SVKM's NMIMS
Mukesh Patel School of Technology Management & Engineering

A REPORT ON

Predictive Machine Failure

By:

Organization: Indian Oil Corporation Limited

A REPORT ON

Predictive Machine Failure

BY

ROHAN MATHUR
(B.Tech C.E.)

UNDER THE GUIDANCE OF
Industry Mentor:

Mr. NPS SIDHU
(DGM(IS), Indian Oil Corporation Limited)

Faculty Mentor:

Prof. UPENDRA VERMA

A report submitted in partial fulfillment of the requirements of 4 years B.Tech Program of Mukesh Patel School of Technology Management & Engineering, NMIMS.

CERTIFICATE

इंडियन ऑयल कॉर्पोरेशन लिमिटेड
गुजरात रिफाइनरी, डाकघर : जवाहरनगर
जिला - वडोदरा - गुजरात - 391 320
Indian Oil Corporation Limited
Gujarat Refinery, P.O. Jawaharnagar,
Dist. : Vadodara, Gujarat - 391 320
दूरभाष : + 91-265-2237106
ई-मेल : gujaratrefinery@indianoil.in वेबसाइट : www.iocl.com

स्वच्छ भारत
एक कदम स्वच्छता की ओर

IndianOil

रिफ़ाइनरी प्रभाग
Refineries Division

JR/TD/2022/0008 **Date: 25/06/2022**

Certificate

This is to certify that ***Mr. Rohan Mathur****, student of* ***B.TECH*** *programme from* ***Mukesh Patel School of Technology Management & Engineering, Mumbai****, has undergone Industrial Training at Indian Oil Corporation Limited, Gujarat Refinery, Vadodara from* ***02/05/2022 to 25/06/2022*** *as a part of the course curriculum.*

He/She has successfully completed the training and submitted a report with his/her overall performance being ***Excellent****.*

Mentor: Mr. NPS SIDHU

DGM (IS)

एन.पी.एस.सिधू
N.P.S. SIDHU
उप महाप्रबंधक (सूचना प्रणाली)
Deputy General Manager (Information Systems)
गुजरात रिफाइनरी, आई.ओ.सी.एल., वडोदरा
Gujarat Refinery, IOCL, Vadodara

MANOJ GAWANDE

DGM (MS, L&D)

MANOJ P. GAWANDE
उप महाप्रबंधक (मानव संसाधन विकास)
DGM (MS, L&D)
गुजरात रिफाइनरी, आई.ओ.सी.एल., वडोदरा
Gujarat Refinery, IOCL, Vadodara

SVKM's NMIMS

Mukesh Patel School of Technology Management & Engineering
Shirpur, Maharashtra

TECHNICAL INTERNSHIP REPORT Semester VII– B.TECH

Submitted in Partial Fulfillment of the requirements for Technical Project/Training for VII Semester B.Tech

Name of Student: Rohan Mathur

Roll No.: B232

SAP ID : 70021119040

Academic Year: 2022

Name of the Discipline: Computer Engineering

Name and Address of Company:
Indian Oil Corporation Limited (IOCL)
Gujarat Refinery, PO Jawaharnagar, Vadodara - 391320, Gujarat, India

Training Period: From 02/05/2021 To 25/06/2021

THIS IS TO CERTIFY THAT
Rohan Mathur has satisfactorily completed his Training/Project work, submitted the training report, and appeared for the presentation and viva as required.

External Examiner	Internal Examiner	Head of Dept.	Chairperson/Dean

Date:
Place:

Seal of University

Table of Contents

Acknowledgement

Without the generous assistance, direction, and support of many people, this endeavor would not have been feasible. To everyone who helped us finish this report, we would like to convey our sincere gratitude. We would like to thank Mr. NPS Sidhu, our industry mentor and Mr. Sandeepan Sen, for giving us the chance to design this project and successfully apply our abilities and technologies to it. We would especially want to express our gratitude to Prof. Upendra Verma, our professor and mentor for our project, for his assistance in motivating comments and encouragement, which helped us to organize our project and assisted us in preparing this report.

Additionally, we want to express our gratitude towards our parents & friends for their kind cooperation and encouragement which helped us in completing this project efficiently.

Student Name: Rohan Mathur
B.Tech C.E.
Roll No.: B232
(70021119040)

Abstract

In this project, we would predict the failure category and failure type of vivid operating machines used in the petroleum sector (specific to Indian Oil Corporation Limited) on the basis of different attributes or features so as to get a point-to-point idea about the machines failure rate according to which further planning can be done for getting maximum throughput taking in considerable maintenance and recovery periods of the predicted failed machines in advance.

The prediction will be performed using multiple machine learning algorithms like Decision Tree, Random Forest, Support Vector Machine (SVM) and etc. The algorithm with the maximum accuracy and efficiency will be the most suitable of all the analyzed algorithms. In fact, the correctness and efficiency will be checked through possible visualization tools and representations so as to get a more detailed output check. Therefore, it is of utmost importance to go through these so that the production and manufacturing is performed accordingly keeping an eye on inventory and buffer stocks simultaneously.

List of the figures:

List of the table:

Chapter 1

Introduction

1.1 Introduction of the Industry

The oil and gas industry is one of the largest sectors in the world in terms of dollar value, generating an estimated $5 trillion in global revenue as of 2022. Oil is crucial to the global economic framework, impacting everything from transportation to heating & electricity to industrial production & manufacturing.

Oil is a truly global industry. It can be transported around the world and refined into gasoline or petroleum which can then be made into a slew of other products like balloons or bandages. Natural gas tends to be more regional in nature, as its transportation and distribution is dependent on pipelines. Natural gas helps create electricity, fuel and heating.

The Oil & Gas industry operates in three segments: upstream, midstream and downstream.

Upstream companies, also known as exploration and production (E&P), find, develop, and produce oil, natural gas, and natural gas liquids. Upstream companies manage their development and production costs and emphasize production volume to generate profit margins, which are sensitive to commodity market prices.

Midstream companies gather, process, store, and transport crude oil, raw and liquified natural gas, and refined petroleum products and chemicals. The midstream business model is similar to a toll road that charges fees for the movement or intermediate processing of Oil & Gas.

