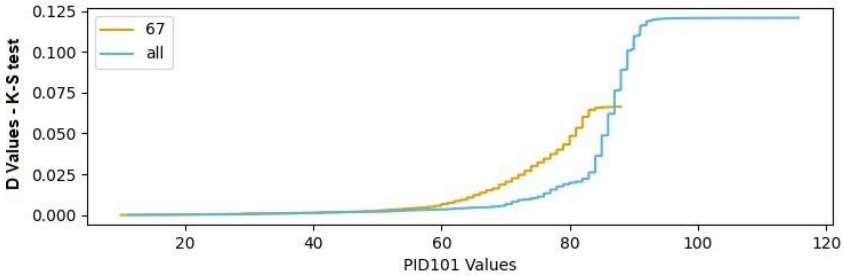

Figure 46 - Kolmogorov-Smirnov (K-S) Test of Coolant Temperature Data from Vehicle 67

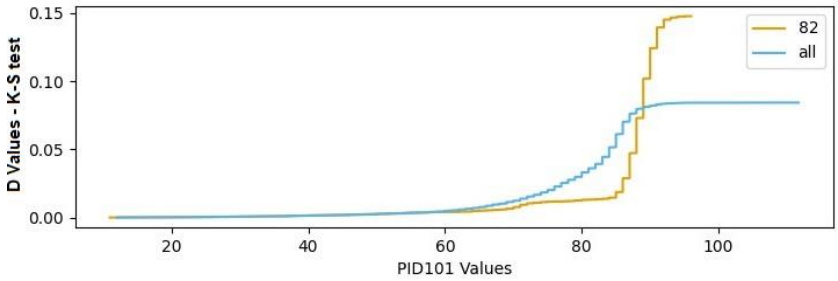

Figure 47 - Kolmogorov-Smirnov (K-S) Test of Coolant Temperature Data from Vehicle 82

When we analyze Figure 45, Figure 46 and Figure 47 we notice that the distances between the distributions which correspond to the D value, reduces in comparison with the values obtained in Figure 42, Figure 43 and Figure 44. This shows that this parameter was configured and has a similar acquisition rate in all selected vehicles. This reduction in value indicates that even when comparing data from vehicles with different brands and models, the data have similarities and proximity in their distributions.

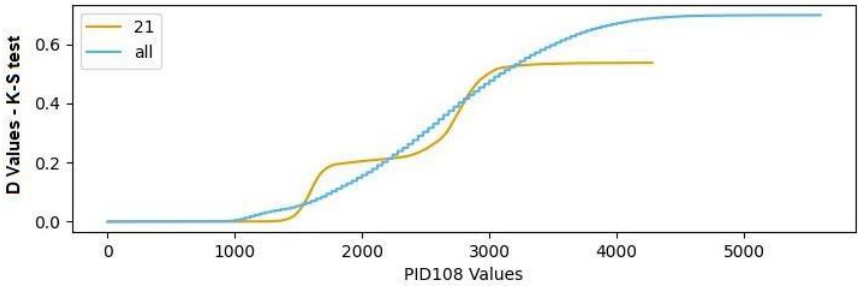

Figure 48 - Kolmogorov-Smirnov (K-S) Test of Oil Pressure Data from Vehicle 21

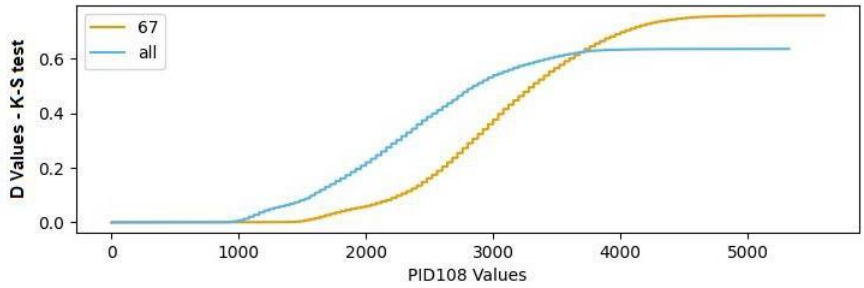

Figure 49 - Kolmogorov-Smirnov (K-S) Test of Oil Pressure Data from Vehicle 67

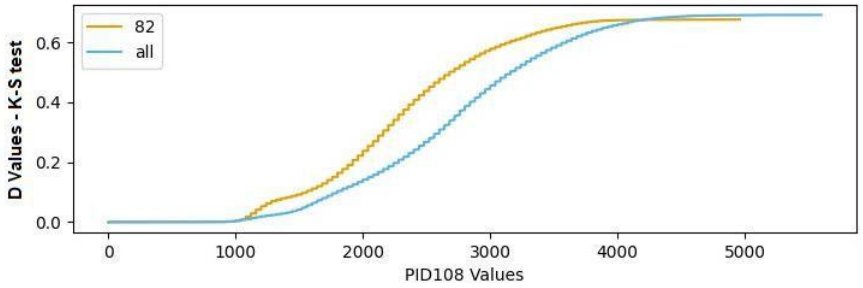

Figure 50 - Kolmogorov-Smirnov (K-S) Test of Oil Pressure Data from Vehicle 82

By analyzing the Figure 48, Figure 49 and Figure 50, which correspond to the comparison of the data distributions for oil pressure, we can indicate the same conclusion obtained from the analysis of Figure 45, Figure 46 and Figure 47, i.e. the data have similarities and closeness in their distributions.

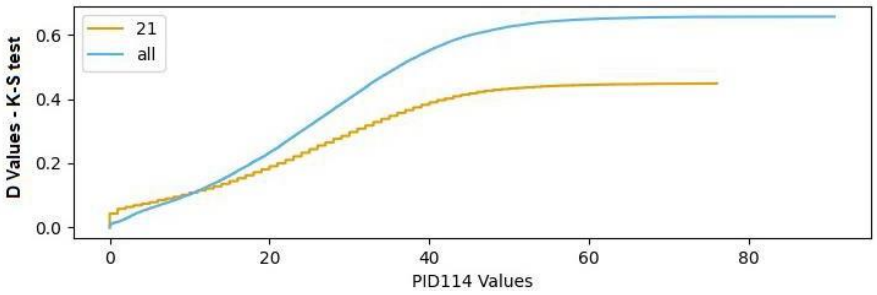

Figure 51 - Kolmogorov-Smirnov (K-S) Test of Vehicle Speed Data from Vehicle 21

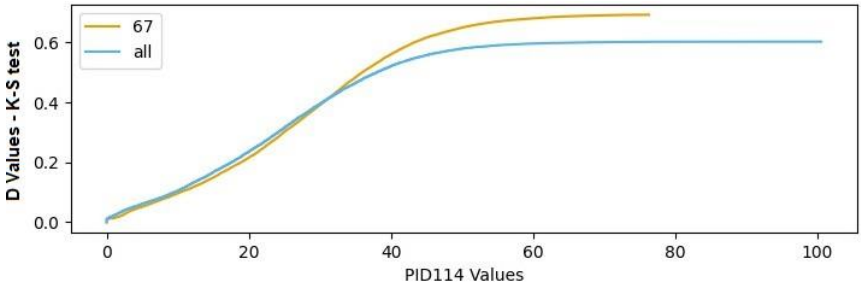

Figure 52 - Kolmogorov-Smirnov (K-S) Test of Vehicle Speed Data from Vehicle 67

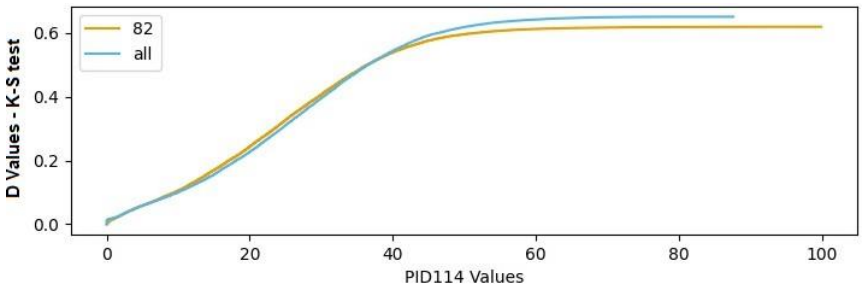

Figure 53 - Kolmogorov-Smirnov (K-S) Test of Vehicle Speed Data from Vehicle 82

Like in the analysis performed on Figure 36, Figure 37 and Figure 38, it is important to mention that when analyzing Figure 51, Figure 52 and Figure 53, we must take into consideration that this is a parameter that suffers direct influence of the type of operation, place of operation and driver. In this way, we can only conclude from the analysis of the D values of vehicle speed, that the distributions are similar to each other and can be considered from the same dataset.

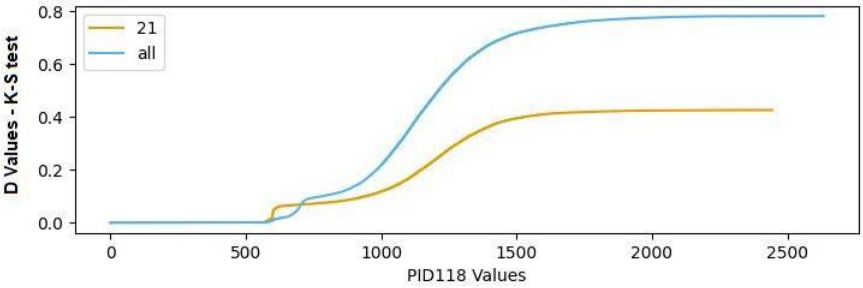

Figure 54 - Kolmogorov-Smirnov (K-S) Test of Engine Speed Data from Vehicle 21

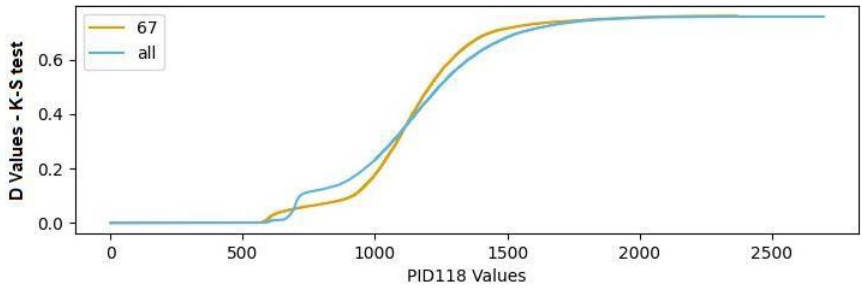

Figure 55 - Kolmogorov-Smirnov (K-S) Test of Engine Speed Data from Vehicle 67

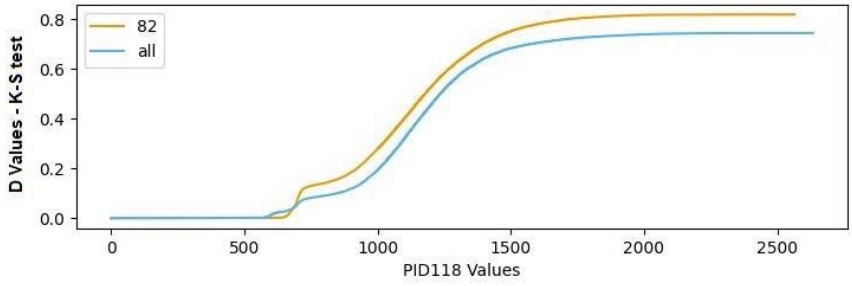

Figure 56 - Kolmogorov-Smirnov (K-S) Test of Engine Speed Data from Vehicle 82

Finally, by analyzing the Figure 54, Figure 55 and Figure 56, we can confirm that the data samples are very close to the same distribution. This type of analysis should be performed for each selected parameter because it indicates if we will have to work on the data to make them follow the same patterns. By having similar data we can apply calculations and metrics that are transversal, without worrying about creating specific analysis means for each parameter and/or vehicle. Thus, it is important to present the Table 30 that indicated all the P and D values for the comparisons made with the vehicle data:

vid	Variable	K-S test P value	K-S test D value
67	Oil Temp.	1.87e-11	0.03
	Coolant Temp.	2.42e-34	0.05
	Oil Pressure	0.00	0.22
	Vehicle Speed	8.74e-94	0.09
	Engine Speed	1.21e-56	0.07
80	Oil Temp.	0.00	0.96
	Coolant Temp.	4.37e-17	0.03
	Oil Pressure	0.00	0.19
	Vehicle Speed	1.84e-36	0.05
	Engine Speed	8.85e-303	0.16
21	Oil Temp.	4.60e-18	0.03
	Coolant Temp.	5.11e-64	0.07
	Oil Pressure	3.13e-298	0.16
	Vehicle Speed	0.00	0.20
	Engine Speed	0.00	0.35
34	Oil Temp.	5.54e-143	0.11
	Coolant Temp.	1.18e-44	0.06
	Oil Pressure	0.00	0.37
	Vehicle Speed	1.97e-262	0.15
	Engine Speed	0.00	0.48
82	Oil Temp.	0.00	0.96
	Coolant Temp.	1.84e-46	0.06
	Oil Pressure	0.00	0.17
	Vehicle Speed	1.90e-12	0.03
	Engine Speed	3.62e-85	0.08

Table 30 - D Value and P Value from the Kolmogorov-Smirnov (K-S) Test

Through the analysis of the Table 30 we can confirm that only Oil Temp. in vehicle 82 has a high D value and all others remain constant. Another important point through the analysis of the table is in relation to the P value, as the p-value is less than 0.05 in all comparisons, we reject the null hypothesis. Thus, we have enough evidence to say that the samples do not come from a normal distribution. This conclusion is important to avoid performing experiments using parametric models. Given this, the assessment of the normality of the data distribution is paramount for the proper description of the sample and its inferential analysis. Furthermore, it is normal that data sets collected

outside controlled testing environments present non-normal distributions, especially in events of high variability, with standard deviation greater than half of the mean value contraindicating the use of statistical techniques intended for normal samples, under penalty of biasing the parameters and the inference of the tests. Even the increase in sample size will not supplant the estimation errors caused by the use of distributions inappropriate to the analysis techniques. In addition to concluding the normality of the data, by analyzing the two values provided by the test (D-value and P-value), we confirm that there are no plausible divergence problems in the data distribution that would justify their removal or replacement in the creation of an intelligent system for predicting the operating state of lubricating oils.

5.3 Sample Distribution

When analyzing the data collected from the lubricating oil one must take into account that most of the time it is in good working conditions, with the vehicle operating without faults over a long periods of timel. This situation results in a highly unbalanced dataset, with our dataset having only 10% of samples where the lubricating oil was not in good conditions (see Figure 57). According to Prytz (2014), the importance of balancing the data lies in the fact that many classification models are designed to work with roughly the same amount of samples for each class Additionally there are several other degrading factors of classification performance that are associated with unbalanced data.

