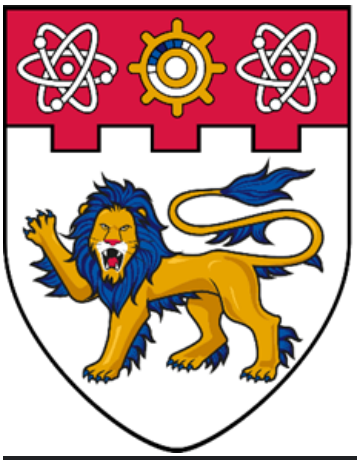

NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

# Optimisation of Operational and Sustainability Practices for Aramco

**Seminar Group 4 Team 3**


| Name | Matriculation No. |
|---|---|
| Villamor Ayesha Marie Martin | U2220739H |
| Ooi Jian Bo | U2110567D |
| Li Liyi | U2220985F |
| Sailesh Sampath | U2221329C |
| Liu Xinyi | U2221219H |

# Table of Contents


Executive Summary 3
Introduction 4
Opportunity statement 4
- Enhancing operational efficiencies 4
- Forecasting future demand of renewable energy sources 4

Value Addition to Aramco 5
Value Addition to Aramco's Stakeholders 6
- Investors 6
- Government 6
- Businesses 7

Methodology 8
- Data collection and cleaning 8
- Machine Predictive Maintenance Classification Dataset 8
- Global Data on Sustainable Energy (2000-2020) dataset 9
- Linear regression 11
- Classification and Regression Trees (CART) Analysis 12

Results & Analysis 14
- Predicting Machine Failure 14
- Forecasting Renewable Energy Share 18
- Limitations 20

How Aramco can adapt and implement our idea 22
References 25
Appendix 26

## Executive Summary

As the oil and gas industry transitions into the use of machine learning applications, the predictive analysis of machine failure and future renewable energy share will be beneficial not only for Aramco, but its various stakeholders such as investors, governments and business in making more informed decisions. A closer look at the changing market landscape is all the more pertinent in providing a comprehensive analysis of the opportunities Aramco can capitalise on.

This report explores possible machine learning applications, specifically linear regression, logistic regression and classification and regression trees (CART) to predict machine failure, as well as renewable energy share as a proxy for global renewable energy demand. For predicting machine failure, both the logistic regression and CART models identified Rotational Speed, Torque and Tool Wear as common predictors in both models. This gives us insight that these variables have a well established relationship with machine failure and has a meaningful contribution in predicting machine failure. Similarly, in predicting renewable energy share, both the linear regression and CART model share the same independent variables, C02 Emissions and Energy Intensity which suggests that the same few variables are useful in the predictive analysis. In comparing the regression models against the CART models, we used misclassification error and root mean square error (RMSE) as performance metrics for comparison. From our analysis, the CART model performed better overall, but the difference in performance between the models was marginal. This could possibly indicate a highly linear relation between the predictor variables and the target variables.

In implementing our results in Aramco's business operations, Aramco can provide a more comprehensive maintenance schedule that will prioritise the repair of machines near failure condition, thereby optimising resource allocation. Moreover, a rule-based alert can be utilised for predictive maintenance, where a monitoring system can be implemented that triggers alerts when key conditions leading to machine failure are met. Furthermore, the insights gained in forecasting renewable energy share can play a pivotal role in Aramco's strategic decision making. Aramco can use this information to craft market diversification strategies and identify possible investment opportunities in the ever-growing renewable energy sector.

As this is a largely exploratory analysis, some limitations of our model include incomprehensive predictor variables and vulnerability to unforeseen circumstances. Future developments to build a more accurate model can bear these limitations in mind to improve overall predictive accuracy.

## Introduction

Aramco is a state-owned and long-established oil company with roots going back to 1933. As oil and gas companies embark on using big data to improve their operations, Aramco must also explore innovative avenues as to how they can utilise predictive analytics in their operations to maintain and enhance their competitive advantage in the industry. This report presents an in-depth analysis of how machine learning and artificial intelligence can empower Aramco in optimising operational efficiency and embracing the principles of Environmental, Social, and Corporate governance (ESG).

## Opportunity statement

In this report, we use analytics to focus on two key opportunities for Aramco:

1. **Enhancing operational efficiencies:** We aim to enhance operational efficiencies by accurately predicting the condition of machines, thereby facilitating accurate maintenance scheduling. This endeavour carries the potential to not only streamline operations but also to significantly reduce overall operational costs.

2. **Navigating the expanding alternative energy market:** The report will delve into the growing alternative energy market, specifically by analysing and forecasting the future demand for other renewable energy sources. This proactive approach aligns with the increasing emphasis on ESG and represents a strategic move for Aramco in moving towards sustainability, diversification and revenue enhancement.

<u>Enhancing operational efficiencies</u>

Given that Aramco has the largest daily oil production of oil-producing companies (Wikipedia, 2023), it likely has extensive operations along with a large number of machines for oil extraction. As such, there may be a multitude of operational efficiencies in their processes. According to our research, we found that using equipment to the point of failure costs ten times more than performing periodic performance (DataNovia, n.d.). Recognising this, we saw an opportunity to produce a data model that can accurately predict a machine's condition and estimate an optimal maintenance frequency schedule. This can improve the efficiency of the maintenance process, reducing the overall downtime of machines and also lead to a reduction in overall operational costs for Aramco.

<u>Forecasting future demand of renewable energy sources</u>

With the growing emphasis on ESG in the industry (Eleven, 2023), Aramco can tackle sustainability issues to promote more efficient, reliable and environmentally friendly operations. In the long run, focusing on ESG can also help Aramco improve its sustainability and efficiency practices. Renewable energy sources are gaining popularity in recent years, this is evident in the recent trends where annual clean energy investments are expected to rise by 24% from 2021 to 2023 as opposed to a lower 15% rise in fossil fuel investment over the same period (Iea, 2023). As such, we aim to create our own predictive model that can forecast the percentage of energy

consumption from renewable energy sources. This allows Aramco to make informed decisions on how fast the renewable energy sector is growing and balance the investments they should allocate into the sector accordingly.

**Value Addition to Aramco**

Our proposal offers significant value to Aramco as well as its stakeholders such as investors, government entities and potential business partners. By leveraging on growing industry opportunities, we aim to enhance Aramco's operations and strategic positioning within the ESG framework.

With our proposal, not only will Aramco further cement itself in the oil and gas industry as a leader, it will also establish itself as a diverse, forward-looking company in the renewable energy sector. Firstly, In terms of operational efficiency, our proposal will help Aramco improve by predicting machine conditions and maintenance needs accurately. This enables proactive maintenance, reducing downtime and avoiding costly machine failures.

Secondly, the predictive analysis of future renewable energy share can be used by Aramco in resource allocation. The detailed information can be used as a guideline as to how much oil Aramco should produce in line with market demand. If the renewable energy share in total energy consumption is increasing drastically, Aramco must take measures to decrease and adjust their oil production accordingly and instead increase their focus on renewable energy production, so as to prevent inefficient resource allocation. This approach minimises inventory storage costs, reduces energy wastage, and maximises revenue by avoiding missed opportunities.

Lastly, this proposal facilitates Aramco's transition towards ESG compliance by identifying potential opportunities in the renewable energy market. Aramco can explore diversifying into renewable energy sources, such as solar panels, for small to medium-sized firms, expanding its revenue streams while aligning with ESG goals.