Downstream companies refine petroleum products and engage in the manufacturing, marketing, and distribution of refined petroleum products such as gasoline, jet fuel, heating oil, asphalt, motor oil, and lubricants.

1.2 About the Company

Indian Oil Corporation Limited (IOCL)

- IOCL is India's no. 1 fortune 500 company.
- Indian Oil Corporation Limited (IOCL) is the largest commercial oil company in India.
- Indian Oil has ventured into alternative energy and globalization of downstream operations. It has subsidiaries in Sri Lanka, Mauritius, the UAE, Singapore, Sweden, USA and The Netherlands.

Indian Oil Corporation Limited (IOCL) business interests overlap the entire hydrocarbon value-chain, including

1. Refining
2. Pipeline transportation
3. Marketing of petroleum products
4. Exploration and production of crude oil, natural gas and petrochemicals

The IS (Information Systems) department handles all IT related infrastructure of IOCL, from procurement of systems to implementation of security related guidelines of the Indian government.

Role of IS department is as follows:

1. **Procurement** - In this purchasing of different items and services takes place through tendering. The items may be different hardwares and softwares, machinery tools and other goods as well.
2. **Software Development** - The IS department develops certain softwares that are used within the company. Here the clients are the employees themselves. These softwares are used for intra-communication and are accessible within the organization.
3. **LAN Laying** - This design involves connecting multiple systems through a well implemented cable architecture. It is primarily laid with an objective of communication and transmission.

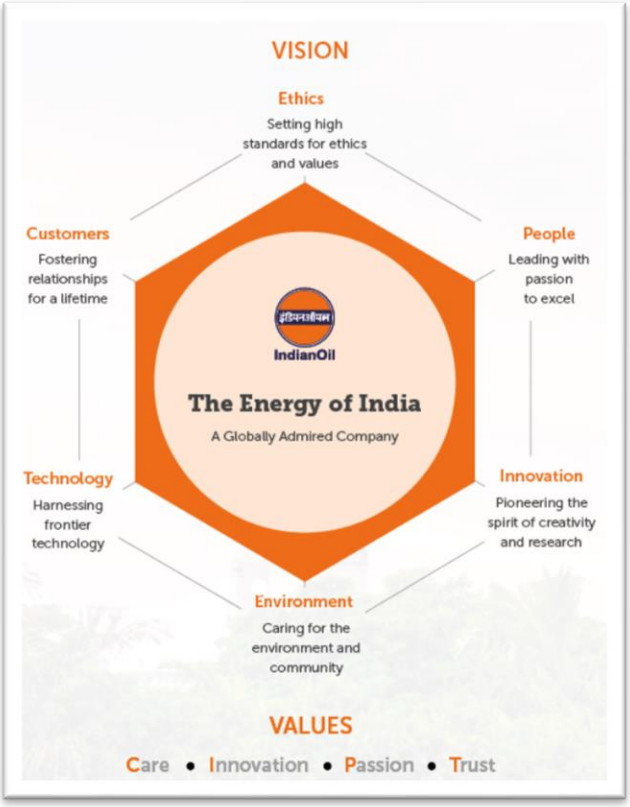

Figure 1: Vision of IOCL

SAP Implementation - Major Activities: The following activities are inbuilt in the implementation of SAP.

- Conceptual Design
- Detailed Design
- Construction
- Implementation
- Data Conversion

3-tier Architecture of SAP:

A. Presentation Layer: Typically installed on a PC, provides the SAP Graphical User Interface.

B. Application Layer: Executes the business logic, process client transactions, print jobs, running reports, coordinate access to the Database.

C. Database Layer: Stores both the business generated data and SAP application programs, which are loaded into application servers from the database.

1.3 Background of the project topic

Machines form a very crucial part in a manufacturing industry and this fact is particularly true in the case of the oil industry. These are required in one form or another at every step of the manufacturing cycle to allow each individual process to be completed. The estimated time that machines may be down due to various contributing factors.

It is technically not possible to accurately and precisely manually observe and calculate estimated time periods for such a large number of machines operating too frequently. Therefore, there must be a way to predict the failure state of machines that may fail at some point in the future. The main reason for carrying out this project was to get an overview of the number of machines that can fail and the reasons for their failure in advance, in order to be able to plan the future development of the company today based on the results generated with a certain efficiency.

1.4 Motivation and scope of the project

The aim of our project is to build a machine learning model that would be trained using the historical data which will be used to predict the type of breakdown that a machine has gone under. This will help to predict the failure easily and hence would save the costs. The goal of this project includes:

- To observe the past data of machine maintenance from the dataset provided by the company and predict the type of failure a machine has undergone so that it could be put under maintenance immediately.
- It will help the organization to minimize their monetary loss and increase the company's profits.

1.5 Salient Contribution

- The Predictive Maintenance of Machines mechanism will help to find the failure status of machines in advance.
- It can predict with a certain level of accuracy that the machine under observation will fail or will work as usual.
- There are a wide range of machines for which the model works efficiently.

1.6 Project Timeline

Week	Tasks
1	• Brief introduction to the organization and basic fire and safety training. • Got to know about the refinery, IS (Information Systems) department • How the ERP network is spread across the country and how it operates.
2	• Brief understanding and working of SAP system.
3	• Project topic discussion with mentor and discussion on various ideas and scope of projects that can be done in this training . • Project Topic allocation
4-5	• Met few employees from mechanical department and discussed the necessary parameters regarding the project • Collected the necessary information for the project. • Found sample dataset
6	• Data preprocessing and visualization • Exploratory Data analysis (EDA)
7	• Data Preprocessing for Prediction 1. Encoding categorical features 2. Splitting test & train data 3. Feature Scaling
8	• Model development • Comparison of multiple models and selecting the best model

Table 1: Project Timeline

Chapter 2

Problem Statement

To predict the failure and failure type of different operating machines so as to estimate the average maintenance period for vivid machines in advance according to which further plans can be made with the intention of maximizing the throughput and production considering all the affecting factors before hand only.

The objective is to estimate the number of machines that may fail anytime in future hampering the manufacturing activities and slowing down the production levels. Prediction will be made by training and testing multiple ML models based on common features which directly or indirectly contribute to machine failure and simultaneously the failure type will be determined for a machine. Out of all the implemented models, the best of all will be chosen and be used for future predictions. One with the highest efficiency and certain accuracy will come out to be the best of all.

