"Predicting Remaining Useful Life of Air-Filters"

Submitted in Partial Fulfillment of the

Requirements for Degree of

MASTER OF TECHNOLOGY

IN

DATA SCIENCE

By

Jainil Patel

21MDS005

Under the Guidance of

Dr. Shilpa Pandey (Assistant Professor)

&

Ms. Nilima Gautam (Industrial Guide)

Department of Computer Science & Engineering,
School of Technology,
Pandit Deendayal Energy University,
Gandhinagar 382007

JULY-MAY, 2022-23

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBRIVIATIONS

Sr. No.	Abbreviations	Full Form
1.	RUL	Remaining Useful Lifeline
2.	PD	Pressure Drop
3.	FEMA	Failure Mode & Effect Analysis
4.	PHM	Prognostics & Health Monitoring
5.	CBM	Condition Based Monitoring
6.	PSO	Particle Swarm Optimization
7.	RBF	Radial Bases Function
8.	PdM	Predictive Maintenance
9.	SMH	Short Maintenance Haul
10.	GPM	Gaussian Process Model
11.	HHE	Hilbert Huang Entropy
12.	SVM	Support Vector Machine
13.	KNN	K-Nearest Neighbors
14.	DT	Decision Tree
15.	ML	Machine Learning
16.	MA	Moving Average
17.	ANN	Artificial Neural Network
18.	AI	Artificial Intelligence
19.	INR	Inner
20.	OTR	Outer
21.	KPA	Kilo-Pascal
22.	CNN	Convolution Neural Network

LIST OF ABBRIVIATIONS

Sr. No.	Abbreviations	Full Form
23.	LSTM	Long Short-Term Memory
24.	LR	Linear Regression
25.	WF	Weibull Fitter
26.	RT	Right
27.	LT	Left
28.	PCA	Principal Component Analysis
29.	OE	Original Equipment

TABLE OF CONTENT

Chapter 1 : INTRODUCTION

1.1 Background

Air filters play a crucial role in air pollution control, but their efficiency deteriorates over time, leading to increased energy consumption, reduced air quality, and higher maintenance costs. Predicting the remaining useful life of air filters can help mitigate these problems by enabling timely replacement and reducing operational costs.

The background section of this thesis provides context and essential information about air filters, their functions, and working principles. It begins by introducing the importance of air pollution control and the role of air filters in maintaining air quality. It then discusses the various types of air filters used in industrial, residential, and commercial settings, and their applications.

Furthermore, the background section explains the working principles of air filters, including the physical and chemical processes involved in capturing and removing airborne contaminants from the air. It also discusses the factors that affect the efficiency and performance of air filters, such as filter material, design, and operating conditions.

Overall, the background section serves as a foundation for the subsequent chapters of the thesis, providing readers with a comprehensive understanding of air filters and their significance in air pollution control.

1.1.1 Air-Filters

An air filter is a device that removes contaminants, such as dust, dirt, and debris, from the air that enters an engine. Air filters play a critical role in ensuring optimal engine performance and fuel efficiency, particularly in the mining industry, where heavy equipment such as mining trucks operate in dusty and harsh environments.

In mining trucks, air filters are typically located in the air intake system, which draws air from outside the vehicle and feeds it into the engine. The air filter traps particles and contaminants before they can enter the engine, preventing them from damaging the engine components or reducing its performance.

The air filter works by using a filter element, usually made of paper or foam, that is designed to capture particles of a certain size. As air passes through the filter element, the particles are trapped on the surface of the filter, forming a layer of dust and debris. Over time, this layer can become thick enough to reduce the flow of air into the engine, leading to reduced engine performance, increased fuel consumption, and elevated emissions.

To prevent this, air filters need to be cleaned or replaced regularly. In mining trucks, air filters are typically cleaned or replaced according to a maintenance schedule, which takes into account factors such as the operating conditions, the environment, and the age of the filter. Predictive maintenance models can help optimize these schedules by predicting when an air filter is close to its end of life and needs replacement or cleaning, minimizing downtime and unnecessary maintenance costs.

Figure 1.1 Different types of Air-Filters

1.1.2 Air-Filtering for Heavy loading vehicles

Off-highway diesel engines, including those used in heavy construction vehicles, mining machines, and agricultural equipment, often operate in environments with severe dust conditions. To ensure these engines run reliably, they require powerful and reliable air filtration systems.

Two-Stage Filtration System:

One solution to address this problem is a two-stage filtration system. The primary filter captures larger particles and prevents them from entering the engine. The secondary filter, located downstream from the primary filter, captures smaller particles that may have passed through the primary filter. This two-stage system provides enhanced filtration and helps protect the engine from dust and other contaminants.

This heavy-duty two-stage filtration system shown in Fig.1.2 is designed to ensure the longevity of off-highway engines operating in harsh environments. By capturing both larger and smaller particles, this system reduces engine wear and extends the time between maintenance intervals. This system is essential for maintaining the reliability and efficiency of diesel engines in heavy-duty off-highway applications.

Figure 1.2 S-Series Air Filter

Application:

- The heavy-duty two-stage filtration system is specifically designed for use in off-road, heavy-duty, and extreme dust conditions.
- The SSG 20 model allows for an airflow throughput of 1700 to 2400 cfm, while the SSG 29 models allow for an airflow throughput of 2580 to 4800 cfm.
- The filtration system is intended to be installed horizontally.
- This system shown in Fig. 1.3 is ideal for use in a variety of heavy-duty vehicles, including scrapers, earth movers, graders, and haul trucks.

Figure 1.3 Dust Cap for filters

1.1.3 Working of Air-Filter

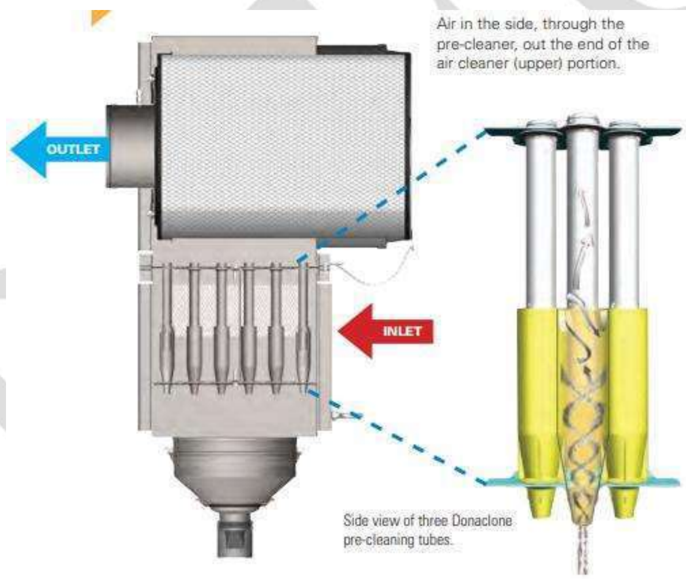


Figure 1.4 Air-filter workflow

The first stage of an air cleaner that utilizes a pre-cleaner system to remove dust and dirt from the incoming air.

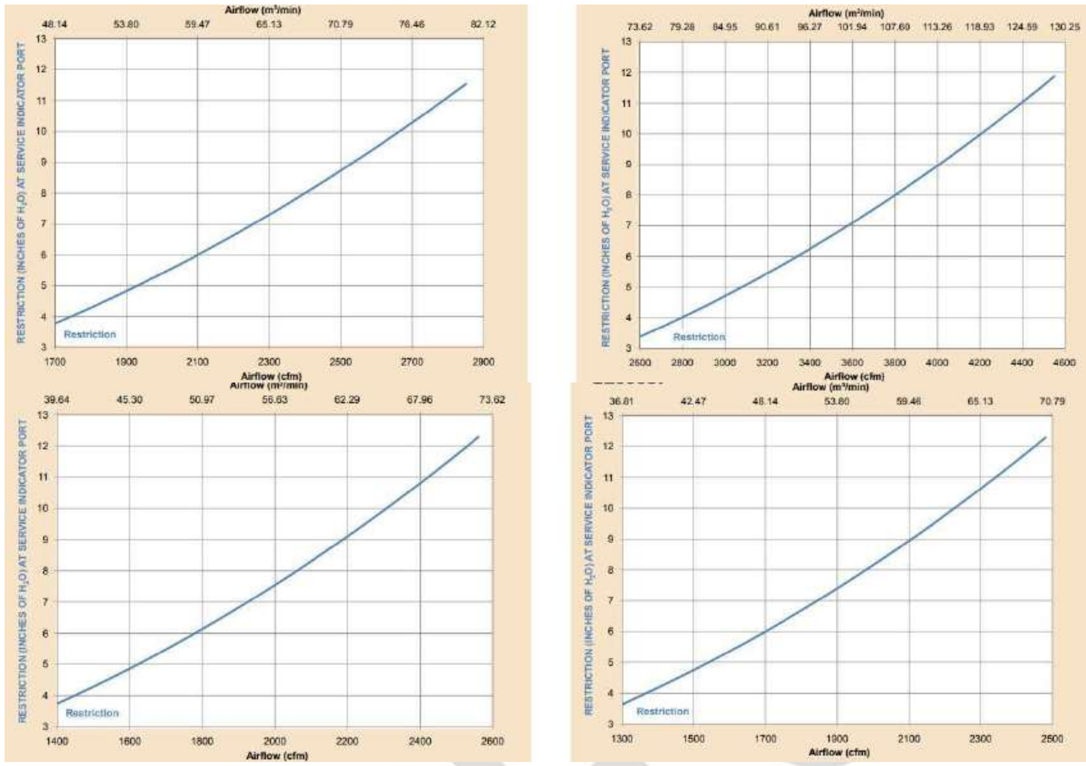


Figure 1.5: Airflow restriction indicators

The pre-cleaner system consists of hundreds of exclusive tubes, which spin the air to create a centrifugal force that separates up to 97% of the dust and dirt in the air.

They have no moving parts, meaning there is nothing to break down or maintain. As long as the engine is running, the pre-cleaner tubes will function properly.

Once the dust and dirt have been separated from the air, they are automatically ejected from the air cleaner using a Vacuator Valve, which is located below the lower housing body as shown in Fig 1.4. This ensures that the air cleaner continues to function effectively, and the air that enters the engine is clean and free of contaminants.

Graph shows that the relationship between the restriction (or pressure drop) across an air filter and the airflow rate through the filter. This plot is commonly used to monitor the performance of air filters in systems or other applications where air quality is important.

The airflow rate is plotted on the x-axis and the pressure drop across the filter is plotted on the y-axis. The plot typically shows a curve that starts at zero pressure drop (i.e., no filter installed) and increases as the airflow rate increases. As the filter becomes dirty and clogged with particulate matter, the pressure drop increases at a faster rate, causing the curve to bend and eventually level off as the filter reaches its maximum capacity.

The airflow graph, on the other hand, is a plot that shows the airflow rate through system or other air handling equipment as a function of time. This graph is used to monitor the overall performance of the system and to identify any changes in airflow that may indicate a problem.

Together, these two graphs can be used to identify when an air filter needs to be replaced. By comparing the current airflow rate to the expected airflow rate based on the restriction at service indicator plot, it is possible to determine if the filter is causing a reduction in airflow and needs to be replaced. This information can help ensure that the air quality in the space being served remains at a high level and that the system is operating efficiently.

1.1.4 Process behind replacement of Air-Filter

It is important to check the restriction level of the air filter regularly. The filter should only be replaced when the maximum recommended restriction level by the engine or equipment manufacturer has been reached. This ensures that the air filter is functioning optimally and prolongs its lifespan, saving you time and money in the long run.

Figure 1.6: Air- Filter restriction level checking

Dust cups and valves are components commonly used in industrial dust collection systems to capture and remove dust and other particulate matter from the air.

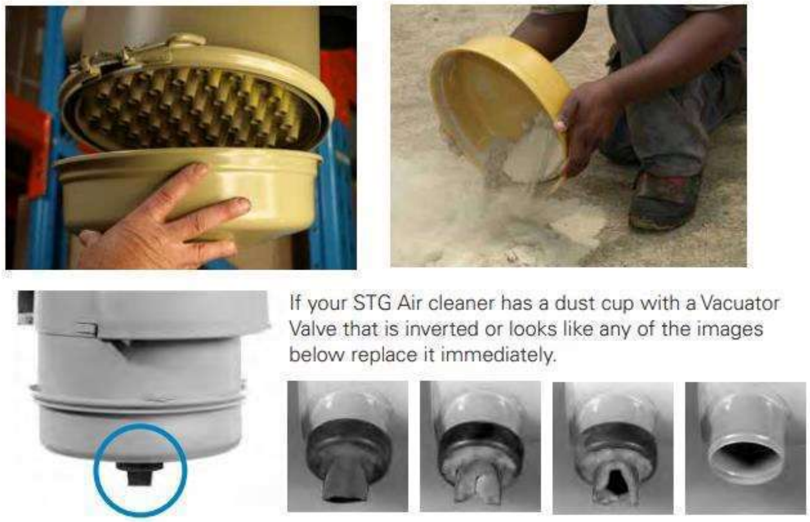


Figure 1.7: Dust cup with vacuator valves

A dust cup shown in Fig 1.7 is a container that is designed to collect dust and other debris that has been captured by a dust collection system. These cups are typically made of metal or plastic and can vary in size depending on the specific application. Once the dust cup is full, it can be emptied and cleaned, allowing the dust collection system to continue operating efficiently.

Installation of air filters is shown below:

Figure 1.8: Air-Filter Installation

Valves are used to control the flow of air and dust through the dust collection system. These valves are typically operated manually or automatically and can be used to isolate specific areas of the system, regulate airflow, or control the discharge of collected dust.

1.1.5 Air-Filter Indicators

A variety of filtration monitoring systems called Filter Minder are created to deliver the best performance and value. Worldwide Original Equipment Manufacturers for both on- and off-road vehicles rely on its sensors, indicators, and switches. Equipment operates more effectively and efficiently when maintenance managers can plan filter changes methodically and are aware of any abrupt changes in restriction thanks to the information provided by Filter Minder devices.

With a variety of indicator kinds, mounting options, and fitting styles, the range provides the market's most extensive portfolio of filtration monitoring solutions. In order to fulfil ever-tighter production schedules, this enables heavy equipment manufacturers, maintenance managers, fleet managers, and owner-operators to maximize equipment efficiency, uptime, and performance.

The newest connectivity technologies, which enable real-time filter data to be acquired, analyzed, and shown with actionable information, are one of Filter Minder's most important features. The devices are utilized by numerous OE platforms and fleets around the world and are composed of sturdy, high-quality materials.

The visual indications, sensors, and switches that are part of the Filter Minder technology. Graduated Indicators, which display when filters are getting close to their limitation limits, and Single Position Indicators, which alert you when it's time to change your air filter, are the two different sorts of visual indicators. The switches include Switch Only, which sends an electrical monitoring signal to a remote "time to service filter" light situated in the equipment cab, and Visual Indicator + Switch, which provides both visual and electrical monitoring.

Figure 1.9: Air Filter Indicators

As a filtration monitoring technology, Filter Minder offers Sensors + LED Displays, which allows electronic limitation monitoring with a continuous signal supplied to a remote in-cab device for display. This system gives equipment operators constant feedback about restrictions, giving them the greatest amount of freedom to keep an eye on their equipment's needs. Standard, Heavy-duty, and Hall-Effect sensors are the three types of sensors offered by Filter Minder.

1.2 Motivation

Air-channel is a piece of the vehicle which assists with decontaminating the air coming from motor. These days, apparatus assumes a significant part in the mining business which utilize various types of air channels for cleaning, and over the long run, these machines and the related gear might weaken, and once in a while even the whole hardware might fizzle. Breakdowns truly influence the exhibition and cost of creation lines and frequently lead to an emotional decrease of accessibility due to the exorbitant upkeep time frame.

For staying away from these disappointment cases, the support of assets is many times arranged ahead of time. In any case, the support cost of certain ventures can increment by up to 70% of the complete expense. Thusly, the decrease of upkeep costs is viewed as an essential and significant benefit to the maker in an exceptionally serious assembling area, such as, the semiconductor business. In numerous modern areas, for example, auto producing, support the executives is an express essential issue to make vital moves on time. Fixing the creation line after the breakdown can be more expensive than directing preventive upkeep in front of the breakdown. Likewise, the update of the creation plan can cause fluctuation in assistance and item quality.

From an AI point of view, there are a few difficulties of preventive upkeep in air channels. To begin with, it is for sure challenging to secure machine glitch information and name the disappointment case by and by in the datasets. Second, there exists an enormous measure of cycle information (i.e., large information) created underway lines, and handling of this huge information requires a unique framework, master information, and custom brilliant programming. Last, many organizations don't share this kind of information openly because of information security, and accordingly, scientists in this field can't approve new models with more datasets. To this end, there is a requirement for additional exploration to think of powerful measures and new models to execute prescient support. So main motivation is to go with RUL.

Remaining useful life (RUL) is a vital measurement and basic to foreseeing the disappointment of a machine in the creation line. The test of RUL expectation is that RUL isn't generally named in the preparation dataset, and thusly, regulated learning calculations of AI can't be applied for this situation. A wellbeing record should be accurately characterized and interjected to plan the connection among highlights and the RUL. After this insertion step, an AI based model can be utilized to foresee the wellbeing record by learning the introduced information precisely. The AI based approach is supposed to deal with the unfavorable effects of clamor in the dataset and conceivable sensor issues (i.e., sensor float) that could emerge during the activity of the creation line.

1.3 Objectives

The principal objective of our review is to limit the unfavourable impacts of breakdowns and fabricate a clever ML based RUL expectation model. We propose and approve another ML based model in foreseeing the disappointment of gear (i.e., RUL forecast) in air channels and dissect the pertinence of ML calculations in anticipating the disappointment of hardware. During our trials, different statistic calculations, pre-handling and component choice procedures, and boundary streamlining approaches are examined to construct a clever model to foresee the gamble of Air channels.

The idea of RUL is utilized to assess the gamble of creation line breakdown. For the contextual analysis, the NASA dataset on super motors has been utilized in this review. The contextual investigation exhibits the viability of our forecast model to foresee the RUL inside the extent of prescient support.

The objectives of this research are as follow:

a) Find the behaviour of the data using statistical modelling
b) Extract the cycle from the data using mathematical equation.
c) Find the rate of increment in differential pressure based on the time using linear model.
d) Apply mathematical equation to estimate the RUL of the filters.

1.4 Dissertation Outline

The outline of the dissertation work is as follow:

- Chapter 1 Introduction provides the information about the problem. This chapter also includes the background information on domain knowledge.
- Chapter 2 Literature review provides information about the existing approaches in context of predicting RUL & statistical models used for air filters.
- Chapter 3 Proposed systems describe the proposed method and algorithm to predict the remaining useful life of air-filters.
- Chapter 4 Result and discussion present the evaluation of the implemented method.
- Chapter 5 Conclusion and future scope summarize and concludes the dissertation while outlining the future scope of the proposed system

1.5 Introduction to Machine learning and its types

1.5.1 Machine learning

Machine learning is a field of study that enables computers to learn from data and make decisions without being explicitly programmed. It automates the process of building analytical models and identifies patterns that would otherwise be difficult or impossible to detect. While it falls under the umbrella of artificial intelligence, machine learning is unique in that it has the ability to evolve and improve upon its previous iterations by learning from new data.

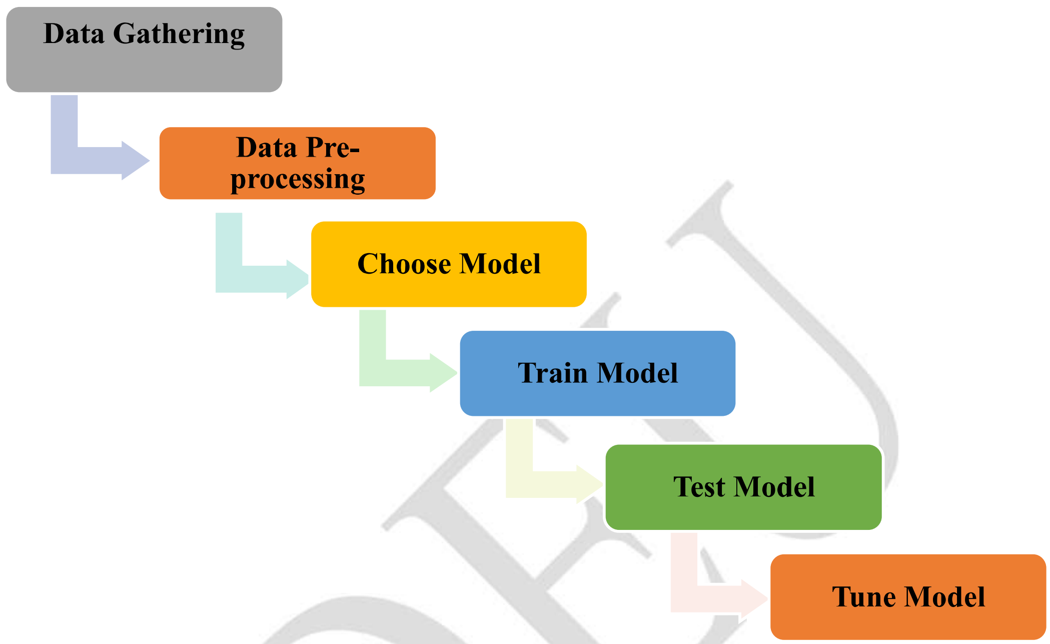


Figure 1.10: Machine Learning flowchart

Using various programming techniques, machine learning algorithms can process large amounts of data and extract valuable insights. This allows them to make predictions and decisions with minimal human intervention. Overall, the goal of machine learning is to enable computers to learn and adapt to new information, ultimately improving their performance and accuracy over time.

1.5.2 Type of machine learning algorithms

Basically, machine learning categorizes in three types as shown in Fig. 1.11:

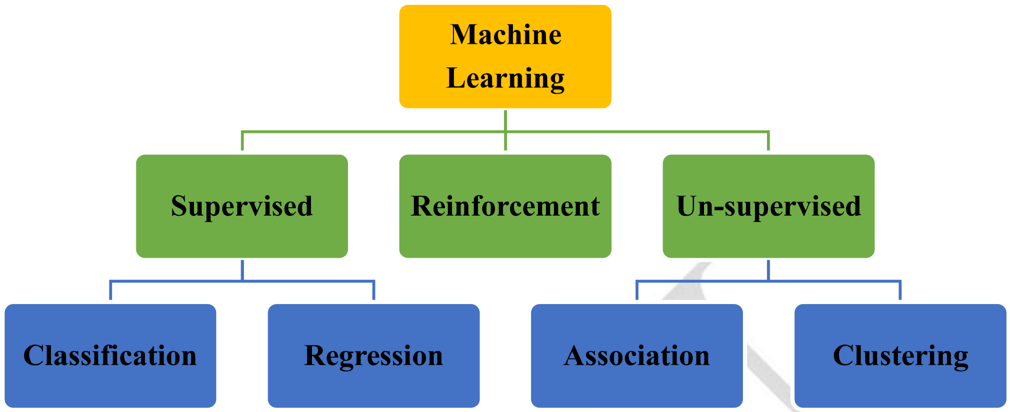


Figure 1.11: Types of Machine Learning

1.5.3 Supervised learning

Supervised learning is a subfield of machine learning in which an algorithm is trained using labelled data, with the aim of producing accurate predictions or decisions when presented with new, unseen data. The name "supervised" learning comes from the fact that the algorithm is supervised by a teacher or supervisor, who provides the correct answers for the labelled data.

Supervised learning algorithms learn by example, using training data consisting of inputs paired with their correct outputs. During training, the algorithm searches for patterns in the data that are correlated with the desired outputs. Once the algorithm has been trained, it can be used to make predictions or decisions for new, unseen inputs by labelling them with the most likely output.

There are two main subcategories of supervised learning: classification and regression. In classification, the algorithm is trained to assign inputs to one of several predefined categories based on their characteristics. In regression, the algorithm is trained to predict a continuous numerical output based on the input data. Both types of algorithms are widely used in various applications, such as image recognition, speech recognition, and financial forecasting.