**Value Addition to Aramco's Stakeholders**

Aramco's stakeholders will benefit from our proposal as it demonstrates a commitment to environmental responsibility. By reducing the carbon footprint and investing in renewable energy, Aramco can fulfil its ESG obligations, contributing to a more sustainable future. Moreover, stakeholders can have confidence in Aramco's long-term viability and stability. Our predictive maintenance models ensure consistent oil production, reducing supply disruptions and maintaining stakeholder trust.

Investors

Investors are particular about the future direction of the company and industry trends. Machine learning and predictive modelling can be useful in this context to provide investors an overview of the growing renewable energy trends in the sector, and the ESG practises Aramco is adopting to adjust to these market trends. These predictive models can equip investors with the knowledge necessary to pinpoint the most promising regions where they can expand their renewable energy investments, in line with the results and analysis of our models that will be discussed later on in this report. This empowers the investors to make informed decisions and allows them to capitalise on the most promising opportunities presented by Aramco's strategic shifts. Moreover, the implementation of our predictive maintenance solutions diminishes the risk of sudden machine failure and disruptions. The reduction of such operational uncertainties portrays Aramco as a more secure and reliable investment, thereby allowing investors to foster greater confidence and trust in Aramco.

Government

Firstly, governments focusing on ESG policies will view Aramco's efforts positively. Our proposal aligns with Saudi Arabia's government objectives to move towards environmental sustainability. This can be seen in their National Renewable Energy Program (NREP) where they plan to transition 50% of their domestic energy supply to renewable energy sources by 2030 (Philipp, 2023). Hence, considering our model in predicting renewable energy share, the Saudi Arabia government can utilise this to predict how fast the renewable energy share is growing in the country. They can then use this information to make strategic decisions on the optimal amount of infrastructure and incentives they ought to put in place towards companies so as to facilitate the transition towards clean energy sources.

Secondly, our models can be used to boost economic efficiency. By studying the growth trajectory of renewable energy sources in Saudi Arabia, this information can be used to minimise energy wastage and resource misallocation. Energy wastage is reduced given that Aramco and the government both have a better understanding of the changing energy landscape and the demand for various energy sources. They can then calibrate their national output accordingly to adjust to the changing trends in energy demand, thereby boosting resource allocation. This results in a more resource-efficient and economically competitive energy sector, fostering sustainability and contributing to the Saudi Arabian government's broader goals of energy diversification and responsible resource management. Moreover, with our predictive maintenance model, Aramco will be able to improve its productive capabilities. Since oil and gas are a vital component of the

country’s economy, achieving greater operational efficiency for Aramco will largely contribute to the country's overall economic efficiency.

Businesses

Machine learning can aid companies in the oil and gas industry to refine their business strategy. As companies in the sector embark on the Industry 4.0 Journey, the pillar of big data generated across all value chain sectors can be effectively utilised to create more efficient, reliable and environmentally friendly operations (Roberty, 2021). With an accurate forecast of future renewable energy share, businesses in the sector can analyse the general trend and supplement it with their own growth projections to come up with business strategies. For instance, Aramco can leverage our proposal to form partnerships with renewable energy companies and suppliers. Our forecasting tools enable Aramco to make informed decisions on which renewable energy sources to invest in, fostering potential collaborations.

Moreover, our proposal has the potential to enhance supply chain reliability. For businesses that depend on Aramco’s oil production, such as petrochemical companies, our predictive maintenance model will ensure consistent and reliable oil supply. It ensures that Aramco can meet its delivery commitments consistently, thereby reducing the chances of supply chain disruptions. This reliability is crucial for Aramco and its business partners to meet their own production schedules and customer demands.

In summary, our proposal offers multifaceted benefits to Aramco and its various stakeholders, aligning the company with ESG goals, improving operational efficiency, and contributing to a more sustainable and responsible business model. This further cements Aramco’s position for success in an evolving market landscape.

**Methodology**

Data collection and cleaning

Accuracy and volume of data are prerequisites before we are able to perform any meaningful predictions of the data regardless of the machine learning model that we choose to use. Therefore, to prepare for this report, we were thorough in searching for high-quality and complete datasets that would result in reliable and meaningful predictions for Aramco. The main considerations we had were the sufficiency of data points and variables for analysis, and the reliability of the sources. Many of the datasets that we found during the initial phase were not relevant and could not fulfil the requirements that we set. Therefore, we decided to explore alternative datasets and discovered two that matched our criteria.

Machine Predictive Maintenance Classification Dataset

This dataset provides information that can give us insights into machine failure and possible variables that may influence machine failure. To better understand our dataset, we conducted data exploration. Data exploration should be purpose driven, and in this case the purpose of our data exploration is to analyse the distributions of our variables. A point to note is that in this dataset, there are two target variables, “failure” and “failure_type”. Hence, in our predictive models, we have purposefully taken out “failure_type” when building the full model with all predictor variables.

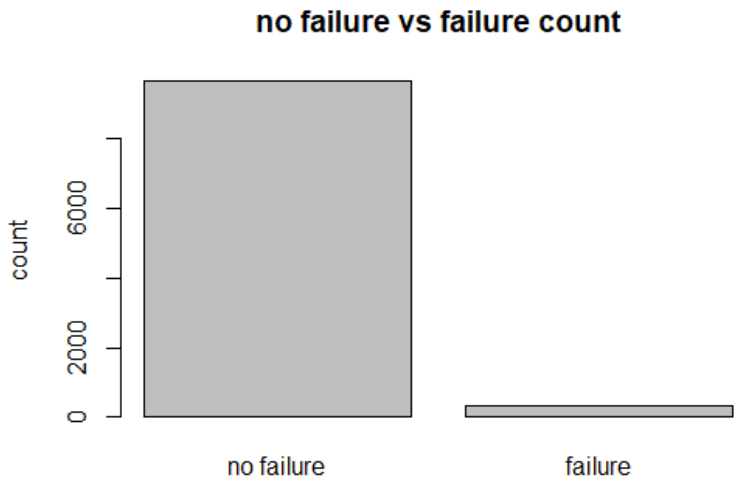


*Figure 1:Failure by Type*

After initial analysis of the data, we determined that the data is imbalanced where most data points had observations of ‘no failure’. To address the imbalance in data, we sampled the number of ‘no failure’ observations in the train set to ensure that the data was more balanced and would result in more relevant and reliable predictions of machine failure.