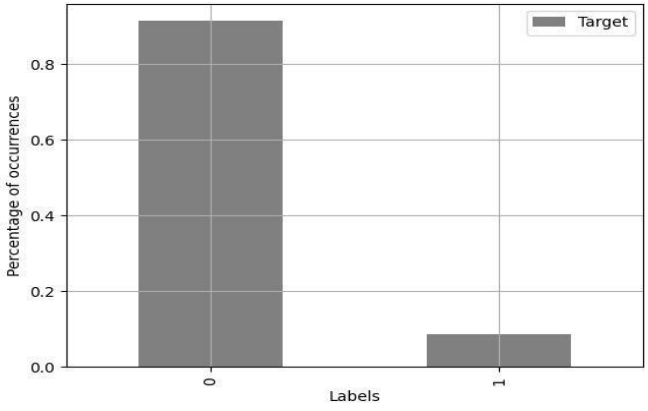

Figure 57 - Percentage of Labels in an Unbalanced Dataset

According to Mitchell (1997), if we develop a system without considering this disproportionality in the data, the model will fall victim to the Accuracy Paradox[6], where the algorithm parameters will not differentiate the minority class from the other categories, despite having an high accuracy. This lack of differentiation can cause serious problems since the identification of these minority cases can be the target of the challenge to be solved. For example example, consider our problem: If our model does not successfully differentiate the positive diagnoses of lubricant system

[6] The Accuracy Paradox is by definition a contradictory situation in which a high accuracy in your classification system may highlight a failure of your own system to make meaningful predictions.

failure, classifying them as negative cases (majority class), we would be developing a system that instead of predicting wear and maintaining the health of the equipment would be opening margins for excessive wear and consequently a catastrophic failure, which would result in additional costs. Thus, according to Prytz (2014), if there is little data from which to learn, a predictive algorithm may have difficulty in extracting and learning important patterns that discriminate problem classes and establish a reliable decision boundary, which directly interferes with the model's performance in making future generalizations about unknown data samples.

According to Dogan & Birant (2021), methods based on sampling (resampling) are the simplest and oldest methods used for balancing datasets. They consist of modifying the structure of the unbalanced data set so that it has equivalent amounts of samples for the classes present, either by removing (undersampling) or adding (oversampling) new samples. These types of methods are also known as data-level approaches, since the entire solution for balancing the data is done directly on the dataset, isolating the classification models, which just receive the already balanced data for training. Since this process usually occurs before that classification models are trained, sampling methods are related to the data pre-processing step.

To address this issue, we use a stratified data balancing technique, which consists in the initial creation of two datasets: a first dataset containing only the data that has the lubricant operating condition label and a second dataset with all the occurrences of the non-operating condition label and nearby situations. Next, we check how many samples we have in the second dataset, which contains the non-operating condition data points, and randomly extract the same amount of situations from the first dataset. Thus, we insert that amount of good lubricant operating condition points extracted from the first dataset into the second dataset that previously only had non-operating condition situations and form a dataset with the same amount of points that determine good lubricant operating condition and non-operating condition. After this step we will have a balanced dataset with two operating situations needed to create the model.

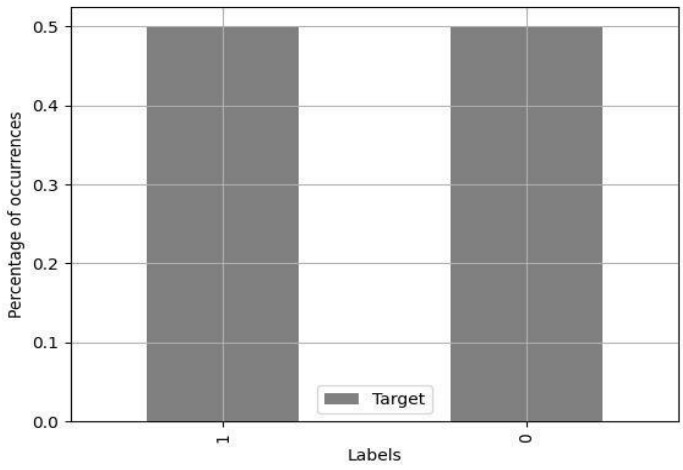

Figure 58 - Percentage of labels in a balanced dataset

According to Figure 58, we can see that the final distribution of samples, bringing a data structure very close to the ideal (50% of each class). It is important to mention that according to Prytz (2014), data balancing, as well as variable selection, missing values imputation and the study of the behavior of variables with verification of noisy samples, are fundamental steps that must be implemented by the data analyst who is concerned with the reliability of his classification products.

5.4 Principal Component Analysis (PCA)

PCA is an unsupervised dimensionality reduction technique that constructs relevant features/variables through linear (linear PCA) or non-linear (central PCA) combinations of the original variables. PCA, is a mathematical procedure to convert a set of observations of possibly correlated variables into a set of linearly uncorrelated variable values called Principal Components (PC's) (Rodrigues et al., 2020). That is, an orthogonal transformation of the data into a series of uncorrelated data that live in the reduced space of the PCA, such that the first component explains the most variance in the data with each subsequent component adding more information. Thus, to evaluate this variation in our dataset we drawn from the PCA analysis the cumulative variance of the parameters, which determines the loss of information with the simplification of variables. According to Rodrigues et al. (2020), the cumulative variance is calculated by decreasing order of importance.

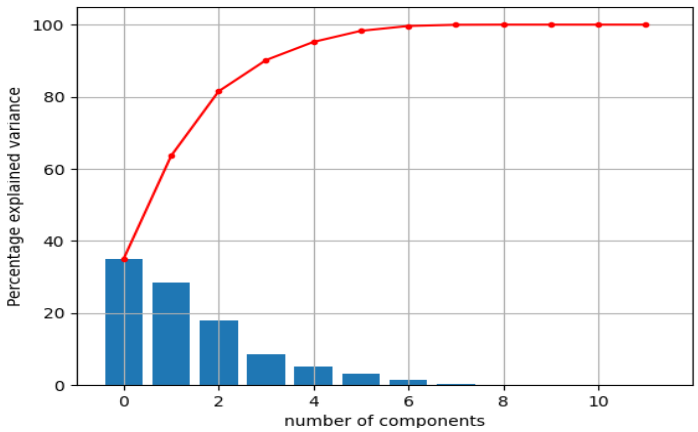

Figure 59 - The Cumulative Explained Variance Ratio as a Function of the Number of Components

According to Figure 59, the blue bars show the percentage of the variance explained by each principal component and the red line shows the cumulative sum. From the graph, we can see that the first principal component explains 35% of the variance in the dataset. When we consider the first and the second, principal components we can explain 63% of the variance, and so on. According to W. Wang & Zhang (2005), we can consider that the cumulative variance calculation is adequate to evaluate the relationships between the variables, since it explains a large part of the variability of the data, and generally an explanation greater than 50% in the first two components is required to use this type of

analysis. So, following this analysis, the chart in Figure 59 shows that we need only 4 of the 9 principal components to explain 90% of the variation in the original data. But it is necessary to evaluate each principal component in detail, examining the magnitude of the coefficients of the original variables. That is, the higher the absolute value of the coefficient, the more important will be the variable corresponding to the principal component calculation.

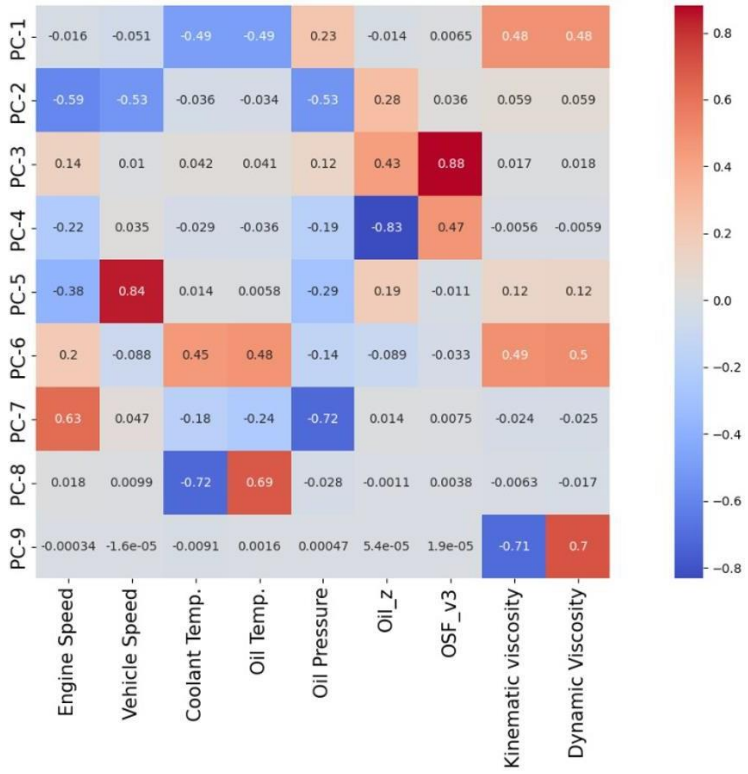

Figure 60 - Effect of Variables on Each Principal Component

Figure 60 shows the relation between each principal component and a variable, which can also be called factor loading. This type of analysis makes it easy to grasp the dimension behind a component. The overall dimension would be the optimal lubricating capacity of the equipment, and a component strongly correlated with these variables can be interpreted as the dimension of determining the rate of degradation of the lubricating oil, i.e., the loss of lubricating capacity of the equipment. The importance of each characteristic is reflected by the magnitude (no matter the sign) of the corresponding values in the eigenvectors.

Looking at values inside the heatmap (Figure 60), for principal component 1 (PC-1) the most important variables are: Kinematic viscosity (visc_cin), Dynamic viscosity (visc_din), Oil Temp. and Coolant Temp. Also, we can see that the characteristics engine speed, vehicle speed and Oil_z are the most important for PC2. This analysis of importance of the variables by component is important because it indicates which groups of variables impact each other and consequently help understanding the lubrication system behavior as a whole. In other words, when we analyze the most important variables for PC-1 (Kinematic viscosity (visc_cin), Dynamic viscosity (visc_din), Oil Temp. and Coolant Temp.)

we notice that these variables can be considered as a group of data from the lubricant. When we perform the same analysis for PC-2 (Engine Speed, Vehicle Speed and Oil Pressure and Oil_z) we can consider that this component concerns the system where the lubricant is inserted.

Looking closer at the way each variable affects the principal components, we can see that PC-1 has a negative association with the lubricant temperature and the coolant temperature. On the other hand, this component correlates positively with the kinematic and dynamic viscosity. In summary, it is intrinsic that these characteristics are extremely important in determining the rate of lubricant degradation and that, going in opposite directions, they show evidence that temperature is the primary influencer of viscosity and that as temperature increases, viscosity decreases, thereby reducing the lubricating capacity of the oil. Following this line of reasoning, PC-2 is more related to the characteristics of the equipment itself, showing the importance of a complete view of the equipment and the lubrication system in general, not just focusing on the product (lubricating oil) to determine the degradation index.

To have a better understanding of the distribution of samples by class, we developed Figure 61 which shows our dataset considering only the two Principal Components (PC-1 and PC-2). In concrete, samples that correspond to lubricant failures were marked in red and samples that correspond to good lubricant operating condition were marked in green.

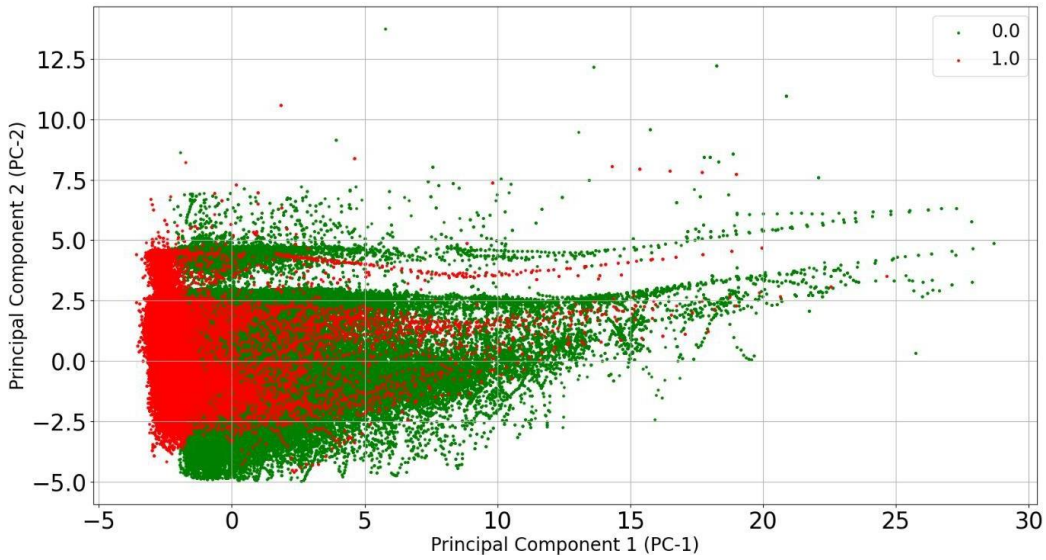

Figure 61 – Principal Component Analysis (PCA) for Two Principal Components

From the analysis of Figure 61, it is possible to identify that in the interval from -4 to -2 of the PC-1 axis, there is a clear separation between labels 0 and 1, which indicates that 18% of the points do not overlap. But for values of PC-1 greater than -2 we can see that some samples from the two classes overlap, representing the largest portion of the data points (about 82%). This situation show that using only these two principal components an ML model will have some difficulties to distinguish between

situations where the oil is good for operation and situation where it is not, indicating that we might need to rely on more variables. Building on this analysis, we can provide the principal axes in the problem space, representing the directions of maximum variance in the data. This means that we can see influence in each of the components by feature.

To help us with this analysis, we will use the Biplot graph, which is a visualization model that contains the angulation and projections of the features with respect to the problem definition:

1. PCA scatter plot which shows the first two components (Figure 61);
2. PCA loading plot which shows how strongly each characteristic influences a principal component (Figure 62).

It is important to mention that all vectors start at the origin of the referential and the projected values on the components corresponds to their weight on that component. Also, the angles between the individual vectors indicate correlation between them. In classical analysis, the size of the vector is proportional to the variance of the variable and the cosine of the angle between two vectors is the correlation between the variables.

