

**Nanyang Business School**

**BC2406 Analytics I: Visual and Predictive Analytics**

**Prepared by:**

**Seminar Group: 3**

**Group Number: 5**

|  |  |
| --- | --- |
| **Name** | **Matriculation** |
| Dong Zhijian | U2222844K |
| Koh Jia Hui Rachel | U2222747H |
| Ang Khay Yee | U2210538H |
| Andy Chan Yan Meng | U2221216B |
| Zoey Chew Jie Ying | U2210369K |
| Muhammad Wisnu Darmawan | U2220618B |

## 

[1. Executive Summary 3](#_Toc150154105)

[2. Introduction 4](#_Toc150154106)

[2.1 Aramco’s Current Methodology 4](#_Toc150154107)

[2.2 Limitations of Current Methodology 5](#_Toc150154108)

[3. Opportunity Statement 6](#_Toc150154109)

[4. Data Description & Selection 6](#_Toc150154110)

[4.1 Key Assumptions 6](#_Toc150154111)

[5. Exploratory Data Analysis (EDA) 7](#_Toc150154112)

[6. Proposed Methodology using Machine Learning 10](#_Toc150154113)

[6.1 Brief Overview 10](#_Toc150154114)

[6.1.1 Dependent Variable 10](#_Toc150154115)

[6.1.2 Independent Variables 10](#_Toc150154116)

[6.2 Data Preparation 11](#_Toc150154117)

[6.2.1 Data Cleaning 11](#_Toc150154118)

[6.2.2 Addressing Outliers 11](#_Toc150154119)

[6.2.3 Addressing Class Imbalance via Oversampling 12](#_Toc150154120)

[6.2.4 Selection of Accuracy Metrics 12](#_Toc150154121)

[6.3 Machine Failure Analysis 13](#_Toc150154122)

[6.3.1 First Model Type - Logistic Regression Model 13](#_Toc150154123)

[6.3.2 Second Model Type - Classification and Regression Tree Model 14](#_Toc150154124)

[6.3.3 Model Results Analysis 16](#_Toc150154125)

[6.4 Machine Failure (Type) Analysis 17](#_Toc150154126)

[6.4.1 First and Second Models - Logistic Regression and CART Model 18](#_Toc150154127)

[6.4.2 Third Model Type: K-Nearest Neighbors (kNN) Model 19](#_Toc150154128)

[6.4.3 Model Results Analysis 20](#_Toc150154129)

[7. Business Value of Proposed Methodology 20](#_Toc150154130)

[8. Recommendations 21](#_Toc150154131)

[8.1 Short Term Approach 21](#_Toc150154132)

[8.2 Long Term Approach 21](#_Toc150154133)

[9. Limitations of recommended solutions 22](#_Toc150154134)

[9.1 Limited Independent Variables 22](#_Toc150154135)

[9.3 Insufficient Observations of Machine Failure 22](#_Toc150154136)

[9.4 Complexity of CART Models 23](#_Toc150154137)

[10. Conclusion 23](#_Toc150154138)

[References 24](#_Toc150154139)

[Appendix 26](#_Toc150154140)

[Appendix A: Datasets 26](#_Toc150154141)

[Appendix B: Logistic Regression 27](#_Toc150154142)

[Appendix C: CART 29](#_Toc150154143)

[Appendix D: kNN 35](#_Toc150154144)

## 

## 1. Executive Summary

Our proposal aims to enhance Aramco’s equipment maintenance systems through predictive maintenance. Aramco faces maintenance challenges due to extreme environmental conditions and offshore operations. To address these challenges, Aramco implemented Preventive Maintenance Optimization (PMO). PMO optimises maintenance schedules based on factors including equipment criticality, legal requirements, and manufacturer recommendations. A set of questions were also introduced to guide the optimization process.

While this approach optimises maintenance schedules, it has its limitations. New machines lack historical data, making it harder to optimise the maintenance of new machines. Human judgement in the PMO process also results in bias and inconsistencies. Insufficient data and lapses in human judgement can lead to over- or under-optimization of maintenance tasks. Lastly, PMO requires periodic monitoring and adjustments to ensure effectiveness.

In the initial exploratory data analysis of the dataset, missing values and anomalies were detected. A significant imbalance exists between functioning and failed machines, potentially affecting predictive model accuracy. Weak correlations between 'Machine failure' and numerical parameters necessitate further scrutiny. The distribution of failure types also gives direction for a more focused analysis of these failure modes.

Given the highly imbalanced data, we employed oversampling to balance the minority class. In our comprehensive evaluation of predictive models for machine failure classification, we analysed two machine learning algorithms: Logistic Regression and Classification and Regression Tree (CART). The purpose was to identify the most effective model for deployment in a predictive maintenance system that can accurately predict machine failures. Our best model turned out to be the CART model trained on “balanced” data, achieving an F1-Score of 95.52%.

We also extended our analysis to identify a model to predict and distinguish between four types of machine failures, by including a third model: k-Nearest Neighbors (kNN). The kNN model emerged as the superior option, exhibiting high precision and recall, with the latter being crucial for avoiding costly maintenance oversights.

We recommend a phased maintenance strategy upgrade for Aramco: initially, a CART model for prompt, data-driven maintenance decisions, followed by a long-term shift to a kNN and IoT-based predictive system to inform maintenance and smarter procurement of durable equipment through data insights.

To address limitations associated with our recommended predictive system, we suggest a pilot IoT sensor program to better analyse equipment failures, controlled experiments, and digital twin technology to clarify cause-effect relationships. Expanding the dataset through industry collaboration will improve model accuracy. Simplifying the CART model with an intuitive digital form to facilitate ease of use by maintenance staff.

## 2. Introduction

Aramco has one of the world’s most extensive energy infrastructures and operates under some of the highest temperatures on Earth. At high temperatures, oil in pipelines and plants becomes more volatile (Wallace, 2020). This presents a major maintenance challenge for Aramco due to complications from material degradation, thermal stress, and overheating. Offshore oil and gas platforms also pose significant challenges for engineers. The natural surroundings of offshore platforms provide a permanently corrosive atmosphere which makes equipment more prone to degradation (Sulzer, 2017). This means more frequent maintenance is needed. Additionally, the limited space and remote location of offshore platforms complicate maintenance procedures. To mitigate these risks, Aramco has implemented rigorous maintenance practices to ensure the safety of its personnel and the efficacy of its equipment.

### 2.1 Aramco’s Current Methodology

Aramco has taken a step beyond preventive maintenance (PM) by implementing Preventive Maintenance Optimization (PMO) to enhance its maintenance practices. The definitions of PM and PMO by Aramcio are as follows:

*“Preventive Maintenance (PM): Serving equipment or a component of equipment at regular intervals regardless of condition for the purpose of bringing the equipment back to a healthy condition before they fail.*

*Preventive Maintenance Optimization (PMO): Application of a systematic approach to review, evaluate and analyse preventive maintenance tasks to make preventive maintenance more cost effective and reliable.*” (Al-Sultan, 2009, p3-4).

Aramco’s optimisation process considered equipment criticality, legal requirements, PM cycle requirements, PM required shutdown, PM duration and PM requirements recommended by the manufacturer. Aramco’s PMO were guided by the following questions:

“*1. Is it based on equipment failure or history analysis?*

*2. Is it based on the equipment’s known failure rate?*

*3. What is my best judgement on this PM?*

*4. Any duplications with other PMs?*

*5. Does the equipment exist? Or has it been removed or upgraded?*

*6. PM tasks for mothballed equipment (equipment kept in working order but not in use)?*

*7. Has the PM task already been performed?*

*8. Is the PM task covered by other means?*” (Al-Sultan, 2009, p8).

These questions guided Aramco in identifying equipment not covered by the PM program, equipment with irrelevant PM and duplication of equipment lubrication. Aramco also found that there were missing critical checks in the PM task sheet and a high PM frequency on non-critical equipment. In response to its findings, the company added missing equipment into its PM schedule, removed non value-adding PM activities, adjusted PM frequencies, and optimised its labour hours. New technologies for PM were also identified to enhance electrician safety, reduce PM hours, and increase accuracy. The PMO program improved Aramco’s PM process by optimising PM schedules and enhancing equipment reliability and availability.

### 2.2 Limitations of Current Methodology

Aramco’s PMO methods require historical data and performance records to determine the machines’ “known failure rate” and conduct “history analysis” as indicated by the guiding questions. However, there is no past performance data available for newer machines since the machines have not been in operation long enough to accumulate failure data. As a result, Aramco would not be able to fully utilise its PMO framework for its newer machines. Instead, the company would have to rely on other sources such as manufacturer specifications or industry knowledge to establish maintenance plans for newer machines.

Next, Aramco’s framework for PMO relies on human judgement to evaluate and optimise machine maintenance. Using human judgement can introduce biases and errors into the optimisation process. The judgement of maintenance professionals can be influenced by their personal experiences and understanding of the equipment. This leads to biases such as confirmation bias where professionals favour information that agree with their pre-existing assessments. Such biases may lead to inaccurate maintenance assessments as professionals may not be able to assess the situation objectively. Moreover, maintenance assessments made by professionals may vary based on their experience. What one engineer considers high risk, another perceives as low risk. This leads to inconsistencies in the PMO process.

Insufficient data on the machine’s personal records and lapses in human judgement can lead to inaccuracy of the optimisation process. If the process is under-optimised, components would be replaced before they reach the end of their useful life, resulting in resource wastage. Moreover, frequently replacing components can introduce risks such as installation errors, mismatched parts and defects in new components (Mobbley, 2002). Such errors are expensive and lead to inefficient use of resources. On the other hand, over-optimisation of the PMO process results in the elimination of critical maintenance tasks, making machines more prone to failures.

Lastly, implementing PMO also requires periodical monitoring and adjustments to ensure effectiveness as equipment and conditions change. At Aramco, PMO processes must be reviewed every 3 years to align the company’s maintenance strategy with evolving equipment needs and government policies (Al-Sultan, 2009). The process of implementing each PMO can be time-consuming and costly.

## 

## 3. Opportunity Statement

Saudi Aramco, a global oil leader, confronts significant challenges with operational disruptions causing an average financial loss of $149 million annually per facility (Siemens, 2023). These disruptions, averaging 20 unscheduled stoppages per month, underscore the broader implications for the global supply chain, where minor issues in Aramco’s operations can lead to extensive setbacks.

Despite Aramco’s adherence to PMO, the strategy is inefficient due to the lack of historical data for new machines, human bias/involvement of human judgement, potential inaccuracies due to under- or over-optimization and compulsory periodical triennial reviews (Al-Hamad, 2011).

The 2007 Haradh-Othmaniya gas pipeline failure disaster, resulting in 28 fatalities, demonstrates the critical need for enhanced safety and maintenance protocols for equipment in the high-risk oil and gas sector, to prevent such catastrophic failures and protect lives (Webb & Karam, 2007).

Keeping this in mind, our group wants to develop a predictive model to **accurately anticipate equipment failures** in order to **implement predictive maintenance and optimise operational safety and efficiency,** while **identifying the significant factors** that contribute to a machine failure.

## 4. Data Description & Selection

The Predictive Maintenance Dataset (AI4I 2020) from Kaggle was chosen to predict Aramco’s machine failures. This source is a synthetic dataset meticulously modelled after an actual milling machine. The dataset draws its inspiration from real-world scenarios and captures the interplay of various parameters leading to machine failure and its associated machine failure type. The data contains temperature variables that are significant in Aramco’s operations which run under high temperatures and have high variability. The dataset also contains sufficient observations and relevant variables that can be analyzed to form predictions using logistic regression and CART models.

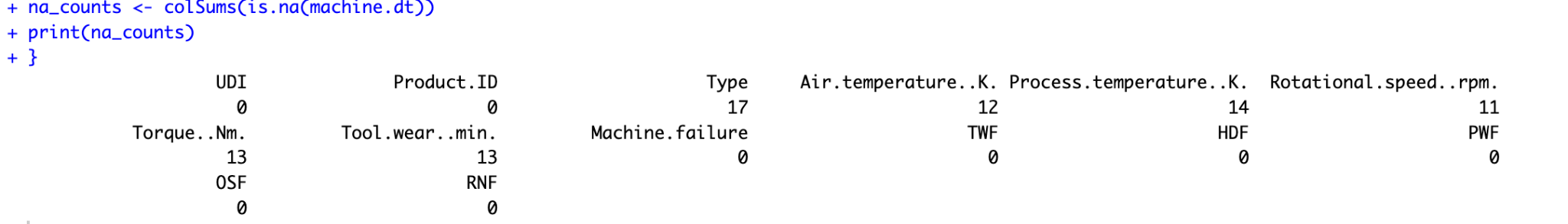
With a collection of 10,000 data entries, each data point in the dataset corresponds to distinct operational conditions and the resultant state of the machine. Within the dataset, 6 variables could be used to predict *‘Machine failure’*. The independent variables are ***‘***Air Temperature [K]’, ‘Process Temperature [K]’, ‘Rotational Speed [rpm]’, ‘Torque [Nm]’, ‘Tool Wear [min]’ and ‘Type’.

### 4.1 Key Assumptions

Although the data was not based on Aramco’s equipment, we assume that this dataset could be an indicator in predicting Aramco’s equipment failures.

