Table of Content

List of Figures

List of Tables

CHAPTER 1
INTRODUCTION

This chapter introduces the project the author worked on alongside the author's team. It also includes the background of the project as well as the aims, vision, and mission of carrying out this project. It will also describe the structure and provide insights into the remaining chapters.

1.1   Background

There is no doubt that oil and gas are vital elements to the growth of the economy. There have been traces of oil trade ever since 1875 BC [1]. In this modern, technologically-advanced society, the demand for oil and gas has only continued to grow stronger. It is used for many modern inventions enjoyed by a vast majority of people, such as vehicles, fuels, medical equipment, agriculture, and many more [2]. Additionally, the oil and gas industry has also provided jobs to thousands of individuals [3].

There are many oil and gas reserves in different corners of the world. In Indonesia, in particular, the Energy Ministry has recorded that in January 2021, there is a total reserve of 2.44 billion barrels of oil and 43.6 trillion cubic feet of gas [4]. However, due to the rapidly increasing population and a growing economy, the demand for oil and gas in Indonesia is rising [5]. Furthermore, 50% of Indonesia's energy is derived from oil [5]. This reliance on oil results in Indonesia importing nearly 350,000 barrels per day (BPD) and 50,000 barrels of fuel per day from other countries [5].

Oil and gas have many uses and have a substantial impact on the economy of a country. Therefore, oil and gas industries often make use of dashboard-based software applications in order to help them manage it, such as a Virtual Data Room (VDR) application. A VDR is an online repository that can store data securely and can be accessed by multiple users simultaneously [6]. These kinds of applications can help the oil and gas industries discover which areas could have more oil and gas. It can also help clients visualize the oil and gas data. Lynx and INTViewer are examples of software applications capable of data visualization [7] [8]. These applications are similar, yet they also have their differences. Lynx offers petroleum data services and geophysical and Geographical Information System (GIS) services [7]. It offers 2D and 3D seismic viewers and costs at least £250 per user per year [7]. On the other hand, INTViewer is a platform that allows users to check seismic data, geospatial integrity, and also process datasets [8]. It can cost up to $4,000/person a year [9]. These types of applications can benefit oil and gas industries greatly; however, they tend to be expensive. Additionally, according to the product owner, Mr. Ardimas Andi Purwita, some oil and gas companies in Indonesia often search for cheaper and custom software similar to Lynx and INTViewer. Therefore, the goal of this project is to develop a VDR application with similar features intended for the oil and gas industries in Indonesia at a lower cost. The proposed VDR application will be developed as a website application as part of the requirements from the author's client, for example, PT Geodwipa Teknika Nusantara [10].

The oil and gas industry contains lots of data, data that can be used to obtain more information. This is where data science comes in. Data science is the method of obtaining meaningful insights from a large set of data [11]. The data will be analysed

and processed so that high-level data analysis can be performed [11]. The data analysis will then reveal patterns in the data, thus enabling users to draw conclusions regarding the data [11]. This is useful for the VDR as users will be able to understand the oil and gas data, thus gaining meaningful insights from it. In a more specific sense, the author's client wishes to have a VDR that contains a predictive model that can predict oil and gas production as shown in the Appendix Figure B.1. This thesis will focus on that issue.

1.2 Scope

This section describes the scope of the author's group as well as the author's individual scope for this project.

1.2.1 Scope of Problem and Solution

The main goal of the project is to develop a VDR application that will benefit the oil and gas industry. As has been mentioned in Section 1.1, there are existing applications for this purpose; however, these applications are expensive. Therefore, the author and the author's team will make use of open-source libraries and hand-pick essential features based on the request of the customer. Additionally, the author and the author's team will also develop custom features requested by the customers. The application will consist of several features, such as uploading and storing files, deleting and downloading files, viewing maps as well as obtaining oil and gas production data with the use of a predictive model.

1.2.2   Scope of Work

In this project, the author and the author's team had different responsibilities, as shown in Table *1.1*. The author's responsibility was to create a predictive model capable of predicting oil and gas production. The author had to collect and scrape valuable data in order to make a dataset. This dataset would then be processed and cleaned to train the model. The author will make use of machine learning algorithms to predict the oil and gas production values. Furthermore, the author will conduct a comparative study of models to see which models perform the best at predicting oil and gas production. From this study, the author will choose the model which will be implemented in the VDR application. After the model has been selected, trained, and evaluated, the author will connect the model to the website created by the other members of the author's team.

Table 1.1: Scope of Activities

Name	Role
Kotrakona Harinatha Sreeya Reddy	Collecting and Processing Data Using the data collected to develop predictive models and evaluate it Develop endpoints in backend for oil and gas prediction
Elizabeth Chan	Design the frontend of the proposed VDR application that uses GIS Testing (e.g., unit test & integration test)
Vicky Vanessa	Designing UI/UX of the frontend of the website application Visualizing the data of oil and gas Testing

1.3   Aim and Benefits

This section will cover the aim and benefits of this project.

1.3.1   Aims

The author aims to build a predictive model capable of predicting oil and gas production values. There are already sensors in place which can measure the value of the pressure and temperature in the wells. Therefore, machine learning can be used to predict oil and gas production through these sensors, eliminating the need to extract the oil and gas. This will help the oil and gas industry by showing areas that are more likely to contain more oil and gas. Hence, the oil and gas industry can focus on the wells which contain more oil and gas, which would save time and money. The model the author builds will be integrated into the website application created by the author's team members. This will be further explained in Chapter 3.

1.3.2   Benefits

The main benefit of this VDR website application is to help the oil and gas industry discover more profitable areas of resources by visualizing oil and gas volume as well as visualizing reserve resources. The author and the team wish to increase the local services in the oil and gas industries as some local companies would like to have their own VDR that matches their needs. Furthermore, this application will help engineers understand complex data.

1.4   Structure

This thesis consists of seven chapters which will be briefly described in this section.

1.4.1 Chapter 1: Introduction

Chapter 1 introduces the author's topic, the scope, objectives, aims, vision, and mission of this project.

1.4.2 Chapter 2: Theoretical Foundation

Chapter 2 describes the fundamental theories behind the predictive models designed by the author. It defines specific terms and provides further insights into the problem.

1.4.3 Chapter 3: Problem Analysis

Chapter 3 will detail the problem even further and describe the works related to the author's project while also briefly describing the model the author intends to train.

1.4.4 Chapter 4: Solution Design

Chapter 4 focuses on the design of the solution devised by the author; it includes data pre-processing as well as how the models will be manipulated.

1.4.5 Chapter 5: Results

Chapter 5 will center on the results obtained from model training and evaluation. It will also analyse the information obtained from the results.

1.4.6 Chapter 6: Discussion

Chapter 6 will describe the key results observed in the thesis and analyse it further.

1.4.7   Chapter 7: Implementation

Chapter 7 will focuses on the API made by the author that is used to connect the

model to the VDR website created by the author's teammates.

1.4.8   Chapter 8: Conclusion and Recommendations

Chapter 8 will conclude all the important results and observations obtained, it will

also discuss recommendations for possible future works

CHAPTER 2

THEORETICAL FOUNDATION

This chapter will delve into the theories and techniques the author used while developing this project. It will probe into how oil and gas are produced in the reservoir. Additionally, it will discuss how the author intends to build the model to predict oil and gas production using machine learning. Afterwards, this chapter will discuss missing data, outliers, and feature correlation in the dataset used to train the model. It will also examine how to evaluate the performance of the model and how the model can be connected to a website. Afterwards, it will delve into the specifics of a VDR, which the author's team intends to develop.

2.1 Oil and Gas in a Reservoir

In order to build an oil and gas predictive model, it is vital to understand how oil and gas are formed in a reservoir and the factors that affect its formation.

2.1.1 Oil Formation

A formula that can be taken into account for oil formation is the oil formation volume factor (B_o). It is the ratio of oil volume and dissolved gas at a specific temperature and pressure that is needed to make one barrel of oil [12]. B_o is either greater than or equal to unity [13].

The equation for the oil formation volume factor is :

$$B_o = \frac{(V_o)pT}{(V_o)_{sc}}.$$

(2.1)

In Equation 2.1, B_o is the oil volume factor, V_o is the volume of oil, $(V_o)sc$ is the volume of oil measured under standard conditions, p is the pressure at the reservoir, whereas T is the temperature at the reservoir [12]. From Equation 2.1, it can be inferred that temperature and pressure are essential factors in the formation of oil. Once the oil reaches the surface, it loses the dissolved gas, which leads to changes in the reservoir oil obtained. First of all, the mass of the oil will reduce as it loses the dissolved gas, then the oil will also contract as temperature decreases on the surface [12]. Afterwards, the oil will again expand as the pressure increases [12]. Often the effect of the temperature and pressure changes when the oil reaches the surface is minimal and will cancel out each other [12].

2.1.2 Gas Formation

A formula that can be taken into account for gas formation is the gas formation volume factor (B_g). It is the ratio of the volume of gas at a specific temperature and pressure that is needed to manufacture one standard volume of gas [14]. This equation for gas formation volume factor can be expressed as :

$$B_g = \frac{V_{p,T}}{V_{sc}} \tag{2.2}$$

In Equation 2.2, B_g is the gas formation volume, $V_{p,T}$ is the volume of gas at the reservoir pressure and temperature and V_{sc} is the volume of gas at standard conditions.

In real life, gases follow the real gas law, which can be expressed mathematically as :

$$pV = znRT, \tag{2.3}$$

where p is the pressure, V is the volume, n is the number of moles of gas, R is the universal gas constant, T is the temperature, and z is the gas compressibility factor [15]. Variable z can be expressed as :

$$z = \frac{V_a}{V_i},$$

(2.4)

where V_a is the actual volume of n-moles of gas at a certain temperature and pressure, and V_i is the ideal volume of n-moles of gas at the same temperature and pressure [15]. Therefore, the equation for real gas law should be applied to Equation 2.2. Equation 2.3 is applied onto Equation 2.2 by substituting for the volume (V), which will result in Equation 2.5.

$$B_g = \frac{zTP_{sc}}{T_{sc}P}.$$

(2.5)

In Equation 2.5, B_g is the gas formation volume, P is the pressure, T is the temperature, P_{sc} is 1 atm, T_s is 60°F, and z is the gas compressibility factor at standard conditions (1.0) [15]. With the assumption that the standard conditions are represented by $P_{sc} =$ 14.7 $psia$ and $T_{sc} = 520$, Equation 2.5 can be reduced to :

$$B_g = 0.0283\frac{zT}{P}.$$

(2.6)

2.1.3 Pressure and Temperature in Oil and Gas Formation

Figure *2.1* shows the phase diagram of oil and gas in a reservoir. As stated previously in Section 2.3.1, when oil is drilled, it also contains dissolved gas. Therefore, in a reservoir, there exist 2 phases, namely liquid and gas. Based on the current pressure and temperature, the phase diagram shows that there is a region where the mixture will be either liquid or gas only and a region where both liquid and gas are at equilibria.

The black line, known as the Bubble Point Line, denotes where both phases begin to appear [16]. Before the bubble point, the only phase that exists is liquid. However, at a constant temperature, as pressure decreases, the total volume of gas increases, whereas the volume of oil decreases [16]. This property is supported by the Le Chatelier's Principle, which states that an increase in volume or decrease in pressure would increase the formation of the gaseous product.
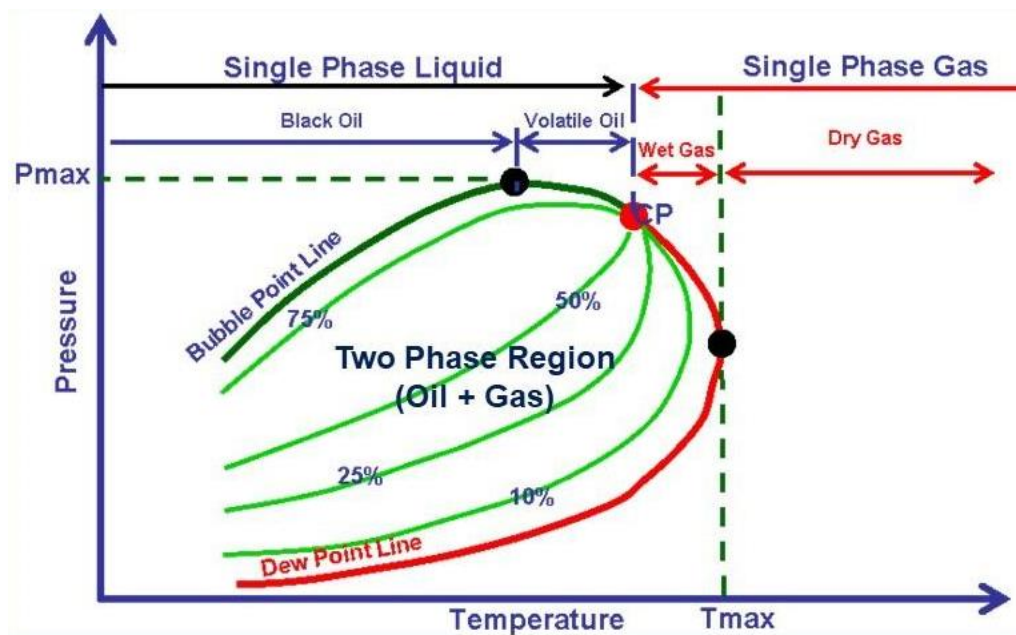
Figure 2.1: Phase Diagram of Oil and Gas *[16]*

As the pressure continues to decrease, more heavier molecules become gaseous, increasing the density and viscosity of the gas [16]. Subsequently, there will be a point where only a small portion of liquid remains; this is called the Dew Point [16]. If the pressure drops below this point, the only phase that exists is gas [16].

2.2 Machine Learning

Upon briefly explaining the oil and gas formation process, this section will delve into machine learning. Machine learning is the central machinery in building a prediction model of the oil and gas production data. It is defined as *the capability of a system to be able to learn from data and algorithms to automate the process of solving certain tasks* [17]. Machine learning is a part of Artificial Intelligence (AI) which centers on using data and algorithms to echo the way humans act and learn [18]. Machine learning helps uncover insights, make classifications, and make predictions from the data given in order to aid users [18]. Machine learning depends on a dataset, which is a collection of data that will be regarded as one unit by the machine [19]. This dataset will act as the "training data" for the machine to learn. It is preferable to have large amounts of data as this means the machines would learn more efficiently and be able to solve problems with better accuracy. However, the quantity of the dataset is not the only significant factor in machine learning; the quality of the dataset is also a notable factor. A machine would perform significantly better with a high-quality dataset in contrast to a poor-quality dataset.

Machine learns in different ways, namely, unsupervised learning, reinforcement learning, and supervised learning. Unsupervised learning aims to derive meaningful information from unlabelled data [20]. It is not as commonly used as supervised learning [20]. On the other hand, reinforcement learning is another complex part of AI where the model is trained to make decisions sequentially [21]. The output is dependent on the state of the current input, and the following input would then be reliant on the output of the previous output [21]. Supervised learning is a part of machine learning and artificial intelligence; it is learning by means of mapping

between a set of input variables and output variables [22]. The input variables are fed into the machine learning model, and after the training phase, it will apply what it learned to unknown data [23]. This type of machine learning is one of the most common methods and is usually used for classification and regression problems [23]. There are several types of supervised machine learning models, namely Naive Bayes, linear regression, support vector machine (SVM), KNN, and others [24]. There is also another new form of supervised machine learning, it is known as self-supervised machine learning. In this method, a machine learning model trains by using a part of of the input data to learn another part of the input data [25]. In terms of a prediction model, supervised learning is ideal, especially with limited computational resources.

2.3   Predictive Models

Predictive modelling is a part of machine learning. It is the process of predicting future outcomes from data gathered beforehand. There are different types of predictive model algorithms that can be used to predict values. These algorithms include classification and regression algorithms. Classification is a type of machine learning algorithm that aims to categorizes the data into specific class labels or categories [26]. Classification algorithms include K-Nearest Neighbours, Naïve Bayes and Support Vector Machines [26]. This algorithm can be used for image classification or email spam classification. On the other hand, regression is a type of machine learning algorithm that aims to discover correlations between the dependent and independent variables and predict the continuous values of the output based on the input [27]. Regression algorithms include Linear Regression, Polynomial Regression, and Decision Tree Regression [26]. This algorithm can be used for production prediction, weather prediction or house price prediction [26]. A part of machine learning that can be used for predictive modelling

is deep learning. Deep learning consists of multiple layers of algorithms known as an artificial neural network (ANN). An ANN is designed to behave similarly to a human brain. The simplest ANN consists of a single neuron, also known as a perceptron [28]. These neurons will be stacked on top of one another, which will create layers [28]. Each layer will learn something new and pass it on to the next layer that will learn something else. There are different types of ANN, such as recurrent neural network (RNN) and convolution neural network (CNN). CNN is a kind of neural network that works well for image and video data, whereas RNN works well with sequential data [28] [29]. An extension of RNN is a Long and Short Term Memory (LSTM) network designed to handle situations where RNNs might not be sufficient [30]. In a study [31], a researcher compared the use of deep learning algorithms and tree-based machine algorithms on a variety of datasets for prediction. The research showed that deep learning tends to perform better on unstructured data, such as images or voice [31]. On the other hand, tree-based algorithms function better with tabular structured data compared to deep learning [31]. Another study where a researcher compared tree-based algorithms and deep learning models on a variety of tabular datasets with different learning objectives also gave the same result [32]. Tabular data is a dataset that consists of a set of rows and columns; it is one of the most common types of datasets.

Tree-based algorithms are a well-known part of machine learning, more specifically, predictive modelling. Tree-based regression algorithms are commonly used for predictive analysis of numerical values [33]. This regression model works by investigating the connection between variables [33]. It will determine the value of one variable based on the other variables present [33].

A commonly used algorithm for predictive models is the random forest algorithm [33]. This is a supervised learning algorithm that is based on the ensemble learning method [33]. Ensemble learning is the process of combing the prediction results of several machine learning algorithms [33]. The goal of this is to make the prediction results more accurate. The random forest algorithm combines the predictive results of several decision trees [33]. The respective decision trees do not interfere with one another [33]. There are two steps for the random forest algorithm; the first step is building n decision tree regressors, where n is the number of decision tree regressors [33]. These trees can be modified by specified hyperparameters, such as the strategy best used to split the node into sub-nodes or the function used to measure the quality of the split [34]. The final step would be to take the average prediction values of the decision tree regressors; this average will serve as the final output of the model [33].

Another algorithm for predictive models is the gradient boosting algorithm. This algorithm is based on the concept of boosting [35]. In terms of regression, boosting is a procedure of building strong regressors by combing weak learners [35]. This algorithm has three requirements, namely loss function, weak learners, and additive model.

A loss function would measure how similar the values predicted by the algorithm are to the actual values. In terms of regression problems, the loss function used could be Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Square Error (RMSE), and Coefficient of Determinant (R^2). [35]. Additionally, this algorithm is based on the idea that combing multiple weak learners would result in an accurate

result. The weak learners used in gradient boosting are typically decision trees [35]. Gradient boosting is also an additive model as it adds the weak learners one by one. Every new predictor would gain new knowledge from the error of the previous predictor, and it would work to correct the error, which would result in a better model [35].

2.4   Data Analytics Pipeline

The predictive models have to be trained on a dataset so that they can learn; however, before training, it is vital to understand and clean the dataset used. The steps that can be taken to understand and clean the dataset before model training are shown in Figure 2.2. These steps will be explained further in the upcoming sections.
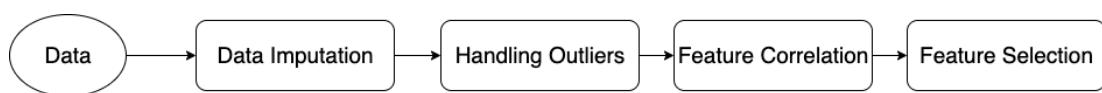
Figure 2.2: Data Analytics Pipeline

2.5   Data Imputation

An essential part of model training is the quality of the dataset. A possible problem in a dataset is missing data. Missing data in a dataset could prove to be problematic as it could affect the model's ability to perform well.

2.5.1   Mechanism of Missingness

There are three possible mechanisms for missing data in a dataset; these mechanisms are Missing Completely at Random (MCAR), Missing at Random (MAR), Missing Not at Random (MNAR).

In the MCAR mechanism, the missing values are unrelated to the other values in the dataset, both missing and present; therefore, these missing values are random. In this situation, the missing values are considered negligible as they would not significantly impact the model performance [36].

In the MAR mechanism, the missing values are also random such as in MCAR; however, there are possibilities of the data in question being dependent on other values in the dataset. In this situation, the missing values should be considered as they could affect the model's performance. However, the effect is not extreme [36].

In the MNAR mechanism, the missing values are strongly dependent on the other values in the dataset, both missing and present. MNAR is the most serious reason mechanism for missing data as it cannot be ignored and could affect the model's performance [36]. In these cases, it is recommended to validate the data collection process [36].

If the mechanism of missingness is either MNAR or MAR, the effects of the missing values could greatly impact the model's performance. The mechanism can be determined by calculating the correlation of missingness between each feature in a dataset. Correlation is known as a statistical measure that describes how one feature is related to another feature [37]. In this situation, a correlation value close to 1 would indicate that if the value of one feature is empty, then the value of the other feature would be empty as well [38]. A correlation value close to -1 would indicate that if the value of one feature is empty, then the value of the other feature would not be empty

[38]. On the other hand, a correlation value close to 0 indicates that there is no or barely any relationship between the features [38].

In order to counteract the effects of the missing values on the dataset, data imputation methods could be used. Data imputation methods include central value imputation and forward filling imputation.

2.5.2 Central Value Imputation

Central value imputation is the process of filling in the missing data in the dataset with their central tendencies [39]. These central tendencies could either be the mean, median, or mode. The mode is typically used to fill in the missing data for categorical variables, whilst the mean and median are often used to fill in for numerical variables [40]. The central tendencies are deemed as reasonable estimates for filling in the missing data. However, this method would not yield ideal results if the missing data follows the MNAR mechanism, and it could also introduce bias in the dataset [41]. Additionally, filling in the missing values with the mean could reduce the variance in the data set [39].

2.5.3 Forward Filling Imputation

Forward filling is the process of filling in the missing data with the value observed before the missing value [42]. For instance, in a dataset such as Table *2.1*, the forward filling imputation method could be used to fill in the missing data. Using this method would change the dataset, as shown in Table *2.2*. This method is generally used for time series datasets and is one of the easiest ways to deal with missing values.

However, a disadvantage of this method is that it will not be able to fill in the missing value if there is no value prior to the missing value.

Table 2.1: Sample Dataset

5	NaN	4
NaN	3	2
3	2	NaN

Table 2.2: Sample Dataset after Forward Filling Imputation

5	NaN	4
5	3	2
3	2	2

2.6 Outliers

Besides missing data, another problem possible in a dataset is the presence of outliers. Outliers can be defined as *a data in a dataset that strays from the other data* [35]. It is necessary to detect these outliers as they could skew the model's training which would reduce the accuracy of the model [43]. The removal of outliers is usually one of the earliest steps in a machine learning problem [43]. There are several methods that can be utilized in order to identify these outliers. One of those methods is to use Tukey's method. The Tukey's Method is based on statistics where data is expected to follow a distribution model such as normal distribution [44]. A data is considered an outlier if it deviates from the model [44]. The Tukey's Method divides the dataset into quartiles; the quartiles commonly used are the lower quartile (Q_1), median (Q_2), and upper quartile (Q_3) [44]. The equation for a quartile is :

$$Q_r = l_1 + \frac{r\left(\frac{N}{4}\right) - c}{f}(l_2 - l_1), \tag{2.7}$$

where Q_r is the rth quartile, l_1 is the lower limit, l_2 is the upper limit, f is the frequency, and c is the cumulative frequency of the class preceding the quartile class [44]. The Tukey's Method involves calculating the Interquartile Range (IQR) between the lower quartile and the upper quartile in a boxplot [44]. The equation for the IQR is

$$IQR = Q_3 - Q_1. \tag{2.8}$$

In order to accurately determine which data is an outlier, the Tukey's Method calculates the upper limit and lower limit of the data distribution. The equation for the upper limit is

$$Upper\ Limit = Q_3 + (1.5 * IQR). \tag{2.9}$$

On the other hand, the equation for the lower limit is

$$Lower\ Limit = Q_1 - (1.5 * IQR). \tag{2.10}$$

The Tukey's Method will remove any data that does not fall between the upper limit and lower limit [44].