Classification

Classification is a subcategory of supervised learning, where the algorithm is trained to assign input values to one of several predefined categories or classes. The classification algorithm's primary task is to assign a class label to an input value based on the training data provided.

A common example of a classification problem is email filtering, where the algorithm is trained to differentiate between spam and non-spam emails. This is known as a binary classification problem, where the algorithm has to choose between two classes. The algorithm is trained using a dataset that includes both spam and non-spam emails, and the features that determine whether an email is spam or not are learned by the algorithm. Once the algorithm is trained, it can be used to classify new, unseen emails as either spam or not spam. Other examples of classification problems include fraud detection and image classification.

Several algorithms can be used to solve classification problems, depending on the data and situation. Some popular algorithms for classification include linear classifiers, support vector machines, decision trees, K-nearest neighbours, and random forests. The choice of algorithm depends on the specific problem and the characteristics of the dataset.

Regression

Regression is a statistical method used to determine the relationship between a dependent variable (represented as Y) and a set of independent variables (represented as X). It is commonly used in finance, investing, and other fields to predict outcomes based on data. Regression algorithms attempt to find the relationship between the dependent and independent variables in order to predict a continuous number, such as sales, income, or test scores.

There are two basic types of regression: simple and multiple linear regression. Simple linear regression uses one independent variable to predict the outcome of the dependent variable Y, while multiple linear regression uses two or more independent variables to predict the outcome. Non-linear regression methods can also be used for more complex data analysis.

Regression is widely used in various applications, such as weather forecasting, risk assessment, and score prediction. In simple linear regression, the relationship between the dependent variable and independent variable is typically represented as a straight line that best fits all the data points. Multiple regression can involve differentiating separate variables using subscripts.

There are different types of regression algorithms, with three of the most common being linear regression, logistic regression, and polynomial regression. Linear regression is used for predicting continuous values, while logistic regression is used for predicting categorical values. Polynomial regression is used for fitting a curved line to the data, which can be useful for more complex relationships between the variables.

1.5.4 Unsupervised Learning

Unsupervised learning is a type of machine learning that differs from supervised learning in that it does not require labeled data. In unsupervised learning, an algorithm is given a large amount of data and is tasked with identifying patterns or relationships within the data. The goal is to group or cluster the data in a way that is meaningful and useful for analysis or decision making.

Unsupervised learning can be a powerful tool for extracting insights from unstructured or unlabelled data, which is common in many industries. There are two main types of unsupervised learning: association and clustering. Association refers to finding patterns or relationships between variables, such as market basket analysis or text mining. Clustering involves grouping data points based on their similarities, such as in medical research or targeted marketing.

Overall, unsupervised learning has the potential to unlock valuable insights and improve decision making in a wide range of applications.

1.5.5 Reinforcement Learning

Reinforcement learning is a type of machine learning that involves taking actions in an environment to maximize a cumulative reward. The reinforcement agent interacts with the environment by taking actions and receiving rewards or punishments based on its actions. The goal of the agent is to learn the best policy or set of actions to maximize the long-term reward.

Unlike supervised learning, where the model is trained on labeled data, in reinforcement learning, the agent learns from trial and error without prior knowledge of the correct output. The agent learns by receiving feedback in the form of rewards or punishments for its actions, which helps it update its policy to improve future decisions.

Reinforcement learning has numerous applications, such as robotics, gaming, finance, and healthcare. For example, reinforcement learning can be used to teach a robot to navigate a maze, or to optimize a portfolio of investments.

Overall, reinforcement learning is a powerful technique that allows machines to learn from their own experiences and make decisions that optimize long-term rewards.

Chapter 2 : LITERATURE REVIEW

The literature review section of this thesis summarizes and analyzes the existing research on predicting the remaining useful life of air filters. It begins by identifying the key research questions and objectives addressed in this field, and then provides an overview of the methods and techniques used in previous studies.

2.1 Remaining Useful Life

Remaining useful life (RUL) is the estimated amount of time that a system or component will continue to function before it fails or requires maintenance. It is a key parameter in asset management, predictive maintenance, and reliability engineering, as it can be used to optimize maintenance schedules, reduce downtime, and extend the lifespan of equipment.

The RUL of a system or component can be estimated using a variety of techniques, including data-driven approaches, model-based approaches, and hybrid approaches. These techniques involve analyzing historical data on the system or component, as well as current operating conditions, to generate a probabilistic estimate of the time remaining until failure.

Accurately estimating the RUL of a system or component is important for optimizing maintenance schedules and reducing maintenance costs. If maintenance is performed too early, it can result in unnecessary downtime and increased costs. If maintenance is performed too late, it can result in unexpected failures and increased repair costs. By accurately estimating the RUL, maintenance can be scheduled at the optimal time to maximize the useful life of the system or component.

2.2 Techniques to Detect Remaining Useful Life

2.2.1 Data-Driven Approach

These approaches involve using historical data on the system or component to train machine learning models to predict the RUL. This can include techniques such as regression, decision trees, support vector machines, and neural networks.

2.2.2 Model Based Approach

These approaches involve using mathematical models to simulate the behaviour of the system or component over time. This can include techniques such as physics-based modelling and system identification.

2.2.3 Hybrid Approach

These approaches combine data-driven and model-based techniques to improve the accuracy of RUL detection.

2.2.4 Failure Mode and Effect Analysis (FEMA)

FMEA is a technique that identifies potential failure modes in a system or component and evaluates their effects on the system or component. This can be used to estimate the RUL by tracking the progression of the identified failure modes over time.

2.2.5 Prognostics and health management (PHM)

PHM is a system-level approach that integrates multiple sensing, monitoring, and diagnostic technologies to assess the health of a system or component and predict its RUL.

2.2.6 Condition-based monitoring (CBM)

CBM is a technique that involves monitoring the condition of a system or component in real-time using sensors and other monitoring technologies. This data can be used to estimate the RUL by detecting changes in the system or component's behavior that indicate impending failure.

2.3 Literature based on ML models

R. Ghimire, K. R. Pattipati and P. B. Luh present Fault diagnosis and augmented reality-based troubleshooting of HVAC systems refer to using advanced technologies to identify and address problems in heating, ventilation, and air conditioning (HVAC) systems.

which involves analyzing data from the HVAC system to identify any issues that may be affecting its performance. Process involves the use of sensors, software, and other tools to monitor and measure various aspects of the system, such as temperature, pressure, and airflow.

By analyzing this data, they pinpoint the source of any problems and take corrective action.
Augmented reality-based troubleshooting involves using technology to overlay digital information on top of the real-world environment. This can include using tools such as smartphones or tablets to display virtual images or instructions that help technicians diagnose and fix HVAC system problems. By providing visual aids and step-by-step instructions, augmented reality-based troubleshooting can simplify complex repair tasks and reduce the risk of errors.

Lei et al. (2018) provides a systematic review of the entire process from data acquisition to RUL prediction. The review includes various approaches such as data-driven methods and model-based methods [1].

Xi et al. (2022) investigates the water hammer characteristic of stalling fluid in eccentric casing-tubing annulus, which is related to the health of machinery [2].

Chao et al. (2022) present an adaptive decision-level fusion strategy for fault diagnosis of axial piston pumps and an integrated slipper retainer mechanism to eliminate slipper wear in high-speed axial piston pumps [3][4].

tTang et al. (2021, 2022) propose improved convolutional neural networks with adaptable learning rates and adaptive deep learning models for fault diagnosis of hydraulic piston pumps using multi-signal data, acoustic images, and pressure signals [5][7][8].

Yuan et al. (2022) presents a theory model of dynamic bulk modulus for aerated hydraulic fluid, which is relevant to the performance and health of hydraulic systems.

Li et al. (2015) and Wang et al. (2018) propose improved models for predicting the RUL of rolling element bearings and a nonlinear degrading system in service, respectively.

Nian (2018) provides different viewpoints on prognostics and health management, which can help provide context for the research in this field.

Liu et al. (2021) and Pei et al. (2019) review the current research and challenges of deep learning-based RUL prediction methods for equipment, which can help identify gaps in the existing literature and future research directions.

Peng et al. (2019) proposes a new nonlinear degradation modeling and residual life prediction method, which is relevant to the RUL prediction of machinery.

2.4 Algorithms used in this research work

2.4.1 Decision Tree (DT)

A decision tree (DT) is an algorithm used for decision support, where a graph in the form of a tree is drawn to model decisions and their possible consequences, such as resource costs, chance event outcomes, and utility. This allows the algorithm to be displayed using only conditional control statements. Each internal node, branch, and leaf node in the flowchart represents a "test" on an attribute, the outcome of the test, and a class label, respectively, and the classification rules are represented by paths followed from the root to the leaf.

DT-based methods offer stable, highly accurate predictive models that are easy to interpret, making them one of the best and most commonly used supervised learning approaches. DTs can also effectively capture non-linear relationships in addition to linear models, and they can be adapted to solve a wide range of problems, whether classification or regression. DT algorithms are known as Classification and Regression Trees (CART), and "the possible solutions to a given problem arise as the leaves of the tree, with each node representing a point of deliberation and decision."

Decision trees come with a set of common terms, including:

- Root Node: Represents the entire sample or population and is divided into two or more homogeneous subsets.
- Splitting: The process of dividing a node into sub-nodes.

- Decision Node: A sub-node resulting from splitting, which is also divided into sub-nodes.
- Leaf/ Terminal Node: A sub-node that does not split further.
- Pruning: The removal of sub-nodes from a decision node, opposite to splitting.
- Branch / Sub-Tree: A sub-section of the entire tree.
- Parent and Child Node: A node that splits into sub-nodes is called the parent, while the sub-nodes are referred to as children.

Decision trees have been widely used in various fields, such as urban planning, education, law, energy, architecture, economics, pharmacy, healthcare, and industry, due to their simplicity and versatility. This technique is commonly used in supervised learning with pre-defined target variables, both continuous and categorical, as input and output parameters. The data sample is partitioned into homogeneous subsets based on the most significant decisive variable, which acts as a splitter or differentiator.

There are two types of decision trees based on the type of target variable: A. Categorical variable decision tree: This type of DT has a categorical target variable. An example of a categorical variable decision tree is "Will children play Ludo?" where the target variable is YES or NO. B. Continuous variable decision tree: This type of DT has a continuous target variable.

The decision tree algorithm uses the tree representation to solve the problem. The best attribute of the dataset is placed on the tree root, and the training dataset is split into subsets in such a way that the data of a subset has the same value of the attribute. Both these steps are repeated on each subset until leaf nodes are found in all the branches of the tree. To predict a class mark for a document in decision trees, one should start from the root of the tree. The value of the root attribute is then compared with the database attribute, and the branch corresponding to the value is followed based on the comparison, moving to the next node. This process of continuous comparison of record's attributes with other internal nodes of the tree is repeated until a predicted class value of a leaf node is obtained.

The following assumptions are taken into consideration while creating a decision tree: a) In the beginning, the entire training set is considered as the root. b) Categorical feature values are preferred, and if the values are continuous, they should be discretized before building the model. c) Records are distributed recursively based on attribute values. d) A statistical approach is used to place attributes as the root or internal node of the tree.

Advantages of Decision Tree:

1. Easy to understand: Decision trees are represented graphically and can be easily understood by people with no statistical knowledge.
2. Useful in Data exploration: Decision trees can be used to identify the most significant variables and their relationships, and can be used to create new variables that may better predict the target variable.
3. Variable screening or feature selection: Decision trees can be used for variable screening or feature selection, indirectly.
4. Requires relatively small efforts for data preparation: Decision trees require less data cleaning as compared to other modelling techniques and are not influenced greatly by missing values and outliers.
5. Handles both numerical and categorical variables: Decision trees can handle both numerical and categorical variables, and can be used for multi-output problems.
6. Non-parametric method: Decision trees have no assumptions about classifier structure and space distribution.
7. Not affected by non-linear relationships: The performance of decision trees is not affected by non-linear relationships between parameters.
8. Few hyper-parameters to be tuned: Decision trees have very few hyper-parameters to be tuned.

Disadvantages of Decision Tree:

1. Over-fitting: Decision trees can over-fit the data resulting in poor generalization on new data. This problem can be overcome by setting constraints on model parameters and pruning.

2. Not suitable for continuous variables: Decision trees may lose information while working with continuous numerical variables when categorizing them into different categories.
3. Unstable: Small variations in the data can result in generation of a completely different tree, making them unstable. Variance needs to be lowered by approaches like bagging or boosting.
4. Not globally optimal: Greedy algorithms used in decision tree construction do not guarantee a globally optimal decision tree. This problem can be mitigated by training various trees where features and samples are arbitrarily sampled with replacement.
5. Biased trees: Decision tree learners can create biased trees if some classes dominate. Therefore, it is recommended to balance the dataset prior to fitting in DT.
6. Biased response: Information gain gives a biased response for attributes with a greater number of categories in a DT with categorical variables.
7. Low accuracy: Generally, accuracy is lower in prediction for a dataset over other machine learning algorithms.
8. Complex calculations: If there are many class labels, calculations can become complex.

2.4.2 K-Nearest Neighbors (KNN)

The K-nearest neighbors (KNN) algorithm is a supervised machine learning algorithm that can be used for both classification and regression predictive problems. However, it is commonly used in industry for classification problems. KNN is a lazy learning algorithm, meaning it does not have a specialized training phase and instead uses all the data for training. It is also a non-parametric learning algorithm, as it does not make assumptions about the underlying data.

To predict values for new data points, KNN uses feature similarity, meaning the value of a new data point is assigned based on how closely it matches points in the training set. The KNN algorithm can be implemented in the following steps:

(1) load the training and test data.

(2) choose a value for K (any integer), which represents the number of nearest data points to consider.

(3) for each point in the test data, calculate the distance between the test data and each row of training data using a distance method such as Euclidean, Manhattan, or Hamming distance. Then, sort them in ascending order based on the distance value, choose the top K rows from the sorted array, and assign a class to the test point based on the most frequent class of these rows.

To choose the right value for K, the KNN algorithm is run multiple times with different K values, and the K value that reduces the number of errors is chosen. When choosing K, it is important to keep in mind that decreasing K to 1 can make predictions less stable, and increasing K beyond a certain point can lead to more errors. K is typically chosen as an odd number to break ties in cases where majority vote is considered (e.g., picking the mode in a classification problem).

KNN has several advantages, including its simplicity and ease of implementation, its ability to handle nonlinear data, and its versatility for regression, classification, and search problems. However, it also has some disadvantages, such as being computationally expensive, requiring high memory storage, and being sensitive to the scale of data and irrelevant features

2.4.3 Support Vector Machine (SVM)

The Support Vector Machine (SVM) algorithm is used to find a hyperplane in an N-dimensional space (N is the number of features) that can accurately classify data points into different classes. The main objective is to find a hyperplane with the maximum margin, which refers to the maximum distance between data points of different classes. This increases the confidence in classifying future data points. The hyperplane is a decision boundary that separates data points into different classes, and its dimension is determined by the number of features. When the number of features is 2, the hyperplane is a line; when it is 3, the hyperplane becomes a two-dimensional plane. As the number of features increases beyond 3, it becomes difficult to visualize.

Support vectors are data points that are closest to the hyperplane and influence the orientation and position of the hyperplane. The margin of the classifier can be maximized by using these support vectors. If the support vectors are removed, the position of the hyperplane will be changed.

In logistic regression, the output of a linear function is squashed using the sigmoid function to obtain a value within the range of [0,1]. If this value exceeds a threshold value of 0.5, the label 1 is assigned; otherwise, the label 0 is assigned. In SVM, the output of the linear function is used to identify the classes based on their output values of 1 and -1. By changing the threshold values to 1 and -1 in SVM, a range of values [-1,1] is obtained that acts as the margin.

The SVM model represents different classes of a hyperplane in a multidimensional space. The objective of SVM is to generate a hyperplane iteratively that can accurately separate the classes and find a maximum marginal hyperplane (MMH). Support vectors are used to define the separating line.

SVM is implemented using a kernel, which transforms a low-dimensional input space into a higher-dimensional space. This technique is known as the kernel trick. By adding more dimensions, non-separable problems can be converted into separable problems, making SVM more powerful, accurate, and flexible. SVM uses different types of kernels such as linear, polynomial, radial basis function (RBF), and sigmoid.

Linear Kernel

The formula of linear kernel is using a dot product between any two observations:

$$\mathbf{K(x, x_i) = sum(x * x_i)}$$

It is clear from the above equation that the product between two vectors say x & x_i is the sum of the multiplication of each pair of input values.

Polynomial Kernel

Polynomial kernel is the more generalized form of linear kernel and differentiate curved or nonlinear input space:

$$\mathbf{K(x, x_i) = 1 + sum(x * x_i)^d}$$

Where d is the degree of polynomial, which need to specify manually in the learning algorithm.

Radial Basis Function (RBF) Kernel

In SVM classification, RBF kernel is mostly used, maps input space in indefinite dimensional space, mathematically it can be expressed as**:**

$$\mathbf{K}(\mathbf{x}, \mathbf{x_i}) = \mathbf{e}^{-\gamma * \mathbf{sum}(\mathbf{x} - \mathbf{x_i}^{2})})$$

Where, γ ranges from 0 to 1, which need to specify manually in learning algorithm. A good default value of gamma is 0.

Chapter 3 : RESEARCH METHODOLOGY & IMPLEMENTATION

Problem Statement: The estimation of remaining useful life of Engine Air-Filters in mining industry is a primary concern. The purpose of using Air-Filters is to remove harm dust particles from entering the engine components. Due to continues usage the filter plug with contamination, thus the restriction to air flow increases. The restriction in airflow can be measured as the pressure inside the engine component. For heavy vehicles like mining trucks, it is assumed the filter gets plugged at 7.5 KPA and it is recommended to change the filter.

In current scenario the operators are not able to identify the life time and an unplanned shutdown occurs which last for 7-10 working days for a complete filter replacement.

Our approach proposes a solution for the problem by predicting the remaining useful life of the filter. The method identifies the cycle change and starts the estimating the RUL based on the rate of KPA increment. The flow diagram of this research work shown in below figure:

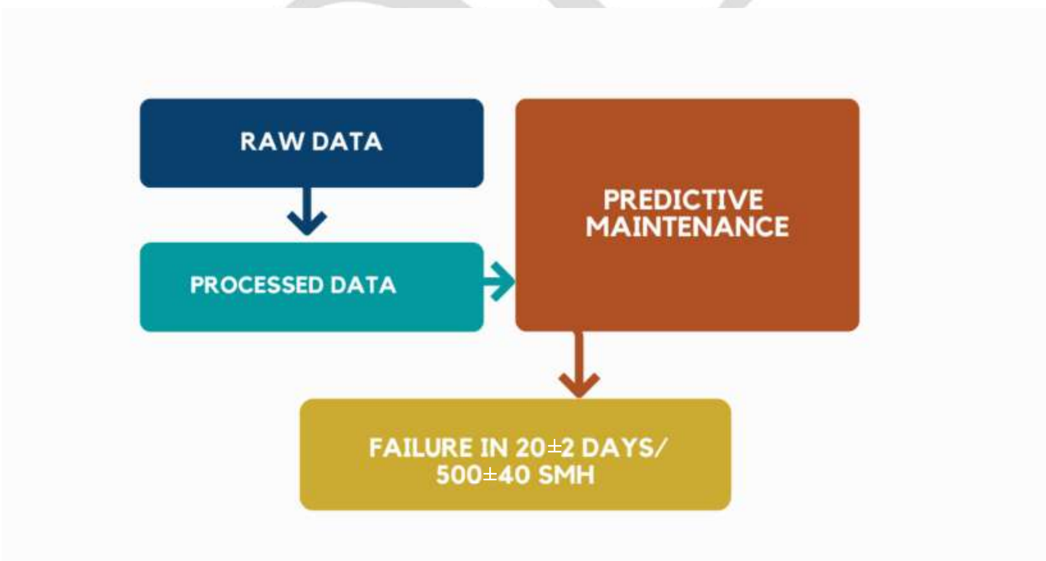


Figure 3.1: Model Workflow

3.1 Predictive Maintenance

Predictive maintenance (PdM) is a maintenance strategy that uses data analysis tools and techniques to monitor the condition of equipment and predict when maintenance is required. The goal of PdM is to maximize the reliability and availability of equipment while minimizing maintenance costs and downtime.

PdM relies on a variety of sensing and monitoring technologies to collect data on equipment performance, including vibration analysis, acoustic analysis, temperature monitoring, and oil analysis. This data is then analyzed using machine learning algorithms, statistical models, and other advanced analytics techniques to identify patterns and trends that can be used to predict when maintenance is required.

PdM is different from other maintenance strategies such as reactive maintenance, where equipment is only repaired when it fails, and preventive maintenance, where maintenance is performed on a schedule regardless of the equipment's condition. PdM is more proactive, as it uses data analysis to predict when maintenance is required before equipment failure occurs.

The benefits of PdM include increased equipment reliability, reduced maintenance costs, increased safety, and improved operational efficiency. By performing maintenance only when it is required, PdM helps to minimize the amount of downtime and lost productivity due to maintenance activities.

Figure 3.2: Mining Vehicle Prototype

3.2 Short Maintenance Haul

"Short maintenance haul" (SMH) is a term used in the mining industry to describe a specific type of mining truck that is designed for short trips between the mining face and the maintenance area. These trucks are typically smaller and more agile than other mining trucks, and are used to transport personnel, equipment, and materials quickly and efficiently.

The term "haul" refers to the transportation of materials, such as ore or waste rock, from one location to another within a mine site. In the context of mining trucks, a haul can refer to a single trip from the mining face to the processing plant, or to a series of trips made by a truck over a period of time.

In contrast to other types of mining trucks, which may be designed for long hauls or heavy loads, SMH trucks are specifically designed to move quickly and efficiently between the mining face and the maintenance area. These trucks may be used to transport workers, tools, and equipment to the mining face, or to bring ore or other materials to the processing plant for further treatment.

SMH trucks may also be used to transport materials from the processing plant back to the mining face for use in operations. Because these trucks are designed for short trips and quick turnarounds, they are typically more agile and maneuverable than other types of mining trucks, making them well-suited for use in tight spaces or areas with limited access.

Overall, the term "short maintenance haul" refers to a specific type of mining truck that is designed for short trips between the mining face and the maintenance area, and is an important tool in the mining industry for moving personnel, equipment, and materials quickly and efficiently.

3.3 KPA in Air Filters

Air filters are an important component of normal vehicles, as they help to protect the engine from harmful contaminants and particles that can cause damage or reduce efficiency. When air enters the engine, it can bring with it dirt, dust, and other debris, which can accumulate over time and cause wear and tear on the engine components.

The air filter is typically located in the air intake system of the vehicle, between the outside air and the engine. As air flows through the filter, contaminants are trapped in the filter media, allowing clean air to pass through to the engine. Over time, the filter can become clogged with dirt and debris, reducing airflow and causing a decrease in engine performance.

Regular maintenance of the air filter is essential to ensure optimal engine performance and longevity. Most vehicle manufacturers recommend that the air filter be replaced every 15,000 to 30,000 miles, depending on driving conditions and other factors. However, if you drive in particularly dusty or dirty conditions, you may need to replace the filter more frequently.

There are several types of air filters available for normal vehicles, including paper filters, foam filters, and cotton gauze filters. Paper filters are the most common type of filter, and are typically the least expensive. They are designed to capture large particles and debris, but may not be as effective at filtering smaller particles.

Foam filters are designed to capture smaller particles and are often used in off-road vehicles or in dusty environments. They require more maintenance than paper filters, as they need to be cleaned and re-oiled periodically.

Cotton gauze filters are designed for high-performance engines, as they offer better airflow and filtration than other types of filters. They are typically more expensive than other types of filters and require periodic cleaning and re-oiling.

Overall, air filters are an essential component of normal vehicles, helping to protect the engine from harmful contaminants and particles. Regular maintenance of the air filter is essential to ensure optimal engine performance and longevity, and choosing the right type of filter for your driving conditions can help to maximize performance and efficiency.

