

**BC2406 – Analytics I**

**Project AramcoGuard: Guarding Against Equipment Failure At Aramco**

**Seminar Group: 05, Team 6**

**CoB (NBS) Michele Nguyen (Dr)**

**4th Nov 2023**

|  |  |  |
| --- | --- | --- |
| **Group Members** | | |
| **SCSE TUA JING YU, JONAS (U2220298J)** | | |
| **NBS NICHOLAS TING JINGJIE (U2211640A)** | | |
| **NBS ASHIQ MAHDIYAT (U2210651J)** | | |
| **NBS ALETHEA SOH MAY TING (U2210332B)** | | |
| **NBS PEH JIA LIN PHOEBE (U2210143H)** | | |
|  | **Pages** | **19** |

**Table of Contents**

[**Executive Summary 5**](#_ac7s62kricda)

[**1. Business Understanding 6**](#_9sokk0smsazu)

[1.1. Background on Saudi Arabian Oil Group (Aramco) 6](#_6lgtkpkuigyn)

[1.2. Business Problem 6](#_q9gqdd4qxve)

[1.3. Opportunity Statement 6](#_tugke9b02osr)

[**2. Methodology 7**](#_gk9gg86yqr0m)

[**3. Data Preparation 7**](#_rlijcw29xtu9)

[3.1. Data Planning 7](#_d6bbn9nthwlt)

[3.2. Data Sources 8](#_jk3ox64s7uc4)

[3.3. Dataset 8](#_nrjrhwsxgp8y)

[3.3.1. Y variable 9](#_il2spmm5l4a5)

[3.3.2. X variable 9](#_h3thywv6zxhk)

[3.4. Data Cleaning 9](#_z4rwc7hr74e9)

[**4. Data Exploration 11**](#_z348zugyzu6m)

[4.1. database.csv 11](#_ox60w7fi5jpx)

[4.1.1. Categorical X Variables against Continuous Y Variables 11](#_g978t8duuqtb)

[4.1.2. Continuous Y Variables against Continuous Y Variables 11](#_8n2bx2fhcvaj)

[4.2. predictive\_maintance.csv 12](#_5iblc571psv2)

[4.2.1. Continuous X variables against continuous X variables 12](#_p7218ykpyagr)

[4.3. Sampling Methods 12](#_l3a6sk6l9kyr)

[4.3.1. Stratified Sampling 12](#_8be9u6qs827y)

[4.3.2. Undersampling 13](#_neo4ne750qqz)

[**5. Model Training & Evaluation Methods 13**](#_ehgics7e766k)

[5.1. Logistic Regression Model 13](#_rbm2p35vbjlo)

[5.1.1. Logistic Regression - Stratified Sampling 13](#_jp8jronnrjvq)

[5.1.2. Using AIC To Determine Optimal Logistic Model 13](#_sfn5qfi11sf0)

[5.1.3. Odds Ratio and Confidence Interval of OR 13](#_c6mlp21kzq23)

[5.1.4. Preventing Multicollinearity Using VIF 14](#_uii8535787wa)

[5.2. Logistic Regression Model - Undersampling 15](#_adn5s2yr2zd6)

[5.3. Classification and Regression Tree Model (CART) 16](#_r364we66h6s8)

[5.3.1. CART Model - Stratified Sampling 16](#_nbiuuonsf969)

[5.4. Performance Evaluation 17](#_stuhz148xm5v)

[5.4.1. Confusion Matrix (Appendix Q) 17](#_8q51ehj1hrff)

[5.4.2. Accuracy Percentage 18](#_8ql8favts0f)

[5.4.3. Cost Sensitive CART Model (Appendix R) 18](#_4hxlcgybvnak)

[5.5. CART Model - Undersampling 18](#_6pu6ezdoq3x2)

[5.6. Performance Evaluation 19](#_2kn376ku42yc)

[5.6.1. Confusion Matrix (Appendix T) 19](#_glrqgztsu0s3)

[5.6.2. Accuracy Percentage 19](#_ar9tuelmeth5)

[5.6.3. Cost Sensitive CART Model (Appendix U) 19](#_ngx0ejm6ub3l)

[**6. Comparison Of All Models 20**](#_760hqlzfkyst)

[6.1. Accuracy Percentage 20](#_okrzytakwjlh)

[6.2. AUC-ROC Curve 21](#_vl3zcejr03cw)

[6.3. False Negative Rate Plots 22](#_6ua2czij2ddz)

[6.4. Conclusion 22](#_95mqgfd8xq5)

[**7. Recommendations and conclusions 23**](#_eorhvft24icr)

[7.1. Collection of Data 23](#_g95v38k6xea6)

[7.2. Better Harnessing of Digital Twin Technology 23](#_xjm701b121k5)

[**Appendices 23**](#_8xqkz85coad9)

[Appendix A: Data Dictionary 24](#_sg1sst3d1or)

[Appendix B: Factorising Data 27](#_mi3jjeydvtzh)

[Appendix C: Checking for NA values 29](#_cd0lf7rqvnqu)

[Appendix D: Checking and removing duplicated rows 29](#_rtiqoyil6pn7)

[Appendix E: Box Plot of Continuous Variables 30](#_u01xe8hb3a78)

[Appendix F: Factorising and Renaming Columns 31](#_yr8vvcs1rw24)

[Appendix H: Restructuring Variables from K to °C 31](#_7qm3haqz0hor)

[Appendix I: Dropping Columns 31](#_ai27w1cinyq8)

[Appendix J: Relationship Between Categorical X Variables & Continuous Y Variable 31](#_es94u0ta9qkk)

[Appendix K: Relationship Between Continuous Y Variables and Continuous Y Variable 34](#_din1dqkuhk0o)

[Appendix L: Relationship Between Continuous X Variables 34](#_kya6oc3mgt7r)

[Appendix M: Histogram on the Frequency of Failure 35](#_zdf64x4e7c4c)

[Appendix N: Stratified Sampling Process 36](#_9ahmdz330w2u)

[Appendix O: Undersampling Process 36](#_mjoi1xivr7kk)

[Appendix P: Optimising CP - Stratified 36](#_u7fj8ohctfxw)

[Appendix Q : Confusion Matrix Comparison before after pruning - Stratified 37](#_xu7b0j6utizw)

[Appendix R: Cost Sensitivity Learning and Confusion Matrix - Stratified 38](#_xu09okuoc4fs)

[Appendix S: Optimising CP - Undersampling 38](#_n0w68yl0dw8j)

[Appendix T: Confusion Matrix before after pruning - Undersampling 39](#_rv5l5p99qvt2)

[Appendix U: Cost Sensitivity Learning and Confusion Matrix - Undersampling 40](#_2g20s7kkwgb5)

[Appendix V: Calculation and Compilation of Confusion Matrix for Evaluation 40](#_6iuqaasnwiju)

[**References 42**](#_bbeyys1c5anv)

# Executive Summary

This report seeks to examine and determine significant causes of pipeline accidents as well as provide predictive models that allow Aramco to detect early signs of identified causes that pose safety concerns. These models will help Aramco reduce unnecessary safety hazards and resource loss brought about by delayed responses to identified causes The models allow us to analyse and suggest business solutions to Aramco to deal with safety concerns.

Dataset and Proposed Proof of Concept (POC)

The team has conducted in-depth exploratory data analysis (EDA) on our dataset and optimised it for our POC models. We came up with six different models throughout this project, utilising two algorithms:

1. **Logistic Regression:** A predictive model that we used to identify the key factors that lead to equipment failure in the oil and drilling industry. This would allow Saudi Aramco to plan out their maintenance efforts, alleviating safety concerns and giving them greater confidence in their operations.
2. **Classification & Regression Tree:** An algorithm which enables us to determine which are the significant variables and predict equipment failure by generating a decision tree.

Our group chose these models as they are useful in helping us comprehend the causes leading to pipeline accidents, helping Aramco take preventive measures and avoid pipeline accidents.

# Business Understanding

## Background on Saudi Arabian Oil Group (Aramco)

The oil and drilling industry is known as one of the most hazardous occupations, with the workers facing inherent safety risks in every step of the process (Katch Kan, 2023). The Deepwater Horizon Oil Spill in 2010 that occurred due to equipment failure leaking oil led to a catastrophic explosion. The explosion not only took 11 lives, but also led to extensive environmental damage, and significant financial costs (Pallardy, 2010). This goes to show that there is a pressing issue in the safety of Aramco’s drilling equipment, which also poses a significant risk to their oil supply.

As a preventive measure, Aramco currently schedules regular maintenance to ensure that their equipment runs smoothly, preventing pipeline leaks. For example, Aramco shuts its Al Jubail refinery off for 45 days for maintenance works, losing 450,000 barrels of oil that could be drilled per day (Oil&Gas, 2023). This is extremely costly in terms of profit that could have been generated, as well as high maintenance costs.

## Business Problem

Aramco currently only has actions that are reactive and corrective in nature (Aramco, 2011). However, it opens up the chances of Aramco incurring high costs from having to deal with equipment failures. For instance, preventive maintenance can lead to inefficiencies, condition-based maintenance is cost-prohibitive, and corrective maintenance results in unplanned downtime. Predetermined maintenance may not accurately reflect equipment condition, and predictive maintenance faces challenges due to the scarcity of skilled personnel in this field for data interpretation (Cox, 2020). Addressing these challenges is crucial to optimise maintenance and enhance operational efficiency.

## Opportunity Statement

Considering the broader industry’s susceptibility to equipment-related failure, there is a pressing need for advanced maintenance strategies. In this context, there exists a prime opportunity for Saudi Aramco. By harnessing advanced technologies and data analytics, predictive maintenance can offer real-time insights into the health and performance of equipment, allowing for timely interventions before potential failures occur.

# Methodology

We hypothesise that the occurrence of pipeline accidents in Aramco’s processes can be predicted through variables such as product quality type, torque, tool wear, rotational speed, air temperature, process temperature etc. We will thus be using the following methodology of preparing our data and conducting exploratory analysis, selecting features to focus on, building different models and evaluating the models to come to a conclusion to our hypothesis.



# Data Preparation

## Data Planning

|  |
| --- |
| 1. dataset.csv |
| We gathered research on the causes of accidents to ensure the dataset we chose included variables related to accidents in the oil and drilling industry. Our research shows that there are five common causes to accidents experienced by oil rig workers: Equipment malfunction, Lack of safety equipment, Weather, Explosions, Equipment maintenance. (Dunn | Sheehan, 2022)  Other causes include: Transportation Incidents, Exposure to Dangerous Substances and Slips and Falls (Escobedo, 2020) and improper processes (BOP Team, 2021). As such, we were able to find a dataset that included as many of the above variables to ensure that this was relevant to accidents found in the oil and drilling industry.  With ‘report.number’ being our Y variable, we explore which cause was the leading cause for accidents in the oil and drilling industry. We would then look for datasets relevant to equipment failure to conduct analysis to build prediction models and compare and predict the most accurate model to predict equipment failure. |
| 1. predictive\_maintenance.csv |
| We gathered research on hypothetical predictive maintenance datasets to test proof of concept of a predictive equipment failure model. Being able to formulate an effective model would allow us to apply the same methodology onto a similar dataset focused on the oil and drilling industry.  With ‘target’ being our categorical Y variable, to effectively model and make predictions based on the data, we conducted Logistic Regression, Random Forest Classifier and CART. |

## Data Sources

We obtained our datasets from Kaggle, a renowned platform within the data science community. However, its openness also necessitates scrutiny of the data sources to ensure their reliability and relevance. Thus, we examined the background of the data provided, ensuring that our selected datasets originated from reputable sources. Leveraging data from credible sources is crucial, as it is the foundation of our analysis.

## Dataset

|  |  |
| --- | --- |
| *database.csv:* | The dataset is named *dataset.csv* and is a database of each oil pipeline leak or spill reported to the Pipeline and Hazardous Materials Safety Administration since 2010 recorded by US Department of Transportation’s Pipeline and Hazardous Materials Safety Administration. |
| *predictive\_maintenance.csv:* | The dataset is named *predictive\_maintenance.csv* and is a dataset reflecting predictive maintenance data from the milling industry, with the data being compiled by University of California Irvine. |

### Y variable

|  |  |
| --- | --- |
| *database.csv:* | The Y variables we are using are continuous in nature. ‘report.count’ is a frequency based variable which allows us to get the count of oil pipeline spills based on the X variable. We also identified other continuous Y variables, such as ‘all.costs’ and ‘net.loss..barrels’, which indicates the cost and loss incurred respectively. |
| *predictive\_maintenance.csv:* | The Y variable we identified is ‘target’. ‘target’ is binary and indicates equipment failure, where ‘0’ = ‘No Failure’, and ‘1’ = a form of equipment failure. ‘1’ can indicate different forms of failure, but is irrelevant as we are simply looking for any form of failure, hence we used ‘target’ as Y variable instead of ‘Failure Type’ |

### X variable

|  |  |
| --- | --- |
| *database.csv:* | The X variable we identified as the relevant predictor is ‘Cause.Category’, a factor with character strings that categorises the reason behind each pipeline spill, which are corrosion, excavation damage, incorrect operation, material/weld/equipment failure, natural force damage, other outside force damage and all other causes. The other variables such as ‘property.damage.costs’ in the dataset are not relevant as the X variable as they merely add further contextual information about the oil pipeline leaks. |
| *predictive\_maintenance.csv:* | We took the rest of the variables in the dataset as X variables as they were all relevant predictors. They are: air temperature, process temperature, rotational speed [rpm], torque [Nm], tool wear [min]. The data dictionary can be found in [**Appendix A**](#_1we4wyswjzvo). The dataset contains variables that cover some conventional factors resulting in equipment degradation and eventual failure. |

## Data Cleaning

Data cleaning is essential in data analysis, as it ensures that the dataset’s quality is upheld and that subsequent analyses are based on accurate and consistent information, to derive meaningful insights.

The table below is a summary of the overall data cleaning process for dataset.csv:

|  |  |  |
| --- | --- | --- |
| **Variable** | **Action** | **Reasoning** |
| Categorical variables | Data type changed to “factor” ([Appendix B](#_mi3jjeydvtzh)). | Factorizing the data allows R to interpret the nature of our data better, helping us facilitate our analysis. |
| All Variables | Checked for NA values ([Appendix C](#_cd0lf7rqvnqu)). | Identifying where the NA values lie facilitates our decision-making process: to replace them or remove their corresponding rows. |
| Checked for and dropped duplicated rows  ([Appendix D](#_dshh3ngsejks)) | Removing duplicate rows ensure each record in our dataset is unique and meaningful to our analysis. |
| Public.Evacuations | Replaced NA with 0 | Given these columns represent counts, an NA likely indicates a count of zero, such as no injuries or no costs incurred. |
| VariousInjury/Fatality/Cost variables |
|
| Accident.City | Remove row with NA value | Dataset contained only one such row. Since its removal has minimal impact on our analysis, we decided to remove it. |
| Pipeline.Shutdown, Liquid.Ignition, Liquid.Explosions | Replaced ‘YES’ and ‘NO’ with “1”, “0” respectively | Formats each of these categorical variables from characters into integers, making them easier to analyse |

For *predictive\_maintenance.csv*:

|  |  |  |
| --- | --- | --- |
| **Variable** | **Action** | **Reasoning** |
| Categorical variables | Data type changed to “factor” ([Appendix B](#_mi3jjeydvtzh)). | Factorises data into a factored data type which allows R to better understand the nature of the data and support us in our analysis. |
| All Variables | Checked for NA values. | Helps us find NA values that may either have to be replaced with the median or removed.([Appendix C](#_cd0lf7rqvnqu)). |
| Checked for and dropped duplicated rows | This helps us find and remove duplicated rows for further analysis.([Appendix D](#_dshh3ngsejks)) |
| Air\_Temperature  Process\_Temperature  Rotational.speed..rpm.  Torque..Nm.  Tool.wear..min. | Created boxplots  ([Appendix E](#_u01xe8hb3a78)). | This would help us to check for potential outliers that may result in distortions in our findings from modelling. |
|
|
|
|
| Type | Replaced L, M, H with “3”, “2”, “1” respectively | Converting Type to categorical labels with numerical values streamlines our analysis, making it more conducive for future modelling techniques ([Appendix E](#_yr8vvcs1rw24)). |
| Air.temperature..K., Process.temperature..K. | Subtracted 273.15 from all values | This formats each of these variables from integers in Kelvin into degree celsius, making for easier visualisation. ([Appendix F](#_dshh3ngsejks)). |
| UDI, Product.ID, Failure.Type | Dropped columns | This removes irrelevant data columns that are not required for analysis ([Appendix G](#_ai27w1cinyq8))  We dropped Failure Type as the focus of our analysis is on predicting machine failure occurrence, not type of failure. |

*Predictive\_maintenance.csv* is well-maintained, with an absence of NA values and duplicate entries. However, there were outliers, as seen in [Appendix E](#_u01xe8hb3a78), that have the potential to skew data distributions. Considering the context of our analysis, which is to predict the occurrence of machine failure, we decided to retain the outliers as they could encapsulate crucial information that could be indicators of machine failure.

# Data Exploration

In order to have a better idea of the relationship between the variables, we have performed various plots between the accident cause category and various Y variables.

## database.csv

### Categorical X Variables against Continuous Y Variables

​​Firstly, utilising a line graph ([Appendix J](#_es94u0ta9qkk)) to examine the data, we observed a consistent upward trend in pipeline incidents from 2011 to 2015, reaching its peak in 2015. It's worth noting that we excluded incidents in 2017 from our analysis due to limited data, as there were only two incidents reported in early January, which do not reflect the entire year.

Subsequently, through bar charts ([Appendix J](#_es94u0ta9qkk)), our analysis identified equipment failure as the primary cause of accidents, resulting in the most shutdowns and significant financial and material losses, including oil.

Clearly, material, weld, and equipment failures are the primary contributors to pipeline incidents, which is why our focus will centre on preventing equipment failure through predictive maintenance.