2.7   Correlations

In order to better understand a dataset, the correlation between features in the dataset could be considered. As mentioned in Section 2.6.1, correlation is known as a statistical measure that describes how one feature is related to another feature [37]. It is often used during Exploratory Data Analysis (EDA) to gain a better understanding

of how a feature affects other features in the dataset. There are different types of correlations, namely positive correlation, negative correlation, and no correlation [37].

A positive correlation denotes that as the value of a certain feature rises, the value of another feature would rise as well [37]. In a graph format, a strong positive correlation would have a positive gradient, as shown in Figure *2.3*.

A negative correlation denotes that as the value of a certain feature falls, the value of another feature would fall as well [37]. A negative correlation would have a negative gradient, as shown in Figure *2.4*.

No correlation indicates that the features being assessed are not related; therefore, a change in one feature would not impact the other feature [37]. In a graph format, features with no correlation would look like Figure *2.5*.
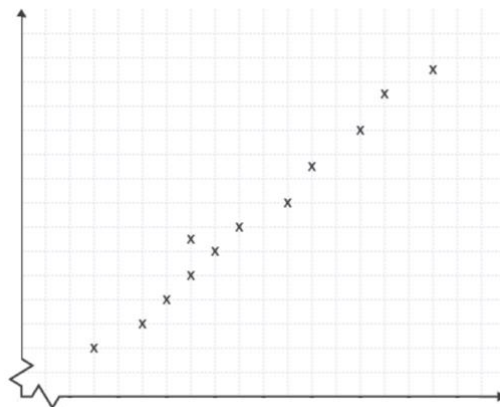
Figure 2.3: Positive Correlation *[45]*
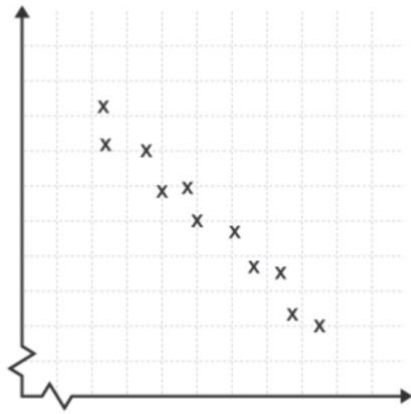
Figure 2.4: Negative Correlation *[45]*
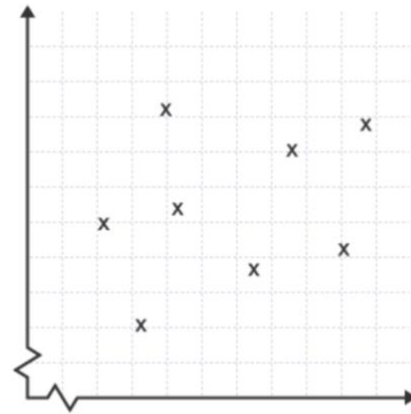
Figure 2.5: No Correlation *[45]*

For numeric features, the commonly used methods for measuring the correlation between features are Pearson Correlation and Spearman Correlation.

2.7.1 Pearson Correlation

In Pearson correlation, the features being compared get assigned a value between -1 and 1 [46]. A correlation value of 1 or – 1 would mean that the features being compared are strongly related to one another. A correlation value of 1 expresses that if one feature is present, then the other feature will unquestionably be present as well [46]. In addition to this, a correlation value of -1 would mean that if one feature is present, then the other feature will undeniably be absent [46]. There are also possibilities of having a correlation value of <1 or >-1. This means that the correlation is almost exactly positive or negative; however, there exists a small number of records that behave differently [46]. On the other hand, a correlation value of 0 would mean that the absence or presence of a feature is in no way related to the presence or absence of another feature [46].

The equation for Pearson correlation is :

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}},$$ *(2.11)*

where r is the Pearson correlation coefficient, x is the values in the first set of data, y is the values in the second set of data, and n is the total number of values.

2.7.2 Spearman Correlation

The Spearman correlation is a method that measures the strength and direction of the relationship between two features in a dataset [47]. Spearman correlation requires continuous data, which has a monotonic relationship. This means that when one feature increases, the other feature could either increase or decrease [47]. However, the relationship between the features does not have to be linear [47]. The correlation values in Spearman correlation follow the same principle as those in Pearson correlation. The values range from -1 to 1 as well. If the correlation value is -1, then as one variable increases, the other variable would decrease [47]. If the correlation value is 0, then a change in a variable would not affect the other variable [47]. On the other hand, if the correlation value is 1, then as one variable increases, the other variable would increase as well [47].

There are two equations that can be used to calculate Spearman's correlation. The first equation is :

$$\rho = 1 - \frac{6\sum_i d_i^2}{n(n^2 - 1)},$$ *(2.12)*

where ρ is the spearman correlation, d_i is the difference between the features, and n is the total number of values [48]. Equation 2.12 can only be used if there are no

duplicates in the dataset. If duplicates exist in the dataset, then the second equation will be used. The second equation is :

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 (y_i - \bar{y})^2}},$$

(2.13)

where ρ is the spearman correlation, x is the value of feature x, $\bar{x}$ is the mean of feature x, y is the value of feature y, and $\bar{y}$ is the mean of feature y [48].

2.8   Feature Selection

After understating the dataset, the process of feature selection could be implemented. Feature selection is the process of cutting down the input variables which will be fed into the models [49]. This is useful as it gets rid of the noise in the dataset so that the model can focus on valuable information [49]. In order to determine which features are ideal to be used in the dataset, the Pearson correlation of the features should be taken into consideration as the values in the dataset are numerical [50]. It is ideal to add highly correlated features for the model's training. However, highly correlated parameters should not be the only features added to the model as they could reduce the model's accuracy [50]. It would lead to a lack of variation in the data or even result in data leakage, which would make the model perform unrealistically well [50].

2.9   Model Training and Hyperparameter Optimization

After the dataset has been cleaned and processed, the model will be trained on that dataset. A model consists of parameters that control the way it learns on the dataset, otherwise known as hyperparameters [51] [52]. Every algorithm has its own hyperparameter that can be defined by the data engineer. The gradient boosting algorithm and random forest algorithm also have several hyperparameters such as

number of trees or maximum depth. These hyperparameters can affect the model's performance, therefore it is imperative to find the best value for each hyperparameter [51] [52]. The process of discovering the ideal hyperparameters for a model is known as hyperparameter optimization or hyperparameter tuning [51] [52]. There are different methodologies for hyperparameter tuning, namely manual tuning, grid search, randomized search and Bayesian optimization [53].

2.9.1 Manual Tuning

Manual tuning is the process of altering the hyperparameters manually to check the performance of the model with the hyperparameters specified [53]. After noting the model's performance, the hyperparameter is altered again manually without using any automation [53]. This is a time-consuming process, however it helps the engineer gain an in-depth understanding of the hyperparameters [53].

2.9.2 Grid Search

In grid search optimization, a set of hyperparameters is defined from the beginning, and the model will train on all of it [53]. Afterwards, the best set of hyperparameters is returned. This is an exhaustive search process as every possible combination defined will be covered [53]. However, it could take a long time to compute, especially if there are a lot of hyperparameters that need to be tuned [53].

2.9.3 Random Search

Random search optimization is similar to grid search, however, the hyperparameters will be chosen randomly [53]. The number of times the process will run can be defined by the data engineer [53]. This method is not as computationally taxing as grid search, however there are chances that the parameters will not be explored properly [53].

2.9.4 Bayesian Optimization

Bayesian optimization differs from other hyperparameter optimization methods as it is capable of remembering previous optimization results [53] [54]. It uses this memory to select the next set of hyperparameter to test on. The idea of Bayesian optimization is to reduce expensive computations by remembering the set of hyperparameters that performed well previously [53] [54]. There are four vital parts in Bayesian optimization, the objective function, domain space, optimization algorithm and the result history [54]. The objective function is the loss or error of a machine learning model based on the hyperparameters [54]. The domain space is the set of hyperparameter values defined beforehand [54]. The optimization algorithm is the method to choose the next set of hyperparameters. The result history is the outcomes of the objective function for every iteration [54].

2.10 Model Evaluation

After the model has been trained, it is time to evaluate the performance of the model. Model evaluation is vital as it allows researchers to determine whether or not the model made is accurate. In order to evaluate models, researchers make use of metrics; the metrics for regression models are MAE, MSE, RMSE and, R^2. The MAE, MSE, and RMSE metrics greatly penalize outliers as their' value increases significantly in the presence of outliers [55]. For these metrics, a higher value indicates poor performance. However, RMSE is generally preferred over MAE and MSE as RMSE uses the same units as the variable in the y-axis [55]. The R^2 metric is also another ideal metric to consider as it is able to explain how well the model can predict the value compared to the original value [55].

2.10.1 Root Mean Square Error

This metric is the root squared average difference between the actual value and the predicted value [55]. RMSE is the square root of the MSE metric. The lower this value, the lower the deviations between the actual and predicted values [55].

The formula for RMSE is,

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_p - y)^2}{n}} \, , \qquad (2.14)$$

where y_p is the predicted value, y is the actual value, and n is the number of values [55].

2.10.2 Coefficient of Determinant

This metric is the measure of how well the regression model has predicted the value based on the actual value [55]. R^2 generally ranges from 0 to 1; however, there are instances when the value could be negative [55]. A R^2 value closer to 1 would mean that the model gives an accurate prediction. The formula for R^2 is

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_p - y)^2}{\sum_{i=1}^{n}(\bar{y} - y)^2} \, , \qquad (2.15)$$

where y_p is the predicted value, y is the actual value, and $\bar{y}$ is the average of the actual values [55].

2.11 REST API

Sometimes the models developed might be used by external applications such as websites. In these cases, REST API could be used to connect the models to the external

application. Representational State Transfer (REST) is a type of architectural style that specifies principles that will act as a guide for website architecture design [56]. The REST API allows users to access web services in a simple manner. Users use HTTP methods, namely GET, POST, DELETE, PUT, and PATCH, to operate the resources such as websites [56]. The GET method is mainly used to read information; this method does not allow information modification [56]. The POST method is used to create new resources which are subordinate to another parent resource [56]. The DELETE method is used to delete an existing resource [56]. The PUT method is used to update a resource that is present; if the resource specified is not present, then a new resource could be generated [56]. The PATCH method is also used to update resources, similar to the PUT method [56]. However, the PATCH method only performs partial updates; it will not wholly change the resource [56]. Unlike the PUT method, the PATCH method is not capable of creating a new resource [56].

REST API is the ideal method to connect to an external application as, based on the adoption trend, REST API is widely accepted by many developers [57]. Furthermore, REST separates the client side and server side, which is advantageous as if one component fails, it would not impact the other components [58]. In addition to this, REST is capable of adapting to any type of schema or platform [58].

There are several frameworks that can be used to create REST APIs such as Flask or FastAPI. Flask is lightweight and modular and is 100% compliant with WSGI which makes it easy to deploy for production [59]. However, it can also be time-consuming to use for large projects [59]. Furthermore, the Flask framework also relies on a lot of dependencies [59]. On the other hand, FastAPI is a new framework that has gain lots

of attention [59]. It is contains all the features Flask contains, however it's speed is one of the features that distinguishes FastAPI from Flask. It can outperform Flask by 100% and many consider FastAPI as the fastest python web framework [59]. Additionally, it is quite convenient to test FastAPI endpoints using the TestClient provided [59]. However, in FastAPI, as everything is tied to the file where the FastAPI app is defined, the main file tends to be crowded [59].

After developing an API, Docker can be used to package it. Docker is an open container-based platform which allows users to deploy and control applications on it. Docker provides a consistent and isolated environment for each container [60] [61]. Each container can access the resource they require without disrupting another container [60] [61]. Therefore, it reduces the chances of issues such as downtime. Furthermore, any application can be removed cleanly by simply deleting the container [60]. Another advantage of using Docker is that it is portable [60] [61]. It can run on any platform as long as the host operating system supports Docker. It can be deployed to any system that supports Docker as all the dependencies will be packaged into a container [60] [61]. In addition to this, Docker also comes with an in-built version control system. It allows users to roll back to a previous version of a Docker image in case there are issues with the current version [60] [61]. Docker is also considered more secure as every application remains isolated from other applications. Therefore, a container cannot access another container without authorization [60] [61].

2.12 Virtual Data Room

As mention in Section 2.11, an API can be used to connect a website to a machine learning model. A VDR is a type of data room that can be developed through a website in which the concept of Software Development Life Cycle (SDLC) will be used during the development process. A data room is a valuable tool for the oil and gas industry as the companies can use it whenever they desire to dilute equity in assets [62]. The company places the data in the data room where it can be assessed [62]. If the company wishes to sell the data, buyers can visit the data room and inspect the data [62]. There are different types of data rooms, namely, Physical Data Room (PDR), VDR, and a PDR – VDR combination [62].

PDR is a physical secure room where the data is placed by the seller [62]. However, PDR has mostly been replaced by VDR as it has several drawbacks. A few of the drawbacks of a PDR are that it is expensive, burdensome, and time-consuming compared to VDR [62]. A VDR is a website where documents are uploaded that users can access and asses at ease [62]. A VDR is as secure as PDR and is also always available for as long as the client desires [62]. Furthermore, as a VDR can be accessed online, there is no need for the client or their representative to travel to access the data [62]. The information in the VDR can also be updated immediately to show new information and is immediately accessible [62]. Based on [63], the features of a VDR could include :

- accessible anywhere and anytime, disregarding operating systems,
- downloading and generating documentations such as reports, and
- petrotechnical solutions such as reservoir analysis or exploration and production tools.

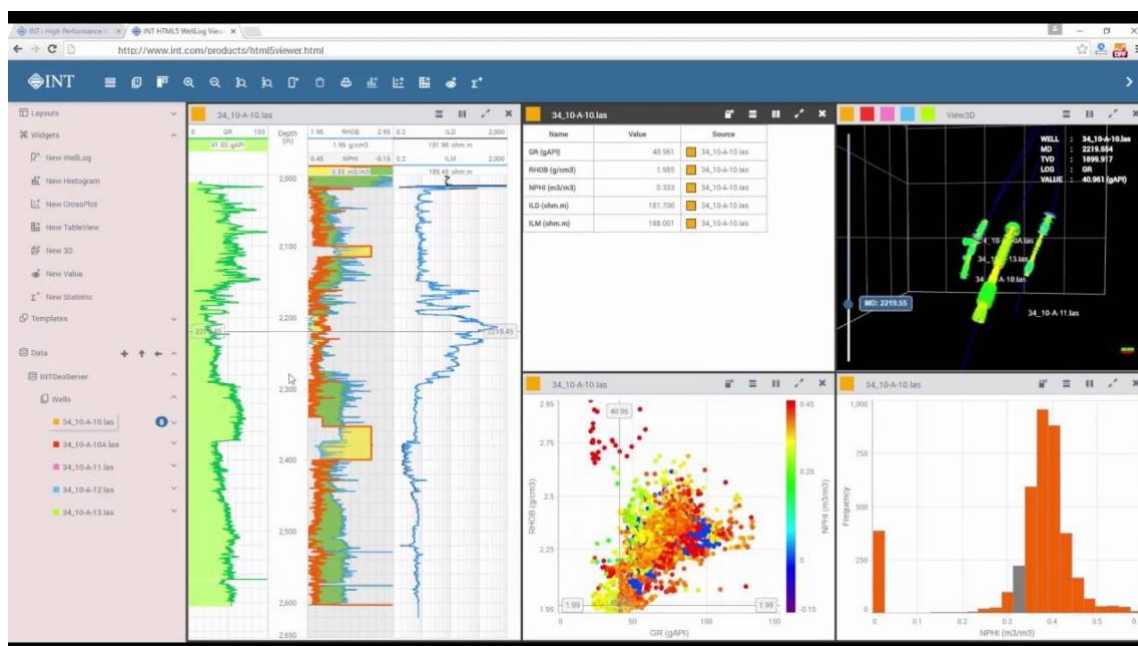Figure 2.6 shows how oil and gas data visualization would look like in a VDR.

Figure 2.6: Dashboard in a Virtual Data Room

CHAPTER 3

PROBLEM ANALYSIS

This chapter will discuss the problem statement of this project as well as the proposed solution for the problem. It will also discuss existing works done in this field.

3.1 Problem Statement

As has been mentioned in Chapter 1, the main goal of this project is to create a VDR website application to help the oil and gas industry. The author wants to enhance the VDR website application by providing a way for users to see which areas contain more oil and gas. Oil and gas companies have to drill exploratory wells to discover whether or not there is a presence of oil and gas [64]. If the presence of oil and gas is detected, then the company would drill more wells, known as development wells, to obtain the oil and gas [64]. However, development wells do not always contain a large amount of oil and gas; therefore these wells end up being abandoned [64]. Abandoning these wells would mean that both the time and money of the company have been wasted.

3.2 Related Works

In [65], Xie, Chao, Qin, and Li made use of 2 models to predict the concentration of gas. Xie et al. used a LSTM model and a random forest model and compared the results. An LSTM model is a variant of RNN, which is capable of remembering past information, which makes it suitable for predicting features. However, as mentioned in Section 2.5, deep learning models do not perform as well as tree-based algorithms when it comes to structured tabular data. On the other hand, the random forest algorithm is a tree-based algorithm that performs exceptionally well on tabular data. In this study, the models were evaluated with the R-squared score, RMSE, and MAE. The LSTM has an R-squared value of 0.31, an RMSE value of 0.45, and a MAE value

of 0.56. On the other hand, the random forest model has a R-squared value of 0.95, RMSE value of 0.23, and MAE value of 0.34. From the values of these evaluation metrics, the researchers concluded that the random forest model was simpler and gave better results than the LSTM model.

In [66], Chauhan made use of Facebook's Prophet model in order to predict gas production. The dataset used by Chauhan was Canadian's natural gas production; the dataset contained two columns which were the date and the volume of gas. The model was evaluated with the R-squared score and the MAE metric. The R-squared value was 0.911, whereas the MAE score was 7782. An advantage of using the Prophet model is that the results are easy to understand [67]. However, it requires a large dataset as it is recommended to have at least two or three years of historic data [67]. The Prophet model works better if the dataset contains daily and weekly observations [67]. Furthermore, though this model performs quickly, the results are often less accurate compared to when other algorithms are used [68].

Chahar in [69] shows the performance of linear regression in predicting oil production. The dataset used was the Volve dataset which is located in the North Sea and was updated on a daily basis from 2005 to 2016. The linear regression model was evaluated with the R-squared score; the value was 0.55. An advantage of linear regression is that there are low chances of the model overfitting [70]. Additionally, it is easy to implement and works exceptionally well on variables that have a linear relationship [70]. On the other hand, linear regression models perform poorly when the relationship between variables is non-linear [70]. Furthermore, there are possibilities for linear regression models to underfit [70].

The performance of polynomial regression in predicting oil production was also shown in [69]. The same dataset used for the linear regression model was used for this polynomial regression. The model was evaluated with the R-squared score; the value was 0.95. An advantage of polynomial regression is that it works well even if the variables do not have a linear relationship [71]. Furthermore, the dataset size does not matter, as polynomial regression works well regardless of dataset size [71]. On the other hand, polynomial regression is extremely sensitive to outliers; the results could change drastically with the presence of one outlier [71].

Table 3.1 summarizes the comparison of the studies mentioned earlier.

Table 3.1: Summary of Research

Title	Model Used	Metrics / Performance	Predicted Feature	Reference
"Using a stochastic forest prediction model to predict the hazardous gas concentration in a one-way roadway"	LSTM	R-squared : 0.31 RMSE : 0.45 MAE: 0.56	Gas Concentration	[60]
"Using a stochastic forest prediction model to predict the hazardous gas concentration in a one-way roadway"	Random Forest	R-squared : 0.95 RMSE : 0.23 MAE: 0.34	Gas Concentration	[60]
"Using Facebook Prophet for Forecasting Natural Gas Production"	Facebook's Prophet	R-squared : 0.911 MAE : 7782	Gas Production Value	[62]
"Prediction of Oil Production by applying Machine Learning on Volve Field Production Data."	Linear Regression	R-squared : 0.55	Oil Production Value	[65]
"Prediction of Oil Production by applying Machine Learning on Volve Field Production Data."	Polynomial Regression	R-squared : 0.95	Oil Production Value	[65]

3.3   Proposed Solution

Machine learning can help determine how likely the well would contain oil and gas. Through machine learning predictions, users can focus on the wells which contain more oil and gas which would save time and money as they would not waste time on wells that contain less oil or gas. The author plans to develop a predictive model that can predict oil and gas production. This section describes the model the author intends to use as well as how the model will be integrated into the website created by the author's team.

3.3.1   Model Selection

As discussed in Section 3.1, there are several models that have been used in the field of oil and gas production prediction. Amongst these models, the random forest algorithm was shown to achieve one of the best results.

Oil and gas datasets are structured and tabular, thus, as mentioned in Section 2.3, tree-based algorithms are more ideal for tabular datasets [31] [32]. Tree-based models find tabular structured data more natural [31] [32]. Additionally, tree-based models are also deterministic which means that the output is determined solely on the input feature values [31] [32]. This makes it ideal for oil and gas prediction as it oil and gas production values depend on the parameter values. Therefore, this project will use make use of tree-based algorithms.

The author will compare the random forest algorithm and gradient boosting algorithms to determine which one performs better. The gradient boosting algorithm has the capability of giving a more accurate result compared to the random forest algorithm.

This is because in the gradient boosting algorithm, the trees are trained one by one; thus, the current tree is capable of correcting the error of the previous one [72]. The author will test the algorithms on different hyperparameters to determine which hyperparameter would give a better result.

3.3.2 Model Integration

Figure 3.1 shows the flowchart of the model integration.
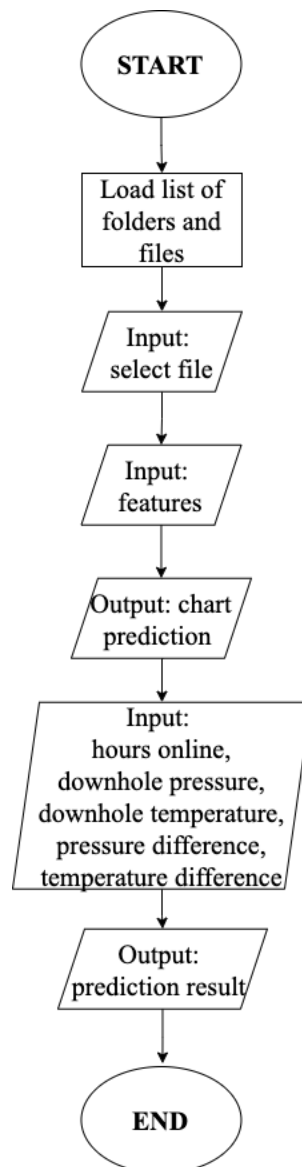

Figure 3.1: Flowchart for Integration

First of all, the premium user will be given a template for the file that contains the data such as the pressure and temperature of the oil and gas wells. The user will be directed to download the template file and then they will be asked upload the file containing the values in the file management section. Afterwards, the user will be asked to choose which feature they wish to predict, either oil or gas. Depending on the feature chosen by the user, the model will predict oil or gas production from this file. The user will then see the visualization of the oil and gas production values. The user will also be able to predict oil and gas production values by manually imputing the values in the website.