In the context of air filters in normal vehicles, KPA (kilo pascal) is a unit of measurement that is used to indicate the pressure drop across the air filter. The pressure drop is the difference in pressure between the air intake and the engine compartment, and is a measure of how effectively the air filter is working.

As air passes through the filter, it encounters resistance from the filter media, which can cause a decrease in pressure on the engine side of the filter. The pressure drop across the filter is measured in units of KPA, and can be used to determine the efficiency of the filter and the amount of airflow that is being restricted.

In general, a lower KPA value indicates that the filter is working more efficiently, as there is less resistance to airflow. A higher KPA value indicates that the filter is becoming clogged with dirt and debris, and may need to be replaced.

Most vehicle manufacturers recommend that the air filter be replaced when the pressure drop across the filter reaches a certain threshold, typically around 10-20 KPA. This threshold may vary depending on the specific vehicle and driving conditions.

In summary, KPA is a unit of measurement that is used to indicate the pressure drop across the air filter in normal vehicles. A lower KPA value indicates that the filter is working more efficiently, while a higher KPA value indicates that the filter may need to be replaced due to clogging with dirt and debris.

Figure 3.3: Air Filter

3.3.1 Calculation of KPA

In mining trucks, the calculation of KPA (kilo pascal) involves measuring the pressure drop across the air filter in the same way as in normal vehicles. However, due to the larger size and higher performance requirements of mining trucks, the pressure drop threshold for replacing the air filter may be different than in normal vehicles.

To calculate KPA in mining trucks, a pressure gauge is required that can measure the pressure drop across the air filter. This can typically be done by connecting the gauge to a port on the air intake system, either before or after the air filter.

To obtain the pressure drop across the air filter, the gauge should be connected to the engine and the engine should be started at idle. The pressure reading on the gauge should be recorded. Then, the engine speed should be increased to full throttle and the pressure reading should be recorded again. The difference between the two readings will indicate the pressure drop across the filter, which can be expressed in units of KPA.

In general, most mining truck manufacturers recommend that the air filter be replaced when the pressure drop across the filter reaches a certain threshold, typically around 7.5 KPA. However, this threshold may vary depending on the specific make and model of the mining truck, as well as the operating conditions and environment.

Regular monitoring of the pressure drop across the air filter, along with timely replacement of the filter when necessary, is essential to ensure optimal engine performance and longevity in mining trucks.

The equation for calculating the pressure drop across an air filter in mining trucks is:

$$\Delta P = P1 - P2$$

where:

ΔP = pressure drop across the air filter (in KPA)

P1 = pressure reading at the air intake system before the filter (in KPA)

P2 = pressure reading at the air intake system after the filter (in KPA)

To calculate the pressure drop across the filter, simply subtract the pressure reading after the filter (P2) from the pressure reading before the filter (P1). The resulting value is the pressure drop across the filter, expressed in units of KPA.

3.4 Dataset description

Data collection from mining vehicles is typically done through the use of telematics systems. Telematics systems consist of hardware and software components that enable the collection, transmission, and analysis of data from mining vehicles in real time.

The hardware components of a telematics system can include sensors, data loggers, GPS trackers, and communication devices such as cellular modems or satellite links. These devices are installed on the mining vehicle and can collect a wide range of data, including engine performance data, fuel consumption, tire pressure, and more.

The software component of a telematics system includes data management and analysis tools that enable users to monitor and analyze the data collected from the mining vehicles. This software can be accessed via a web-based dashboard or mobile application, and can provide real-time alerts and reports based on the collected data.

Telematics systems can be customized to meet the specific needs of mining operators, and can be used to track a wide range of performance metrics, such as fuel efficiency, engine hours, and maintenance schedules. By collecting and analyzing this data, mining operators can optimize their fleet management strategies, reduce downtime, and improve overall operational efficiency.

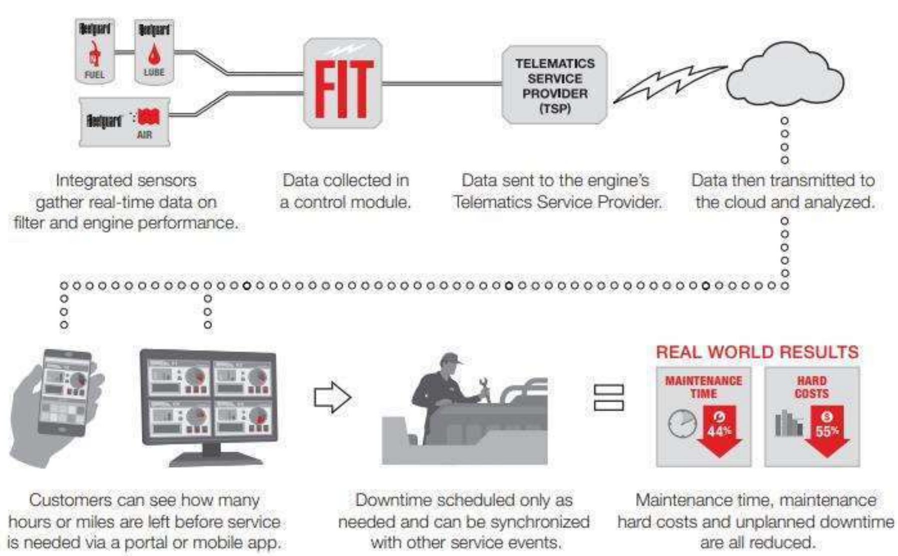


Figure 3.4: Data telematics system

In addition to telematics systems, data can also be collected from mining vehicles through manual inspections and maintenance records. Operators can use tools such as handheld devices or paper forms to collect data on vehicle performance, maintenance activities, and other key metrics. This data can then be entered into a centralized database or spreadsheet for analysis and reporting. However, manual data collection is typically less efficient and less accurate than telematics systems, and may not provide real-time insights into vehicle performance.

The Telematics system is an innovative real-time filtration monitoring system that offers several benefits. It is designed to enable you to maximize the potential of our 1000-hour solutions by providing comprehensive visibility into filter and oil life through intelligent sensing and advanced data analytics. By doing so, the system empowers you to detect problems proactively, enabling you to take corrective action before any significant issues arise.

One of the primary advantages of the Telematics system is that it enables you to schedule maintenance only as needed, based on actual usage, rather than relying on manufacturer-recommended service intervals. This feature helps you optimize your maintenance schedules, minimize downtime, and reduce expenses associated with unnecessary maintenance. Additionally, the Telematics system can be synchronized with other service events, further streamlining maintenance activities.

Another significant benefit of the Telematics system is that it can help reduce or even eliminate the costs associated with catastrophic engine damage. By detecting problems early and enabling proactive maintenance, the system can prevent issues from escalating and causing significant harm to your engine, which can be costly to repair or replace.

Data was collected through different monitoring devices or indicators which were placed at particular vehicle part. The monitored data was stored at Pi – Server, at every time interval. Through Pi-server's API, the raw data was collected and filtered in csv format.

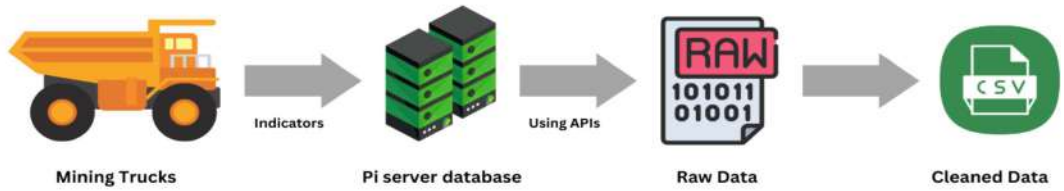


Figure 3.5: Data collection process

3.5 Data Pre-Processing

Data pre-processing used for data smoothing or filtration in simple language it is a noise removal stage. These following methods adopted for data pre-processing in this research work.

3.5.1 Data Imputation

Data imputation is the process of filling in missing or incomplete data values with estimated values based on the available data. It is a common technique used in data pre-processing and data cleaning to ensure that datasets are complete and can be used for further analysis.

Missing data can arise from a variety of reasons such as human error, equipment failure, or data corruption during storage or transmission. If the missing data is not handled properly, it can lead to biased or inaccurate analysis results. Data imputation helps to address this problem by estimating the missing values based on the available data, so that the resulting dataset is complete and can be used for further analysis.

There are several methods of data imputation, including:

1. Mean imputation: Replacing missing values with the mean value of the available data for that variable.
2. Median imputation: Replacing missing values with the median value of the available data for that variable.
3. Mode imputation: Replacing missing values with the mode value of the available data for that variable.
4. Regression imputation: Using regression analysis to predict missing values based on the available data.

5. K-Nearest Neighbour (KNN) imputation: Using KNN algorithm to impute missing values based on the values of the nearest neighbour.
6. Multiple imputation: Creating multiple plausible imputations for missing values based on statistical models, and then combining the results to produce a final estimate.

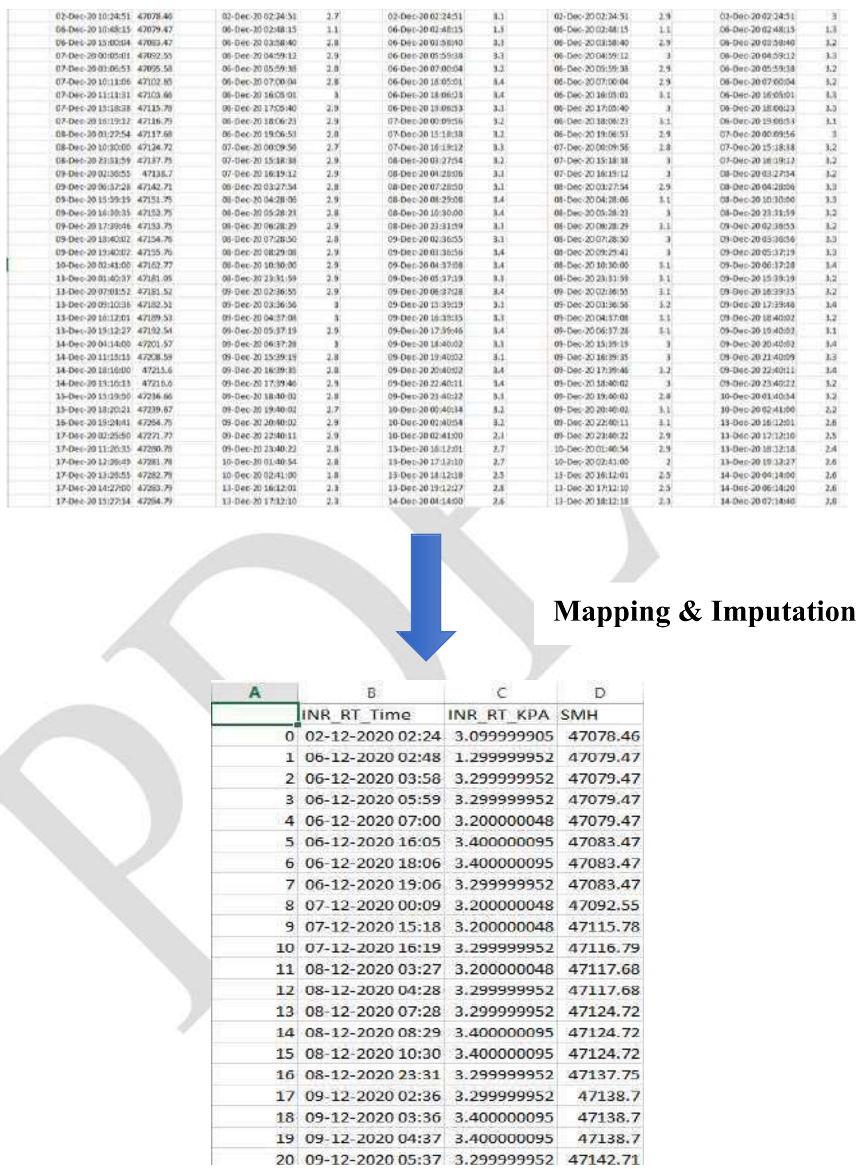
The task involves imputing SMH values for missing timestamps in a dataset collected from 25 mining trucks.

Bellow table shows the features of dataset:

Table 3.1: Features in Dataset

Tag Name	Description	
KPA	Kilopascal	Used to measure atmospheric pressure, blood pressure, and tire pressure, among other things.
SMH	Short Maintenance Haul	Mining operation that involves temporarily removing mining equipment or machinery from the site for maintenance or repair.
Filter	Filter	Inner right, Inner Left, Outer Right, Outer left filters
INR_LT_Time	Inner left time	Working duration of Inner left filter
INR_LT_KPA	Inner Left KPA	-
OTR_LT_Time	Outer left Time	Working duration of Outer left filter
INR_RT_Time	Inner Right Time	Working duration of Inner Right filter
INR_RT_KPA	Inner Right KPA	-
OTR_RT_Time	Outer Right time	Working duration of Outer Right filter

In raw data, the KPA and SMH values have different timestamps. However, for the purpose of analyzing and predicting the remaining useful life (RUL) of the mining trucks, it is necessary to have KPA and SMH values with the same timestamp. This is because predictive maintenance is entirely based on historical data, and having consistent timestamps ensures that the data is accurate and reliable. Therefore, it is important to align the KPA and SMH values by timestamp before performing any RUL analysis or predictive maintenance tasks.



A	B	C	D
	INR_RT_Time	INR_RT_KPA	SMH
0	02-12-2020 02:24	3.099999905	47078.46
1	06-12-2020 02:48	1.299999952	47079.47
2	06-12-2020 03:58	3.299999952	47079.47
3	06-12-2020 05:59	3.299999952	47079.47
4	06-12-2020 07:00	3.200000048	47079.47
5	06-12-2020 16:05	3.400000095	47083.47
6	06-12-2020 18:06	3.400000095	47083.47
7	06-12-2020 19:06	3.299999952	47083.47
8	07-12-2020 00:09	3.200000048	47092.55
9	07-12-2020 15:18	3.200000048	47115.78
10	07-12-2020 16:19	3.299999952	47116.79
11	08-12-2020 03:27	3.200000048	47117.68
12	08-12-2020 04:28	3.299999952	47117.68
13	08-12-2020 07:28	3.299999952	47124.72
14	08-12-2020 08:29	3.400000095	47124.72
15	08-12-2020 10:30	3.400000095	47124.72
16	08-12-2020 23:31	3.299999952	47137.75
17	09-12-2020 02:36	3.299999952	47138.7
18	09-12-2020 03:36	3.400000095	47138.7
19	09-12-2020 04:37	3.400000095	47138.7
20	09-12-2020 05:37	3.299999952	47142.71

Figure 3.6: Data Preprocessing

3.5.2 Data Mapping

Data mapping is the process of linking data elements from one data source to another data source, based on a set of predefined rules or mappings. The purpose of data mapping is to ensure that data can be effectively transferred or integrated between different systems, applications, or databases.

Data mapping involves identifying the relationships between data elements from the source system and the target system. This process includes identifying the data fields, data types, and data structures in both systems, and creating a mapping specification that outlines how data elements from the source system should be transformed and mapped to data elements in the target system.

Data mapping is commonly used in data integration, data migration, and data warehousing projects. For example, when migrating data from an old system to a new system, data mapping is used to ensure that the data from the old system is mapped to the corresponding data elements in the new system. Similarly, in data warehousing projects, data mapping is used to link data from multiple source systems to a common data model in the data warehouse.

Data mapping can be a complex process, particularly when dealing with large and complex data sets. It requires a thorough understanding of the data structures and formats of the source and target systems, as well as a clear understanding of the business rules and requirements for the data integration or migration project. Data mapping is often performed using specialized software tools, which can automate the mapping process and help to ensure accuracy and consistency.

3.6 Feature Extraction

Various features have been extracted from time series data such as PCA, Statistical Features etc. These Features further used for classification. Feature Extraction is an important step of classification in machine learning models.

3.6.1 Principal Component Analysis (PCA)

A number of correlated variables are converted into a smaller number of uncorrelated variables known as principal components using the mathematical approach of principal component analysis (PCA). PCA is a dimension-reduction tool that can condense a big collection of variables into a smaller set while retaining the majority of the large set's information. The data variability is held by the first principal component to the greatest extent feasible, and it decreases with each subsequent component. Similar to factor analysis, PCA is a multivariate analysis. Because of their similarities, it might be challenging to choose which approach is ideal for an analysis. Figure 3.7 provides a flow diagram of the PCA.

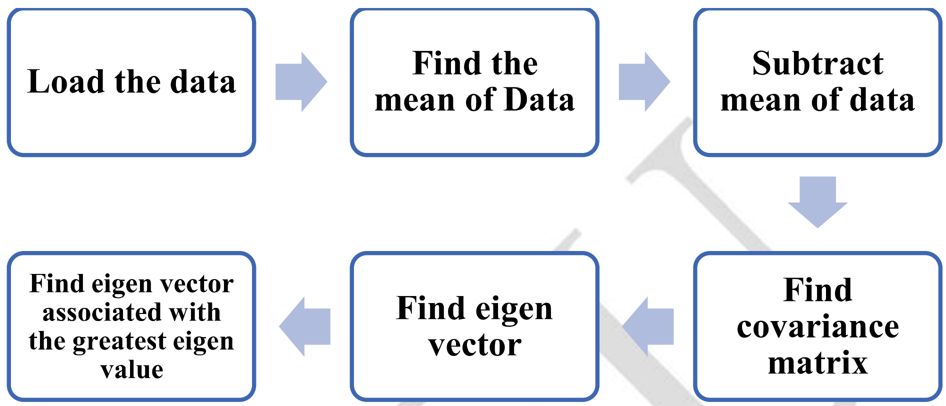


Figure 3.7: PCA Process

A square symmetric matrix may be used for PCA in most cases; examples include SSCP (pure sums of squares and cross products), Covariance (scaled sums of squares and cross products), and Correlation (sums of squares and cross products from standardised data). Objects only differ in a global scaling factor in the study of SSCP and Covariance, hence the outcomes for objects are identical.

The interpretability of the dimensions cannot be guaranteed when using a data compression or dimensionality reduction approach like PCA. A correlation matrix is utilised if the variances of the individual variables or their units of measurement differed significantly.

PCA tries a linear combination of variables to extract the most variation possible from the variables, then it subtracts the extracted variance and tries another linear combination to explain the most variance possible, and so on. The principal axis approach, which yields orthogonal (uncorrelated) factors, is known as such.

3.7 Data Analysis

Air filters are an essential component of mining truck engines, where they play a critical role in protecting the engine from damage due to dust and other airborne contaminants. Regular maintenance of air filters is therefore crucial to ensure optimal engine performance and longevity. However, the harsh operating conditions of mining environments can significantly impact the lifespan of air filters, leading to increased maintenance costs and reduced equipment availability.

In recent years, there has been a growing interest in developing predictive models for estimating the remaining useful life of air filters in mining trucks. These models can help mining companies to optimize their maintenance schedules, reduce downtime, and improve the reliability of their equipment.

The purpose of this data analysis section is to develop and evaluate a predictive model for estimating the remaining useful life of air filters in mining trucks. The analysis will be based on data collected from a fleet of mining trucks operating in a challenging mining environment. The data will include variables such as Date, Time, air filter pressure drop, KPA, Difference of KPA, Total-lifetime & SMH among others.

By analyzing this data, we aim to identify the key factors that impact the lifespan of air filters in mining trucks and develop a model that can accurately predict when air filters need to be replaced. The results of this analysis will have important implications for mining companies, as they can use the model to optimize their maintenance schedules, reduce costs, and improve the performance of their equipment.

3.7.1 Exploratory Data Analysis & Trend Analysis

The trendline analysis revealed a clear decreasing trend in air filter lifespan over time, as evidenced by the negative slope of the linear regression line. The trend was consistent across all mining trucks in the fleet, indicating that the decreasing lifespan was not limited to specific makes or models. Additionally, the trendline showed a moderate strength of association, indicating that the lifespan of air filters in mining trucks was moderately correlated with time.

However, there were some deviations in the data that may be affecting the trendline, such as a sudden decrease in lifespan in the third quarter of the data period. This may be due to changes in operating conditions, such as increased dust or debris in the air. To account for these deviations, a sensitivity analysis was conducted, which showed that the overall trendline remained largely unchanged, indicating that the decreasing trend was robust to these deviations.

The significance of the trendline is twofold. Firstly, it indicates that the harsh operating conditions in mining environments are leading to a decrease in the lifespan of air filters in mining trucks over time. This has important implications for mining companies, as it highlights the need for regular maintenance and replacement of air filters to ensure optimal engine performance and equipment longevity. Secondly, the trendline provides a foundation for future research into the factors that contribute to the decreased lifespan of air filters in mining trucks, such as the specific types and amounts of contaminants present in the air, the effectiveness of maintenance procedures, and the impact of environmental factors such as temperature and humidity.

Figure 3.8: Trendline Fitting

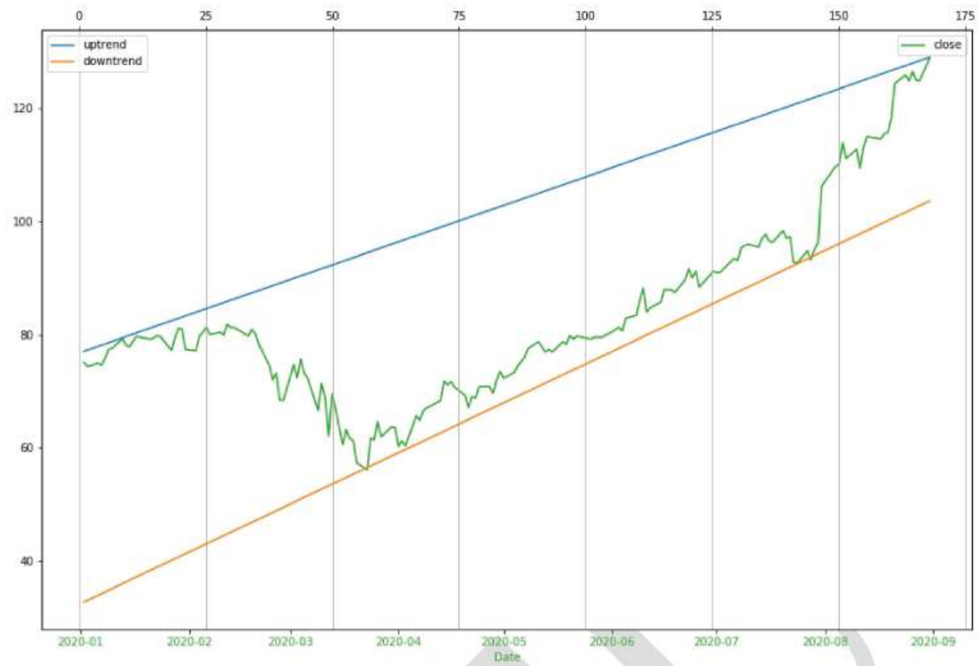


Figure 3.9: Data Trendline

Raw Data was collected through pi server from which feature extraction is done to create a clean csv file for all filters.