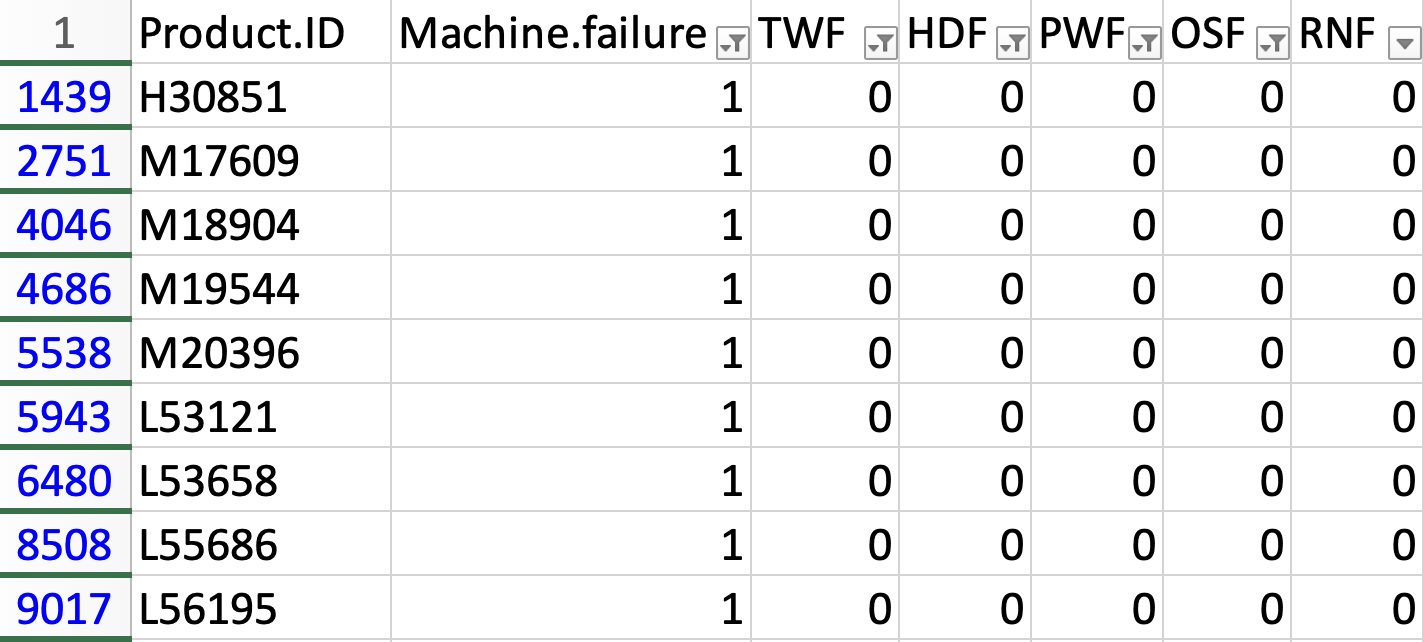
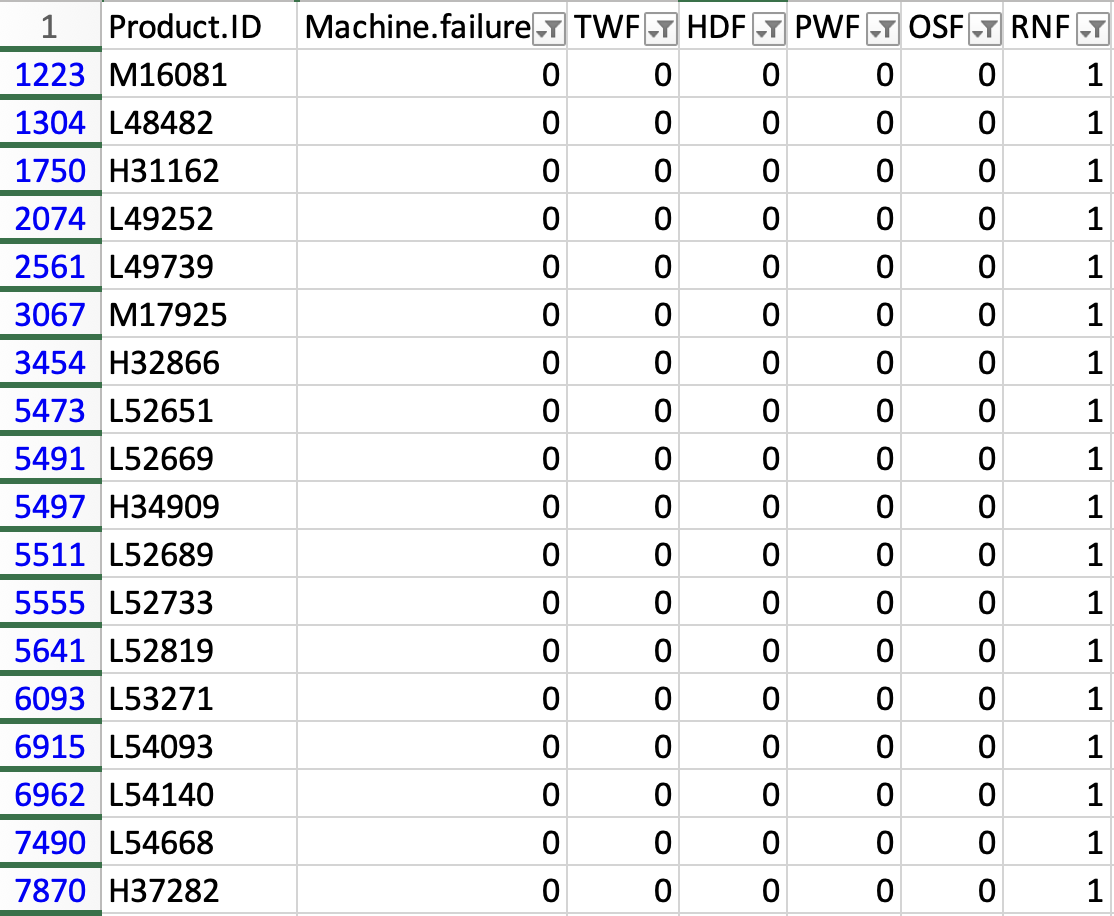
## 

## 5. Exploratory Data Analysis (EDA)



*Figure 5.1: Missing values in some of the variables*

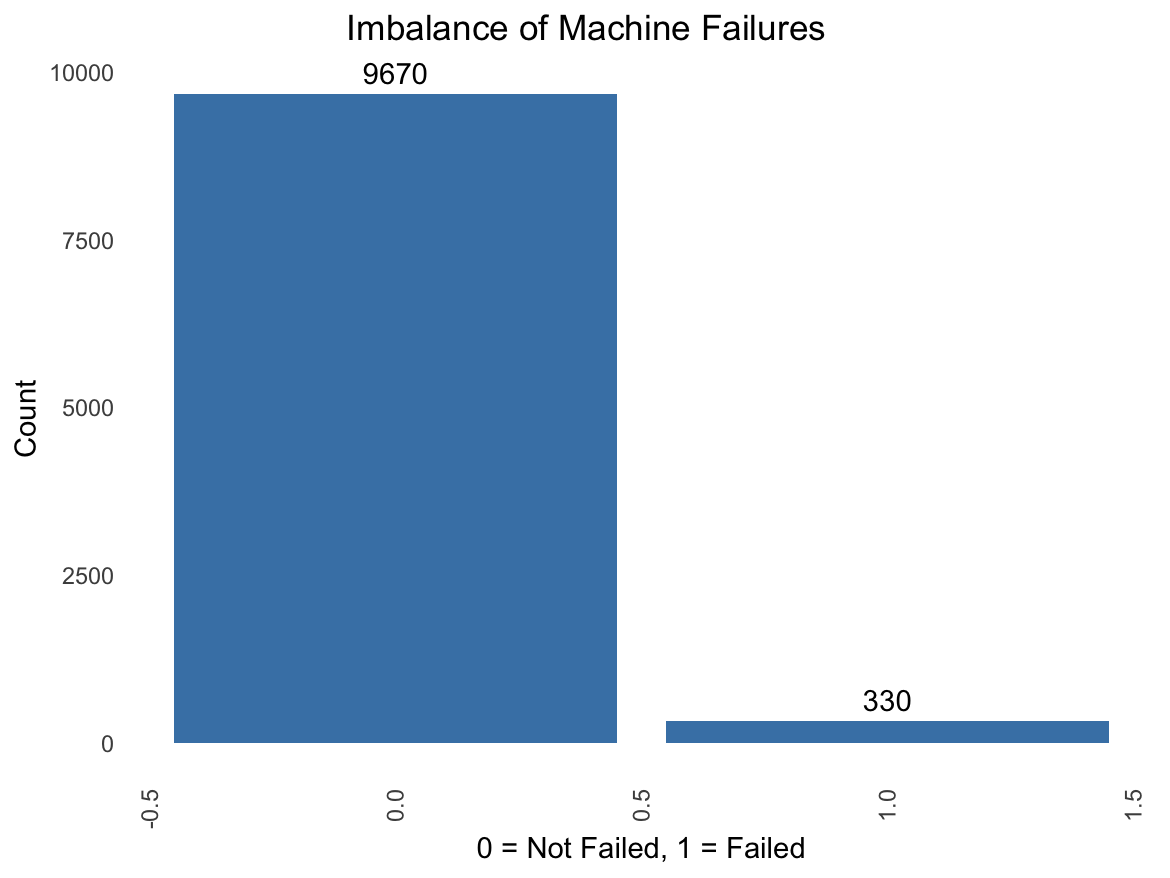
To begin the EDA of the dataset, we first checked the existence of missing values. From Figure 5.1, we identified missing values for variables ‘Type’, ‘Air Temperature’, ‘Process Temperature’, ‘Rotational speed’, ‘Torque’ and ‘Tool Wear’. We discuss the handling of these missing values in Section 6.2.1.



*Figure 5.2: Anomalous data points*

Furthermore, upon exploring the current dataset, we identified 9 data points where the rows have a 'Machine failure' value of 1, yet they are not categorised under any of the 5 failure types depicted in Figure 8. There are also another 18 data points where `Machine failure` is 0 but the RNF value is 1. We will also discuss how to handle these anomalies in Section 6.2.1.

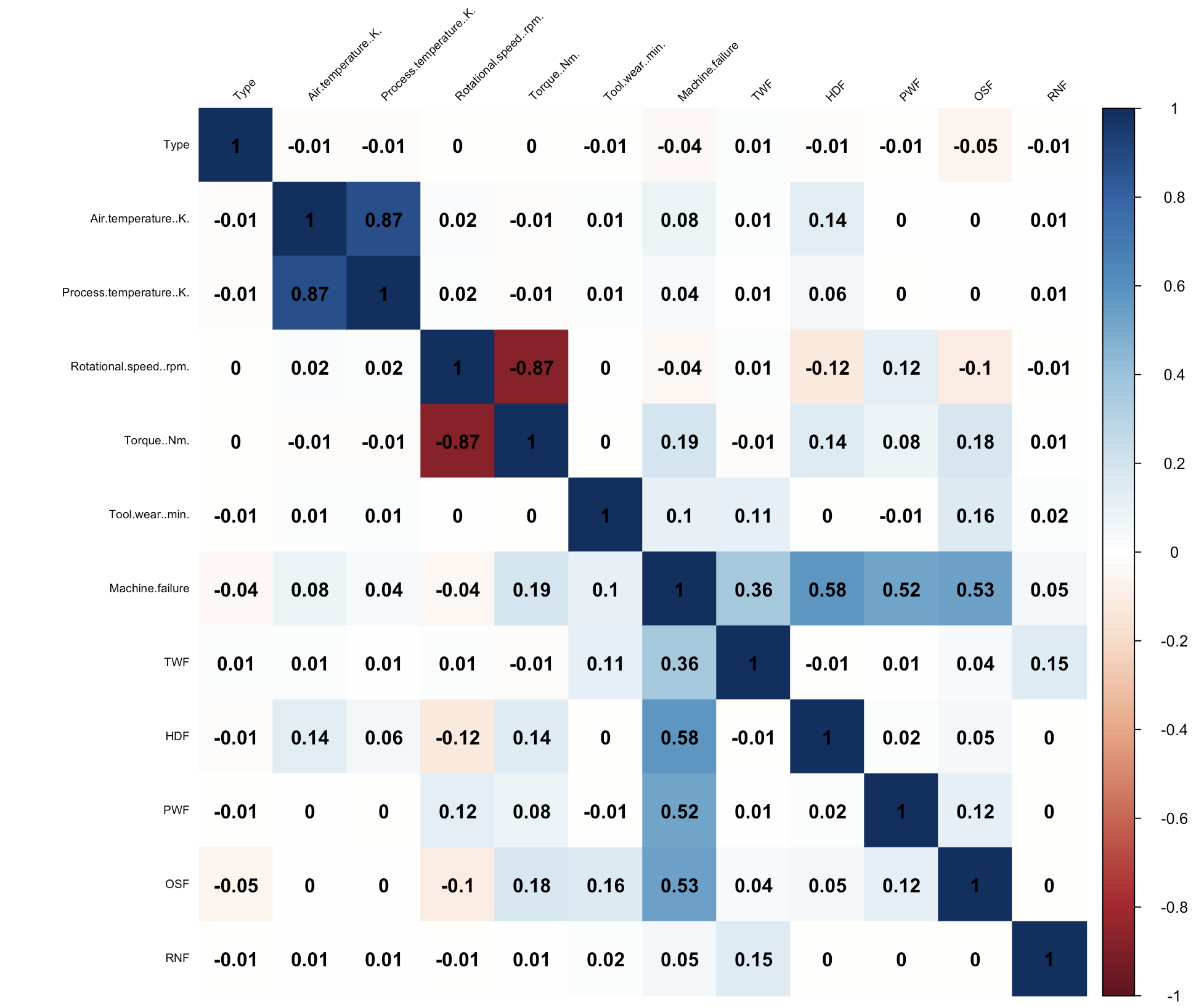
Before performing further data visualisations, our team rectified these discrepancies.