CHAPTER 4

SOLUTION DESIGN

Figure 4.1 shows the overall steps that will be taken while building the oil and gas production prediction model. The processes are data cleaning and pre-processing, model training, and model evaluation. This chapter will discuss the overall methodology and delve into the process taken for data cleaning and pre-processing. It will also discuss the experiments that will be conducted on the models in order to determine which would give a better performance.
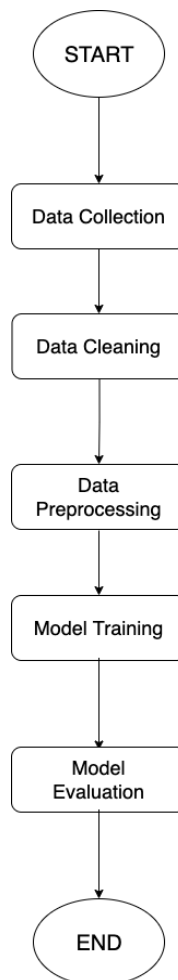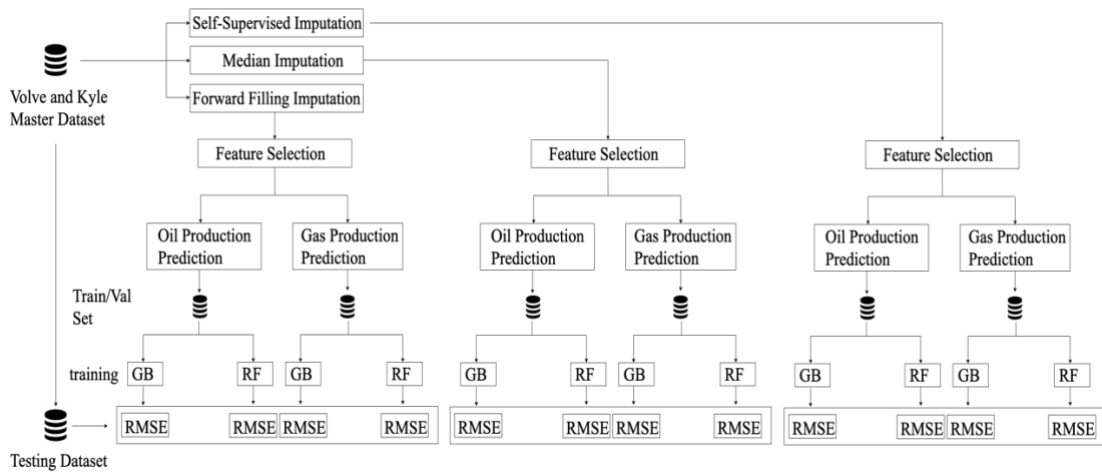

Figure 4.1: Flowchart

Figure 4.2: Methodology

Figure 4.2 describes the overall methodology for data preparation, model training, and evaluation.

4.1 Data Collection

For the project, the author utilized two open-sourced datasets. The first dataset is entitled Volve, whilst the second dataset is entitled Kyle Master. The Volve dataset contained 15,634 rows of data and was obtained from *Kaggle*[1]. On the other hand, the Kyle Master dataset contained 27,324 rows of data and was obtained from the online data centre of the *Oil and Gas Authority*[2]. It is ideal to use a large dataset as it would lead to lower estimation variance, which means the model will be able to predict more accurately. Both Volve and Kyle Master datasets contain valuable information. The columns, their unit of measurement, as well as their meanings are shown in Table 4.1.

---

[1] https://www.kaggle.com/datasets/nazarmahadialseied/volve-field-production-dataset-oil-and-gas

[2] https://experience.arcgis.com/experience/50b61d215bff4072bf0649efe6e8d845/page/Page/?views=by-Field

Table 4.1: Columns in Volve and Kyle Master dataset

Volve Dataset	Kyle Master Dataset	Unit of Measurement	Feature Description
DATEPRD	Date	-	Date
WELL_BORE_CODE	Wellbore ID	-	ID of the wellbore
NPD_WELL_BORE_CODE	-	-	ID of the wellbore per Norwegian Petroleum Directorate
NPD_WELL_BORE_NAME	-	-	Name of the wellbore per Norwegian Petroleum Directorate
NPD_FIELD_CODE	-	-	Field code per Norwegian Petroleum Directorate
NPD_FIELD_NAME	-	-	Field name per Norwegian Petroleum Directorate
NPD_FACILITY_CODE	-	-	Facility code per Norwegian Petroleum Directorate
NPD_FACILITY_NAME	-	-	Facility name per Norwegian Petroleum Directorate
ON_STREAM_HRS	Hours Online	hours	How long the machine has been operating
AVG_DOWNHOLE_PRESSURE	Av. DHP (bar)	bar	Pressure measured at the bottom of the well
AVG_DOWNHOLE_TEMPERATURE	Av. DHT (Deg C)	$^{\circ}$C	Temperature measured at the bottom of the well
AVG_DP_TUBING	-	bar	Pressure build-up in the tubing
AVG_ANNULUS_PRESS	-	bar	Pressure between the tubing and the casing
AVG_CHOKE_SIZE_P	Platform Choke %	%	Size of choke
AVG_CHOKE_UOM	-	%	Unit of measurement
AVG_WHP_P	Av. WHP (bar)	bar	Pressure difference measured at the top of the well
AVG_WHT_P	Av. WHT (Deg C)	$^{\circ}$C	Temperature difference measured at the top of the well
DP_CHOKE_SIZE	-	-	Size of choke
BORE_OIL_VOL	Oil (m3)	m^3	Volume of oil produced
BORE_GAS_VOL	Gas (m3)	m^3	Volume of gas produced
BORE_WAT_VOL	Produced Water (m3)	m^3	Volume of water produced
BORE_WI_VOL	-	-	Volume of injected water
FLOW_KIND	-	-	What kind of well is it (production/injector)
WELL_TYPE	-	-	Type of well

4.1.1 Correlation

The author made use of the concept of Pearson's correlation, which was described in Section 2.9.1, to calculate the correlation between the features. Table 4.2 describes the features with high correlation values in Volve and Kyle Master datasets, before data cleaning. On the other hand, Table 4.3 describes the features with low correlation values in Volve and Kyle Master datasets, before data cleaning.

Table 4.2: Features with High Pearson Correlation in Volve and Kyle Master

Volve		Kyle Master	
Features	**Correlation**	**Features**	**Correlation**
BORE_OIL_VOL and BORE_GAS_VOL	0.999	Av. WHT (Deg C) and Oil (m3)	0.565
AVG_DOWNHOLE_PRESSURE and AVG_ DP _TUBING	0.949	Av. WHT (Deg C) and Gas (m3)	0.552
BORE_WAT_VOL and AVG_CHOKE_SIZE_P	0.760	Av. DHP (bar) and Av. DHT (Deg C)	0.577

Table 4.3: Features with Low Pearson Correlation in Volve and Kyle Master

Volve		Kyle Master	
Features	**Correlation**	**Features**	**Correlation**
BORE_GAS_VOL and BORE_WAT_VOL	-0.009	Platform Choke % and Gas (m3)	-0.0585
BORE_WI_VOL and NPD_WELL_BORE_CODE	-0.055	Produced Water (m3) and Av. DHP (bar)	-0.009
DP_CHOKE_SIZE and AVG_DP_TUBING	0.093	Platform Choke % and Produced Water (m3)	-0.17

4.1.2 Feature Selection

This section will explain and justify which features will be used for the model's training. Table 4.4 shows the features that are selected for model training. As the goal is to create a model that can predict oil and gas production, it is essential to include their production values. In the Volve dataset, the first two features selected for model training are *BORE_OIL_VOL* and *BORE_GAS_VOL*. As mentioned in Section 2.3, oil and gas formation are also reliant on pressure and temperature. Therefore,

AVG_DOWNHOLE_PRESSURE, *AVG_DOWNHOLE_TEMPERATURE,*
AVG_WHP_P and AVG_WHT_P are also included. *ON_STREAM_HRS* will also be
added as this column shows how long the machine operates. In the Kyle Master
dataset, the first two features selected are *Oil (m3)* and *Gas (m3),* as these features
contain the production value of oil and gas. Additionally, as oil and gas production is
reliant on the pressure and temperature of the reservoir, the features *Av. DHT (Deg C),*
Av. DHP (bar), AV. WHT (Deg C), and AV. WHP (bar) are added for the model's
training. Lastly, *Hours Online* will also be added for training the model. Table 4.4
shows the features that have been selected for model training.

Table 4.4: Features selected for training in Volve and Kyle Dataset

Volve Dataset	Kyle Master Dataset	Unit of Measurement	Selected for Model Training
DATEPRD	Date	-	No
WELL_BORE_CODE	Wellbore ID	-	No
NPD_WELL_BORE_CODE	-	-	No
NPD_WELL_BORE_NAME	-	-	No
NPD_FIELD_CODE	-	-	No
NPD_FIELD_NAME	-	-	No
NPD_FACILITY_CODE	-	-	No
NPD_FACILITY_NAME	-	-	No
ON_STREAM_HRS	Hours Online	hours	Yes
AVG_DOWNHOLE_PRESSURE	Av. DHP (bar)	bar	Yes
AVG_DOWNHOLE_TEMPERATURE	Av. DHT (Deg C)	ºC	Yes
AVG_DP_TUBING	-	bar	No
AVG_ANNULUS_PRESS	-	bar	No
AVG_CHOKE_SIZE_P	Platform Choke %	%	No
AVG_CHOKE_UOM	-	%	No
AVG_WHP_P	Av. WHP (bar)	bar	Yes
AVG_WHT_P	Av. WHT (Deg C)	ºC	Yes
DP_CHOKE_SIZE	-	-	No
BORE_OIL_VOL	Oil (m3)	m^3	Yes
BORE_GAS_VOL	Gas (m3)	m^3	Yes
BORE_WAT_VOL	Produced Water (m3)	m^3	No

BORE_WI_VOL	-	-		No
FLOW__KIND	-	-		No
WELL_TYPE	-	-		No

These datasets both have similar columns even though the names are different. For instance, the features *Av. DHT (Deg C)* and *Av. DHP (bar)* in the Kyle Master dataset has the same meaning as the features *AVG_DOWNHOLE_PRESSURE* and *AVG_DOWNHOLE_TEMPERATURE* in the Volve dataset. Additionally, *Oil (m3)* and *Gas (m3)* in the Kyle Master dataset have the same meaning as *BORE_OIL_VOL* and *BORE_GAS_VOL* in the Volve dataset.

4.1.3 Feature Statistics

In order to better understand the selected features in the dataset, several techniques were employed to understand how the data is distributed. Table 4.5 describes the selected features of the Volve dataset, whereas Table 4.6 describes the selected features for the Kyle Master dataset. The histogram, boxplot with and without outliers for these features can be seen in Appendix A in Figure A.7 – A.20.

Table 4.5: Feature Statistics for Volve Dataset

Feature	Range	Outlier Count	Mean	Standard Deviation	Variance
ON_STREAM_HRS	25 hours	715	23 hours	3 hours	9 hours
AVG_DOWNHOLE_PRESSURE	307 bar	144	240 bar	22 bar	484 hours
AVG_DOWNHOLE_TEMPERATURE	107.7 ºC	156	104 ºC	4 ºC	16 ºC
BORE_OIL_VOL	5,900 m^3	283	1,476 m^3	1,464 m^3	2,143,296 m^3
BORE_GAS_VOL	86,863 m^3	182	215,541 m^3	207,094 m^3	42,887,924, 836 m^3
AVG_WHP_P	120 bar	44	48 bar	20 bar	400 bar
AVG_WHT_P	86 ºC	352	73 ºC	18 ºC	324 ºC

Table 4.6: Feature Statistics for Kyle Master Dataset

Feature	Range	Outlier Count	Mean	Standard Deviation	Variance
Hours Online	1,912 hours	1,326	23 hours	27 hours	729 hours
Av. DHP (bar)	1,122 bar	3	111 bar	39 bar	1,521 bar
Av. DHT (Deg C)	245 ºC	645	94 ºC	9 ºC	81 ºC
Oil (m3)	3,509 m^3	447	380 m^3	328 m^3	107,584 m^3
Gas (m3)	1,304,298,362,420 m^3	226	178,525,800,000 m^3	175,599,300,000 m^3	30,835,114,160,490,000,000,000 m^3
Av. WHP (bar)	325 bar	48	57 bar	35 bar	875 bar
Av. WHT (Deg C)	228 ºC	597	62 ºC	19 ºC	361 ºC

In Table 4.5 and Table 4.6, range denotes the range of the specified feature; more specifically, it is the difference between the lowest value up to the highest value of the feature. Outlier count is the number of outliers in the feature. The mean is the average of the feature. Standard deviation is the measure of how varied the feature is relative to the mean.

From Table 4.5 and Table 4.6, it can be seen that the range of values for all the selected features in the Kyle dataset is larger than the features in the Volve dataset. This denotes that the data in the Kyle dataset is more dispersed compared to the Volve dataset. In addition to this, the standard deviation and variance of the features in the Kyle dataset are much larger than the features in the Volve dataset. This observation further supports the fact that the features in the Kyle dataset are more spread out than the features in the Volve dataset.

4.2   Data Cleaning and Pre-processing

Volve and Kyle Master contained missing data; therefore, it is imperative to check the relationship between the features in the dataset. This is done so that it can be

determined whether or not the presence of the missing value is correlated to other values in the dataset. In order to check this, a heatmap was used to see the correlation values on both datasets. The heatmap can be seen in Appendix A in Figure A.1 and Figure A.2. Table 4.7 describes the observations derived from the heatmaps. As stated in Table 4.7, the Volve dataset follows the MNAR mechanism, whereas the Kyle Master dataset follows the MAR mechanism. Section 2.7.1 states that these missing mechanisms imply that the missing values are dependent on one another. Thus, it should not be ignored and should either be deleted or filled in using data imputation methods.

Table 4.7: Observations for Missing Data

Dataset	Volve	Kyle Master
Observation	Contains mainly "<1" and "1" feature correlation values, meaning the features are highly dependent on one another. A value of "<1" denotes that the correlation is almost exactly 1.	Feature correlation values are mostly 0.1, and some features have a correlation value of 1, meaning most of the features do not show much correlation, however, few features are highly correlated.
Missing Data Mechanism	MNAR	MAR

For this project, the author will use three methods and compare them to see which method would make the model perform better. The first method the author will use is forward filling imputation, where the empty value is replaced by the last observed record. The second method used is central value imputation, where the author will fill in the missing values with the median value of the feature. The third method used is a self-supervised imputation method where the missing values will be filled in by a machine learning model. There are data in the dataset where the pressure and temperature values are filled in. Therefore, a baseline gradient boosting and random forest model is used to predict these values. Afterwards, the author placed these values into the original dataset.

For the model's training, the author combined both the Volve and Kyle Master datasets. The selected columns in the Volve dataset and the Kyle Master datasets have the same meaning. Therefore when combining the datasets, the columns in the Volve dataset were renamed to match the columns in the Kyle Master dataset. Table *4.8* describes the selected features of the combined dataset after forward filling imputation is used. Additionally, Table *4.9* describes the selected features of the combined dataset after median imputation is used. On the other hand, Table 4.10 describes the selected features of the combined dataset after self-supervised imputation is used. The boxplot, histogram, and kernel density function for these features can be seen in Appendix A in Figure A.21 – A.34.

Table 4.8: Feature Statistics for Volve and Kyle Master Dataset after forward filling imputation

Feature	Range	Mean	Standard Deviation	Variance
Hours Online	1,912 hours	23 hours	22 hours	484 hours
Av. DHP (bar)	162 bar	102 bar	19 bar	361 bar
Av. DHT (Deg C)	344 ℃	158 ℃	74 ℃	5,476 ℃
Oil (m3)	5,900 m^3	801 m^3	1,117 m^3	1,247,689 m^3
Gas (m3)	1,164,213 m^3	176,375 m^3	178,282 m^3	31,784,471,524 m^3
Av. WHP (bar)	325 bar	49 bar	29 bar	841 bar
Av. WHT (Deg C)	228 ℃	64 ℃	18 ℃	324 ℃

Table 4.9: Feature Statistics for Volve and Kyle Master Dataset after median imputation

Feature	Range	Mean	Standard Deviation	Variance
Hours Online	1,912 hours	22 hours	19 hours	361 hours
Av. DHP (bar)	307 bar	182 bar	69 bar	4,761 bar
Av. DHT (Deg C)	108 ℃	100 ℃	4 ℃	16 ℃
Oil (m3)	5,900 m³	733 m³	927 m³	859,329 m³
Gas (m3)	13,044,298 m³	155,145 m³	162,085 m³	26,271,547,225 m³

Av. WHP (bar)	325 bar	45 bar	23 bar	529 bar
Av. WHT (Deg C)	228 ºC	69 ºC	17 ºC	289 ºC

Table 4.10: Feature Statistics for Volve and Kyle Master Dataset after self-supervised imputation

Feature	Range	Mean	Standard Deviation	Variance
Hours Online	1,912 hours	23 hours	22 hours	484 hours
Av. DHP (bar)	308 bar	163 bar	70 bar	4,900 bar
Av. DHT (Deg C)	105 ºC	99 ºC	5 ºC	25 ºC
Oil (m3)	5,888 m^3	827 m^3	1109 m^3	1,229,881 m^3
Gas (m3)	13,044,297 m^3	194,119 m^3	190,655 m^3	36,349,329,025 m^3
Av. WHP (bar)	325 bar	49 bar	28 bar	784 bar
Av. WHT (Deg C)	228 ºC	66 ºC	18 ºC	324 ºC

Table *4.8*, Table *4.9*, and Table 4.10 show that the mean and standard deviation of the datasets after data imputation have changed slightly. Most of the mean and standard deviation of the features in Table 4.10 is greater than the features in Table *4.8* and Table *4.9*. This shows that the distribution of the dataset after self-supervised imputation is more dispersed compared to the dataset after forward filling and median imputation.

4.3   Data Splitting

The combined dataset will then be split into three sets, namely a training set, testing set and a validation set. The purpose of the training set will be to train the model, additionally, the testing set will be used to evaluate the model's performance. The validation set will be used during hyperparameter optimization. The ratio for splitting is 80% for training, 10% for testing, and 10% validation.

4.4   Hyperparameter Tuning

As mentioned in Section 3.3.1, the author will test the algorithms on different hyperparameters to determine which would result in a better performance. The hyperparameters that will be tuned for the algorithms are described in Section 4.4.1 and Section 4.4.2.

4.4.1   Gradient Boosting

In the gradient boosting algorithm, the hyperparameters that will be tuned are the learning rate, number of trees, and the maximum depth. Learning rate is an important hyperparameter as it controls how quickly the model learns [73]. The learning rate hyperparameter is closely related to the number of trees parameters. The number of trees denotes the number of trees that will be used [73]. A fine balance has to be achieved between the learning rate and the number of trees, as the smaller the learning rate, the higher the number of trees should be. It is ideal to use a low learning rate because it would let the model train slower, which makes it more efficient [73]. However, the number of trees should not be too high as it would result in overfitting [73]. The third hyperparameter that the author will tune is the maximum depth. The maximum depth is how deep the tree is allowed to be. The deeper the tree, the more it will split and learns information about the dataset. However, if the depth is too high, it could lead to overfitting [73].

4.2.2   Random Forest

In the random forest algorithm, the hyperparameters that will be tuned are the number of trees, max depth, and the minimum number of samples needed to split the leaf node, also known as min_samples_split. The number of trees and max depth hyperparameters have been explained in the previous section. The other

hyperparameter that the author will tune for this model is the min_samples_split. The lower the value of this parameter, the more the tree will split [74]. If the tree splits too much, it could lead to overfitting [74]. However, a high value for this parameter is not ideal as well. A high value would mean that the tree would not split as much, which would result in the model underfitting [73] [74]. Therefore, it is necessary to find the right number for this parameter.

4.5   Evaluation Technique

After the model has been trained, the model will be evaluated with the RMSE and R-squared metric. As shown in Figure 4.2 , there will be six different datasets, thus there will be twelve models. These twelve models will be evaluated and the best model will be selected and deployed as an API.

4.6   Integration to Website

The system architecture of the VDR website is shown in Figure 4.3. The author will be responsible for developing the API that will be connected to the website. There will be two models selected, one for oil production and another for gas production. The best models will be saved into a pickle file and the author will be using the FastAPI framework to deploy the models. The API will then be packaged into a docker container in the backend.
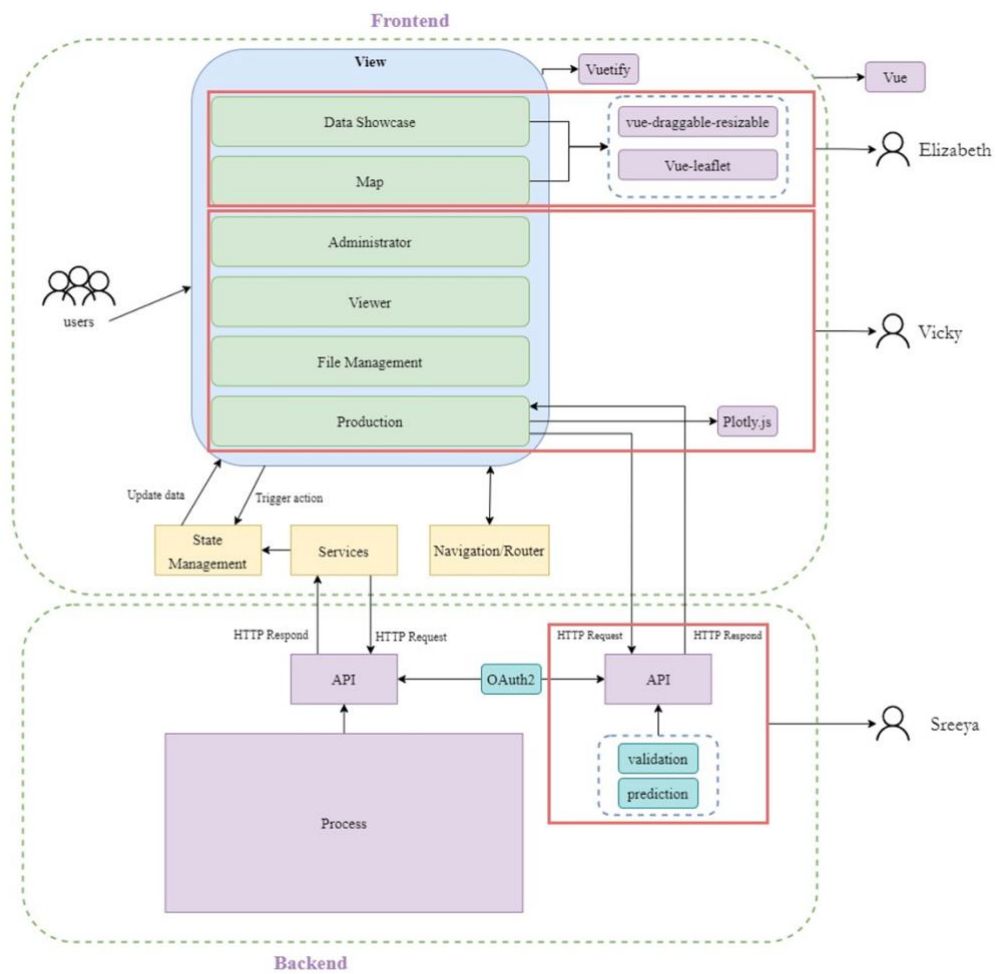
Figure 4.3: System Architecture

CHAPTER 5

RESULTS

This chapter will discuss the results obtained during this project. It will show the results of conducting hyperparameter optimization using the Bayesian optimisation method. Afterwards, it will show the performance of each model that the author compared.

5.1   Hyperparameter Optimization

In order to optimise the models to deliver optimal results, the Bayesian Optimization method could be used. This method was used on the different datasets. For the gradient boosting model, the range for number of trees is 5 to 100, the range for maximum depth is 2 to 50, whereas the range for learning rate is 0 to 1. For the random forest model, the range for number of trees is 5 to 100, the range for maximum depth is 2 to 50, whereas the range for min_samples_split is 2 to 50. Every model went through 1,000 iterations during the hyperparameter optimization before the final hyperparameters were obtained. During the hyperparameter optimization, for each model, the RMSE metric was used on the validation dataset as the performance metric. The results of the hyperparameter optimization for oil production is shown in Table 5.1, whereas the results for gas production is shown in Table 5.2.

Table 5.1: Hyperparameter Optimization for Oil Production

Imputation Method	Model Used	Obtained Hyperparameters	Validation Set RMSE / m3
Forward Filling Imputation	Gradient Boosting	Learning rate : **0.29268132226068777** Maximum depth : **10.0** Number of trees: **27.0**	138
	Random Forest	Maximum depth : **28** Min_samples_split : **9** Number of trees : **120**	128

Median Imputation	Gradient Boosting	Learning rate : **0.19885221540658993** Maximum depth : **11** Number of trees: **41**	173
	Random Forest	Maximum depth : **20** Min_samples_split : **3** Number of trees : **25**	181
Self-supervised Imputation	Gradient Boosting	Learning rate : **0.17000031038373792** Maximum depth : **11** Number of trees: **40**	156
	Random Forest	Maximum depth : **25** Min_samples_split : **2** Number of trees : **34**	160

Table 5.2: Hyperparameter Optimization for Gas Production

Imputation Method	Model Used	Obtained Hyperparameters	Validation Set RMSE / m3
Forward Filling Imputation	Gradient Boosting	Learning rate : **0.20604737829083644** Maximum depth : **20** Number of trees: **50**	64618
	Random Forest	Maximum depth : **50** Min_samples_split : **9** Number of trees : **120**	55816
Median Imputation	Gradient Boosting	Learning rate : **0.19965130013395532** Maximum depth : **11** Number of trees: **41**	52815
	Random Forest	Maximum depth : **24** Min_samples_split : **2** Number of trees : **100**	52793
Self-supervised Imputation	Gradient Boosting	Learning rate : **0.13470450722751402** Maximum depth : **9** Number of trees: **119**	59339
	Random Forest	Maximum depth : **80** Min_samples_split : **2** Number of trees : **100**	54153

5.2   Model Performance and Evaluation

In this section, the performance of each model will be shown and the model will be

evaluated on the test dataset using RMSE as the primary performance metric and R-

Squared as the secondary performance metric

5.2.1 Gradient Boosting

For the gradient boosting model, the model was evaluated by plotting the performance of the training set against the testing set. The graphs are shown in Figure 5.1, Figure 5.2 and Figure 5.3.