Chapter 3

Literature Survey

[5]Ana Rita Santiago proposed a Big Data Service architecture devised for heating, ventilation and air-conditioning monitoring and optimization. The experimentation has shown that the system shall have adjustments regarding the data processing and the models precision. First, time processing increases based on data expansion, second the precision is below 50%. However, the low precision can be related to the fact that the models are developed using an unbalanced dataset. Although some of the results were not perfect, an impact evaluation was conducted to understand how the results can respond to the problem that was meant to be solved. This evaluation has demonstrated that the user comfort has risen since the solution is able to identify more than 30% of the failures.

[7]Liqing Qiu established the mathematical model based on maximization of averaging service rate of machines and was able to improve the efficiency of the machine, and the average service rate of preventive maintenance period from 15 days to 35 days. Since the machine failure rate in the short preventive maintenance period under 24 days is relatively low, it is not an optimal strategy, besides the time cost of short period preventive maintenance reduces the average service rate of the machine. The long preventive maintenance period above 24 days increases the machine failure rate, therefore increases the breakdown maintenance time cost, which reduces the average service rate.

[8]Mohammed Lafi, Bilal Hawashin and Shadi AlZu’bi used Random Forest Classifier, AdaBoot and Decision Tree for predicting machine maintenance in which they achieved 78% precision, 83% recall, and 79%F1-score.

[9]Qian Cao, Hongsheng Su and Xueqian Wang used the newly introduced the TBM strategy and CBM strategy, analyzed the advantages and disadvantages of the two strategies, and verified by actual operation data

Chapter 4

Methodology

Importing the dataset

A dataset that contains a list of various industrial machines along with factors that impact their working conditions. These factors are important as multiple failure categories depend differently on these factors. Dataset also contains the target attributes that tell us whether for the given values of these factors, has the machine failed or not, and if failed then the type of failure is also mentioned.

	UDI	Product ID	Type	Air temperature [K]	Process temperature [K]	Rotational speed [rpm]	Torque [Nm]	Tool wear [min]	Target	Failure Type
0	1	M14860	M	298.1	308.6	1551	42.8	0	0	No Failure
1	2	L47181	L	298.2	308.7	1408	46.3	3	0	No Failure
2	3	L47182	L	298.1	308.5	1498	49.4	5	0	No Failure
3	4	L47183	L	298.2	308.6	1433	39.5	7	0	No Failure
4	5	L47184	L	298.2	308.7	1408	40.0	9	0	No Failure
5	6	M14865	M	298.1	308.6	1425	41.9	11	0	No Failure
6	7	L47186	L	298.1	308.6	1558	42.4	14	0	No Failure
7	8	L47187	L	298.1	308.6	1527	40.2	16	0	No Failure
8	9	M14868	M	298.3	308.7	1667	28.6	18	0	No Failure
9	10	M14869	M	298.5	309.0	1741	28.0	21	0	No Failure

Figure 2: Dataset View

Data Preprocessing:

Data preprocessing is the process of transforming raw data into an understandable format. It is also an important step in data mining as we cannot work with raw data. The quality of the data should be checked before applying machine learning or data mining algorithms.

Preprocessing of data is mainly to check the data quality. The quality can be checked by the following

- **Accuracy**: To check whether the data entered is correct or not.
- **Completeness**: To check whether the data is available or not recorded.

- **Consistency**: To check whether the same data is kept in all the places that do or do not match.
- **Timeliness**: The data should be updated correctly.
- **Believability**: The data should be trustable.
- **Interpretability**: The understandability of the data.

Steps performed for data preprocessing:

1. The first step in Data Preprocessing was to understand the dataset properly and to analyze the same so as to get an overview about data.
2. Further, we used the statistical methods or pre-built libraries in python so as to visualize the dataset and get a clear image of how the data looks in terms of distribution.
3. Next, we summarized the data in terms of the number of duplicates, missing values, and outliers present in the data.
4. Then we dropped the fields that were irrelevant to our model implementation and development. Dimensionality reduction is important so as to analyze and perform modeling.
5. Lastly, we proceeded with some feature engineering activities and figured out which attributes contribute most towards model training.

Key purposes to use EDA:

- So in this dataset, we have removed unwanted columns as a part of data preprocessing since they don't carry any useful information.
- Apart from that no missing values were found in the dataset, thus we can now perform Exploratory Data Analysis.

Exploratory Data Analysis (EDA)

It is an approach to perform initial investigations on data to discover patterns, spot anomalies, test hypotheses and check assumptions with the help of statistics and graphical representations.

- The main purpose of EDA is to examine the data before making assumptions. This helps identify obvious errors, better understand patterns in the data, find

outliers or anomalous events, and find interesting relationships between variables.

- Therefore, data scientists can use exploratory analysis to ensure that the results they generate are valid and applicable to their desired business outcomes and goals. EDA also assists stakeholders by ensuring that they are asking the right questions. EDA helps answer questions about standard deviations, categorical variables, and confidence intervals. Once EDA is complete and insights are gained, its capabilities can be used for more advanced data analysis or modeling, including machine learning.
- Data Visualization is also a part of EDA to represent categorical data with graphical representation.

Data Preprocessing for prediction

These are some of the important statistic values that describe the data-

- **count** - The number of not-empty values.
- **mean** - The average (mean) value.
- **std** - The standard deviation.
- **min** - The minimum value.
- **25%** - No. of values less than 25% percentile.
- **50%** - No. of values less than 50% percentile.
- **75%** - No. of values less than 75% percentile.
- **max** - The maximum value

Skewness Analysis

- Skewness measures the deviation of a random variable's given distribution from the normal distribution, which is symmetrical on both sides. Skewness Analysis is performed to see whether the numerical features are severely skewed or not and this will help us in creating better linear models.
- If the skewness is between -0.5 & 0.5, the data are nearly symmetrical.
- If the skewness is between -1 & -0.5 (negative skewed) or between 0.5 & 1(positive skewed), the data are slightly skewed.
- If the skewness is lower than -1 (negative skewed) or greater than 1 (positive skewed), the data are extremely skewed.