*Figure 5.3: Imbalance of the Dataset (Machine failure)*

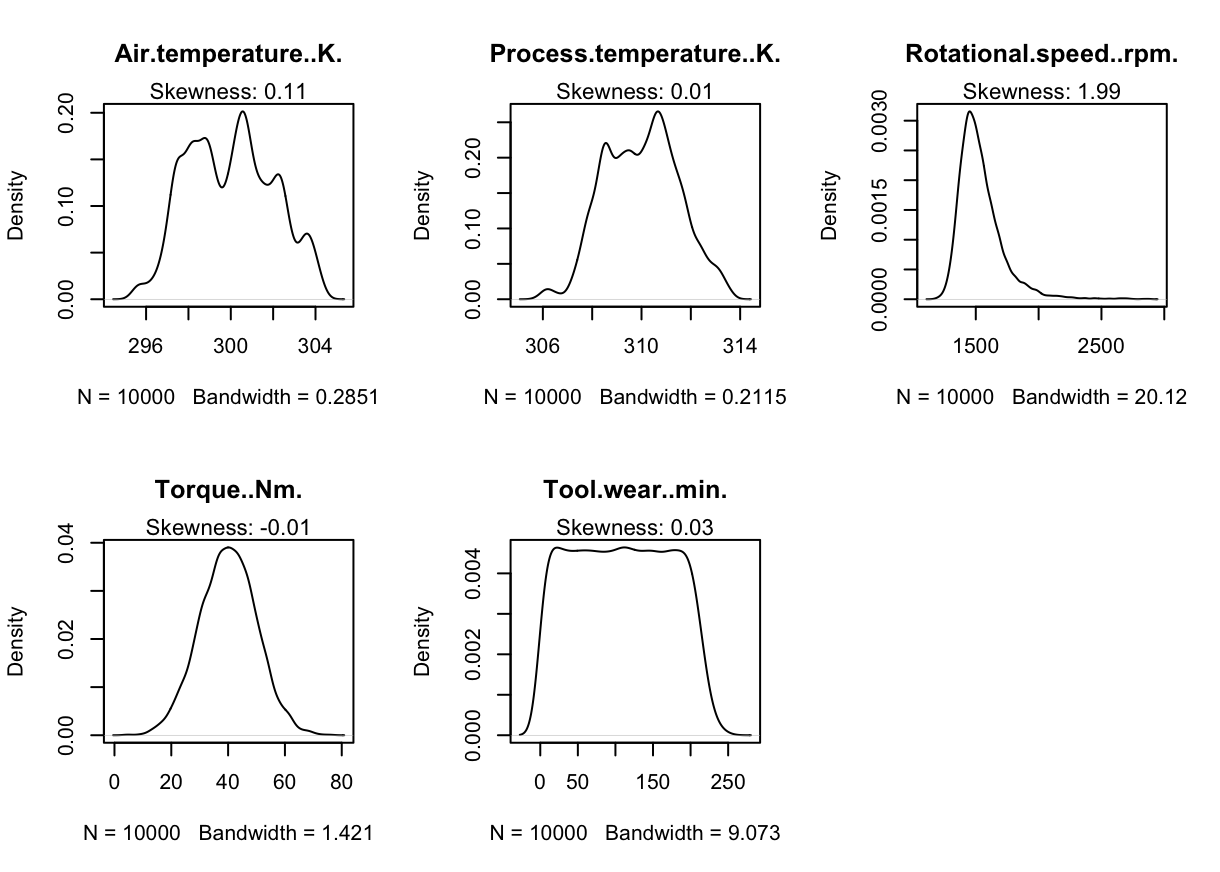
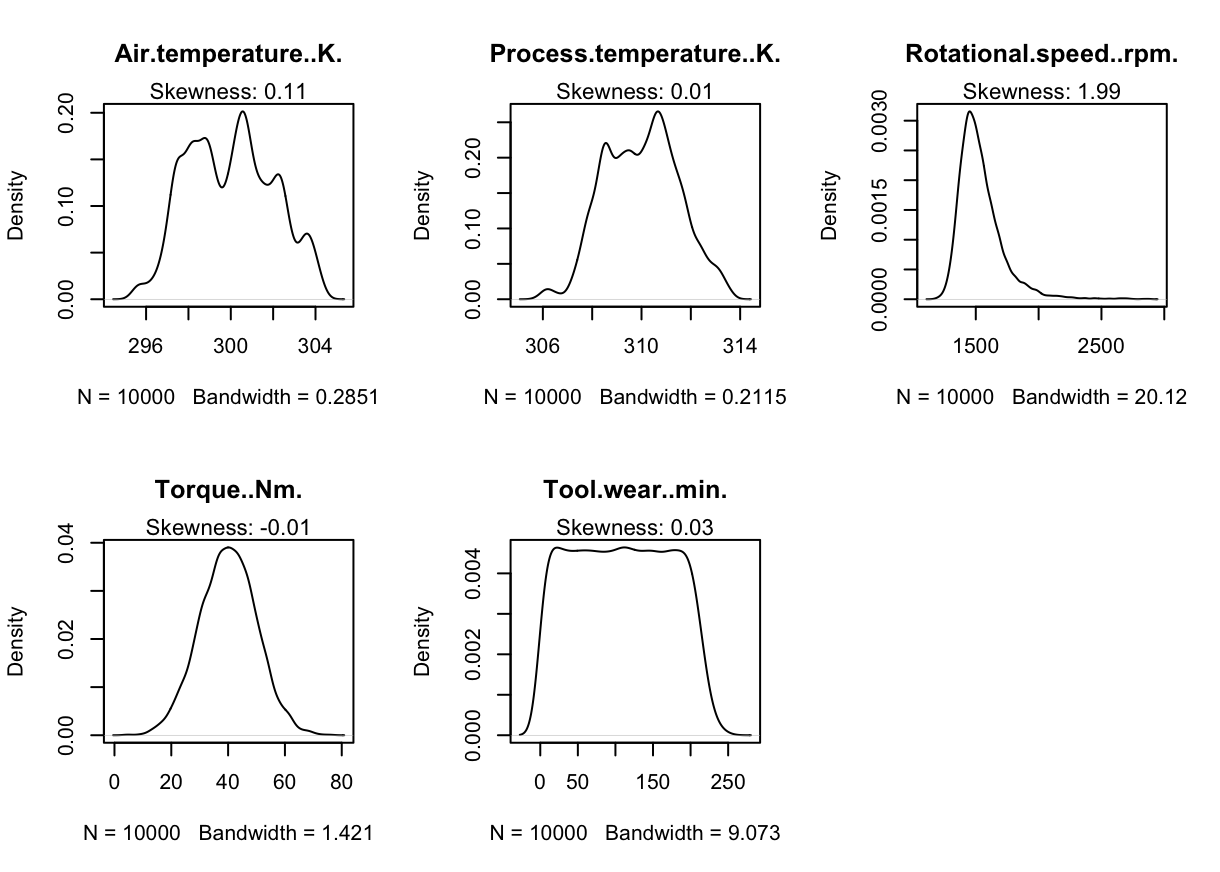
The 'Machine failure' variable in the dataset of 10,000 instances indicates an imbalance with 9,670 machines functioning normally and only 330 experiencing failures. While such an imbalance reflects real-world machine performance, it poses challenges to the accuracy of predictive modelling.

The dataset also provides 5 numeric variables that consist of measurement of different parameters that might explain the reason why a certain machine failed. These 5 variables are Air Temperature [K], Process Temperature [K], Rotational speed [rpm], Torque [Nm] and Tool Wear [min]. Our team decided to do some explorations focusing on these 5 variables.



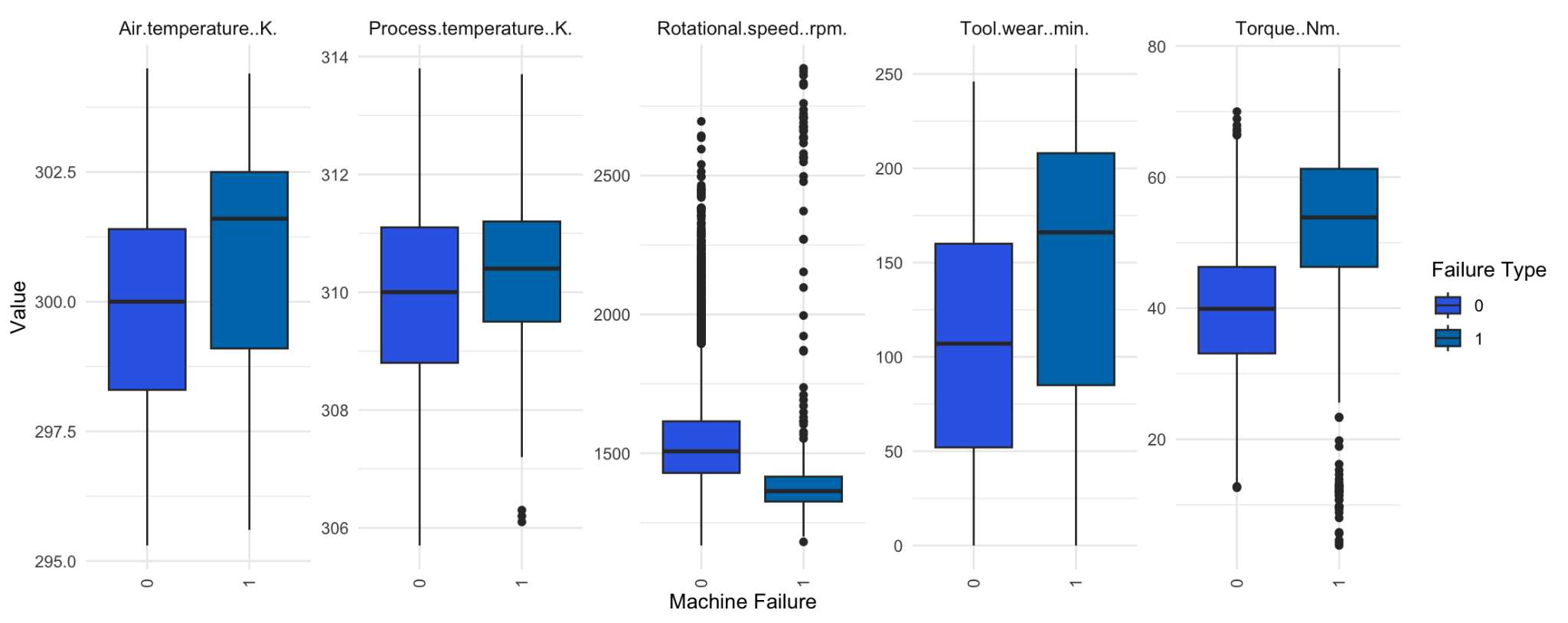
*Figure 5.4: Correlation Plot of the dataset*

Figure 5.4 indicates that the five numerical parameters generally have a weak correlation with 'Machine failure.' 'Torque' has the highest correlation at only 0.19, followed by 'Rotational Speed' at 0.10, with the remaining parameters exhibiting even weaker correlations of less than 0.1. This suggests a need for further analysis to understand how variations in these parameters influence the likelihood of machine failure.



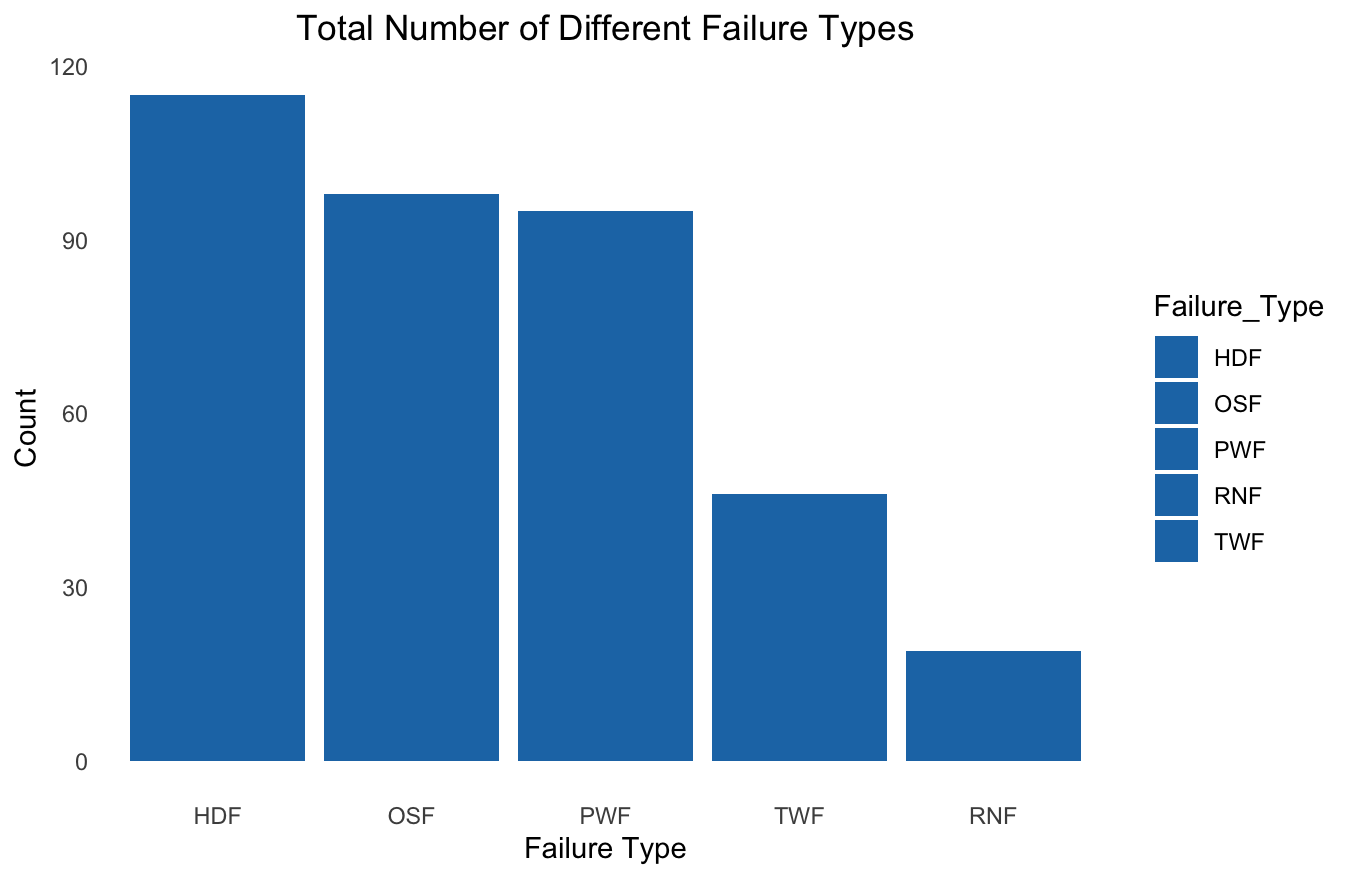
*Figure 5.5: Density Plot of the 5 numeric variables*

Figure 5.5 shows that "Rotational speed" is highly right-skewed, with a skewness value of 1.99.