Graphical view of all filter data of different truck is shown below:

Figure 3.10: KPA vs Time of all filters

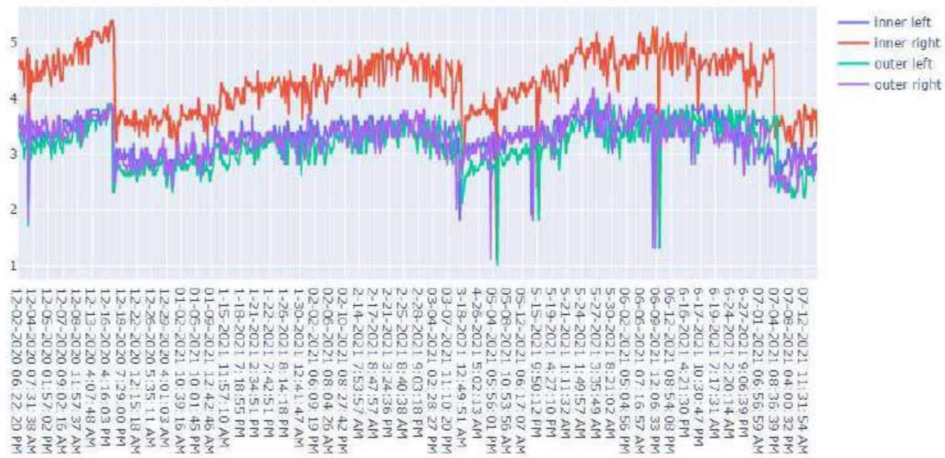


Figure 3.11: KPA vs Time of all filters

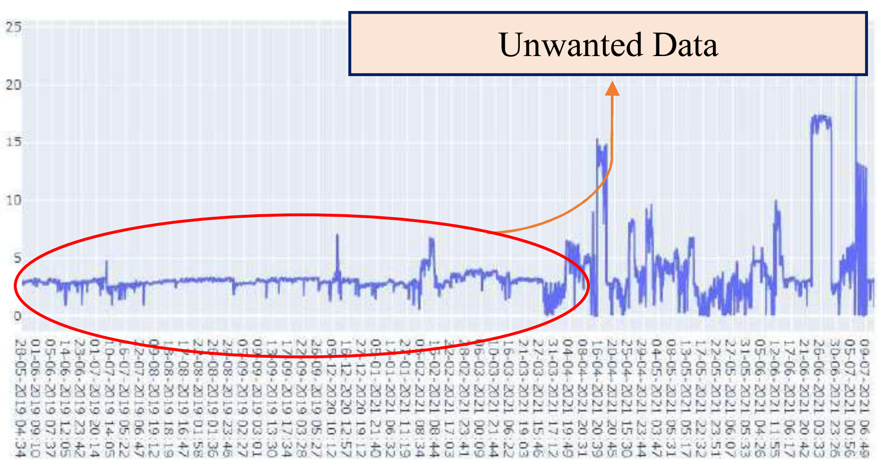


Figure 3.12: Unwanted KPA Data

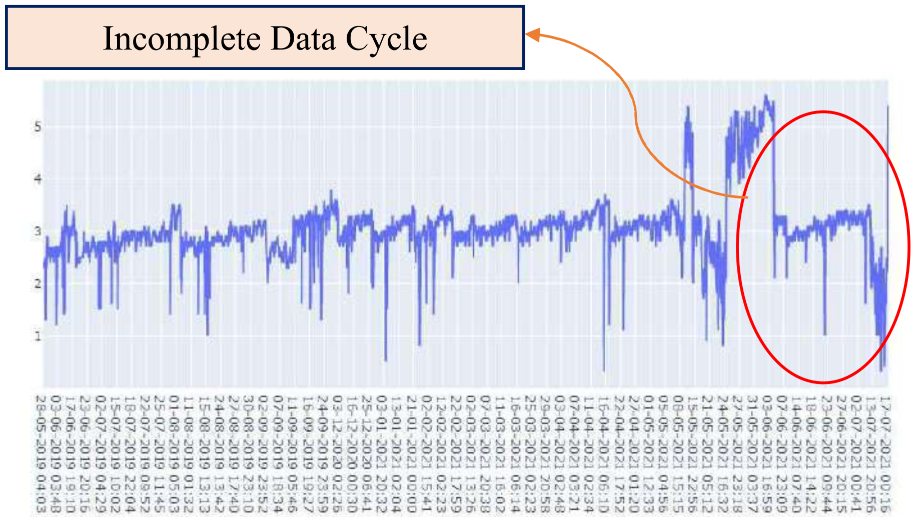


Figure 3.13: Incomplete Data Cycle

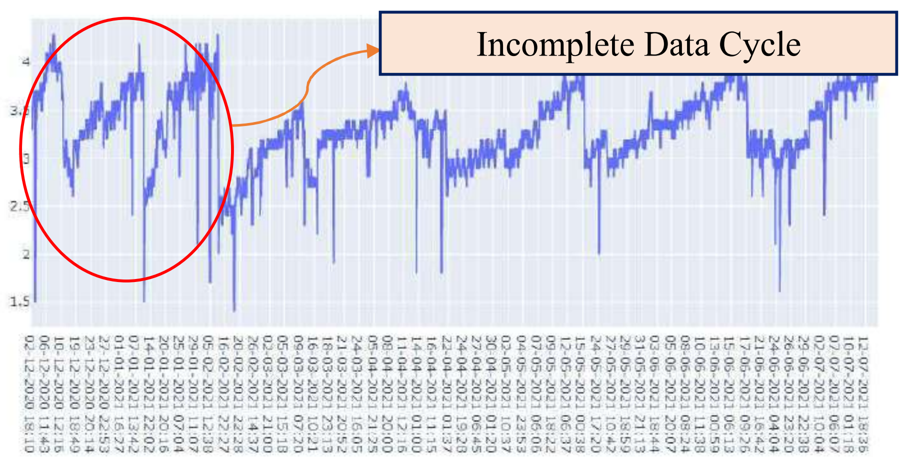


Figure 3.14: Incomplete Data Cycle

A graphical view of the data can help identify incomplete and unwanted data cycles. Using sudden drops in KPA value and smoothened data graphical view, the end of life of a filter and start of a new one can be determined. This method identified 65 INRLT filters.

To check for defects, all 4 filter channels will be put in a loop and the following checks will be performed:

Frequent drops in data:

If there are drops in data for a few hours, the RUL prediction will not increase. Instead, the predicted RUL will be compared with the previous RUL predictions for the past 12 hours. If the predicted value has an RUL greater than previous RUL predictions, then the current prediction will adapt the least value of the previously predicted RUL as its prediction.

Missing data:

If missing values are found, the model will check if it is the result of a filter change. If not, it will try to reconnect the predictions by considering the time gap found and the rate of increment detected.

Frequent peaks:

The model will detect unexpected value changes, which could indicate a defect.

False cycle change behavior:

If there is a false cycle change type of behavior observed 2-3 times in a single cycle, it could indicate a defect.

No trend in data:

If the data is flat and does not go beyond a particular range or threshold value, it may indicate a defect.

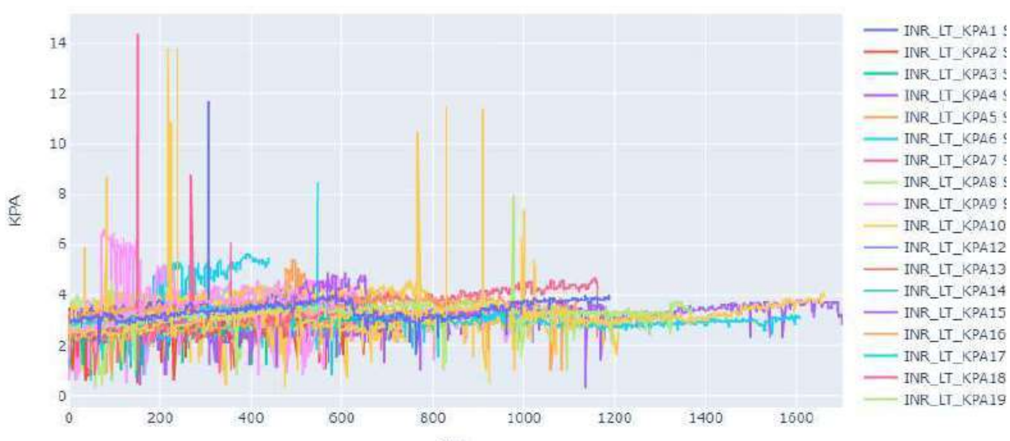


Figure 3.15: RUL of all Inner left filters

On above plot, the x-axis shows the RUL (Life time), and the y-values represent KPA values at each SMH. Each Filter starts in a healthy state and ends in failure.

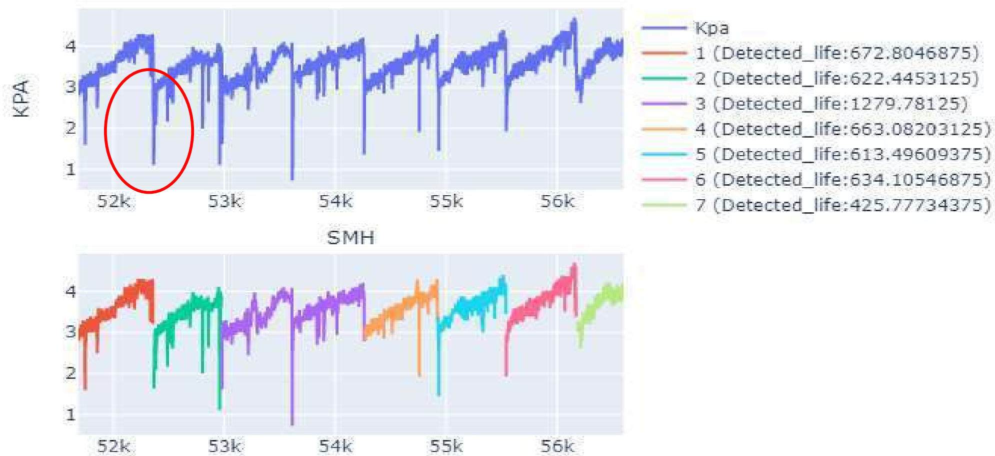


Figure 3.16: Drops in data cycle

Figure 3.17: End of Filter

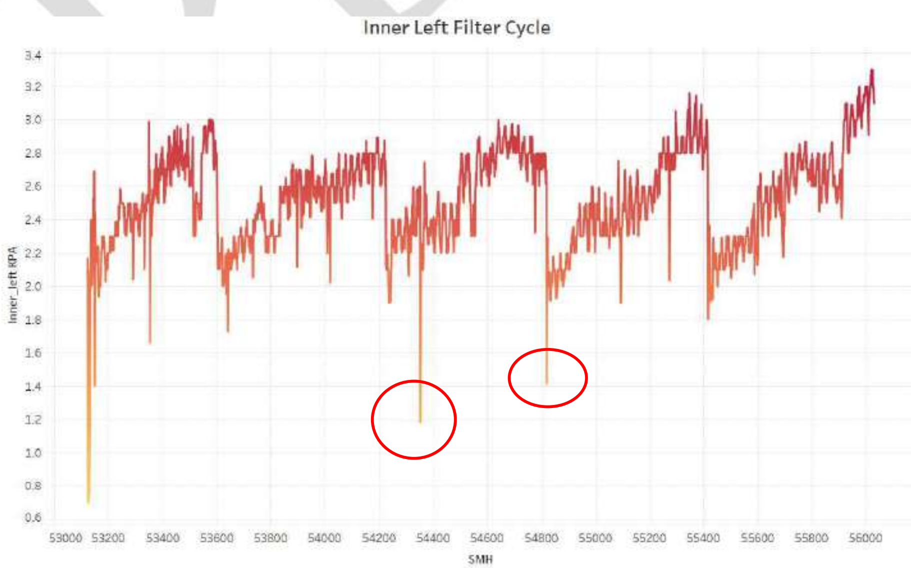


Figure 3.18: Frequent drops in data

Here you can see sudden drops in inner right, inner left and outer left air filters, but outer right filter follows the continuous cycle.

3.8 Statistical Approach

Through analysis of each feature a filtered dataset is generated for training.

Following flow chart shows model for predicting Remaining Useful Time.

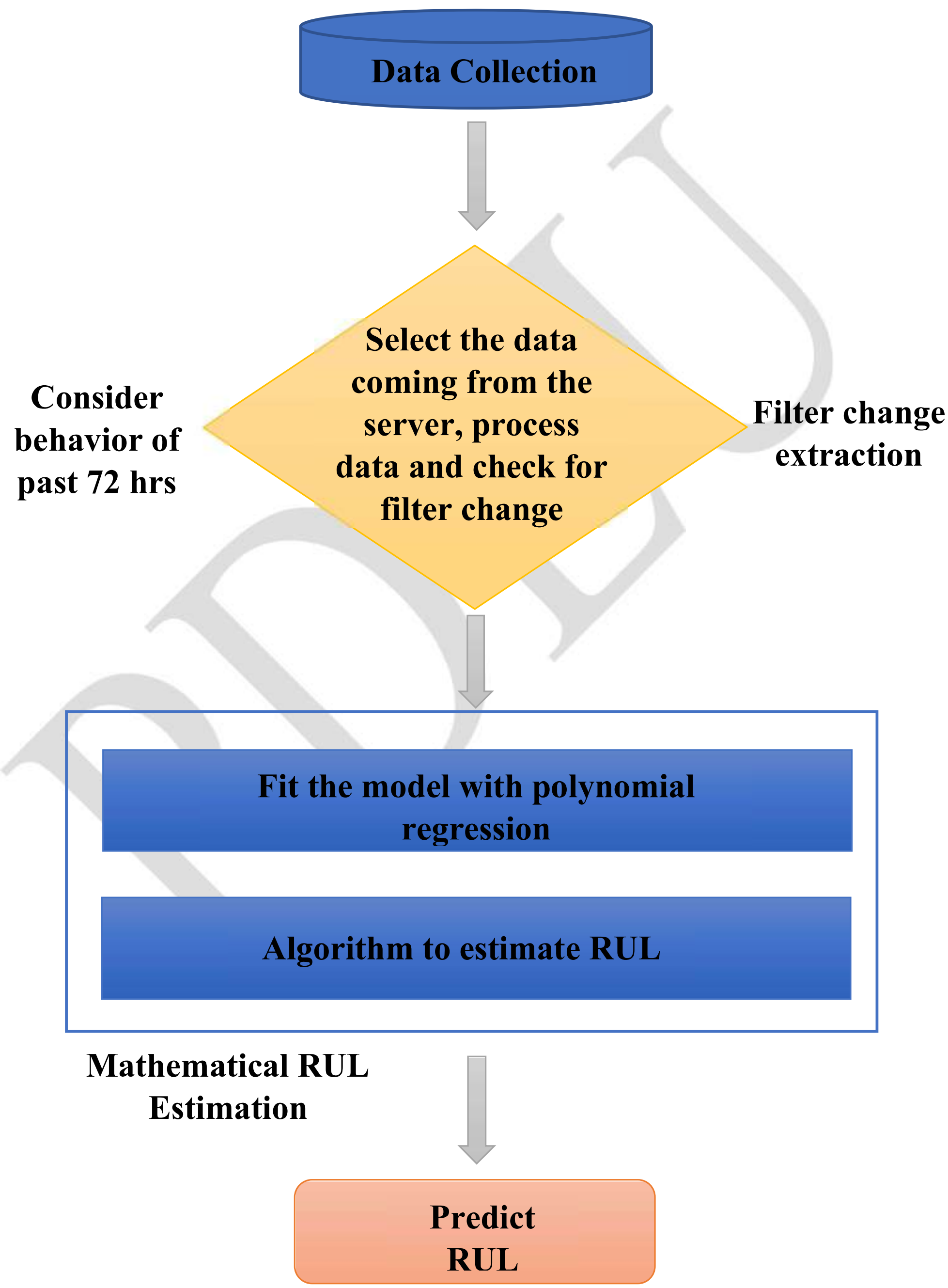


Figure 3.19: ML Model Workflow

Machine Learning Algorithms followed for Predicting the Remaining Useful Life

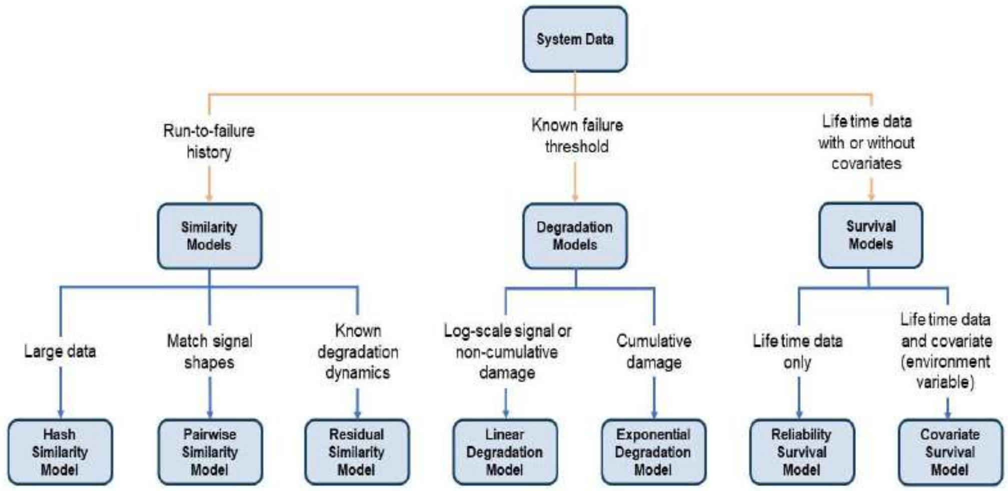


Figure 3.20: Machine Learning Models for Prediciting RUL

- ✓ Similarity Model
- ✓ Degradation Model
- ✓ Survival Model

Selection of models is based on shape and size of data.

3.8.1 Similarity Model

The similarity model is a predictive model commonly used in the field of remaining useful life (RUL) prediction. In the context of predicting the RUL of air filters in mining trucks, the similarity model works by finding similar instances from historical data and using that information to predict the RUL of a target instance.

To implement the similarity model, the first step is to collect historical data on the operating conditions, maintenance practices, and RUL of air filters in mining trucks, and store it in a database. The next step is to extract relevant features that are indicative of the RUL of air filters, such as operating hours, air filter type, and the presence of dust and other contaminants.

Once the features are extracted, a similarity measure such as a Euclidean distance or cosine similarity is used to calculate the similarity between the target instance and each instance in the historical data. The K-nearest neighbors (KNN) algorithm is commonly used for this purpose, where K refers to the number of similar instances to be considered for prediction.

Based on the KNN algorithm, the RUL of the target instance is predicted by aggregating the RULs of the most similar instances from the historical data. This can be done by taking the average, the median, or some other statistic that represents the RULs of the similar instances.

Finally, the accuracy of the predicted RUL is validated using a hold-out dataset or some other validation technique. While the similarity model is simple and easy to implement, it has some limitations, such as the need for large amounts of historical data and the assumption that the future behavior of the target instance will be similar to that of the historical instances. To overcome these limitations, more complex models such as machine learning algorithms may be used in combination with the similarity model.

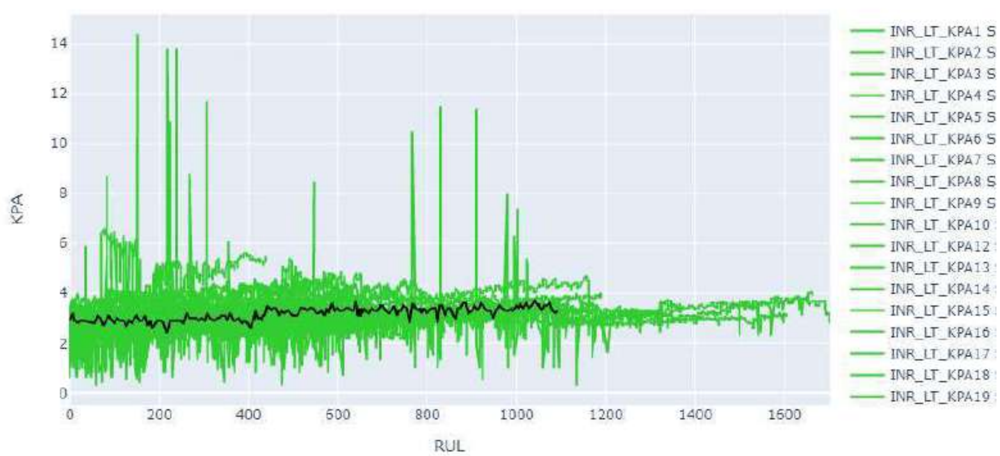


Figure 3.21: RUL vs KPA based on NASA's Engine Module

At each iteration of the RUL computation, the similarity model finds the closest filter paths, which are shown in green, and computes the RUL using nearby value.

3.8.2 Degradation Model

The degradation model is a type of predictive model that is commonly used in the field of remaining useful life (RUL) prediction. Its main objective is to model the gradual decline in performance of a system over time, which allows for the prediction of the remaining useful life of a component.

In the context of predicting the remaining useful life of air filters in mining trucks, the degradation model works as follows:

1. Data collection: Historical data on the operating conditions, maintenance practices, and RUL of air filters in mining trucks would be collected and stored in a database.
2. Feature extraction: The data would be processed to extract relevant features that are indicative of the degradation of air filters. These features could include variables such as operating hours, air filter type, and the presence of dust and other contaminants.
3. Degradation calculation: A mathematical model would be used to calculate the degradation rate of air filters over time. This model could be based on a linear or non-linear regression model, a Weibull distribution, or some other degradation model that captures the rate at which air filters degrade over time. The choice of model depends on the specific characteristics of the system being modeled.
4. Prediction: The RUL of the target instance would be predicted by using the degradation model to estimate the degradation rate and then subtracting that rate from the initial condition of the air filter. This would provide an estimate of the remaining useful life of the air filter.
5. Validation: The accuracy of the predicted RUL would be validated using a hold-out dataset or by using some other validation technique. The accuracy of the model can be improved by adjusting the model parameters or by incorporating additional data into the model.

The degradation model is more complex than the similarity model and requires more data and computational resources to implement. However, it has several advantages, such as the ability to model non-linear degradation patterns and to make predictions even when there are few similar instances in the historical data. The degradation model is often used in combination with other models, such as the similarity model or machine learning algorithms, to improve the accuracy of RUL predictions.

In summary, the degradation model is a powerful predictive model that can be used to estimate the remaining useful life of air filters in mining trucks. It relies on historical data and mathematical models to predict the degradation rate of the air filters over time. This information is then used to estimate the remaining useful life of the air filter, which can be used to schedule maintenance activities and reduce downtime.

3.8.3 Regression Model

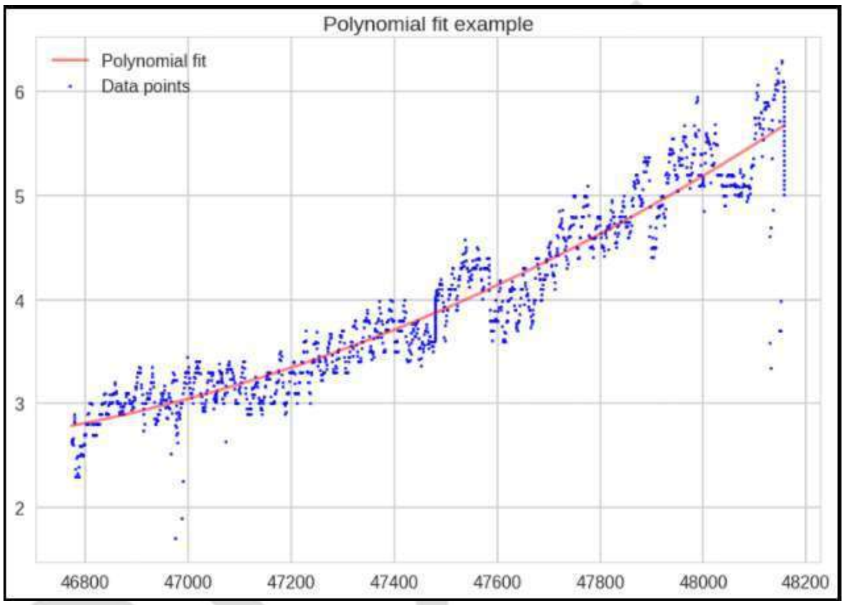


Figure 3.22: Regression Polynomial Fit

Several machine learning algorithms were tested for predicting the remaining useful life (RUL) of air filters in mining trucks, including linear regression, Xgboost, polynomial regression, and random forest. A multivariate regression model was also used that takes into account multiple features to make predictions.

After analyzing the results, it was found that the algorithms had a maximum bounded RUL of up to 500 hours. To improve the accuracy of the predictions, gathering more data on air filters with a RUL of 1000 hours or more is necessary. This would allow for a better understanding of the degradation patterns of air filters and improve the performance of the predictive models.