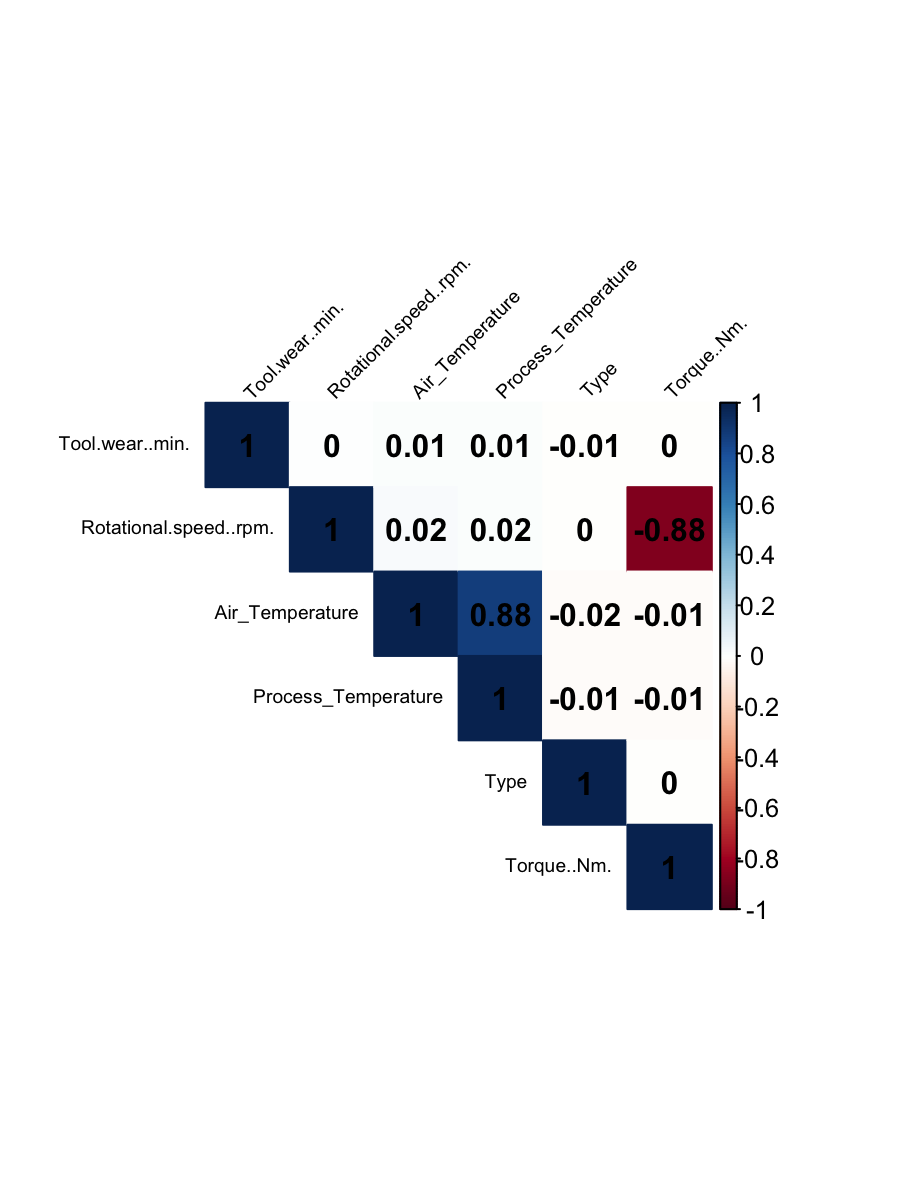
### Continuous Y Variables against Continuous Y Variables

We used a scatter plot to compare the relationship between continuous Y variables and other continuous Y variables. From our charts, it shows that all 3 graphs ([Appendix K](#_din1dqkuhk0o)), net loss (barrels) vs all costs, pipeline shutdowns vs all costs and liquid ignition vs all fatalities, all have a positive relationship. Pipeline shutdowns vs all costs has the least correlation out of the 3 graphs while net loss (barrels) and all costs was the most correlated.

## predictive\_maintance.csv

### Continuous X variables against continuous X variables

To discern relationships among continuous variables, we used scatter plots, with each plot aimed to examine the relationship dynamics between each pair of variables ([Appendix L](#_kya6oc3mgt7r)). We observed a strong negative relationship between torque and rotational speed. Also, there was a strong positive correlation between process temperature and air temperature. This is consistent with our correlation matrix heatmap, as shown below.



## Sampling Methods

During data preparation, we plotted a histogram on the frequency of equipment failure ([Appendix M](#_zdf64x4e7c4c)), which showed a clear imbalance in data. There was a stark contrast of 9,661 entries of failure, and only 339 entries of No failure. In order to balance the data and allow for a fairer representation in the training process, we went with 2 methods- Stratified Sampling which creates balanced subsets of the data, and Undersampling to reduce the number of instances in the majority class (Brownlee, 2021). We also maintained a 70:30 train-test ratio to allow for a better estimate of performance of predictive models, as studies showing that the split yields the best results (Gholamy et al., 2018).

### Stratified Sampling

Using the ‘createDataPartition’ function in ‘caret’, we split the dataset into a training set and a test set while maintaining a similar distribution of the target variable in both. ([Appendix N](#_9ahmdz330w2u))

After splitting the data, the results show a similar distribution between training and test sets, indicating a good stratification of the split.

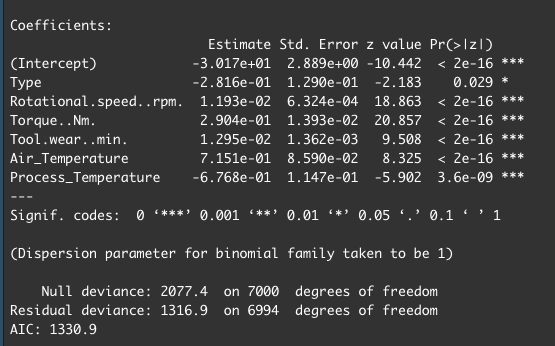
### Undersampling

By sampling rows from the majority class to match the number of rows in the minority class, we are able to get the same number of rows to resolve the data imbalance issue. Using the ‘rbind’ function , we combine the 2 sets and do a sample split of 70:30. ([Appendix O](#_mjoi1xivr7kk))

# ModelTraining & Evaluation Methods

## Logistic Regression Model

### Logistic Regression - Stratified Sampling

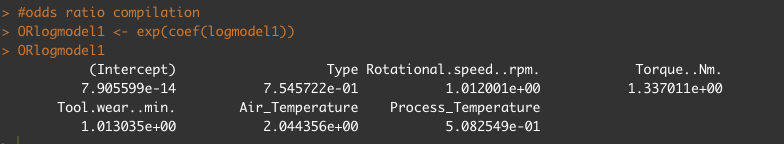


Our team decided to use logistic regression because of its suitability in predicting binary outcomes (machine failure). Since logistic regression is adept at identifying significant relationships between predictor variables and a binary target variable, by leveraging the model, we aim to pinpoint the most influential factors that in machines that could lead to machine failure. The insights gained will then guide our predictions.

#### Using AIC To Determine Optimal Logistic Model

After conducting AIC-based backward stepwise elimination to refine our logistic regression model, the model remained unchanged, suggesting that all the predictors we used were statistically significant and contribute meaningfully to predicting the outcome, indicating that our model is not overfitted with redundant predictors.

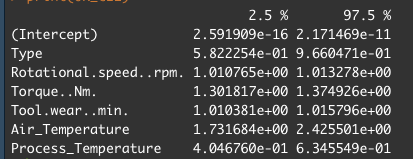
#### Odds Ratio and Confidence Interval of OR



Computing the odds ratio (OR) offers a comprehensive understanding of the influence each predictor has on the outcome.

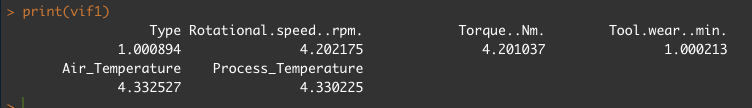
* An odds ratio greater than 1 suggests that as the predictor’s value rises, the likelihood of machine failure occurring also increases
* Conversely, an odds ratio less than 1 suggests that as the predictor’s value increases, the likelihood of machine failure occurring decreases.

From our analysis of OR, in this case, Air Temperature has the highest odds ratio, with a value of 2.044, indicating that with each unit increase, the odds of machine failure more than doubles, all else being equal.

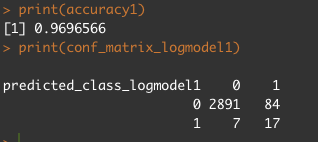


The confidence interval of the odds ratio gives us an expected range for the true odds ratio for the population to fall within.

#### Preventing Multicollinearity Using VIF



To ensure no multicollinearity in our logistic models, we employed the Variance Inflation Factor (VIF) test, by fitting a linear model first. Generally, for continuous predictors, when VIF is higher than 10, there is significant multicollinearity that needs to be corrected (Kim, 2019). Given the VIF values of our predictor variables, there is no need for correction.



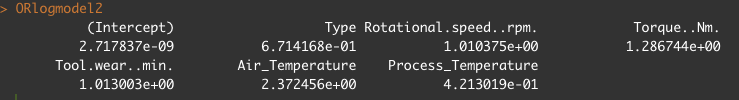
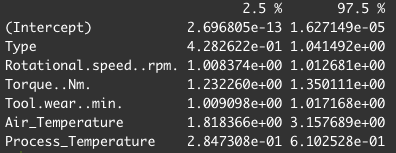
We proceeded to test the accuracy of the model on the testset data, giving us an accurate evaluation of the model. Our model has a 97.0% accuracy of predicting equipment failure.

### Logistic Regression Model - Undersampling

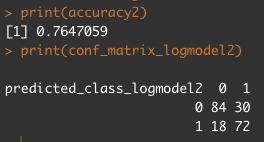
## 

To ensure that our model was not overfitting to the imbalanced data set, we also tested the model with balanced data, using undersampling. We also performed AIC-based backward step elimination, and used VIF to ensure no multicollinearity. We found that our model was robust, and all predictor variables were significant in predicting the target.

* + 1. Odds Ratio and Confidence Interval of OR

From our analysis of the Odds Ratio, Air Temperature still had the highest influence on predicting machine failure.



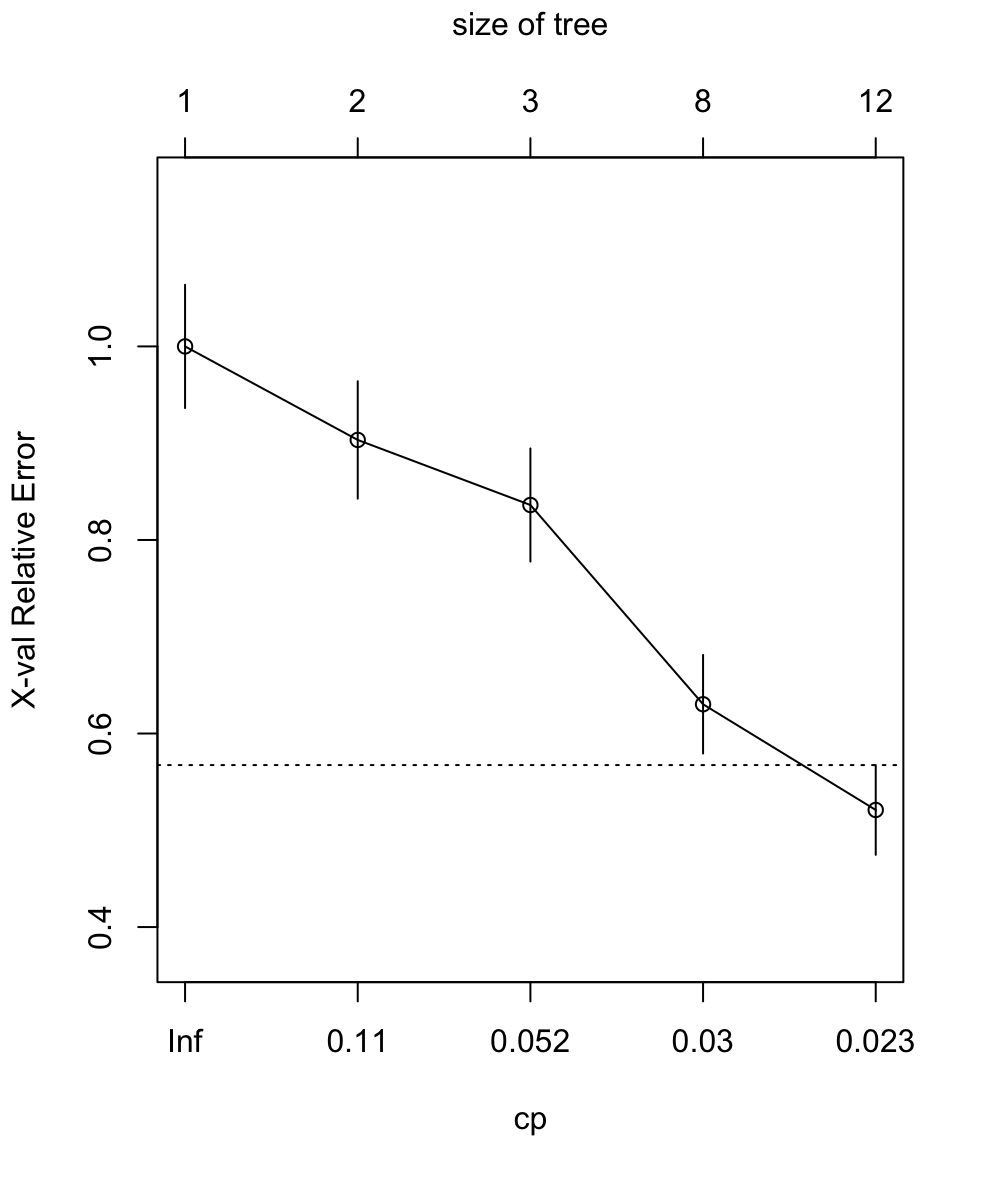
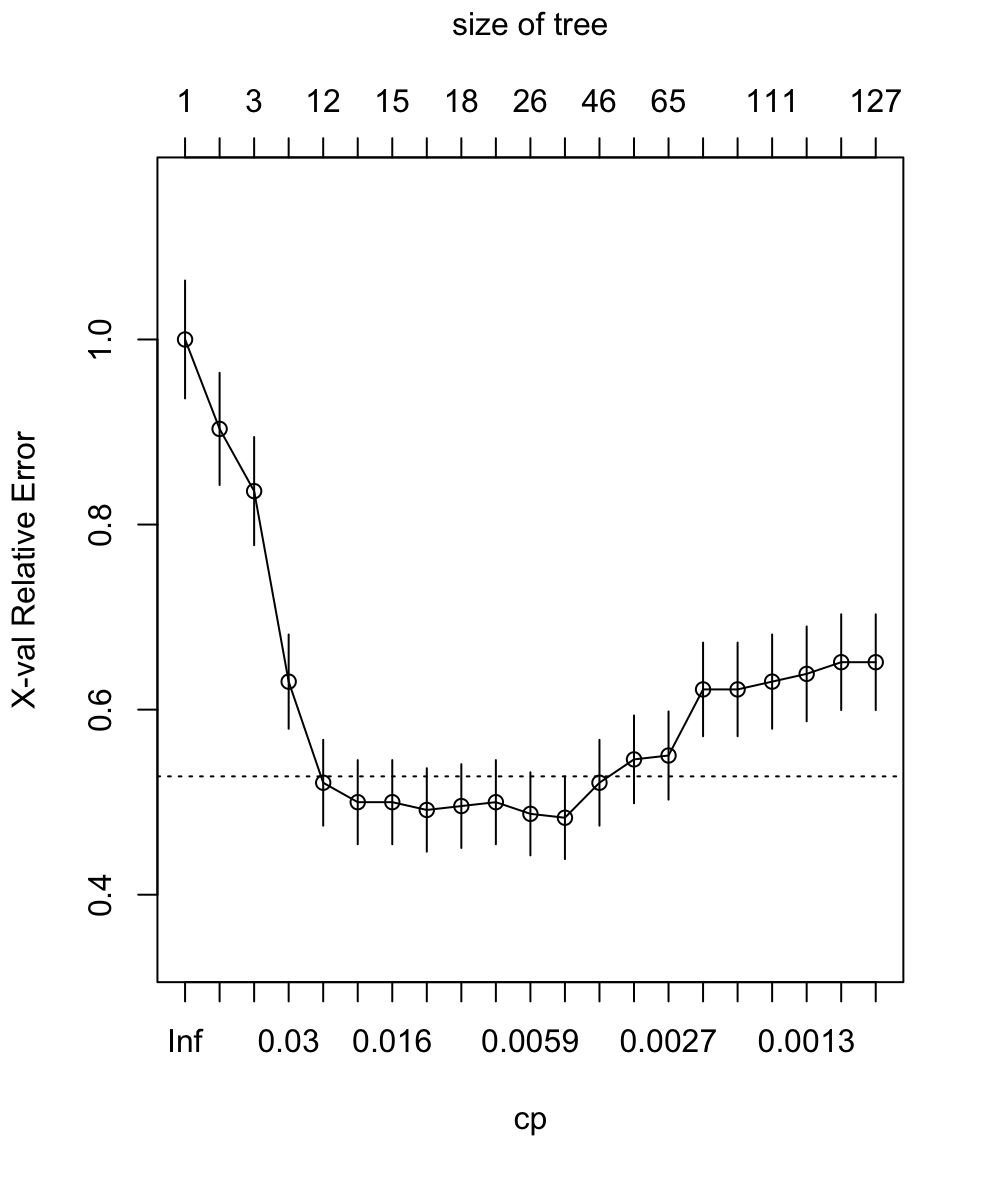
We proceeded to test the accuracy of the model on the b\_testset, giving us an accurate evaluation of the model. Our undersampled model has a 76.5% accuracy of predicting equipment failure.