Figure 5.1: Gradient Boosting Model with Forward Filling Imputation Dataset; (a) oil and (b) gas

Forward Filling Imputation
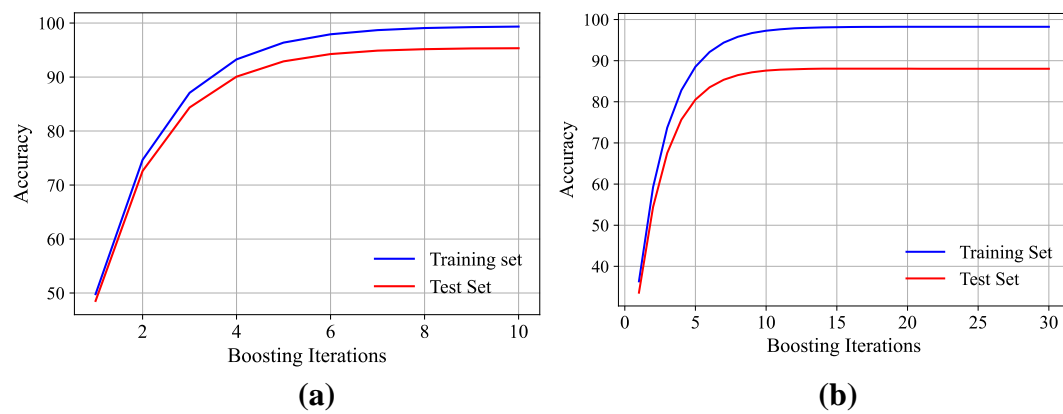
(a) (b)

Figure 5.1: Gradient Boosting Model with Forward Filling Imputation Dataset; (a) oil and (b) gas

Median Imputation

(a) (b)
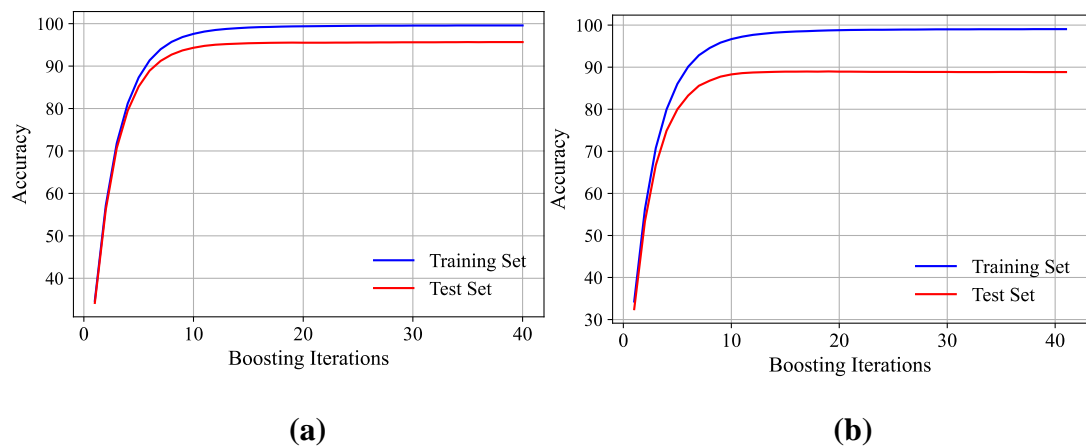
Figure 5.2: Gradient Boosting Model with Median Imputation Dataset; (a) oil and (b) gas

Self-Supervised Imputation

(a)   (b)
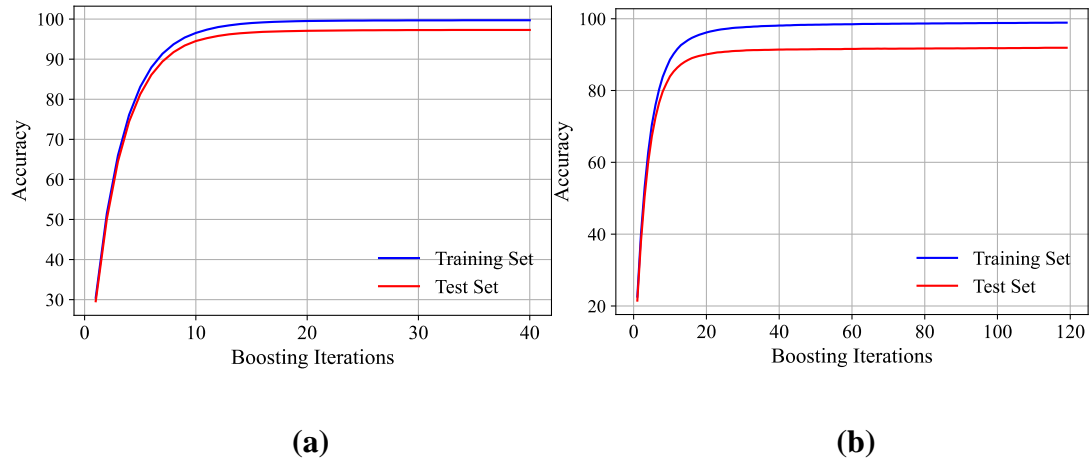
Figure 5.3: Gradient Boosting Model with Self-Supervised Imputation Dataset; (a) oil and (b) gas

Figure 5.1, Figure 5.2, and Figure 5.3 show that for the oil production, the model with the self-supervised imputation dataset gave the best result. It had the highest accuracy and the gap between the training set and the testing set is the smallest when the self-supervised imputation dataset was used. However, the model with forward filling imputation dataset gave the worst result.

For gas production, the model with the self-supervised imputation dataset also gave the best result. It also had the highest accuracy and the gap between the training and testing set is the smallest with the self-supervised imputation dataset. Similar to the oil production model, the model with forward filling imputation dataset performed the worse.

5.2.2 Random Forest

For the random forest model, the model was evaluated by plotting the predicted values against the actual values. The diagrams are shown in Figure 5.4, Figure 5.5, and Figure 5.6.

Forward Filling Imputation

<center>(a)   (b)</center>
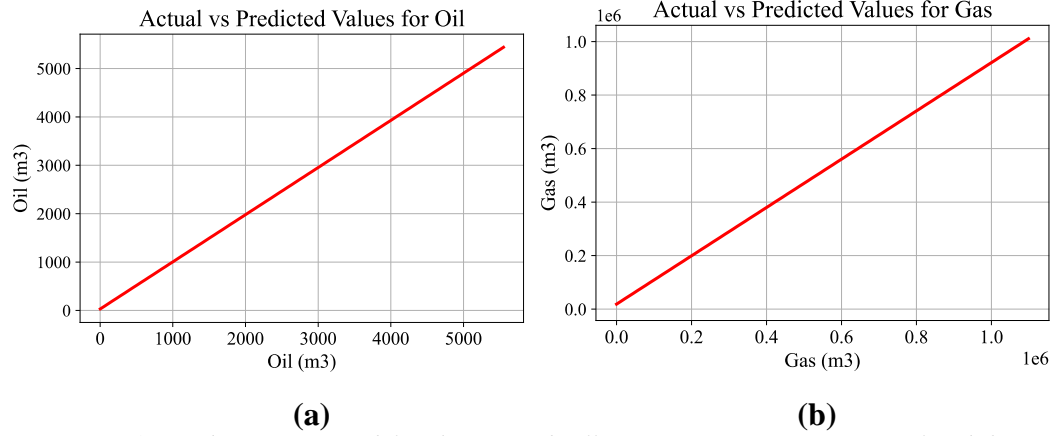
Figure 5.4: Random Forest Model with Forward Filling Imputation Dataset; (a) oil and (b) gas

Median Imputation

<center>(a)   (b)</center>
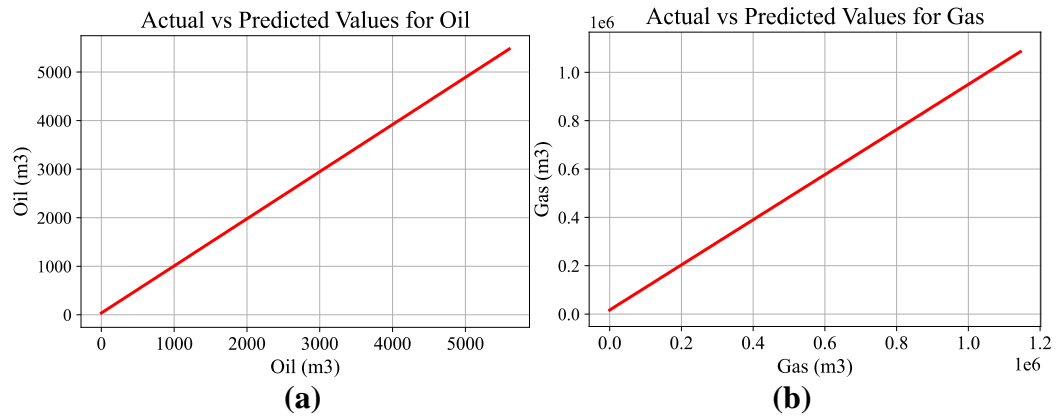
Figure 5.5: Random Forest Model with Median Imputation Dataset; (a) oil and (b) gas

Self-Supervised Imputation

<center>(a)   (b)</center>
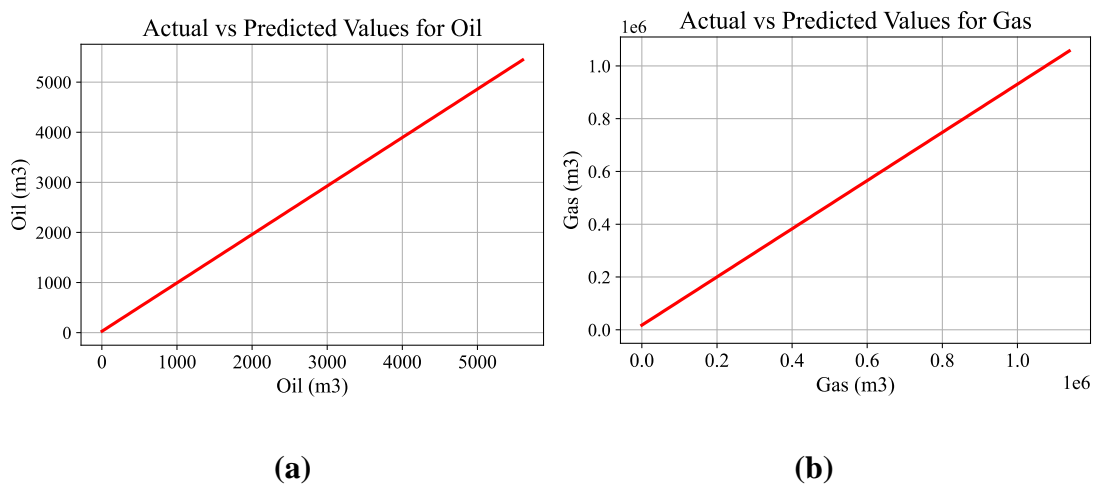
Figure 5.6: Random Forest Model with Self-Supervised Imputation Dataset; (a) oil and (b) gas

For oil production, it can be seen that Figure 5.4, Figure 5.5, and Figure 5.6 are quite similar as the models predicted values that are similar to the actual values. However, it can be seen that the blue dots are somewhat more scattered for the model with the median imputation dataset. This indicates that this model performed the worst.

For gas production, Figure 5.4, Figure 5.5, and Figure 5.6 are quite similar as well. From this it can be seen that all the models predicted values that are similar to the actual values. However, the blue dots are slightly more scattered for the model with the median imputation dataset. Therefore, for gas production the model with the median imputation dataset performed the worse.

5.2.3   Inference time

This section describes the inference time for each model to predict oil and gas production. The inference time was calculated 10 thousand times on a laptop with RAM 16 GB and Intel Core i5. Table 5.3, Table 5.4, and Table 5.5 show the mean inference time and standard deviation for each model.

Forward Filling Imputation

Table 5.3: Inference Time for Models with Forward Filling Imputation Dataset; GB denotes gradient boosting, RF denotes random forest

	GB Oil	GB Gas	RF Oil	RF Gas
Mean Inference Time / sec	0.00449	0.01744	0.0105	0.0415
Standard Deviation / sec	0.000488	0.00147	0.0017	0.00595

Median Imputation

Table 5.4: Inference Time for Models with Median Imputation Dataset; GB denotes gradient boosting, RF denotes random forest

	GBM Oil	GBM Gas	RF Oil	RF Gas
Mean Inference Time / sec	0.00579	0.00571	0.00235	0.0496
Standard Deviation / sec	0.000666	0.000885	0.000578	0.00843

Self-Supervised Imputation

Table 5.5: Inference Time for Models with Self-Supervised Imputation Dataset; GB denotes gradient boosting, RF denotes random forest

	GB Oil	GB Gas	RF Oil	RF Gas
Mean Inference Time / sec	0.00642	0.00812	0.01669	0.0494
Standard Deviation / sec	0.000830	0.000647	0.00205	0.00600

Figure 5.7 shows the inference time for each model against the model's RMSE values for oil and gas production respectively.

(a)   **(b)**
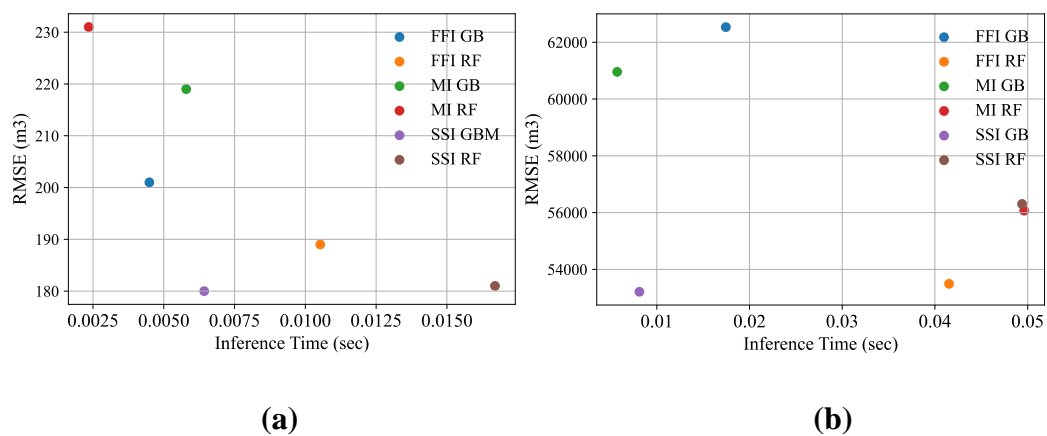
Figure 5.7: Inference Time for Models; FFI denotes forward filling imputation, MI denotes median imputation, SSI denotes self-supervised imputation, GB denotes gradient boosting, RF denotes random forest

For oil production, Figure 5.7 shows that the random forest model with the median imputation dataset performed the fastest while the random forest model with the self-supervised imputation dataset took the longest to predict the oil production value. The models with the highest RMSE values is the random forest model with median imputation dataset. On the other hand, the model with the lowest RMSE value is the gradient boosting model with the self-supervised imputation dataset. From Figure 5.7, it can be inferred that the gradient boosting model with the self-supervised imputation dataset performed the best. There are two reasons why this model is deemed as the best performing model based on Figure 5.7. First of all, although this model is not as fast as the random forest model with the median imputation dataset, it is still one of the fastest performing models. In addition to being fast, it also has the lowest RMSE score, meaning it is more accurate than the other models. Therefore, the gradient boosting model with the self-supervised imputation dataset is the best model for oil production.

For gas production, Figure 5.7 shows that the gradient boosting model with the median imputation performs the fastest while the random forest model with the median imputation dataset took the longest to predict the gas production value. Figure 5.7 also showed that the gradient boosting model with the forward filling imputation dataset had the highest RMSE value. On the other hand, the gradient boosting model with the self-supervised imputation dataset had the lowest RMSE value. Similar to oil production, Figure 5.7 shows that the gradient boosting model with the self-supervised imputation dataset is the best performing model. The reasoning behind this observation is because the model has the lowest RMSE value while still performing fast. Therefore,

the gradient boosting model with the self-supervised imputation dataset is the best

model for gas production as well.

5.2.4   Model Evaluation

The result of the evaluation for oil production is shown in Table 5.6. On the other hand,

the evaluation result for gas production is shown in Table 5.7.

Table 5.6: Evaluation for Oil Production Model

Imputation Method	Model Used	RMSE / m3	R-Squared
Forward Filling Imputation	Gradient Boosting	201	96.3
	Random Forest	189	96.7
Median Imputation	Gradient Boosting	219	95.6
	Random Forest	231	95.1
Self-supervised Imputation	Gradient Boosting	180	97.3
	Random Forest	181	97.2

Table 5.7: Evaluation for Gas Production Model

Imputation Method	Model Used	RMSE / m3	R-Squared
Forward Filling Imputation	Gradient Boosting	62532	88.0
	Random Forest	53490	91.2
Median Imputation	Gradient Boosting	60957	88.8
	Random Forest	56578	90.5
Self-supervised Imputation	Gradient Boosting	53211	91.9
	Random Forest	56367	90.9

From Table 5.6 and Table 5.7, it can be seen that for both oil and gas production, the gradient boosting model with the self-supervised imputation dataset performed the best.

In terms of oil production with the random forest model, the RMSE values indicates that the model with the self-supervised imputation dataset performed 22% better than the model with the median imputation dataset. Additionally, it performed 5% better than the model with the forward filling imputation dataset. On the other hand, the R-Squared values indicate that the model with the self-supervised imputation dataset performed nearly 1% better than the model with the forward filling dataset imputation and around 2% better than the model with the median imputation dataset.

In terms of oil production with the gradient boosting model, the RMSE values showed that the model with the self-supervised imputation dataset performed 18% better than the model with the median imputation dataset. It also performed 11% better than the model with the forward filling imputation dataset. In addition to this, the R-Squared values show that the model with the self-supervised imputation dataset performed 1% better than the model with the forward filling dataset imputation and around 2% better than the model with the median imputation dataset.

In terms of gas production with the random forest model, the RMSE values denotes that the model with the self-supervised imputation dataset performed 1% better than the model with the median imputation dataset. It also performed 5% worse than the model with the forward filling imputation dataset. Additionally, the R-Squared values denotes that the model with the self-supervised imputation dataset performed nearly

1% better than the model with the median imputation dataset and almost 1% worse than the model with the forward filling imputation dataset.

In terms of gas production with the gradient boosting model, the RMSE values conveys that the model with the self-supervised imputation dataset performed 13% better than the model with the median imputation dataset. Furthermore, it performed 15% better than the model with the forward filling imputation dataset. Additionally, the R-Squared values conveys that the model with the self-supervised imputation dataset performed 0.8% better than the model with the forward filling dataset imputation and around 3% better than the model with the median imputation dataset.

CHAPTER 6

DISCUSSION

This chapter will evaluate and clarify the key results obtained in this project and what it means for this thesis.

6.1   Discussion

The results showed that the models were capable of predicting oil and gas production values. Both the gradient boosting and random forest models performed well for predicting oil and gas production values. This shows that tree-based models can perform quite well on tabular structured data.

The results also showed that data imputation method has an effect on the model's performance. The performance of the model changes depending on the dataset used. For oil production, the models with the self-supervised imputation dataset performed better than the models with the median imputation dataset. However, the models with the forward filling imputation dataset performed worse than the median imputation dataset. From this, it can be inferred that the self-supervised imputation method improved the dataset better than the conventional imputation methods e.g. forward filling imputation and median imputation. Therefore, this method could be used in domains where open-sourced datasets are difficult to find such as in oil and gas production. Furthermore, as the model was able to predict oil and gas production values with high accuracy, the dataset can be used by other researchers to work in the field of oil and gas production prediction.

For oil and gas production, amongst all the models that the author evaluated, the models with the self-supervised imputation dataset outperformed the models with the other datasets. It had the lowest RMSE value and the highest R-squared value, making it the best model. The chosen gradient boosting model for oil production had an inference time of 0.00642 seconds measured over 10,000 iterations on a laptop with RAM 16 GB and Intel Core i5. On the other hand, the chosen gradient boosting model for gas production has an inference time of 0.00812 seconds. The inference time was also measured over 10,000 iterations on a laptop with RAM 16 GB and Intel Core i5. The inference time of the model should not be slow as the model will be deployed into an API and then used to predict the oil and gas production values. It would not be ideal if it takes a long time for the users to obtain the results. The results show that the inference time for both models are fast, this shows that model is suitable to be deployed into an API that will be used in the website as it predicts quickly.

There were constraints faced in this thesis. The main constraint faced was the lack of good quality open source datasets for oil and gas production values. These types of data tend to be private company information, hence it was challenging to find good quality datasets. Furthermore, the dataset that was available contained a lot of missing values.

CHAPTER 7
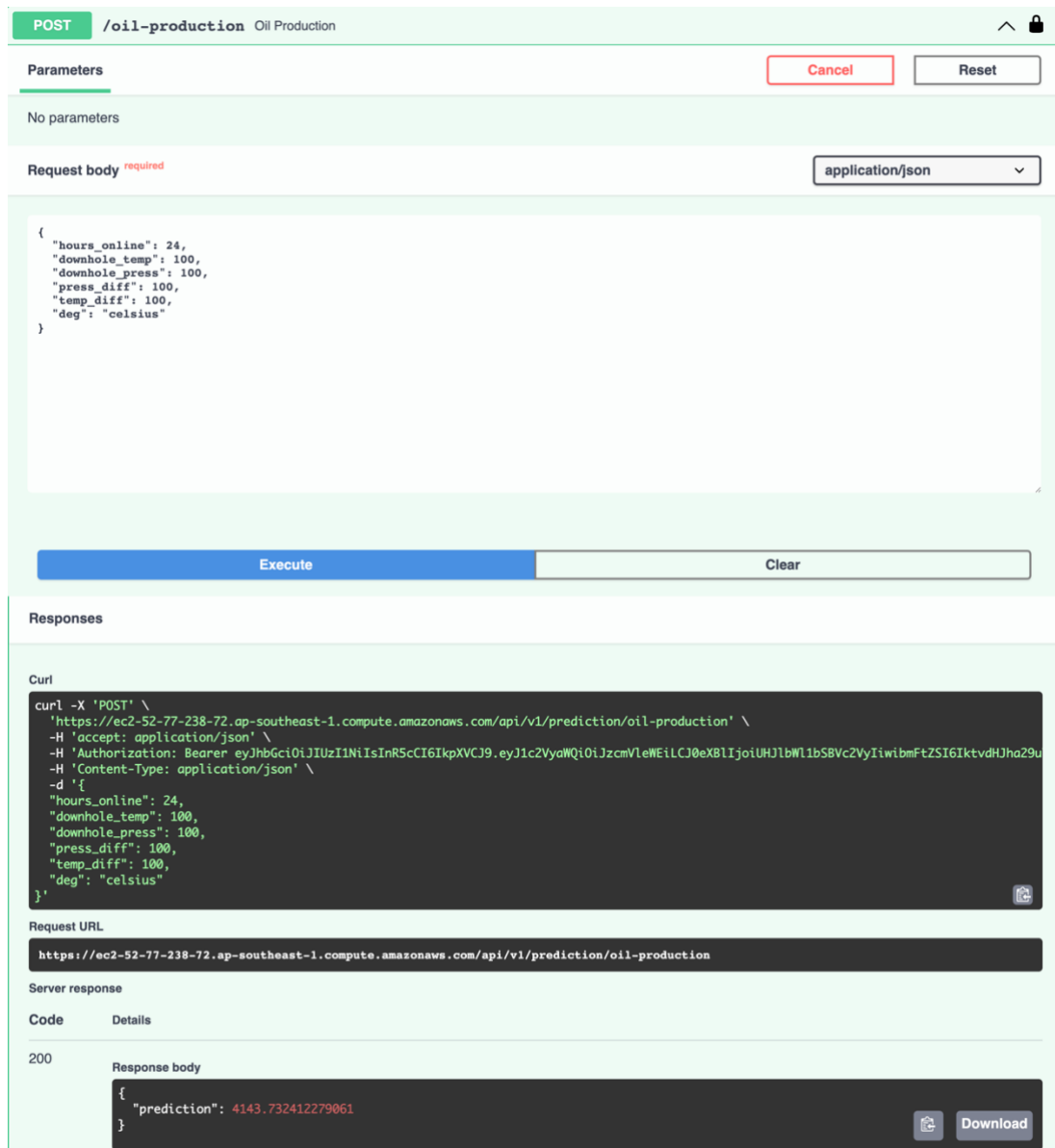
IMPLEMENTATION

This chapter will discuss the endpoints made by the author that is used to connect the model to the VDR website created by the author's teammates.