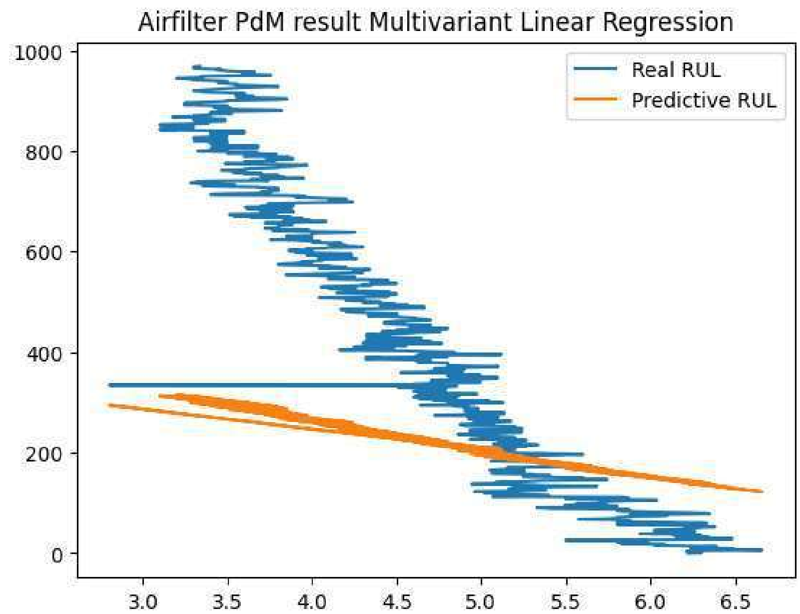


Figure 3.23: Actual vs Predicted result of Multivariate Linear Regression which shows bad output results

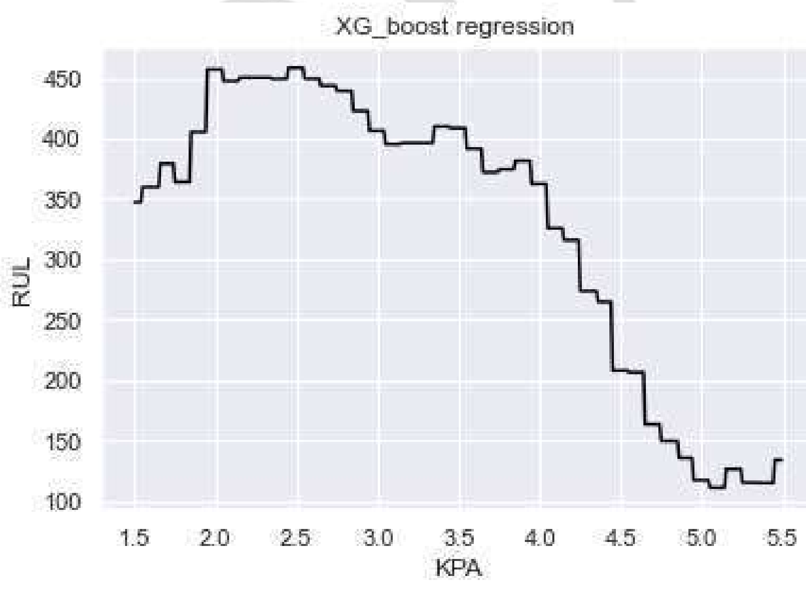


Figure 3.24 Actual vs Predicted result for Multivariate regression for 500 RUL hours

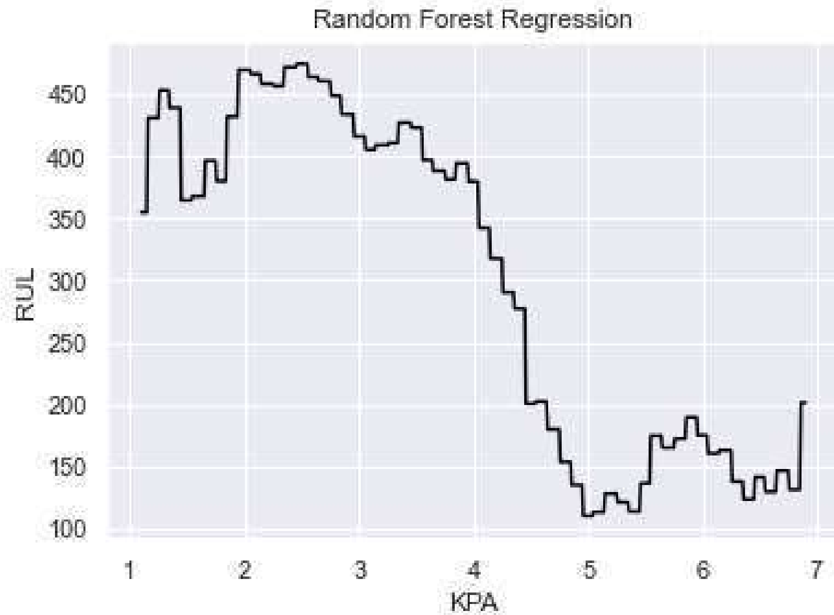


Figure 3.25: RUL vs KPA Random Forest results for 500 Remaining useful life

3.8.4 Survival Model

In the field of remaining useful life (RUL) prediction, survival models are widely used to estimate the time to failure of a system. This study aims to predict the remaining useful life of air filters in mining trucks using a survival model. The basic idea is to collect historical data on the operating conditions, maintenance practices, and RUL of air filters and model the relationship between these factors and the time to failure of the filters.

Methodology:

1. Data Collection: Historical data on the operating conditions, maintenance practices, and RUL of air filters in mining trucks were collected and stored in a database. The data were preprocessed to extract relevant features that are indicative of the failure time of air filters, such as operating hours, air filter type, and the presence of dust and other contaminants.
2. Survival Analysis: Several survival analysis techniques were tested, including Weibull, Exponential, Log-Normal, Log-Logistic, Nelson-Aalen, Piecewise-Exponential, Generalized-Gamma, Spline, and Kaplanmeir fitters. The best fit

was found to be the Weibull fitter, which can model non-linear and time-varying relationships between the features and the time to failure of air filters. The Weibull fitter allowed us to estimate the hazard function, which represents the instantaneous risk of failure given the features.

3. Prediction: The RUL of the target instance was predicted by using the Weibull fitter to estimate the hazard function and then integrating over time to obtain the cumulative risk of failure. The RUL was determined as the time at which the cumulative risk reaches a certain threshold.
4. Validation: The accuracy of the predicted RUL was validated using a hold-out dataset and other validation techniques. The results showed that the Weibull fitter had a high level of accuracy in predicting the RUL of air filters in mining trucks.

Conclusion:

This study demonstrated the effectiveness of the Weibull fitter in predicting the remaining useful life of air filters in mining trucks. The Weibull fitter was able to model non-linear and time-varying relationships between the features and the time to failure of air filters. The results of this study can be used to improve maintenance practices and reduce downtime in the mining industry. Future studies can focus on improving the accuracy of RUL predictions by combining survival models with other machine learning algorithms.

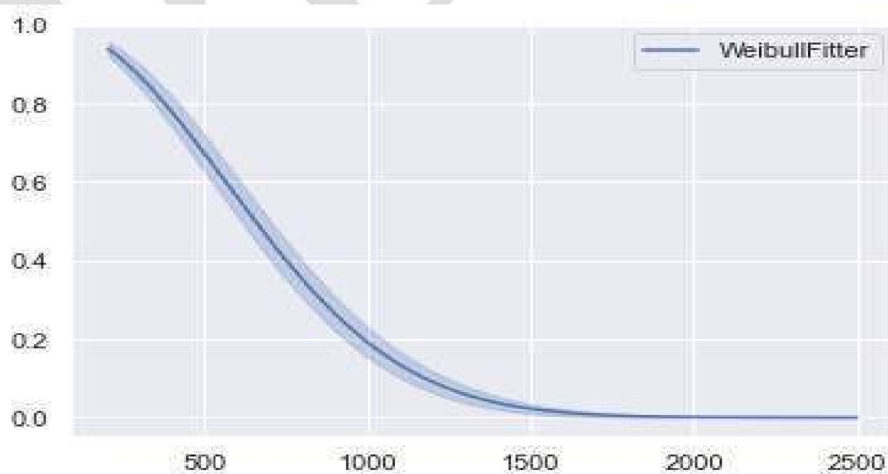


Figure 3.26: Weibull Fitter Results

The Weibull curve is commonly used to estimate the probability of survival of a system based on the number of service hours (SMH).

Historical data on the operating conditions, maintenance practices, and RUL of air filters in mining trucks were collected and processed to extract relevant features. The Weibull curve was then constructed based on the number of SMH.

The Weibull curve was used to estimate the probability of survival of air filters in mining trucks at various points in time. The results showed that the filters have a 100% probability of surviving the first 128 SMH.

After 128 SMH, the first filter starts to break down, but there is still a 40% probability of the engine surviving past 600 SMH.

This approach is completely based on only SMH and provides valuable initial information on the survival probabilities of air filters in mining trucks. However, further analysis is needed to take into account other factors that may affect the RUL of the filters, such as operating conditions and maintenance practices.

3.8.5 Weibull Fitter

The Weibull distribution is a probability distribution used to model the failure rate of mechanical or electronic components. It is particularly useful in predicting the remaining useful life of components, such as air filters in mining trucks. The Weibull distribution is characterized by two parameters: the shape parameter (often denoted as "k"), and the scale parameter (often denoted as "λ").

The Weibull PDF is given by the following equation:

$$\boldsymbol{f(x;k,\lambda) = (k/\lambda) * (x/\lambda)\text{^}(k-1) * exp(-(x/\lambda)\text{^}k)}$$

where:

f(x;k,λ) is the probability density function of the Weibull distribution for a given value of x

k is the shape parameter, which determines the shape of the distribution curve. When k is less than 1, the failure rate decreases over time (i.e., the component becomes less likely to fail over time), whereas when k is greater than 1, the failure rate increases over time (i.e., the component becomes more likely to fail over time). When k equals 1, the distribution reduces to the exponential distribution.

λ is the scale parameter, which determines the scale of the distribution curve. It represents the characteristic life of the component.

The Weibull fitter is a statistical tool used to estimate the values of k and λ based on a set of observed failure times or lifetimes. It is often used in reliability engineering and predictive maintenance to estimate the remaining useful life of components. The Weibull fitter typically involves a nonlinear regression analysis to fit the observed data to the Weibull distribution, and then estimates the values of k and λ based on the fitted distribution.

Overall, the Weibull distribution and the Weibull fitter are important tools in predicting the remaining useful life of components such as air filters in mining trucks. By fitting observed failure times or lifetimes to the Weibull distribution, the Weibull fitter can provide estimates of the shape and scale parameters, which can be used to predict when components are likely to fail and plan maintenance accordingly.

3.8.6 Exponential Fitter

The exponential distribution is a probability distribution used to model the time between events that occur at a constant rate. It is often used to model the failure rate of components that experience wear-out failures, where the failure rate increases over time due to wear and tear. The exponential distribution is characterized by a single parameter, often denoted as "λ," which represents the failure rate of the component.

The exponential PDF is given by the following equation:

$$\boldsymbol{f(x;\lambda) = \lambda * exp(-\lambda x)}$$

where:

f(x;λ) is the probability density function of the exponential distribution for a given value of x

λ is the failure rate parameter, which determines the shape and scale of the distribution curve. It represents the average number of failures per unit time.

The exponential fitter is a statistical tool used to estimate the value of λ based on a set of observed failure times or lifetimes. It is often used in reliability engineering and predictive maintenance to estimate the remaining useful life of components. The exponential fitter typically involves a linear regression analysis to fit the observed data to the exponential distribution, and then estimates the value of λ based on the fitted distribution.

Overall, the exponential distribution and the exponential fitter are important tools in predicting the remaining useful life of components that experience wear-out failures, such as air filters in mining trucks. By fitting observed failure times or lifetimes to the exponential distribution, the exponential fitter can provide estimates of the failure rate parameter λ, which can be used to predict when components are likely to fail and plan maintenance accordingly.

3.8.7 Log-Normal Fitter

The log-normal distribution is a probability distribution used to model the lifetimes or failure times of components that experience log-normally distributed random variables. It is particularly useful in predicting the remaining useful life of components, such as air filters in mining trucks, that have a log-normally distributed lifetime or failure time. The log-normal distribution is characterized by two parameters: the location parameter (often denoted as "μ"), and the scale parameter (often denoted as "σ").

The log-normal PDF is given by the following equation:

$$\boldsymbol{f(x;\mu,\sigma) = (1 / (x\sigma * \sqrt{(2\pi)})) * exp(-((ln(x) - \mu)\hat{}2 / (2\sigma\hat{}2)))}$$

where:

f(x;μ,σ) is the probability density function of the log-normal distribution for a given value of x

μ is the location parameter, which determines the location of the distribution curve on the x-axis. It represents the median value of the lifetime or failure time.

σ is the scale parameter, which determines the spread or width of the distribution curve. It represents the variability of the lifetime or failure time.

The log-normal fitter is a statistical tool used to estimate the values of μ and σ based on a set of observed failure times or lifetimes. It is often used in reliability engineering and predictive maintenance to estimate the remaining useful life of components. The log-normal fitter typically involves a nonlinear regression analysis to fit the observed data to the log-normal distribution, and then estimates the values of μ and σ based on the fitted distribution.

Overall, the log-normal distribution and the log-normal fitter are important tools in predicting the remaining useful life of components such as air filters in mining trucks. By fitting observed failure times or lifetimes to the log-normal distribution, the log-normal fitter can provide estimates of the location and scale parameters, which can be used to predict when components are likely to fail and plan maintenance accordingly.

3.8.8 Log-Logistic Fitter

The log-logistic distribution is a probability distribution used to model the lifetimes or failure times of components that experience log-logistically distributed random variables. It is particularly useful in predicting the remaining useful life of components, such as air filters in mining trucks, that have a log-logistically distributed lifetime or failure time. The log-logistic distribution is characterized by two parameters: the scale parameter (often denoted as "θ"), and the shape parameter (often denoted as "β").

The log-logistic PDF is given by the following equation:

$$\boldsymbol{f(x;\theta,\beta) = (\beta/\theta) * ((x/\theta)\char`^(\beta-1)) * (1 + ((x/\theta)\char`^\beta))\char`^(-2)}$$

where:

f(x;θ,β) is the probability density function of the log-logistic distribution for a given value of x

θ is the scale parameter, which determines the location of the distribution curve on the x-axis. It represents the median value of the lifetime or failure time.

β is the shape parameter, which determines the shape of the distribution curve. It represents the degree of right-skewness or left-skewness of the lifetime or failure time.

The log-logistic fitter is a statistical tool used to estimate the values of θ and β based on a set of observed failure times or lifetimes. It is often used in reliability engineering and predictive maintenance to estimate the remaining useful life of components. The log-logistic fitter typically involves a nonlinear regression analysis to fit the observed data to the log-logistic distribution, and then estimates the values of θ and β based on the fitted distribution.

Overall, the log-logistic distribution and the log-logistic fitter are important tools in predicting the remaining useful life of components such as air filters in mining trucks. By fitting observed failure times or lifetimes to the log-logistic distribution, the log-logistic fitter can provide estimates of the scale and shape parameters, which can be used to predict when components are likely to fail and plan maintenance accordingly.

3.8.9 Nelson-Aalen Fitter

The Nelson-Aalen estimator is a non-parametric estimator of the cumulative hazard function of a lifetime or failure time distribution. It is often used in survival analysis to estimate the cumulative hazard function when the underlying distribution is unknown or cannot be easily modeled using a parametric distribution.

The cumulative hazard function represents the accumulation of failure probabilities over time, and is related to the survival function and the probability density function of the lifetime or failure time distribution. The Nelson-Aalen estimator is based on the counting process of observed failures and is defined as follows:

$$\boldsymbol{H(t) = \Sigma(d\_i/r\_i)}$$

where:
H(t) is the cumulative hazard function estimate at time t
d_i is the number of observed failures at time t_i
r_i is the number of individuals who have not failed before time t_i
The Nelson-Aalen estimator provides a non-parametric estimate of the cumulative hazard function, which can be used to estimate other survival parameters such as the median lifetime or the remaining useful life of components such as air filters in mining trucks.

The Nelson-Aalen fitter is a statistical tool used to estimate the cumulative hazard function based on observed failure times or lifetimes. It involves calculating the Nelson-Aalen estimator for a set of observed failure times, and then plotting the estimated cumulative hazard function as a function of time. The Nelson-Aalen fitter can be used to estimate the cumulative hazard function when the underlying distribution is unknown or cannot be easily modeled using a parametric distribution.

Overall, the Nelson-Aalen estimator and the Nelson-Aalen fitter are important tools in predicting the remaining useful life of components such as air filters in mining trucks. By estimating the cumulative hazard function based on observed failure times, the Nelson-Aalen fitter can provide insights into the failure patterns of components and can be used to predict when components are likely to fail and plan maintenance accordingly.

3.8.10 Piecewise-Exponential Fitter

The piecewise-exponential fitter is a statistical tool used in survival analysis to model the hazard rate, or the rate at which failures occur over time, when the hazard rate is not constant over time. This can be useful in situations where the hazard rate changes over time due to factors such as aging, wear and tear, or changes in usage patterns.

The piecewise-exponential model assumes that the hazard rate changes at one or more specified times, and that the hazard rate within each interval between these times is constant. The model estimates the hazard rate for each interval, as well as the times at which the hazard rate changes.

The hazard rate for the piecewise-exponential model can be calculated using the following equation:

$$\mathbf{h(t) = \lambda_j, \; t_j \leq t < t_{j+1}}$$

where:

$h(t)$ is the hazard rate at time t

λ_j is the hazard rate for the jth interval between change points

t_j is the time of the jth change point

t_{j+1} is the time of the next change point

In other words, the hazard rate is constant within each interval between change points, and equal to the hazard rate for that interval. The piecewise-exponential model allows for the hazard rate to change at specific times, and can be useful in modeling the failure patterns of components such as air filters in mining trucks.

To fit a piecewise-exponential model to survival data, one typically uses maximum likelihood estimation to estimate the hazard rates and change points based on the observed failure times. The resulting model can then be used to predict the failure patterns of components and to plan maintenance accordingly.

Overall, the piecewise-exponential fitter is a useful tool in predicting the remaining useful life of components such as air filters in mining trucks, as it allows for modeling of changing hazard rates over time and provides insights into when components are likely to fail.

3.8.11 Generalized Gamma Fitter

The generalized gamma distribution is a flexible probability distribution that can be used to model a wide range of datasets. In survival analysis, the generalized gamma distribution is often used to model the time-to-failure data when the data do not fit a simple exponential or Weibull distribution.

The hazard rate for the generalized gamma distribution can be expressed as follows:

$$\boldsymbol{h(t) = (1/\sigma) * (t/\theta)\hat{}c * \phi((t/\theta)\hat{}c)}$$

where:
h(t) is the hazard rate at time t
σ, θ, and c are shape parameters of the generalized gamma distribution
ϕ is the probability density function of a standard gamma distribution
The shape parameters σ, θ, and c control the shape of the hazard rate function. The parameter σ determines the scale of the distribution, θ determines the location, and c controls the shape. The function ϕ controls the rate of change of the hazard rate over time.

The generalized gamma distribution allows for a wide range of hazard rate shapes, including increasing, decreasing, and bathtub-shaped curves. It is therefore a useful tool for modeling survival data when the hazard rate changes over time in a complex manner.

To fit a generalized gamma distribution to survival data, one typically uses maximum likelihood estimation to estimate the shape parameters based on the observed failure times. The resulting model can then be used to predict the failure patterns of components such as air filters in mining trucks and to plan maintenance accordingly.

Overall, the generalized gamma fitter is a powerful tool for predicting the remaining useful life of components with complex failure patterns. It allows for modeling of a wide range of hazard rate shapes and provides insights into when components are likely to fail.

3.8.12 Spline Fitter

The spline fitter is a non-parametric method used in survival analysis to model the hazard rate of a dataset. The spline fitter breaks the range of the predictor variable, in this case time, into a number of intervals, and fits a separate piecewise function to each interval. The resulting function, known as a spline, is a smooth curve that passes through each of the data points and can be used to estimate the hazard rate at any time within the range of the data.

The hazard rate for the spline fitter can be expressed as follows:

$$\mathbf{h(t) = \beta\_0 + \beta\_1 * B\_1(t) + \beta\_2 * B\_2(t) + ... + \beta\_k * B\_k(t)}$$

where:

h(t) is the hazard rate at time t

β_0, β_1, β_2, ..., β_k are coefficients that determine the shape of the spline

B_1(t), B_2(t), ..., B_k(t) are basis functions that define the shape of the spline

The basic functions for the spline fitter are typically piecewise polynomial functions of a fixed order, such as cubic splines. The number and position of the knots, or points at which the polynomial segments connect, are chosen by the user based on the shape of the data.

The spline fitter allows for flexible modeling of the hazard rate function, without making any assumptions about its shape. It is particularly useful for data that exhibit complex or irregular patterns over time, such as those arising from repairable systems like mining trucks.

To fit a spline to survival data, one typically uses a method known as penalized likelihood estimation, which balances the trade-off between fitting the data well and avoiding overfitting. The resulting spline model can then be used to predict the hazard rate at any point in time and to estimate the remaining useful life of components such as air filters in mining trucks.

Overall, the spline fitter is a powerful tool for predicting the remaining useful life of components with complex failure patterns. It allows for flexible modeling of the hazard rate function and can capture irregular patterns that other methods may miss.

3.8.13 Kaplanmeir Fitter

The Kaplan-Meier fitter is a non-parametric method used in survival analysis to estimate the survival function of a dataset. The survival function gives the probability that an event has not occurred up to a given time, and is defined as $S(t) = P(T > t)$, where T is the time to event.

The Kaplan-Meier estimator is used to estimate the survival function based on censored data, where some of the observed times are right-censored, meaning that the event of interest has not yet occurred by the end of the study period. The estimator takes the form of a step function that increases at the observed event times and remains constant between them.

The survival function for the Kaplan-Meier estimator can be expressed as follows:

$$\mathbf{S(t) = \Pi[i: t\_i \leq t]\ (1 - d\_i/n\_i)}$$

where:

t_i is the time of the ith event

d_i is the number of events at time t_i

n_i is the number of individuals at risk just before time t_i

The product in the equation represents the cumulative probability of surviving up to time t, taking into account the number of individuals who have experienced an event and those who are still at risk. The Kaplan-Meier estimator assumes that the hazard rate is not constant over time, but may change in response to external factors such as aging or wear and tear.

The Kaplan-Meier estimator is a popular method for analyzing survival data because it is simple to use and can handle censored data without requiring any assumptions about the underlying distribution of the data. It is often used to estimate the survival function for a specific population or subgroup, such as mining truck air filters, and can be used to compare survival curves between different groups.

To fit a Kaplan-Meier estimator to survival data, one typically uses a statistical software package that provides the necessary tools for data input, analysis, and visualization. The resulting estimator can then be used to estimate the survival function at any point in time and to calculate the median survival time, which is a common measure of the remaining useful life of components such as air filters in mining trucks.

Overall, the Kaplan-Meier estimator is a powerful tool for estimating the survival function of a dataset and can be used to predict the remaining useful life of components with complex failure patterns. It is particularly useful for analyzing censored data and does not require any assumptions about the underlying distribution of the data.

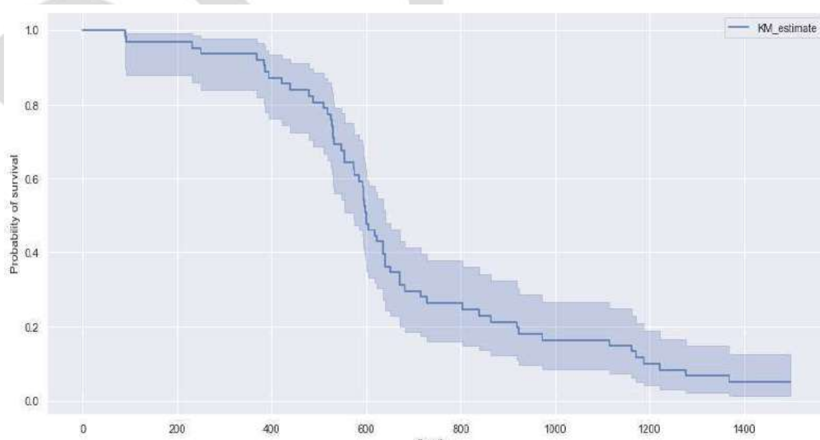


Figure 3.27: Kaplan-Meier Results

3.9 Machine Learning Model

Machine learning approach refers to the use of algorithms and statistical models that enable computer systems to learn and improve their performance on a specific task by analyzing and learning from data, without being explicitly programmed.