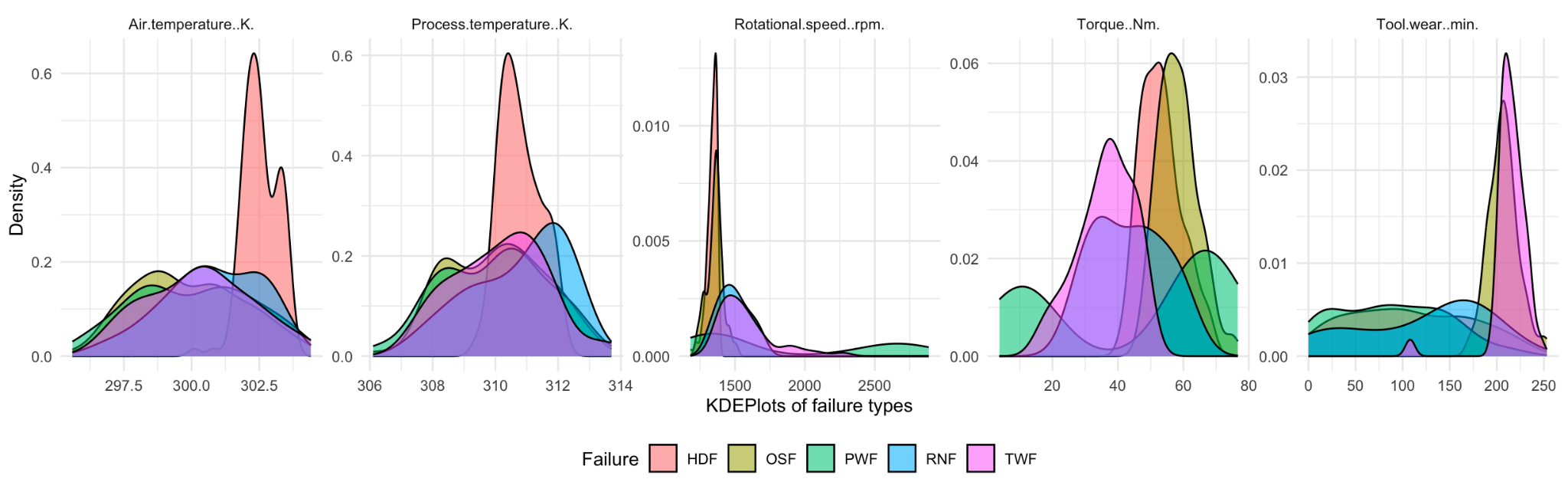
*Figure 5.6: Box plots of the 5 numeric variables for the normal vs failed machines*

Figure 5.6 presents box plots for five numerical variables, comparing normally functioning machines (left) with failed ones (right). It is apparent that median values for "Air Temperature," "Process Temperature," "Torque," and "Tool Wear" are higher in failed machines, while "Rotational Speed" exhibits a lower median value. These variations could indicate factors contributing to machine failures. Additionally, the figure reveals numerous outliers for "Rotational Speed" at the higher end of the distribution, and for "Torque," there are many outliers below the lower limit in failed machines.



*Figure 5.7: Total Number of Different Failure Types*

Figure 5.7 shows the distribution of the 5 different failure types. The five failure types consist of Heat Dissipation Failure (HDF), Over-speed Failure (OSF), Power Failure (PWF), Tool Wear Failure (TWF), and Random Failure (RNF). It is evident that HDF, OSF, and PWF are the top 3 failure types in terms of number of occurrences, and this partly explains why these 3 variables have the highest absolute correlation value with “Machine failure” in Figure 5.4. For a more detailed explanation of these failure types, please refer to section 6.4.



*Figure 5.8: Faceted Kernel Density Estimate (KDE) Plots of the 5 numeric variables*

From Figure 5.8, we can deduce that higher-than-usual "Air temperature" and "Process temperature" are likely to cause the machine to experience HDF. A higher-than-normal "Tool wear" also suggests that the machine would likely encounter TWF. Conversely, a lower-than-usual "Rotational speed" is likely to lead to HDF and/or OSF.

## 6. Proposed Methodology using Machine Learning

Machine learning (ML) is an artificial intelligence technique where computers are trained to derive conclusions from large sets of data. ML plays a critical role in the oil and gas industry by helping to extract and analyse the enormous data generated (Bitstrapped, 2023). The extensivity and volatility of these data makes it hard for manual analysis by data scientists. This is where ML helps to process the data to make inferences and decisions for the company. ML can also carry out predictive analytics based on historical data to make predictions. This is useful in the area of maintenance as ML can predict possible failures in the equipment, and prompt maintenance of the system. This reduces the occurrence of reactive maintenance and prevents unnecessary cost, safety hazards and environmental pollution.

### 6.1 Brief Overview

Our main deliverable is to build machine learning models meticulously designed and trained to enhance ARAMCO's operational efficiency through the predictive analysis of machine failure. The model will leverage quantitative data metrics, such as the type of machine and tool wear time, to predict equipment malfunctions as well as the type of equipment malfunctions. This system will possess the capability to ingest new, real-time data points from any of the company's machines and yield a reliable forecast of potential failures, thereby allowing maintenance teams to act proactively.

#### 6.1.1 Dependent Variable

For the dependent variable, or “focus” of our quantitative research into Aramco’s operational efficiency, we have chosen ‘Machine Failure’ as it serves as a vital indicator of the equipment’s operational status at any given data point. An observation with ‘Machine Failure’ marked as 1 would signify a breakdown in the equipment due to one or more failure modes being true.

#### 6.1.2 Independent Variables

For the independent variables, we obtained a list of operational metrics from our dataset, which we believe would provide essential insights into the factors that might influence machine failures. Our list of independent variables are as follows:

| **Variable Name** | **Description** | **Rationale** |
| --- | --- | --- |
| Type | A letter L, M, or H for low, medium and high product quality variants respectively. | Different product qualities might have varying material compositions, densities, and properties which may affect the wear and tear of a machine differently. |
| Air Temperature [K] | The temperature of the ambient air surrounding the equipment or machinery. | Extreme temperatures can influence a machine's efficiency, cooling mechanisms, and lead to machine failure (Zahid, 2021). |
| Process Temperature [K] | The temperature within a specific process or operation. | If the process temperature veers too far from the optimal range, it can lead to issues like overheating and excessive wear. |
| Rotational speed [rpm] | The number of complete rotations the machine makes in one minute. | If the machine operates at abnormal speeds, it can result in premature failure. |
| Torque [Nm] | The rotational force applied on the machine. | Abnormal torque levels can lead to undue stress, causing parts to wear out prematurely. |
| Tool Wear [min] | The gradual failure of tools due to regular operation. The quality variants H/M/L add 5/3/2 minutes of tool wear to the used tool in the process. | As tools wear down, their performance can degrade, leading to subpar product quality or even damaging other parts of the machine. |

*Table 6.1: Interpretation of Dependent Variables*

### 6.2 Data Preparation

#### 6.2.1 Data Cleaning

To address the missing values found in EDA Section 5, we chose to impute the continuous variables with their respective medians and fill the categorical variables with the mode. This strategy was chosen in light of the presence of outliers within our dataset, which we elected to retain, as outlined in Section 6.2.2.

We also observed anomalies in the dataset where ‘Machine Failure’ was recorded as ‘1’, but all 5 failure modes were marked as ‘0’ (refer to EDA Section 5). Since these records did not fall under any particular failure category, we decided to correct the value of ‘Machine Failure’ to 0. This ensures that the dataset remains consistent and does not give rise to any false positives.

Moreover, we further identified an inconsistency when 'Machine Failure' was recorded as '0', yet RNF was labelled as '1'. Given that RNF is a category signifying the occurrence of Random Failures (found in 18 rows out of 10,000), we opted to drop these data points. Removing these will not substantially impact our analysis since the remaining data count remains notably close to the original dataset (9,982 rows vs 10,000 rows). Due to the unpredictable nature of these random failures, predicting such events is challenging. Consequently, we chose to remove these 18 rows from our analysis.

#### 6.2.2 Addressing Outliers

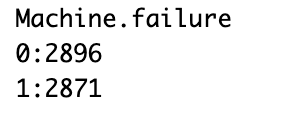
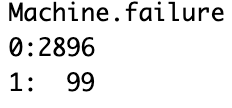
From the dataset, we have identified some variables where outliers exist. For instance, in Figure 5.5, the "Rotational speed" variable is significantly right skewed, displaying numerous data points that deviate from the central cluster. Similarly, according to Figure 5.6, the “Torque” variable displays outliers at both tails.

Our team chose to retain these outliers within our dataset due to the nature of predictive maintenance, where it is likely that these outliers could be the signals of impending machine failures. In this context, our outliers are not statistical anomalies, but rather, they could potentially represent critical instances that result in machine failure. Removing them could inadvertently strip our models of crucial data points and affect predictive capabilities.

#### 6.2.3 Addressing Class Imbalance via Oversampling

As mentioned in Section 5, given the significant difference in data size between instances of machine failure = 1 and machine failure = 0, our group has opted to apply oversampling to the instances of machine failure. Without such a measure, machine learning models could develop a bias toward predicting the more prevalent class — machine failure = 0. This imbalance could result in suboptimal generalisation when it comes to accurately identifying instances of machine failure, which are the minority class in the dataset.

We employed the SMOTE (Synthetic Minority Over-sampling Technique), which is designed to augment the minority class. SMOTE generates new, synthetic instances of the minority class by interpolating between existing minority instances. This procedure was facilitated by the utilisation of the “smotefamily” library within Rstudio, enabling us to increase the minority class representation in our dataset. As a result, the adjusted dataset comprises 19,222 data points. The distribution of Machine Failure instances has been substantially equalised, achieving a near 1:1 ratio between the two classes.

*Figure 6.1: Machine Failure before (left) and after oversampling (right)* 

#### 6.2.4 Selection of Accuracy Metrics

We used 4 error metrics to measure the accuracy of our model (Sharma, 2022).

Using a combination of metrics to determine the accuracy of our models allows us to avoid being misled by the accuracy paradox, where a model has high accuracy but is poor in predicting the minority class (Afonja, 2017).

| Metric | Formula | Value for Predictive Maintenance |
| --- | --- | --- |
| Accuracy |  | Gauges the overall correctness of the model by calculating the proportion of true results (true positives and true negatives) |
| Precision |  | Assesses the model's exactness by measuring the ratio of true positives to the total number of predicted positives |
| Recall |  | Evaluates completeness by determining the ratio of true positives to the actual number of positives |
| F1 Score |  | Provides a harmonic mean of precision and recall, offering a balance between the two and serving as a single measure of a model’s precision. |

\*TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative

*Table 6.2: Formula and Interpretation behind accuracy metrics*

### 6.3 Machine Failure Analysis

#### 6.3.1 First Model Type - Logistic Regression Model

The first machine learning model that we experimented with was logistic regression. To understand the relationship between our selected parameters and the machine failure outcome in our dataset, our model formulated an equation that probabilistically relates the parameters to the likelihood of machine failure.

Recognising the potential challenges posed by imbalanced datasets (refer to Section 6.2.4) we developed two distinct models. The first model did not apply any oversampling treatment and is hence referred to as the “imbalanced” model (see Appendix A2). Conversely, the second model incorporated oversampling to address the imbalance, termed the “balanced model” (see Appendix A3). This approach allowed us to comparatively assess the predictive accuracy of each model, ensuring that our findings were both robust and reliable.

To ensure the reliability of our model estimates, we monitored the multicollinearity amongst the dependent variables using Variance Inflation Factor (VIF) analysis. We verified that none of the continuous variables had a VIF value exceeding 10, and the categorical variable 'Type' did not exceed a GVIF value of 2 (see Appendix B3), thereby affirming the statistical significance and independence of each model component.

To quantify the impact of each feature, we also calculated the Odds Ratios and their corresponding Confidence Intervals. Notably, none of the calculated Odds Ratios equals 1, and the Confidence Intervals exclude 1 (see Appendix B4), therefore we can conclude that all of our independent variables are statistically significant.

We ran the model against all 6 independent variables (see Appendix B1, B2) and validated our model's performance using a confusion matrix, where the rows represent the model's predictions, and the columns represent the actual instances of machine failures. From our confusion matrix, we were able to obtain the necessary accuracy metrics as mentioned in Section 6.2.4.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F1-Score |
| Logistic Regression Model Using Imbalanced Dataset | 97.23% | 75.00% | 24.24% | 36.64% |
| Logistic Regression Model Using Balanced Dataset | 84.19% | 83.93% | 84.40% | 84.16% |

*Table 6.3: Machine Failure Prediction using Logistic Regression*

At first glance, it may appear that the imbalanced model has a higher accuracy at 97.23% compared to the balanced dataset at 84.19%. However, in the context of our dataset, it is important to note that our final goal is to successfully predict machine failures. Thus, a high accuracy might merely reflect the model's adeptness at predicting the majority class (i.e., no machine failure) while largely overlooking the instances of actual machine failures.