There are four endpoints created by the author that will be accessed by the author's teammates. Two of these endpoints were made for predicting oil production value while the other two endpoints were made for predicting gas production value. There two methods that the users can use to predict the values, namely singular data prediction and excel file prediction.

7.1 Singular Data Prediction

These endpoints make the user input the feature values manually. Additionally, based on the request from the client, the user will also be able to choose the unit for the temperature values. The user could make the temperature in the form of Celsius or Fahrenheit. Figure 7.1 shows a sample response from the oil production endpoint. The hours_online, downhole_temp, downhole_press, press_diff and temp_diff are values that will be used by the machine learning model to predict the oil production value. On the other hand, deg is where the user can decide whether they wish the temperature values to be in Celsius or Fahrenheit. The gas production endpoint has the same request as the oil production endpoint. In addition to this, as this prediction feature is only accessible by premium users, the status of the users will be checked beforehand. If the user is not a premium user, then it will return an error code, preventing the user from accessing the feature.

Figure 7.1: Oil Production Endpoint

7.2 Excel File Prediction

These endpoint will accept an Excel file as an input and predict oil and gas production values from the file. The file should be uploaded in the file management section, therefore, the user would just have to select the file. The file has to follow the template provided to the premium user. If the file uploaded by the user differs from the template provided, then it will return an error code. Furthermore, the feature values should only consists of numerical values, or else it will return an error code. Additionally, there

should not be any missing values in the excel file, or else it will return an error code.
If all the conditions are met properly, then the endpoint will predict the oil and gas
production values. Figure 7.2 shows a sample response from the oil production excel.
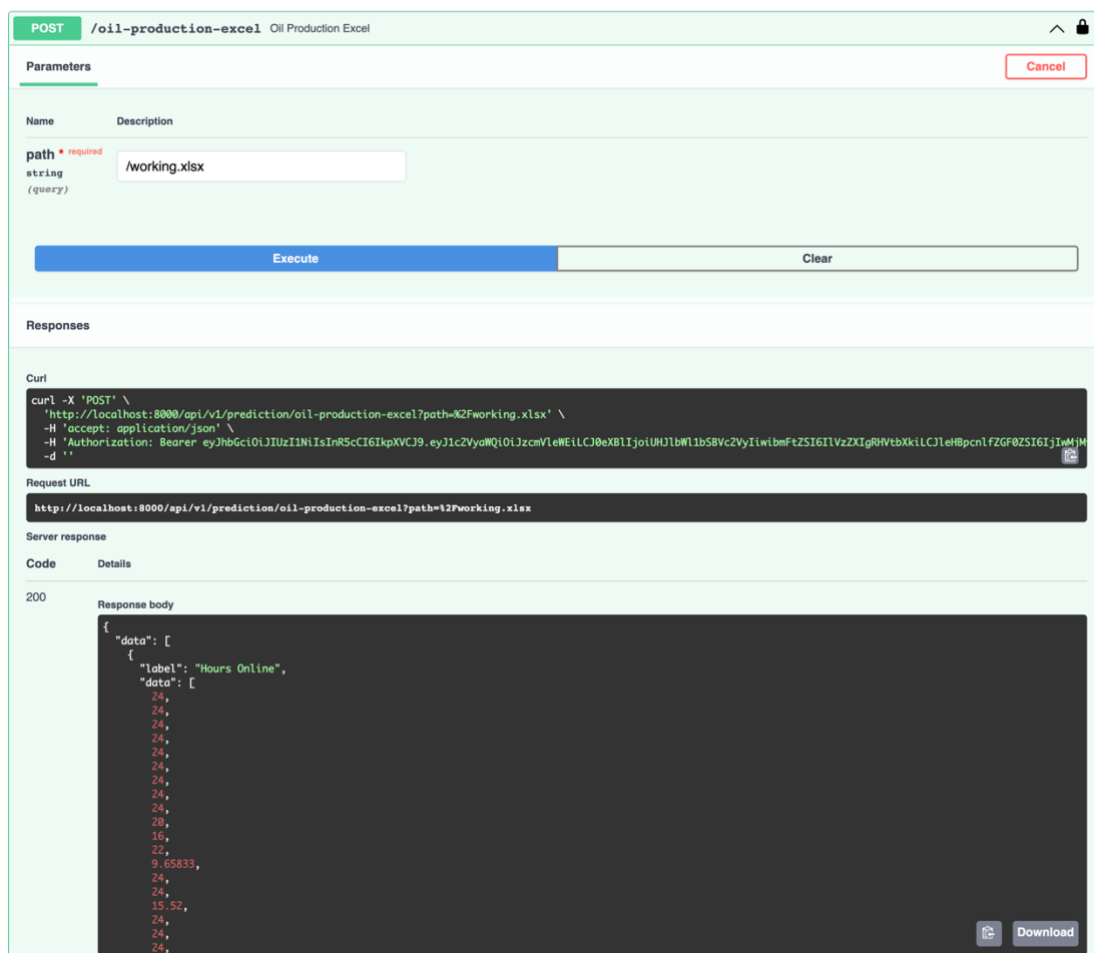
Figure 7.2: Oil Production Excel Endpoint

Table 7.1 contains the summary of the endpoints created by the author.

Table 7.1: Endpoints

Endpoints	Method	Description	Request	Response (200)
/oil-production	POST	Predicting oil production value based	{ "hours_online": float,	{

		on singular imputation from user	"downhole_temp": float,
			
			  "downhole_press": float,
			
			  "press_diff": float,
			
			  "temp_diff": float,
			
			  "deg": string
			
			}
				"prediction": float
				
				}

/oil-production-excel	POST	Predicting oil production value(s) based on a excel file	{
			
			"path": string
			
			}
				{
				
				"data": array[
				
				  "label": string,
				
				  "data": array[]
				
				  ]
				
				}

/gas-production	POST	Predicting gas production value based on singular imputation from user	{
			
			  "hours_online": float,
			
			  "downhole_temp": float,
			
			  "downhole_press": float,
			
			  "press_diff": float,
			
			  "temp_diff": float,
			
			  "deg": string
			
			}
				{
				
				"prediction": float
				
				}

/gas-production-excel	POST	Predicting gas production value(s) based on a excel file	{
			
			"path": string
			
			}
				{
				
				"data": array[
				
				"label": string,
				
				"data": array[]
				
				]
				
				}

CHAPTER 8

CONCLUSION AND RECOMMENDATION

This chapter will conclude the results obtained in this thesis and provide recommendation that can be implemented for future works.

8.1   Conclusion

The results show that the tree-based models can perform well on tabular structured data as the gradient boosting and random forest models gave good results. In addition to this, the results showed that the model with the self-supervised imputation dataset gave better results than the other conventional imputation methods. For oil production, the best performing model was the gradient boosting model with the self-supervised imputation dataset. On the other hand, the worst performing model for oil production is the random forest model with the median imputation dataset. The results shows the best oil production model had a RMSE value of $180\,\mathrm{m}^3$ and R-Squared value of 97.3%. On the other hand, the worst oil production model had a RMSE value of 231 m^3 and R-Squared value of 95.1%. For gas production, the best performing model is also the gradient boosting model with the self-supervised imputation dataset while the worst model is gradient boosting model with the forward filling imputation dataset. The best model gave a RMSE value of 53,211 m^3 and R-squared value of 91.9%. Whereas, the worse model gave a RMSE value of 62532 m^3 and R-Squared value of 88.0%. The best oil and gas models were then deployed as an API and was used in the VDR website application created by the author's teammate.

8.2   Recommendation

The model was able to predict oil and gas production with high accuracy, however there is still room for improvement. Although oil and gas production is mainly

influenced by pressure and temperature, there are other factors that could play a role in its production. A factor that could affect production is the state of the equipment used to obtain the oil and gas. Perhaps a dataset that also contains the state of equipment, e.g. new or old, could improve the performance of the models.

Another method that could improve the performance of the model is by improving the self-supervised imputation method. For this thesis, the author only made use of a baseline gradient boosting and random forest model to fill in the missing values. Perhaps by optimizing the models with Bayesian optimization would improve the dataset, hence improving the prediction results.

REFERENCES

[1]   M. S. Vassiliou, Historical dictionary of the petroleum industry, 2018.

[2]   International Association of Oil & Gas Producers, "Oil and gas in Everyday Life," International Association of Oil & Gas Producers, [Online]. Available: https://www.iogp.org/oil-natgas-in-everyday-life/. .

[3]   W. P. Council, "Why are oil and gas important?," [Online]. Available: https://www.world-petroleum.org/edu/221-why-are-oil-and-gas-important#:~:text=Oil%20is%20one%20of%20the%20most%20important%20raw,about%20two%20million%20tonnes%20of%20oil%20and%20gas. . [Accessed March 2022].

[4]   R. Ranggasari, "Oil and gas reserves potential in eastern Indonesia reaches 9.8bn barrels," Tempo, [Online]. Available: https://en.tempo.co/read/1536679/oil-and-gas-reserves-potential-in-eastern-indonesia-reaches-9-8bn-barrels#:~:text=Overall%2C%20the%20Energy%20Ministry%20recorded%20there%20are%2070,2.44%20billion%20barrels%20and%20gas%20of%2043.6%20TCF. .

[5]   Indonesia Investment, "Crude Oil Indonesia," [Online]. Available: https://www.indonesia-investments.com/business/commodities/crude-oil/item267..

[6]   W. Kenton, "Virtual Data Room (VDR)," 23 June 2021. [Online]. Available: https://www.investopedia.com/terms/v/virtual-data-room-vdr.asp#:~:text=Virtual%20Data%20Rooms%2C%20or%20VDRs%2C%20exist%20as%20a,joint%20venture%20that%20requires%20access%20to%20shared%20data.. [Accessed 19 03 2022].

[7]   Lynx, "License Pricing - Lynx Information Systems," Lynx Information System, [Online]. Available: http://www.lynxinfo.co.uk/download-pricing.html.

[8]   Intviewer, "Intviewer - Fast Geoscience Visualization, Analysis & QC,," Intviewer, 02 August 2021. [Online]. Available: https://www.int.com/products/intviewer/#:~:text=INTViewer%20is%20a%20platform%20and%20application%20that%20allows,to%20a%20desktop%20or%20remotely%20via%20the%20cloud..

[9]   INTViewer, "INTViewer. Geoscience Analysis and QC, Simplied.," INTViewer, [Online]. Available: https://www.int.com/products/intviewer/.

[10]   Geodwipa Teknika Nusantara, "Geodwipa Teknika Nusantara," Geodwipa Teknika Nusantara, [Online]. Available: https://ptgtn.com/.

[11]   IBM, "Data Science," IBM, 15 May 2020. [Online]. Available: https://www.ibm.com/cloud/learn/data-science-introduction.

[12]   A. El-Banbi, A. Ahmed and E.-M. Ahmed , "Black Oils," in *PVT Property Correlations*, Elsevier, 2018, p. 147–182.

[13]   S. Mokhatab, W. A. Poe and J. Y. Mak, "Natural Gas Fundamentals," in *Handbook of Natural Gas Transmission and Processing*, Elsevier, 2019.

[14]   A. El-Banbi, A. Ahmed and E.-M. Ahmed, "Dry Gases," in *PVT Property Correlations*, Elsevier, 2018.

[15]   T. Ahmed, "Reservoir-Fluid Properties," in *Reservoir-Fluid Properties* , Elsevier, 2010, pp. 29-135.

[16]   I. Fetoui, "Hydrocarbon Phase Behavior," [Online]. Available: https://production-technology.org/category/pvt/.

[17]   P. Z. a. K. H. C. Janiesch, "Machine learning and deep learning," *Electronic Markets,* vol. 31, no. 3, pp. 685-695, 2021.

[18]   IBM Cloud Education, "What is machine learning," IBM, 15 July 2020. [Online]. Available: https://www.ibm.com/cloud/learn/machine-learning..

[19]   I. Sydorenko, "What is dataset in Machine Learning," High quality data annotation for Machine Learning, 5 April 2021. [Online]. Available: https://labelyourdata.com/articles/what-is-dataset-in-machine-learning.

[20]   D. N. Dimid, "Unsupervised learning algorithms cheat sheet,," 17 February 2022. [Online]. Available: https://towardsdatascience.com/unsupervised-learning-algorithms-cheat-sheet-d391a39de44a. .

[21]   R. S. Sutton and A. G. Barto, "Reinforcement learning: An introduction," *Cambridge, MA,* 2020.

[22]   W. X. P. C. M. C. a. S. D. Y. Gong, "Supervised Learning," *Machine learning techniques for multimedia,* p. 21–49., 2008.

[23]   IBM Cloud Education, "What is Supervised Learning," IBM, 2022. [Online]. Available: https://www.ibm.com/cloud/learn/supervised-learning. .

[24]   A. M. J. A. V. M. A. A. L. a. A. A. S. A. R. van Loon, "Understanding supervised, unsupervised, and reinforcement learning," *Big Data Made Simple,* 2019.

[25]   A. Jaiswal , A. R. Babu, M. Z. Zadeh, D. Banerjee and F. Makedon, "A Survey on Contrastive Self-Supervised Learning," in *PETRA*, 2020.

[26]   Javatpoint, "Regression vs. Classification in Machine Learning," Javatpoint, [Online]. Available: https://www.javatpoint.com/regression-vs-classification-in-machine-learning.

[27]   "Classification vs. Regression Algorithms in Machine Learning," 16 February 2022. [Online]. Available: https://www.projectpro.io/article/classification-vs-regression-in-machine-learning/545#mcetoc_1fp6av4s69.

[28]   z_ai, "Deep Learning for NLP: ANNs, RNNs and LSTMs explained!," 8 July 2019. [Online]. Available: https://towardsdatascience.com/deep-learning-for-nlp-anns-rnns-and-lstms-explained-95866c1db2e4.

[29]   aravindpai, "CNN vs. RNN vs. ANN – Analyzing 3 Types of Neural Networks in Deep Learning download Share," 17 February 2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/02/cnn-vs-rnn-vs-mlp-analyzing-3-types-of-neural-networks-in-deep-learning/.

[30]   GeeksforGeeks, "Understanding of LSTM Networks," GeeksforGeeks, 25 June 2021. [Online]. Available: https://www.geeksforgeeks.org/understanding-of-lstm-networks/.

[31] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep Learning is not all you need," *Information Fusion,* vol. 81, pp. 84-90, 2022.

[32] V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawlczyk and G. Kasneci, "Deep Neural Networks and Tabular Data: A Survey," 2021.

[33] Elzain, Hussam Eldin, Chung, Sang Yong, Senapathi, Venkatramanan, Sekar, Selvam, Lee, Seung Yeop, Roy, Priyadarsi D., Hassan, Amjed and Sabarathinam, Chidambaram, "Comparative study of machine learning models for evaluating groundwater vulnerability to nitrate contamination," *Ecotoxicology and Environmental Safety,* vol. 229, pp. 61-113, 2022.

[34] Scikit, "ScikitLearn," [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html.

[35] A. Kumar, "Introduction to the Gradient Boosting Algorithm," 20 May 2020. [Online]. Available: https://medium.com/analytics-vidhya/introduction-to-the-gradient-boosting-algorithm-c25c653f826b.

[36] S. Rosenthal, "Data Imputation," in *The International Encyclopedia of Communication Research Methods*, Wiley, 2017, pp. 1-12.

[37] F. Malik, "Understanding value of correlations in data science projects," Medium, 10 June 2019. [Online]. Available: https://medium.com/fintechexplained/did-you-know-the-importance-of-finding-correlations-in-data-science-1fa3943debc2#:~:text=Correlation%20is%20a%20statistical%20measure.%20Correlation%20explains%20how,%28variables%29%20can%20be%20positively%20correlated%.

[38] A. McDonald, "Using the missingno Python library to Identify and Visualise Missing Data Prior to Machine Learning," 10 June 2021. [Online]. Available: https://towardsdatascience.com/using-the-missingno-python-library-to-identify-and-visualise-missing-data-prior-to-machine-learning-34c8c5b5f009.

[39] A. Swalin, "How to handle missing data," Medium, 19 March 2018. [Online]. Available: https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4..

[40] H. Kang, "The prevention and handling of the missing data," in *Korean Journal of Anesthesiology*, vol. 64, 2013, p. 402.

[41] W. Badr, ""6 different ways to compensate for missing data (data imputation with examples)," 12 January 2019. [Online]. Available: https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779..

[42] K. N, "Part-1 : Data Preparation Made Easy with python!!," Medium, 09 May 2020. [Online]. Available: https://medium.com/analytics-vidhya/part-1-data-preparation-made-easy-with-python-e2c024402327..

[43] Á. Fernández, J. R. Dorronsoro and J. Bella, "Supervised outlier detection for classification and regression," *Neurocomputing,* vol. 486, pp. 77-92, 2022.

[44] C. M. Salgado, C. Azevedo, H. Proença and S. M. Vieira, "Noise Versus Outliers," in *Secondary Analysis of Electronic Health Records*, Cham, Springer International Publishing, 2016, pp. 163-183.

[45] BBC Bitesize, "Types of correlation - scattergraphs - national 4 application of Maths Revision," BBC News, [Online]. Available: https://www.bbc.co.uk/bitesize/guides/zmt9q6f/revision/2..

[46] Nettleton, David, "Selection of Variables and Factor Derivation," in *Commercial Data Mining*, Elsevier, 2014, pp. 79-104.

[47] "Spearman correlation coefficient: Definition, formula and calculation with example," QuestionPro, 15 January 2020. [Online]. Available: https://www.questionpro.com/blog/spearmans-rank-coefficient-of-correlation/. .

[48] Swapnilbobe, "Spearman's correlation," 13 April 2021. [Online]. Available: https://medium.com/analytics-vidhya/spearmans-correlation-f34c094d99d8#:~:text=Here%2C%20we%20are%20calculating%20spearman%E2%80%99s%20correlation%20using%20the,of%20relationship%20between%20ranks%20of%20two%20individual%20features. .

[49] V. Kumar and S. Minz, "Feature Selection: A literature Review," *Smart Computing Review,* vol. 4, pp. 1-19, 2014.

[50] K. Menon, "Feature selection in machine learning," Simplilearn, 16 September 2021. [Online]. Available: https://www.simplilearn.com/tutorials/machine-learning-tutorial/feature-selection-in-machine-learning#what_is_feature_selection..

[51] Javatpoint, "Hyperparameters in Machine Learning," [Online]. Available: https://www.javatpoint.com/hyperparameters-in-machine-learning.

[52] M. Feurer and F. Hutter, "Hyperparameter Optimization," in *Automated Machine Learning*, Springer, 2019, pp. 3-35.

[53] M. Dei, "Hyperparameter Tuning Explained — Tuning Phases, Tuning Methods, Bayesian Optimization, and Sample Code!," 13 December 2019. [Online]. Available: https://towardsdatascience.com/hyperparameter-tuning-explained-d0ebb2ba1d35.

[54] W. Koehrsen, "Automated Machine Learning Hyperparameter Tuning in Python," 3 July 2018. [Online]. Available: https://towardsdatascience.com/automated-machine-learning-hyperparameter-tuning-in-python-dfda59b72f8a#8811.

[55] A. Chugh, "MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better?," 8 December 2020. [Online]. Available: https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e.

[56] Zhao, Yunwei and Wan, Xin, "The Design of Embedded Web System based on REST Architecture," in *IEEE*, 2019.

[57] D. Bryant, "GraphQL-ultimate-guide," [Online]. Available: https://www.infoq.com/articles/GraphQL-ultimate-guide/.

[58] BBVA API Market, "REST API: What is it, and what are its advantages in project development?," 2016. [Online]. Available: https://www.bbvaapimarket.com/en/api-world/rest-api-what-it-and-what-are-its-advantages-project-development/.

[59] R. Naushad, "Comparison of FastAPI and Flask. Simple Explanation!," 24 July 2020. [Online]. Available:

https://medium.datadriveninvestor.com/comparison-of-fastapi-and-flask-simply-explanation-c8c075f6aa80.

[60]   Docker, "Docker," [Online]. Available: https://docs.docker.com/.

[61]   GeeksforGeeks, "Why Should You Use Docker – 7 Major Reasons!," 26 April 2021. [Online]. Available: https://www.geeksforgeeks.org/why-should-you-use-docker-7-major-reasons/.

[62]   Harrison, Bob, "The data room," 2020, pp. 21-26.

[63]   Schlumberger, "Secure, Remote Access to Field Datasets Enables Potential Investors to Complete Asset Evaluations," Schlumberger, [Online]. Available: https://www.slb.com/resource-library/case-study/dss/delfi-virtual-data-room-generic-asia-pacific-cs.

[64]   M. Taylor, "Machine Learning in the Oil and Gas Industry," 21 January 2021. [Online]. Available: https://newengineer.com/blog/machine-learning-in-the-oil-and-gas-industry-1507752.

[65]   C. Xie, L. Chao, Y. Qin, J. Cao and Y. Li, "Using a stochastic forest prediction model to predict the hazardous gas concentration in a one-way roadway," *AIP Advances,* vol. 10, no. 11, 2020.

[66]   R. Sharma, "Using Facebook Prophet for Forecasting Natural Gas Production," Medium, 13 March 2021. [Online]. Available: https://medium.com/mlearning-ai/forecast-using-prophet-canadian-natural-gas-production-dataset-b1f9c57548d8.

[67]   S. Goled, "Why Are People Bashing Facebook Prophet," 18 October 2021. [Online]. Available: https://analyticsindiamag.com/why-are-people-bashing-facebook-prophet/.

[68]   L. Menculini, A. Marini, M. Proietti, A. Garinei, A. Bozza, C. Moretti and M. Marconi, "Comparing Prophet and Deep Learning to ARIMA in Forecasting Wholesale Food Prices," *Forecasting,* vol. 3, no. 3, pp. 644-662, 2021.

[69]   J. Chahar, "Prediction of Oil Production by applying Machine Learning on Volve Field Production Data.," 17 December 2020. [Online]. Available: https://www.linkedin.com/pulse/prediction-oil-production-applying-machine-learning-volve-chahar/.

[70]   "Advantages and Disadvantages of Linear Regression," [Online]. Available: https://iq.opengenus.org/advantages-and-disadvantages-of-linear-regression/.

[71]   A. Pant, "Introduction to Linear Regression and Polynomial Regression," 13 January 2019. [Online]. Available: https://towardsdatascience.com/introduction-to-linear-regression-and-polynomial-regression-f8adc96f31cb#:~:text=Disadvantages%20of%20using%20Polynomial%20Regression%20The%20presence%20of,analysis.%20These%20are%20too%20sensitive%20to%20the%20outliers..

[72]   EDUCBA, "Difference Between Random forest vs Gradient boosting," [Online]. Available: https://www.educba.com/random-forest-vs-gradient-boosting/.

[73]   S. Dash, "Gradient Boosting – A Concise Introduction from Scratch," 21 October 2020. [Online]. Available:

https://www.machinelearningplus.com/machine-learning/gradient-boosting/#:~:text=Using%20a%20low%20learning%20rate%20can%20dramatically%20improve,iterations%20to%20converge%20to%20a%20final%20loss%20value..

[74] R. Meinert, "Optimizing Hyperparameters in Random Forest Classification," 6 June 2019. [Online]. Available: https://towardsdatascience.com/optimizing-hyperparameters-in-random-forest-classification-ec7741f9d3f6#:~:text=Hyperparameter%20tuning%20can%20be%20advantageous%20in%20creating%20a,values%20can%20be%20very%20time%20consuming%20as%20well..