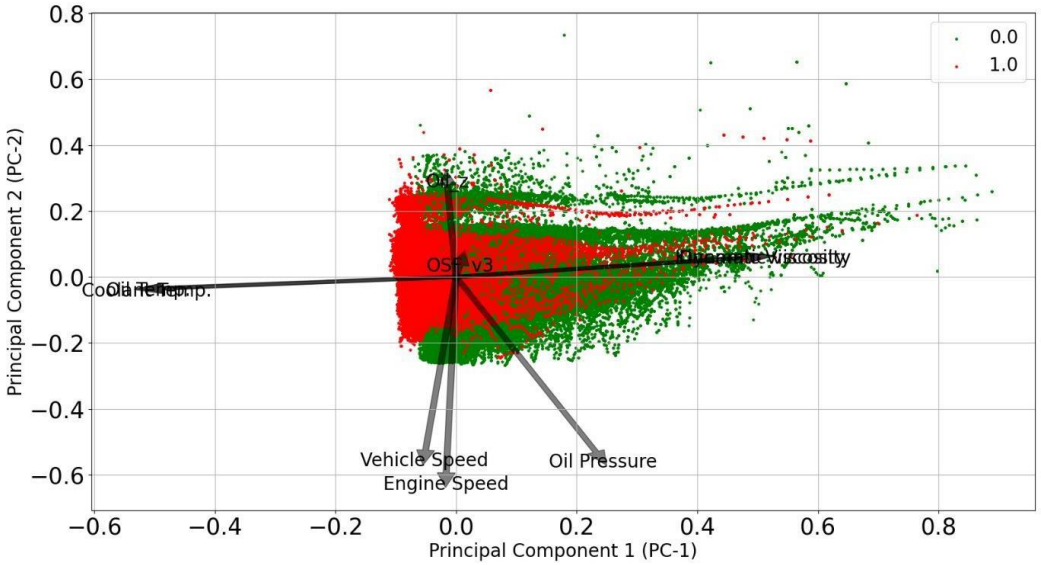

Figure 62 - Principal Component Analysis (PCA) Biplot Graph

Before starting the analysis of Figure 62, it is noticeable that there is a difference in the positioning of the data points in relation to Figure 61. This is due to the fact that to improve the visualization of vectors and variables, it was necessary to use factors of scale following the equation:

$$PC_{Factor} = \frac{1}{x_{max} - x_{min}} \tag{7}$$

Where, PC_{Factor} is the scale factor of the principal component, x_{max} is the maximum value that this component reached during the PCA analysis and x_{min} is the minimum value.

Looking at Figure 62, we can see that the arrows (variables/characteristics) pointing in the same direction indicate a correlation between the variables they represent, while the arrows pointing in opposite directions indicate a contrast between the variables they represent. It is possible to notice that the length of the arrows is different for each variable and this represents how much variance this variable explains in the factorial plane. According to W. Wang & Zhang (2005), this can be called the representation quality of that variable in the plane. It is also important to mention that the angle between the variables gives an indication of how well the variables are correlated.

1. A small angle indicates that the representation of the two variables in the factorial plane are positively correlated. Above we see that oil temperature and liquid temperature are positively correlated in this factorial plane;
2. An angle of 90 degrees indicates that there is no correlation. Above we can see that temperatures and Oil_z have little correlation;
3. An angle of 180 degrees indicates a negative correlation, just as we see to temperatures and viscosities;
4. The variables considered most representative, are those that form smaller angles with the line that presents the circle formed with the arrow, that is, temperatures and viscosities.

We can conclude from this analysis that if we are viewing a multivariate dataset in a high dimensional space, with 1 axis per variable, PCA can be used to provide a lower dimensional view of the same data, a shadow of the original object when viewed from its most informative point. This is done using only the first principal components, so that the dimensionality of the transformed data is reduced. By performing the PCA analysis we were able to determine how many components would be necessary (see Figure 59) to represent the information in the dataset, the importance of the selected variables (see Figure 60), and ascertain if these variables could be used for the development of an automatic lubricant condition determination system (see Figure 61 and Figure 62). All these conclusions are important because they will determine the obtaining of valid results of the system and its implementation in the real environment.

5.5 Correlation Analysis

To further understand the existing relationships between the variables being used, we performed a correlation analysis. The correlation coefficients are helpful in studies with many related variables because they provide some information that help us understand how the variability of one variable affects the other.

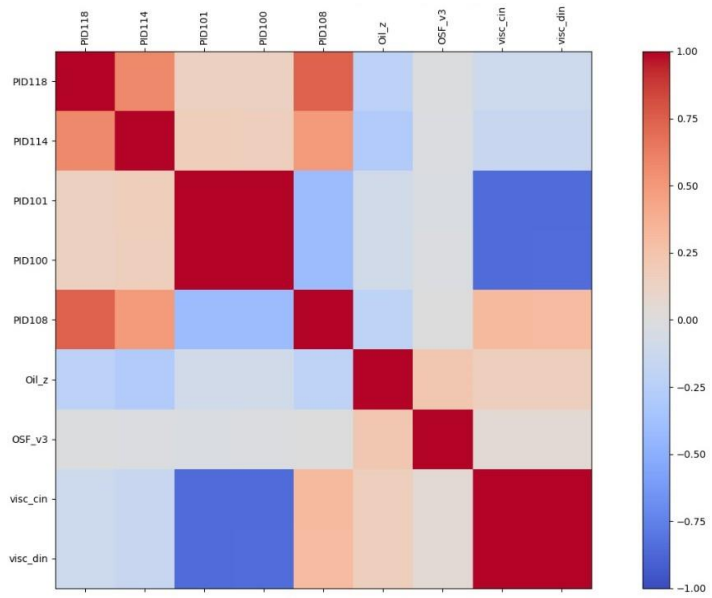

Figure 63 - Pearson Correlation for the Lubricant Operation Condition (Label 0)

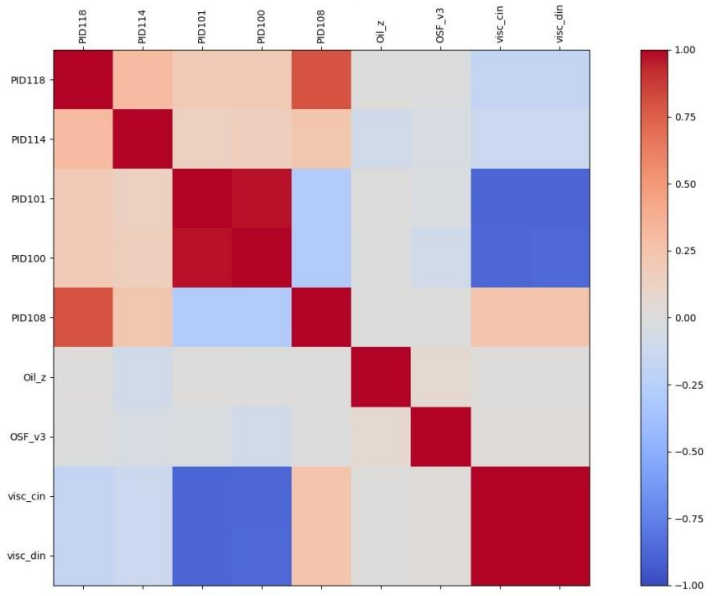

Figure 64 – Pearson Correlation for the Non-Operational Condition of the Lubricant (Label 1)

Figure 63 and Figure 64 show a visual representation of the Pearson correlation between every pair of variables. The former represents the correlation for situations where the oil is in good conditions (Label 0), whilst the latter shows the correlation between the variables when the oil is not in good conditions (Label 1). Looking at the results presented in the figures, we can observe the following:

1. Oil_z in relation to all the other variables, has correlation values for label 0, something that is almost not identified for Label 1. This information determines that the calculated variable Oil_z is an important and decisive variable for determining the oil's non-operating condition, since it is related to all the variables selected for solving the problem;
2. Oil pressure in relation to oil and liquid temperature, decreases its correlation value when it change from label 0 to label 1. This information indicates that the control for the

relation of the amount of oil necessary for complete lubrication of the engine, according to its operating temperature, is no longer fully functioning and may not be sufficient;

3. Oil pressure versus engine speed increases its correlation value when it changes from label 0 to label 1. This information determines that the engine needs to rotate faster to try to meet the lubrication demand;
4. The viscosities in relation to the engine speed decrease their correlation values when transiting from label 0 to label 1. This information determines that the film required for optimal engine lubrication at that speed is no longer being achieved;
5. Oil pressure versus viscosities increase their correlation values when transitioning from label 0 to label 1. This information determines the need for greater lubricant pressure to meet lubrication needs, since oil degradation causes it to lose its properties, viscosity being one of the most important oil properties.

Even though Pearson correlation values are relatively higher, we decided to apply one more correlation method to analyze the differences. According to Baak et al. (2020), we applied the recently developed Phik correlation (Figure 65 and Figure 66), which has been shown to efficiently capture non-linear dependencies. The Phik correlation is obtained by inverting the statistics of the chi-square contingency test, thus allowing users to also analyze the correlation between numeric, categorical, interval and ordinal variables.

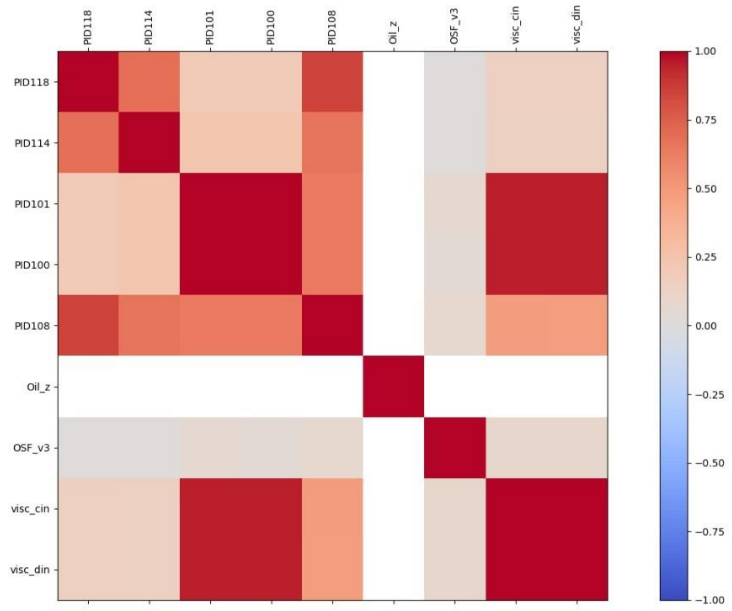

Figure 65 – Phik Correlation for the Lubricant Operation Condition (Label 0)

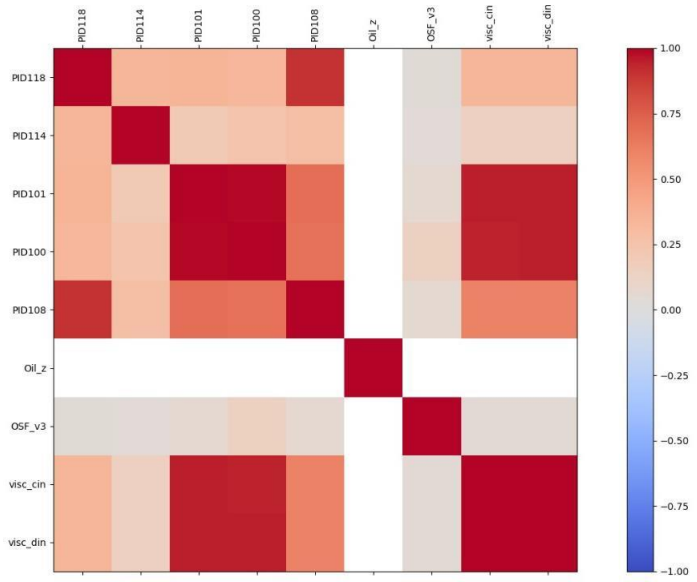

Figure 66 – Phik Correlation for the Non-Operational Condition of the Lubricant (Label 1)

Unlike Pearson's correlation coefficient which range from -1 to +1, where ±1 indicates perfect agreement or disagreement, and 0 indicates no relationship, the coefficient phik has a value in the interval [0, 1], which is determined by the distribution of the pair of variables. The values for correlation levels are 0 for no association and +1 for complete association. Through the analysis of Figure 65 and Figure 66, we can see an increase in the correlation values of most variables. These results suggest that there is an association between the variables, but some of them could be non-linear. This correlation method, since these variables, as seen in the other methods, are of paramount importance for the construction of the system for identifying the operating condition of lubricating oils in diesel engines.

We are also interested in looking into which variables may or may not be relevant as input to develop a model (see Figure 67).

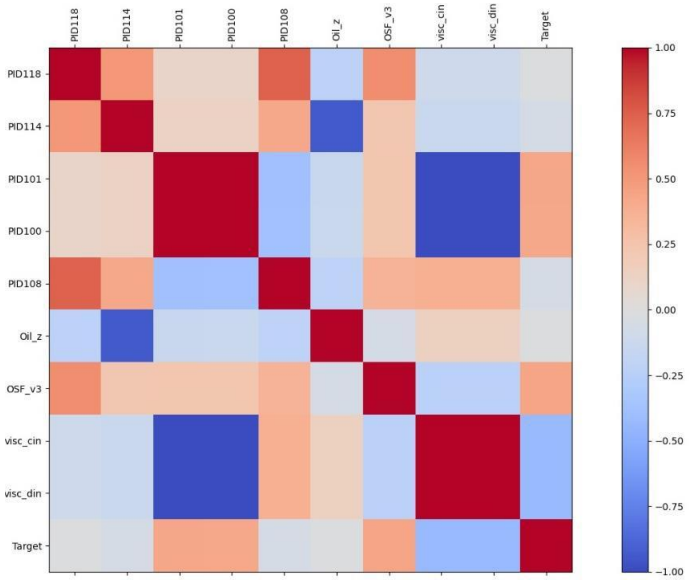

Figure 67 - Correlation Between the Input Variables with the Output Variable

By looking at Figure 67, we can see from this correlation graph that the variables oil temperature and liquid temperature have a high correlation value with the target, which is identified because the operating temperatures are directly linked to the calculations made by the electronic control unit for the engine's lubrication needs, and the higher the engine's operating temperature, the lower the kinematic viscosity and the worse the oil's lubrication power. Another important point to mention is that due to the high temperature relationship, there is consequently a correlation value of the kinematic and dynamic viscosity with the target, even if to a lesser extent. These viscosity correlation values are due to the lubricant properties, and viscosity control is one of the most important factors in determining the lubricant's operating condition. It is worth remembering that this correlation analysis is important for the selection of the algorithm that will be used. For example, a feature may be considered important for a model that considers linear relationships such as Linear Regression, while it is not important for a model that can identify non-linear relationships such as Decision Trees and Random Forest.

Chapter 6 - Automatic Oil Classification: Model Design and Validation

An important aspect to be considered when developing an intelligent system is the algorithm that will serve as the basis for solving the problem. To answer this question, we need to take into account:

- What kind of answer do we want, i.e., do we want a real number or a category?
- How is the input data scaled?
- Do we have labels?
- Do we want black or white box model that allows an expert to be able to determine if the decision path followed is correct?
- What are the computational resources that we have at our disposal?