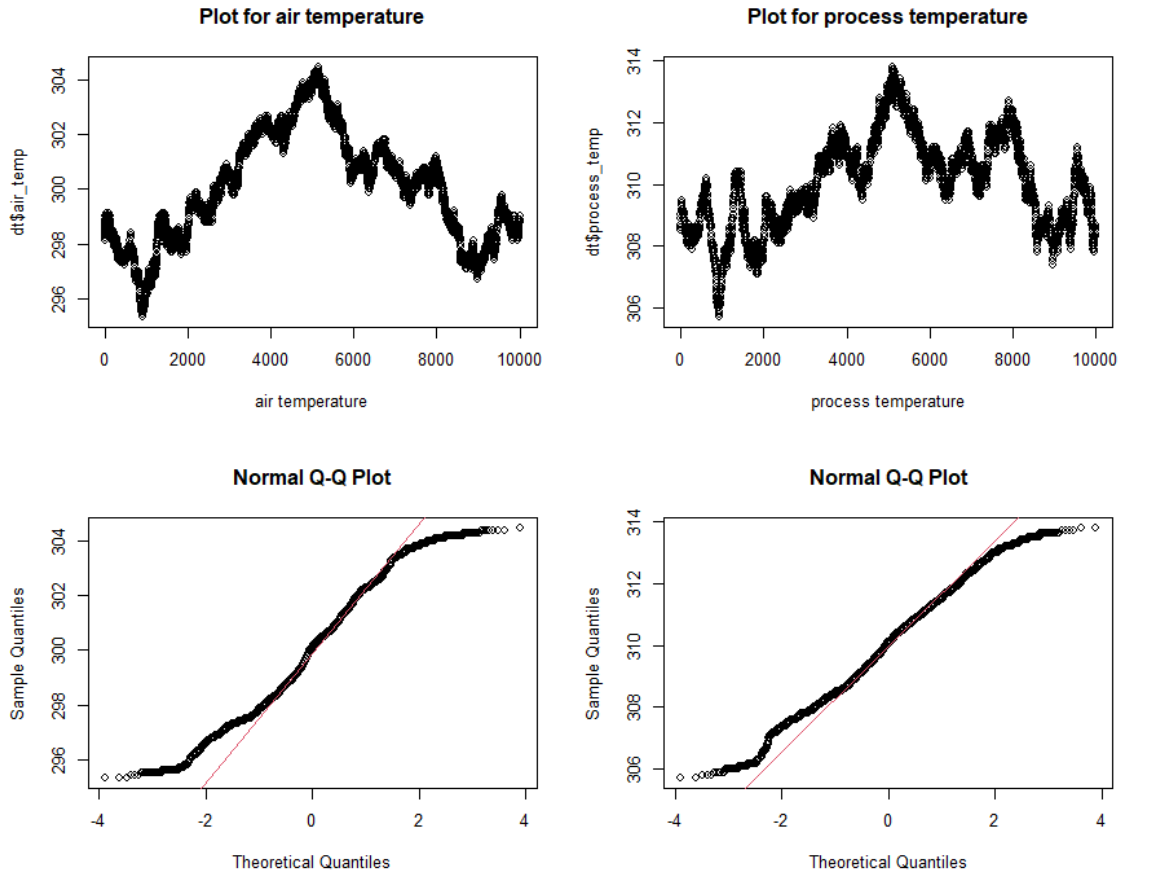


Figure 2: Plots for air and process temperature

Upon further analysis of our variables' distributions, we found that the distributions of air and process temperature seem to resemble that of a normal distribution. This is further confirmed by the Q-Q (Quantile-Quantile) plot, a graphical method for assessing a variable's normality. As can be seen from the QQ plots at the bottom, the data points align closely to the diagonal line, providing graphical confirmation that the distributions of these variables are approximately normal.

Global Data on Sustainable Energy (2000-2020) dataset

Our team has identified that there is a current increase in sentiment towards companies that comply with Environmental, Social and Corporate Governance standards (ESG). As we performed analysis to predict the future renewable energy demand, we chose this dataset given that it had the most relevant variables that could be used in predicting future renewable energy demand.

After looking across multiple sources for datasets, we could only find datasets that contained a limited amount of data where there were many missing values or where total renewable energy consumption was shown against a time period. We decided to settle on this current dataset as it discusses all the different energy consumption and production methods as well as how they all

relate to GDP growth as well as their emissions. As we were performing our analysis, we realised that if we focused on different unique countries for analysis because data was only recorded on a yearly basis for 20 years. We were unable to perform any meaningful analysis of the data. We decided to change our approach to not look at the change in demand for the top few consumers of oil from Saudi Arabia but instead focus on predicting the global demand for renewable energy.

Looking at this dataset, we realised that there were numerous variables that were directly related to the proportion of renewable energy that is being consumed within a country, therefore, we decided to look into 3 specific variables to predict the renewable energy share of a country. As we wanted to predict the renewable energy demand for the future year, we also needed to ensure that the data contained all countries. Therefore, we decided to perform the analysis on the different countries in the year 2019.

The final variables that we kept are as shown below:

| | |
|---|---|
| Renewable energy share in total final energy consumption (%) | Percentage of renewable energy in final energy consumption. |
| Energy intensity level of primary energy (MJ/$2011 PPP GDP) | Energy use per unit of GDP at purchasing power parity. |
| Value_co2_emissions (metric tons per capita) | Carbon dioxide emissions per person in metric tons. |
| GDP per capita | Gross domestic product per person. |

*Table 3: Variables chosen for second dataset*

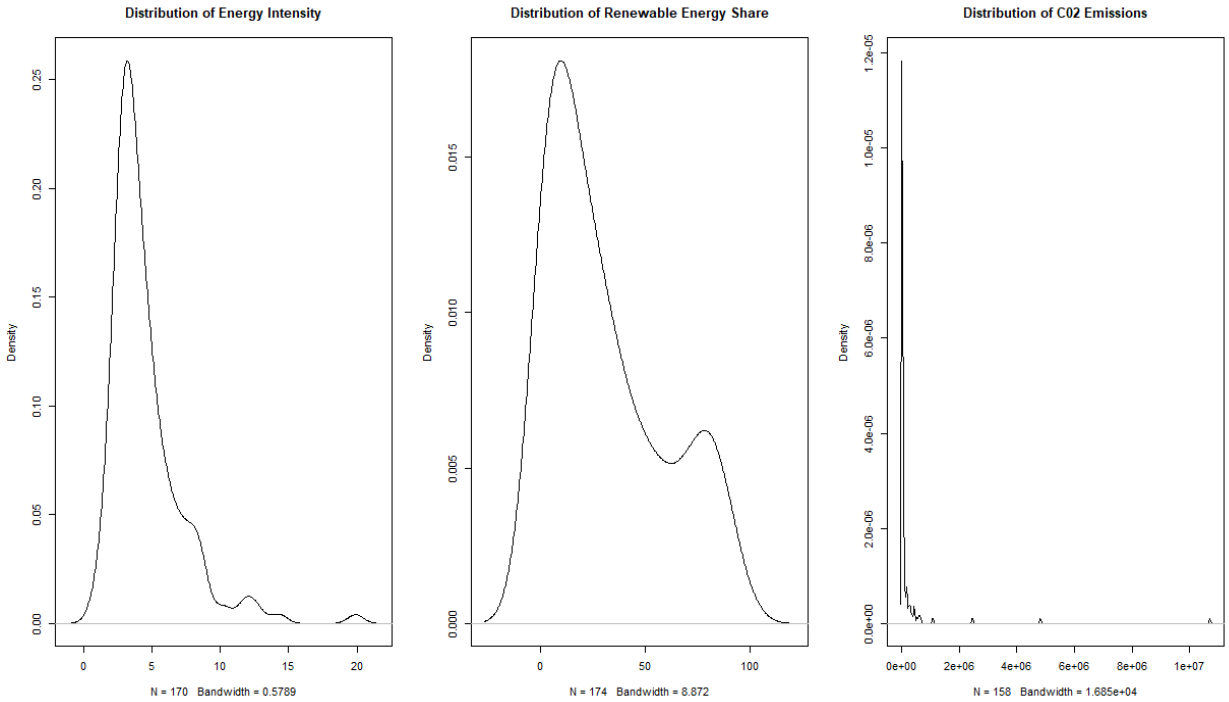


*Figure 4: Distribution plots of Energy intensity, Renewable Energy Share and C02 Emissions*

We also conducted data exploration to gain a better understanding of the dataset. After further analysis, we noticed that the distributions of Energy intensity, Renewable Energy Share, as well as C02 Emissions are right-skewed. The skewness was evident visually, as well as from their positive skew coefficients: 2.21, 0.75, 9.32 for Energy intensity, Renewable Energy Share and C02 Emissions respectively. Given that the skew coefficient of Renewable Energy Share is less than 1, this indicates moderate skewness while a skew coefficient of more than 1 indicates high skewness.