## Classification and Regression Tree Model (CART)

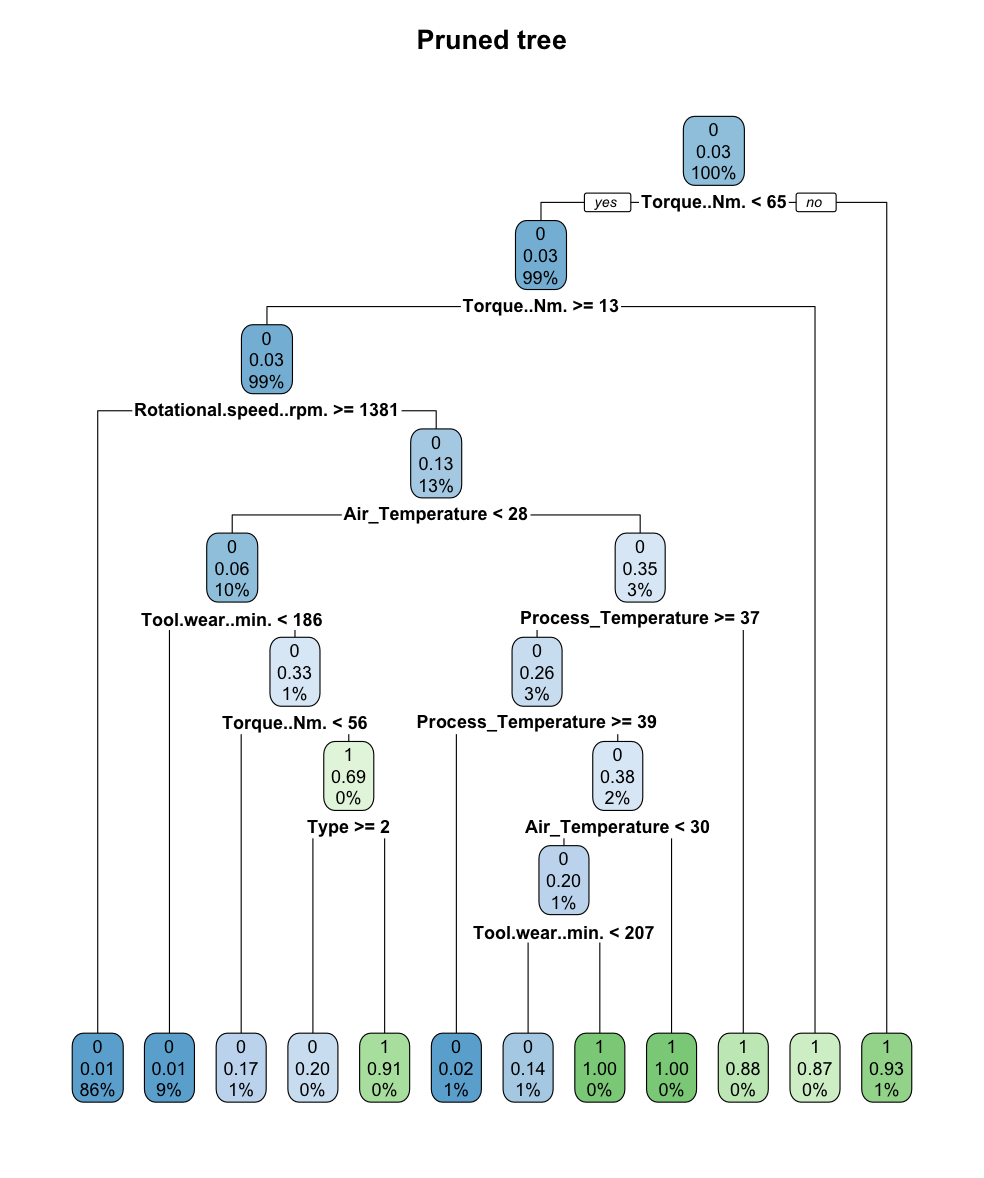
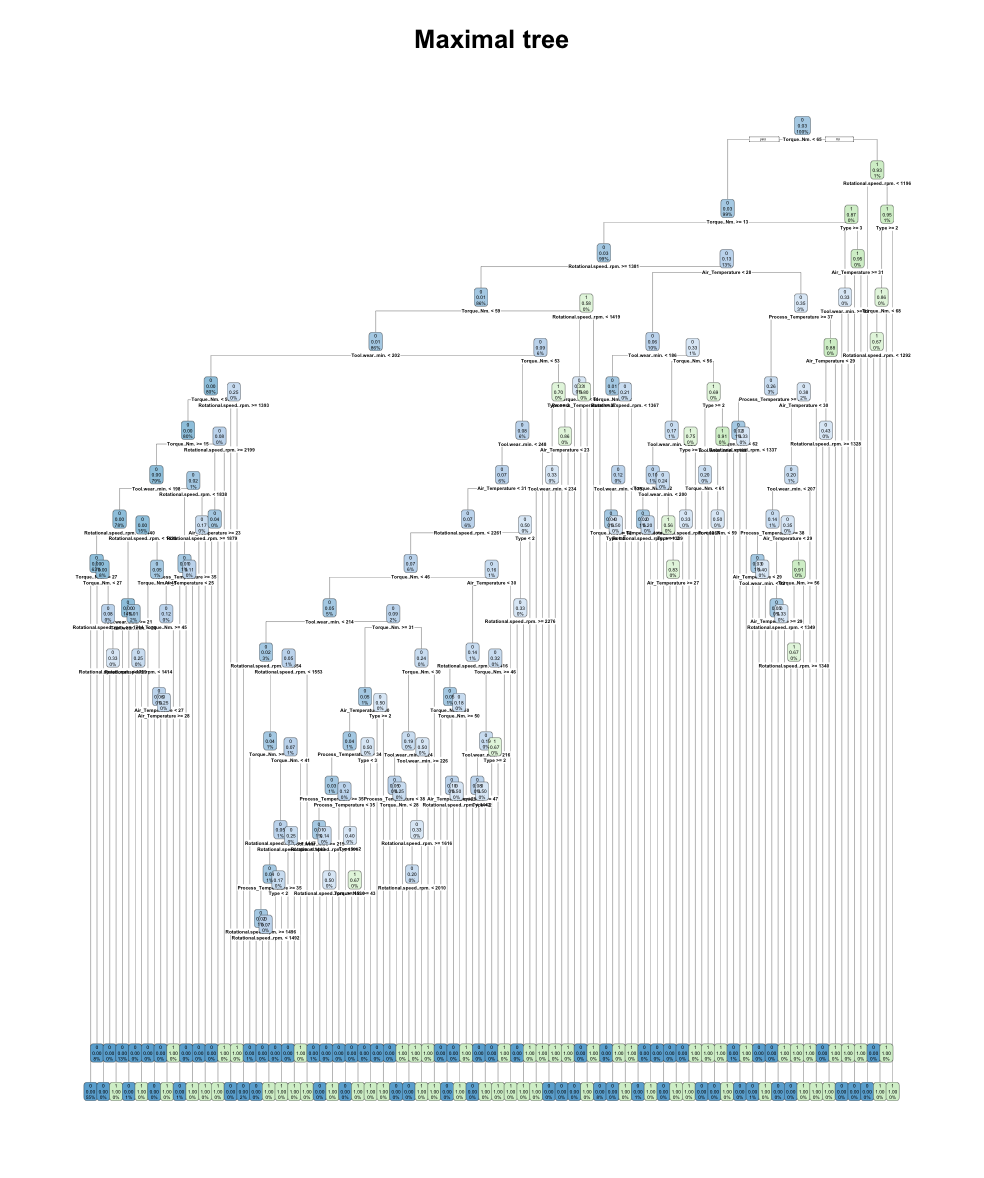
We chose the CART model for our analysis to uncover key variables that contribute to equipment failure, aiming to enhance the predictive maintenance strategy. The CART model is well-suited for this purpose due to its ability to handle complex, non-linear interactions among variables. Moreover, the model's tree structure allows us to easily interpret and explain the decision-making process, which is vital for implementing practical maintenance interventions.

### CART Model - Stratified Sampling

We started our analysis by using the trainset and testset from stratified sampling for the CART model, then fully expanding the decision tree, limiting premature pruning by setting the complexity parameter (cp) to 0. We then used 'class' as the method as our variables are categorical. We then set minsplit as 2 considering our dataset's size, designed to strike a balance between the risks of overfitting and underfitting. Upon constructing the tree, we proceeded to prune it to optimize model simplicity without significantly compromising its discriminative capacity. This pruning reduced the tree's complexity to an optimal level, as indicated by the chosen cp value of 0.02183257 ([Appendix P](#_u7fj8ohctfxw)).



The plotcp graphs reveal the cross validation results. The first graph shows that the 5th node will give us the most optimal tree where it first crosses below the line. From the second plot, we can see that there are only 5 nodes left after deleting the child nodes of a branch node.



Visualisation of full tree vs pruned tree

We can observe that the unpruned tree is much larger than the pruned version. The final pruned tree has 12 terminal nodes, compared to the many in the maximal tree which has the possibility of being overfitted.

### Performance Evaluation

### Confusion Matrix ([Appendix Q](#_xu7b0j6utizw))

After pruning the tree, we finalise our model. By using the confusionMatrix() function, we are able to find the following:

|  |  |  |  |
| --- | --- | --- | --- |
| True Positives (TP): | Decreased | True Negatives (TN): | Increased |
| False Positives (FP): | Increased | False Negatives (FN): | Decreased |

We can see that the pruning has improved the model in this regard as there are less false negatives. However, the increase in false positives indicates a trade-off as the model now predicts more instances as potential failures. The pruned model's reduced specificity also indicates that it may have lost some ability to correctly classify the minority class (positives)

### Accuracy Percentage

After pruning the tree, we finalise our model. By using the trainset to rest the accuracy on the accuracy, we have a reduced accuracy from 98% to 97.8%. ([Appendix Q](#_xu7b0j6utizw))

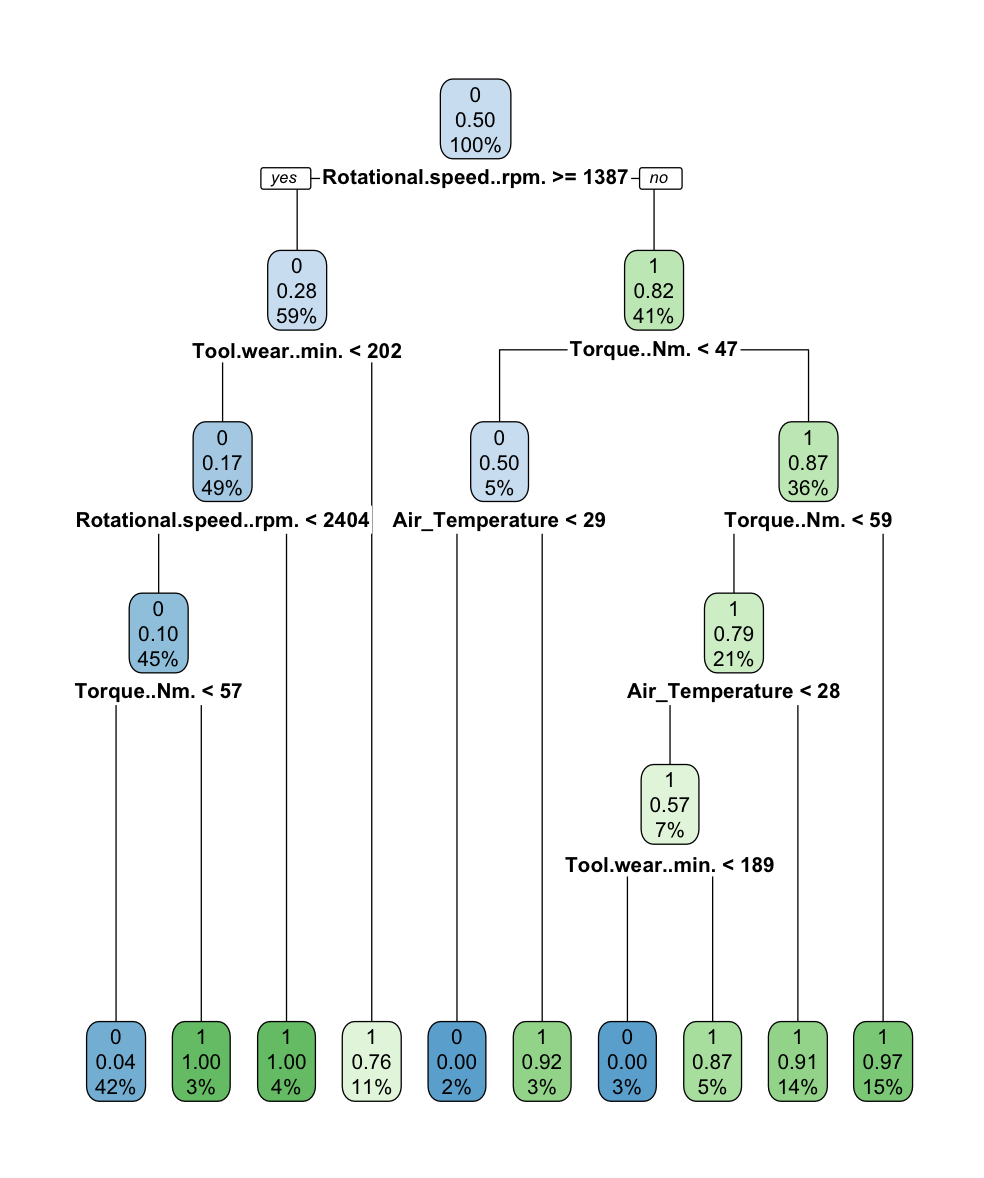
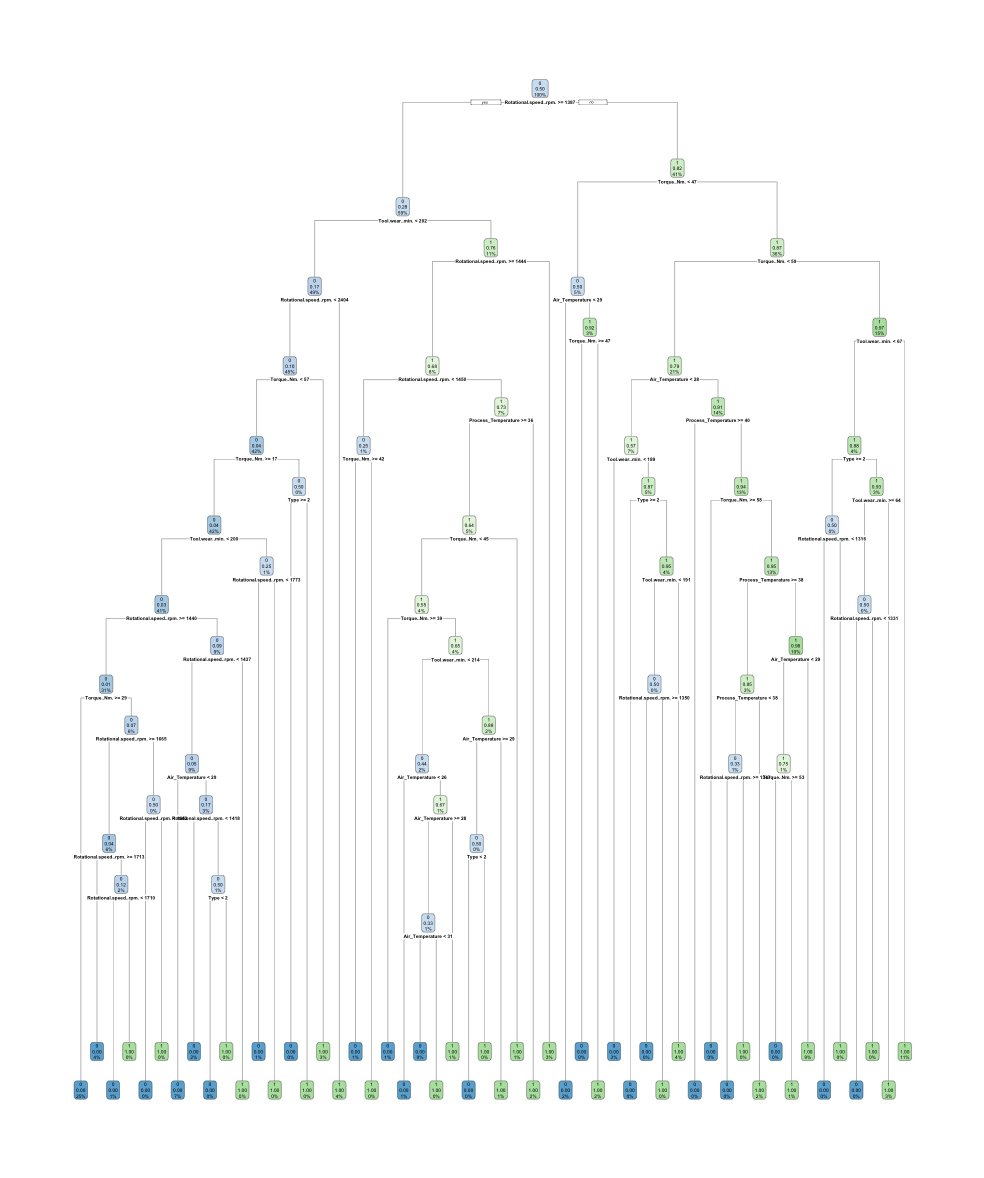
### Cost Sensitive CART Model ([Appendix R](#_xu09okuoc4fs))

Due to the low sensitivity towards the minority class, we decided to utilise Cost Sensitivity learning to increase said sensitivity towards positives. From applying the cost-sensitivity learning, we are able to see that while accuracy reduces further to 96.9%, there is an increase in specificity as well as the number of true and false positives identified. There is a tradeoff as there are more false positives found, it is more justifiable as less false negatives are identified. Furthermore, the weightages can be changed to vary the effect of the cost sensitive learning model.