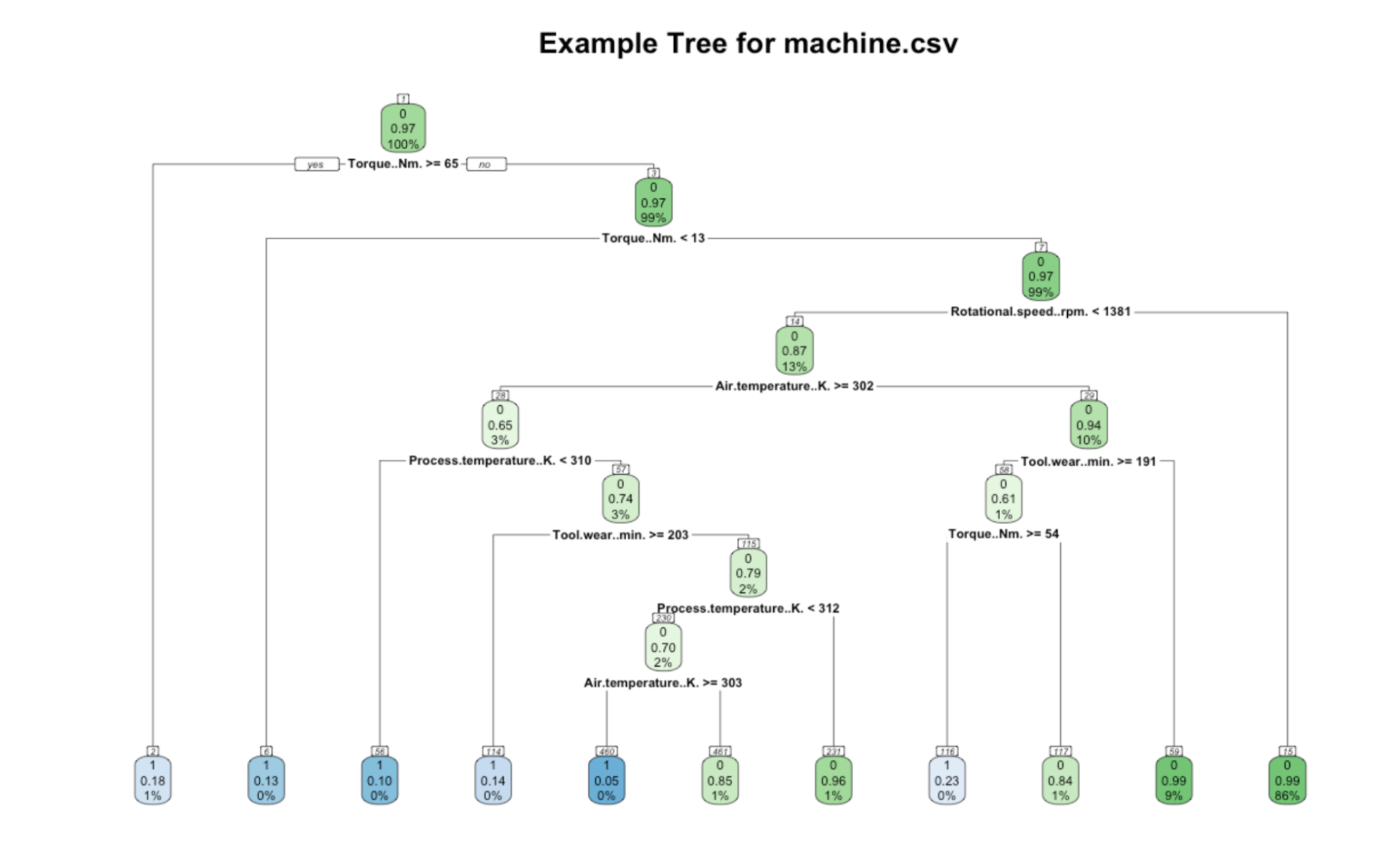
This is verified when we look at the recall. The imbalanced model's recall of 24.24% indicates that it only correctly identified about a quarter of actual machine failures. In industrial settings, missing out on three-quarters of potential machine failures could lead to costly downtimes (see Section 3).

In contrast, the balanced model, despite a dip in accuracy to 84.19%, showcased a significant leap in recall, precision and F1-Score to 84.40%, 83.93%, and 84.16% respectively. This is indicative of the model's heightened sensitivity to machine failures without excessively compromising on false alarms.

In this scenario, as the cost of a false negative is significantly higher, our balanced model is the better choice for predicting machine failures, since it offers a more holistic performance.

#### 6.3.2 Second Model Type - Classification and Regression Tree Model

The second component of our machine learning approach employs a Classification and Regression Tree (CART). This method produces a decision tree that serves as the model's output. In this tree structure, each branch point signifies a division based on a feature variable, while the terminal nodes represent a forecast for the target variable. Below, we provide a sample decision tree to illustrate this concept:



*Figure 6.2: Example decision tree trained on our under sampled dataset*

At each juncture of the decision tree, we make a calculated choice to proceed to either the right or left branch, depending on the criteria at that split. The first division is predicated on the variable 'Torque (Nm)' and its relation to the value of 65. This bifurcation of data creates two distinct paths:

* For instances where Torque is equal to or exceeds 65 Nm, we have a subset of 51 instances. Within this group, the probability of the outcome being labelled as "1", or machine failure is 0.176, and no machine failure being 0.824.
* Conversely, when Torque falls below 65 Nm, it encompasses a significantly larger portion of 6,950 instances. The labelling probabilities are markedly skewed towards "0" with a probability of 0.972, compared to a mere 0.028 for "1" (due to dominance of class "0" within this subset).

The decision tree depicted here is heavily simplified and has a broad error margin. For actionable insights and predictive accuracy, our team refined our model to balance the minimization of error against the risk of overfitting to our dataset.

Similar to our regression methodology, we produced 2 separate models, one without the application of oversampling treatment (see Appendix C.1, C,3), and the other with the application of oversampling (see Appendix C.2, C.4) to compare for predictive accuracy.

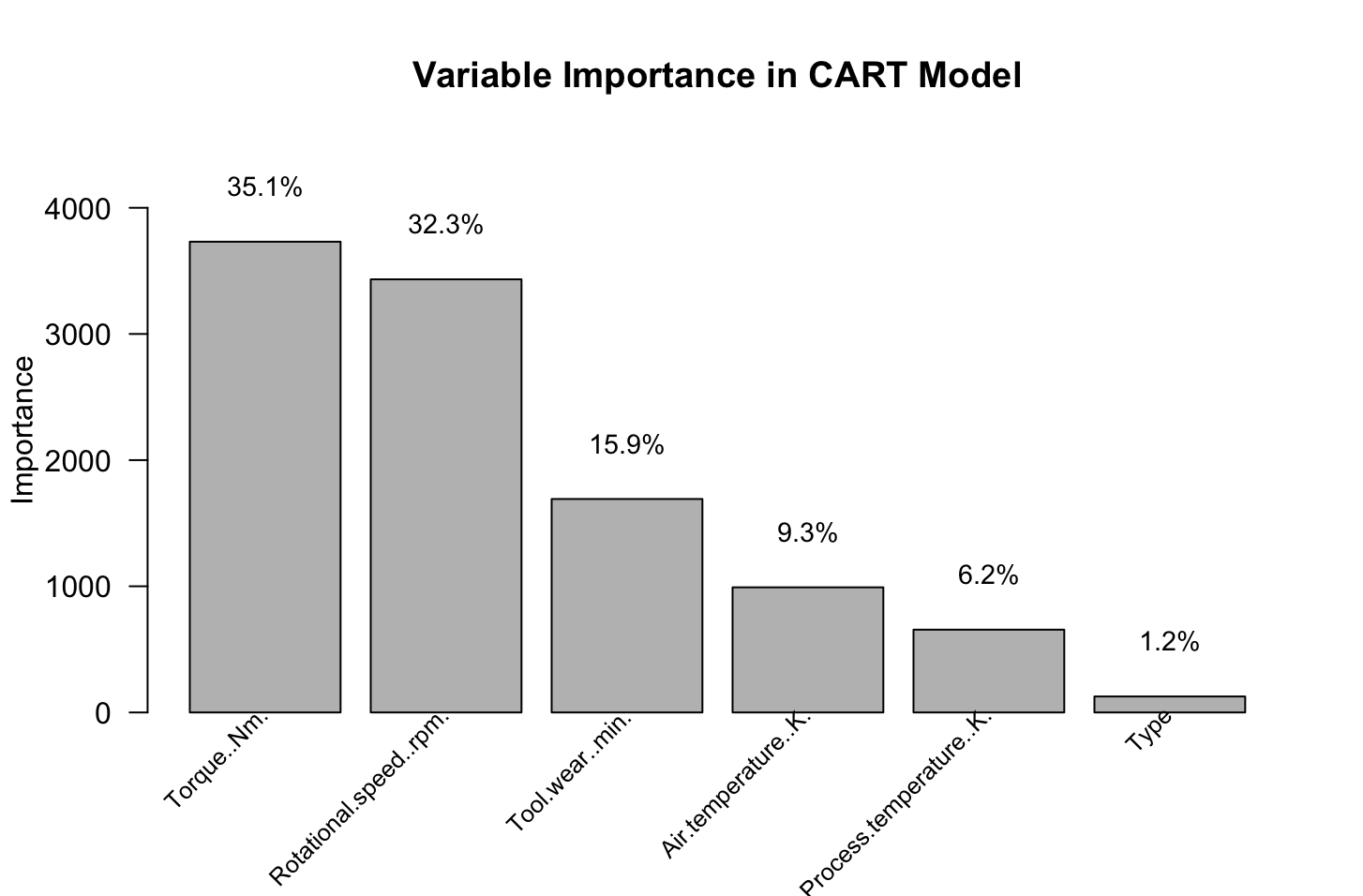
We followed standard procedures to obtain an optimal decision tree (see Appendix C.5, C.6) which entails achieving the minimal error in our prediction whilst maintaining the most simplicity. Predictive accuracy was then measured using the various accuracy metrics (see Section 6.2.4).

*Table 6.4: Machine Failure Prediction using CART*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Accuracy | Precision | Recall | F1-Score |
| CART Model Using Imbalanced Dataset | 98.36% | 84.72% | 61.62% | 71.35% |
| CART Model Using Balanced Dataset | 95.51% | 94.82% | 96.24% | 95.52% |

Similar to the logistic regression model, the CART model shows an improved F1 score, precision and recall, though the increase is less than that of logistic regression. This is because CART is inherently more adept in handling imbalanced datasets.

Besides failure prediction, we also wanted to see which were the variables which were crucial to determining machine failure. CART models provide such a metric for measuring the magnitude of the impact of an independent variable on machine failure called variable importance. The higher the proportion of variable importance, the more significant in predicting machine failure.

*Figure 6.3: Variable Importance Barplot from the CART Model trained on Balanced Dataset*

From Figure 6.3, torque has a 35.1% variable importance proportion, indicating that it accounts for 35.1% of the predictive power in machine failures. This metric is followed closely by rotational speed, with 32.3% importance, revealing these as the most significant factors in determining machine failure.

The results demonstrate that our CART model, applied to a balanced dataset, stands as a useful tool for Aramco, offering a reliable and user-friendly machine learning solution for predicting equipment failures. Utilising the CART model, we have also extracted valuable insights regarding the primary factors contributing to machine failures.

#### 6.3.3 Model Results Analysis

Comparing the best results from Logistic Regression and CART (models performed on the oversampled dataset):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Balanced Model | Accuracy | Precision | Recall | F1-Score |
| Logistic Regression Model | 84.19% | 83.93% | 84.40% | 84.16% |
| CART Model | 95.51% | 94.82% | 96.24% | 95.52% |

*Table 6.5: Best Results Comparison from CART and Logistic Regression Model*

The CART model outperforms the logistic regression model by a significant margin. With the CART model achieving an accuracy of 95.51% compared to the 84.19% of the logistic regression model, the CART model can better classify both machine failure and non-machine failure instances.

The higher precision of the CART model suggests that 94.82% of the instances it predicts as failures are actual failures, in contrast to the 83.93% precision of the logistic regression model. The CART model also has a higher recall of 96.24%, indicating that only a small percentage of machine failures were undetected by the model. This F1-score also jumps from 84.16% to 95.51%

The superior results of the CART model imply that the predictors and the target variable may share non-linear relationships. CART models, unlike logistic regression that assumes a linear relationship, are adept at identifying non-linear interactions within a dataset by successively splitting it into more uniform subsets. This ability to deal with nonlinearity without the need for variable transformation is a powerful asset in predicting complex outcomes like machine failure.

Moreover, the CART model offers a buffer against the influence of outliers, which our group decided to retain in our dataset. While outliers can skew the coefficients and predictions in logistic regression, potentially leading to misleading results, the branching nature of CART models tends to isolate these outliers, significantly mitigating their impact.

Given the significantly improved metrics, the CART model is better suited for predicting machine failure. We will further elaborate on the business value of our findings to Aramco in Section 7.

### 6.4 Machine Failure (Type) Analysis

Predictive models should not only be able to predict machine failures but also identify the type of failure that is likely to occur. This would enable maintenance teams to prepare more effectively and respond with precise interventions. Hence, we extended our predictive modelling to distinguish among four different machine failure types: TWF, HDF, PWF, and OSF.

We chose to exclude RNF from the analysis. Random failures typically occur without warning and are not directly related to the machine's condition or operating environment. Unlike systematic failures such as TWF, HDF, PWF, and OSF, which can be anticipated based on machine data and operational parameters, RNF is inherently stochastic and cannot be predicted precisely.

| Failure Type | Reason for Machine Failure |
| --- | --- |
| Tool Wear Failure (TWF) | The degradation or breakage of cutting or shaping tools results in diminished machine performance or the inability to continue operation. |
| Heat Dissipation Failure (HDF) | Overheating due to inadequate cooling can cause material or component damage, leading to machine malfunction or complete breakdown. |

*Table 6.6: Interpretations behind each machine failure type*

| Failure Type | Reason for Machine Failure |
| --- | --- |
| Power Failure (PWF) | Fluctuations in the machine's power supply can cause abrupt stoppages, damage to electronic components, or loss of operational control. |
| Over-speed Failure (OSF) | Excessive operational speed can induce mechanical stress and potential failure due to the increased forces on the machine's components. |

#### 6.4.1 First and Second Models - Logistic Regression and CART Model

Following a similar methodology as the models used in Section 6.3, we obtained the following accuracy metrics as well as the variable importance proportions for the respective failure types.

*Table 6.7: Machine Failure Type Prediction using Logistic Regression Model*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Failure Type Prediction | Accuracy | Precision | Recall | F1-Score |
| TWF using LR | 94.88% | 68.03% | 24.48% | 36.01% |
| HDF using LR | 97.10% | 90.51% | 93.88% | 92.16% |
| PWF using LR | 97.10% | 91.06% | 87.50% | 89.25% |
| OSF using LR | 96.97% | 91.43% | 88.89% | 90.14% |

*Table 6.8: Machine Failure Type Prediction using CART Model*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Failure Type Prediction | Accuracy | Precision | Recall | F1-Score |
| TWF using CART | 98.18% | 84.01% | 83.23% | 83.62% |
| HDF using CART | 99.31% | 97.44% | 98.61% | 98.02% |
| PWF using CART | 98.11% | 93.31% | 93.65% | 93.48% |
| OSF using CART | 98.49% | 95.66% | 94.71% | 95.19% |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Torque [Nm] | Rotational Speed [rpm] | Tool Wear [min] | Air Temperature [K] | Process Temperature [K] | Type |
| TWF | 20.0% | 19.2% | 23.8% | 17.0% | 15.1% | 4.9% |
| HDF | 20.0% | 26.3% | 4.9% | 27.4% | 20.4% | 1.0% |
| PWF | 49.8% | 35.7% | 5.0% | 4.1% | 3.8% | 1.6% |
| OSF | 29.2% | 20.6% | 32.2% | 5.7% | 5.5% | 6.9% |

*Table 6.9: Variable Importance for Each Machine Failure from CART Model (See Appendix C.7)*

From Table 6.9, Torque emerges as a universally significant factor across all failure types, with its greatest influence on Power PWF, where it accounts for almost 50% of variable importance. This underscores torque as a pivotal element in machine performance and integrity, indicating that monitoring and controlling torque could be vital in preventing failures.