3.9.1 NASA’s Engine

NASA's Engine Degradation Simulation (EDS) dataset is a publicly available dataset that can be used for predicting the remaining useful life (RUL) of aircraft engines. The EDS dataset was collected as part of a NASA research program aimed at developing and testing new methods for RUL prediction in aircraft engines.

The EDS dataset consists of simulation results from a physically-based engine degradation model, which simulates the effects of different operating conditions and component failures on the health and performance of aircraft engines over time. The simulation results include time series data on engine performance and health parameters, as well as information on the operating conditions and component failures that occurred during the simulation.

The EDS dataset can be used for various RUL prediction tasks, such as:

Regression-based RUL prediction: In this approach, regression models are trained on the engine performance and health parameters from the EDS dataset to predict the RUL of aircraft engines.

Prognostic and health management (PHM): In this approach, the EDS dataset is used to develop and evaluate algorithms for real-time monitoring and prediction of engine health and RUL, based on engine performance and health parameters.

Deep learning-based RUL prediction: In this approach, deep learning algorithms are used to analyze the time series data from the EDS dataset to predict the RUL of aircraft engines.

The EDS dataset is a valuable resource for researchers and engineers working in the field of RUL prediction, as it provides a large and diverse set of data on engine degradation, which can be used to develop, test, and validate RUL prediction algorithms.

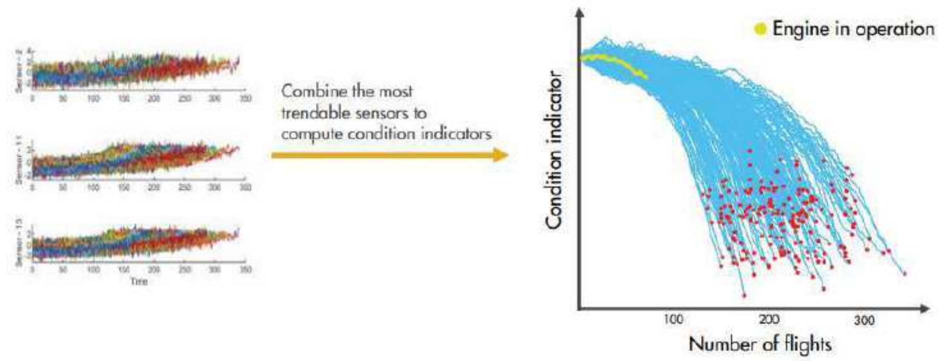


Figure 3.28: NASA's Engine Module

3.9.2 Binary Classification

Binary classification is a machine learning technique that can be used for predicting the remaining useful life (RUL) of aircraft engines. In this approach, the task of RUL prediction is framed as a binary classification problem, where the goal is to predict whether an engine will fail within a specified time frame or not.

To perform binary classification, the available data on engine performance and health parameters is divided into two classes: positive (failure) and negative (non-failure). A machine learning algorithm is then trained on this labelled data to learn the patterns and relationships between engine performance and health parameters and the likelihood of engine failure.

Once the model has been trained, it can be used to predict the class label (failure or non-failure) for new, unseen data points. The prediction can be used to estimate the remaining useful life of the engine and to make maintenance decisions based on the predicted risk of failure.

Binary classification can be useful for RUL prediction in aircraft engines because it provides a simple and straightforward approach for predicting the risk of engine failure. This approach can be particularly useful in cases where there is limited data available on engine performance and health parameters, or where the data is noisy and unreliable.

However, it's important to note that binary classification has some limitations and may not be the best approach for RUL prediction in all cases. For example, if the goal is to estimate the exact remaining useful life of an engine, rather than simply predicting whether it will fail or not, regression-based approaches may be more appropriate.

Multiple binary classification algorithms were used to predict which filter will fail in the current period, e.g., remaining cycles or life time in the range 0-1700 SMH. Evaluated algorithms include Logistic Regression, Decision Trees, Support Vector Machines, K Nearest Neighbors, Naive Bayes, and Random Forests. A short definition of these algorithms.

3.9.3 Linear Model

Linear models are a type of machine learning algorithm that can be used for predicting the remaining useful life (RUL) of aircraft engines. In a linear model, the relationship between the predictor variables (engine performance and health parameters) and the response variable (RUL) is modeled as a linear combination of the predictor variables.

Linear models are popular in RUL prediction because they are simple to implement, fast to train, and interpretable, making them a good choice for cases where interpretability is important. In addition, linear models can be used for both regression and classification tasks, making them flexible for a range of RUL prediction scenarios.

To train a linear model for RUL prediction, the available data on engine performance and health parameters is used to fit the model to the data. The model parameters are then estimated using optimization techniques, such as least squares estimation or maximum likelihood estimation.

Once the model has been trained, it can be used to make predictions for new, unseen data points. The predicted RUL can be used to estimate the remaining useful life of the engine and to make maintenance decisions based on the predicted risk of failure.

It's important to note that linear models have some limitations and may not be the best approach for RUL prediction in all cases. For example, if the relationship between the predictor variables and the response variable is non-linear, a linear model may not be able to capture the complex relationships in the data, leading to poor performance. In such cases, non-linear models, such as polynomial regression or neural networks, may be more appropriate.

Linear models estimate RUL by predicting KPA using currant value based on mathematical degradation equation.

$$\frac{\boldsymbol{MaxValue - CurrentValue}}{\boldsymbol{Mean(KPA)}} * \boldsymbol{Mean\_hour}$$

Where $MaxValue$ stands for max KPA value, $CurrentValue$ stands for current KPA of filter & $Mean(KPA)$ stands for mean value of KPA.

3.9.4 Growth Model

A growth model is a type of machine learning algorithm that can be used for predicting the remaining useful life (RUL) of aircraft engines. A growth model reflect the degradation of an engine over time, allowing the prediction of the RUL based on the current state of the engine and its degradation rate.

Growth models are particularly useful in RUL prediction because they take into account the time-dependent nature of engine degradation, making them well suited for tasks where time is an important factor. Some common types of growth models used for RUL prediction include exponential growth models, logistic growth models, and Gompertz growth models.

To train a growth model for RUL prediction, the available data on engine performance and health parameters is used to fit the model to the data. The model parameters are then estimated using optimization techniques, such as least squares estimation or maximum likelihood estimation.

Once the model has been trained, it can be used to make predictions for new, unseen data points. The predicted RUL can be used to estimate the remaining useful life of the engine and to make maintenance decisions based on the predicted risk of failure.

It's important to note that growth models have some limitations and may not be the best approach for RUL prediction in all cases. For example, if the relationship between the predictor variables and the response variable is not well modeled by a growth function, a growth model may not be able to capture the complex relationships in the data, leading to poor performance. In such cases, other models, such as regression models or neural networks, may be more appropriate.

Growth model Equation is as below:

$$\boldsymbol{predicted\_kpa(t) = first\_kpa * exp(k * t)}$$

The air filter degradation process will be affected by uncertain factors, such as complex working conditions, and so on.

3.10 Hybrid Model

One approach to predicting the remaining useful life (RUL) of air filters in mining trucks could be to combine a linear model with a survival analysis model, such as a Weibull fitter.

The linear model could be used to predict the air filter degradation based on features such as the age of the air filter, the type of material used, and environmental factors such as humidity and temperature. This model would essentially provide a baseline prediction of the remaining useful life of the air filter.

The survival analysis model, such as the Weibull fitter, could then be used to estimate the probability distribution of the remaining useful life, based on the degradation data obtained from the mining trucks. The Weibull distribution is commonly used in reliability engineering and can model the failure rate of a system over time, making it a suitable choice for modeling the remaining useful life of air filters.

Combining the predictions from the linear model and the survival analysis model could result in a more accurate prediction of the RUL. For example, the linear model could provide an estimate of the remaining useful life based on the known factors, while the Weibull fitter could adjust the prediction based on the degradation data obtained from the mining trucks.

Overall, the combined approach of using a linear model and a survival analysis model could provide a more accurate and reliable prediction of the remaining useful life of air filters in mining trucks, which would be valuable for optimizing maintenance schedules and reducing downtime.

We have developed our custom hybrid model, which is the combination of linear model & survival model.

Chapter 4 : RESULT AND DISCUSSIONS

In this chapter the trained model is going to evaluate. The training and testing accuracy on the trained model would be shown here. Results of all the models which are tested is shown in this chapter.

Experimentational Results:

This section includes the experimental results of training. Many experiments are performed while training different ML & statistical models.

4.1 RUL & KPA results

Data consist of more than 1000 remaining useful life of air-filters, a particular KPA limit was set which shows the filter gets plugged out.

Null hypothesis states that there is no correlation between air-filter differential pressure and operating time of the truck. Rate of increment of kpa does not affect RUL prediction.

Alternative hypothesis states that there is a linear correlation between the Air filter differential pressure and the operating time of the truck.

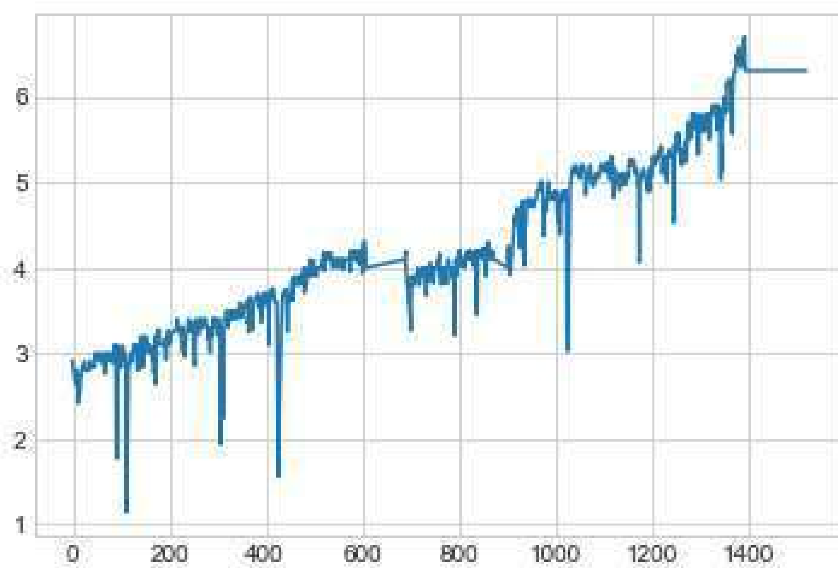


Figure 4.1: RUL vs KPA

A rolling mean with a window size of 10 & 50 would be calculated by taking the average of the first five data points, then the average of the second through sixth data points, and so on until the end of the dataset is reached. This creates a new dataset that is smaller than the original, with fewer fluctuations and smoother trends.

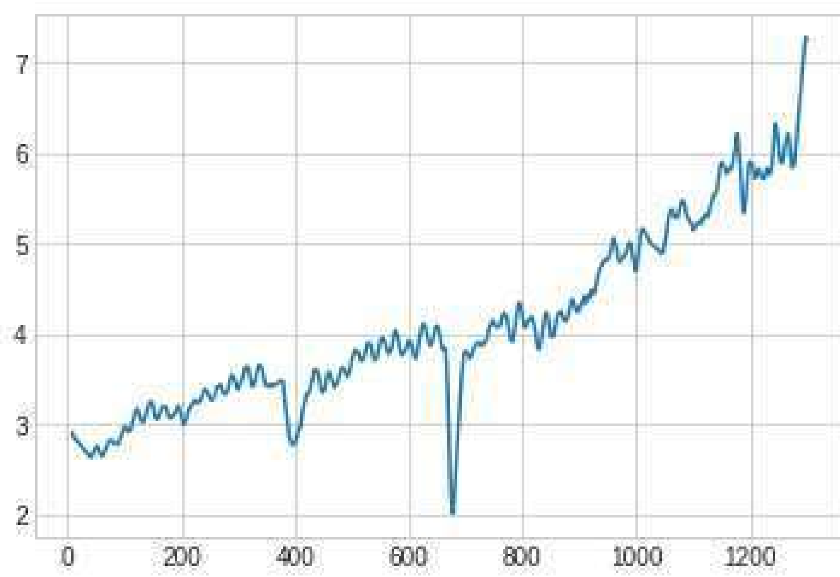


Figure 4.2: RUL Vs KPA with Rolling Mean_10

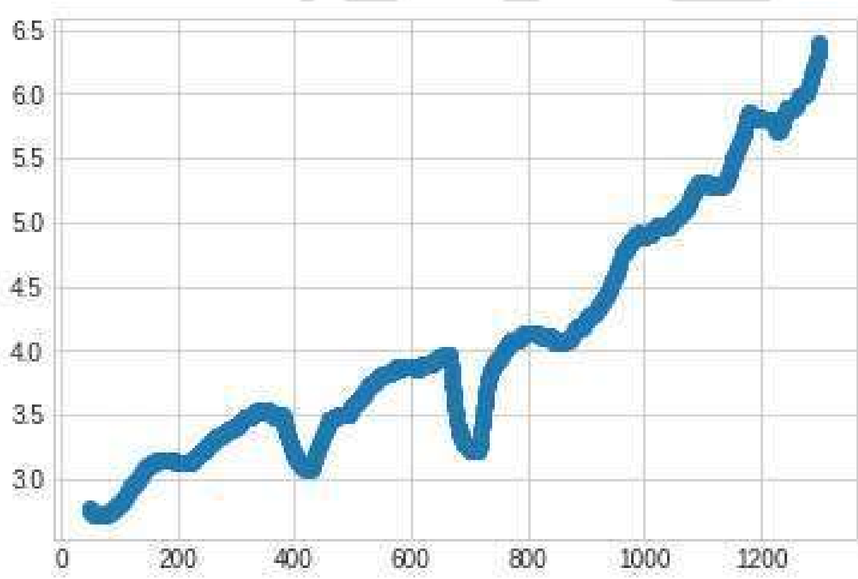


Figure 4.3: KPA Vs RUL with Rolling Mean_50

Significant difference between the predicted remaining useful life and the actual remaining useful life, this can indicate a problem with the model or a change in the operating conditions that was not accounted for. In some cases, it may be possible to improve the model or adjust the maintenance schedule to better reflect the actual conditions of the asset. In other cases, the difference may be due to unforeseen failures or unexpected changes in the operating environment, and may require additional investigation to identify the root cause.

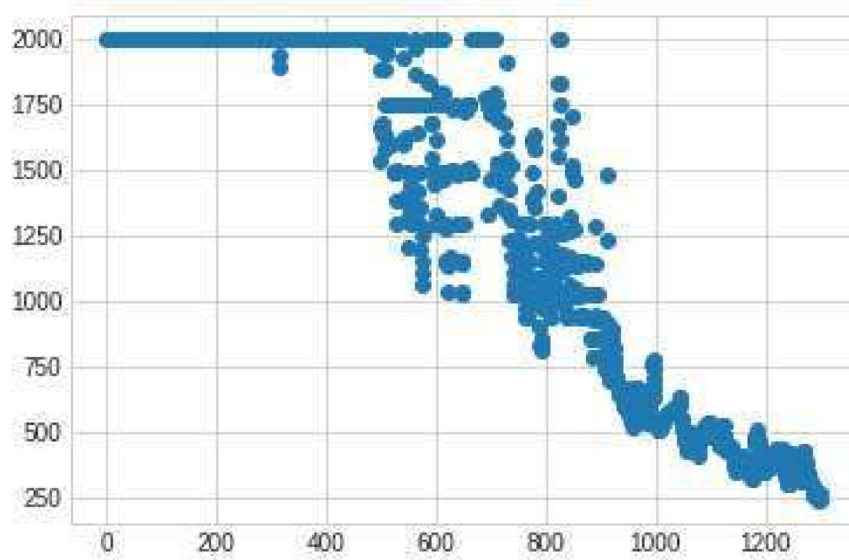


Figure 4.4: RUL Difference over time

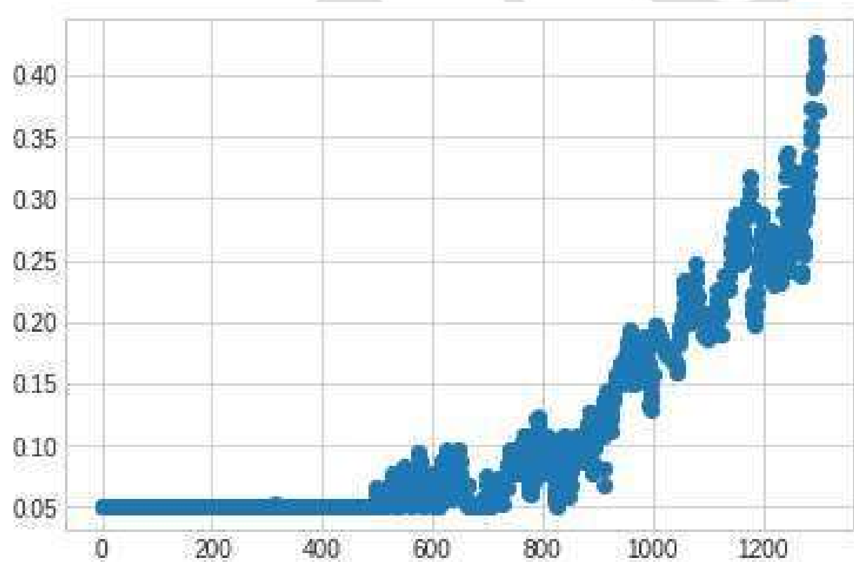


Figure 4.5: KPA & RUL Difference

Least squares is a statistical method used to estimate the parameters of a linear regression model by minimizing the sum of the squares of the residuals, which are the differences between the observed values and the predicted values. In other words, it is a method for finding the line of best fit for a given set of data.

The least squares method involves finding the values of the regression coefficients (intercept and slope) that minimize the sum of the squared differences between the observed data points and the predicted values of the dependent variable. This is done by calculating the partial derivatives of the sum of the squared residuals with respect to the regression coefficients, setting them equal to zero, and solving the resulting system of equations.

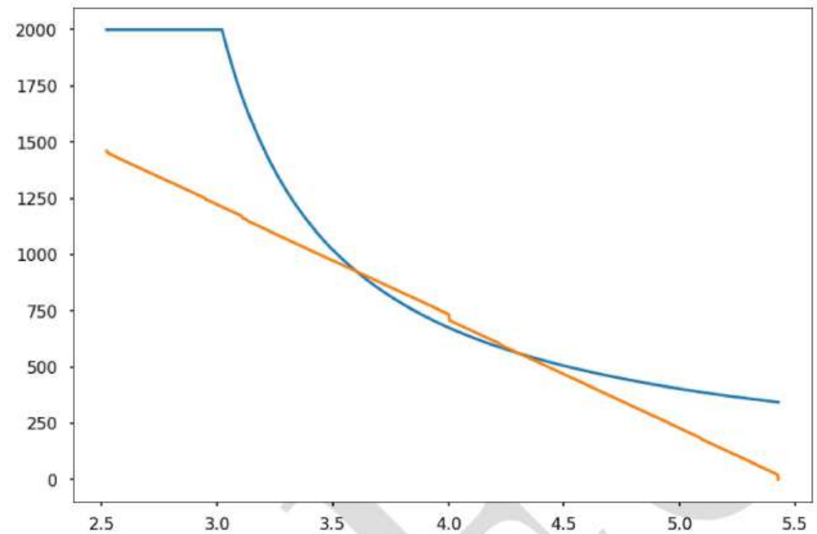


Figure 4.6: Least Square-RUL

Applied exponential curve fitting to check the parameters fit the given set of data. Exponential curve fitting is a statistical method used to estimate the parameters of an exponential function that best fits a given set of data points. Exponential functions have the form y = ab^x, where a and b are constants and x is the independent variable.

The curve fitting process involves finding the values of a and b that minimize the sum of the squared differences between the observed data points and the predicted values of the dependent variable. This is typically done using nonlinear regression techniques, which iteratively adjust the values of a and b until the sum of the squared residuals is minimized.

Exponential curve fitting is commonly used in many fields, including biology, physics, and finance, to model processes that exhibit exponential growth or decay over time. Examples of such processes include population growth, radioactive decay, and the decay of chemical compounds. Exponential curve fitting can also be used to make predictions about future values of the dependent variable based on the fitted model.

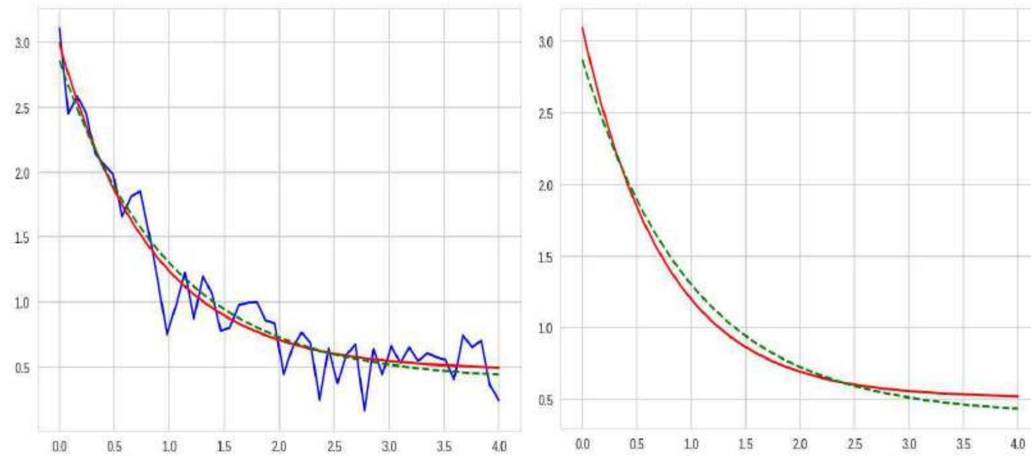


Figure 4.7: Exponential Curve Fitting

4.2 Statistical Model Results

Similarity Model: To determine the expected failure time of a current filter, the similarity model identifies the filter profiles that are most similar to the current filter's key performance attributes (KPA) up to the current state of the system. By examining the failure times of the closest matching filters in historical truck data, the model can estimate the anticipated failure time of the current filter.

The similarity model employed the K-Nearest Neighbours (KNN) algorithm to identify the most similar filter paths. During each iteration of the Remaining Useful Life (RUL) computation, the model locates the closest filter paths and calculates the RUL based on nearby values.

Survival Model: To calculate the survival probability, the number of filters that have reached the end of their useful life is divided by the total number of filters in a "bad condition." In this context, "bad condition" refers to filters that have failed, as opposed to those that have been removed, dropped out, or had their subject status changed. The overall probability of survival up to a given time interval is obtained by multiplying the probabilities of survival at all preceding time intervals using the law of multiplication of probabilities to determine the cumulative probability.