Data Visualization

Data Visualization is also a part of EDA to represent categorical data with graphical representation.

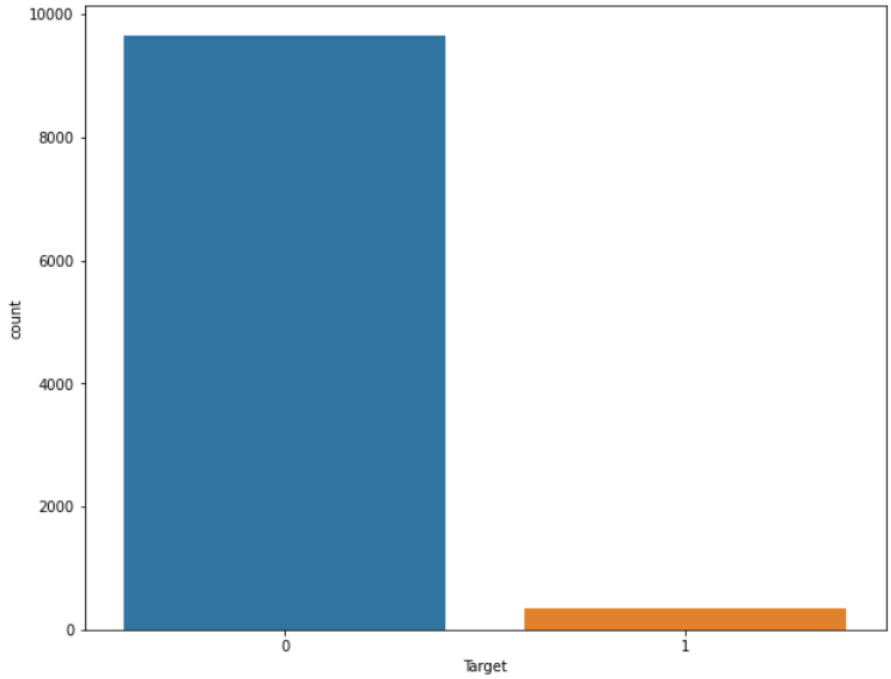

Figure 3: Countplot representing total no of machines not failed(0) and failed(1)

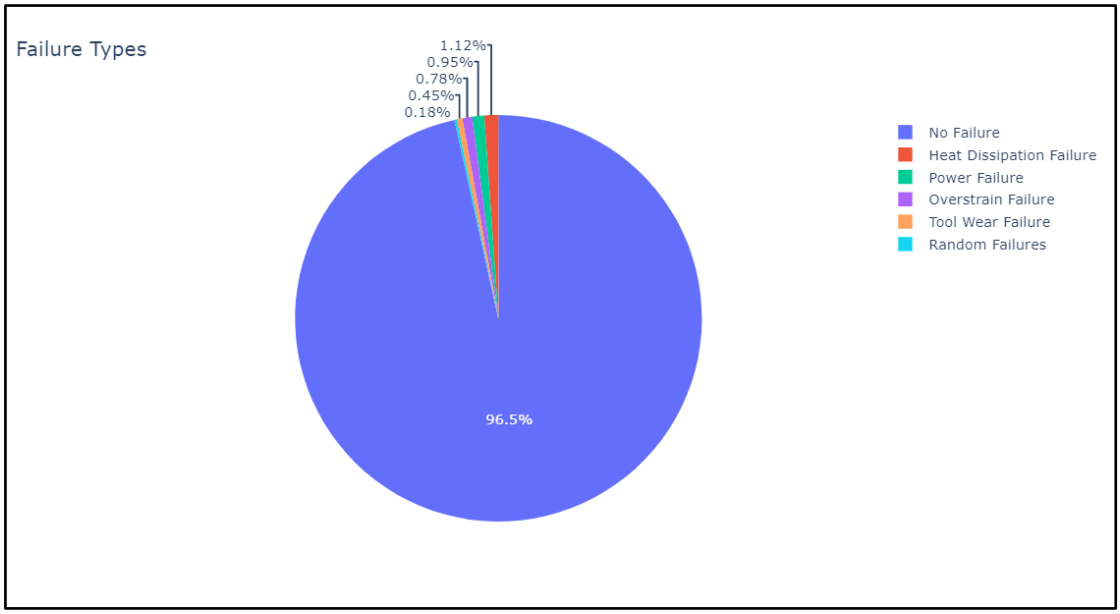

Figure 4: Pie Chart depicting percentage of different failure types of machines

Box Plots

- Box plot are generated to observe the relationship between categorical features with the Target and Failure Type

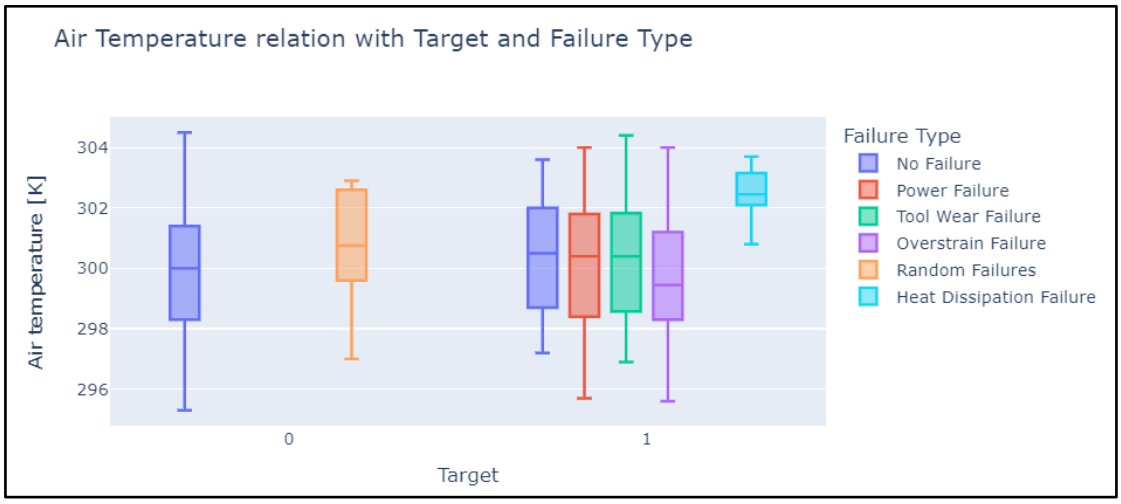

Figure 5: Boxplot representing range values of different failure types vs Air Temperature