We need a model that can handle the characteristics of the data we have, but we also have to take into consideration the specific situation in which it will be deployed, namely assisting a human operator in the decision making process. As such, we need to choose models that, are able to provide the human expert with a possible explanation of the decision process. Finally, and not less important, we have to take into consideration the computational resources available to run the model.

Thus, taking these requirements into account, we selected the supervised machine learning algorithm RF (Breiman, 2001) as our first prediction model. According to Keartland & Van Zyl (2020), RF is an ensemble classifier that results from the combination of multiple Decision Trees (DT). In the standard RF model, we select a limited random number of features from the training set, and use this subset (i.e., a bootstrap) to create a complete DT. This procedure is repeated for as many DTs as we include in the RF. The RF decision is usually performed using the most popular class predicted by each DT. By limiting the number of features that can be selected for the bootstrap and by creating shallower DTs, we reduce the computational load required allowing the RF model to be deployed in environments with modest computational capabilities.

According to Jayabharathi & Ilango (2021), DTs and RFs are considered to be some of the simplest algorithms to implement and fastest to obtain results. Although an RF is a set of decision trees, there are differences between each algorithm: while decision trees generate rules and nodes by calculating information gain and gini index (a measure of inequality) using all features, RFs generate decision trees using a random subset of characteristics, thus increasing responsiveness. Furthermore, while larger DTs (deep trees) may suffer from overfitting problems, RFs avoid this by working with random subsets of features, and by building smaller trees from these subsets. Thus, our choice of the RF algorithm as the first evaluation is based on the following:

1. RFs can be used for both classification and regression and is also one of the most flexible and easy-to-use algorithms;
2. RFs shows good results on various types of problems;
3. RFs have implementations in the most diverse artificial intelligence libraries available today.

In addition to the points mentioned, RFs have shown good results in a variety of applications, such as recommendation engines, image classification, feature selection, rank loan applicants, identify fraudulent activities, and predict failures in a mechanical system (Hanafy & Ming, 2021).

6.1 Model Evaluation Metrics

Assessing the performance of a ML algorithm is an essential part of any project. A model may achieve satisfying results when evaluated using a specific metric, but otherwise could be not enough for application in real world. For this type of situation, it is necessary to select a group of metrics that are representative of the problem one is dealing with and that reflects the real conditions of the environment where the model will be deployed (Sharma & Gandhi, 2008). Given our specific situation, and the imbalanced nature of the problem at hand, we selected the following evaluation metrics to assess the performance (Hanafy & Ming, 2021):

1. Precision score;
2. Recall score;
3. F1-score;
4. ROC curve with AUC.

The result obtained by the evaluation metrics reflects the quality of the model developed, so if they are poorly chosen, it will not be possible to evaluate whether the model is in fact meeting the necessary requirements. For example, there are cases in determining the condition of the lubricant where different errors have different costs, and the metric calculation must reflect this difference.

6.2 Results of the First Model

Before presenting the results, It is important to mention that in our experimental study we rely on the Python module scikit-learn (sklearn) implementation (Pedregosa et al., 2011) of the RF classification algorithm, with 200 individual DTs of maximum size 4 and with 4 features in each bootstrap. For all the other remaining parameters we used the sklearn default values (scikit-learn.org, 2021).

For our validation procedure, we use the cross-validation method called K-fold, which consists of dividing the total data set into k mutually exclusive subsets of the same size, and from there, one subset is used for testing and the remaining subsets are used for validation and parameter estimation,

calculating the accuracy of the model on the errors found. This process is performed k times by alternating the test subsets in a circular fashion.

Therefore, we take our original dataset and divide it into subsets, varying k from 2 to 6 and comparing the associated average error, in order to determine the best number of subsets or k values:

K-fold	Average Error
2	17.30%
4	14.70%
6	13.60%

Table 31 – K-Fold Average Error

According to the Table 31 the model that had the lowest associated average error was with 6-Fold, thus justifying the choice of the optimal split value for testing. This method allows us to evaluate the performance of the model using different training and test datasets keeping the computational costs affordable. The results obtained by the prediction model using the RF algorithm, with the parameters mentioned above, are summarized in Table 32, where each row shows the average values obtained for Recall, F1-Score and Precision.

Metrics	Value
Recall	95.90% (+/- 0.90)
Precision	97.50% (+/- 0.69)
F1-score	96.70% (+/- 0.50)

Table 32 - Results Obtained with the Random Forest Model Using 6-Cross Validation

Looking at the results shown in Table 32, we can identify some important signals that the model could be used to determinate the operation condition of the lubricant oils in diesel engines. The high score of the recall (95.90%) demonstrate the ability of a classification model to identify a large portion of relevant instances, whilst the high score of the precision (97.50%) show the ability of a classification model to return only relevant instances and the F1 score (96.70%) that combines recall and precision using the harmonic mean confirm our good perception about the performance of the model. In summary, a detailed inspection of how each instance is being classified by the model reveals that it is able to correctly identify all the instances where the oil is good for operating conditions (label 0). However, it fails to correctly classify a small number of instances with label 1, i.e., it is not able to identify some situations when the lubricating oil should be changed because it is no longer good for operating conditions. This is not surprising when one takes the imbalanced nature of the problem at hand and the results of the PCA analysis (Figure 61), which revealed that for some cases there is a small overlap between samples of the two classes. Looking at the results obtained in the precision metric, we can see that they are slightly higher which confirms that the model is identifying most of the situations where a vehicle has an oil in the lubricating system that is not in good conditions. Finally, it is important to refer the low values of standard deviation for all the metrics, which are an indication of robustness

model. In Figure 68 we show the average ROC curve obtained using the same 6-fold cross validation discussed earlier.

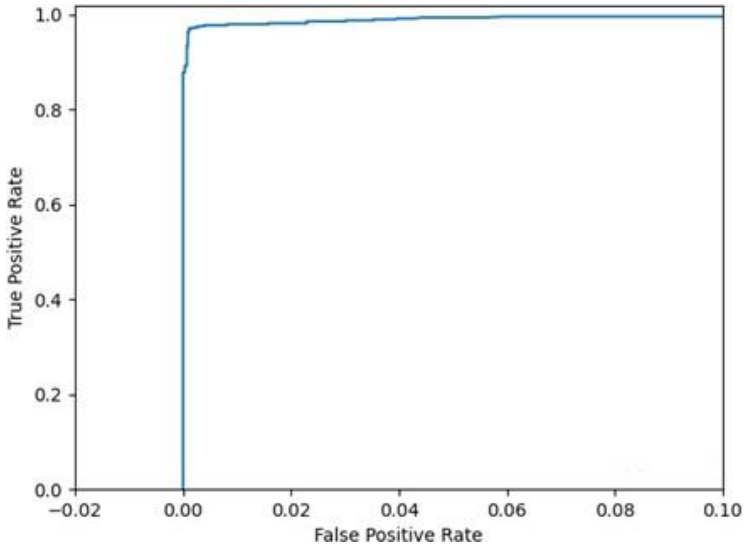

Figure 68 - ROC Curve Obtained by our Model, Using a 6-Fold Cross-Validation

By looking at Figure 68, we can confirm what we have discussed previously, showing that we have a high true positive rate. This is a remarkable result (AUC value of 0.98), given the nature of our problem, where accurately identifying the situations where the oil is gone bad is of the utmost importance to avoid catastrophic failures. We also conducted an analysis to understand which of the used features are the most relevant to distinguish the operating conditions. This study will allow us to verify if the additional features that we developed (i.e, Oil_z, OSF, Kinematic Viscosity and Dynamic Viscosity) are useful. To compute the importance of the features, we used the normalized Gini importance metric. Figure 69 depicts the results of the 9 most important features.

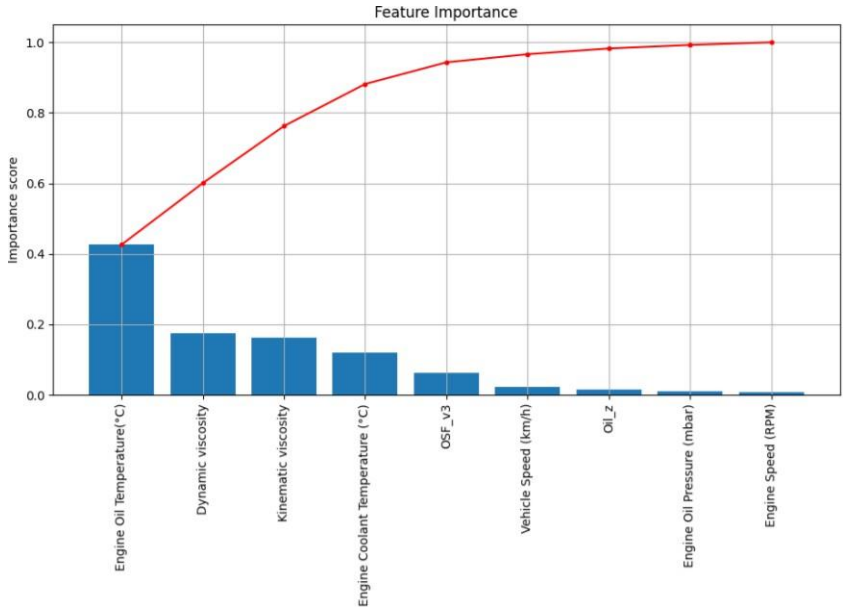

Figure 69 - Cumulative Sum of the Features Importance

Each bar in Figure 69 represents the importance score, between 0 and 1, of the corresponding feature. For example, the Oil Temperature (ºC) accounts for an importance score of roughly 0.4. The red line represents the cumulative sum of the importance of the different features. A brief perusal of the results confirm that the proposed features have a high importance score. In concrete, the Dynamic Viscosity accounts for almost 0.20 of importance, the Kinematic Viscosity for approximatly 0.15 and the OSF v3 for about 0.10. Another interesting result is that the first 4 variables (Oil Temperature (ºC), Dynamic Viscosity, Engine Coolant Temperature (°C) and Kinematic Viscosity) account for 0.90 of the importance. The temperature is used by the ECU of the vehicle to compute the lubrification needs of the engine. Higher temperatures result in smaller values of the kinematic and dynamic viscosity which decrease the lubrication power of the oil. Since viscosity control is one of the most important factors in determining the lubricant's operating condition, it is not a surprise that these variables play an important role in the model's decisions. Despite not having the same level of importance score as the temperatures and viscosities we can verify in the fifth position of importance, the calculated feature OSF v3. This feature represents the operating conditions of the vehicle and although the selected vehicles are of different brands and models they tend to follow the same patterns of operation. Nevertheless, it is important to take this descriptor into account in the implementation of the model in a business environment where vehicles may have different operating conditions.

6.3 Models Validation

According to Jayabharathi & Ilango (2021), in developing a model to answer a given problem, one must compare different approaches and variants and determine through his expertise and the results obtained, which algorithm is the most appropriate for the problem at hand. We acknowledge that tuning an algorithm with all its characteristics takes time and can become a non-compensatory task, since changing one parameter of the algorithm can indicate an improvement or worsening in the final results. As such certain conditions must be met to choose an adequate algorithm and solve the problem of determining the operating condition of lubricating oils in this study, namely:

1. Ensure that the amount of data is sufficient to apply a given machine learning algorithm to solve the highlighted problem;
2. The studies and analysis performed previously should determine a particular pattern in the data in order to help the algorithm understand the information contained in the data and improve decision making;
3. Performing mathematical approximation analysis assists in extracting knowledge from the data and can be applied to certain structured learning algorithms.

Given these requirements, we perform a cross-validation analysis of results using the algorithms described below:

- Logistic Regression (LR): Is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labeled 0 and 1. In the logistic model, the log-odds (the logarithm of the odds) for the value labeled 1 is a linear combination of one or more independent variables ("predictors"); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value);
- Perceptron (PER): Is a linear ML algorithm for binary classification tasks. It may be considered one of the first and one of the simplest types of artificial neural networks. It is definitely not "deep" learning but is an important building block. Like logistic regression, it can quickly learn a linear separation in feature space for two-class classification tasks, although unlike logistic regression, it learns using the stochastic gradient descent optimization algorithm and does not predict calibrated probabilities;
- Decision Tree Classifier (CART): Is a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piece wise constant approximation;
- Random Forest (RF): These are ensemble classifiers, which result from combining multiple DT. In the standard RF model, we select a limited random number of features from the training set, and use this subset (i.e., a bootstrap) to create a full DT. This procedure is repeated for as many DTs as we include in the RF. The RF decision is performed using the most popular class predicted by each DT. By limiting the number of features that can be selected for the bootstrap, and by not pruning the DTs, we reduce the computational burden required allowing the RF model to be deployed in environments with modest computational capabilities;
- Gradient Boosting Classifier (GBM): are a group of ML algorithms that combine many weak learning models together to create a strong predictive model. Decision trees are usually used when doing gradient boosting. Gradient boosting models are becoming popular because of their effectiveness at classifying complex datasets.

In our experimental study we rely on the Python module scikit-learn implementation of the Logistic Regression algorithm, Perceptron algorithm, Decision Tree Classifier algorithm, RF classification algorithm and Gradient Boosting Classifier algorithm. Although the initial model results are promising, it is also important to mention that for the final validations of the system developed in

this study, we performed 3 different experiments: The first experiment is a cross-validation between models and the number of features used in each model; in the second experiment we used the best model with the number of features selected in the first experiment and trained the model with data from 4 vehicles and validated if with one vehicle remaining from the original dataset. In the third experiment, we selected the two temperature variables (oil and coolant temperature), which serve as the basis of the calculation for some of the other variables and trained and tested our best model (selected in the first experiment) with only one of these variables at a time.

6.3.2 Cross Validation of Models and Features

Selecting the best model, as previously mentioned, is a complex task. For instance, it is necessary to take into account the computational capabilities, the type of input data and the different algorithms that can be applied to solve a problem. Thus, although we obtained worthy results with the use of the RF (see Table 32), we performed a systematic study using different models and input variables. This experiment had the assumptions of obtaining a robust final model capable of adapting to the needs of each vehicle fleet and still verify from a list of algorithms which would provide the best results.