Tool Wear takes precedence in TWF and OSF, implying that the physical degradation of tools is a critical predictor for these failure modes. Its prominence in OSF also suggests a possible interplay between tool condition and machine speed regulation, necessitating attentive maintenance and regular inspection schedules.

Air Temperature is a standout variable for HDF, with a notable percentage, pointing to the sensitivity of this failure type to thermal conditions. It indicates that external temperature conditions can significantly impact machine components, and that managing thermal dynamics could be a key strategy in reducing the risk of HDF.

Lastly, the 'Type' variable is relatively less important across the board, suggesting that the nature of failures is less dependent on the machine's model or category and more on the operational and environmental conditions it encounters. This observation may lead to more generalised, across-the-board strategies for failure prevention, rather than type-specific approaches.

#### 6.4.2 Third Model Type: K-Nearest Neighbors (kNN) Model

While our CART model results (Table 6.8) showed better prediction capabilities than our logistic regression models, the 'No Free Lunch' theorem suggests that there is no single best machine learning algorithm for predictive modelling problems such as classification and regression (Brownlee, 2021). As such, we decided to extend our methodology to include the k-Nearest Neighbors (kNN) model. This decision is grounded in our pursuit to identify the most suitable predictive model.

While logistic regression provides a solid baseline and CART offers insights from a decision tree perspective, kNN's instance-based learning approach could be more relevant in our prediction of machine failure type. This is because the kNN model's ability to adapt to new changes rapidly without requiring retraining can be particularly advantageous in a production environment where conditions often vary and new data becomes available.

*Table 6.10: Machine Failure Type Prediction using KNN Model*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Failure Type Prediction | Accuracy | Precision | Recall | F1-Score |
| TWF | 97.97% | 78.03% | 91.15% | 84.08% |
| HDF | 99.39% | 97.11% | 99.62% | 98.35% |
| PWF | 99.12% | 95.91% | 97.73% | 96.81% |
| OSF | 97.40% | 87.88% | 96.67% | 92.06% |

#### 6.4.3 Model Results Analysis

Overall, both the kNN and CART models demonstrate comparable predictive abilities. The kNN model generally presents higher recall rates, particularly with TWF and OSF. This indicates that the kNN model has a heightened sensitivity to actual failures, making it a potentially more suitable option for situations where the repercussions of overlooking a failure are substantial.

Conversely, the CART model exhibits a slight improvement in precision for HDF and OSF failures, suggesting that its predictions of failures are more often correct. This characteristic may be advantageous in contexts where false positives entail significant consequences. The CART model also maintains more uniform performance across various metrics for each type of failure, which may indicate that it can offer more consistent predictions under different operational conditions.

Despite the strengths of both models, the group has chosen the kNN model as the more appropriate one for predicting machine failure types. This decision is based on the priority to minimise missed failure detections over the occasional false alarm, especially considering the specific needs of Aramco, where the costs of undetected failures are likely higher than those of false positives.

## 7. Business Value of Proposed Methodology

Adopting predictive maintenance marks, a pivotal shift for Aramco from preventive to proactive equipment management. This shift ensures timely repair or replacement of equipment before a costly failure occurs, streamlining operations and minimising downtime costs with an average reduction of 20% in downtime reported in the oil and gas industry (Decaix et al., 2021). This approach not only preserves equipment integrity but also supports operational excellence and financial stability.

Integrating predictive maintenance into Aramco's operations enhances safety and environmental conservation. By predicting and preventing equipment failures without human biases before they occur, this ensures a safer working environment by reducing the risks of accidents caused by equipment failure by approximately 50% (Gager, 2017). Moreover, this also leads to a reduction in environmental risks such as oil spills, which could result from equipment failure at Aramco (Aramco, 2023). Through predictive maintenance, Aramco can better align with global safety and environmental standards, thereby reinforcing its position as a leader in responsible energy production practices (Aramco, 2023).

Our predictive maintenance model also improves Aramco's maintenance efficiency by accurately forecasting machine failures and its associated machine failure type, which streamlines diagnostics and accelerates the repair process. This accuracy in pinpointing issues reduces maintenance costs by approximately 10% (Deloitte, 2017), optimises interventions, minimises delays, and increases equipment lifespan (Raj, 2020). Furthermore, this model facilitates a just-in-time inventory system, ensuring the availability of spare parts exactly when needed and reducing the necessity for large storage spaces. This approach promotes efficient use of resources, lowers storage expenses, and reduces the likelihood of surplus inventory, yielding cost savings.

By leveraging predictive maintenance data, Aramco can also tailor equipment design and procurement strategies by enabling the development of more robust machinery and the selection of assets that are less susceptible to frequent failures. For instance, insights from milling machines show that torque irregularities significantly contribute to equipment failures. Aramco could proactively integrate torque monitoring and control systems into critical equipment such as pumps and compressors, strengthening operational resilience, and affirming its leadership in innovative and efficient machinery management.

## 8. Recommendations

### 8.1 Short Term Approach

**CART Model Implementation for Maintenance Enhancement:** We recommend the short-term implementation of a CART model to bolster Aramco’s maintenance strategy. The CART model is adept at utilising existing data to derive valuable insights, even in the absence of data specifically tailored for machine failure prediction. By harnessing these insights, maintenance personnel can transition from preventive to a more proactive, equipment-specific maintenance approach, making informed on-the-spot decisions regarding equipment replacement or repairs. This interim solution does not necessitate an extensive overhaul of existing systems or the immediate integration of IoT sensors, thus emerging as a practical and cost-effective enhancement to the current maintenance framework.

### 8.2 Long Term Approach

**Transitioning to Predictive Maintenance with kNN Model:** In the long term, advancing towards a predictive maintenance strategy, underpinned by the kNN model and IoT sensor data, is recommended. This enables real-time equipment monitoring and failure prediction, building upon the foundational insights garnered from the short-term CART model implementation. The goal is to establish a centralised maintenance hub that facilitates timely, precise, and informed maintenance decisions. This strategic shift addresses not only the immediate inefficiencies but also instils a resilient and adaptive maintenance framework, aligning with the evolving demands of modern machinery management.

**Procurement Strategy Enhancement through Machine Failure Analysis:** We propose that Aramco integrates insights from predictive maintenance analytics into its procurement process. Analysis of failure trends can grant Aramco a deeper understanding of machinery performance and maintenance requirements, which aids the strategic selection of reliable and maintainable machinery. Using predictive data, Aramco could strategically negotiate improved terms with vendors to ensure optimal benefits for operational uses. For instance, if predictive maintenance analytics indicate that a certain type of pump frequently fails due to overheating after a specific number of operating hours, Aramco could use this information to procure new pumps with enhanced cooling features or those made from heat-resistant materials. Conversely, Aramco could negotiate for pumps that come with an extended warranty for overheating issues, or they could work with vendors to design a custom solution that addresses this specific problem.

## 9. Limitations of recommended solutions

### 9.1 Limited Independent Variables

In our analysis for proof of concept, we have only considered a few independent factors in the construction of our models. These factors may not fully explain equipment failure in Aramco’s operation.

**Proposed solution:** We suggest Aramco's engineers initiate a pilot study using IoT sensors on selected equipment. This approach will enable real-time monitoring and predictive analytics, offering a clearer understanding of equipment failure precursors and improving maintenance strategies.

**9.2 Correlation Does Not Imply Causation**

As with all machine learning models that attempt to find hidden patterns and relationships between variables, correlations between any two variables do not imply causation. For example, in figure 5.4 there is a strong negative correlation (-0.87) between 'Rotational speed\_rpm' and 'Torque\_Nm.' While these two variables move in opposite directions, this correlation does not confirm that a change in rotational speed directly causes a change in torque, or vice versa.

**Proposed solution:** Aramco should pursue a two-pronged approach by conducting controlled experiments and leveraging digital twin technology. Controlled experiments will allow for the manipulation of individual variables, such as 'Rotational speed\_rpm' and 'Torque\_Nm,' in a monitored environment to establish causation. Concurrently, developing digital twins of critical machinery will enable Aramco to simulate various scenarios, validate the findings from controlled experiments, and understand the broader impact of operational changes (IBM, 2022).

#### 

#### 9.3 Insufficient Observations of Machine Failure

Our predictive models are constrained by a limited dataset where only 330 out of 10,000 entries represent machine failures, which may not adequately capture the full spectrum of failure causes. To address this, we utilised SMOTE to oversample the minority class and enhance model training. However, this technique introduces its own limitations by potentially biasing the model towards these synthetic instances, thus affecting the model's ability to generalise to real-world scenarios accurately.

**Proposed solution:** Aramco could significantly improve the precision of its predictive models by participating in data sharing partnerships with industry counterparts. This collaboration would create a more extensive and varied machine failure dataset, leading to a more comprehensive grasp of equipment reliability across shared operational contexts. Additionally, integrating incremental learning algorithms would ensure the model evolves and adapts continuously, incorporating new data as it becomes available and keeping the predictive insights current and increasingly reliable over time.

### 9.4 Complexity of CART Models

The complexity of our current CART model is due to its extensive number of decision nodes and deep tree structure, reflecting a nuanced interpretation of the data. This intricacy can make the model difficult for business stakeholders to interpret, leading to potential trust issues, higher maintenance costs due to the specialised upkeep required, and hurdles in training staff to use it effectively. Moreover, the model’s complexity may result in decision-making delays as users navigate the dense array of rules, which underscores the need for a simplified approach or a more intuitive user interface to facilitate broader adoption and real-world application.

**Proposed solution:** The process can be streamlined by creating an intuitive digital form for maintenance workers to submit relevant machine data, which is instantly analysed by the CART model. The system would generate immediate, actionable feedback in a simple format that workers can easily interpret and apply. Training sessions would be provided to ensure workers are well versed with the form and results, with continuous refinements to the system based on user feedback.

## 10. Conclusion

Adopting the outlined predictive maintenance methodology could significantly advance Aramco's operational efficiency, safety, and equipment management. The proposed short-term and long-term strategies, underpinned by CART and kNN models, aim to mitigate unexpected equipment failures by enabling accurate forecasting of machine breakdowns and identification of potential failure types. This proactive approach not only enhances operational efficiency but also optimises inventory management and informs strategic procurement decisions, thus reducing maintenance costs and fostering continuous improvement. As Aramco transitions from a preventive approach to predictive maintenance practices, it stands to redefine industry standards, solidifying its position as a leader in maintenance innovation and operational excellence.

## References

Afonja, T. (2017). Accuracy paradox. *Medium.*

https://towardsdatascience.com/accuracy-paradox-897a69e2dd9b

Al-Hamad, T. (2011). Saudi Aramco Suppliers Safety Management System. *Saudi Aramco.*

<https://www.aramco.com/-/media/downloads/working-with-us/saudi-aramco-suppliers-safety-management-system-ssms.pdf?la=en&hash=9620BCE26DE8648B32F58E09D5A9A41D363CD63E>

Al-Sultan, T. (2009). Preventive Maintenance Optimization (PMO) at ShGP. *Saudi Aramco.*

<http://www.gpa-gcc-chapter.org/Conferences/spevents/MediaHandler/GenericHandler/documents/SpecialEvents/PreventiveMaintenanceOptimization-CaseStudy.pdf>

Aramco. (2023). *sustainability report 2022*. Aramco. https://www.aramco.com/-/media/downloads/sustainability-report/report-2022/2022-minimizing-environmental-impact-en.pdf?la=en&hash=EF6ADCCB08339239F02A97FE1467C9CD6A338CAC

*Bitstrapped*. (2023). Machine Learning in the Oil and Gas Industry: ML Roles and Applications.

<https://www.bitstrapped.com/blog/machine-learning-oil-and-gas-industry>

Brownlee, J. (2021). No free lunch theorem for machine learning. *MachineLearningMastery*.

https://machinelearningmastery.com/no-free-lunch-theorem-for-machine-learning/

Decaix, G., Gentzel, M., Luse, A., & Thibert, J. (2021, April 23). *A smarter way to digitize maintenance and reliability | McKinsey*. Www.mckinsey.com. https://www.mckinsey.com/capabilities/operations/our-insights/a-smarter-way-to-digitize-maintenance-and-reliability

Deloitte. (2017). *Predictive Maintenance Taking pro-active measures based on advanced data analytics to predict and avoid machine failure*. https://www2.deloitte.com/content/dam/Deloitte/de/Documents/deloitte-analytics/Deloitte\_Predictive-Maintenance\_PositionPaper.pdf

Garger, A. (2017). Predicting Equipment Failures. Facilitiesnet.

https:// facilitiesnet.com/facilitiesmanagement/article/Predicting-Equipment-Failures--17256

IBM. (2022). What Is a Digital Twin. *IBM*.

https://www.ibm.com/topics/what-is-a-digital-twin

Mobley, R. K. (2002). An Introduction to Predictive Maintenance. In Google Books. *Elsevier.*

<https://books.google.com.sg/books?hl=en&lr=&id=SjqXzxpAzSQC&oi=fnd&pg=PP1&ots=iGxROGEjfg&sig=zC3mbI8Q1sEFo1J3UN5gmHbaTWk&redir_esc=y#v=onepage&q&f=false>

Raj, R. (2020, September 30). *How preventive maintenance can backfire and harm your assets*. Www.ey.com. https://www.ey.com/en\_us/consulting/how-preventive-maintenance-can-backfire-and-harm-your-assets

Sharma, P. (2022). Decoding the confusion matrix. *Medium.*

<https://towardsdatascience.com/decoding-the-confusion-matrix-bb4801decbb>

*Siemens AG.* (2023). SENSEYE PREDICTIVE MAINTENANCE.

<https://assets.new.siemens.com/siemens/assets/api/uuid:3d606495-dbe0-43e4-80b1-d04e27ada920/dics-b10153-00-7600truecostofdowntime2022-144.pdf>

*Sulzer.* (2017). The Challenges of Maintaining Equipment on Offshore Platforms.