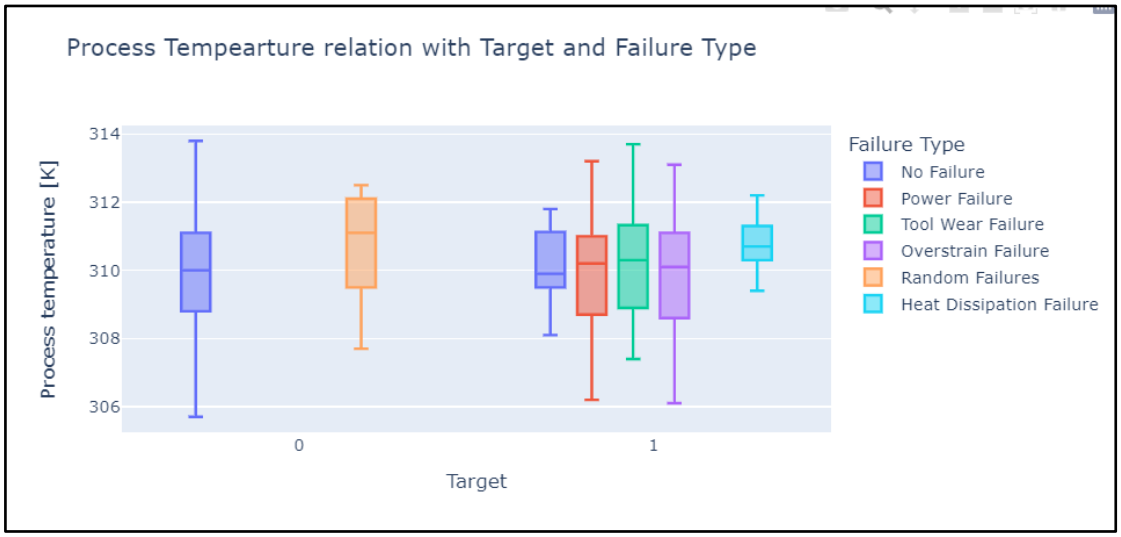

Figure 6: Boxplot representing range values of different failure types vs Process Temperature

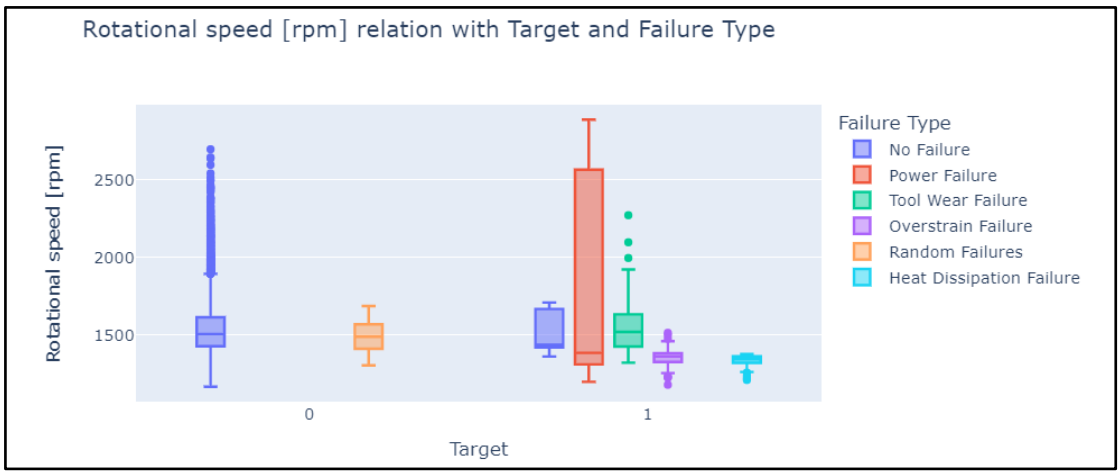

Figure 7: Boxplot representing range values of different failure types vs Rotational speed

Figure 8: Boxplot representing range values of different failure types vs Torque

Figure 9: Boxplot representing range values of different failure types vs Tool wear

Data Preprocessing for Prediction

Before using ML model the data is processed again in 3 steps -

1. Encoding categorical features
2. Splitting test & train data
3. Feature Scaling

Encoding

Label Encoding refers to converting the labels into a numeric form so as to convert them into the machine-readable form. We use it on our training data, and using "fit" it will figure out the unique values and assign a value to it, returns the encoded labels.

Test-Train split

The test-train split is used to estimate the performance of machine learning algorithms using training data and validate with test data.

- Test size: This parameter specifies the size of the testing dataset.
- Random state: Determine the randomness of the splits

Feature Scaling

Feature scaling is a method used to normalize the range of independent variables or features of data and make them closer to each other. Feature scaling is essential for machine learning algorithms that calculate distances between data (Ex- KNN).

Model Development

Model development involves collecting data from multiple trusted sources, process the right data to build the model, choose algorithms to build the model, build the model, calculate performance metrics, and get the best performing model.

The following algorithms are used to develop model to predict failure status of machines:

A. Logistic Regression

- Logistic regression is one of the most popular machine learning algorithms, belonging to supervised learning techniques. It is used to predict the categorical dependent variable using a certain set of independent variables.
- Logistic regression predicts the output of categorical dependent variables. Therefore, the result must be a categorical value or a discrete value. It can be yes or no, 0 or 1, true or false, but instead of specifying the exact values as 0 and 1, specify a probability value in the range 0 to 1.
- Logistic regression is very similar to linear regression except for its usage. Linear regression is used to solve regression problems, and logistic regression is used to solve classification problems.
- In logistic regression, instead of fitting a regression line, we fit an "S" -shaped logistic function that predicts two maximums (0 or 1).
- The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.
- Logistic regression is an important machine learning algorithm because it provides probabilities and can classify new data using continuous and discrete datasets.