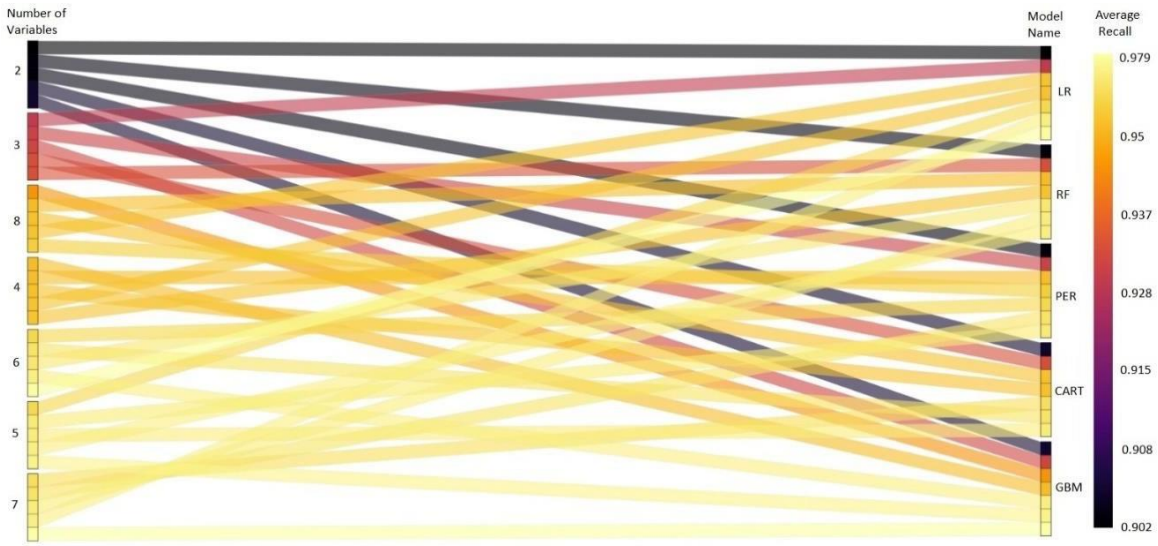

Figure 70 - Recall: Parallel Coordinate Graph

To facilitate the visualization of the results, we selected the parallel coordinates chart, which is commonly used when we have more than one metric and a single cluster. Looking at the results shown in Figure 70 we can see that the Recall ranges from 90.20% to 97.90%, showing that in general, all models are able to assess the lubricant's operating conditions well.

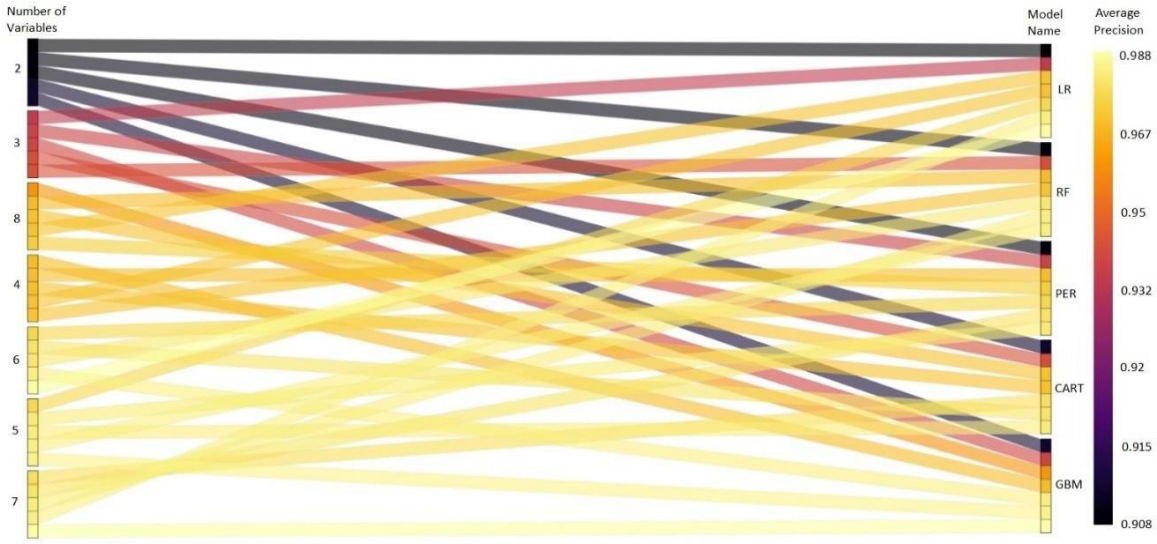

Figure 71 - Precision Parallel Coordinate Graph

By analysing Figure 71 we can see that our evaluation metric precision ranges from 90.80% to 98.80%. These results show that all models have the ability to avoid False Positives when segmenting the point cloud. i.e., for every 100 classifications of the lubricating oil's non-condition, it is expected that 90 to 98 are really true.

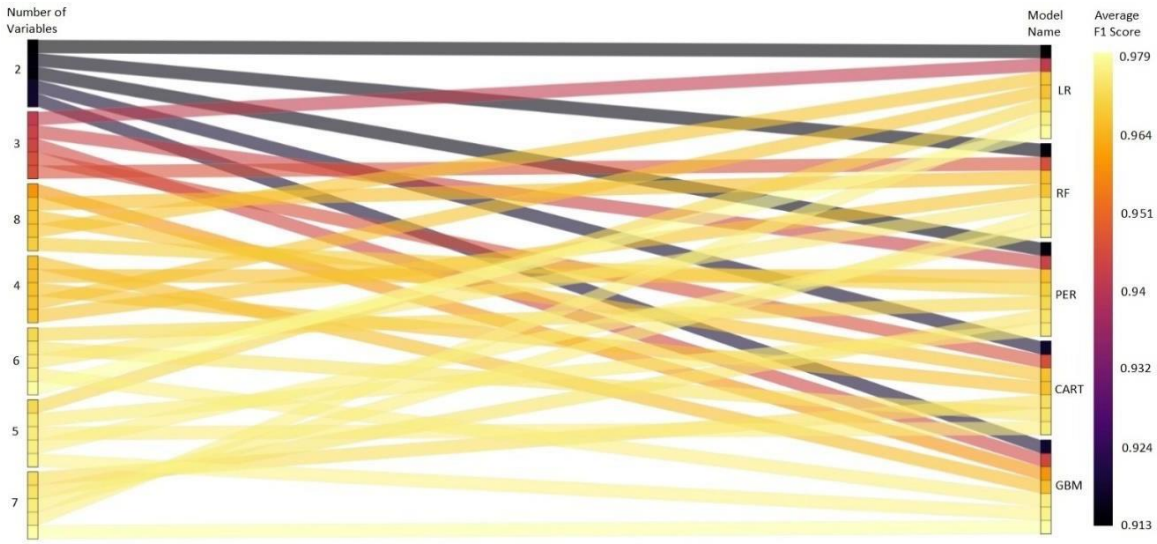

Figure 72 - F1 Score: Parallel Coordinate Graph

By analysing Figure 72 we can see that our evaluation metric F1-score ranges from 91.30% to 97.90%, This metric takes both precision and recall into account. Thus, according to Hanafy & Ming (2021), for the F1-score to be high, both precision and recall must also be high. That is, a model with a good F1-score is a model that is capable of both getting its predictions right (high precision) and retrieving the examples of the class of interest (high recall). Therefore, this metric tends to be a better summary of model quality, showing that all models are able to assess the lubricant's operating conditions well. Looking at the results one can see that, with the insertion of more variables the model

becomes more stable and gives better results. Through the combined analysis of the Figure 70, Figure 71 and Figure 72 we can see that the model with the best results was the one based on the GBM algorithm with 7 variables. Taking into account the result obtained in this analysis, we selected the best performing model, i.e., the GBM, and used it to study the impact that the different features have on the performance. In concrete, and since we have 9 features, we create all the possible sets that result of the combination of 7 features. After, we rely on the 6-fold cross-validation method to assess how different feature sets affect the performance of the GBM models.

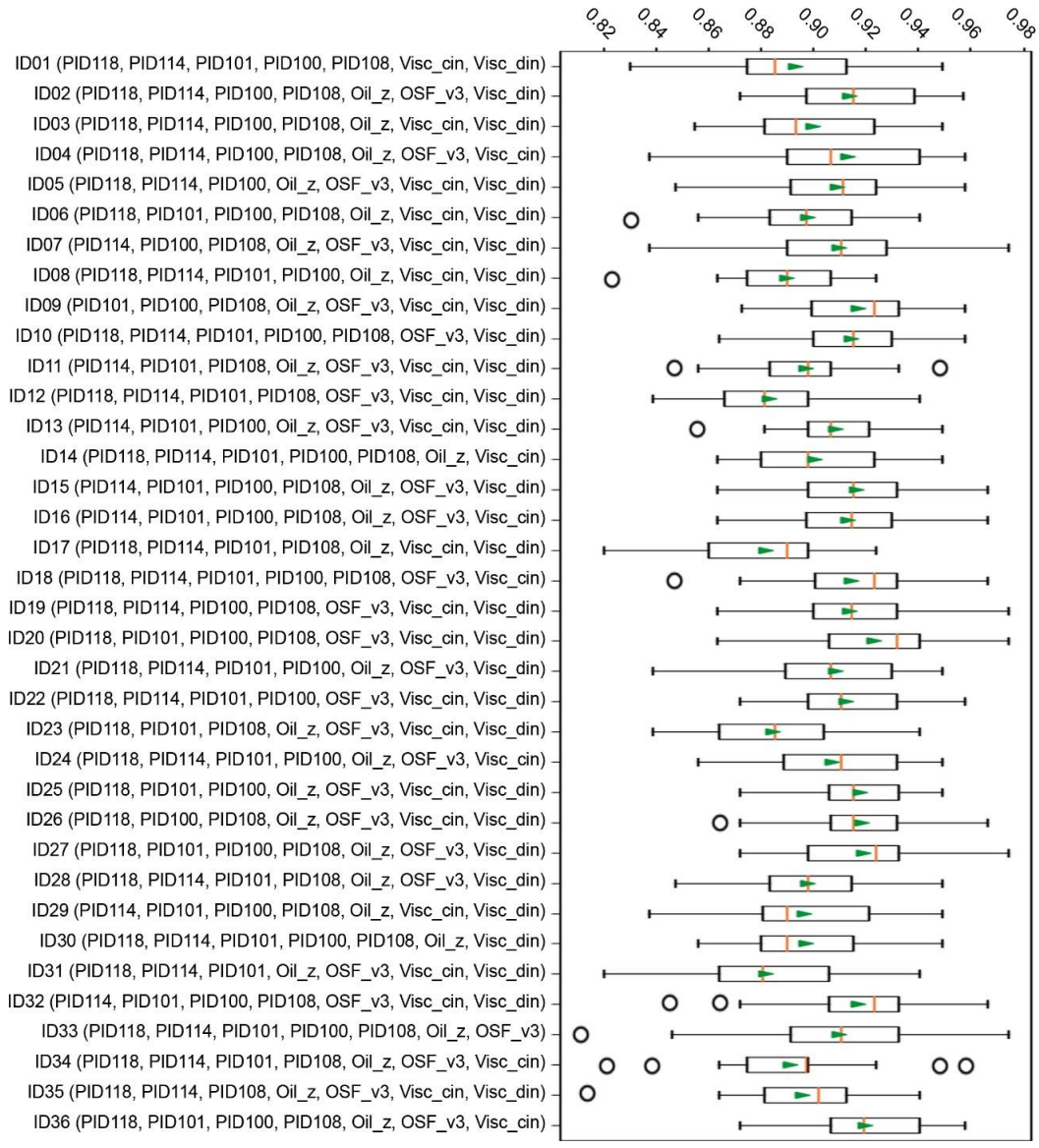

Figure 73- Box plot obtained by cross-validation of the models and number of features

The results of this analysis are shown in the Boxplot of Figure 73. A detailed analysis of the results reveals that the range of precision variation is between 82% and 98% with an average 92% among the various combinations. Additionally, we can highlight some important points from the analysis of the figure: looking at the combinations of features presented in lines with ID15 and ID16,

we can see that the variable Visc_cin and Visc_din seem to have little impact on the model results, i.e., even if we do the replacement of the variable Visc_cin by the variable Visc_din or vice-versa in the set of input variables the model results obtained are the same. Analyzing the variable combinations present in lines with ID8, ID11, ID13, ID25 and ID34, we can see that the combination of variables Oil_z and Coolant Temperature (PID101), cause less variance. Looking at the overall results, we can highlight that the best combination of variables is: ('Engine Speed - PID118' + 'Coolant Temperature - PID101' + 'Oil Temperature - PID100' + 'Oil Pressure - PID108' + 'OSF_v3' + 'Kinematic Viscosity - Visc_cin' + 'Dynamic Viscosity - Visc_din') and that the features engineered play a key role in the variability and dispersion of the results of each combination. The average results of the selected model, show in the line with ID20 are promising and can be seen in Table 33.

Metrics	Value
Recall	97.9% (+/- 0.51)
Precision	98.8% (+/- 0.72)
F1-score	97.9% (+/- 0.68)

Table 33 - Results Obtained with Best Model with Original Dataset – ID20

According the evaluation metrics show in Table 33, we can identify some important signals that the model could be deployed in a real world scenario to determinate the operation condition of the lubricant oils in diesel engines. The high score of the recall (97.90%) demonstrate the ability of a classification model to identify all relevant instances, the high score of the precision (98.80%) show the ability of a classification model to return only relevant instances and the F1 score (97.80%) that combines recall and precision using the harmonic mean confirm our good perception about the accuracy of the model. Even though we had good results with the previous metrics, we still performed a ROC curve analysis.

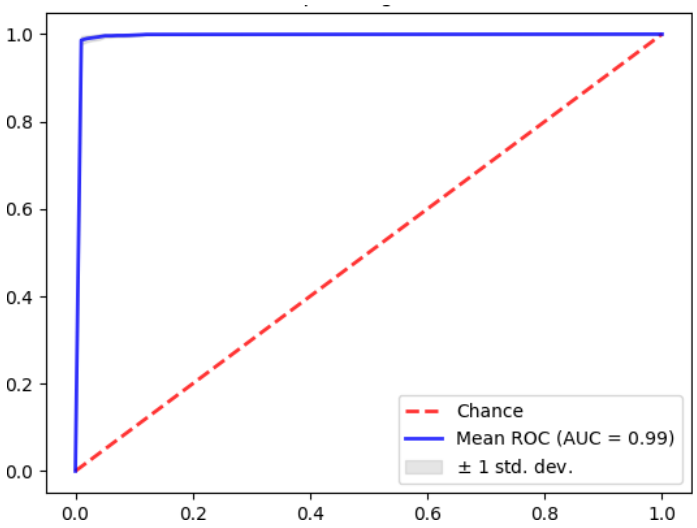

Figure 74 - ROC Curve Best Model – ID20

The Figure 74 shows the ROC response of different data sets, created from K-fold cross-validation. By taking all these curves, it is possible to calculate the average area under curve, and to see the variance of the curve when the training set is divided into different subsets. This roughly shows how the output of the classifier is affected by changes in the training data, and how different the splits generated by K-fold cross-validation are. Still analyzing Figure 74 we can see the AUC results of 0.99, given the nature of our problem, where the accurate identification of situations where the lubricating oil has gone bad is of utmost importance to avoid catastrophic failures. It is important to mention that this evaluation presented in the ROC curve was also performed from a 6-fold cross validation exactly as presented in Figure 68. The difference between these two analyses is given by the algorithm used, since in Figure 68 we used RF and in Figure 74 we used GBM. This comparison of results is important because it demonstrates that the variation of the input data set and the algorithm used can imply differences in the results, even if minimal. Thus, we can conclude that the model developed from the GBM algorithm with 7 variables does not lose its potential to indicate the lubricant status with a change in the input data, and we have confirmed that this model can be used for further analysis in a truly uncontrolled environment, but further analysis is still needed to determine if the model developed here can use previously unknown vehicle data outside the scope of this project.