APPENDICES
Appendix A

I. **Heatmap to show the correlation of missing values in the Volve dataset**
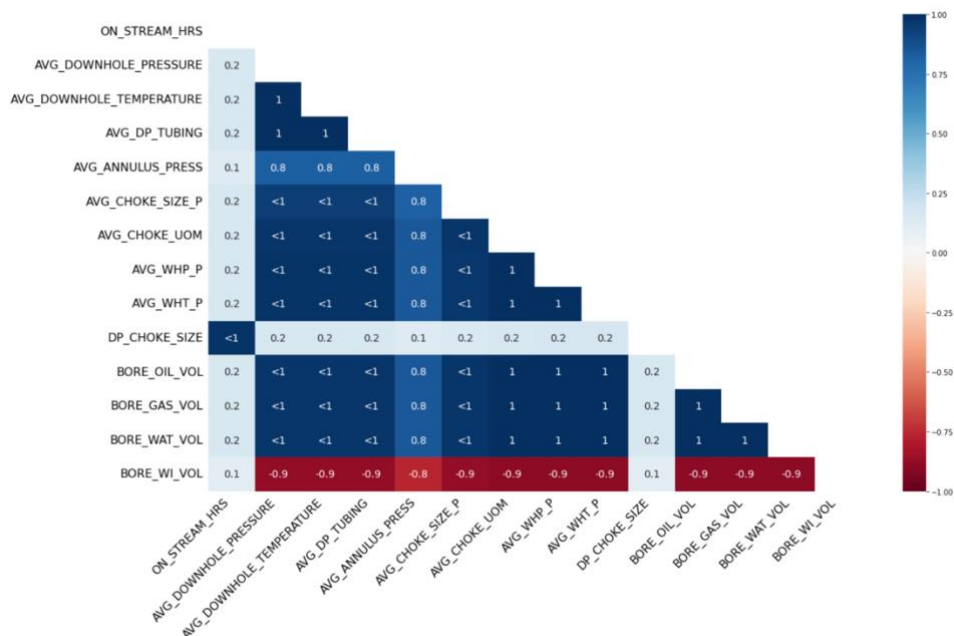
Figure A.1: Missing value correlation in Volve

II. **Heatmap to show the correlation of missing values in the Kyle Master**
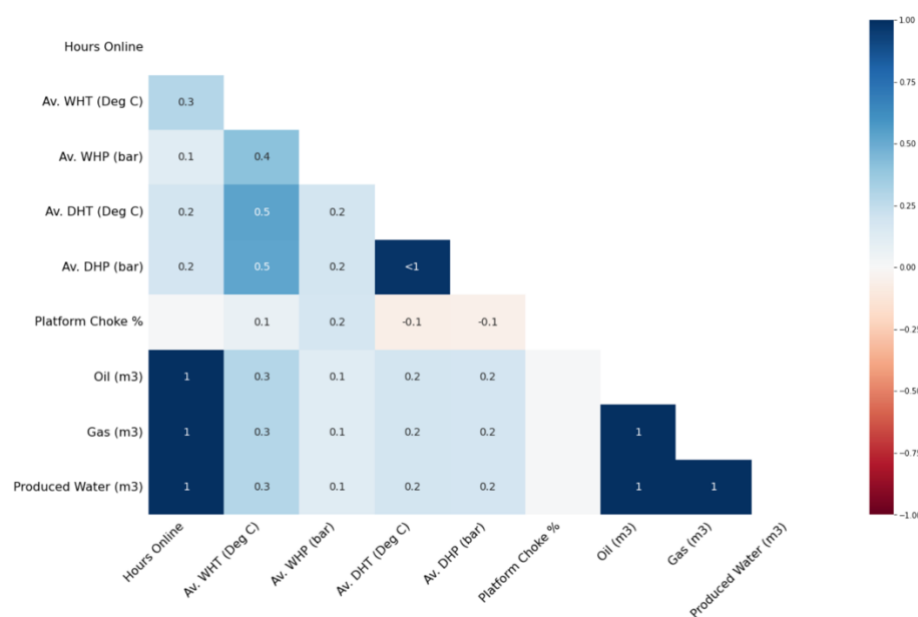
Figure A.2: Missing value correlation in Kyle Master

III.   Feature correlation in Volve and Kyle Master datasets
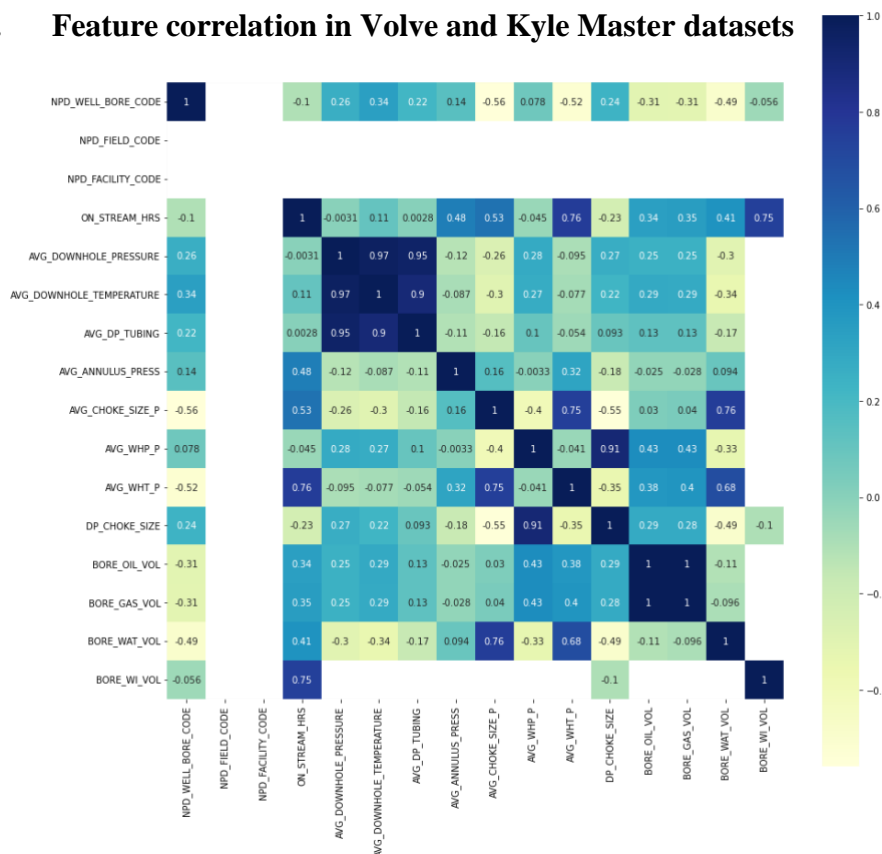
Figure A.3: Feature correlation for Volve dataset
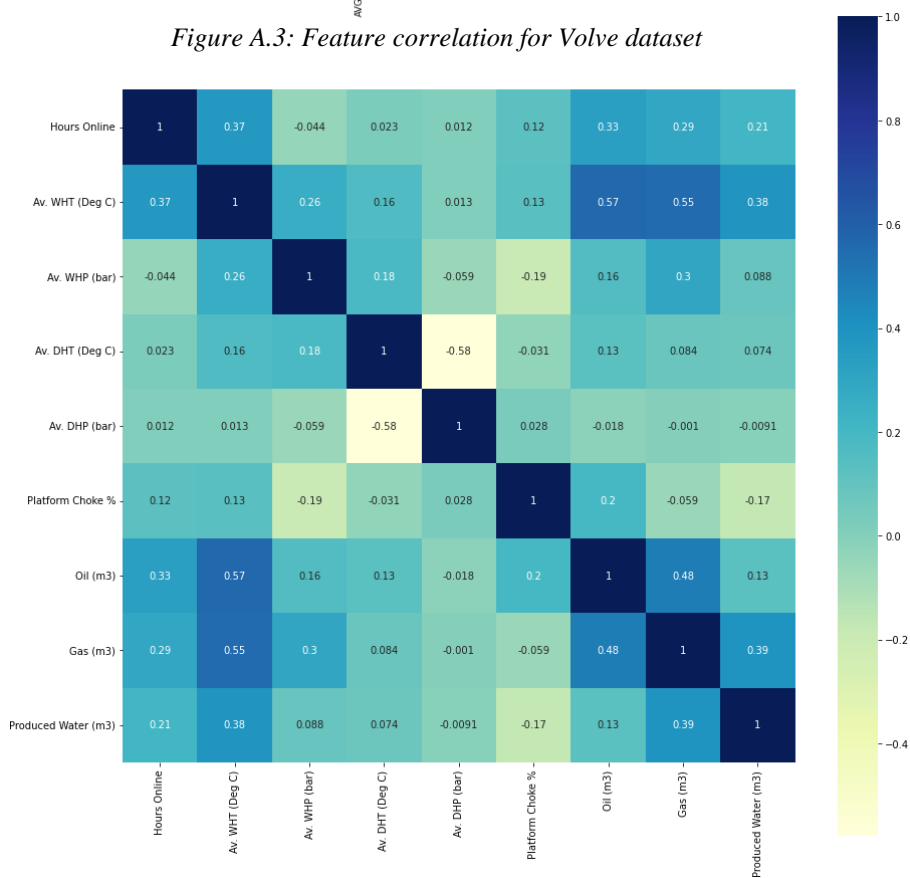
Figure A.4: Feature correlation for Kyle Master dataset

IV. Feature statistics in Volve dataset

(a)

(b)

(c)
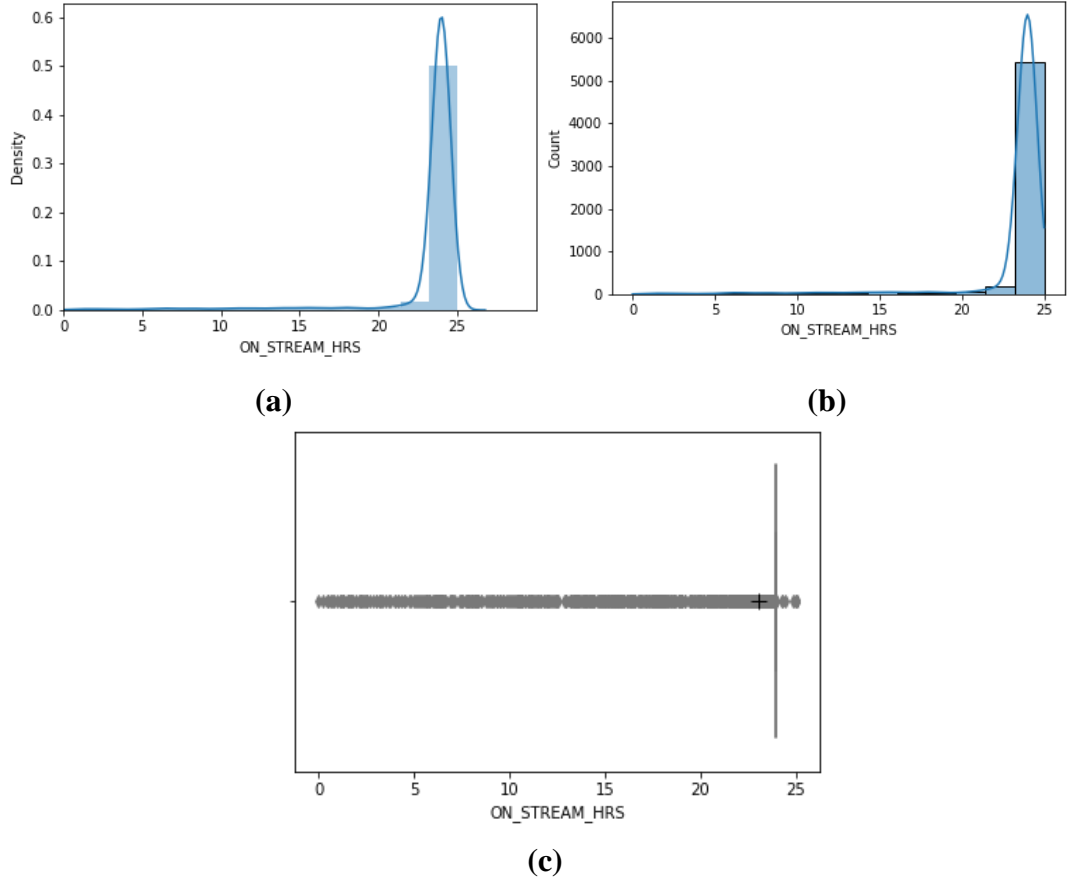
Figure A.5: (a) Kernel Density Estimation plot for ON_STREAM_HRS (b) Histogram for ON_STREAM_HRS (c) Boxplot for ON_STREAM_HRS

AVG_DOWNHOLE_PRESSURE

(a)

(b)

(c)

(d)
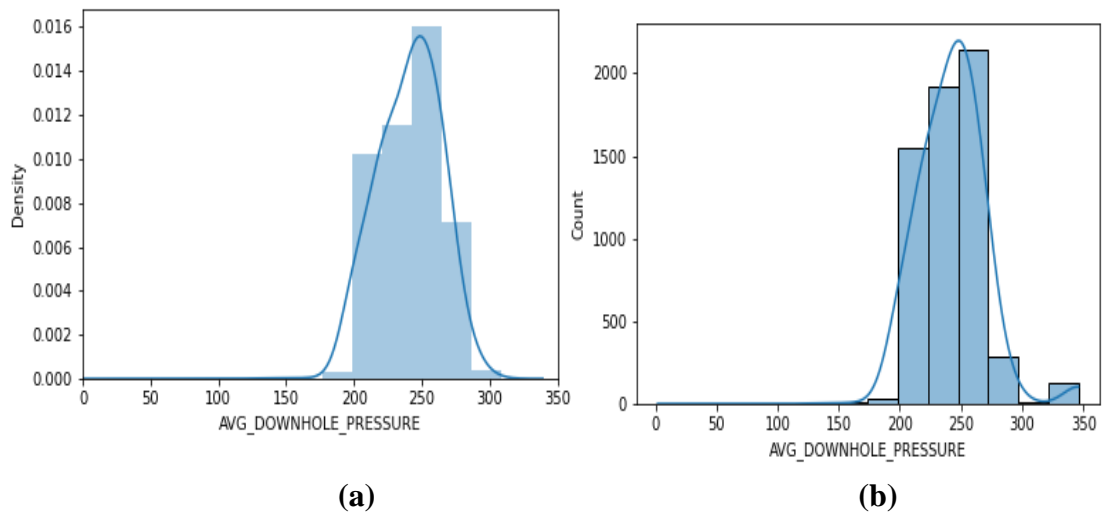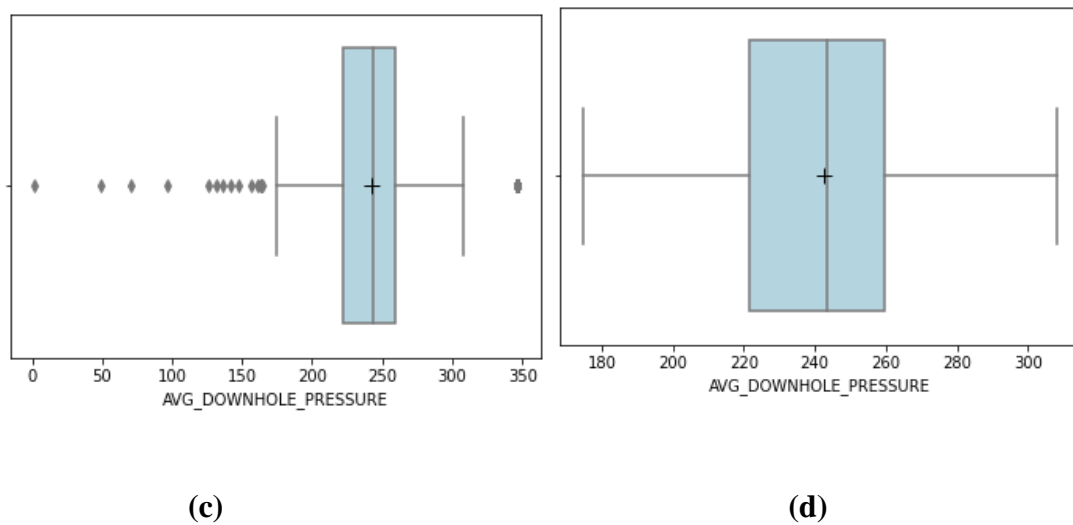
Figure A.6: (a) Kernel Density Estimation plot for AVG_DOWNHOLE_PRESSURE (b) Histogram for AVG_DOWNHOLE_PRESSURE (c) Boxplot for AVG_DOWNHOLE_PRESSURE (d) Boxplot without outliers for AVG_DOWNHOLE_PRESSURE

AVG_DOWNHOLE_TEMPERATURE

(a)
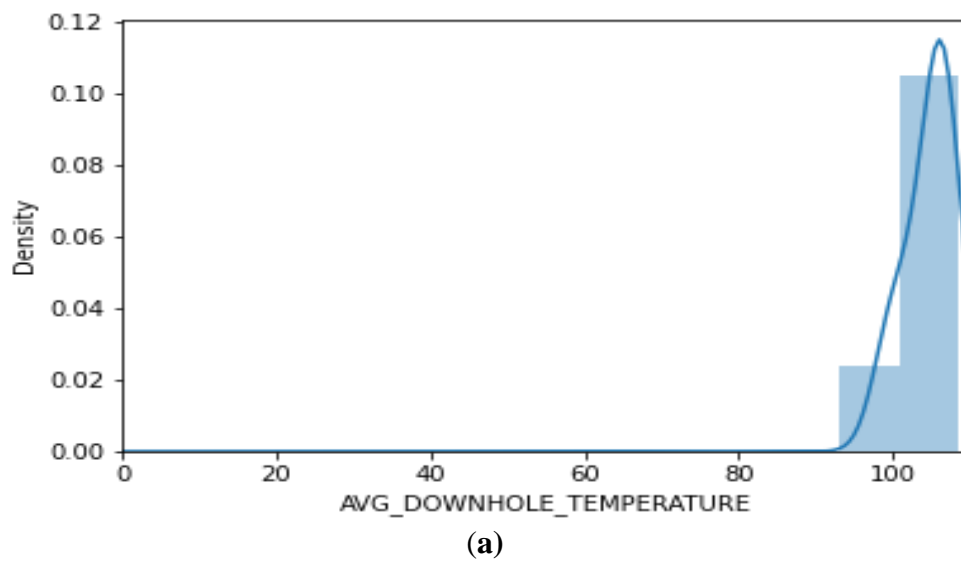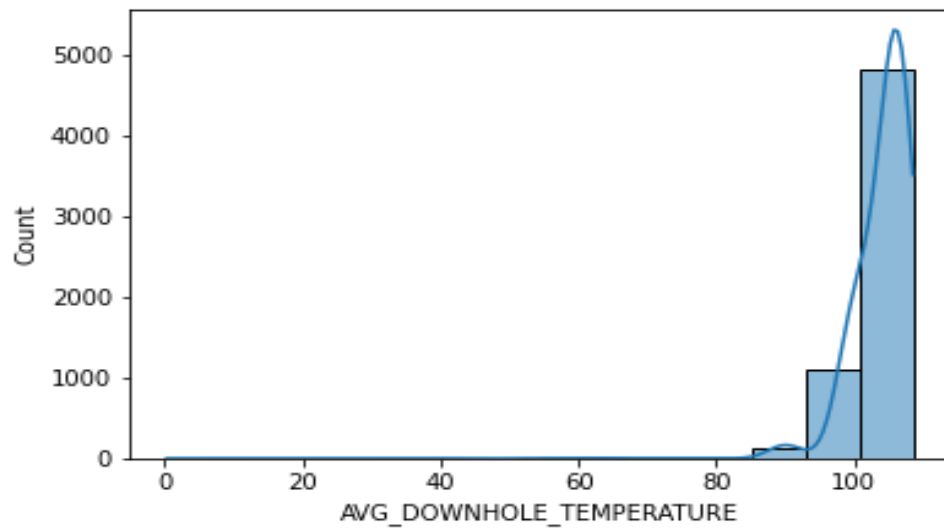
(b)
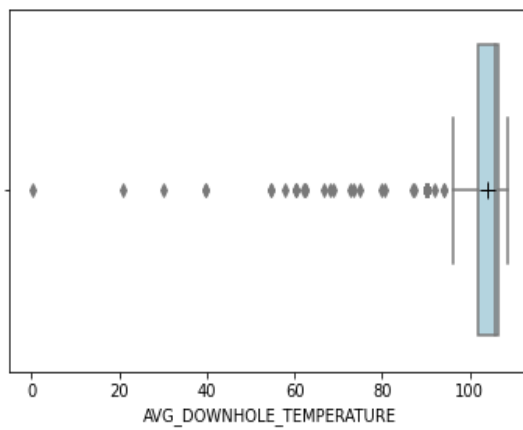
(c)
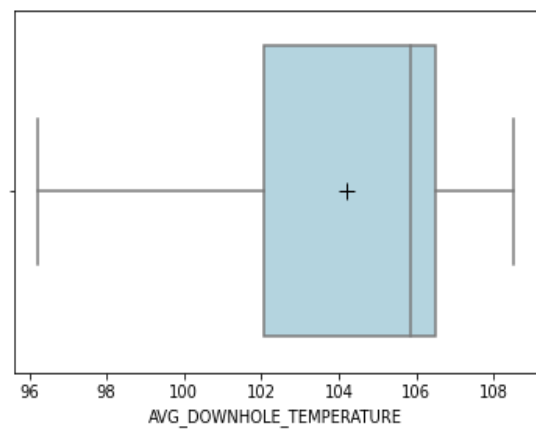
(d)

Figure A.7: (a) Kernel Density Estimation plot for AVG_DOWNHOLE_TEMPERATURE (b) Histogram for AVG_DOWNHOLE_ TEMPERATURE (c) Boxplot for AVG_DOWNHOLE_ TEMPERATURE (d) Boxplot without outliers for AVG_DOWNHOLE_ TEMPERATURE

BORE_OIL_VOL
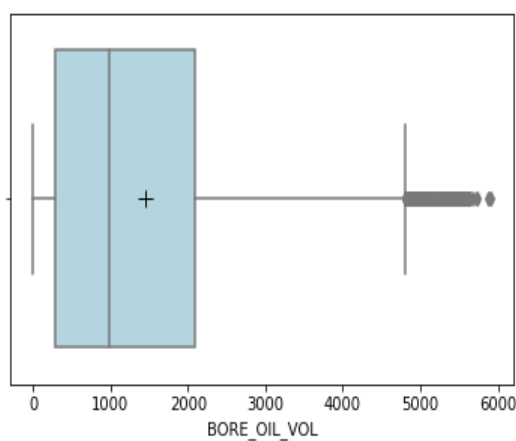
(a)
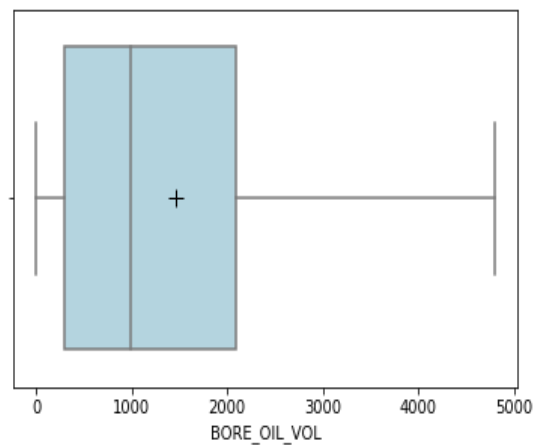
(b)
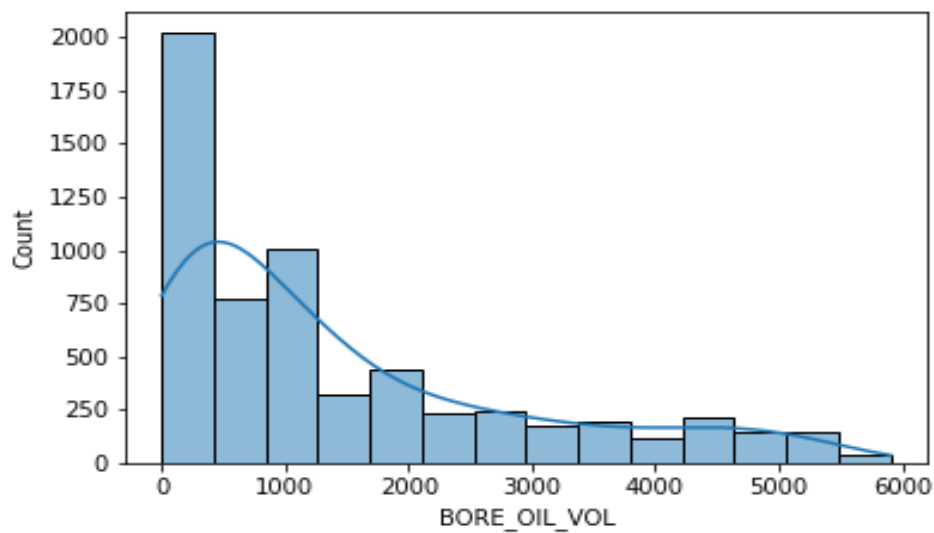
(c)
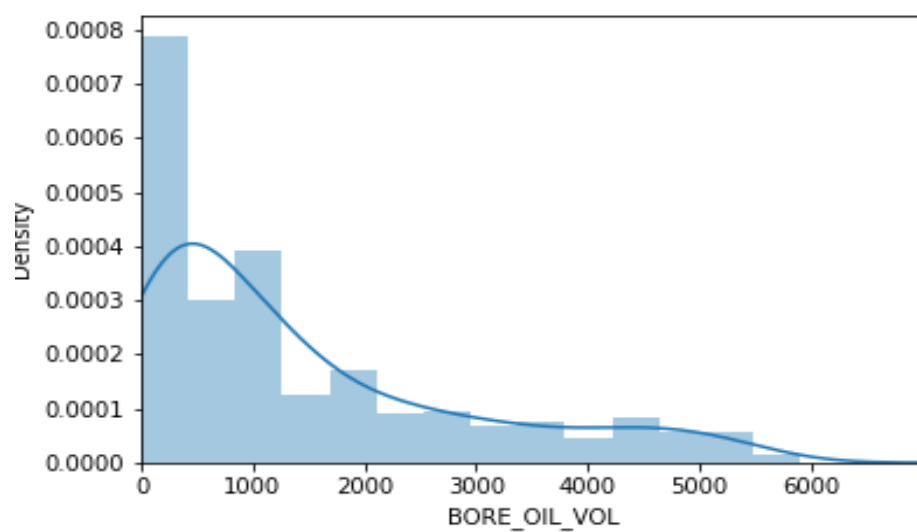
(d)