### CART Model - Undersampling

For this model, we utilised the trainset and testset from undersampling instead, we then conduct the same process of expansion of decision tree as well as pruning process as the CART modelling using stratified dataset. The pruning reduced the tree's complexity to an optimal level, as indicated by the chosen cp value of 0.01193429 ([Appendix S](#_n0w68yl0dw8j)).

The plotcp graphs reveal the cross validation results. The first graph shows that the 7th node will give us the most optimal tree where it first crosses below the line. From the second plot, we can see that there are only 7 nodes left after deleting the child nodes of a branch node.



Visualisation of full tree vs pruned tree

Similarly, we can observe that the unpruned tree is much larger than the pruned version. The final pruned tree has 10 terminal nodes, compared to the many in the maximal tree which has the possibility of being overfitted.

### Performance Evaluation

### Confusion Matrix ([Appendix T](#_rv5l5p99qvt2))

After pruning the tree, we finalise our model. By using the confusionMatrix() function, we are able to find the following:

|  |  |  |  |
| --- | --- | --- | --- |
| True Positives (TP): | Increased | True Negatives (TN): | Decreased |
| False Positives (FP): | Decreased | False Negatives (FN): | Increased |

We can see that the pruning has improved the model considering that apart from true negatives, the model was able to correctly identify more actual positives, and had fewer instances of incorrectly predicting cases. The reduction in true negatives was also only by 1 instance. In this case, specificity increased from 82.35% to 95.10% indicating an improvement at correctly identifying negative cases.

### Accuracy Percentage

After pruning the tree, we finalise our model. By using the trainset to rest the accuracy on the accuracy, we have a reduced accuracy from Increased from 85.78% to 91.67%. ([Appendix T](#_rv5l5p99qvt2))

### Cost Sensitive CART Model ([Appendix U](#_2g20s7kkwgb5))

While there was a reduction in false negatives identified by the model after pruning, it is imperative that we try to minimise the number of false negatives falsely identified. As such, we decided to utilise Cost Sensitivity learning. Through this, we were able to see that while accuracy reduces further to 90.6%, there is a significant reduction of occurrences of false negatives identified. We were able to hence optimise the model to only falsely identify negatives 2 times.

# Comparison Of All Models

In order to decide the best model for predicting machine failure, we ran both logistic regression and CART in 3 different scenarios.

|  |  |
| --- | --- |
| Scenario 1 | CART and Logistic Regression with stratified training-test split  (logmodel1 and prunedcart1) |
| Scenario 2 | CART and Logistic Regression with undersampling  (logmodel2 and prunedcart2) |
| Scenario 3 | Both CART Models with Cost-Sensitive learning  (costsensitive1 and costsensitive2) |

### Accuracy Percentage

First, we used R to evaluate our performance metrics - the accuracy, sensitivity, specificity, positive predictive value and negative predictive value.

These were done with the following formulas from our confusion matrices.

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Sensitivity = TP / (TP + FN)

Specificity = TN / (TN + FP)

Positive Predictive Value = TP / (TP + FP)

Negative Predictive Value = TN / (TN + FN)

Below is the table, derived from our calculations in R ([Appendix V](#_6iuqaasnwiju))

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Accuracy: | Sensitivity: | Specificity: | Positive Predictive Value: | Negative Predictive Value: |
| (LogModel1) | 96.97% | 70.83% | 97.18% | 16.83% | 99.76% |
| (LogModel2) | 76.47% | 80.00% | 73.68% | 70.59% | 82.35% |
| PrunedCart1 | 97.80% | 79.66% | 98.16% | 46.53% | 99.59% |
| PrunedCart2 | 91.67% | 88.99% | 94.75% | 95.10% | 88.24% |
| CostSensitive1 | 96.93% | 53.72% | 98.75% | 64.36% | 98.07% |
| CostSensitive2 | 90.69% | 85.47% | 97.70% | 98.04% | 83.33% |

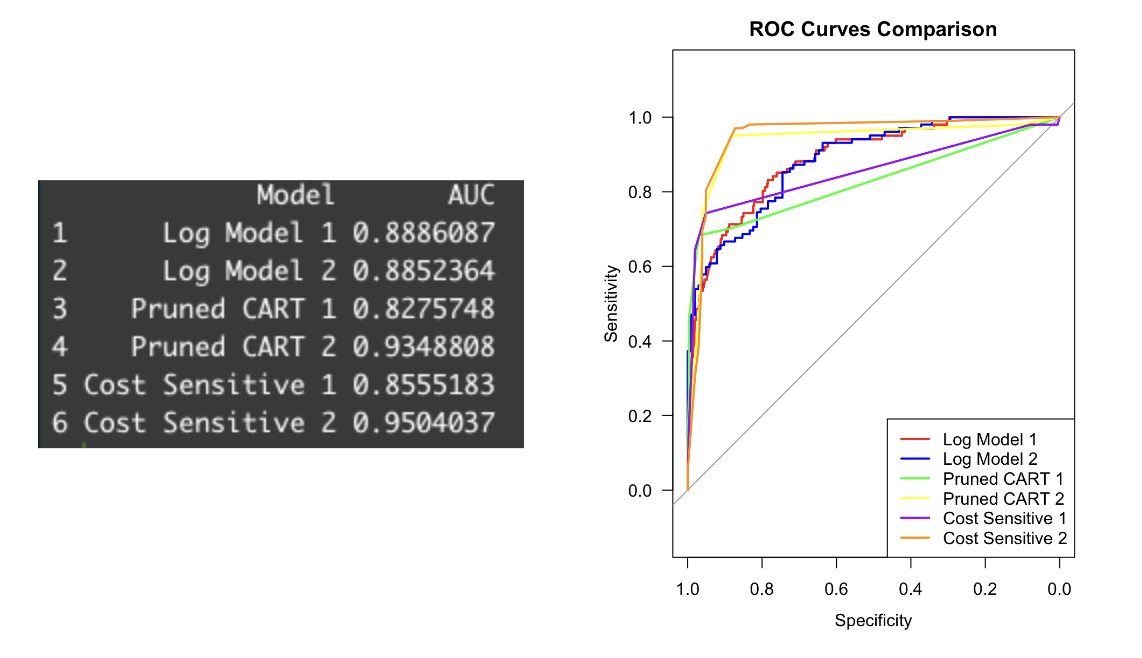
We observe that based on accuracy, LogModel1, PrunedCart1 and CostSensitive1 performed well, with accuracy values of over 96%. However, these values may be misleading, as in an imbalanced dataset, a naive model that predicts the majority class for every instance can achieve a high accuracy. This is especially prevalent in our dataset, with only 3.39% of our observations being positive, and is confirmed by the poor positive predictive values of all 3 models.

In the context of our business problem, due to the high costs associated with equipment failures, the detection of positive cases is the most important factor in our analysis, as failure to predict machine failure can lead to catastrophic consequences.

Thus, we need to look at more comprehensive metrics to evaluate our models.

### AUC-ROC Curve

To do further comprehensive evaluation of our models, we used the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) curve. The ROC curve will provide us with the representation of our classifiers’ performance across different thresholds, plotted with the True Positive Rate (TPR) on the y-axis and the False Positive Rate (FPR) on the x-axis. We computed the AUC for all 6 models, and tabulated them as shown below.

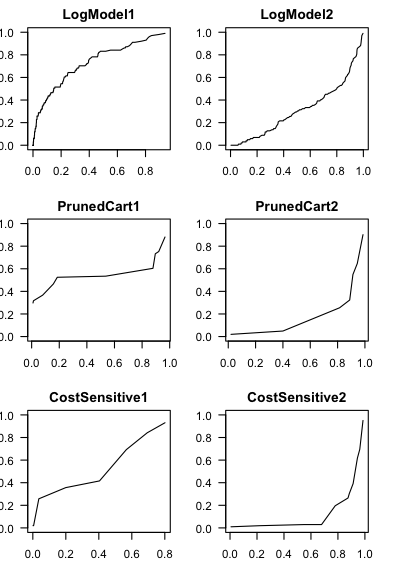


An AUC of 1.0 indicates a perfect classifier, indicating that the model has achieved perfect separation between the positive and negative cases. From our AUC computation, we observe that while all models performed fairly well, cost-sensitive2, the CART model made utilising undersampling and cost-sensitive learning performed the best, as it excels in distinguishing between the two classes - machine failure and non machine failure, with an AUC value of 0.9504037.

The above plot serves as the graphical representation of the diagnostic ability of our classifier systems as the discrimination threshold is varied. This corroborates our findings, as cost-sensitive 2 had the best performance when we analyse the ROC curves.

### False Negative Rate Plots

Using the ROC Curve, we plot the False Negative Rate rate against the thresholds, by using the formula FNR = 1 - TPR. Below is the graphical representation of our results. The x-axis represents the different thresholds, and the y-axis represents the FNR.



Our models exhibit variability in their FNR across different thresholds. The CART models have relatively stable FNRs across the range of thresholds, while the logistic regression models exhibit more pronounced fluctuations.

In the context of our business problem, we would prefer a model with consistently lower FNR across most thresholds, as minimising false negatives is crucial for lowering pipeline accidents. As such, from the analysis of the plots, costsensitive2 is a prime candidate for predicting machine failure.

### Conclusion

Machine failures, especially in the oil and drilling industry, carry significant implications, making prediction of machine failures critical. While several of our models showcased commendable performances in their own aspects, the cost-sensitive2 model, leveraging both undersampling and cost-sensitive learning, proved to be the most robust and reliable. Its performance, evidenced by the various metrics and visual analyses, positions it as the most suitable in our business context.

# Recommendations and conclusions

## Collection of Data

First and foremost, we recommend Aramco to gather their own data and create a dataset in order to prepare the data input for prediction. Currently, Aramco only has sensors that monitor the build-up of pressure in pipelines and processing facilities to predict flaring of hydrocarbon gases (Aramco, n.d.-a). The data collected are condition-based, which essentially monitors current signs of combustion. However, to solve the issue, Aramco should switch to more predictive-based sensors that can monitor the factors leading to accidents that our team has explored. The predictive-based sensors can detect anomalies in equipment before they turn into equipment failures, compared to condition-based sensors detecting anomalies in performance which means that there is already equipment failure (Infineon Technologies, n.d.).

Aramco may completely change their entire system of sensors but that is highly impractical as they currently hold 18,000 data sources (Aramco, n.d.-a) and replacing them is extremely costly. Hence, they should instead upgrade the sensors such that they can gather relevant data on the factors that contribute most to equipment failure, and form a dataset to use for the next stage of our proposed solution.

## Better Harnessing of Digital Twin Technology

A digital twin is a digital representation of a physical object, person, or process, contextualised in a digital version of its environment (McKinsey, 2023). Aramco currently uses this technology but is mainly used to predict equipment breakdown through simulations ran to execute predictive maintenance (Aramco, n.d.-b). However they can improve its usage of this technology by collaborating more with equipment manufacturers. Sharing comprehensive operational data as well as predictive analysis from the digital twin technology gives crucial insights about how the equipment works under field conditions. Manufacturers are then able to develop new solutions that improve equipment design tailored specifically to the challenges faced by Aramco. Furthermore, the shared data helps in preparing specific maintenance schedules that reduce lag time, thereby improving the overall lifecycle of machinery.

Hence, if Aramco were to follow our proposed solutions, it will help them take huge steps towards increased efficiency in equipment maintenance as well as predictive data analysis, allowing them to accurately predict and prevent equipment failures leading to accidents.

# Appendices

## Appendix A: Data Dictionary

|  |  |  |
| --- | --- | --- |
| **Dataset 1. dataset.csv** | | |
| **Name** | **Type** | **Description** |
| Accident Year | Numeric | Year of accident. Ranging from 2010 - 2018. |
| Accident Date / Time | Character | Date and time of accident in the format:  YYYY-MM-DD, Hours:Minutes:Seconds |
| Pipeline Location | Character | Factor with 2 levels. Either Offshore or Onshore |
| Pipeline Type | Character | Factor with 5 levels. Type of pipeline for onshore pipeline |
| Liquid Type | Character | Factor with 5 levels. |
| Liquid Subtype | Character | Factor with 8 levels. |
| Liquid Name | Character | Factor with 68 levels. Only available when Liquid Subtype = OTHER HVL / OTHER. Otherwise the cell has NA. |
| Accident Latitude | Numeric | Measures the location’s distance north or south of the equator. |
| Accident Longitude | Numeric | Measures the location’s distance east or west of the equator. |
| Cause Category | Character | Category of cause of accident |
| Cause Subcategory | Character | Specific category of cause |
| Unintentional Release (Barrels) | Numeric | Amount of liquids released intentionally due to accident. Measured in barrels. |
| Intentional Release (Barrels) | Numeric | Amount of liquids released due to accident. Measured in barrels |
| Liquid Recovery (Barrels) | Numeric | Amount of liquid recovered in barrels |
| Updated Net Loss (Barrels) | Numeric | Corrected net loss measured in barrels.  Net Loss = Intentional Release + Unintentional Release - Liquid Recovery |
| Net Loss (Barrels) | Numeric | Uncorrected net loss measured in barrels from the dataset. |
| Liquid Ignition | Character | Factor with 2 levels. |
| Liquid Explosion | Character | Factor with 2 levels. |
| Pipeline Shutdown | Character | Factor with 2 levels. |
| Shutdown Duration (Hours) | Numeric | Shutdown Duration = Restart Date/Time - Shutdown Date/Time.  Restart Date/Time and Shutdown Date/Time are from the original dataset. |
| Public.Evacuations | Numeric | Factor with 2 levels. Identifies if evacuation was carried out. |
| All.Injuries | Numeric | Factor with 2 levels. Identifies if there were injuries. |
| All.Fatalities | Numeric | Factor with 2 levels. Identifies if there were fatalities. |
| Property Damage Costs | Numeric | Costs related to property damage. |
| Lost Commodity Costs | Numeric | Costs related to raw materials lost. |
| Public / Private Property Damage Costs | Numeric | Costs related to damage caused to properties that do not belong to the corporation. |
| Emergency Response Costs | Numeric | Costs related to emergency response to the accident. |
| Environmental Remediation Costs | Numeric | Costs related to restoring the environment after the accident. |
| Other Costs | Numeric | Other costs. |
| All Costs | Numeric | Summation of all the other costs mentioned above. |

|  |  |  |
| --- | --- | --- |
| **Dataset 2. predictive\_maintenance.csv** | | |
| **Name** | **Type** | **Description** |
| UDI | Numeric | Unique identifier of product |
| Product.ID | Character | Specific serial number to each product that also serves as a unique identifier |
| Type | Character | Described as L (Low: 50% of all products), M (Medium : 30%), or H (High: 20%) as product quality variants and a variant-specific serial number |
| Air temperature [K] | Numeric | Temperature generated by using a random walk process later normalised to a standard deviation of 2 K around 300 K |
| Process temperature [K] | Numeric | Temperature generated by using a random walk process normalised to a standard deviation of 1 K, added to the air temperature plus 10 K |
| Rotational speed [rpm] | Numeric | The rotational speed calculated from machine power of 2860 W, overlaid with a normally distributed noise |
| Torque [Nm] | Numeric | Amount of rotational force the motor develops.  The values are then normally distributed around 40 Nm with an infinite limit of 10Nm and no negative values. |
| Tool wear [min] | Numeric | The amount of time in the gradual failure of cutting 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 respectively. |
| Target | Numeric | ‘0’s and ‘1’s indicating failure of equipment |
| Failure Type | Character | Indication of failure, specifically type of failure, or the lack thereof |