To normalise these distributions, we applied log transformation to Energy Intensity and C02 Emissions, given that they exhibited high skewness. We decided not to conduct log transformation to Renewable Energy Share given that it exhibited moderate skewness only and we planned on utilising models like linear regression which were more robust to moderate skewness. Log transformation is a common technique in data processing pipelines to handle skewed data, transforming the distribution to be more normal while preserving the relationships between data points. After log transformation, the distributions were more normal in nature:

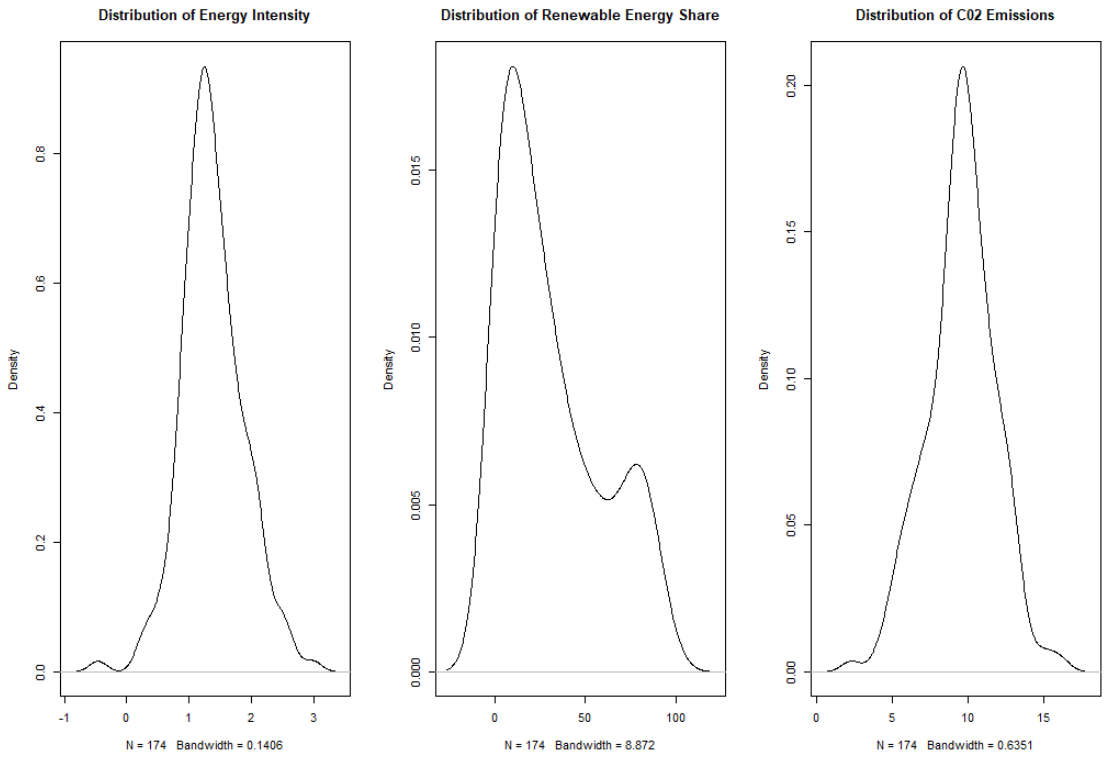


*Figure 5: Distribution plots of Energy intensity, Renewable Energy Share and C02 Emissions after log transformation*

Linear regression

The first machine learning method we used was linear regression. This applies a linear approach to model the relationship between the dependent variable and all independent variables. By plotting the dependent variable against all the independent variables, we will be able to determine the significance of the independent variables in predicting the value of the dependent variable. This is based on the assumption that the variables chosen are linearly related. This allows us to build a model to predict a possible outcome.

One potential problem that could occur when using linear regression on a large data set with a large number of independent variables is multicollinearity, or high intercorrelation, between independent variables. In this case, we would calculate the variance inflation factor (VIF), which is the ratio of the overall model variance to the variance of the model that includes only that single independent variable for every variable that we have chosen. A large VIF suggests that that specific variable is highly collinear with other independent variables in the model which would result in the coefficient of the variables being inaccurate.

In the case of our dataset, since there are many columns, it is very likely that there is multicollinearity between the predictor variables. Therefore, we decided to use backward elimination, a feature selection technique. During backward elimination, all independent variables are first entered into the equation, before being individually evaluated for their impact on the

equation. Step by step, the least significant variables are dropped from the equation until removing any variables from the equation results in the entire model becoming less efficient

One of the main assumptions made in linear regression models is that there is a linear association between variables and errors that are independent of the independent variables and have a constant standard deviation. These are important assumptions that we need to consider when deciding on the feasibility of the model, whether it accurately and realistically illustrates the relationships between variables. In our case, for the prediction of renewable energy demand, there may not be a completely linear relationship between renewable energy share and many of the independent variables.

Finally, the linear regression model will provide us with a model that predicts the dependent variable by estimating it using known values of the independent variables. This is achieved by putting the values of the independent variables into an equation where each independent variable has a coefficient, which reflects how the dependent variable changes when each independent variable changes.

Since the purpose of our machine learning models is to create a model that can accurately and reliably predict our dependent variables, we need to use a trainset - testset split to determine the accuracy of the model. We used a train set-test set split ratio of 70-30 with a seed value of 1, where the train set was used to obtain the linear regression model and the test set was used to determine model accuracy.

Classification and Regression Trees (CART) Analysis

To further enhance the reliability of our predictions, we wanted to make use of another machine learning model such that we are able to obtain a model which is a better fit for the data points. The CART model is also beneficial in tackling the issue where the independent and dependent variables are not linearly related. The CART algorithm constructs the decision tree by considering the sum of squared errors (SSE) at a particular node if the independent variable is taken into account, before using it as a splitting criterion. CART is a two-stage process, followed by a generation of the optimal tree.

In the first stage, the decision tree is grown to the maximal tree using the highest Gini index. This is achieved through testing all the independent variables to determine the best split. Each node is then divided into two child nodes multiple times to obtain the largest possible tree. After that in the second stage, the tree will be pruned to a minimum. Pruning is performed by calculating the contribution of a leaf node to the total classification error to find the weakest decision node and removing all the terminal nodes connected to the weakest decision node.

The optimal tree lies somewhere between the maximum and minimum tree and it is chosen using 10-fold cross-validation with the one standard error rule. In 10-fold cross-validation, the dataset is divided into ten random and equal train test sets and the model will be fitted to the different is tested against each other to prevent the model from overfitting. This helps to generate a better estimate of the model. To aid in minimising the error, the sum of squared errors (SSE) of the

model is calculated, and the index of the first node where the mean error is lower than the lowest sum of squared errors will be used to generate the optimal model.

Using CART, we can also rank independent variables in order of importance. This allows us to determine which variables are the most significant in determining the value of the dependent variable. CART is also a good machine learning model as it is able to generate a replacement list for missing values. They act as a replacement for the best major split in cases where variables that are supposed to be used for splitting are not present in the data set. One differentiating factor between linear regression and CART is that CART is able to deal with missing or NA values through the use of surrogates to split the data, instead of just skipping over it. While our dataset is chosen with data completeness in mind, this assures us that the data will be fully utilised and potentially help to correct any missing values that we may have missed

Ultimately, the final result of CART is a decision tree where the range of the dependent variable can be determined using a certain set of independent variables.
As we have two hypotheses, for the first model of predicting machine failure, there will be different data generated as there are different factors that may result in machine failure for different years. For our second model, the decision tree that we generate will change for different years of data as in different years, there will be different variables that will be significant in contributing to the changes in renewable energy share. This means that the CART model should be performed regularly to ensure that Aramco will be able to obtain more meaningful results for their decision rules.