6.3.3 Model Validation with New Vehicle

In order to identify the ability of the chosen model to adapt to unknown data, we conduct an experiment where we evaluate the performance of the ML model on classifying the oil conditions on vehicles that were not used in it development or training. To perform this analysis, and taken into account that our dataset is composed by data of 5 vehicles, we will use the data from 4 vehicles to design, develop and train the model, and then we will evaluate the performance on the remaining vehicle. Note that the data of this fifth vehicle will never be used during the first phase. In Table 34, column "Vehicles Training" correspond to the vehicles that were present in the training dataset and the column "Vehicles Testing" the vehicles correspond to the vehicles that was present in the test dataset, with their respective results.

Vehicles Training	Vehicles Testing	AUC	Precision	Recall	F1 Score
[67, 34, 80, 82]	[21]	0.72	0.75	0.72	0.71
[21, 82, 80, 34]	[67]	0.76	0.80	0.76	0.75
[82, 21, 67, 34]	[80]	0.96	0.97	0.96	0.96
[80, 21, 67, 34]	[82]	0.95	0.95	0.95	0.95
[82, 80, 21, 67]	[34]	0.51	0.62	0.51	0.37

Table 34 - Results Obtained by Model Validation with New Vehicle

Looking at the results of Table 34, it is possible to see that the GBM algorithm with 7 variables (see ID20 Figure 73), looses some performance when determining the lubricant oil conditions on a new vehicle, obtaining a variation of precision from 62% to 97%, of recall from 51% to 96% and of F1-

score from 37% to 96%. According to Table 34, one can see that there are good results when the test dataset has data from vehicles 80 and 82. These good results shown in these interactions happen because of the difference in the years of manufacture and Km performed on the vehicles that make up the original dataset. That is, according to Table 14 vehicles 80 and 82 started their operation in 2017, which represents 15 years of service less than the oldest vehicle (Vehicle Identification equal 21 - Mercedes Citaro O530 - 2002 in Table 14). Another point that substantiates the good results using these vehicles, is due to the fact that vehicles 80 and 82 had an average of 3 years of total operation by the time of data collection used in this study (data collection performed between January 1, 2020 to August 1, 2020), which represented an average of 151200 km at the end of this period. This average mileage value of the 80 and 82 vehicles compared to the oldest vehicle present in the dataset indicates a difference of almost 1M km, which corresponds in the lubrication systems, differences in terms of lubricant consumption and operating profile. For example, the older vehicles need almost daily lubricant refills to maintain the correct lubricant level and the adjustments between the engine parts have already experienced high wear. Still analyzing the results obtained in this experiment, we know vehicles 34 and 67 have very similar characteristics as well as vehicles 80 and 82. So it was expected that if we had one of them in the test set, having similar data in the training set would result in a good classification performance. But the overall metrics of the experiments performed with the 34 and 67 vehicles are in the top 3 worst metrics unlike the experiments performed with the 80 and 82 vehicles. This difference in results presented in Table 34 is caused by the age of the vehicles and the technologies employed. In other words, the average difference of almost 10 years between the 34, 67 and 80, 82 vehicles is not only related to the number of lubricant non-operating condition situations encountered, but also represents a leap in the availability of information. As mentioned earlier, the older vehicles have more wear and are more susceptible to failures, so the data from the older vehicles, in addition to having a lower quantity in the dataset as a whole, have a higher weight in relation to the data from the newer cars because they have more situations of non-condition of lubricant operation. This situation is confirmed when we verify the results of removing vehicle 34 in comparison to removing vehicle 67, where we can see that vehicle 34 has a higher impact than vehicle 67, being a little older, with approximately 2 years of difference. In relation to the availability of information, the current communication protocols allow the acquisition of data in larger quantities and with higher resolution, which results in a larger portion of samples in the training and test dataset by the vehicles 80 and 82 in comparison with vehicles 34 and 67. In this case this situation is not confirmed with vehicle 21 because it has the slowest data transmission rate among all the other vehicles and during the data collection period it had most of its time without operating. In summary, the removal of vehicles 34 and 67 from the experiment influences much more by the amount of data available and by the number of situations of no lubricant operation condition in comparison to the data removed from vehicles 80 and 82.

6.3.4 Model Validation with Only One Temperature Variable

The main goal of this last experiment is to understand how the temperature variables, namely oil and coolant, affect the performance of the models. In concrete, we want to understand if we really need both temperatures, or if one would be enough to the models perform well. The main reason behind this study is concerned with the fact that in certain situations the vehicles only have information regarding one temperature. The experiment consisted of creating two new datasets, divided as follows: the first dataset had the variables, 'Engine Speed - PID118' + 'Vehicle Speed - PID114' + 'Coolant Temperature - PID101' + 'Oil Pressure - PID108' + 'Oil_z' + 'OSF_v3' + 'Kinematic Viscosity - Visc_cin' + 'Dynamic Viscosity - Visc_din') and the second dataset had the variables ' Engine Speed - PID118' + ' Vehicle Speed - PID114' + 'Oil Temperature - PID100' + 'Oil Pressure - PID108' + 'Oil_z + 'OSF_v3' + "Kinematic Viscosity - Visc_cin' + 'Dynamic Viscosity - Visc_din'). Note that in this division, each dataset has one temperature variable; the first dataset contains the PID101 that is the Coolant Temperature, and the second dataset has the variable PID100 is the Oil Temperature.

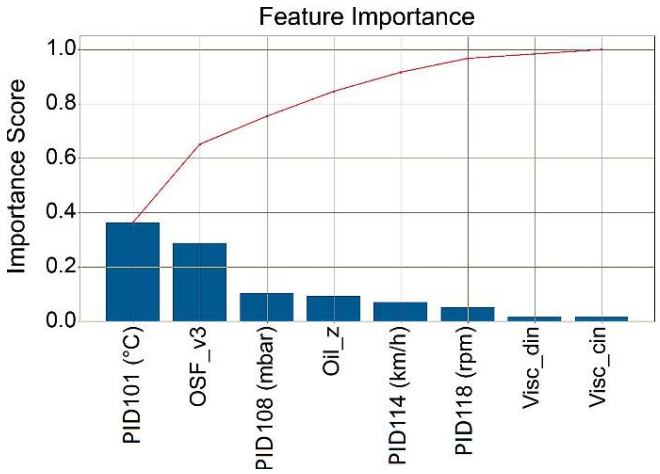

Figure 75 - Feature Importance - Coolant Temperature

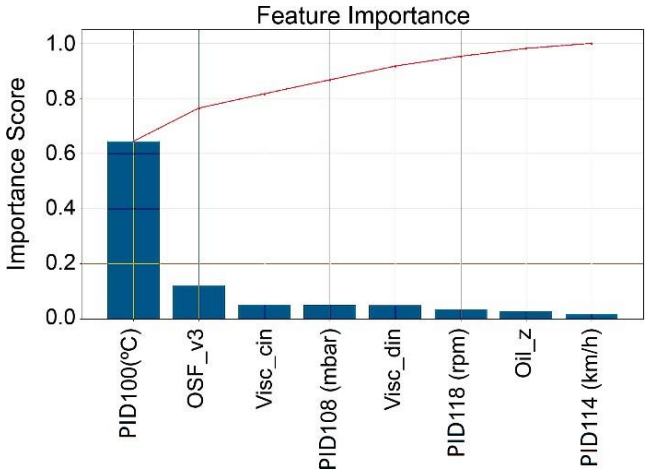

Figure 76 - Feature Importance - Oil Temperature

After creating these two datasets, we ran the GBM algorithm using each dataset as input data. It is possible to see through the analysis of Figure 75 and Figure 76 that there are differences in the distribution of variable importance. In the case of the dataset with the Coolant Temperature (Figure 75), the values of each characteristic are closer and the distribution of importance is more evenly distributed, whereas in the case of the dataset with the Oil Temperature (Figure 76), the importance of this variable is almost twice the sum of the importance of all the other variables, which shows that this feature is important for the success of the model. This discrepancy in importance values is due to the fact that the Oil Temperature is directly related to all the other variables and is a crucial factor in the analysis of lubricant degradation. The average results of this experiment were: Precision of 97%, Recall of 96% and F1-score of 96%. With these values we see that the performance of the model is extremely acceptable if applied only to situations with Oil Temperature. Thus, we should not exclude either of the two variables from our problem, since, from our own experience, there are vehicles that can have in their communication bus the information of both temperatures or only one of them. For the model to be as agnostic as possible, it should have good results with data from both temperatures or just one of them.

Chapter 7 - Conclusion

Modern industries rely on maintenance techniques and methods to anticipate failures in their equipment. The transportation industry is not an exception, and nowadays they are investing more and more in the research and development of methodologies that will allow them to predict vehicles breakdown signatures, preventing catastrophic failures and large costs. One of such signatures is concerned with the engine maintenance, in which they want to establish the most economical oil change and overhaul intervals in terms of cost, wear and failure diagnosis efficiency.

In this work, we propose an automated method to identify the conditions of lubricating oils in diesel engines. We collect data from a fleet of heavy passenger vehicles, and using real-time information from in-vehicle sensors, as well as engineered features, we proposed an intelligent system to assist in vehicle maintenance. For this development, it was necessary to create a work architecture based on CBM and machine learning concepts, which were fundamental to structure a system capable of fitting into the predictive maintenance scope. We started by performing an analysis of databases of common failures in diesel vehicles, to determine failures in the lubrication systems were capable of predicting the operating conditions of the vehicles. After determining the problem that this system would solve, we structured it into 3 phases: "Data collection", "Modelling" and "Deployment". For the success of each of these phases it was necessary to follow defined steps, which were registered and catalogued in the data collection protocol presented, in the experiments performed, and in the system evaluations, in order to make possible its application in a real environment.

To validate the use of the features selected, we carried out a study through correlation and PCA, which indicated that the selected variables can determine the answer to the stated problem, but it is necessary to have at least 4 variables for the variance of the data is not lost. The kinematic viscosity variable proved to be important to characterize the condition of the lubricant. Also 3 of the 4 variables that explain 90% of the variance in the original data were engineered by us, showing that it is necessary to calculate other variables based on the data to improving the results. Furthermore, the correlation analysis identified relationships that confirm that confirmed the relevance of the collected variables, consequently, paved the way for the development of a model to identify the operating conditions of lubricants in diesel vehicles. The best performing model was the Gradient Boosting Classifier (GBM), which successfully predicts whether the oil is good for running conditions or not. Given the reported results, the GBM revealed to be effective obtaining a Recall of 94%, a Precision of 96% and an F1-score of 95%, which shows that it is capable of identifying most situations in which oil lubricant is not good for operating conditions and needs to be changed. Even with good results, to identify potential problems in the implementation of the model proposed in this study in a real fleet environment, we performed an additional set of experiments. One of the studies performed was designed to identify the possibility of the model to generate good results with unknown data. The results of this experiments showed that it can be identified that the performance of the model degrades for older vehicles with long

service life, obtaining Recall results of up to 51%, an Precision of up to 62%, and an F1-score of up to 37%. However it had good results for newer vehicles and with more current variable operation profiles, getting Recall results of up to 96%, an Precision of up to 97%, and an F1-score of up to 96%. The last experiment performed concerned the impact of the temperature variables, since certain vehicles may not have this data available on the communication bus. The results of this study show that the model is able to identify the operating condition of the lubricating oil even with only one of the temperature variables, obtaining a Recall of 96%, a Precision of 97%, and an F1-score of 96%. However, it is important to mention in this experiment the feature importance distribution becomes more uniform with the use only of the coolant temperature and is uneven with the use only of the oil temperature, indicating importance values of 62% for the oil temperature, which shows a difference of almost twice the importance of the other variables added together. Finally, it can be said that the results obtained in this work are encouraging and constitute a step forward for the automation of monitoring the conditions of vehicle lubrication systems in real time. So, we can say that through the system presented and the results obtained in this work, the hypothesis presented that learning models allow the automatic prediction of the state of the lubricant and the equipment, thus reducing downtime and maintenance costs associated with diesel vehicles, is really correct and define that this system can be used in new workshops as a tool to assist in maintenance management.

7.1 Summary of Research Contributions

A fundamental part in the delivery of a dissertation is the presentation of its contributions. Thus, this section is divided into three points that address the contributions to theory, methodology and practice in the environment that guide the development of this automatic system for detection of the degradation index of lubricating oils in diesel engines.

- Theoretical Contributions: Both theory and empirical findings contribute to our understanding of the interrelationship between a maintenance and fleet management environment with development of new technologies and techniques for failure detection. This study also contributes to our understanding of the question of how profit-making fleet-based organizations can adopt and utilize initiatives related to data analytics in order to improve old maintenance concepts.

The conclusions of this study suggest that institutionalized initiatives related to lubricant analytics are able to predict failures and serve as a basis for continuous maintenance improvement, creating the means for continued development in asset management.

Although there have been some research studies on how organizations should deal with asset management and maintenance, this work presents a clear and directed path for implementation focused on vehicle fleets. Thus, it can be said that this thesis serves as a

guide for theoretical developments in the maintenance of this type of equipment (diesel vehicles);

- Methodological Contributions: The main methodological contribution of the research has been the combination and application of concepts of vehicle fleet maintenance management, machine learning techniques, data analysis and development of failure prediction models. Another methodological contribution lies in the experience gained through the application of a case study strategy based on a real fleet, which relies on interpretative techniques applied for data collection. This experience may be useful for further studies on the adoption and use of initiatives related to failure prediction in organizations and in asset maintenance management.

Finally, a methodological contribution concerns the appropriateness of applying theoretical concepts and theories developed in other contexts. The applicability of some theories and research models developed in other types of assets developed for studies in the context of an extremely controlled environment and with unlimited data support have been questioned due to the differences that exist in the technological structures employed in the equipment. The successful use of these theories in this study contributes to providing examples of the interpretation of case studies on older vehicle fleets with a reduced input of information and data collection.

- Practical Contributions: One of the practical contributions of this research is the detailed insight into the development of the system. The selection of a fleet of vehicles operating in a real uncontrolled environment reveals that initiatives related to data analysis and automatic asset management are neither distant visions nor high cost. This implies that for a wider range of failure prediction systems covering a larger number of fleets, emphasis should be placed on the contexts in which failures occur, so as to avoid spending unnecessary time on the development of purely theoretical models.

This will help increase developments towards implementation in a real environment, and the case study still shows that maintenance professionals and managers need to acquire new skills in data analysis in order to realize the development needs in failure prediction systems.