<https://www.sulzer.com/-/media/files/campaigns/res/offshorewhitepapercampaignthechallengesofmaintainingequipmentonoffshoreplatformswhitepaperpdfrender.ashx?la=en>

*The New Arab.* (2019). Two killed during 'incident' at Saudi Aramco oil facility.

https://www.newarab.com/news/two-killed-during-incident-saudi-aramco-oil-facility

Wallace, R. (2020). Big data, big insights. *Aramco.*

<https://singapore.aramco.com/en/magazine/elements/2020/big-data-big-insights>

Watkins, E. (2019). Saudi gas pipeline fire kills at least 28. *Oil & Gas Journal.*

<https://www.ogj.com/pipelines-transportation/article/17286212/saudi-gas-pipeline-fire-kills-at-least-28>

Webb, S., & Karam, S. (2007). Saudi gas pipeline fire kills 28. *Reuters.*

<https://www.reuters.com/article/idINIndia-30564220071118>

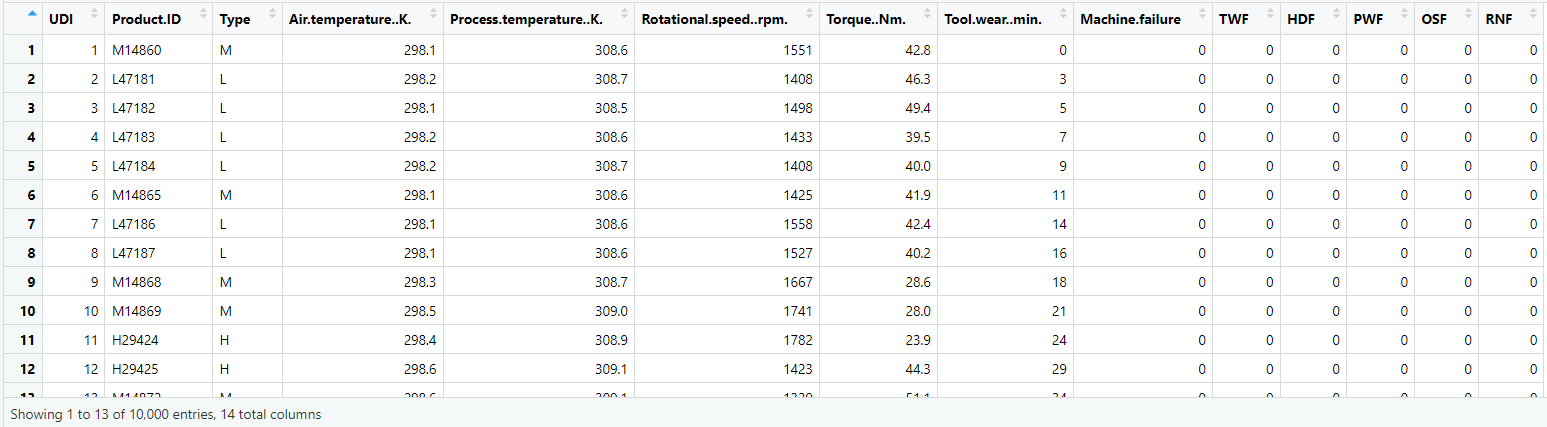
Zahid, H. B. (2021). A textbook of machine design by R.S.KHURMI and J.K.GUPTA. *Academia.edu.*

https://www.academia.edu/45663236/A\_Textbook\_of\_Machine\_Design\_by\_R\_S\_KHURMI\_AND\_J\_K\_GUPTA

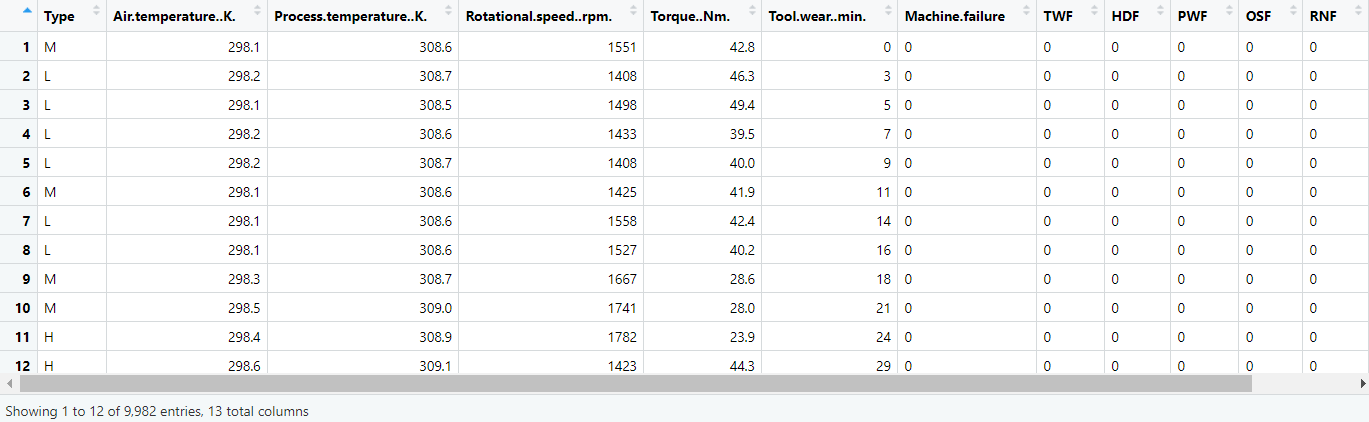
## Appendix

### Appendix A: Datasets

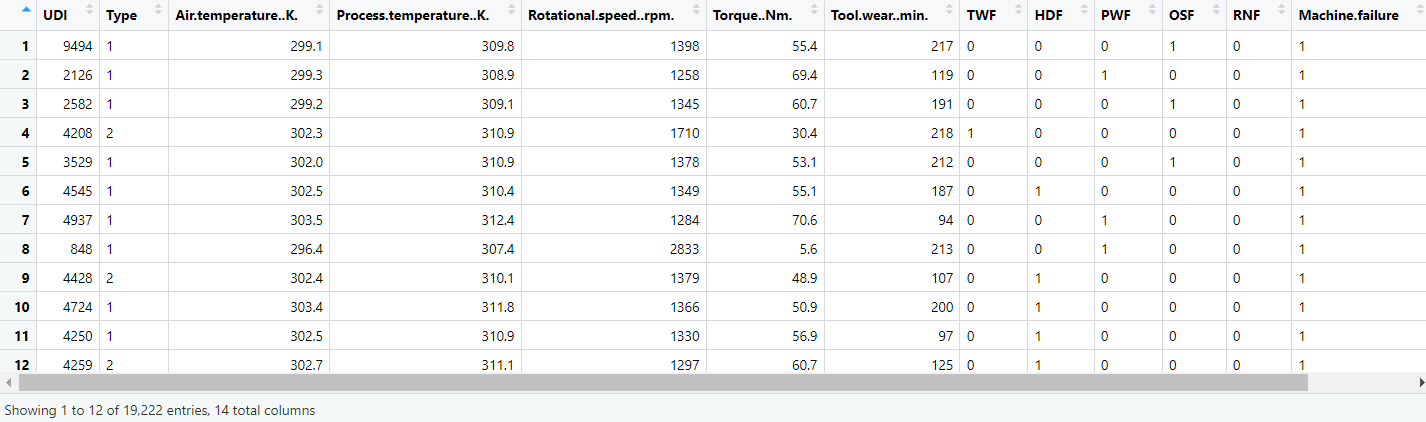
Appendix A1: Snapshot of Raw Dataset



Appendix A2: Snapshot of Imbalanced Dataset (After Data Cleaning)

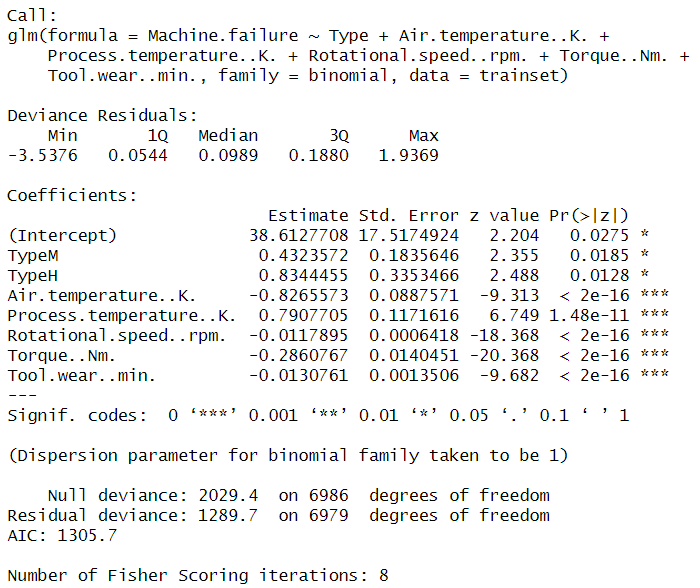


Appendix A3: Snapshot of Balanced Dataset (After Oversampling)

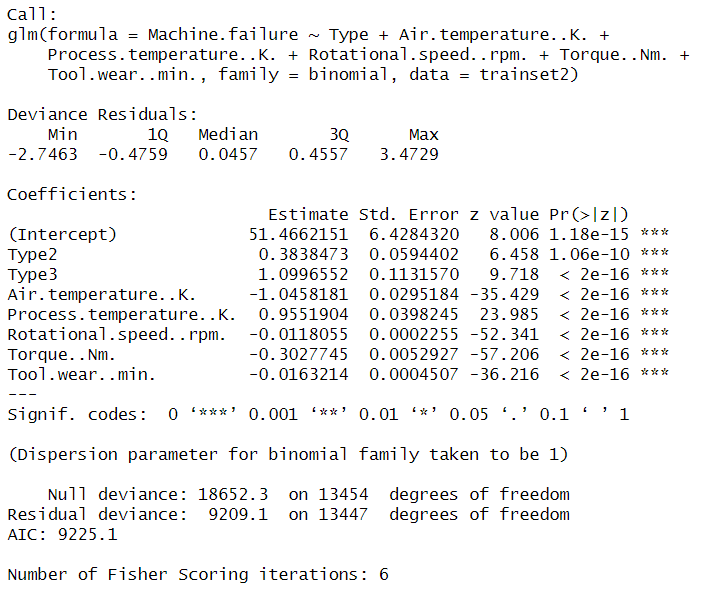


### Appendix B: Logistic Regression

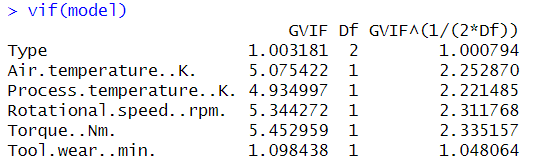
Appendix B1: Summary of Logistic Regression Model (Before Oversampling)



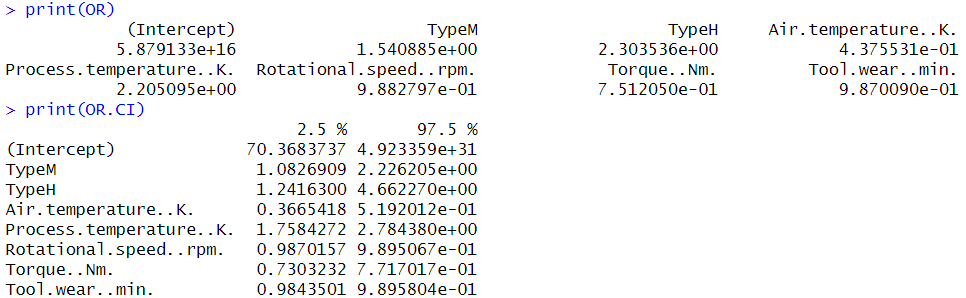
Appendix B2: Summary of Logistic Regression Model (After Oversampling)



Appendix B3: VIF Analysis



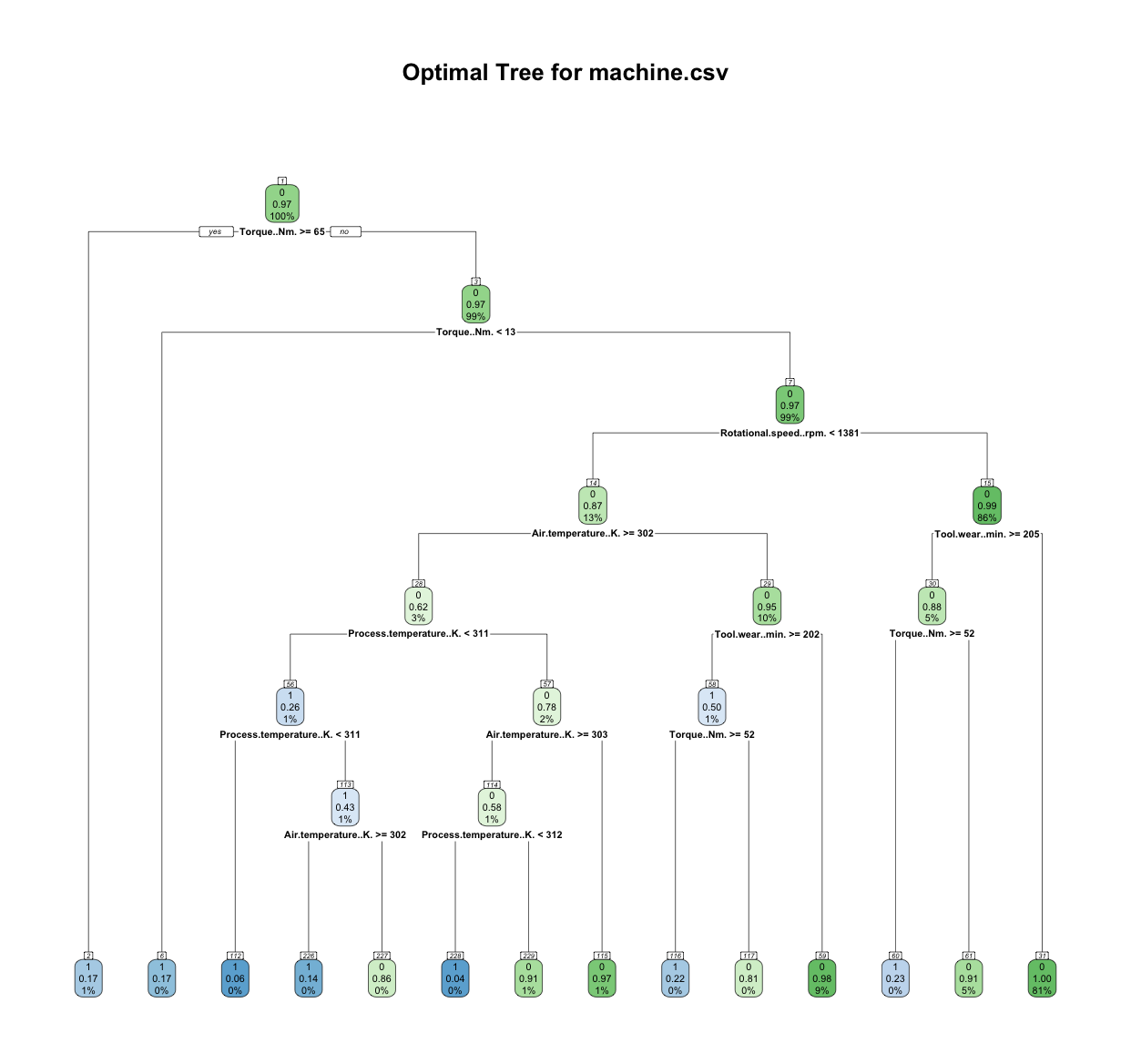
Appendix B4: Odds Ratio and Confidence Interval