**Results & Analysis**

Predicting Machine Failure

In predicting machine failure, we constructed a logistic regression model and a CART model.

Logistic Regression

| | |
|---|---|
| **No. of variables dropped** | 1 |
| **Statistically significant variables** | Air Temperature<br>Process Temperature<br>Rotational Speed<br>Torque<br>Tool Wear |

*Table 6: Significant variables*

| **Variable** | **Vif values of full model** | **Vif values of optimal model** |
|---|---|---|
| **Type** | 1.074982 | - |
| **Air Temperature** | 4.941280 | 4.913303 |
| **Process Temperature** | 4.624677 | 4.612885 |
| **Rotational Speed** | 4.253244 | 4.248324 |
| **Torque** | 4.306814 | 4.294268 |
| **Tool Wear** | 1.206703 | 1.186918 |

*Table 7: VIF values of variables in the full and reduced model*

| | **Imbalanced train set** | **Balanced train set** |
|---|---|---|
| **Confusion matrix** | Reference<br>Prediction no failure failure<br>no failure 2882 80<br>failure 16 22 | Reference<br>Prediction no failure failure<br>no failure 2503 24<br>failure 395 78 |
| **True negative rate** | 22/(22+80) = 0.2157 | 78/(78+24) = 0.7647 |
| **False positive rate** | 80/(80+22) = 0.7843 | 24/(24+78) = 0.2353 |
| **Accuracy** | 0.9680 | 0.8603 |
| **Misclassification error** | 0.032 | 0.1397 |

Positive class: ‘no failure’, Negative class: ‘failure’

*Table 8: Confusion matrix results of the logistic regression model*

The first row of Table 6 indicates that one variable has been dropped to obtain the model. We used backwards elimination to obtain the optimal model to help the logistic regression model in better identifying an appropriate model fit.

VIF was also calculated to measure multicollinearity in the model. As can be seen from the VIF values in Table 7, the VIF values are consistently low across all variables, indicating low multicollinearity within the full model. After performing feature selection through backwards elimination, the 'Type' variable was identified as a less influential predictor and removed. The removal of 'Type' further reduced the VIF values of the variables, reinforcing the notion of low multicollinearity in the model. This makes for a more suitable model for analysis, which provides greater stability in making predictions.

To measure the accuracy of our model, we trained the optimal logistic regression model and ran it against the test set to derive the misclassification error. Misclassification error refers to the rate at which the model's prediction does not match the actual outcomes. It is calculated using the following equation:

**Misclassification error = (false positives + false negatives) / total no. of observations**

CART

| | |
|---|---|
| **Decision rules** | 7: If rotational_speed < 1381 AND air_temp >= 302, failure<br><br>13: If rotational_speed < 1381 AND air_temp < 302 AND torque >= 60, failure<br><br>25: If rotational_speed < 1381 AND air_temp < 302 AND torque < 60 AND tool_wear >= 190, failure<br><br>24: If rotational_speed < 1381 AND air_temp < 302 AND torque < 60 AND tool_wear < 190, no failure<br><br>5: If rotational_speed >= 1381 AND tool_wear > 202, failure<br><br>9: If rotational_speed >= 1381 AND tool_wear < 202 AND rotational_speed > 2149, failure<br><br>17: If rotational_speed >= 1381 AND tool_wear < 202 AND rotational_speed < 2149 AND torque > 57, failure<br><br>16: If rotational_speed >= 1381 AND tool_wear < 202 AND rotational_speed < 2149 AND torque < 57, no failure |
| **Important variables** | Rotational_speed<br>Torque<br>Tool_wear<br>Air_temp<br>Process_temp |

*Table 9: Decision rules and important variables in the CART model*

| | **Imbalanced train set** | **Balanced train set** |
|---|---|---|
| **Confusion matrix** | Reference<br>Prediction no failure failure<br>no failure 2889 35<br>failure 9 67 | Reference<br>Prediction no failure failure<br>no failure 2608 9<br>failure 290 93 |
| **True negative rate** | 67/(67+35) = 0.6569 | 93/(93+9) = 0.9118 |
| **False positive rate** | 35/(35+67) = 0.3431 | 9/(9+93) = 0.0882 |
| **Accuracy** | 0.9853 | 0.9003 |
| **Misclassification error** | 0.0147 | 0.0997 |

Positive class: 'no failure' , Negative class: 'failure'

*Table 10: Confusion matrix results* of the CART model

| **Model** | **Misclassification error** | |
|---|---|---|
| | **Imbalanced train set** | **Balanced train set** |
| **Logistic regression** | 0.032 | 0.1397 |
| **CART** | 0.0147 | 0.0997 |

*Table 11: Comparison of misclassification errors between logistic regression and CART*

We have found several key insights from our logistic regression and CART models. An interesting finding is that the misclassification error when using the imbalanced train set is consistently lower as compared to when using the balanced train set in both models. This could occur because the distribution of the test set is similar to the distribution of the imbalanced train set. Hence, training the model on a more imbalanced train set would yield a lower misclassification error if the test set is imbalanced as well.

In both models, the same 5 variables appeared as predictors in the optimal model. These variables include Rotational speed, Torque, Tool wear, Air temperature and Process temperature. Additionally, an interesting finding is that the most statistically significant variables in the logistic regression model were similar to the most important variables in the CART model. This can be observed in the table below:

| **Logistic regression: Statistically Significant Variables** | **Coefficient** | **P-value** |
|---|---|---|

| | | |
|---|---|---|
| Torque | 0.2390 | 2e-16 |
| Rotational Speed | 0.0094 | 2e-16 |
| Tool Wear | 0.0148 | 5.15e-14 |
| **CART: Important Variables** | **Variable importance** | |
| Rotational Speed | 38 | |
| Torque | 33 | |
| Tool Wear | 21 | |

*Table 12: Comparison of significant and important variables between logistic regression and CART*

Statistical significance and variable importance serve distinct purposes in the context of predictive modelling. While a statistically significant variable suggests the presence of a relationship with the target variable, it does not inherently imply the variable's importance in predicting the target. Assessing variable importance is centred on understanding the actual contribution of each variable to the model's predictive accuracy. However from our results in the above table, there seems to be an agreement between statistically significant variables and important variables in predictive accuracy. Considering the definitions of the variables as previously discussed, they are indeed meaningful in contributing towards the models' predictive accuracy. As such, their relationship with machine operations and wear-and-tear aligns with their statistical significance and importance in the model.

Next, we analysed the True Negative and False Positive rates of both models in Tables 8 and 10. Our negative class refers to 'failure' and our positive class refers to 'no failure'. To better meet our purpose of predicting machine failure, it is critical that our model accurately classifies when a machine is predicted to fail. Hence we examine the true negative rates, the rate at which our model predicts 'failure' and the machine is indeed failing. We observed high true negative rates of 76.47% and 91.18% for our logistic regression and CART models respectively. We also examined the false positive rates, which occurs when a machine is predicted as 'no failure' but it is actually failing. We observed relatively low false positive rates of 23.53% and 8.82% for our logistic regression and CART models respectively. It is a good sign that our false positive rates are lower than our true negative rates. This indicates that given a failing machine, there is a high chance of detection using our predictive models. Moreover, comparing the rates between logistic regression and CART, it can be observed that CART has a higher true negative rate and lower false positive rate than logistic regression, indicating better performance in predicting machine failure.

In the logistic regression model, we also considered the coefficients of the variables. The positive coefficients identified above signify that as torque, rotational speed and tool wear increase, there

is a higher likelihood of machine failure. The positive coefficient means that an increase in these variables results in an increase in the probability of resulting in a 'failure' result. It is also important to note that in a logistic regression model, each coefficient assumes that every other predictor remains constant. However, from our VIF values as shown in table 7, it can be seen that there is a low level of multicollinearity hence this assumption holds true.