Assumptions of Logistic Regression

The dependent variable must be categorical in nature.

The independent variable should not have multicollinearity.

The Logistic regression equation can be obtained from the Linear Regression equation. The mathematical steps to get Logistic Regression equations are given below:

- We know the equation of the straight line can be written as:

$$y= b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

- In Logistic Regression y can be between 0 and 1 only, so for this let's divide the above equation by (1-y):

$$\frac{y}{1-y}$$; 0 for y= 0, and infinity for y=1

- But we need range between -[infinity] to +[infinity], then take logarithm of the equation it will become:

$$log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

The above equation is the final equation for Logistic Regression.

B. K-Nearest Neighbor (KNN)

- K-nearest neighbors (KNN) could be a style of supervised learning rule used for each regression and classification. KNN tries to predict the proper category for the check knowledge by shrewd the gap between the check knowledge and every one the coaching points. Then choose the K variety of points that is closest to the check knowledge.
- The KNN rule calculates the likelihood of the check knowledge happiness to the categories of 'K' coaching knowledge and sophistication holds the very best likelihood are going to be designated. Within the case of regression, the worth is the mean of the 'K' designated coaching points.
- To validate the accuracy of the KNN classification, a confusion matrix is used. Other statistical methods such as the likelihood-ratio test are also used for validation.
- The KNN working can be explained on the basis of the below algorithm:
  - ➢ Select the number K of the neighbors
  - ➢ Calculate the Euclidean distance of K number of neighbors
  - ➢ Take the K nearest neighbors as per the calculated Euclidean distance.

- ➢ Among these k neighbors, count the number of the data points in each category.
- ➢ Assign the new data points to that category for which the number of the neighbor is maximum.
- ➢ Our model is ready.

C. Support Vector Machine (SVM)

- Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, used for both classification and regression problems. However, it is mainly used for classification problems in machine learning.
- The goal of the SVM algorithm is to create the best decision line or boundary that can separate the n-dimensional space into classes so that you can easily place the new data point in the appropriate category in the future. This best decision boundary is called a hyperplane.
- SVM chooses the extreme points / vectors that help create the hyperplane. These extreme cases are called Support Vectors and hence the algorithm is called Support Vector Machine.
- There are two types of SVMs':
  - ➢ **Linear SVM**: Linear SVM is used for linearly separable data, i.e. H if a set of data can be classified into two classes using a single straight line, such data is called linearly separable data and the classifier is called linear SVM classifier.
  - ➢ **Nonlinear SVM**: Nonlinear SVM is used for nonlinearly separated data i.e. H if a set of data cannot be classified using a straight line, that data is called nonlinear data and the classifier used is said to be nonlinear. linear SVM classifier.

D. Decision Tree Classifier

- Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too.

- The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data(training data).
- In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

  Important Terminology related to Decision Trees

  1. **Root Node**: It represents the entire population or sample and this further gets divided into two or more homogeneous sets.
  2. **Splitting**: It is a process of dividing a node into two or more sub-nodes.
  3. **Decision Node**: When a sub-node splits into further sub-nodes, then it is called the decision node.
  4. **Leaf / Terminal Node**: Nodes that do not split are called Leaf or Terminal nodes.
  5. **Pruning**: When we remove sub-nodes of a decision node, this process is called pruning. You can say the opposite process of splitting.
  6. **Branch / Sub-Tre**e: A subsection of the entire tree is called branch or sub-tree.
  7. **Parent and Child Node**: A node, which is divided into sub-nodes is called a parent node of sub-nodes whereas sub-nodes are the child of a parent node.

  The ID3 algorithm builds decision trees using a top-down greedy search approach through the space of possible branches with no backtracking. A greedy algorithm, as the name suggests, always makes the choice that seems to be the best at that moment.

Steps in ID3 algorithm:

1. It begins with the original set S as the root node.
2. On each iteration of the algorithm, it iterates through the very unused attribute of the set S and calculates Entropy(H) and Information gain (IG) of this attribute.
3. It then selects the attribute which has the smallest Entropy or Largest Information gain.
4. The set S is then split by the selected attribute to produce a subset of the data.
5. The algorithm continues to recur on each subset, considering only attributes never selected before.

E. Random Forest Classifier

- Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.
- One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. It performs better results for classification problems.

Features of Random Forest algorithm-

- ➢ It's more accurate than the decision tree algorithm.
- ➢ It provides an effective way of handling missing data.
- ➢ It can produce a reasonable prediction without hyper-parameter tuning.
- ➢ It solves the issue of overfitting in decision trees.
- ➢ In every random forest tree, a subset of features is selected randomly at the node's splitting point.

Working of Random Forest classifier:

Random Forest works in two-phase. First is to create the random forest by combining N decision trees, and second is to make predictions for each tree created in the first phase.

The working process can be explained in the below steps and diagram:

Step-1: Select random K data points from the training set.

Step-2: Build the decision trees associated with the selected data points (Subsets).

Step-3: Choose the number N for decision trees that you want to build.

Step-4: Repeat Step 1 & 2.

Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

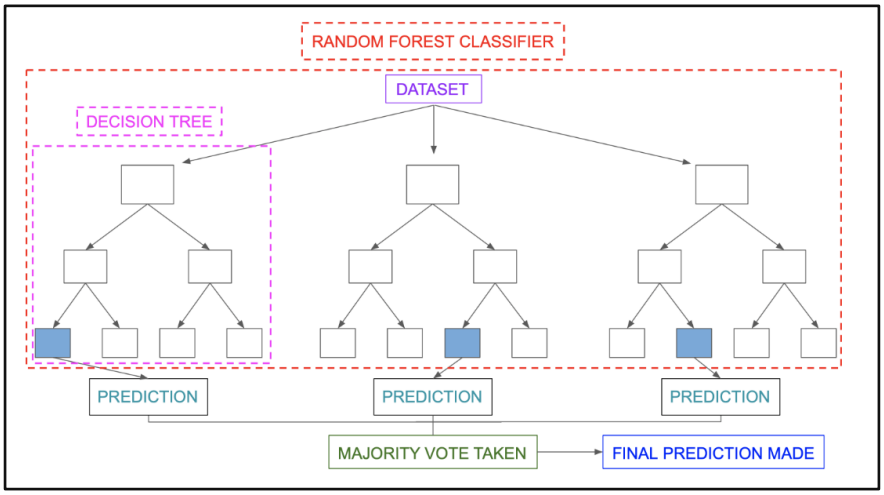

Figure 10: Diagram representing random forest classifier

F. Naïve Bayes Classifier

The Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.