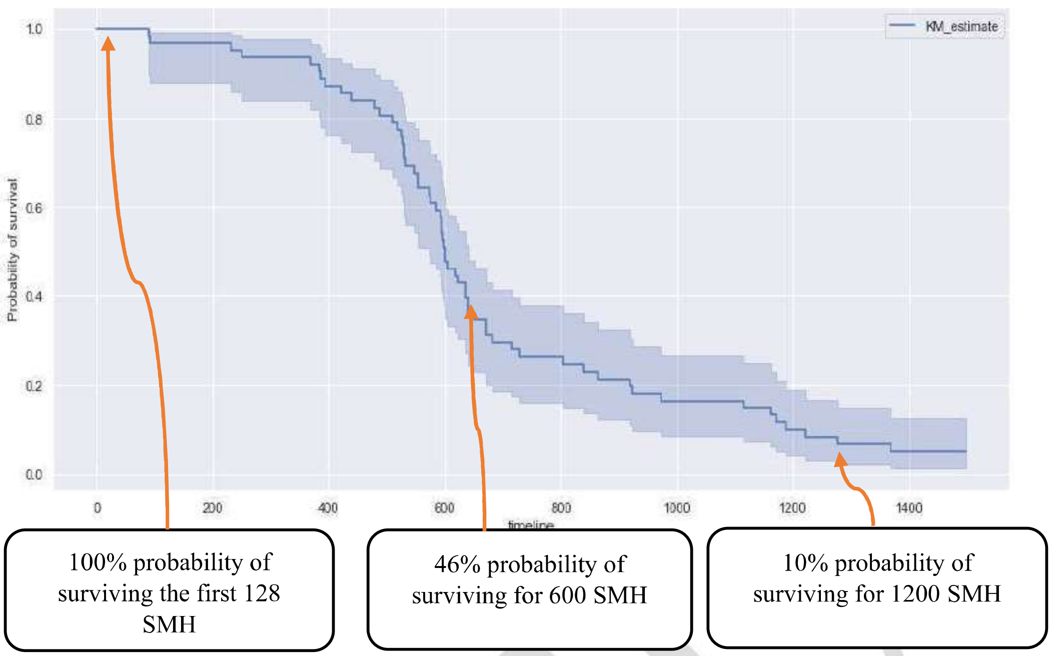


Figure 4.8: Kaplan-Meier Probability of Survival

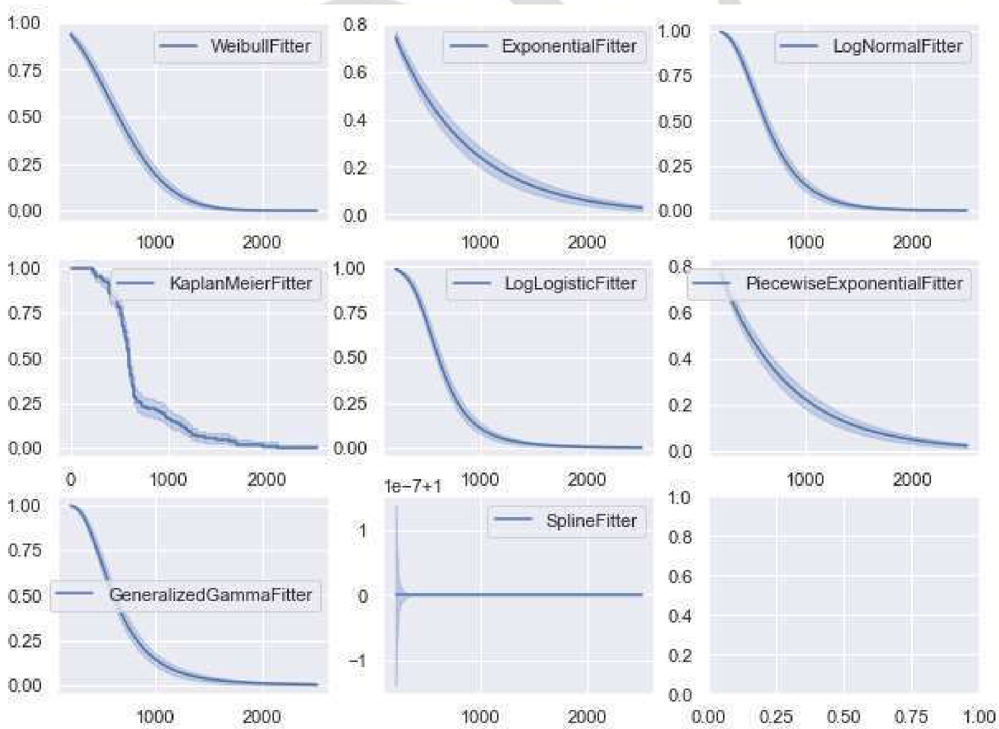


Figure 4.9: Survival Statistical Models

The WeibullFitter was found to be the most effective fitter for survival analysis among other fitting algorithms. The cumulative probability of survival up to a specific time interval is calculated by multiplying the probabilities of survival at all preceding time intervals, in accordance with the law of multiplication of probabilities.

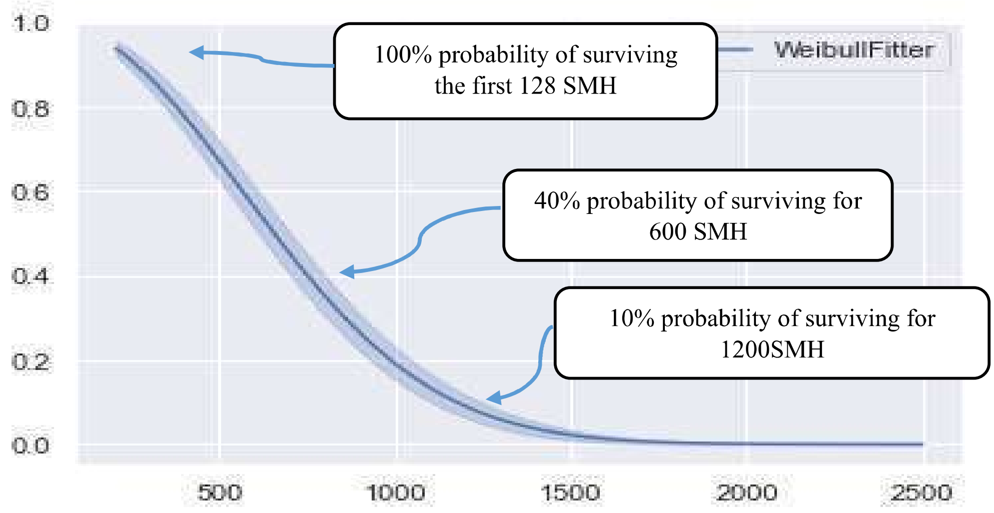


Figure 4.10: Results of Weibull Fitter

4.3 Binary Model Results

Various binary classification algorithms were employed to predict the failure of a filter within the current period, based on its remaining cycles or remaining life within the range of 0-1700 cycles. These algorithms included Logistic Regression, Decision Trees, Support Vector Machines, K Nearest Neighbors, Naive Bayes, and Random Forests. Here is a brief definition of these algorithms:

- Logistic Regression: a statistical method used to model the relationship between a binary dependent variable (in this case, filter failure) and one or more independent variables (such as KPA data).
- Decision Trees: a tree-based model that splits the data into smaller subsets based on the values of the input variables, and uses these subsets to make predictions.
- Support Vector Machines (SVM): a model that separates the data into different classes using a hyperplane, with the goal of maximizing the margin between the classes.
- K Nearest Neighbors (KNN): a non-parametric model that uses the nearest K neighbors of a given data point to classify it.

- Naive Bayes: a probabilistic model that makes predictions based on the probability of each class given the input variables.
- Random Forests: an ensemble model that combines multiple decision trees to improve the accuracy of the predictions.

In binary classification, the goal is to determine whether a given filter will fail or not based on its remaining life and KPA data.

Table 4.1: ML Model Results

	LR A	DT B	DT A	RF B	RF A	SVC B	SVC A
Accuracy	0.84	0.99	0.991	0.97	0.99	0.99	0.99
Precision	1	1	1	0.88	1	1	1
Recall	0.09	0.95	0.95	0.95	0.95	0.95	0.95
F1-Score	0.16	0.97	0.97	0.92	0.97	0.97	0.97
ROC AUC	0.98	0.99	0.98	0.99	0.98	0.98	0.97

Table 4.2: ML Model Results

	SVC B	SVC A	KNN B	KNN A	Gaussian NB B	Gaussian NB A
Accuracy	0.83	0.99	0.85	0.99	0.81	0.94
Precision	0.57	1	1	1	0.44	0.98
Recall	0.08	0.95	0.15	0.95	0.34	0.66
F1-Score	0.14	0.97	0.26	0.97	0.39	0.79
ROC AUC	0.75	0.99	0.96	0.98	0.79	0.98

To assess the performance of the binary classification algorithms, several key metrics were calculated, including the Area under the Receiver Operating Characteristics Curve (AUC ROC), Recall, Precision, F1 Score, and Accuracy. During hyper-parameter tuning using Grid Search, the AUC ROC was used as the score to evaluate the models.

The models' performance was evaluated on a test dataset, and the results were compared between the original feature set (referred to as "B" or "Before features extraction") and the modified feature set (referred to as "A" or "After feature extraction"). The evaluation showed how the different models performed with each feature set.

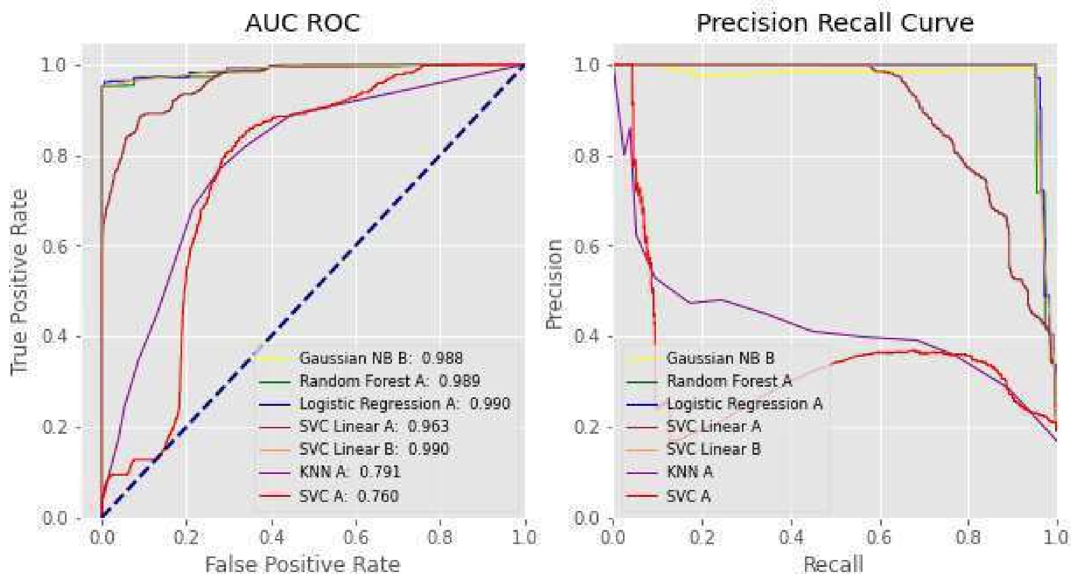


Figure 4.11: AUC-ROC & Precision Recall Curves

The evaluation of the binary classification algorithms revealed that both Logistic Regression and Random Forests achieved the best AUC ROC scores. Moreover, it was observed that feature extraction led to an improvement in the performance metrics of most models.

To further visualize the performance of the best models, AUC for ROC and Precision-Recall curves were plotted. These curves provide a graphical representation of the trade-off between the true positive rate and the false positive rate, as well as the precision and recall of the models.

4.4 Growth Model Results

A growth model in machine learning refers to a model that predicts the growth of a system over time based on historical data. It is often used in time-series analysis to forecast the growth of a variable or process over time, such as predicting the future revenue of a company, the number of customers for a product, or the traffic on a website.

Growth models use various statistical and machine learning techniques, such as regression analysis, autoregressive integrated moving average (ARIMA), and exponential smoothing. These models can be trained on historical data to identify patterns, trends, and seasonality in the growth of the system, which can be used to make predictions about its future behavior.

One of the challenges in building growth models is to capture the non-linear and dynamic nature of the growth process, which may involve factors such as external events, changes in market conditions, and feedback loops. Therefore, machine learning techniques such as neural networks and decision trees are also used to model the complex relationships between different variables and to improve the accuracy of the growth predictions.

Growth model is estimate RUL using current KPA and current SMH

Growth model Equation is as below.

$$\boldsymbol{predicted\_kpa(t) = first\_kpa * exp(k * t)}$$

Example :

Current Kpa =3.2, Current SMH=420,

first_kpa=1.67(initial value of filter),

Test_kpa(max kpa)=7.5

Using the given value Ve determine K

$$\boldsymbol{k = log(current\_kpa/first\_kpa)/t}$$

$$\boldsymbol{k = log(3.2/1.67)/420}$$

$$\boldsymbol{k = 0.005419393194808477}$$

$$\boldsymbol{RUL \ = log(test\_kpa/first\_kpa)/k}$$

$$\boldsymbol{RUL = 970.0065\ Hours}$$

The air filter degradation process will be affected by uncertain factors, such as complex working conditions, and so on.

4.5 Hybrid Model Results

A custom hybrid model has been developed that combines different models to improve its accuracy. The hybrid model is a combination of two models: the survival model and Multivariate Linear Regression (MLR). The output of the MLR model is bounded at 600 hours, and Weibull weights have been added to the hybrid model. The Weibull model is dependent only on SMH (the time since the last maintenance), so it has been incorporated into the hybrid model to enhance its accuracy.

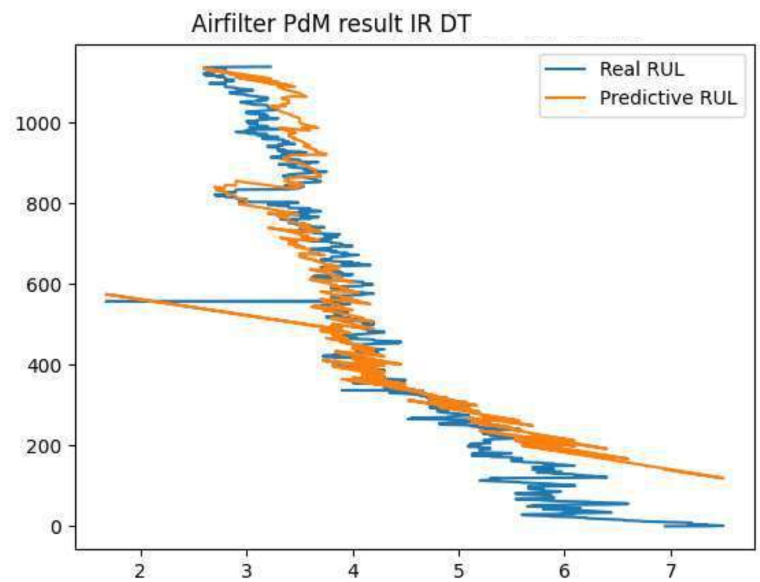


Figure 4.12: RUL results before applying hybrid model

To remove the gap at the end of the RUL prediction, a Multivariate Linear Regression (MLR) model was trained to include values greater than 5 kPa. As shown in the graph, the RUL near 7 kPa was approximately 200, and the MLR model was used to improve the accuracy of the prediction at this range.

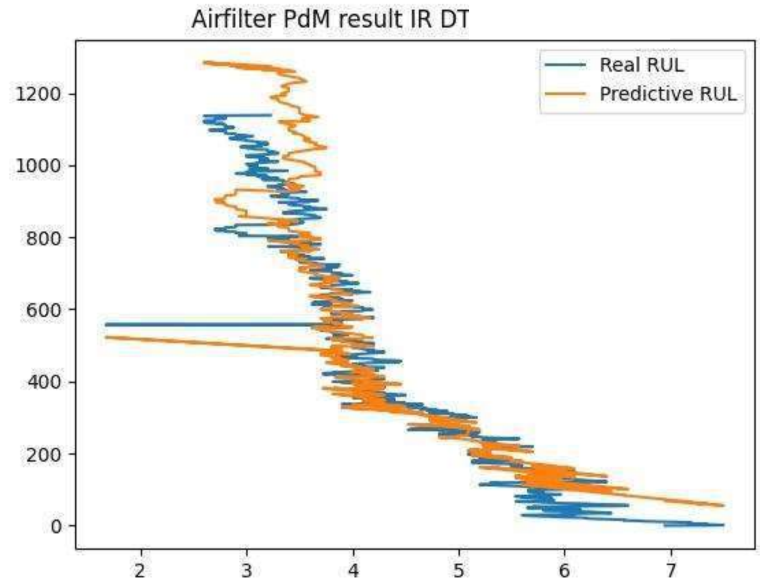


Figure 4.13: RUL results after applying hybrid model

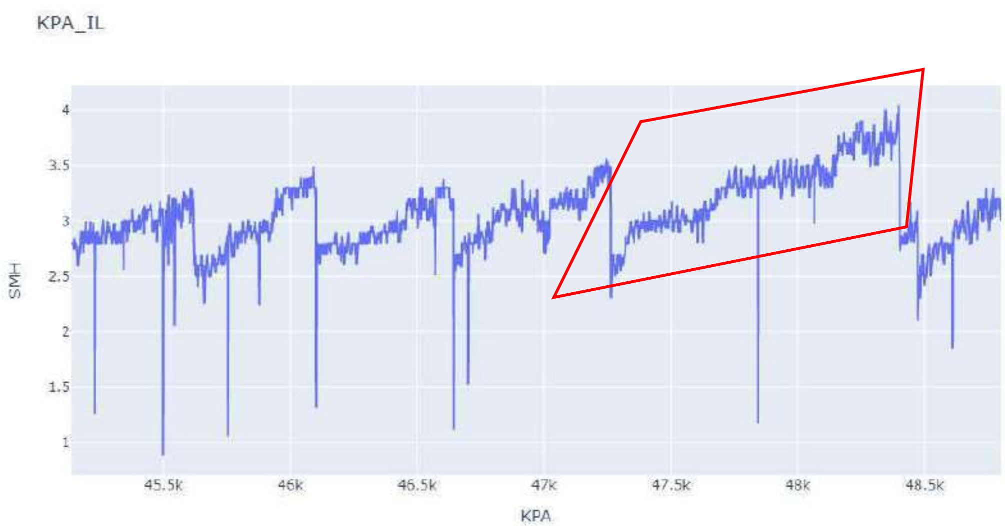


Figure 4.14: RUL cycle consider for analysis

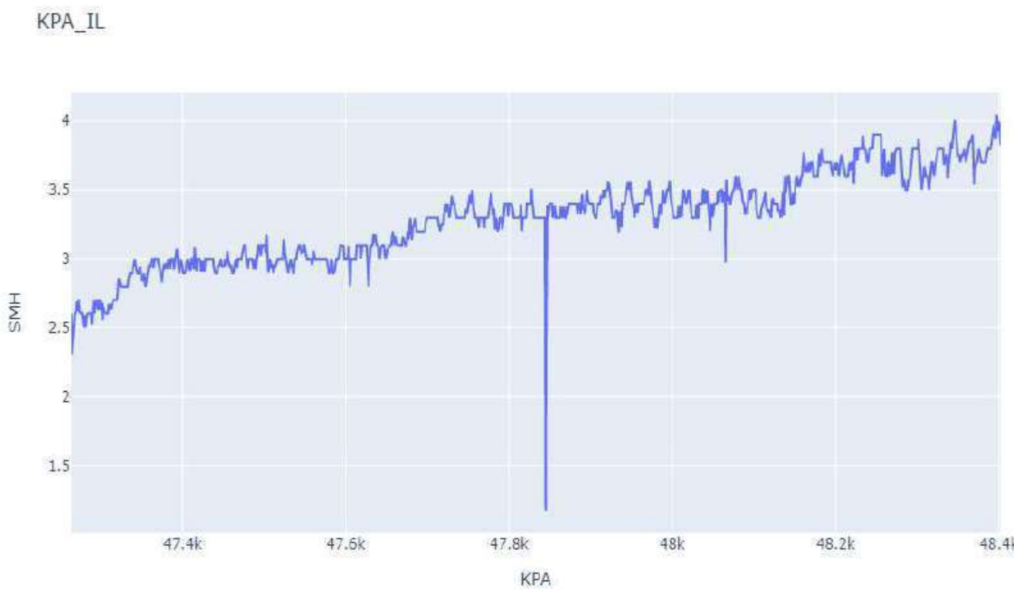


Figure 4.15: Clear view of RUL cycle consider for analysis

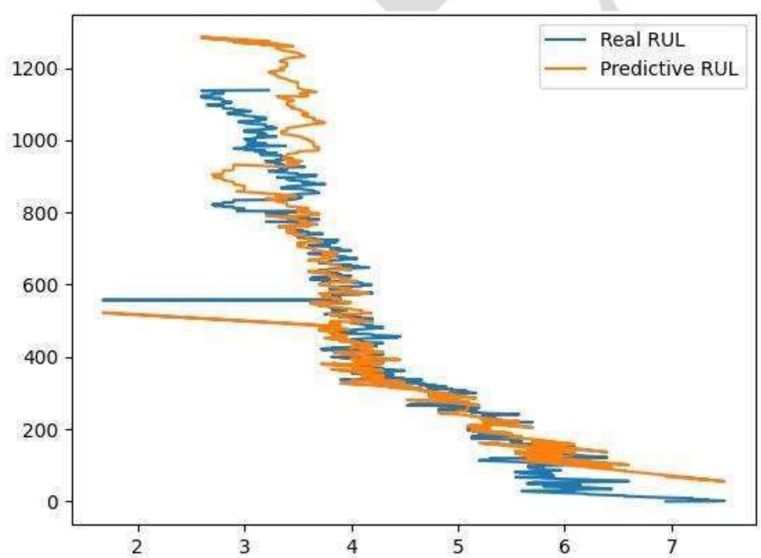
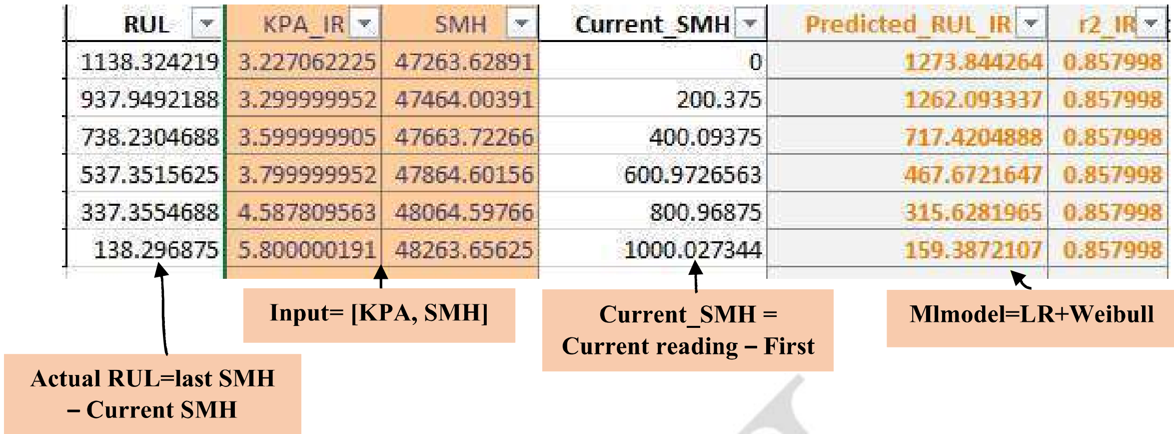


Figure 4.16: Actual vs Predicted RUL



RUL	KPA_IR	SMH	Current_SMH	Predicted_RUL_IR	r2_IR
1138.324219	3.227062225	47263.62891	0	1273.844264	0.857998
937.9492188	3.299999952	47464.00391	200.375	1262.093337	0.857998
738.2304688	3.599999905	47663.72266	400.09375	717.4204888	0.857998
537.3515625	3.799999952	47864.60156	600.9726563	467.6721647	0.857998
337.3554688	4.587809563	48064.59766	800.96875	315.6281965	0.857998
138.296875	5.800000191	48263.65625	1000.027344	159.3872107	0.857998

Figure 4.17: Hybrid Model predicted results

ML model used in this calculation is a combination of a linear model and a Weibull fitter. A linear model is a statistical model that assumes a linear relationship between the input variables and the output variable. A Weibull fitter is a statistical model used to analyze the failure rates of a system or process.

To reformulate the calculation in more detail:

1. Calculate the current SMh reading: Subtract the first SMh reading from the current SMh reading.

   Current SMh = Current reading - First reading

2. Calculate the actual rule (AR): Subtract the current SMh reading from the last SMh reading.

   AR = Last SMh - Current SMh

3. Use a ML model to predict the AR based on the input variables KPA and SMh.

- The ML model used is a combination of a linear model and a Weibull fitter.
- The linear model assumes that there is a linear relationship between KPA and AR.
- The Weibull fitter analyzes the failure rates of the system and predicts the probability of failure based on the SMh.