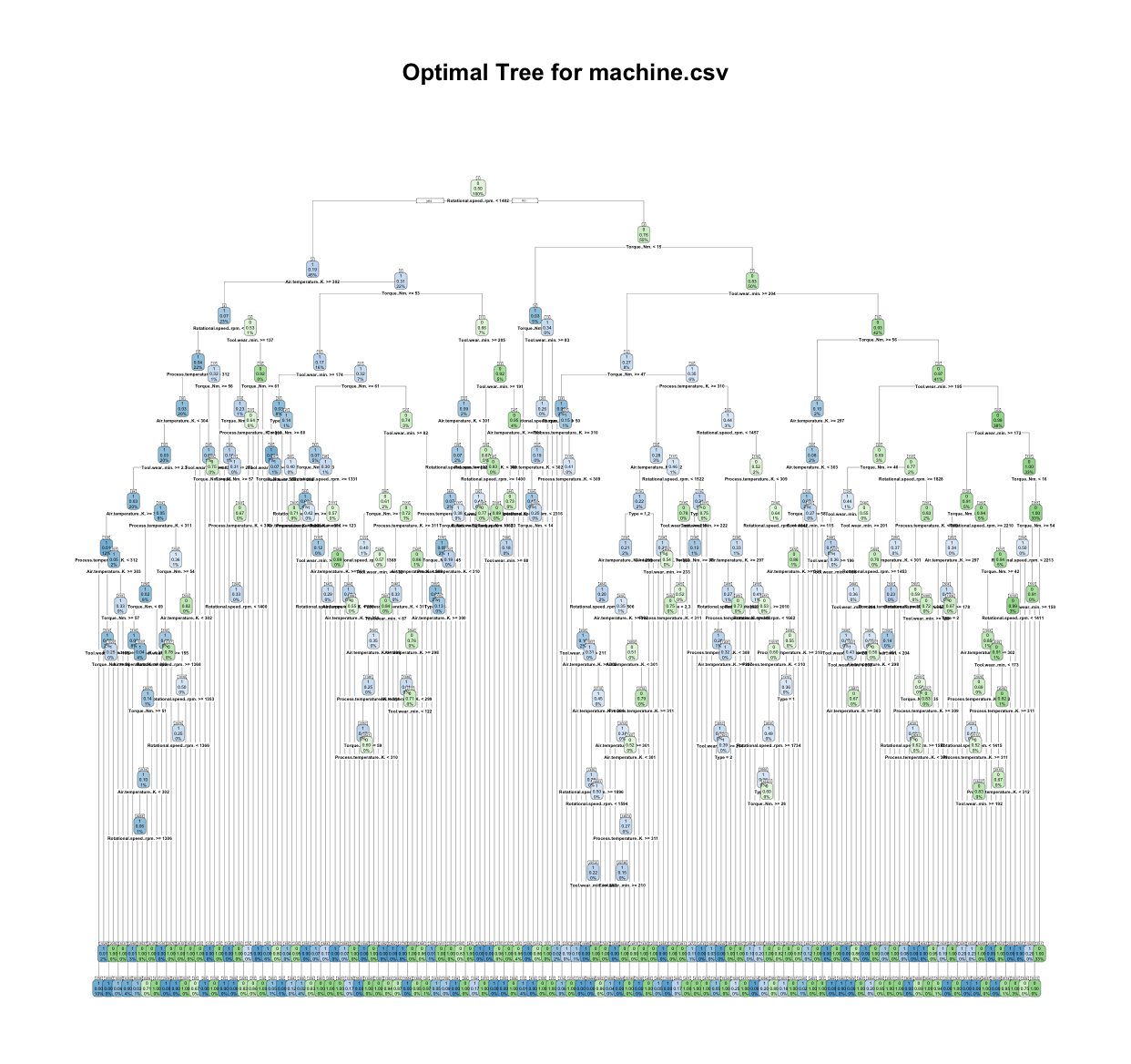


Appendix B5: Summary of Logistic Regression Model (TWF)

### Appendix C: CART

Appendix C1: Optimal Tree for “Raw” Dataset



Appendix C2: Optimal Tree for “Oversampled” Dataset

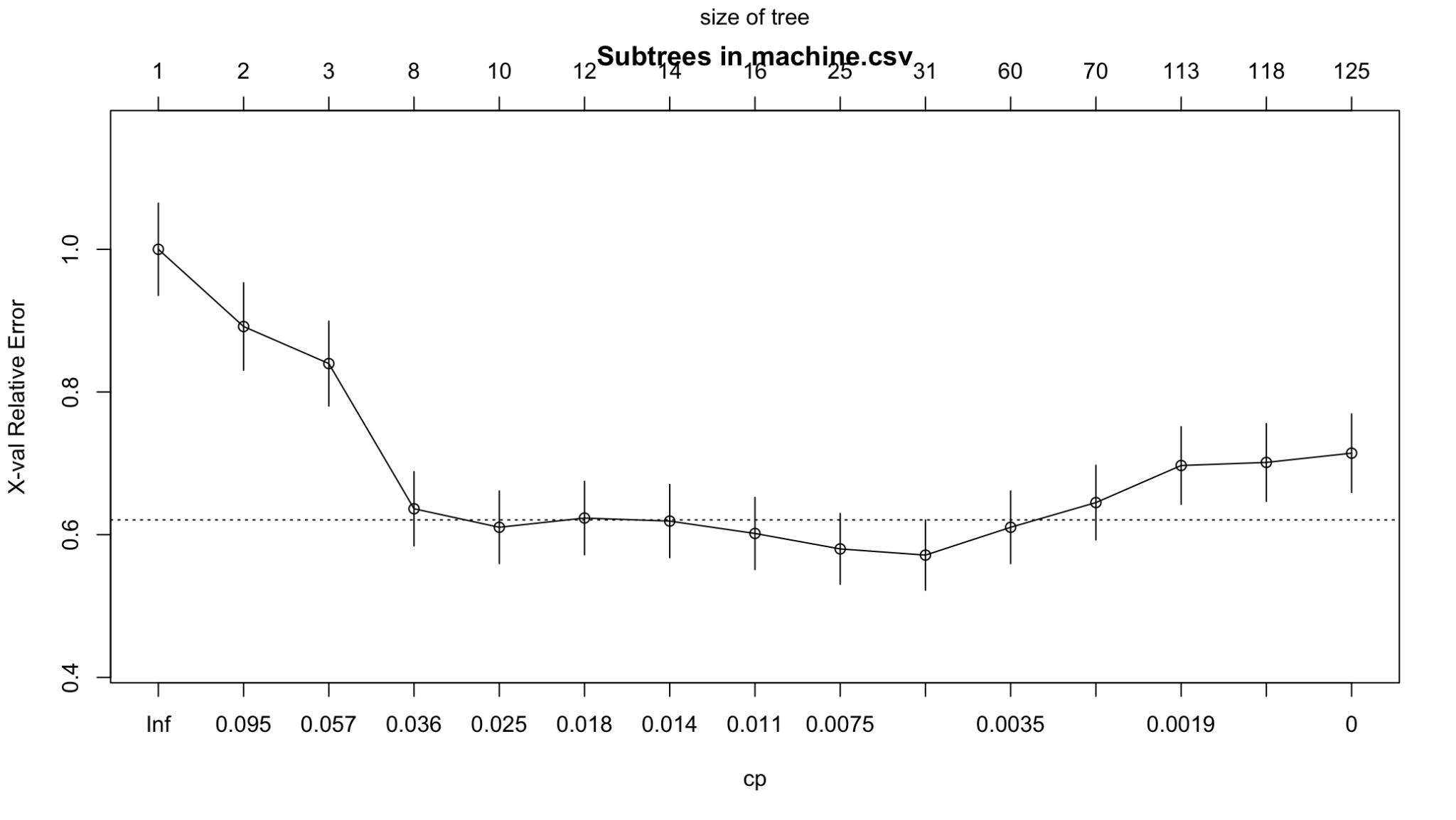
Appendix C3. Results of Optimal printcp () using “Raw” Dataset

|  |
| --- |
|  |
|  |

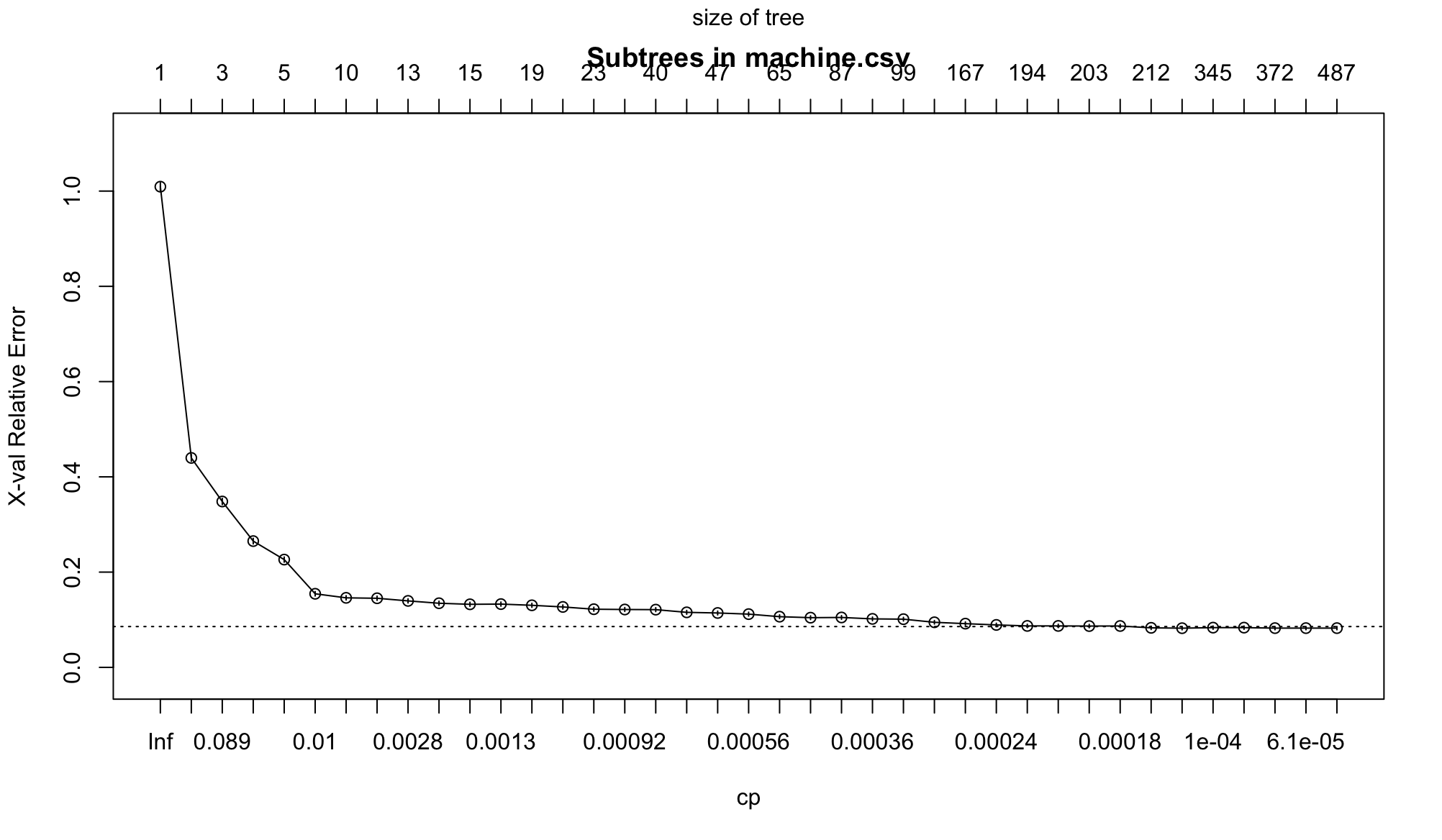
Appendix C4. Results of Optimal printcp () using “Oversampled” Dataset

|  |
| --- |
|  |
|  |

Appendix C5. Relative Error against Complexity Parameter Plot (“Raw” CART Model)



Appendix C6. Relative Error against Complexity Parameter Plot (“Oversampled” CART Model)



Appendix C.7: Variable Importance Barcharts for TWF, HDF, PWF, OSF.’

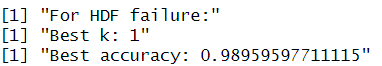
|  |
| --- |
|  |
|  |
|  |
|  |

### Appendix D: kNN

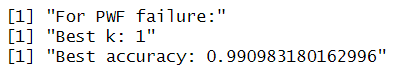
Appendix D1: Optimal k for TWF



Appendix D2: Optimal k for HDF



Appendix D3: Optimal k for HDF



Appendix D4: Optimal k for HDF