In the CART model, Table 9 shows the decision rules of the optimal tree. It can be observed that the variables Rotational speed, Torque and Tool wear appeared more than once in the decision tree, further reinforcing their importance in the decision tree. Considering rotational speed as a decision rule, when rotational speed >= 1381, 51% of the observations resulted in 'no failure', indicating that rotational speeds above 1381 have a high probability of 'no failure'. Similarly, when Tool wear < 202, 51% of the observations also resulted in 'no failure', indicating that low tool wear is associated with a high likelihood of 'no failure'. For Torque, when Torque < 57 most of the observations result in 'no failure' indicating that a lower torque value below 57 is associated with a higher probability of 'no failure'.

Forecasting Renewable Energy Share

| | |
|---|---|
| **No. of variables dropped** | 1 |
| **Statistically significant variables** | C02 Emissions<br>Energy Intensity |
| **R-Squared** | 0.2659 |
| **Adjusted R-Squared** | 0.2535 |
| **Root Mean Square Error (trainset)** | 23.9808 |
| **Root Mean Square Error (testset)** | 23.6242 |

Linear Regression

*Table 13: Results of linear regression model*

| **Variables** | **VIF values for full model** | **VIF values for optimal model** |
|---|---|---|
| **GDP Growth** | 1.031851 | - |
| **C02 Emissions** | 1.032057 | 1.000634 |
| **Energy Intensity** | 1.031851 | 1.000634 |

*Table 14: VIF values of variables in the full and reduced model*

| **Variable** | **Coefficient** | **P Value** |
|---|---|---|

| | | |
|---|---|---|
| **C02 Emissions** | -3.9868 | 5.87e-0.5 |
| **Energy Intensity** | 22.3183 | 2.71e-0.6 |

*Table 15: Coefficient and p-values of variables in the linear regression model*

CART

| | |
|---|---|
| **Decision rules** | 2: If CO2Emissions < 9.7, there will likely be 21% renewable energy share of total energy consumption<br><br>6: If CO2Emissions >=9.7 AND EnergyIntensity <1.3 there will likely be 23% renewable energy share of total energy consumption<br><br>7: If CO2Emissions >=9.7 AND EnergyIntensity >=1.3 there will likely be 68% renewable energy share of total energy consumption |
| **Important variables** | Energy Intensity<br>CO2Emissions<br>GDP Growth |
| **Root Mean Square Error (trainset)** | 20.2783 |
| **Root Mean Square Error (testset)** | 24.8717 |

*Table 16: Results of the CART model*

To measure the accuracy of our models, the trained models were run against the test set, where the root mean square error (RMSE) was subsequently calculated. The RMSE is the root squared difference between the actual and predicted values of the model. In comparing the RMSE between the models, it can be observed that both models are similar in their performance, owing to their similar RMSE values.

| **Linear regression: Statistically Significant Variables** | **Coefficient** | **P value** |
|---|---|---|
| C02 Emissions | -3.9868 | 5.87e-0.5 |
| Energy Intensity | 22.3183 | 2.71e-0.6 |
| **CART: Important Variables** | **Variable importance** | |
| C02 Emissions | 43 | |

| | |
|---|---|
| Energy Intensity | 43 |

*Table 17: Comparison of significant and important variables between the linear regression and CART model*

In comparing the optimal CART and linear regression models, we have found some similarities in the predictor variables. From our results, there seems to be an agreement between statistically significant variables in the linear regression model and important variables in the CART model. This means that the variables C02 emissions and Energy Intensity are meaningful and have a significant relationship between renewable energy share consumption. In a more detailed analysis, from Table 17 we can observe that Energy Intensity is a more significant variable than C02 emissions in the linear regression model given its smaller p-value of 2.71e-0.6. However, from the CART model it is observed that they have equal variable importance. The disparity in results between the two models suggest that there may be non-linear relationships at play that the linear regression model cannot capture. Conversely, the CART model, with its ability of identifying non-linear patterns, offers a more comprehensive overview of the variables' influence on the dependent variable.

In the linear regression model, the negative coefficient of C02 emissions indicate that as C02 emissions increase, the average percentage of renewable energy share consumption decreases, holding all other factors constant. This is possible as higher C02 emissions are associated with a higher reliance on fossil fuels, which leads to a reduced percentage of renewable energy. However, while there is an observed relationship between C02 emissions and renewable energy share, it is essential to exercise caution as correlation does not imply causation. It is possible that there are external factors influencing this relationship.

In the CART model, energy intensity is observed to be a decision rule. When Energy Intensity > 1.3, the mean value for the renewable energy share is larger then when Energy Intensity < 1.3. This means that a higher Energy Intensity is associated with a larger renewable energy share. This relationship aligns with the logic that a higher energy intensity refers to a higher energy use per unit of GDP. As such if there is a higher energy use, it is likely that additional energy sources such as renewable energy will be employed to meet the greater demand, resulting in an increase in the renewable energy share.

Overall, From our analysis, Aramco can consider using the CART model to predict when machines would require maintenance since the CART model has consistently derived a lower misclassification error. In forecasting future renewable energy share, the performances of both models were comparable. However, CART would provide a more comprehensive analysis overall since it is adept at identifying non-linear relationships between variables as opposed to regression models which assume normality.

<u>Limitations</u>

*Incomprehensive predictor variables*

In predicting renewable energy share, our model only considers quantitative variables such as C02 Emissions and Energy Intensity. While these factors are indeed significant in predicting renewable energy share, it is also important to acknowledge that qualitative aspects such as heightened environmental awareness and increasing preference for renewable energy sources can also be important factors in predicting renewable energy share. However, our dataset did not include such qualitative aspects. As such, it is likely that we miss out on key information that could have further solidified and provided a more comprehensive overview of our analysis. Moreover, in predicting machine failure, there were only 5 predictor variables available in the dataset. Although there was a low level of multicollinearity as previously discussed, the model may benefit from the inclusion of more factors that may affect machine failure, such as maintenance costs or frequency of machine maintenance.

*Vulnerability to unforeseen circumstances*

In analytics, machine learning models are subject to the constraint of relying on past data to predict future events. In this case, the pasts data utilised does not account for natural disasters and other unforeseen circumstances that can have a large influence in machine conditions and renewable energy share. Hence, these forecasts should not be used as the sole determinant in predicting machine failure and renewable energy share, but should instead be used as guiding values that a local expert can refer to in the forecasting process before making a more comprehensive judgement that accounts for unforeseen circumstances that may be looming in the future.

**How Aramco can adapt and implement our idea**

Operational efficiency is the ability of an organisation to reduce waste in time, effort and materials as much as possible, while still producing a high-quality service or product (Gillis, 2021)**.** According to Kimberlite, a market research firm that specialises in collecting market data for the oil and gas industry, less than four days of unplanned downtime costs an oil and gas company over $5 million, with most companies averaging 27 days of annual unscheduled downtime (Christiansen, 2023).

Therefore, Aramco can use the given data and the rules derived from our Logistic Regression and CART model for predictive maintenance in the following ways:

Rule-Based Alerts through Real-time Monitoring

Using the CART model, a monitoring system can be implemented that triggers alerts when the conditions specified in the rules are met. For example, if the rotational speed drops below 1381 and the air temperature is above 302, an alert for potential failure is issued. This can be integrated with the logistic regression model that identifies Torque, Rotational Speed, and Tool Wear as the most statistically significant variables. The logistic regression model can provide probabilities of failure, which can be used alongside the rule-based system to make more nuanced decisions. This probability provides a more granular risk assessment than the binary output of the rule-based system. Each piece of equipment can be assigned a risk score based on the logistic regression model's output. This score reflects the current probability of failure and can be updated in real-time as the input variables change. The risk score can be used to prioritise maintenance actions. Equipment with higher scores would be attended to first, as they are at a higher risk of failure.