The output of the ML model is the predicted AR.

Overall, this calculation is used to predict the remaining life of a system or process based on its current KPA and SMh readings, and to predict the probability of failure based on the Weibull fitter.

Below figures shows the result of predicted RUL of each filters

Table 4.3: Hybrid Model Predicted Results for Truck_1

RUL	Predicted_RUL_IL	Predicted_RUL_IR	Predicted_RUL_OL	Predicted_RUL_OR
1383.0273	1279.63	1283.65	1284.52	1286.105574
1382.8476	1279.86	1283.79	1284.79	1286.397279
1382.6679	1280.08	1283.93	1285.06	1286.688983
1382.4882	1280.31	1284.08	1285.33	1286.980688
1382.1328	1280.53	1284.21	1285.59	1287.263184
1381.7304	1280.74	1284.34	1285.85	1287.543229
1381.1171	1280.95	1284.46	1286.10	1287.812229

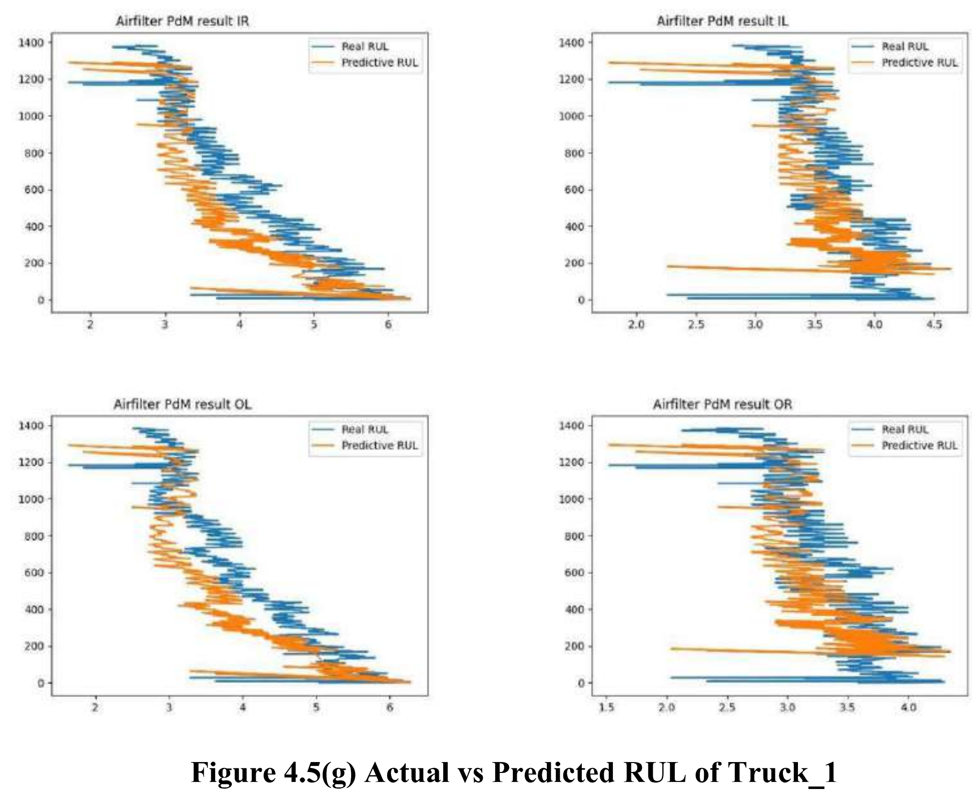


Figure 4.5(g) Actual vs Predicted RUL of Truck_1

Table 4.4: Hybrid Model Predicted Results for Truck_2

RUL	Predicted_RUL_IL	Predicted_RUL_IR	Predicted_RUL_OL	Predicted_RUL_OR
1328.98	1285.80	1280.043	1290.783682	1287.333003
1328.94	1285.56	1279.57	1290.464646	1287.013972
1328.89	1285.32	1279.09	1290.145409	1286.694735
1328.83	1285.08	1278.61	1289.826173	1286.375495
1328.79	1284.844714	1278.13	1289.507141	1286.056463
1328.7	1284.604621	1277.66	1289.187901	1285.737227
1328.69	1284.364733	1277.18	1288.868869	1285.418195

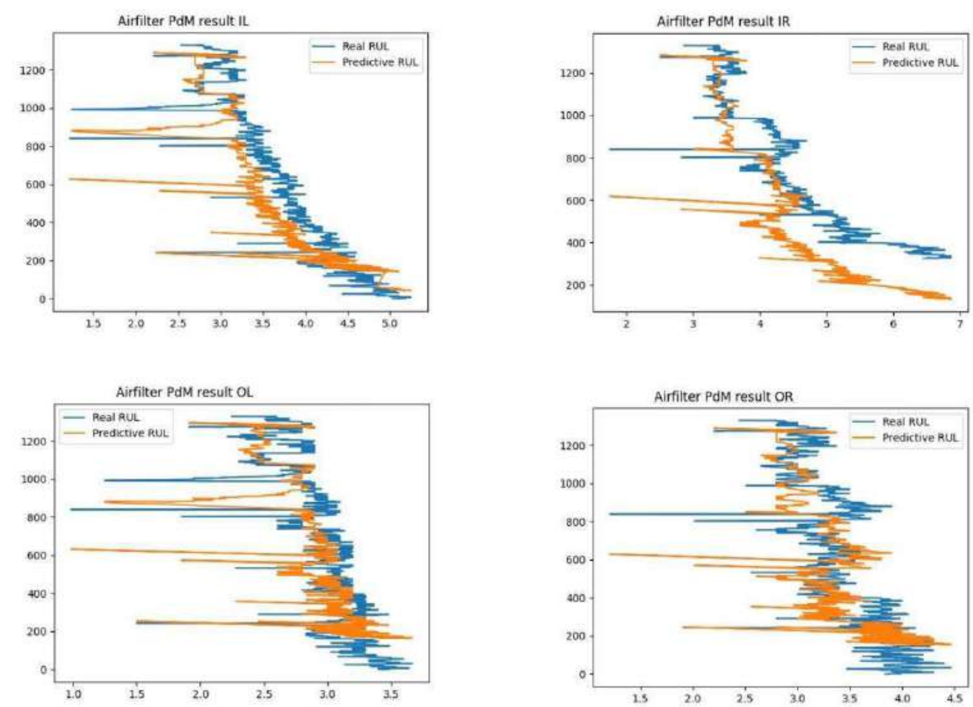


Figure 4.18: Actual vs Predicted RUL of Truck_2

Table 4.5: Hybrid Model Predicted Results for Truck_3

RUL	Predicted_RUL_IL	Predicted_RUL_IR	Predicted_RUL_OL	Predicted_RUL_OR
1270.3125	1283.698908	1279.290142	1288.875978	1282.554902
1270.160156	1284.100682	1280.101672	1289.277752	1283.132284
1270.011719	1284.502665	1280.913407	1289.679735	1283.709875
1269.863281	1284.904644	1281.725141	1290.081714	1284.287466
1269.714844	1285.306626	1282.536876	1290.483697	1284.865053
1269.566406	1285.708605	1283.348607	1290.885675	1285.442644
1269.003906	1286.088905	1284.138659	1291.265975	1285.998552
1268.015625	1285.499966	1283.770226	1291.439451	1285.478701

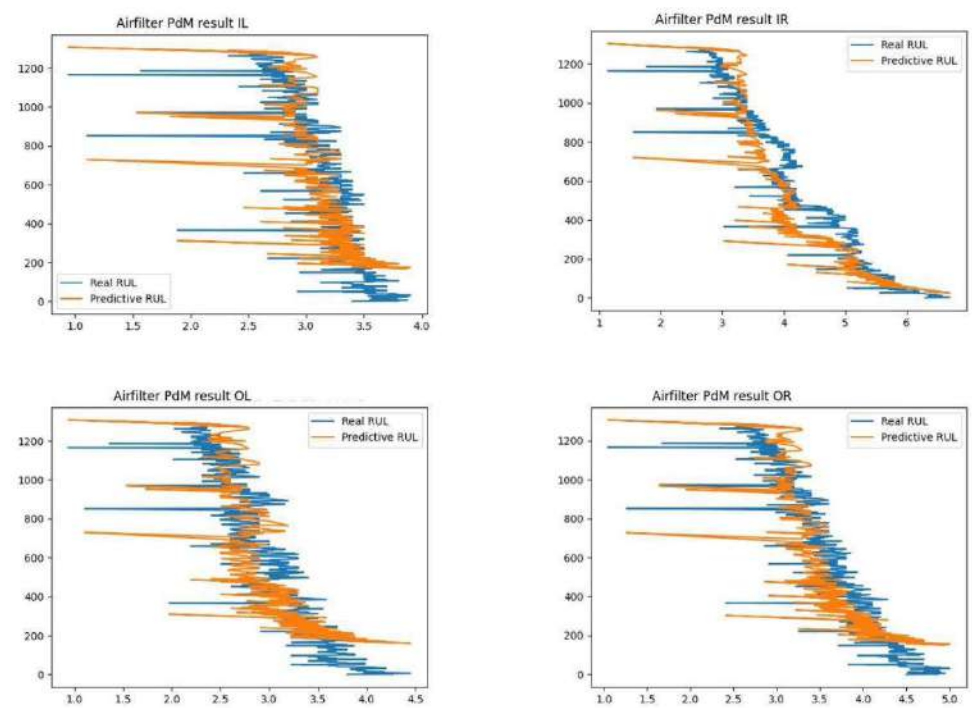


Figure 4.19: Actual vs Predicted RUL of Truck_3

A dashboard that provides important insights into the performance of various filters in a cycle detection system was developed. The dashboard displays the number of cycles detected in each filter, namely inner right, inner left, outer right, and outer left. Additionally, the dashboard presents key metrics such as mean kilopascals (kPA), model error rate, total truck hours, and total hours required to reach 7.5KPA.

The number of cycles detected in each filter is an important indicator of the filter's effectiveness in identifying cycles. This information can be used to optimize the performance of the cycle detection system and improve the accuracy of the system in detecting cycles.

The mean kPA is a measure of the average pressure exerted on the filter during a cycle. This metric provides valuable insights into the performance of the cycle detection system and can be used to identify potential issues with the system.

The model error rate is a metric that indicates the accuracy of the cycle detection model. This metric is crucial in ensuring that the cycle detection system is reliable and effective in identifying cycles.

The total truck hours metric provides an overview of the total operating time of the truck. This information can be used to track the performance of the truck and identify potential issues that may impact the accuracy of the cycle detection system.

Finally, the total hours required to reach 7.5KPA metric is an important measure of the time required for the cycle detection system to accurately detect cycles. This metric can be used to optimize the performance of the system and improve its accuracy in detecting cycles.

Overall, our dashboard provides valuable insights into the performance of the cycle detection system, enabling operators to identify potential issues and optimize the system for improved performance.

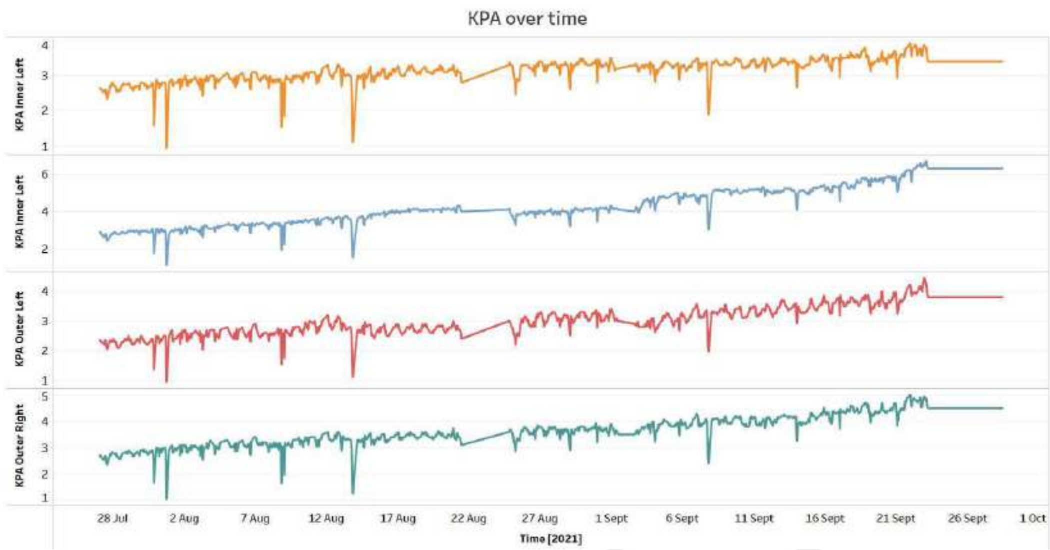


Figure 4.20: Dashboard Section_1

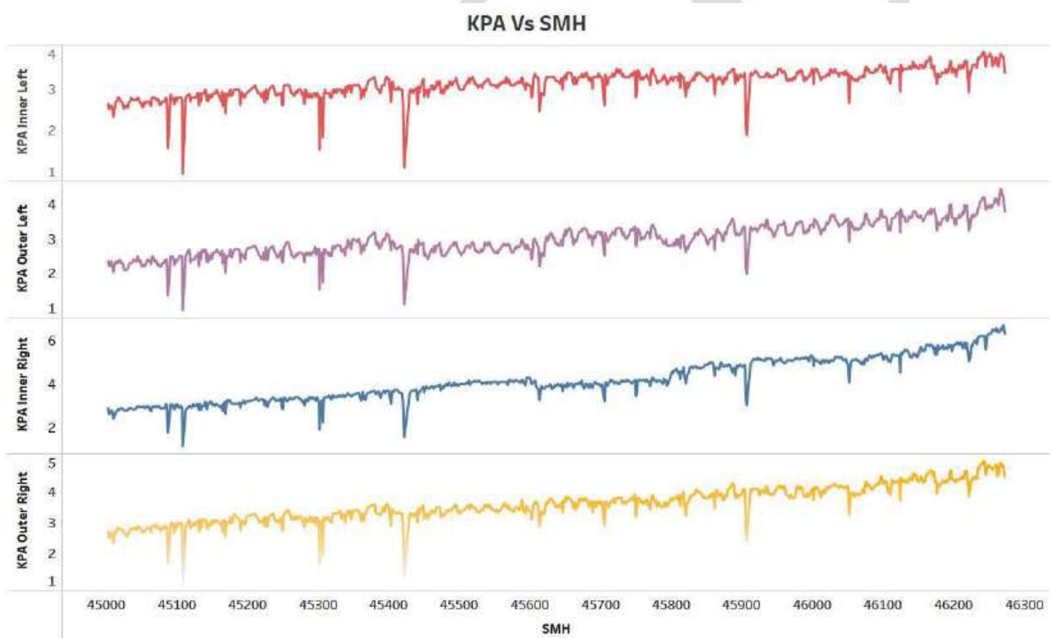


Figure 4.21: Dashboard Section_2

Filtered out last cycle from the SHM slider point

Good Data Pattern Found

Model Accuracy in Range

RUL results of four filters

RUL parameters	Inner Left Air Filter	Inner Right Air Filter	Outer Left Air Filter	Outer Right Air Filter
Mean KPA	2.86	3.02	3.34	2.96
Rate of KPA increment per data point	0.19	0.16	0.21	0.19
Total remaining useful life	539.81	620.25	469.95	540.03
Total truck run hours	185.5	185.69	185.68	185.68
Hours remaining to reach 7.5 KPA	232	224	208	227
Total hour to rrach 7.5 KPA	409.68	417.68	397.68	412.36
Model error rate	0.05	0.04	0.08	0.06
Recommendation	NIL	NIL	NIL	NIL
Comments	NIL	NIL	NIL	NIL

Figure 4.22: Dashboard Section_3 showing results of data fetch and KPA limit with model accuracy

Chapter 5 : CONCLUSION & FUTURE WORK

5.1 Conclusion:

i. **KNN**: This model's accuracy improves with time as more accurate RUL values are obtained. It relies solely on historical data with RUL values, and better accuracy can be achieved with more complete cycle data.

ii. **Regression**: Linear, Polynomial, Random Forest Regression, and Miller Regression models have an accuracy range of 600 to 150 hours, with errors being higher than the KNN model. More complete cycle data is required for better accuracy.

iii. **Survival model:** Several models are available, including Weibull Fitter, Exponential Fitter, Log Normal Fitter, Log Logistic Fitter, Nelson Aalen Fitter, Piecewise Exponential Fitter, Generalized-Gamma Fitter, Spline Fitter & Kaplan-Meier Fitter from which Weibull Fitter is the best fir for filter lifeline.

iv. **Binary classification**: Predicting whether the filter is alive or not is not useful in RUL estimation. The data is heavily imbalanced, with 97% of class 0 and only 3% of class 1.

v. **Statistical model:** The linear model is a data-driven equation that automatically changes its constants as more data is added. However, sometimes the constants are too low, resulting in very large RUL values.

vi. **Growth model:** The model's accuracy improves over time and follows the forecast curve. However, it produces large errors in the initial hours.

vii. **Hybrid model:** This model combines two linear regression models and WeibullFitter. Among all the above models, this model gives the best results with an accuracy of over 90%.

5.2 Limitations:

i. At some point, algorithm fails to detect the cycle properly and need to fix it.

ii. Sometimes model fails to detect proper RUL cycle.

iii. Getting bad data or more error rate in data.

iv. Hardware Failures.

v. Due to RUL limitation from client side, needs more data containing more than 1400 remaining useful life of air-filters.

5.3 Future scopes:

Different literature survey, tells that there are multiple techniques which can be used to predict RUL.

At some point, algorithm fails to detect the cycle properly and need to fix it.
There are several potential future directions for research in the area of predicting the remaining useful life of air filters in mining trucks:

Improved prediction models: There is room for improvement in existing prediction models, such as by incorporating more advanced techniques, such as deep learning or reinforcement learning, to make more accurate predictions.

Incorporation of additional data sources: Currently, air filter degradation is typically estimated based on data from a single sensor or a limited number of sensors. Incorporating additional data sources, such as weather data or data on the operating conditions of the mining truck, could provide a more complete picture of the factors that influence air filter degradation and lead to more accurate predictions.

Integration with maintenance decision-making: The ultimate goal of RUL prediction is to support maintenance decision-making, so future research could focus on developing models that can be integrated into existing maintenance decision-making processes and provide real-time predictions to support proactive maintenance.

These are just a few potential directions for future research in this field, and there may be other areas of focus as well. The important thing is to identify the areas where the current state of knowledge is lacking and where improvements can be made to better predict the remaining useful life of air filters in mining trucks.

Chapter 6 : REFERENCES

[1] Lei, Y.; Li, N.; Guo, L.; Li, N.; Yan, T.; Lin, J. Machinery health prognostics: A systematic review from data acquisition to RUL prediction. Mech. Syst. Signal Process. 2018, 104, 799–834.

[2] Xi, B.; Wang, C.; Xi, W.; Liu, Y.; Wang, H.; Yang, Y. Experimental investigation on the water hammer characteristic of stalling fluid in eccentric casing–tubing annulus. Energy 2022, 253, 124113.

[3] Chao, Q.; Gao, H.; Tao, J.; Wang, Y.; Zhou, J.; Liu, C. Adaptive decision–level fusion strategy for the fault diagnosis of axial piston pumps using multiple channels of vibration signals. Sci. China Technol. Sci. 2022, 65, 470–480.

[4] Chao, Q.; Zhang, J.; Xu, B.; Wang, Q.; Lu, F.; Li, K. Integrated slipper retainer mechanism to eliminate slipper wear in high–speed axial piston pumps. Front. Mech. Eng. 2022, 17, 1.

[5] Tang, S.N.; Zhu, Y.; Yuan, S.Q. An Improved Convolutional Neural Network with an Adaptable Learning Rate towards Multi–signal Fault Diagnosis of Hydraulic Piston Pump. Adv. Eng. Inform. 2021, 50, 101406.

[6] Yuan, X.M.; Wang, W.Q.; Zhu, X. Theory Model of Dynamic Bulk Modulus for Aerated Hydraulic Fluid. Chin. J. Mech. Eng. 2022, 35, 121.

[7] Tang, S.N.; Zhu, Y.; Yuan, S.Q. A Novel Adaptive Convolutional Neural Network for Fault Diagnosis of Hydraulic Piston Pump with Acoustic Images. Adv. Eng. Inform. 2022, 52, 101554.

[8] Tang, S.N.; Zhu, Y.; Yuan, S.Q. An adaptive deep learning model towards fault diagnosis of hydraulic piston pump using pressure signal. Eng. Fail. Anal. 2022, 138, 106300.

[9] Li, N.; Lei, Y.; Lin, J.; Ding, S.X. An Improved Exponential Model for Predicting Remaining Useful Life of Rolling Element Bearings. IEEE Trans. Ind. Electron. 2015, 62, 7762–7773.

[10] Wang, Z.Q.; Hu, C.H.; Fan, H.D. Real–Time Remaining Useful Life Prediction for a Nonlinear Degrading System in Service:Application to Bearing Data. IEEE/ASME Trans. Mechatron. 2018, 23, 211–222.

[11] Nian, F. Viewpoints about the prognostic and health management. Chin. J. Sci. Instrum. 2018, 39, 1–14.

[12] Liu, H.; Liu, Z.; Jia, W.; Zhang, D.; Tan, J. Current research and challenges of deep learning for equipment remaining useful life prediction. Comput. Integr. Manuf. Syst. 2021, 27, 34–52.

[13] Pei, H.; Hu, C.; Si, X.; Zhang, J.; Pang, Z.; Zhang, P. Review of Machine Learning Based Remaining Useful Life Prediction Methods for Equipment. J. Mech. Eng. 2019, 55, 1–13.

[14] Peng, C.; Qian, F.; Du, X. New nonlinear degradation modeling and residual life prediction. Comput. Integr. Manuf. Syst. 2019,25, 1647–1654.

[15] Y. Zheng, "Predicting remaining useful life based on Hilbert–Huang entropy with degradation model," Journal of Electrical and Computer Engineering, vol. 2019, 2019.

[16] R. Celaya, A. Saxena and K. Goebel, "Uncertainty representation and interpretation in model-based prognostics algorithms based on kalman filter estimation," 2012

[17] Prediction of Remaining Useful Life of Equipment based on Condition Monitoring and Expert Knowledge Using Neuro-Fuzzy Inference System

[18] "Remaining Useful Life (RUL) Prediction of Equipment in Production Lines Using Artificial Neural Networks"

[19] Tamara Fernández, Nicolás Rivera, Yee Whye Teh "Gaussian Processes for Survival Analysis"

[20] "Understanding survival analysis: Kaplan-Meier estimate" Manish Kumar Goel,Pardeep Khanna, Jugal Kishore

[21] International Organization for Standardization 2006, ISO/TS 16610-22:2006, Geometrical Product Specifications (GPS) – Filtration – Part 22: Linear Profile Filters: Spline Filters, ISO. Geneva, Switzerland.

[22] Krystek, M. 1996b, 'Discrete L-spline filtering in roundness measurements', Measurement, vol. 18, no. 2, pp. 129–138.

[23] Krystek, M. 2000, 'Discrete linear profile filters', Proceedings of the X International Colloquium on Surfaces, Chemnitz University of Technology, Chemnitz, pp.145–152.

[24] "Lognormal Kalman filter for assimilating phase space density data in the radiation belts", D. Kondrashov, M. Ghil, Y. Shprits

[25] "Fitting the log-logistic distribution by generalized moments", Fahim Ashkar, Smail Mahdi

[26] Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. IEEE Trans. Pattern Anal. and Machine Intel., 24(4):509–522, April 2002

[27] Borg and P. J. F. Groenen. Modern Multidimensional Scaling: Theory and Applications. Springer, New York, 2nd edition, 2005

[28] Cazzanti. Generative Models for Similarity-based Classification. PhD thesis, Dept. of Electrical Engineerng, University of Washington, 2007.

[29] E. Driskell and T. McDonald. Identification of incomplete networks. Technical report, Florida Maxima Corporation, 2008.