Another practical contribution is the framework for analyzing common failures in a fleet in order to gain an understanding of the real needs between maintenance management and development in a specific context. The contribution of this research is to understand, based on theoretical assumptions, how the initiative for creating failure prediction models can be and also how it contributes to development. To this end, the due process model can be used as a practical tool for creating new detection and prediction systems.

Based on these three areas of contribution (theoretical, methodological, and practical), it can be determined that this work has a complete structure for development and implementation. In other words, starting with the theoretical contributions it is possible to glimpse paths for further development and thereby apply the methodologies that are presented. Finally, through the practical contribution it is possible to confirm that this study not only focuses on theoretical developments, but also on the implementation in a real environment focused on the active maintenance management of diesel vehicle fleets through the analysis of lubricant data.

7.2 Limitations and Future Work

The following work had its limitations overcome without any interference in the final results to which it was proposed. However, it is necessary to emphasize the importance of stimulating organizations that rely on vehicle fleets for profit generation to create digital and structured databases. The level of information that exists today in the vast majority of fleets is mostly in physical format and without great possibilities for analysis. This type of format entails a huge effort to transform and structure the information into a digital database before starting any development.

Following this line of structuring and organization of a digital database, we initially suggest as future and further work, the development of a system capable of storing and organizing data of failures, repairs and fault identification tests in vehicle fleets. This database will serve for new failure prediction models and can be used across fleets, since in part of the study of this thesis, we identified that it is possible to establish a relationship between the behavior of vehicle data, not caring about brand and model as well as the failures that are similar among different fleets.

A second direction for future work is the development of new failure prediction models for different vehicle systems other than the lubrication system. For these developments other tests can be used to acquire failure labels, as well as real-time data collection can be performed from other systems. It is important to mention that in this document there is a table, even if initial, that indicates systems that may be the target of further studies.

Even with the possibility of developing new models for other vehicle systems, we can still cite as a third suggestion for future work, the development of the system described in this document based on other machine learning algorithms. These algorithms may not follow the classification lines as the ones used in this document, and may not use labels for failure prediction.

Besides the development of these studies with the use of other algorithms, which opens range of comparative options that serve as suggestions, we can also suggest as a fourth option, the development of solutions for capturing and organizing labels necessary for supervised structure algorithms. It is important to mention that during the development of this study, the developer's expertise regarding maintenance data and vehicle behavior information was extremely important to be able to analyze the spot tests and visually determine the failure or not of the lubricant. This type of

expertise can be transferred to an automatic system that can use a photo of the test to determine whether or not the lubricant fails, thus removing the possibility of human error in the final analysis.

In view of the options presented, it is possible to identify numerous opportunities for future work as well as further developments that have already been started in this document. These plausible options may be in important areas that range from database structuring to the development of support equipment to obtain labels to improve the existing systems.

References

Abbott-mccune, S., & Shay, L. A. (2016). *Intrusion Prevention System of Automotive network CAN bus*.

ACAP. (2021). *MERCADO AUTOMÓVEL EM PORTUGAL. MERCADO GLOBAL MANTÉM TENDÊNCIA DE QUEDA*. Web Page. https://www.acap.pt/pt/noticia/560/mercado-automovel-em-portugal-mercado-global-mantem-tendencia-de-queda

Afefy, I. H. (2010). Reliability-Centered Maintenance Methodology and Application: A Case Study. *Engineering*, *02*(11), 863–873. https://doi.org/10.4236/eng.2010.211109

Agoston, A., Ötsch, C., & Jakoby, B. (2005). Viscosity sensors for engine oil condition monitoring - Application and interpretation of results. *Sensors and Actuators, A: Physical*, *121*(2), 327–332. https://doi.org/10.1016/j.sna.2005.02.024

Alvine, B. B., Lethbridge, T. C., Garzón, M., & Opeyemi, O. A. (2018). Design and implementation of distributed expert systems: On a control strategy to manage the execution flow of rule activation. *Expert Systems with Applications*, *96*, 129–148. https://doi.org/10.1016/j.eswa.2017.11.033

Aucélio, R. Q., de Souza, R. M., de Campos, R. C., Miekeley, N., & da Silveira, C. L. P. (2007). The determination of trace metals in lubricating oils by atomic spectrometry. *Spectrochimica Acta - Part B Atomic Spectroscopy*, *62*(9), 952–961. https://doi.org/10.1016/j.sab.2007.05.003

Automotive, S. (2021). *Stratio Automotive - web page*. https://stratioautomotive.com/

Axelsson, J., Fröberg, J., Hansson, H., Norström, C., Sandström, K., & Villing, B. (2004). *A Comparative Case Study of Distributed Network Architectures for Different Automotive Applications*. *478*, 846–865. https://doi.org/10.1201/9781420036336.ch57

Baak, M., Koopman, R., Snoek, H., & Klous, S. (2020). A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics. *Computational Statistics and Data Analysis*, *152*, 107043. https://doi.org/10.1016/j.csda.2020.107043

Beynon-Davies, P. (1991). *Expert Database Systems - A Gentle Introduction*. McGraw-Hill.

Bloom, N. B. (2005). *Reliability Centered Maintenance - Implementation made simple* (McGraw-Hill Education, Ed.; illustrate). https://doi.org/10.1036/0071460691

Bordatchev, E., Aghayan, H., & Yang, J. (2014). Object shape-based optical sensing methodology and system for condition monitoring of contaminated engine lubricants. *Optics and Lasers in Engineering*, *54*, 128–138. https://doi.org/10.1016/j.optlaseng.2013.10.009

Bousdekis, A., Lepenioti, K., Apostolou, D., & Mentzas, G. (2021). A review of data-driven decision-making methods for industry 4.0 maintenance applications. *Electronics (Switzerland)*, *10*(7). https://doi.org/10.3390/electronics10070828

Bozdal, M., Samie, M., Aslam, S., & Jennions, I. (2020). Evaluation of can bus security challenges. *Sensors (Switzerland)*, *20*(8). https://doi.org/10.3390/s20082364

BPstat. (2021). *Análise das empresas do setor dos transportes*. Web Page. https://bpstat.bportugal.pt/conteudos/publicacoes/1343

Butdee, S., & Kullawong, T. (2015). Integrating Reliability Centered Maintenance with Statistical Forecasting Techniques and Cost Engineering on Machine in Casting Plant of Automotive Parts. *KMUTNB International Journal of Applied Science and Technology*, *8*(2), 1–15. https://doi.org/10.14416/j.ijast.2015.04.002

Cabral, J. P. S. (2013). *Gestão da manutenção de equipamentos, instalações e edifícios* (3ª). Lidel - Edições Técnicas.

Carvalho, T. P., Soares, F. A. A. M. N., Vita, R., Francisco, R. da P., Basto, J. P., & Alcalá, S. G. S. (2019). A systematic literature review of machine learning methods applied to predictive maintenance. *Computers and Industrial Engineering*, *137*(August), 106024. https://doi.org/10.1016/j.cie.2019.106024

Charissis, V., Falah, J., Lagoo, R., Alfalah, S. F. M., Khan, S., Wang, S., Altarteer, S., Larbi, K. B., & Drikakis, D. (2021). Employing emerging technologies to develop and evaluate in-vehicle intelligent systems for driver support: Infotainment AR hud case study. *Applied Sciences (Switzerland)*, *11*(4), 1–28. https://doi.org/10.3390/app11041397

Chen, H., H.L.Chiang, R., & C. Storey, V. (2018). Business Intelligence and Analytics: From Big Data To Big Impact. *MIS Quarterly*, *36*(4), 1165–1188.

Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile Networks and Applications*, *19*(2), 171–209. https://doi.org/10.1007/s11036-013-0489-0

Chiaburu, D. S. (2016). Analytics: A Catalyst for Stagnant Science? *Journal of Management Inquiry*, *25*(1), 111–115. https://doi.org/10.1177/1056492615601342

Chiang, R. H. L., Grover, V., Liang, T. P., & Zhang, D. (2018). Special Issue: Strategic Value of Big Data and Business Analytics. *Journal of Management Information Systems*, *35*(2), 383–387. https://doi.org/10.1080/07421222.2018.1451950

Clark, R. J., & Fajardo, C. M. (2012). Assessment of the Properties of Internal Combustion Engine Lubricants Using an Onboard Sensor. *Tribology Transactions*, *55*(4), 458–465. https://doi.org/10.1080/10402004.2012.670892

Collacott, R. A. (Ralph A. (1977). Mechanical fault diagnosis and condition monitoring / R.A. Collacott. In Mechanical fault diagnosis and condition monitoring. Chapman and Hall.

Delgado Ortiz, J., Saldivia Saldivia, F., & Fygueroa Salgado, S. (2014). Sistema para la determinación de la degradación del lubricante basado en el tratamiento digital de manchas de aceite de motores diesel. *Revista UIS Ingenierías*, *13*(1), 55–61.

Delprete, C., & Razavykia, A. (2018). Piston ring–liner lubrication and tribological performance evaluation: A review. *Proceedings of the Institution of Mechanical Engineers, Part J: Journal of Engineering Tribology*, *232*(2), 193–209. https://doi.org/10.1177/1350650117706269

Diaby, M., Singhal, P., Ousmane, M., Sablier, M., Le Négrate, A., El Fassi, M., & Zymla, V. (2013). Impact factors for the degradation of engine oil causing carbonaceous deposits in the piston's grooves of Diesel engines. *Fuel*, *107*, 90–101. https://doi.org/10.1016/j.fuel.2012.12.021

Diesel & gas turbine engineers working costs & annual report,analysis of stoppages, (1983–1993)

Dogan, A., & Birant, D. (2021). Machine learning and data mining in manufacturing. *Expert Systems with Applications*, *166*, 114060. https://doi.org/10.1016/j.eswa.2020.114060

Dunn, S. (2015). *Big data, predictive analytics and maintenance.* https://www.assetivity.com.au/articles/maintenance-management/big-data-predictive-analytics-and-maintenance/

Dv-, A. P., & Dv-, A. P. (2014). *Vojtěch Kumbár Jiří Votava Differences in engine oil degradation in spark-ignition and compression-ignition engine Rozbieżna degradacja oleju silnikowego przy zastosowaniu w silniku o zapłonie wymuszonym i w silniku wysokoprężnym*. *16*(4), 622–628.

European Union. (2021). *Safe, sustainable and connected transport*. Web Page. https://european-union.europa.eu/priorities-and-actions/actions-topic/transport_en

Feigenbaum, E. A. (1979). *THE ART OF ARTIFICIAL INTELLIGENCE.*

Felipe Lima Bronté, Carlos Renato Pagotto, & Victor Bicalho Civinelli de Almeida. (2015). Predictive maintenance techniques applied in fleet oil management. *Proceedings of the 23rd ABCM International Congress of Mechanical Engineering*, *October*. https://doi.org/10.20906/cps/cob-2015-0469

Green, D. A., & Lewis, R. (2008). The effects of soot-contaminated engine oil on wear and friction: A review. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, *222*(9), 1669–1689. https://doi.org/10.1243/09544070JAUTO468

Guan, L., Feng, X. L., Xiong, G., & Xie, J. A. (2011). Application of dielectric spectroscopy for engine lubricating oil degradation monitoring. *Sensors and Actuators, A: Physical*, *168*(1), 22–29. https://doi.org/10.1016/j.sna.2011.03.033

Hanafy, M., & Ming, R. (2021). Machine learning approaches for auto insurance big data. *Risks*, *9*(2), 1–23. https://doi.org/10.3390/risks9020042

Hoang, A. T., & Pham, V. V. (2019). A study of emission characteristic, deposits, and lubrication oil degradation of a diesel engine running on preheated vegetable oil and diesel oil. *Energy Sources, Part A: Recovery, Utilization and Environmental Effects*, *41*(5), 611–625. https://doi.org/10.1080/15567036.2018.1520344

HUGHES RV. (1969). Diesel Engines. *Mar Engr & Naval Architect*, *92*(1124), 462–464. https://doi.org/10.1016/b978-1-4832-1355-2.50021-2

Ibrahim, F., Mahmood, W. M. F. W., Abdullah, S., & Mansor, M. R. A. (2013). A Review of Soot Particle Measurement in Lubricating Oil. *SCIENCE & TECHNOLOGY RESEARCH INSTITUTE FOR DEFENCE*, *8*, 141–152.

Idros, M. F. M., Ali, S., & Islam, M. S. (2012). Optical analysis for condition based monitoring of oxidation degradation in lubricant oil. *ICIAS 2012 - 2012 4th International Conference on Intelligent and Advanced Systems: A Conference of World Engineering, Science and Technology Congress (ESTCON) - Conference Proceedings*, *2*, 735–740. https://doi.org/10.1109/ICIAS.2012.6306110

Jayabharathi, S., & Ilango, V. (2021). A Brief Revolution of Evolution and Resurgence on Machine Learning. *2021 Asian Conference on Innovation in Technology, ASIANCON 2021*, 1–5. https://doi.org/10.1109/ASIANCON51346.2021.9544706

Keartland, S., & Van Zyl, T. L. (2020). Automating predictive maintenance using oil analysis and machine learning. *2020 International SAUPEC/RobMech/PRASA Conference, SAUPEC/RobMech/PRASA 2020*. https://doi.org/10.1109/SAUPEC/RobMech/PRASA48453.2020.9041003

Konar, P., Sil, J., & Chattopadhyay, P. (2015). Knowledge extraction using data mining for multi-class fault diagnosis of induction motor. *Neurocomputing*, *166*, 14–25. https://doi.org/10.1016/j.neucom.2015.04.040

Kumar, A., & Ghosh, S. K. (2016). Oil condition monitoring for HEMM - A case study. *Industrial Lubrication and Tribology*, *68*(6), 718–722. https://doi.org/10.1108/ILT-09-2015-0124

Kumar, N., & Parrek, S. (2013). *Artificial Intelligence and Expert Systems* (First). Genius Publication.

Kumbár, V., & Sabaliauskas, A. (2013). Lowerature behaviour of the engine oil. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, *61*(6), 1763–1767. https://doi.org/10.11118/actaun201361061763

Lee, P. M., Stark, M. S., Wilkinson, J. J., Priest, M., Smith, J. R. L., Taylor, R. I., & Chung, S. (2005). The degradation of lubricants in gasoline engines: Development of a test procedure to evaluate engine oil degradation and its consequences for rheology. *Tribology and Interface Engineering Series*, *48*, 593–602. https://doi.org/10.1016/s0167-8922(05)80061-6

Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. *Taylor & Francis, Ltd. American Statistical Association*, *62*(318), 399–402.

Malaguti, R., Lourenço, N., & Silva, C. (2021a). A Well Lubricated Machine: A Data Driven Model for Lubricant Oil Conditions. In G. Marreiros, F. S. Melo, N. Lau, H. Lopes Cardoso, & L. P. Reis (Eds.), *Progress in Artificial Intelligence* (pp. 549–560). Springer International Publishing.