- It is mainly used in text classification that includes a high-dimensional training dataset.
- Naïve Bayes Classifier is one of the simplest and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.
- It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.
- Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.

The Naïve Bayes algorithm is comprised of two words Naïve and Bayes, Which can be described as:

- Naïve: It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features. Such as if the fruit is identified on the basis of color, shape, and taste, then red, spherical, and sweet fruit is recognized as an apple. Hence each feature individually contributes to identify that it is an apple without depending on each other.
- Bayes: It is called Bayes because it depends on the principle of Bayes' Theorem.

Chapter 5

System Analysis

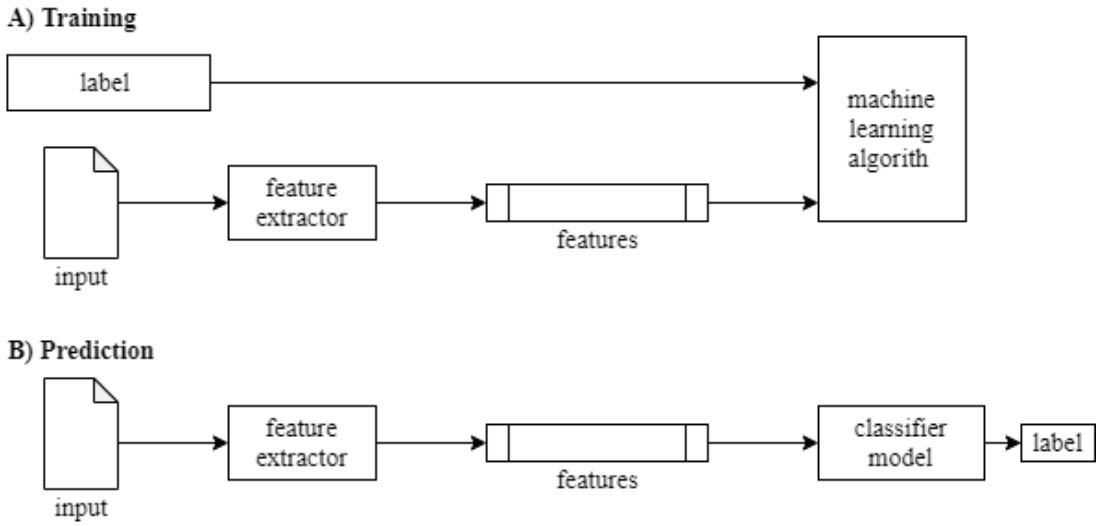

Figure 11: Frameworks for Supervised Learning (Classification and Prediction)

Selecting the appropriate class label for an input is what supervised categorization is all about. Each input is viewed as being fractured from all other inputs in basic classification, and the collection of labels needs to be defined first. There are several appealing variations to the fundamental classification task. For instance, in multi-class classification, each case is taken into consideration with multiple labels; the set of labels is not formally determined in advance, as in the case of open classification, and in sequence classification, the list of inputs are ordered in a mutually supportive manner. Based on a training corpus that has the appropriate label for each input, a supervised classifier is constructed.

In the process of supervised classifier, the feature extractor extracts the various attributes contributing to the working of machines. Then machine learning algorithms classify precisely and allocate the label as 1 for failure and label 0 for non-failure. Once the machine is trained with a huge compendium of data, then it will envisage precisely test data based on their knowledge learnt. During prediction, the feature extractor will be used to find whether the particular machine under consideration will fail (label 1) or will not fail (label 0).

Chapter 6

Software Description

6.1 Python

6.1.1 IDE Used:

1. **VS code-** Visual Studio Code is a lightweight and powerful source code editor that runs on your desktop and works on Windows, macOS, and Linux. Provides built-in support for JavaScript, TypeScript, Node.js, and is eco-friendly with extensions for other languages (C ++, C #, Java, Python, PHP, Go, etc.) and runtimes (.NET, Unity, etc.) It has a system).

   Python extensions to Visual Studio Code that fully support the Python language, including features such as IntelliSense (Pylance), linting, debugging, code navigation, code formatting, and refactoring (all actively supported versions of the language). :> = 3.6), Variable Explorer, TestExplorer, etc. A Jupyter Visual Studio Code extension that provides basic notebook support for the language kernel currently supported by Jupyter Notebook. Many language cores work without change. You may need to change the VS Code language extension to enable advanced features.

2. **Jupyter notebook**- It is an open-source IDE that is used to create Jupyter documents that can be created and shared with live codes. Also, it is a web-based interactive computational environment. The Jupyter notebook can support various languages that are popular in data science such as Python, Julia, Scala, R, etc.

6.1.2 Python Packages Used:

1. **Numpy:** NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

2. **Pandas:** Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.

3. **Matplotlib:** Matplotlib is a cross-platform, data visualization and graphical plotting library for Python and its numerical extension NumPy. As such, it offers a viable open-source alternative to MATLAB. Developers can also use matplotlib's APIs (Application Programming Interfaces) to embed plots in GUI applications.

4. **Scikit-learn:** Scikit-learn is a free machine learning library for Python. It features various algorithms like support vector machine, random forests, and k-neighbors (KNN), and it also supports Python numerical and scientific libraries like NumPy and SciPy.

5. **Seaborn:** Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

Chapter 7

Testing and Results

In this project we trained our data with different machine learning algorithms.
Then these models were tested and the accuracy of these models were calculated using the confusion matrix for each model.

Here is a table that displays the accuracy obtained for each classifier model:

	Models	Accuracy	Runtime (s)
0	RandomForestClassifier	98.367	0.456921
1	DecisionTreeClassifier	97.833	0.055104
2	SVC	97.633	0.725999
3	LogisticRegression	97.367	0.047671
4	KNeighborsClassifier	97.033	0.237892
5	GaussianNB	96.367	0.005000

Figure 12: Accuracy and runtime representation of each model

It was found that the model trained using 'Random Forest Classifier' gave the highest accuracy of about 98.533%.