## 

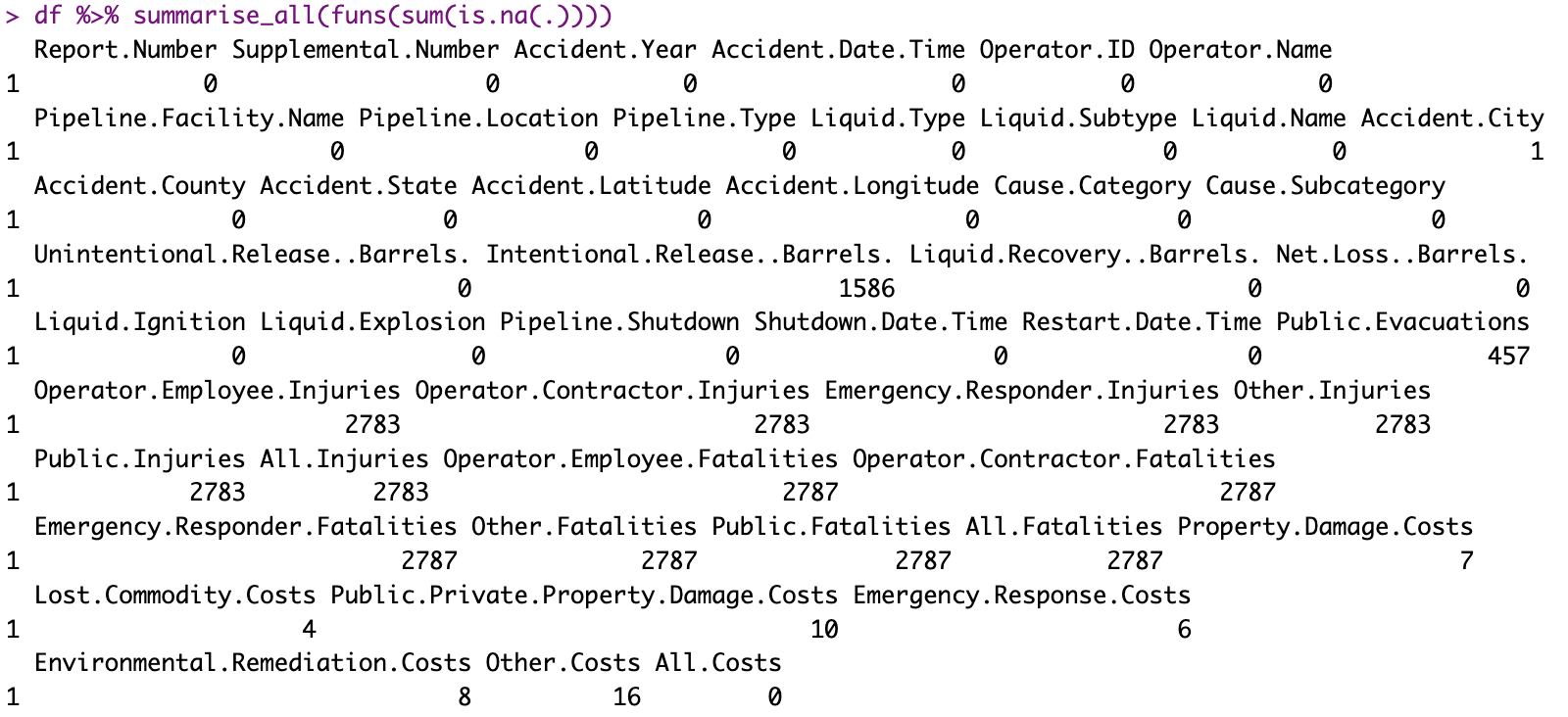
## Appendix B: Factorising Data

For *predictive\_maintenance.csv*:



|  |  |  |
| --- | --- | --- |
| Variables | Trend | Correlation coefficient |
| tool wear vs air temperature | As tool wear increases, air temperature increases slightly since it has a low correlation coefficient of 0.01 | positive |
| tool wear vs process temperature | With a correlation coefficient of 0.2, as tool wear increases, process temperature increases with gentle gradient | positive |
| tool wear vs type | As tool wear increases, type decreases with very slightly increasing gradient | negative |
| rotational speed vs air temperature | As rotational speed increases, air temperature increases with a gentle gradient | positive |
| rotational speed vs process temperature | As rotational speed increases, process temperature increases with a gentle gradient | positive |
| rotational speed vs torque | As rotational speed increases, torque decreases almost linearly | highly negative |
| air temperature vs process temperature | As air temperature increases, process temperature increases almost linearly | highly positive |
| air temperature vs type | As air temperature increases, type decreases | negative |
| air temperature vs torque | As air temperature increases, torque decreases | negative |
| process temperature vs type | As process temperature increases, type decreases | negative |
| process temperature vs torque | As process temperature increases, torque decreases | negative |

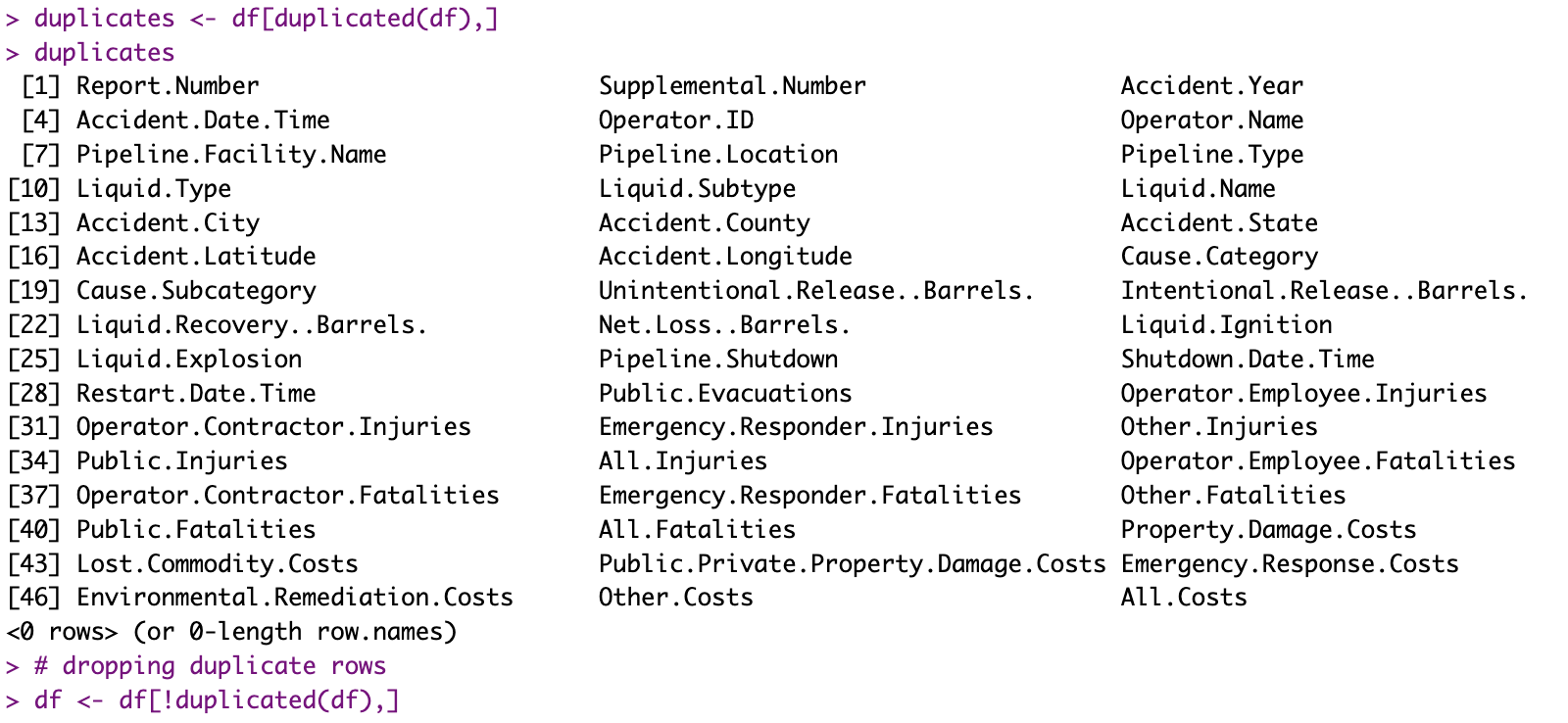
## Appendix C: Checking for NA values

For *dataset.csv*: 

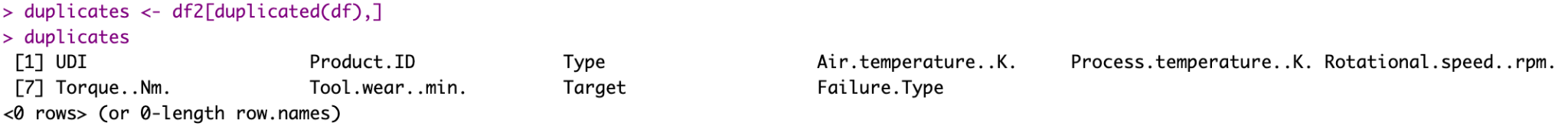
For *predictive\_maintenance.csv*:



## Appendix D: Checking and removing duplicated rows

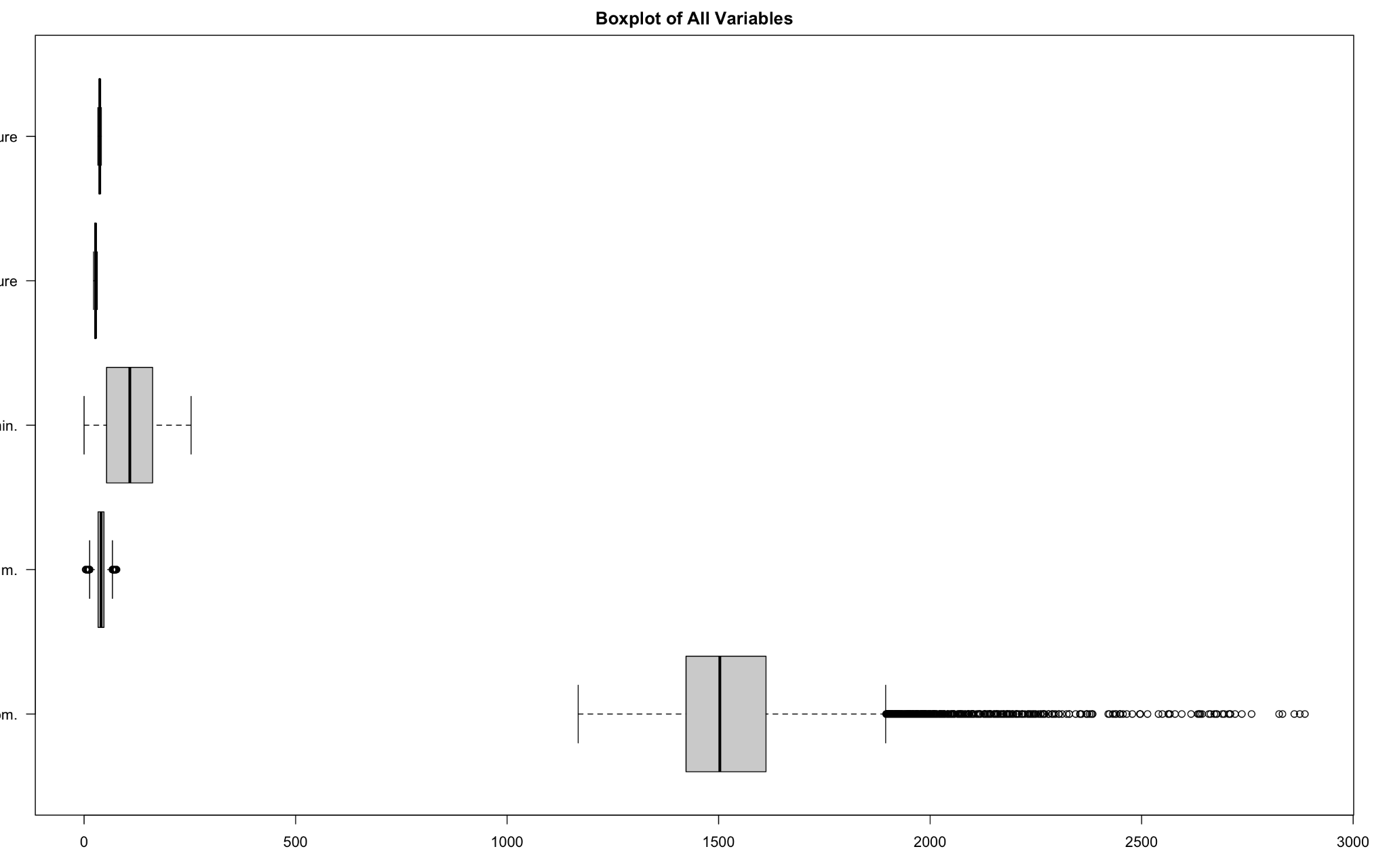
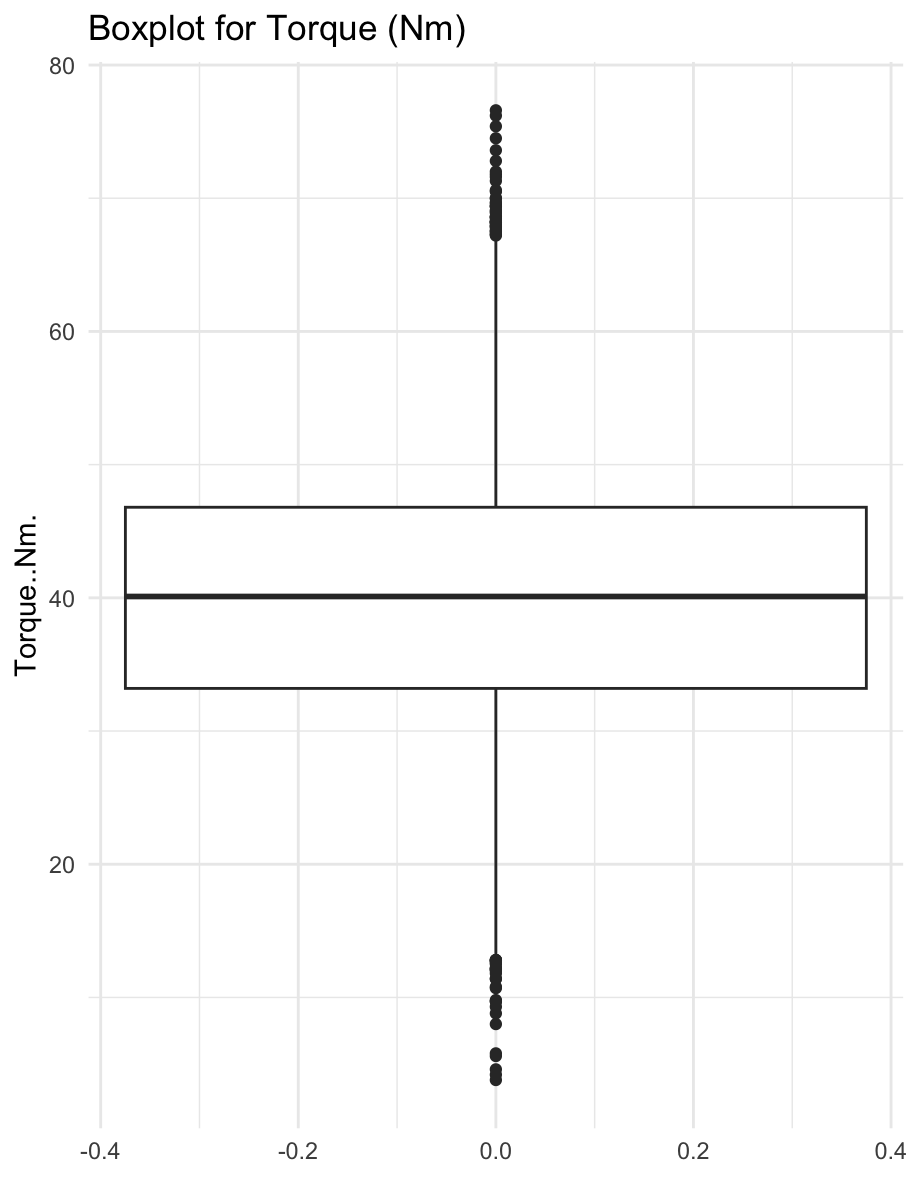
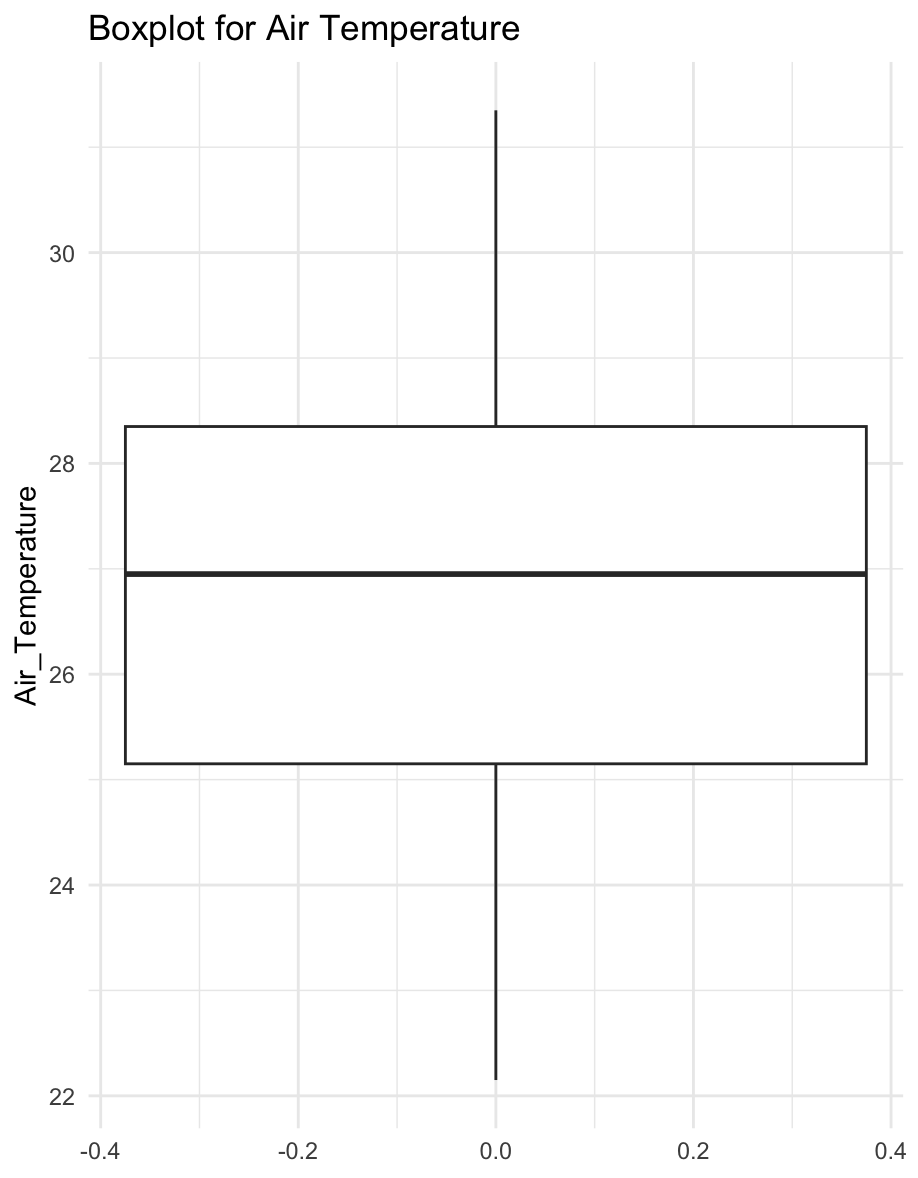
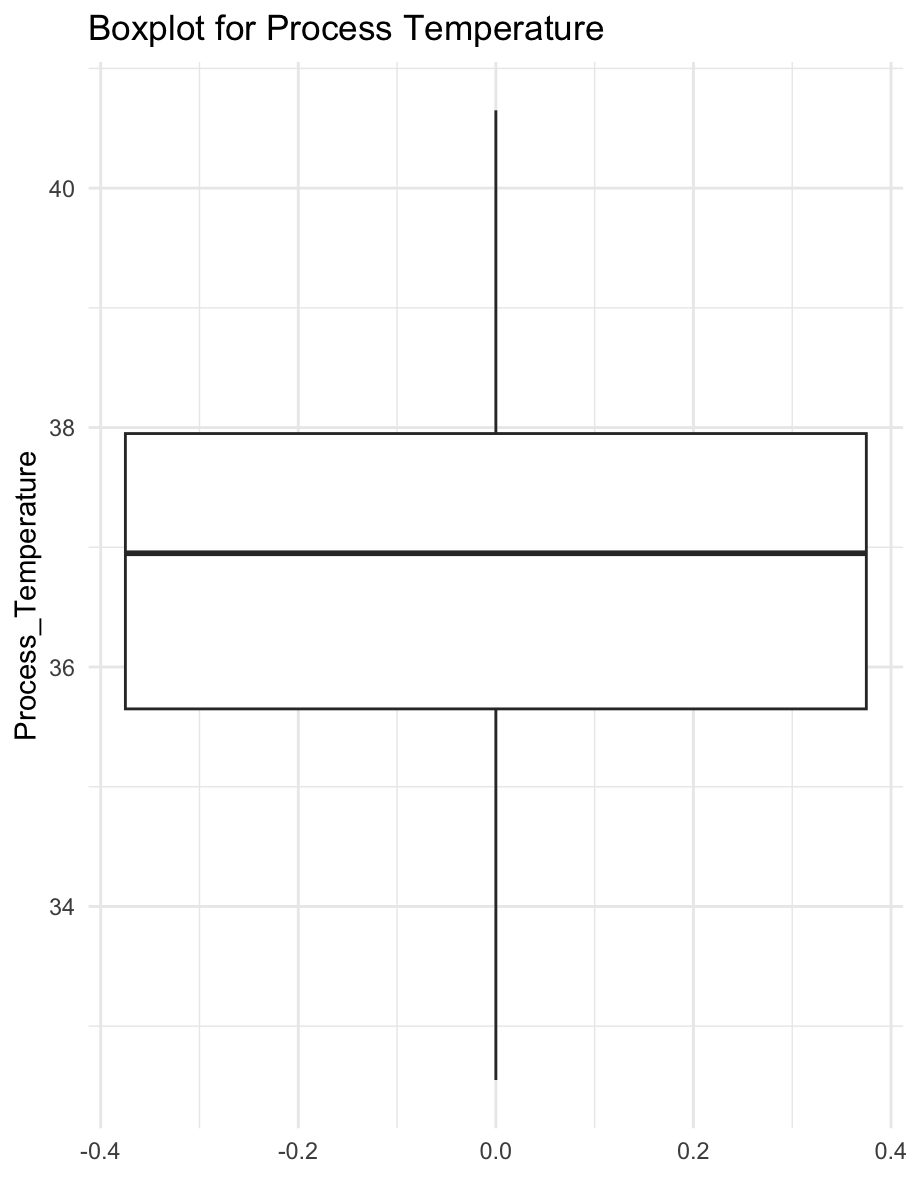
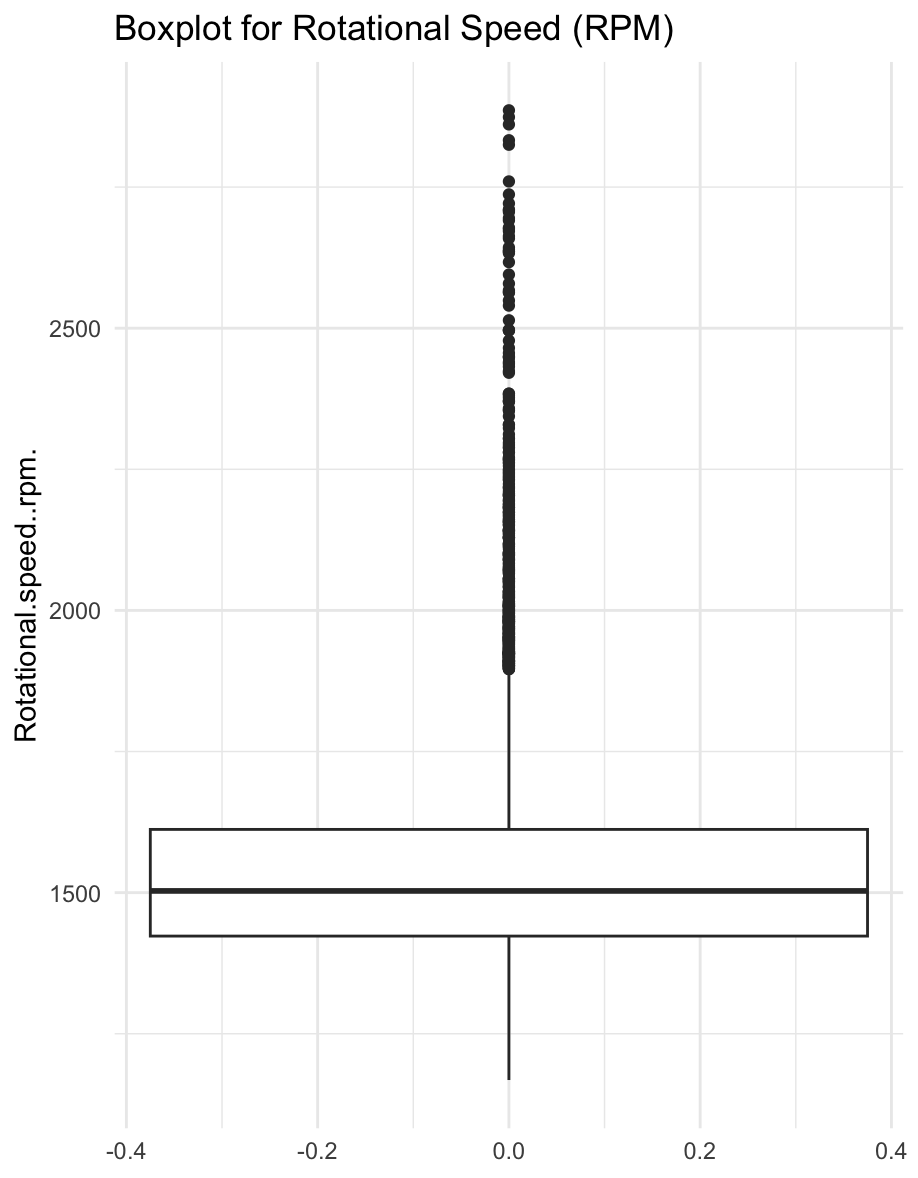
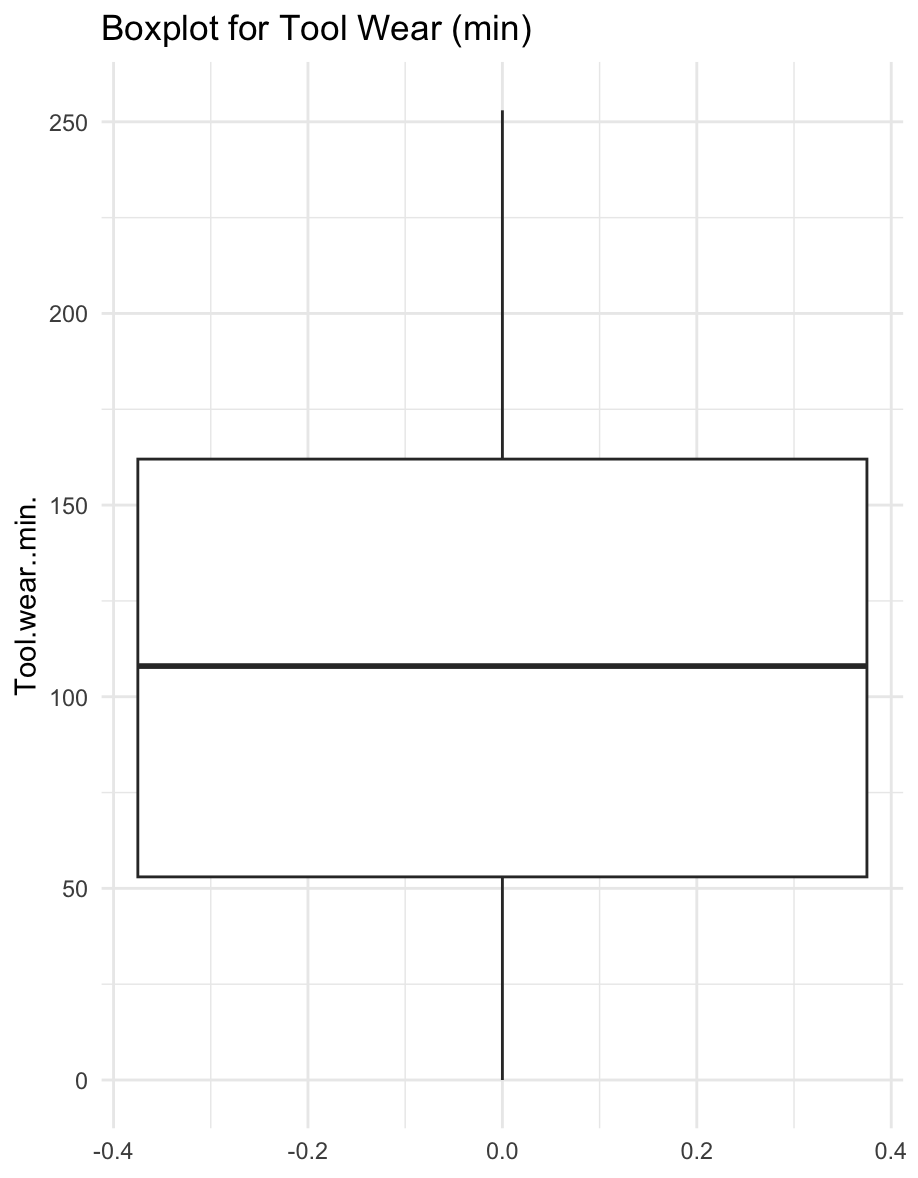
For *dataset.csv*: 