Malaguti, R., Lourenço, N., & Silva, C. (2021b). Wear and Tear: A Data Driven Analysis of the Operating Condition of Lubricant Oils. In A. Dolgui, A. Bernard, D. Lemoine, G. von Cieminski, & D. Romero (Eds.), *Advances in Production Management Systems. Artificial Intelligence for Sustainable and Resilient Production Systems* (pp. 217–225). Springer International Publishing.

Malaguti, R., Lourenço, N., & Silva, C. (2022). A supervised machine learning model for determining lubricant oil operating conditions. *Expert Systems*, *n/a*(n/a), e13116. https://doi.org/https://doi.org/10.1111/exsy.13116

Mell, P., & Grance, T. (2011). The NIST-National Institute of Standars and Technology- Definition of Cloud Computing. *NIST Special Publication 800-145*, 7.

Mikut, R., & Reischl, M. (2011). Data mining tools. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(5), 431–443. https://doi.org/10.1002/widm.24

Mitchell, T. M. (1997). Machine learning. In *EEG Signal Processing and Feature Extraction*. McGraw-Hill Science/Engineering/Math.

Młynarczyk, A., & Sikora, G. (2014). Analysis of the Modern Oil Viscosity Changes During Their Operation in Combustion Engines. *Journal of KONES. Powertrain and Transport*, *21*(4), 361–368. https://doi.org/10.5604/12314005.1130484

Morgan, I., & Liu, H. (2009). Predicting future states with n-dimensional markov chains for fault diagnosis. *IEEE Transactions on Industrial Electronics*, *56*(5), 1774–1781. https://doi.org/10.1109/TIE.2008.2011306

Motamen Salehi, F., Morina, A., & Neville, A. (2017). The effect of soot and diesel contamination on wear and friction of engine oil pump. *Tribology International*, *115*(May), 285–296. https://doi.org/10.1016/j.triboint.2017.05.041

MOTORcheckUP. (2022). *FLUIDcheckUP EVALUATION*. https://motorcheckup.com/evaluation-2/?lang=en

Muchiri, P. P. N., Adika, C. O., & Wakiru, J. M. (2018). *Development of Maintenance Decision Guidelines from Used Oil Data of a Thermal Powerplant* . *15*(2), 63–71. https://doi.org/10.9790/1684-1502026371

Mujahid, A., & Dickert, F. L. (2012). Monitoring automotive oil degradation: Analytical tools and onboard sensing technologies. *Analytical and Bioanalytical Chemistry*, *404*(4), 1197–1209. https://doi.org/10.1007/s00216-012-6186-1

Nahim, H. M., Younes, R., Shraim, H., & Ouladsine, M. (2016). Oriented review to potential simulator for faults modeling in diesel engine. *Journal of Marine Science and Technology (Japan)*, *21*(3), 533–551. https://doi.org/10.1007/s00773-015-0358-6

Ndustries, I. (2012). Reliability engineering. In *Lees' Loss Prevention in the Process Industries*. https://doi.org/10.1016/B978-0-12-397189-0.00007-0

Neumann, A., Mytych, M. J., Wesemann, D., Wisniewski, L., & Jasperneite, J. (2017). Approaches for in-vehicle communication - An analysis and outlook. *Communications in Computer and Information Science*, *718*, 395–411. https://doi.org/10.1007/978-3-319-59767-6_31

Nunes, P., Pinheiro, F., & Brito, M. C. (2019). The effects of environmental transport policies on the environment, economy and employment in Portugal. *Journal of Cleaner Production*, *213*, 428–439. https://doi.org/10.1016/j.jclepro.2018.12.166

Pawashe, R. A., Kalkundri, S. S., Chavan, C. B., & Rammohan, A. (2017). Fault diagnosis of engine lubrication system. *2017 International Conference on Microelectronic Devices, Circuits and Systems, ICMDCS 2017*, *2017-Janua*, 1–6. https://doi.org/10.1109/ICMDCS.2017.8211690

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in {P}ython. *Journal of Machine Learning Research*, *12*, 2825–2830.

Plumley, M. J., Wong, V., & Martins, T. (2018). Oil Degradation in a Diesel Engine with Dual-Loop Lubricating System. *Tribology Transactions*, *61*(4), 596–603. https://doi.org/10.1080/10402004.2017.1378396

Prajapati, A., & Bechtel, J. (2012). *Condition based maintenance: a survey*. https://doi.org/10.1108/13552511211281552

Pranjoto, H., Agustine, L., & Mereditha, M. (2018). OBD-II-based vehicle management over GPRS wireless network for fleet monitoring and fleet maintenance management. *Journal of Telecommunication, Electronic and Computer Engineering*, *10*(2–3), 15–18.

Prasad, M. V. G., & Lakshminarayanan, P. A. (2012). Estimation of Oil Drain Life of Engine Oils in New Generation Diesel Engines in Tropical Conditions. *SAE International Journal of Fuels and Lubricants*, *5*(2). https://doi.org/10.4271/2011-01-2405

Prytz, R. (2014). Machine learning methods for vehicle predictive maintenance using off-board and on-board data. In *Thesis* (Vol. 9, Issue 9).

Pucher, H. (2010). *Klaus Mollenhauer Á Helmut Tschoeke Handbook of Diesel Engines*.

Raadnui, S. (2005). Machinery Health Monitoring Through Low Cost Maintenance Tribology Techniques. *The International Journal of Applied Mechanics …*, *15*(2).

Raadnui, S., & Kleesuwan, S. (2005). Low-cost condition monitoring sensor for used oil analysis. *Wear*, *259*(7–12), 1502–1506. https://doi.org/10.1016/j.wear.2004.11.009

Raposo, H., Farinha, J. T., Fonseca, I., & Ferreira, L. A. (2019). Condition monitoring with prediction based on diesel engine oil analysis: A case study for urban buses. *Actuators*, *8*(1), 1–15. https://doi.org/10.3390/act8010014

Rezende, S. O. (2003). *Sistemas Inteligentes - Fundamentos e Aplicações*. Manole.

Ricardo, L. O. (2014). *O que é Big Data*. http://luizricardo.org/2014/05/o-que-e-big-data-e-por-que-voce-deveria-estar-desesperadamente-interessado-nisso/

Rodrigues, J., Costa, I., Farinha, J. T., Mendes, M., & Margalho, L. (2020). Predicting motor oil condition using artificial neural networks and principal component analysis | Prognozowanie stanu oleju silnikowego za pomocą sztucznych sieci neuronowych i analizy składowych głównych. *Eksploatacja i Niezawodnosc*, *22*(3), 440–448.

Salgueiro, J. M., Peršin, G., Hrovatin, J., Juricic, D., & Vižintin, J. (2015). On-line detection of incipient trend changes in lubricant parameters. *Industrial Lubrication and Tribology*, *67*(6), 509–519. https://doi.org/10.1108/ILT-09-2013-0097

scikit-learn.org. (2021). *sklearn.ensemble.RandomForestClassifier*. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html#sklearn.ensemble.RandomForestClassifierl

Scott, D. W. (2019). Averaged Shifted Histogram. *Wiley StatsRef: Statistics Reference Online*, 1–9. https://doi.org/10.1002/9781118445112.stat00406.pub2

Sejkorová, M., Hurtová, I., Glos, J., & Pokorný, J. (2017). Definition of a motor oil change interval for high-volume diesel engines based on its current characteristics assessment. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, *65*(2), 481–490. https://doi.org/10.11118/actaun201765020481

Shadroo, S., & Rahmani, A. M. (2018). Systematic survey of big data and data mining in internet of things. *Computer Networks*, *139*, 19–47. https://doi.org/10.1016/j.comnet.2018.04.001

Shafi, U., Safi, A., Shahid, A. R., Ziauddin, S., & Saleem, M. Q. (2018). Vehicle remote health monitoring and prognostic maintenance system. *Journal of Advanced Transportation*, *2018*. https://doi.org/10.1155/2018/8061514

Sharma, B. C., & Gandhi, O. P. (2008). Performance evaluation and analysis of lubricating oil using parameter profile approach. *Industrial Lubrication and Tribology*, *60*(3), 131–137. https://doi.org/10.1108/00368790810871057

Sheng, C., Wu, T., & Zhang, Y. (2012). Non-destructive testing of marine diesel engines using integration of ferrographic analysis and spectrum analysis. *Insight: Non-Destructive Testing and Condition Monitoring*, *54*(7), 394–398. https://doi.org/10.1784/insi.2012.54.7.394

Siebert, J., Joeckel, L., Heidrich, J., Trendowicz, A., Nakamichi, K., Ohashi, K., Namba, I., Yamamoto, R., & Aoyama, M. (2021). Construction of a quality model for machine learning systems. *Software Quality Journal*, *0123456789*. https://doi.org/10.1007/s11219-021-09557-y

Silveira, E. L. C., Coelho, R. C., Neto, J. M. M., De Moura, C. V. R., & De Moura, E. M. (2010). Determinação de metais em óleos lubrificantes, provenientes de motores de ônibus urbano, utilizando a faas. *Quimica Nova*, *33*(9), 1863–1867. https://doi.org/10.1590/S0100-40422010000900008

Systems, C. P., Andronie, M., George, L., Iatagan, M., Ut, C., & Roxana, S. (2021). *Artificial Intelligence-Based Decision-Making Algorithms, Internet of Things Sensing Networks, and Deep Learning-Assisted Smart Process Management in Cyber-Physical Production Systems*.

Tauzia, X., Maiboom, A., Karaky, H., & Chesse, P. (2019). Experimental analysis of the influence of coolant and oil temperature on combustion and emissions in an automotive diesel engine. *International Journal of Engine Research*, *20*(2), 247–260. https://doi.org/10.1177/1468087417749391

Taylor, R. I., Mainwaring, R., & Mortier, R. M. (2005). Engine lubricant trends since 1990. *Proceedings of the Institution of Mechanical Engineers, Part J: Journal of Engineering Tribology*, *219*(5), 331–346. https://doi.org/10.1243/135065005X9718

Vigneswaran, E. E., Sujitha, S., & Praveen, R. (2021). *Model Based Design for Automobiles by using Modeling and Optimization of Engine Cooling System*. 821–827.

Viskup, R. (2019). Diesel and Gasoline Engines. In *Diesel and Gasoline Engines*. https://doi.org/10.5772/intechopen.75259

Vrachkov, D. G., & Todorov, D. G. (2018). Automotive Diagnostic Trouble Code (DTC) Handling over the Internet. *2018 IX National Conference with International Participation (ELECTRONICA)*, *15031*, 1–3.

Wakiru, J. M., Pintelon, L., Muchiri, P. N., & Chemweno, P. K. (2019). A review on lubricant condition monitoring information analysis for maintenance decision support. *Mechanical Systems and Signal Processing*, *118*, 108–132. https://doi.org/10.1016/j.ymssp.2018.08.039

Wakiru, J., Pintelon, L., Chemweno, P., & Muchiri, P. (2017). Analysis of lubrication oil contamination by fuel dilution with application of cluster analysis. *XVII International Scientific Conference on Industrial Systems*, 252–257.

Wang, F., Cheng, H., Zhou, K., Hu, Y., Zhang, C., & Kong, L. (2021). *Missing data filling method based on linear interpolation and lightgbm*. https://doi.org/10.1088/1742-6596/1754/1/012187

Wang, S., Wu, T., Wu, H., & Kwok, N. (2017). Modeling Wear State Evolution Using Real-Time Wear Debris Features. *Tribology Transactions*, *60*(6), 1022–1032. https://doi.org/10.1080/10402004.2016.1243746

Wang, T., Liu, J., Wan, C., & Wang, Z. (2021). Remote supervision strategy based on in-use vehicle OBD data flow. *E3S Web of Conferences*, *268*, 1–15. https://doi.org/10.1051/e3sconf/202126801007

Wang, W. (2007). A prognosis model for wear prediction based on oil-based monitoring. *Journal of the Operational Research Society*, *58*(7), 887–893. https://doi.org/10.1057/palgrave.jors.2602185

Wang, W., & Zhang, W. (2005). A model to predict the residual life of aircraft engines based upon oil analysis data. *Naval Research Logistics*, *52*(3), 276–284. https://doi.org/10.1002/nav.20072

Wang, Z., Xue, X., Yin, H., Jiang, Z., & Li, Y. (2018). Research progress on monitoring and separating suspension particles for lubricating oil. *Complexity*, *2018*. https://doi.org/10.1155/2018/9356451

Wei, L., Duan, H., Jin, Y., Jia, D., Cheng, B., Liu, J., & Li, J. (2021). Motor oil degradation during urban cycle road tests. *Friction*, *9*(5), 1002–1011. https://doi.org/10.1007/s40544-020-0386-z

Whitaker, D. A., Egan, D., OBrien, E., & Kinnear, D. (2018). *Application of Multivariate Data Analysis to machine power measurements as a means of tool life Predictive Maintenance for reducing product waste*.

Wikipedia. (2022). *ROC Curve*. https://upload.wikimedia.org/wikipedia/commons/1/13/Roc_curve.svg

Zadorozhnaya, E., Levanov, I., & Oskina, O. (2016). Study of HTHS Viscosity of Modern Motor Oils. *Procedia Engineering*, *150*, 602–606. https://doi.org/10.1016/j.proeng.2016.07.051

Zhang, C. (2003). A Micro-acoustic Wave Sensor for Engine Oil Quality Monitoring. *Proceedings of the Annual IEEE International Frequency Control Symposium*, 971–977. https://doi.org/10.1109/freq.2003.1275222

Zhang, Y., Zhou, B., Cai, X., Guo, W., Ding, X., & Yuan, X. (2021). Missing value imputation in multivariate time series with end-to-end generative adversarial networks. *Information Sciences*, *551*, 67–82. https://doi.org/10.1016/j.ins.2020.11.035

Zhu, J., He, D., & Bechhoefer, E. (2012). A survey of lubrication oil condition monitoring, diagnostics and prognostics techniques and systems. *Technical Program for MFPT 2012, The Prognostics and Health Management Solutions Conference - PHM: Driving Efficient Operations and Maintenance*, *July 2015*.

Zhu, J., Yoon, J., He, D., Qiu, B., & Bechhoefer, E. (2013). Online Condition monitoring and remaining useful life prediction of particle contaminated lubrication Oil. *PHM 2013 - 2013 IEEE International Conference on Prognostics and Health Management, Conference Proceedings*, *2008*, 1–15. https://doi.org/10.1109/ICPHM.2013.6621415

Zhu, X., Zhong, C., & Zhe, J. (2017). Lubricating oil conditioning sensors for online machine health monitoring – A review. *Tribology International*, *109*, 473–484. https://doi.org/10.1016/j.triboint.2017.01.015