Figure A.8: (a) Boxplot for BORE_OIL_VOL (b) Boxplot without outliers for BORE_OIL_VOL (c) Histogram for BORE_OIL_VOL (d) Kernel Density Estimation plot for BORE_OIL_VOL

BORE_GAS_VOL
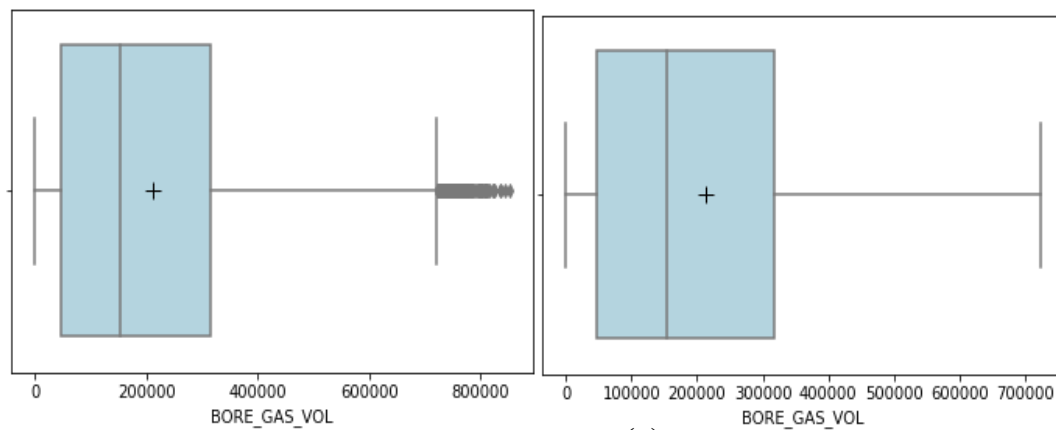
(b)   (a)
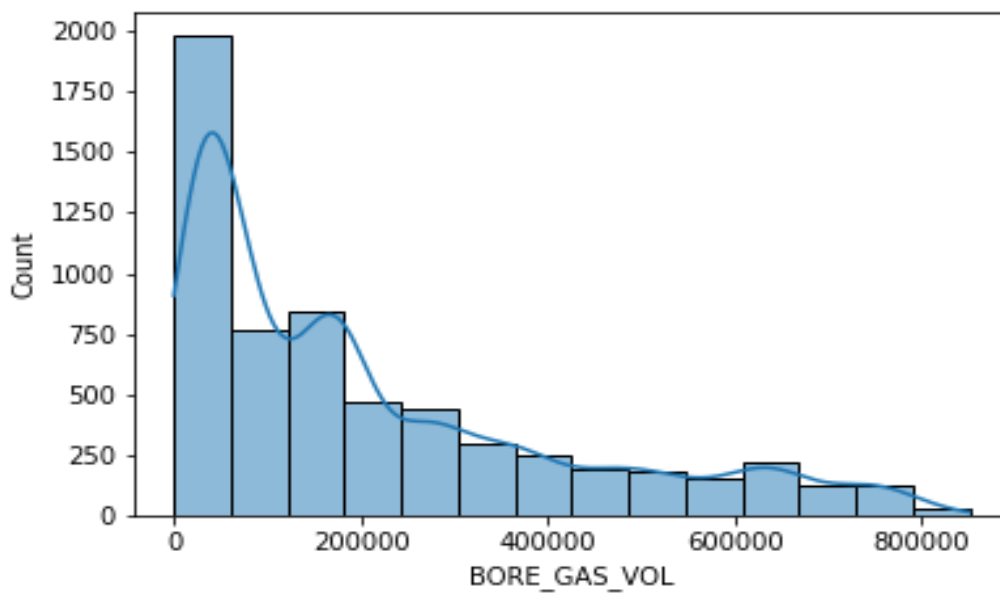
(c)
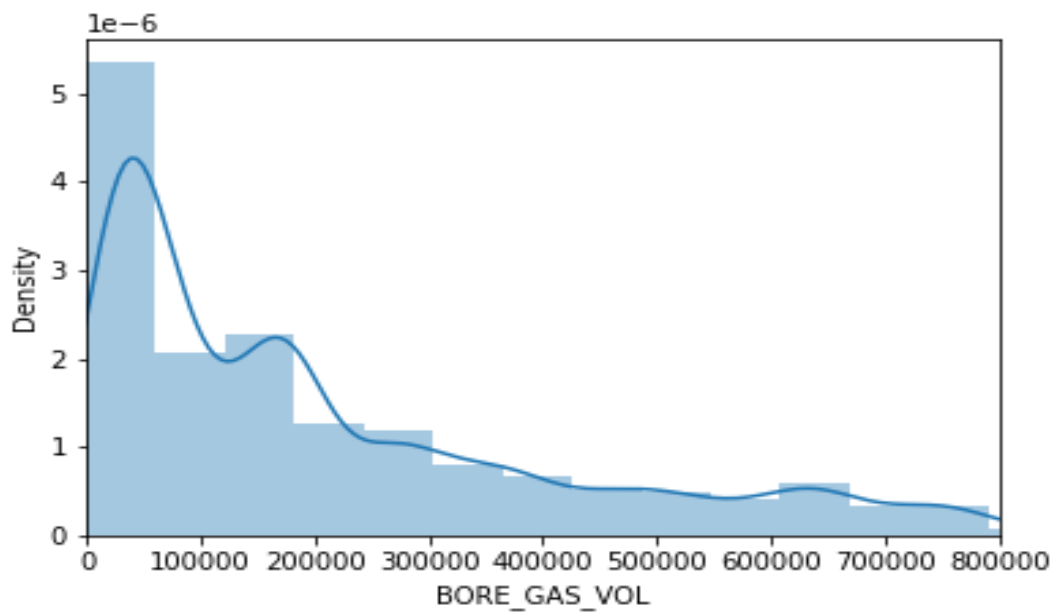
(d)

Figure A.9: (a) Boxplot for BORE_GAS_VOL (b) Boxplot without outliers for BORE_GAS_VOL (C) Histogram for BORE_GAS_VOL (c) Kernel Density Estimation plot for BORE_ GAS _VOL

AVG_WHP_P
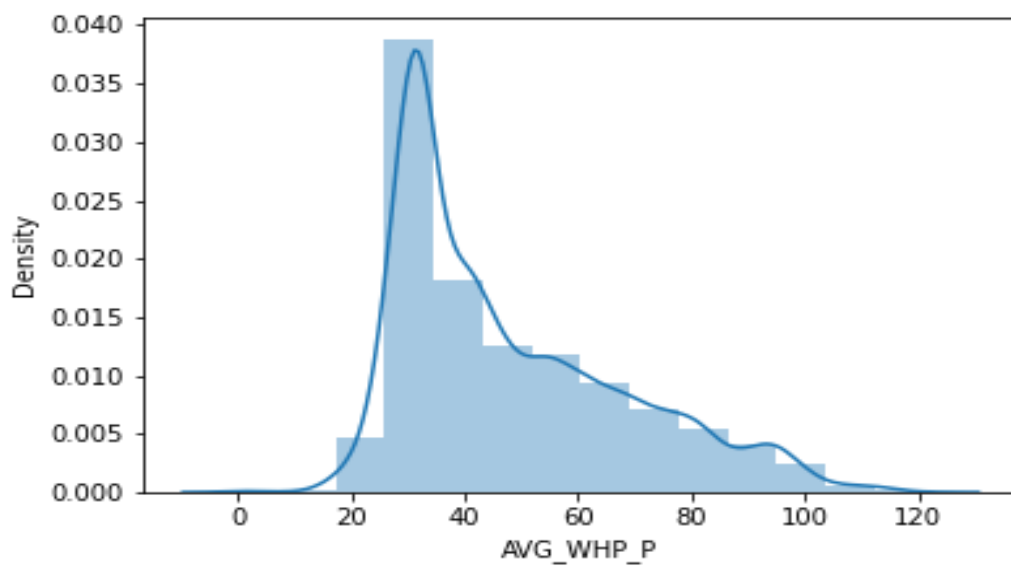
(a)
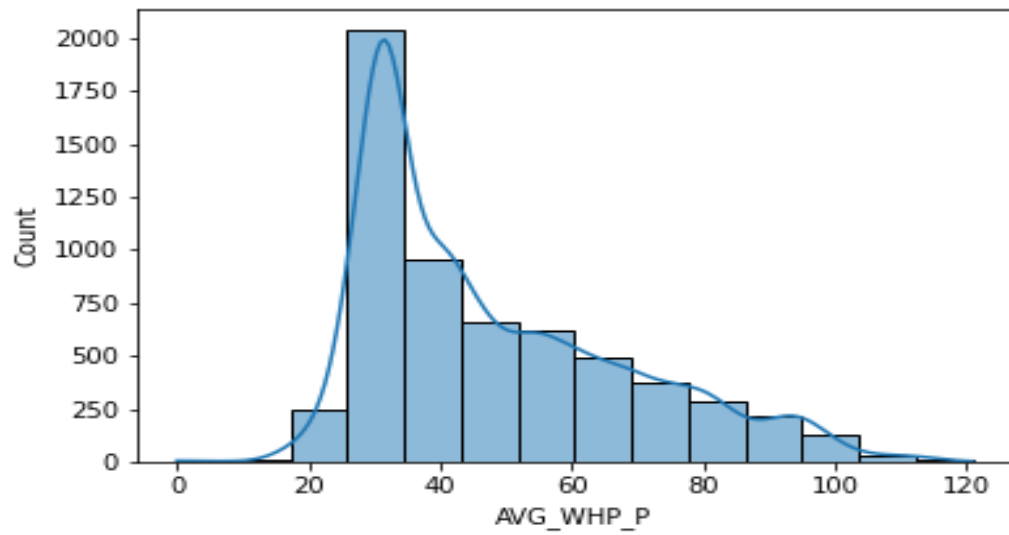
(b)
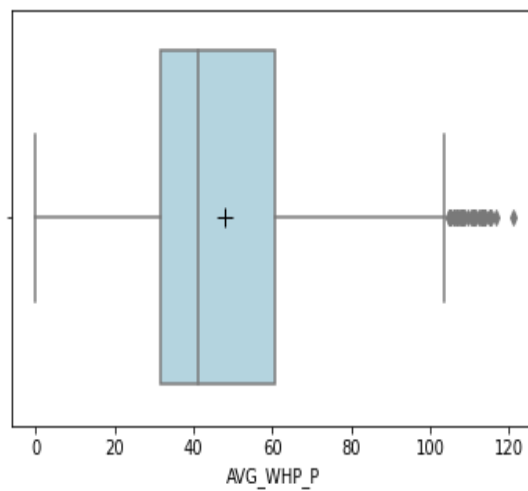
(c)
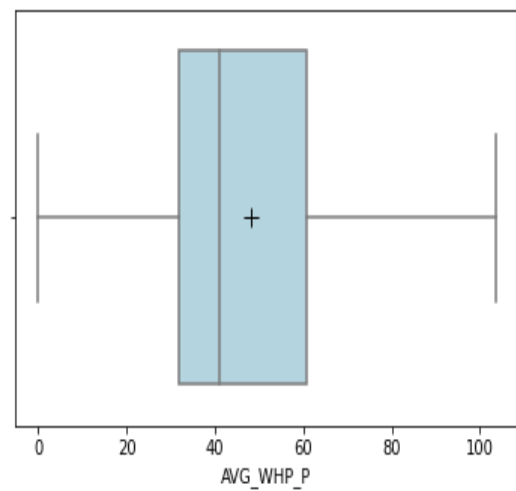
(d)

Figure A.10: (a) Kernel Density Estimation plot for AVG_WHP_P (b) Histogram for AVG_WHP_P
(c) Boxplot for AVG_WHP_P (d) Boxplot without outliers for AVG_WHP_P

AVG_WHT_P
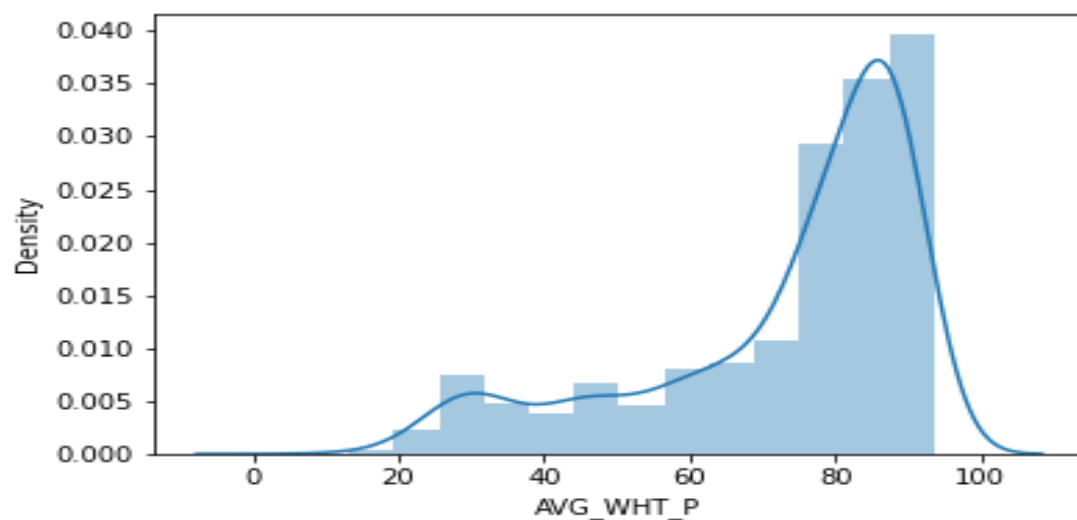
(a)
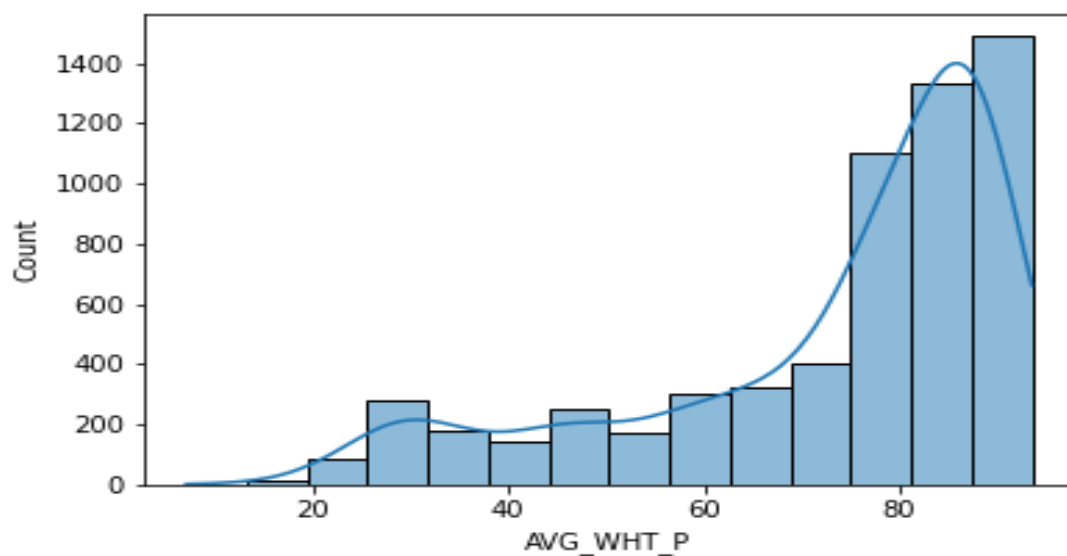
(b)
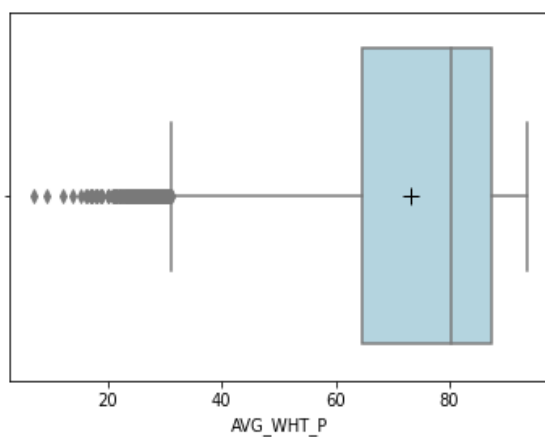
(c)
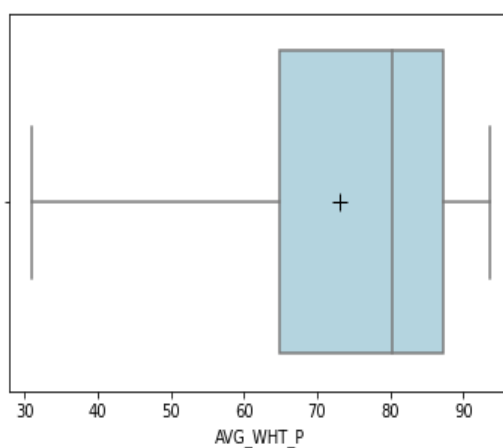
(d)

Figure A.11: (a) Kernel Density Estimation plot for AVG_WHT_P (b) Histogram for AVG_WHT_P
(c) Boxplot for AVG_WHT_P (d) Boxplot without outliers for AVG_WHT_P

V. Feature statistics in Kyle Master dataset

Hours Online

(a)

(b)

(c)
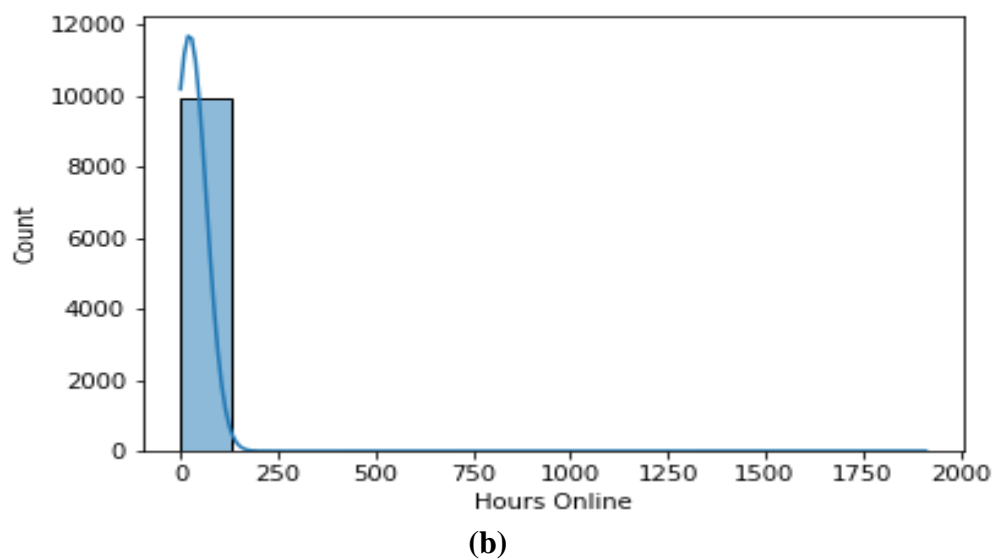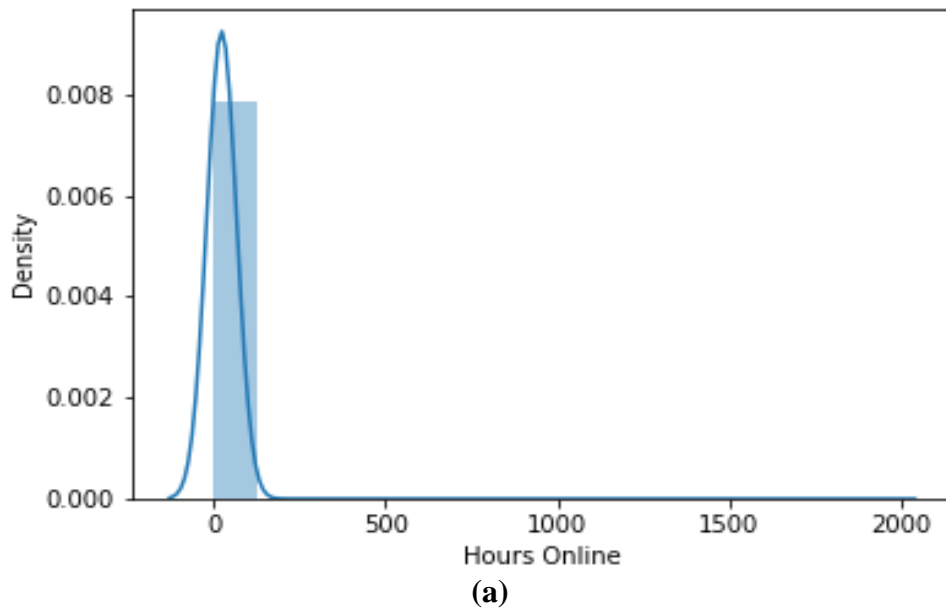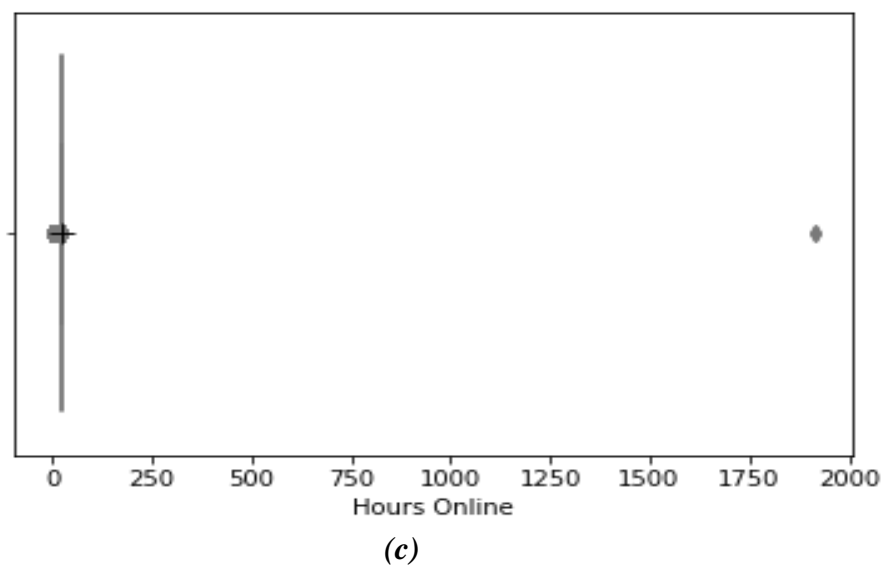
Figure A.12: (a) Kernel Density Estimation plot for Hours Online (b) Histogram for Hours Online (c) Boxplot for Hours Online