For *predictive\_maintenance.csv*:



## 

## Appendix E: Box Plot of Continuous Variables

For *predictive\_maintenance.csv*:

## Appendix F: Factorising and Renaming Columns

For *predictive\_maintenance.csv*:

## 

## 

## Appendix H: Restructuring Variables from K to °C

For *predictive\_maintenance.csv*:

## Appendix I: Dropping Columns

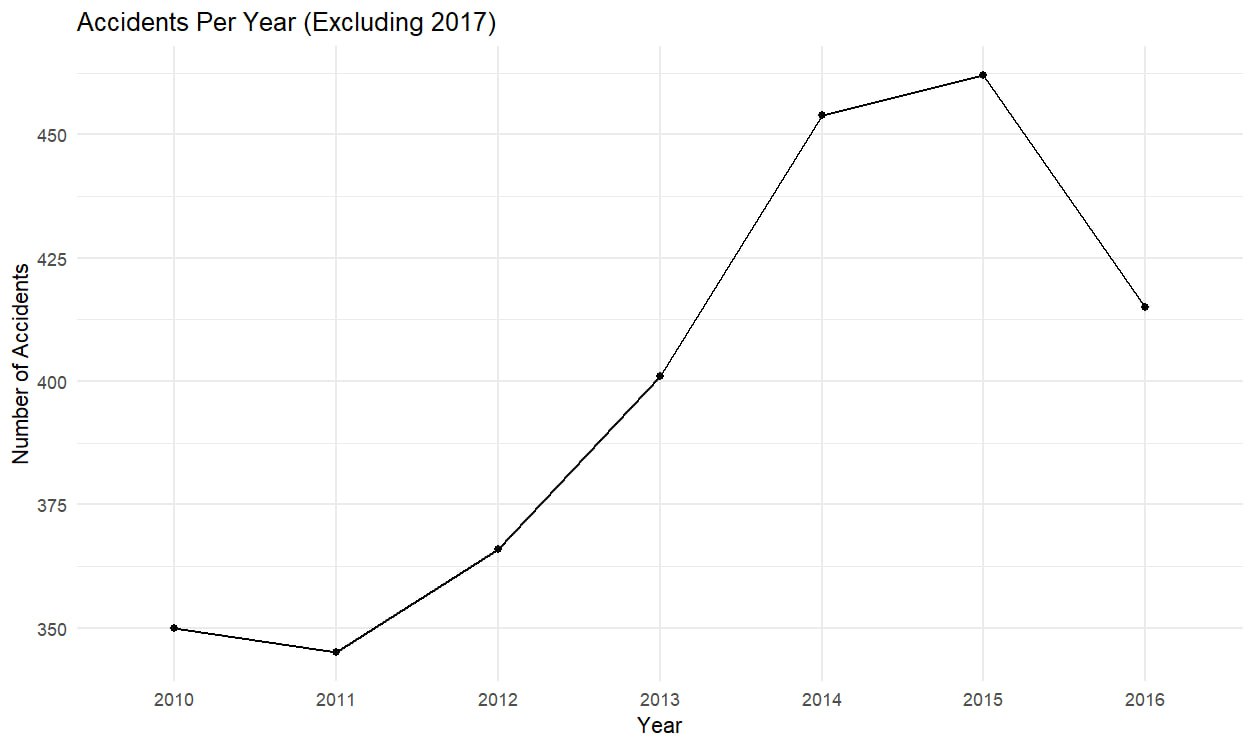
For *predictive\_maintenance.csv*:

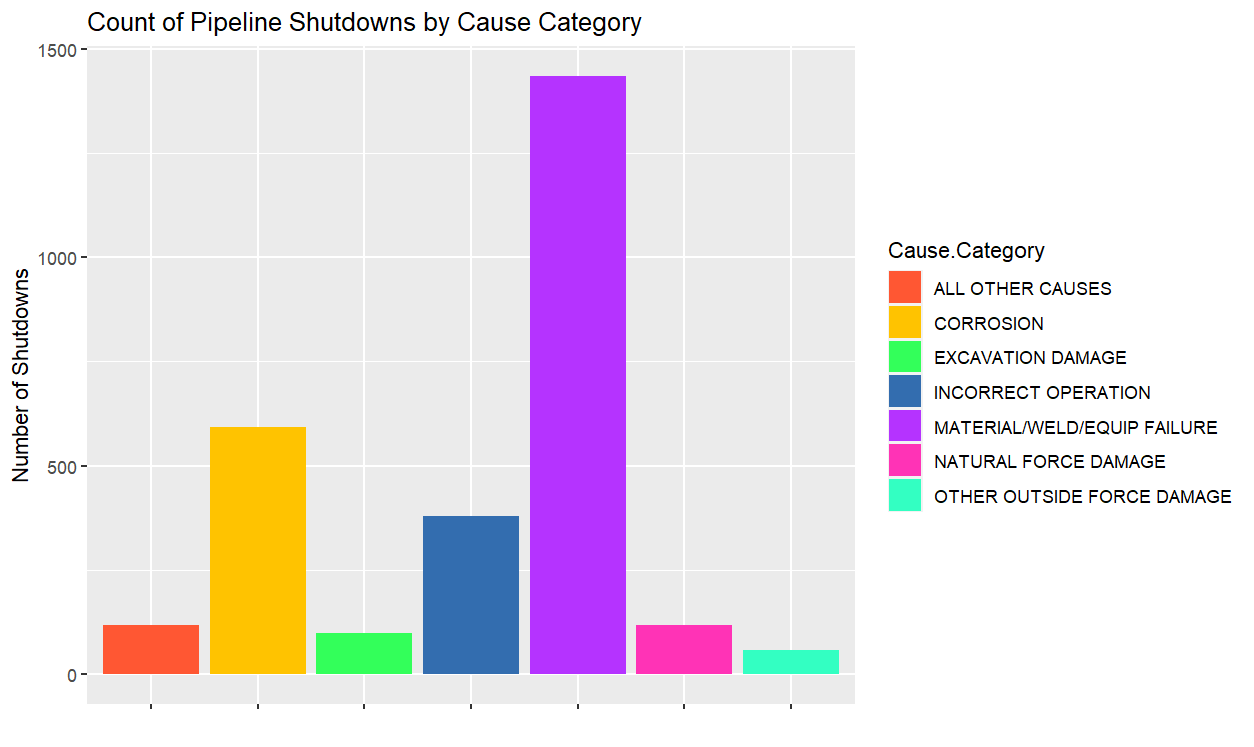
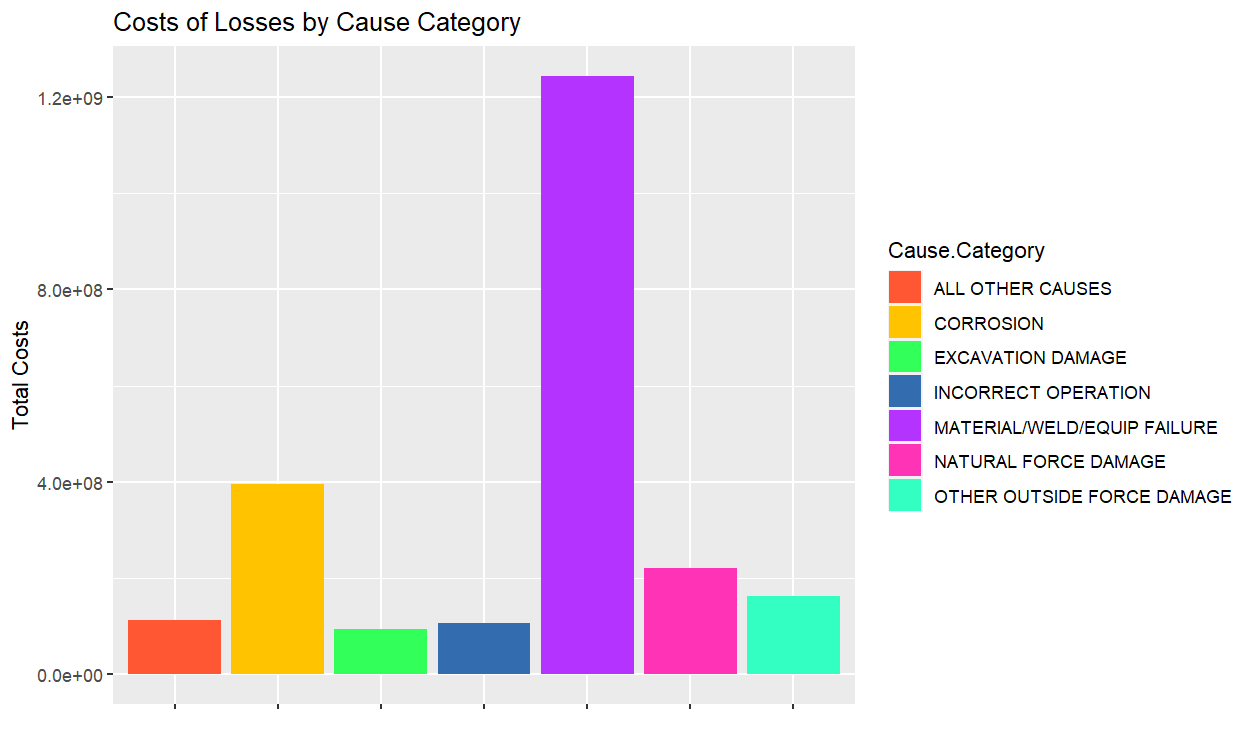
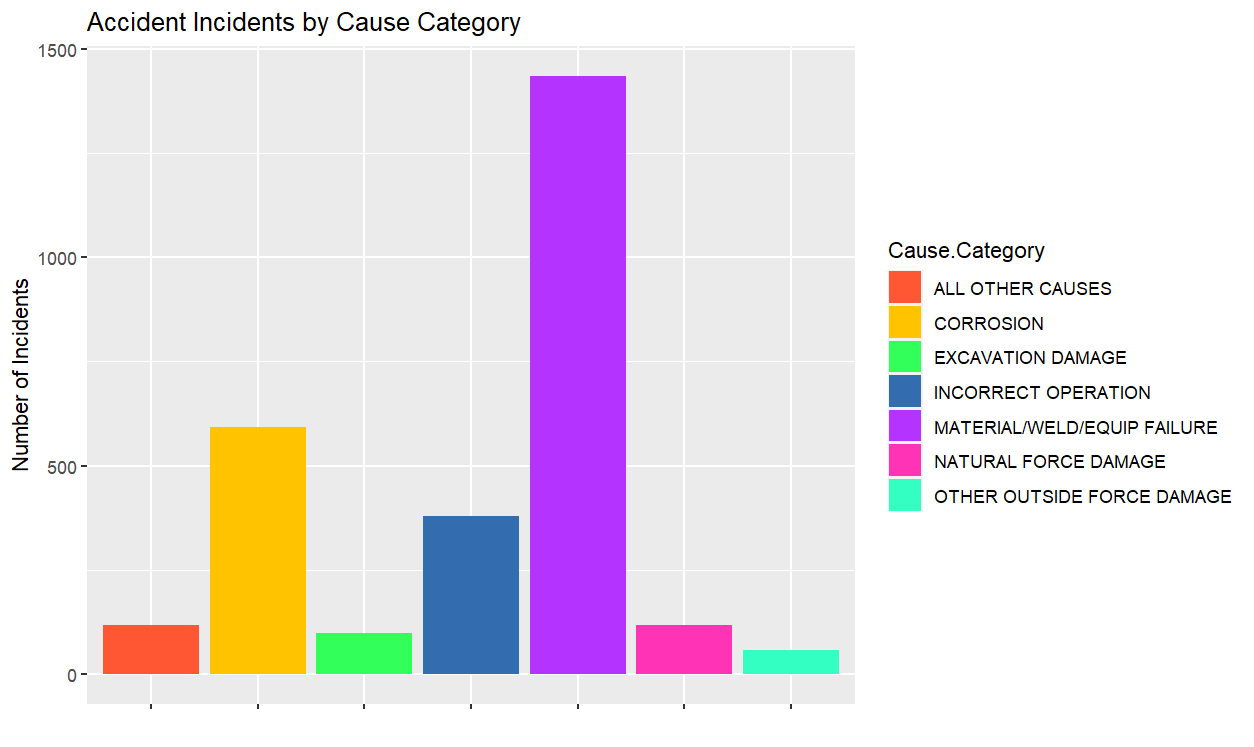


## 

## Appendix J: Relationship Between Categorical X Variables & Continuous Y Variable

For *dataset.csv*:

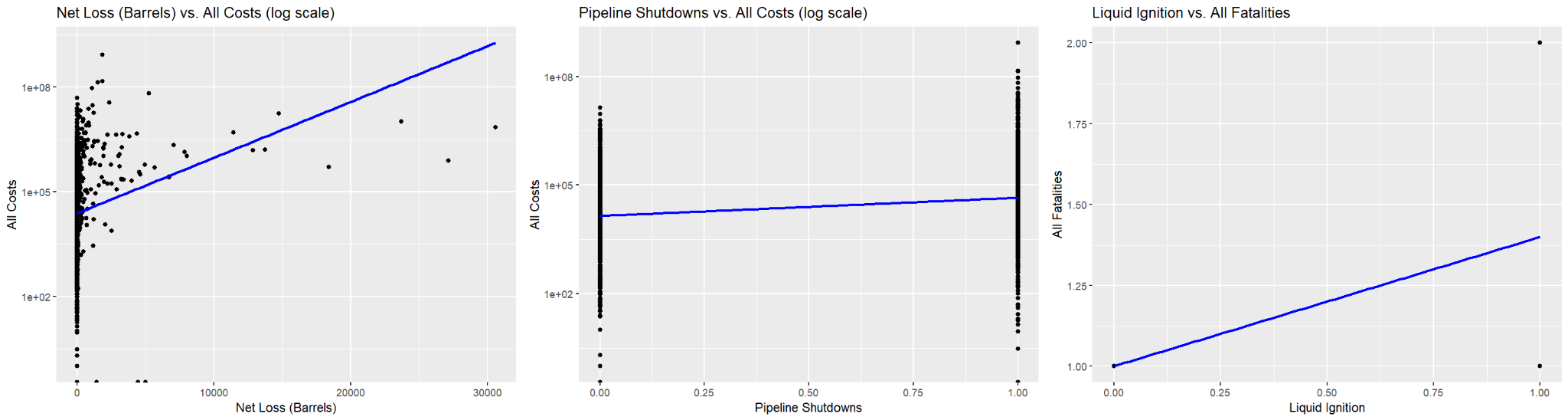




## 

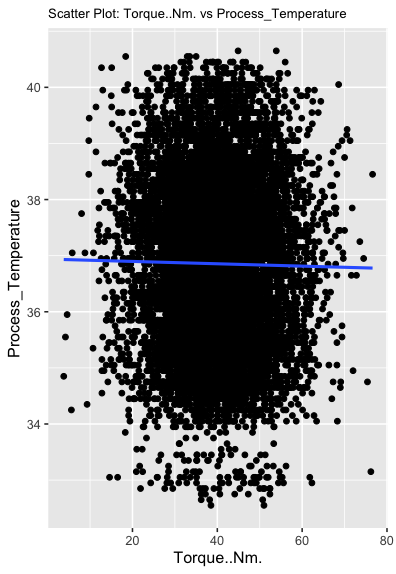
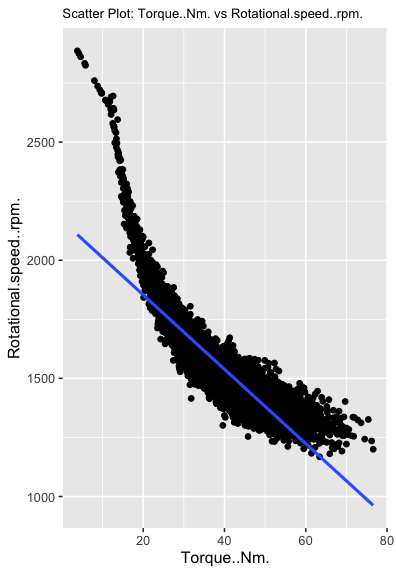
## Appendix K: Relationship Between Continuous Y Variables and Continuous Y Variable

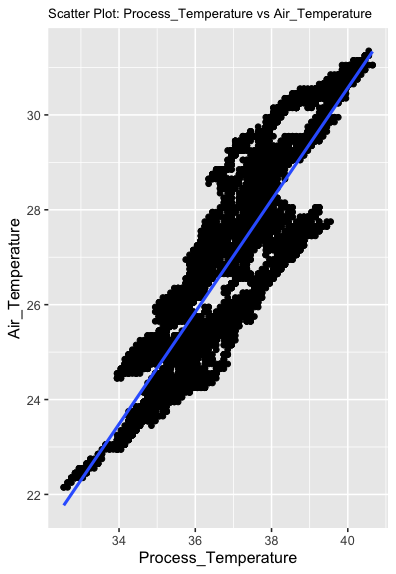
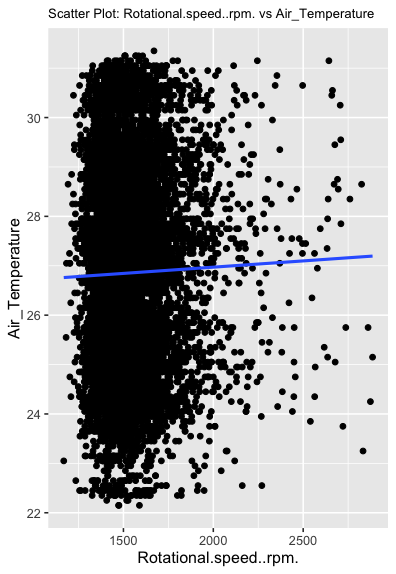
For *dataset.csv*:



## 

## Appendix L: Relationship Between Continuous X Variables



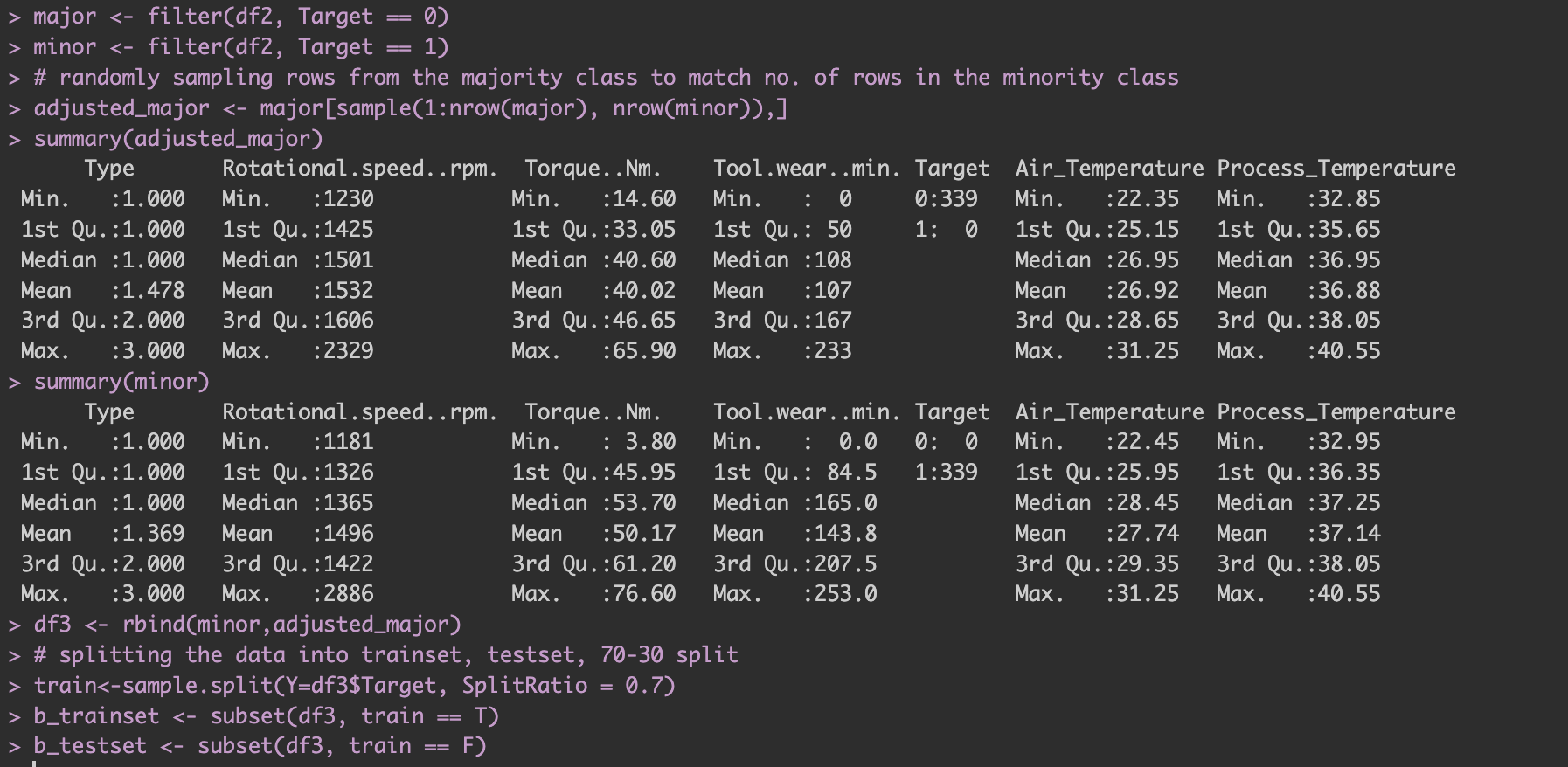


## Appendix M: Histogram on the Frequency of Failure

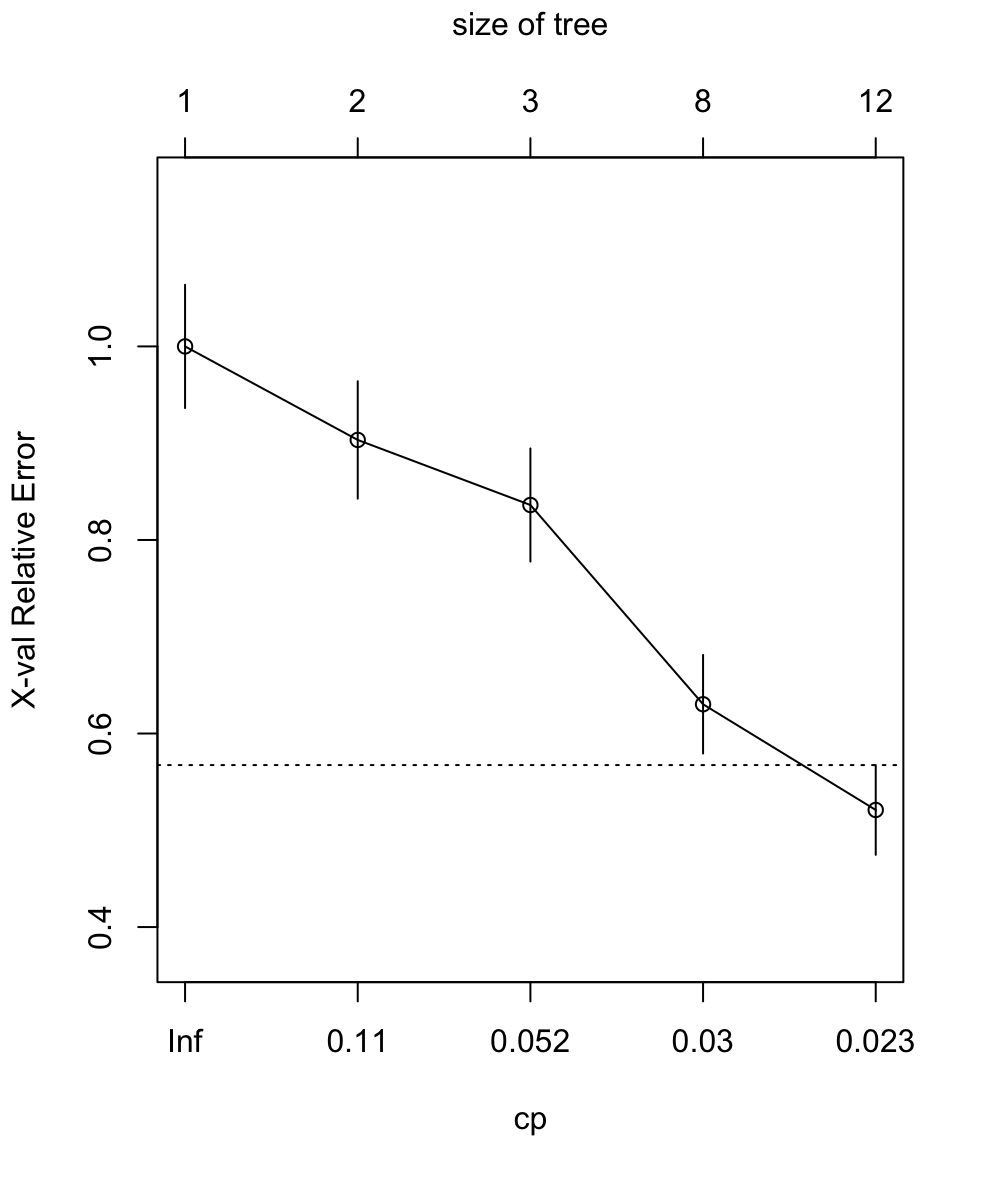
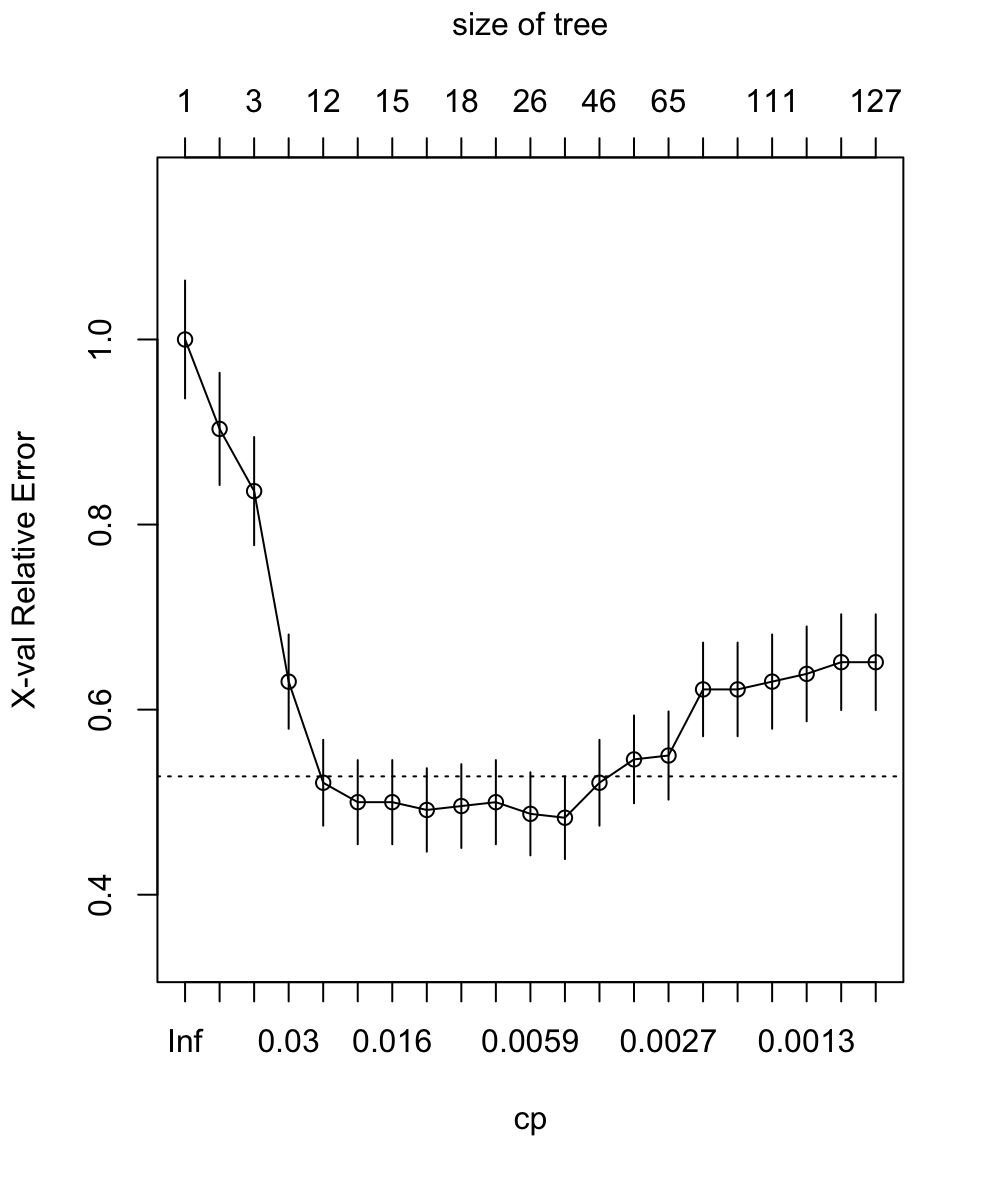
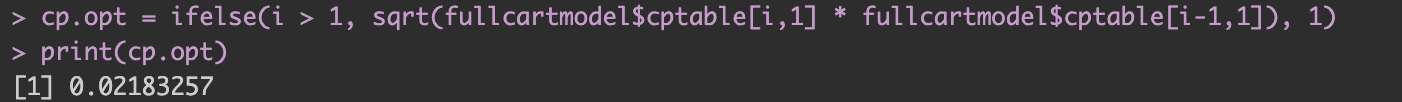
## 

## Appendix N: Stratified Sampling Process

## Appendix O: Undersampling Process



## Appendix P: Optimising CP - Stratified

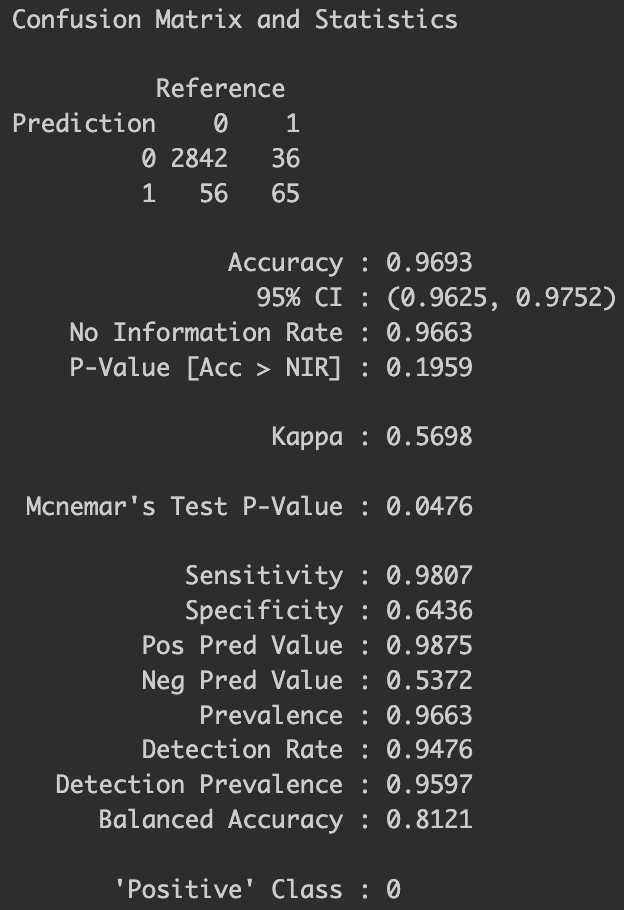


## Appendix Q : Confusion Matrix Comparison before after pruning - Stratified

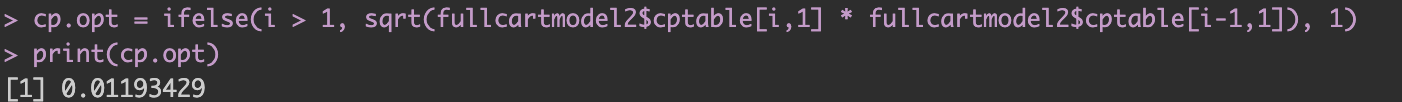
## 

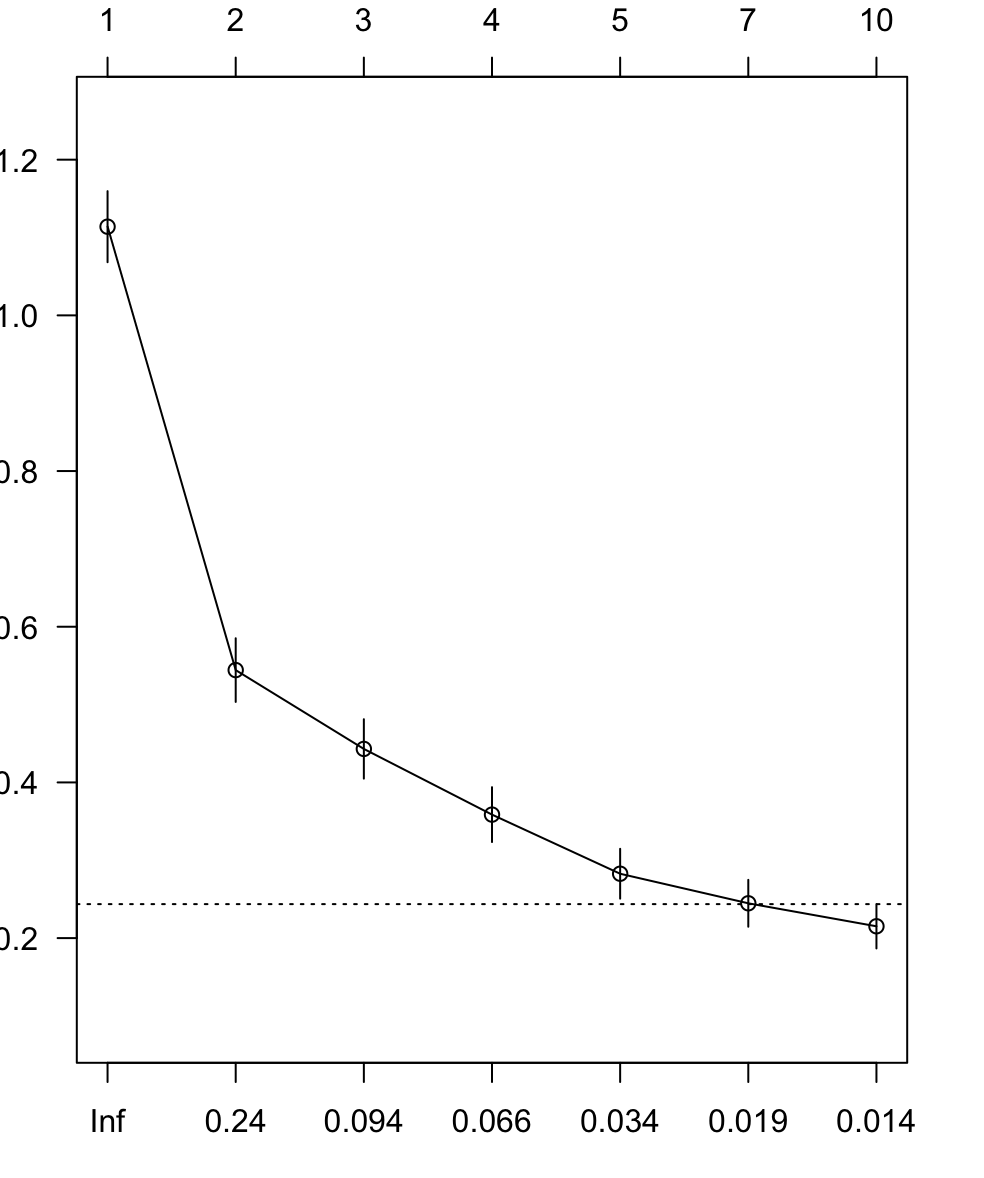
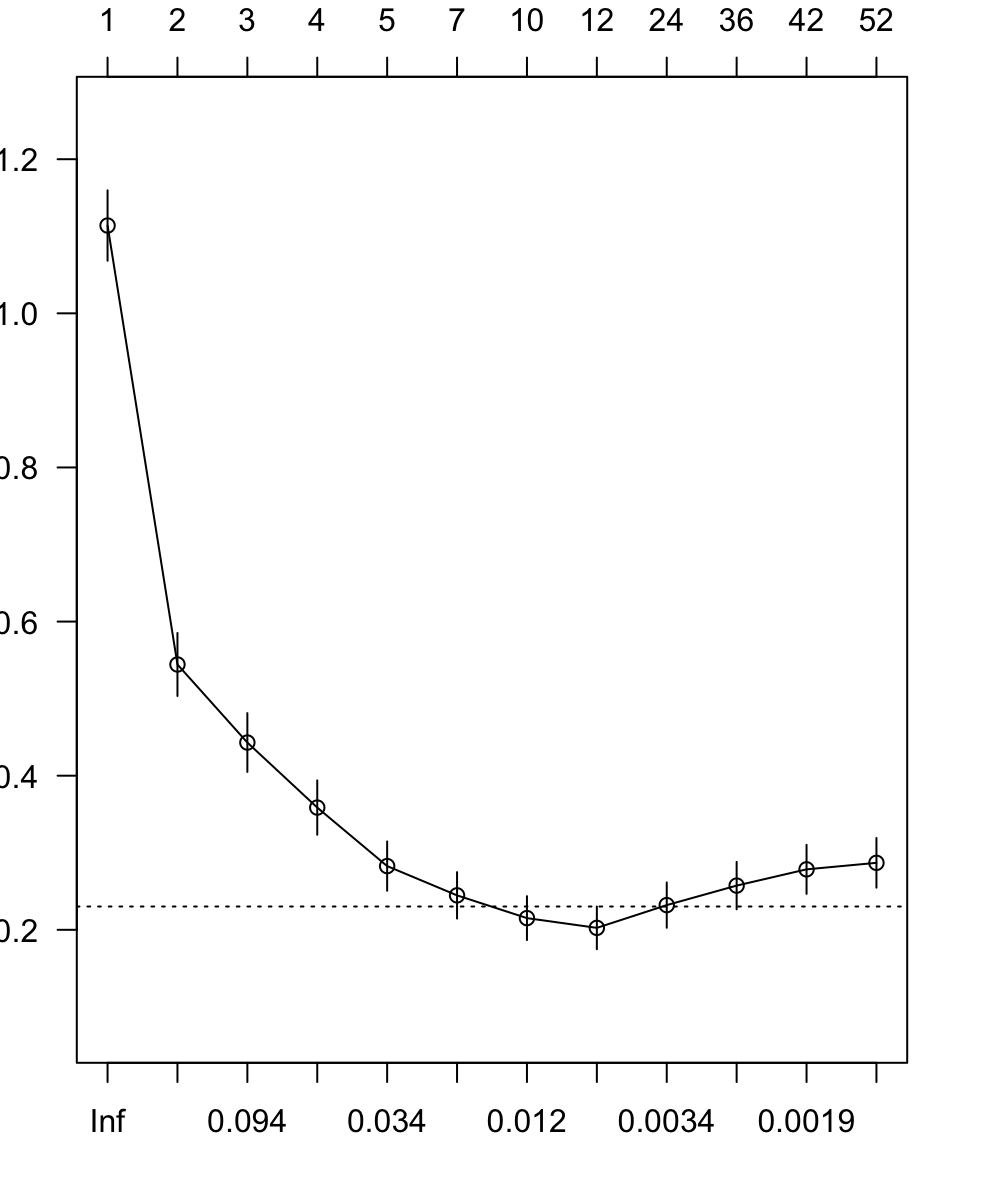
Before After

## Appendix R: Cost Sensitivity Learning and Confusion Matrix - Stratified



## Appendix S: Optimising CP - Undersampling





## Appendix T: Confusion Matrix before after pruning - Undersampling

# 

Before After

## Appendix U: Cost Sensitivity Learning and Confusion Matrix - Undersampling

## 

## Appendix V: Calculation and Compilation of Confusion Matrix for Evaluation

## 

# References

Aramco. (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&amp;hash=9620BCE26DE8648B32F58E09D5A9A41D363CD63E>

Aramco. (n.d.-a). AI and Big Data. <https://www.aramco.com/en/creating-value/technology-development/in-house-developed-technologies/digitalization/ai-and-big-data>

Aramco. (n.d.-b). Digital Technologies. <https://www.aramco.com/en/creating-value/technology-development/in-house-developed-technologies/digitalization/digital-technologies>

Brownlee, J. (2021, January 26). Undersampling algorithms for imbalanced classification. MachineLearningMastery.com. <https://machinelearningmastery.com/undersampling-algorithms-for-imbalanced-classification>

BOP Team. (2021, July 12). 5 common oilfield accidents: Bop Products, Houston Texas. BOP Products. <https://www.bop-products.com/blog/oilfield-safety/5-common-oilfield-accidents/>

Cox, J. (2020, February 19). It’s never been this hard for companies to find qualified workers. CNBC. <https://www.cnbc.com/2020/02/19/its-never-been-this-hard-for-companies-to-find-qualified-workers.html>

Dunn Sheehan. (2022, December 14). Common causes of oil rig accidents. Dunn Sheehan, LLP. <https://www.dunnsheehan.com/news-insights/what-causes-oil-rig-accidents>

Escobedo, G. P. (2020, February 17). Top 5 causes of offshore oil rig accidents. San Antonio Workers’ Compensation Lawyer. <https://www.georgeescobedo.com/blog/2020/february/top-5-causes-of-offshore-oil-rig-accidents/>

Gholamy , A., Kreinovich , V., &amp; Kosheleva, O. (2018). Departmental Technical Reports (CS). <https://scholarworks.utep.edu/cgi/viewcontent.cgi?article=2202&amp;context=cs_techrep>

Infineon Technologies. (n.d.). Predictive maintenance and condition monitoring. <https://www.infineon.com/cms/en/applications/industrial/condition-monitoring-and-predictive-analytics/>

Kim, J. H. (2019). Multicollinearity and misleading statistical results. Korean Journal of Anesthesiology, 72(6), 558–569. <https://doi.org/10.4097/kja.19087>

McKinsey . (2023, July 12). What is digital-twin technology?. McKinsey &amp; Company. <https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-digital-twin-technology>

Oil&Gas. (2023, July 19). Aramco’s SATORP refinery to shut down for maintenance. <https://www.oilandgasmiddleeast.com/news/aramcos-satorp-refinery-to-shut-down-for-maintenance>

Pallardy, R. (2023, September 21). Deepwater Horizon Oil Spill. Encyclopædia Britannica. <https://www.britannica.com/event/Deepwater-Horizon-oil-spill>

Safety on oil rigs: What you need to know. Katch Kan. (2023, August 24). <https://www.katchkan.com/rsrc/safety-on-oil-rigs-what-you-need-to-know/#:~:text=The%20drilling%20industry%20is%20rife,paramount%20in%20the%20oil%20industry>.