- **Accuracy** - Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. (Accuracy = TP+TN/TP+FP+FN+TN)
- **Precision** - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. (Precision = TP/TP+FP)
- **Recall** - Recall is the ratio of correctly predicted positive observations to the all observations in actual class. (Recall = TP/TP+FN)
- **f1-score** - f1-score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.
- **Support** - Support is the number of actual occurrences of the class in the specified dataset.

The **Confusion Matrix** is a matrix used to determine the performance of the classification models for a given set of test data. It is used to visualize important predictive analytics like recall, specificity, accuracy, and precision. Confusion matrices are useful because they give direct comparisons of values like True Positives, False Positives, True Negatives and False Negatives.

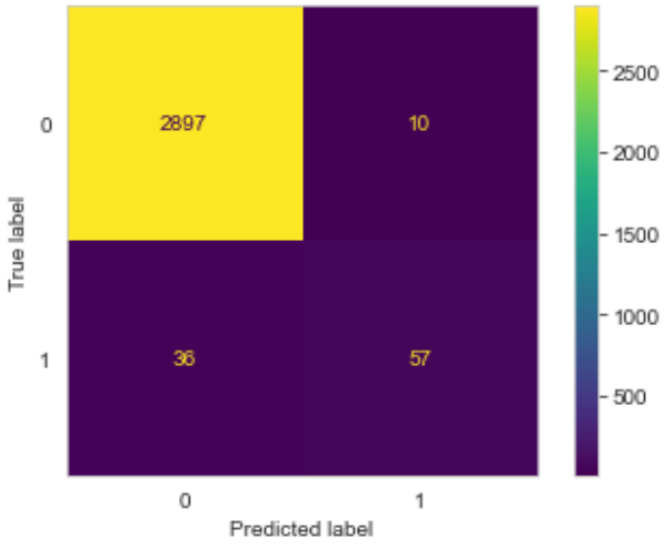

Figure 13: Confusion Matrix of Random Forest Classifier Model

Chapter 8

Advantages, Limitations and Applications

8.1 Advantages:

- Predict the failure type and failure category of machines with a certain level of accuracy.
- Minimal chances of incorrect prediction in any case.
- All the important contributing factors taken in predicting failure.
- The model produces an unbiased output.
- Model predicts the failure well in advance so that accordingly all the arrangements can be made.
- The system excludes human efforts and thus probability of efficiency is enhanced with use of a minimum number of resources.

8.2 Limitations:

- The model must be given correct inputs otherwise the model may produce inaccurate results.

8.3 Applications of project:

- The model implemented can be a great asset to manufacturing industries as these can predict the failure status of machines in use and accordingly plan their activities that too before time.
- This model can be incorporated in industries lacking multiple resources as the system can be used with a minimum number of factors predicting the failure scenario of machines.

Chapter 9

Conclusion and Future Scope

Predicting machine failure category and failure type is very crucial in industries, especially the manufacturing industries as in these the machines are required at each step of the process directly or indirectly depending on their usage and applicability. sIt is of utmost importance that the failure prediction be made accurately with sufficient percentage of efficiency so that the results are considered and considering them ,other plans are made and worked upon.

Therefore, we designed and implemented models based on different machine learning algorithms to predict the failure status of vivid machines incorporating multiple useful factors with the aim of finding the best model of all having the highest efficiency.The goal was to predict failure well in advance so that industries can make productive decisions beforehand enabling enhanced throughput.

The model designed has minimal chances of incorrect prediction in any case as it considers all important contributing factors in predicting failure. In Fact, the model produces an unbiased output and excludes human efforts. Therefore, the probability of efficiency is enhanced with use of a minimum number of resources.
Some key future scopes:

1. More factors can be incorporated for prediction so as to enhance accuracy of the model.
2. Prediction of estimated time of failure will be a major modification
3. The system can be merged for multiple access across different refineries and also within departments at any future time.
4. Model may be embedded into the SAP system so as to view real time data and do modifications accordingly.

References

[1] https://www.softwebsolutions.com/resources/predictive-maintenance-for-oil-and-gas-industry.html

[2] https://www.semanticscholar.org/paper/Predictive-Maintenance-Using-Machine-Learning-in-Suursalu/b55ee5ed0e09e1749a7da49e0a28eec487baebf4

[3] https://databricks.com/session_na20/efficiently-building-machine-learning-models-for-predictive-maintenance-in-the-oil-gas-industry-with-databricks

[4] https://www.offshore-technology.com/comment/predictive-maintenance-oil-gas/

[5] https://ieeexplore.ieee.org/document/8848208

[6] https://www.semanticscholar.org/paper/Scalability-Analysis-of-Predictive-Maintenance-in-Refineries/1dfefd95652a6a6bf8db0cb29e9c2ae3cbf78423

[7] https://ieeexplore.ieee.org/document/7013630

[8] https://ieeexplore.ieee.org/document/9143895

[9] https://ieeexplore.ieee.org/document/9676783

Appendix A: Data Sheets

Machine Predictive Maintenance Classification Dataset

Since real predictive maintenance datasets are generally difficult to obtain and in particular difficult to publish, we present and provide a synthetic dataset that reflects real predictive maintenance encountered in the industry to the best of our knowledge.

The dataset consists of 10 000 data points stored as rows with 14 features in columns

- UID: unique identifier ranging from 1 to 10000
- productID: consisting of a letter L, M, or H for low (50% of all products), medium (30%), and high (20%) as product quality variants and a variant-specific serial number
- air temperature [K]: generated using a random walk process later normalized to a standard deviation of 2 K around 300 K
- process temperature [K]: generated using a random walk process normalized to a standard deviation of 1 K, added to the air temperature plus 10 K.
- rotational speed [rpm]: calculated from power of 2860 W, overlaid with a normally distributed noise
- torque [Nm]: torque values are normally distributed around 40 Nm with an Ïf = 10 Nm and no negative values.
- tool wear [min]: The quality variants H/M/L add 5/3/2 minutes of tool wear to the used tool in the process.
- “Machine failure” label that indicates whether the machine has failed in this particular data point for any of the following failure modes is true.