Operational Adjustments

Aramco can adjust operational parameters that are significant based on the regression model and CART model to avoid entering high-risk zones. For instance, if tool wear is approaching 190 and other conditions are close to the failure thresholds, operations can be modified to reduce stress on the equipment.

Resource Optimization

Prioritise maintenance resources to equipment that frequently operates near the failure conditions. This helps focus efforts where they are most needed, potentially reducing downtime and maintenance costs.

Equipment Design and Purchasing

Aramco can use the data to make informed decisions in the design and purchase of new equipment. If certain operational parameters are associated with higher failure rates, new equipment can be designed or selected to operate more reliably under those conditions.

Moreover, companies using predictive maintenance achieve a 5 to 10 per cent improvement in production efficiency, and a 20 to 30 per cent decrease in maintenance costs compared to contemporaries using time-driven maintenance philosophies (Christiansen, 2023). Additionally,

they claim a 30 to 50 per cent reduction in downtime on critical machines. As such, using our analysis can help Aramco create a dynamic predictive maintenance program that not only prevents failures but also optimises the performance and lifespan of its equipment. This approach can lead to significant cost savings, improved safety, and increased operational efficiency.

For the second goal of our solution, Aramco's navigation through the alternative energy market can be enhanced by leveraging the insights provided by the Linear Regression and CART models:

Demand Forecasting

Linear Regression indicates a negative coefficient for C02 Emissions (-3.9868) and a positive coefficient for Energy Intensity (22.3183). Aramco can implement a demand forecasting system using these variables as key inputs to model, as well as predict the future demand for oil versus renewable energy sources. Aramco's navigation through the alternative energy market can be enhanced by leveraging the insights provided by the Linear Regression and CART models.

Market Diversification Strategies

The CART model predicts a 68% renewable energy share when C02 Emissions are greater than 9,7 and Energy Intensity is greater than 1.3. Aramco can target investment in renewable energy in regions where C02 Emissions and energy intensity cross the identified thresholds, which indicate a high potential for renewable adoption. This will allow Aramco to make more precise investment decisions. Moreover, the high positive coefficient of Energy Intensity in the Linear Regression model indicates that regions with higher energy intensity are more likely to adopt renewable energy sources more aggressively. Through analysing regions which have a higher or growing energy intensity, Aramco can better predict future hotspots for renewable energy demand and adjust its strategies accordingly.

Policy Analysis

The significant p-values of C02 Emissions (5.87e-0.5) and Energy Intensity (2.71e-0.6) in the Linear Regression model highlight that policy changes affecting these variables are likely to have a significant impact on the renewable energy market. Aramco can engage in policy analysis and advocacy in line with their ESG goals, to anticipate and influence certain regulations that may affect the energy market, as well as prepare mitigation strategies for scenarios where policy changes could potentially impact the oil and energy market unfavourably.

With this adaptive approach, Aramco gains a more comprehensive understanding of oil and renewable energy's projected demands, allowing the company to ensure their revenue streams align with global trends. Additionally, by implementing the predictive maintenance of machines, Aramco can minimise equipment downtime and operational costs. Together, these applications of machine learning demonstrate how data-driven approaches can translate to more profitable business operations overall, further cementing Aramco's position as an industry leader.

**References**

Christiansen, B. (2023, June 15). *Important implications and benefits of predictive maintenance in the oil and gas industry*. MRO Magazine. https://www.mromagazine.com/features/important-implications-and-benefits-of-predictive-maintenance-in-the-oil-and-gas-industry/

DataNovia. (n.d.). *Transform data to normal distribution in R: Easy guide*. Datanovia. https://www.datanovia.com/en/lessons/transform-data-to-normal-distribution-in-r/

Gillis, A. S. (2023, June 15). *Important implications and benefits of predictive maintenance in the oil and gas industry*. MRO Magazine. https://www.mromagazine.com/features/important-implications-and-benefits-of-predictive-maintenance-in-the-oil-and-gas-industry/

Iea. (2023, May 25). *Clean Energy Investment is extending its lead over fossil fuels, boosted by energy security strengths - news*. IEA. https://www.iea.org/news/clean-energy-investment-is-extending-its-lead-over-fossil-fuels-boosted-by-energy-security-strengths

Philipp, J. (2023, January 4). *Recent developments for renewable energy in Saudi Arabia*. The Borgen Project. https://borgenproject.org/developments-for-renewable-energy-in-saudi-arabia/#:~:text=Saudi%20Arabia's%20Circular%20Carbon%20Economy&text=Through%20NREP%2C%20the%20Saudi%20government,core%20aims%20of%20Vision%202030.

Roberty. (2021). Predictive analytics applications for oil and gas processing facilities. https://dspace.mit.edu/bitstream/handle/1721.1/140083/machadoroberty-emachado-sm-sdm-2021-thesis.pdf?sequence=1

Team, 12:Eleven. (2023, May 1). *The importance of ESG in the oil and Gas Industry*. 12. https://www.12eleven.com/news/the-importance-of-esg-in-the-oil-and-gas-industry#:~:text=ESG%20can%20help%20companies%20become,partners%20in%20oil%20and%20gas.

**Appendix A - Dataset Dictionaries**

“machinefailure_data.csv”

| | |
|---|---|
| UID | unique identifier ranging from 1 to 10000 |
| productID | consisting of a letter L, M, or H for low (50% of all products), medium (30%), and high (20%) as product quality variants and a variant-specific serial number |
| Type | |
| air temperature [K] | generated using a random walk process later normalised to a standard deviation of 2 K around 300 K |
| process temperature [K] | generated using a random walk process normalised to a standard deviation of 1 K, added to the air temperature plus 10 K. |
| rotational speed [rpm] | calculated from horsepower of 2860 W, overlaid with a normally distributed noise |
| torque [Nm] | torque values are normally distributed around 40 Nm with an Ï$f$ = 10 Nm and no negative values. |
| tool wear [min] | The quality variants H/M/L add 5/3/2 minutes of tool wear to the used tool in the process. and a 'machine failure' label that indicates, whether the machine has failed in this particular data point for any of the following failure modes are true. |
| Failure | The target variable, indicates whether the machine has failed or not. |
| Failure Type | Type of Failure |

“renewableenergy_data.csv”

| | |
|---|---|
| Entity | The name of the country or region for which the data is reported |
| Year | The year for which the data is reported, ranging from 2000 to 2020 |
| Access to electricity (% of population) | The percentage of population with access to electricity. |

| | |
|---|---|
| Access to clean fuels for cooking (% of population) | The percentage of the population with primary reliance on clean fuels. |
| Renewable-electricity-generating-capacity-per-capita | Installed Renewable energy capacity per person |
| Financial flows to developing countries (US $) | Aid and assistance from developed countries for clean energy projects. |
| Renewable energy share in total final energy consumption (%) | Percentage of renewable energy in final energy consumption. |
| Electricity from fossil fuels (TWh) | Electricity generated from fossil fuels (coal, oil, gas) in terawatt-hours. |
| Electricity from nuclear (TWh) | Electricity generated from nuclear power in terawatt-hours. |
| Electricity from renewables (TWh) | Electricity generated from renewable sources (hydro, solar, wind, etc.) in terawatt-hours. |
| Low-carbon electricity (% electricity) | Percentage of electricity from low-carbon sources (nuclear and renewables). |
| Primary energy consumption per capita (kWh/person) | Energy consumption per person in kilowatt-hours. |
| Energy intensity level of primary energy (MJ/$2011 PPP GDP) | Energy use per unit of GDP at purchasing power parity. |
| Value_co2_emissions (metric tons per capita) | Carbon dioxide emissions per person in metric tons. |
| Renewables (% equivalent primary energy) | Equivalent primary energy that is derived from renewable sources. |
| GDP growth (annual %) | Annual GDP growth rate based on constant local currency. |
| GDP per capita | Gross domestic product per person. |
| Density (P/Km2) | Population density in persons per square kilometer. |
| Land Area (Km2) | Total land area in square kilometers. |

| | |
|---|---|
| Latitude | Latitude of the country's centroid in decimal degrees. |
| Longitude | Longitude of the country's centroid in decimal degrees. |

## Appendix B - Models

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -6.186e+01  2.844e+01  -2.175   0.0296 *
air_temp          9.019e-01  1.234e-01   7.311 2.66e-13 ***
process_temp     -7.654e-01  1.716e-01  -4.462 8.13e-06 ***
rotational_speed  9.929e-03  9.797e-04  10.134  < 2e-16 ***
torque            2.476e-01  2.111e-02  11.728  < 2e-16 ***
tool_wear         1.424e-02  1.891e-03   7.528 5.15e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 797.01  on 591  degrees of freedom
Residual deviance: 448.08  on 586  degrees of freedom
AIC: 460.08

Number of Fisher Scoring iterations: 5
```

Figure 1: Logistic regression model of predicting machine failure

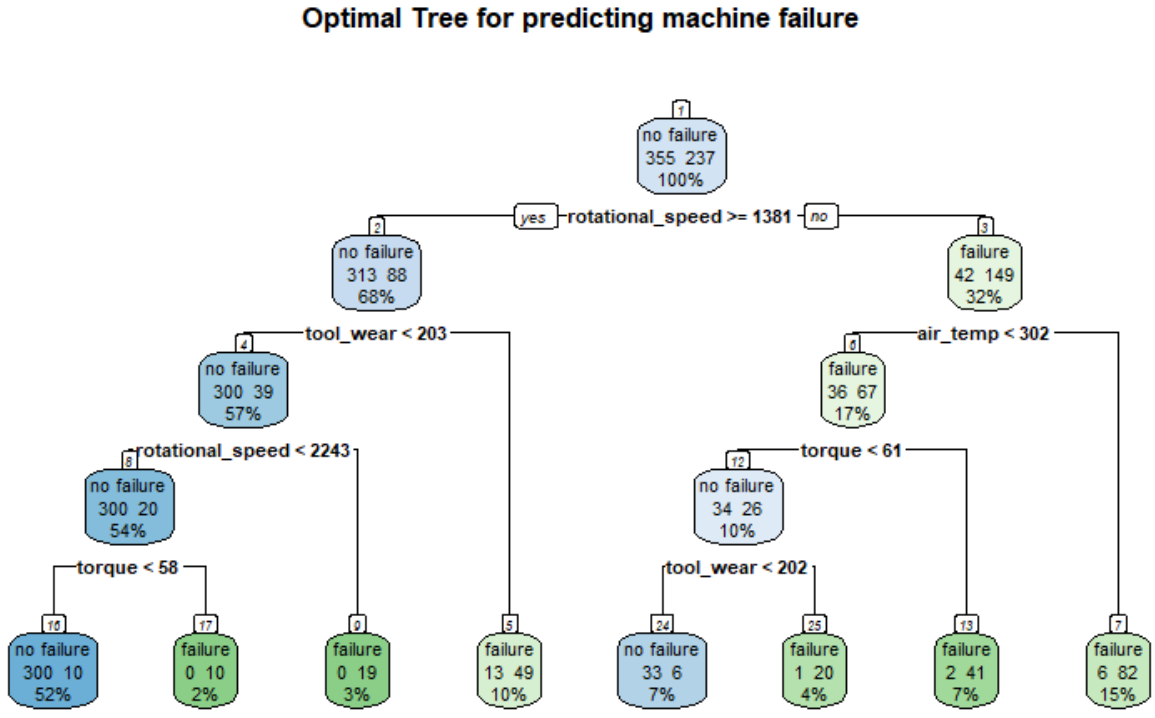


Figure 2: CART model for predicting machine failure

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        39.9967    11.4851   3.482 0.000698 ***
EnergyIntensity_log 22.3183     4.5265   4.931 2.71e-06 ***
CO2Emissions_log   -3.9868     0.9564  -4.169 5.87e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24.45 on 118 degrees of freedom
Multiple R-squared:  0.2659,    Adjusted R-squared:  0.2535
F-statistic: 21.37 on 2 and 118 DF,  p-value: 1.198e-08
```

Figure 3: Linear regression model of predicting renewable energy share

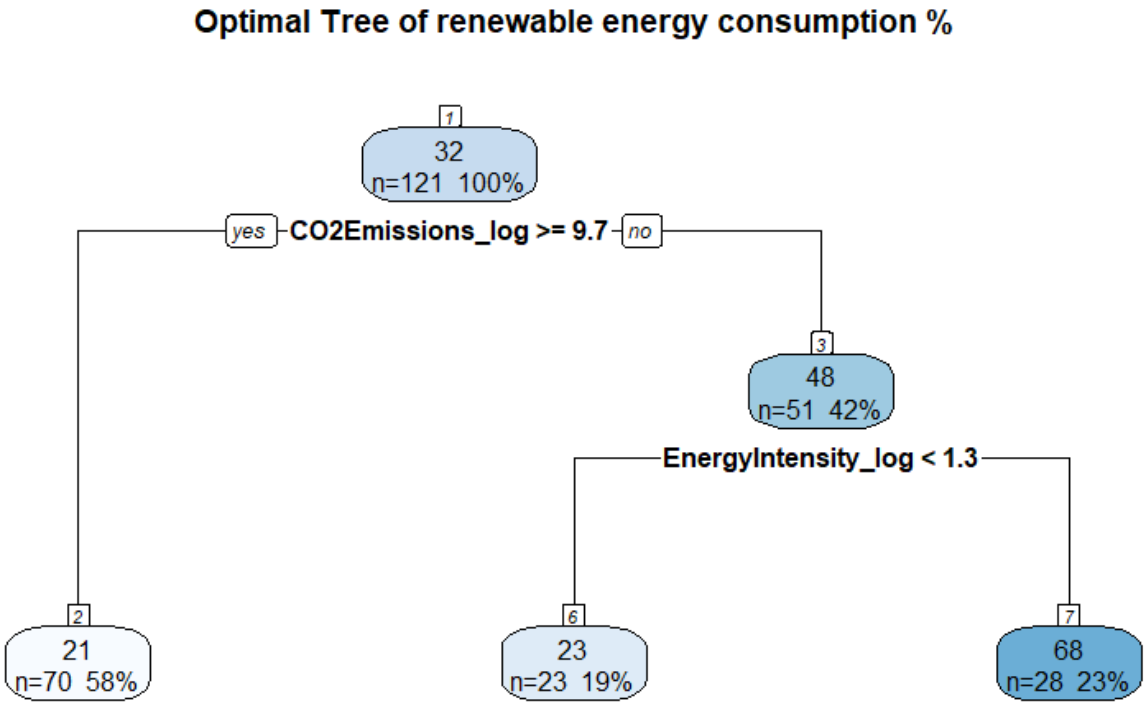


Figure 4: CART model of predicting renewable energy share