Av. DHP (bar)

(a)

(b)

(c) **(d)**
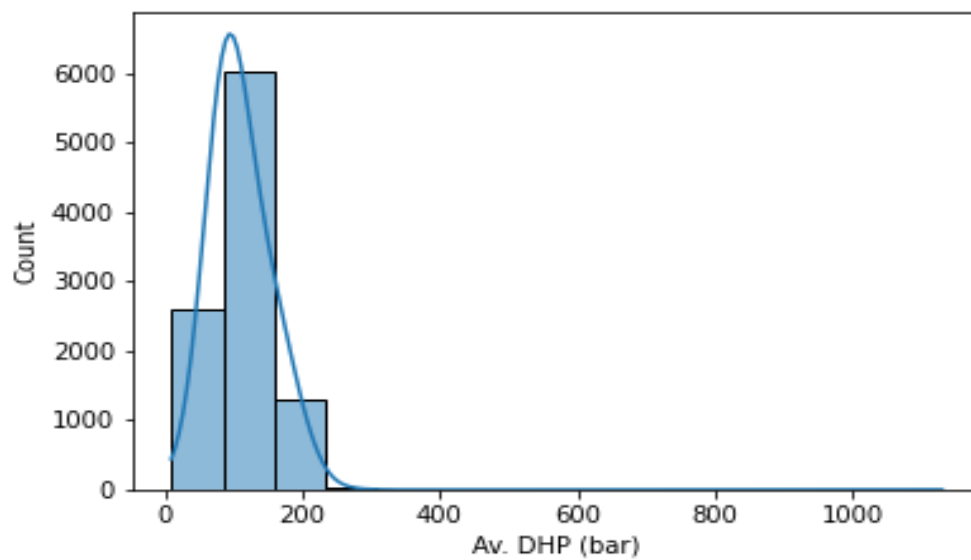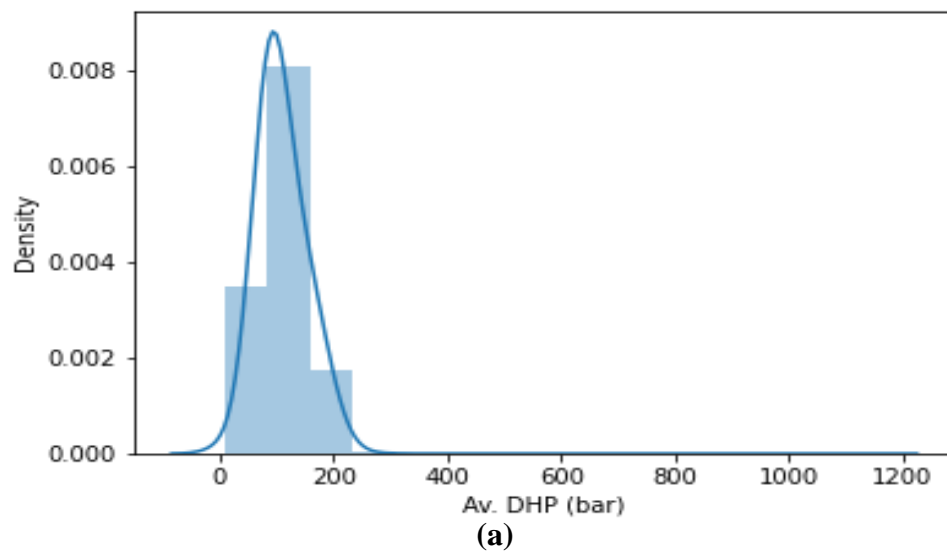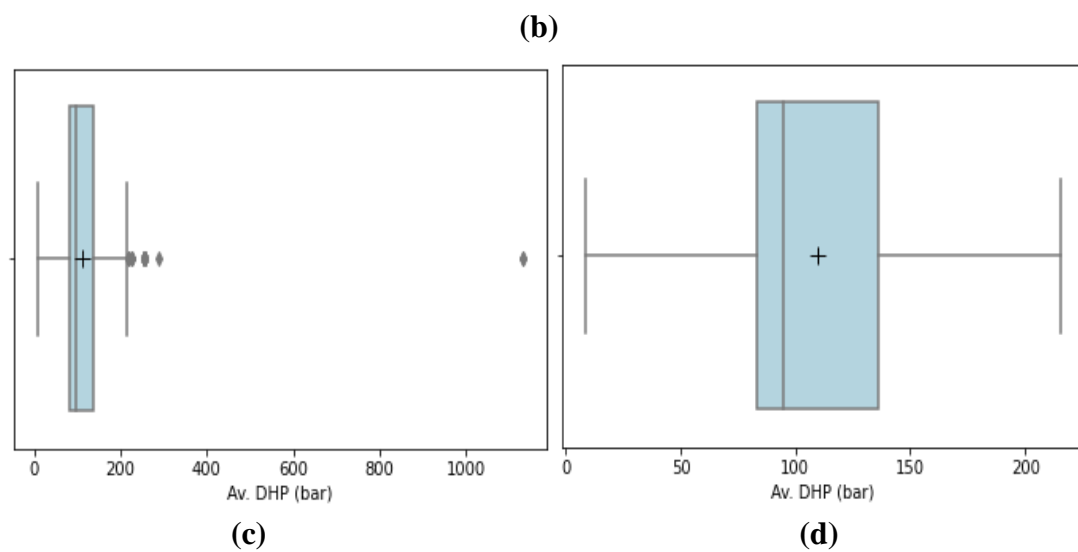
Figure A.13: (a) Kernel Density Estimation plot for Av. DHP (bar) (b) Histogram for Av. DHP (bar) (c) Boxplot for Av. DHP (bar) (d) Boxplot without outliers for Av. DHP (bar)

Av. DHT (Deg C)

(a)
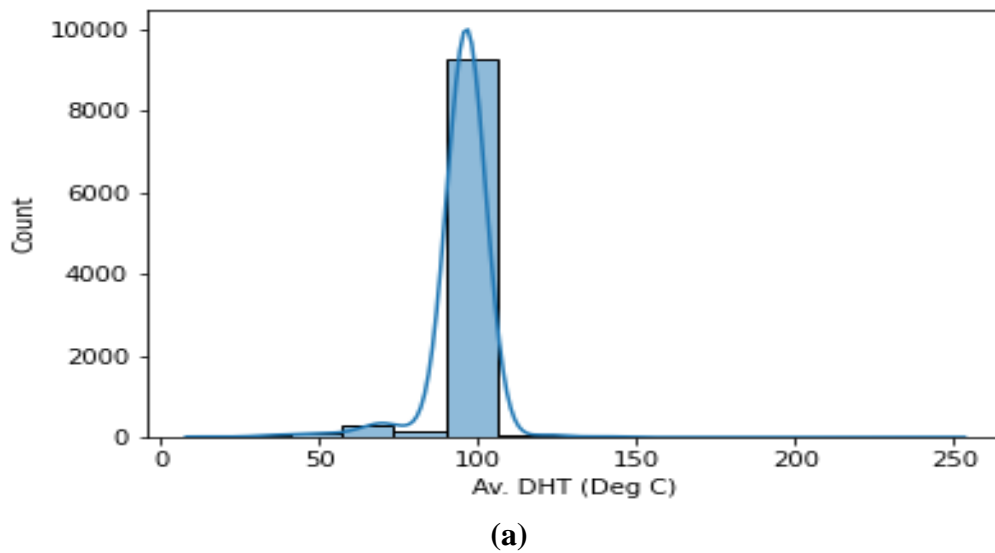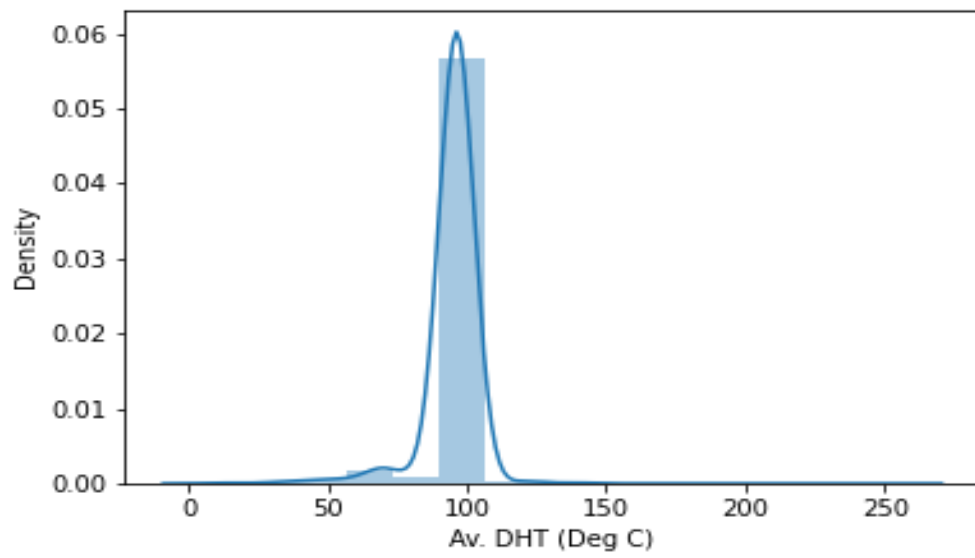
(b)
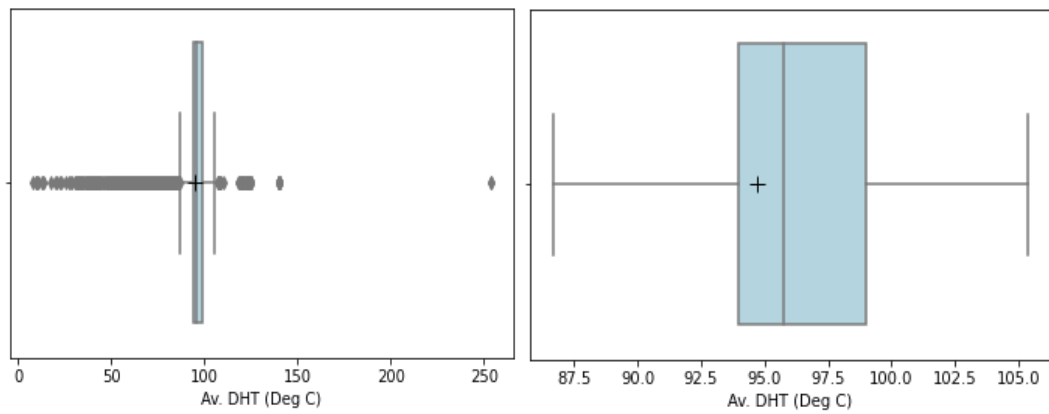
(c)

(d)

Figure A.14: (a) Histogram for Av. DHT (Deg C) (b) Kernel Density Estimation plot for Av. DHT (Deg C) (c) Boxplot for Av. DHT (Deg C) (d) Boxplot without outliers for Av. DHT (Deg C)

Oil (m3)

(a)

(b)

(c) **(d)**
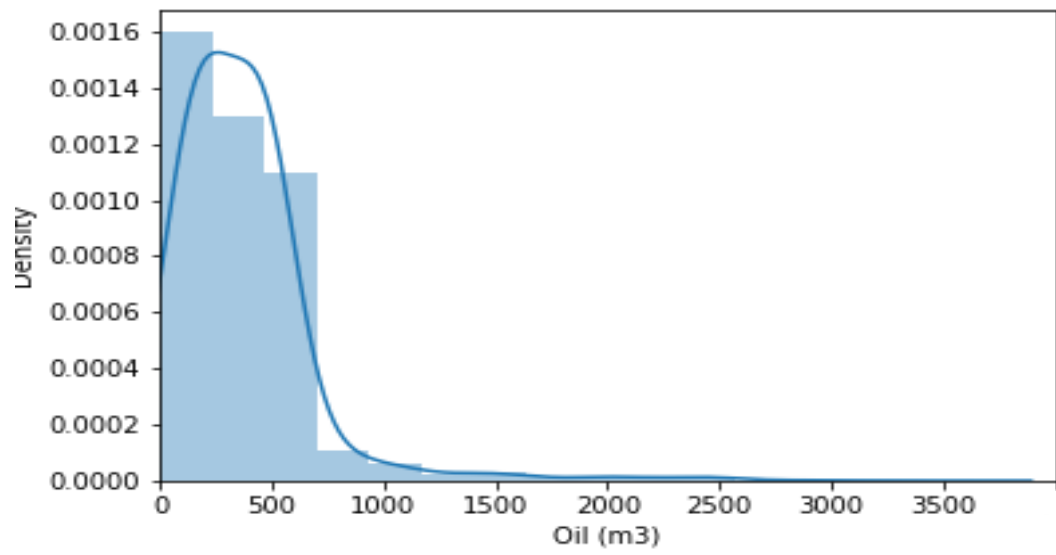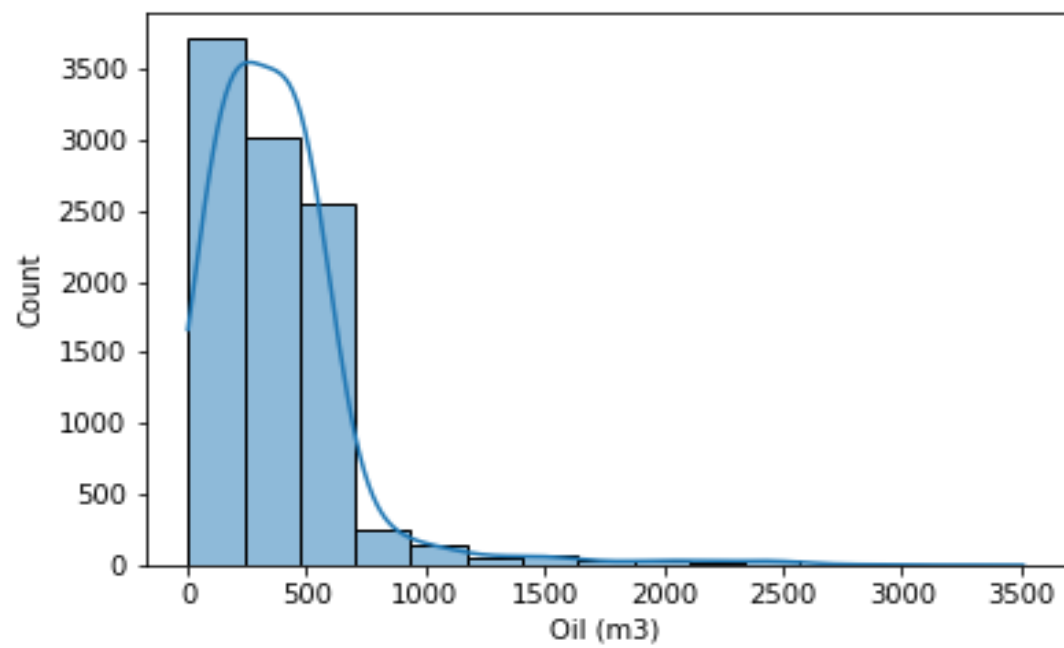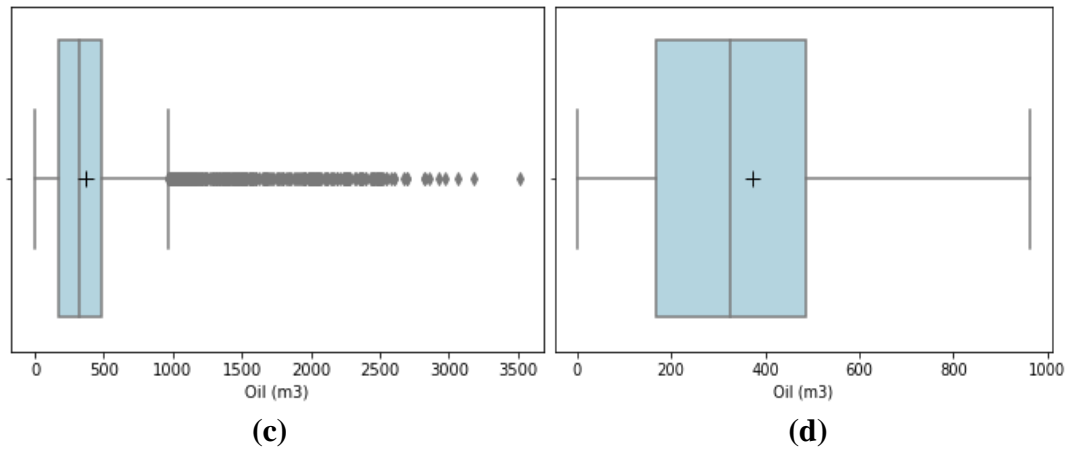
Figure A.15: (a) Kernel Density Estimation plot for Oil (m3) (b) Histogram for Oil (m3) (c) Boxplot for Oil (m3) (d) Boxplot without outliers for Oil (m3)

Gas (m3)

(a)
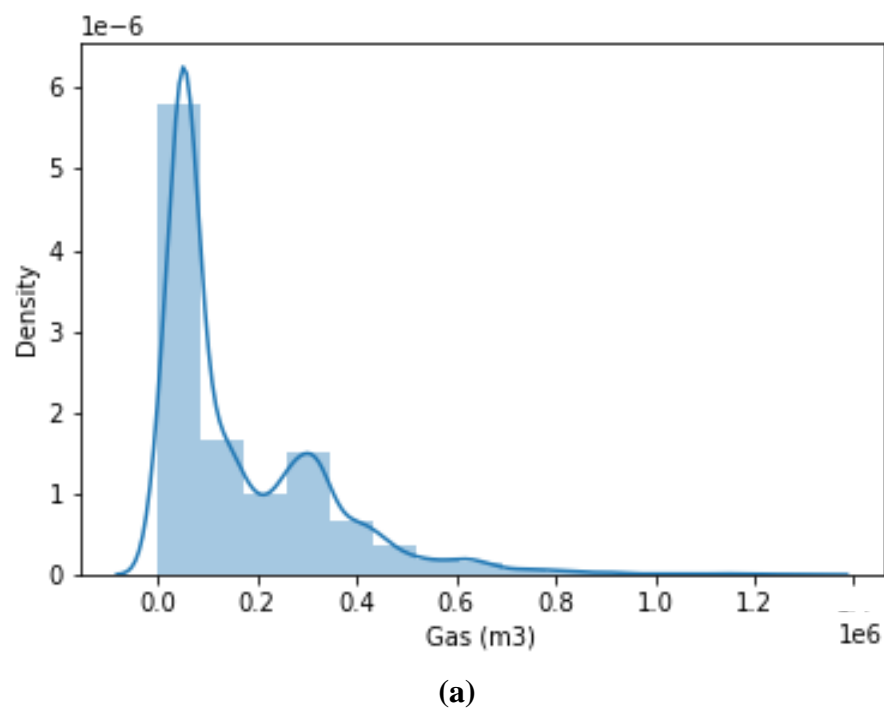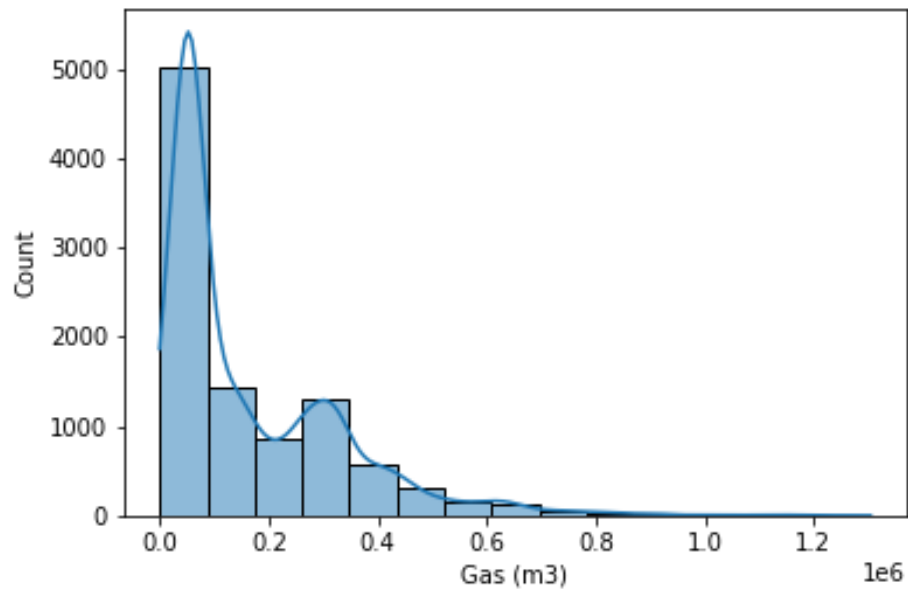
(b)
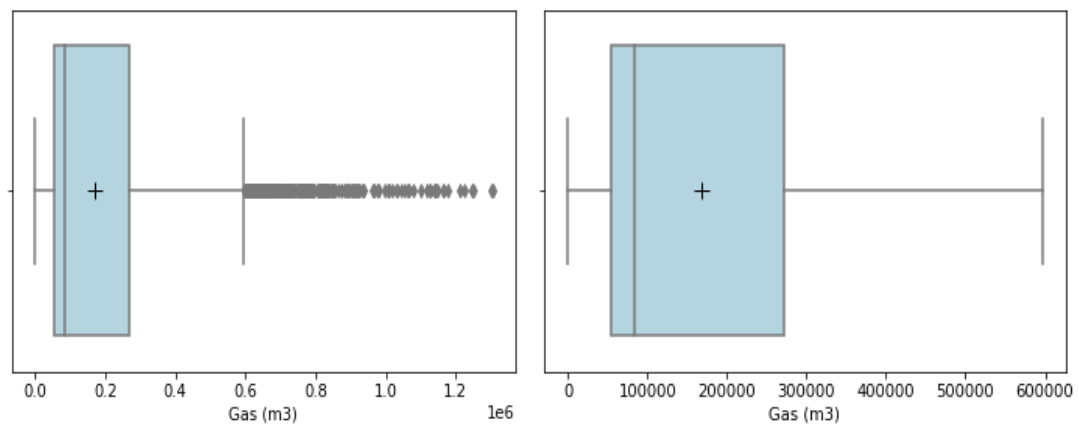
(c)

(d)

Figure A.16: (a) Kernel Density Estimation plot for Gas (m3) (b) Histogram for Gas (m3) (c) Boxplot for Gas (m3) (d) Boxplot without outliers for Gas (m3)

Av. WHT (Deg C)
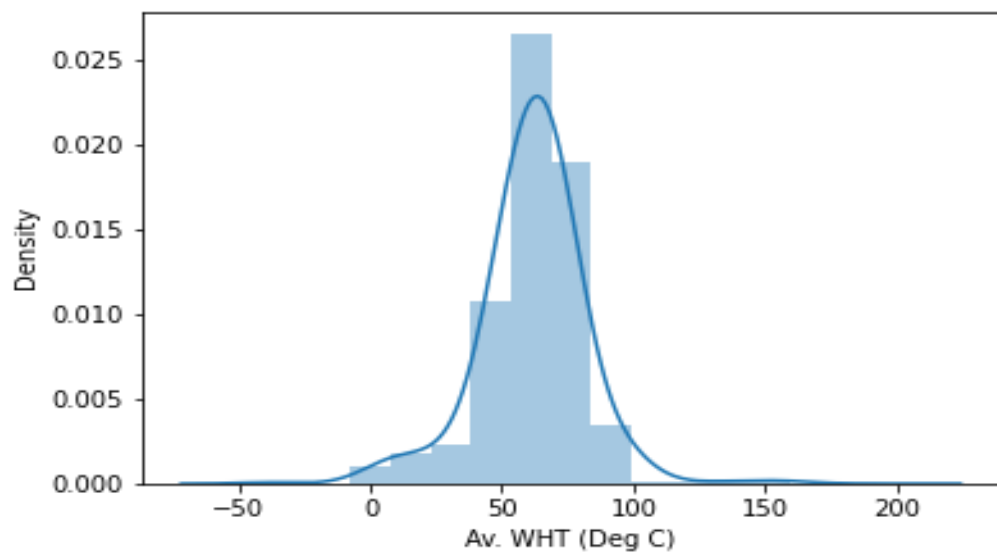
(a)
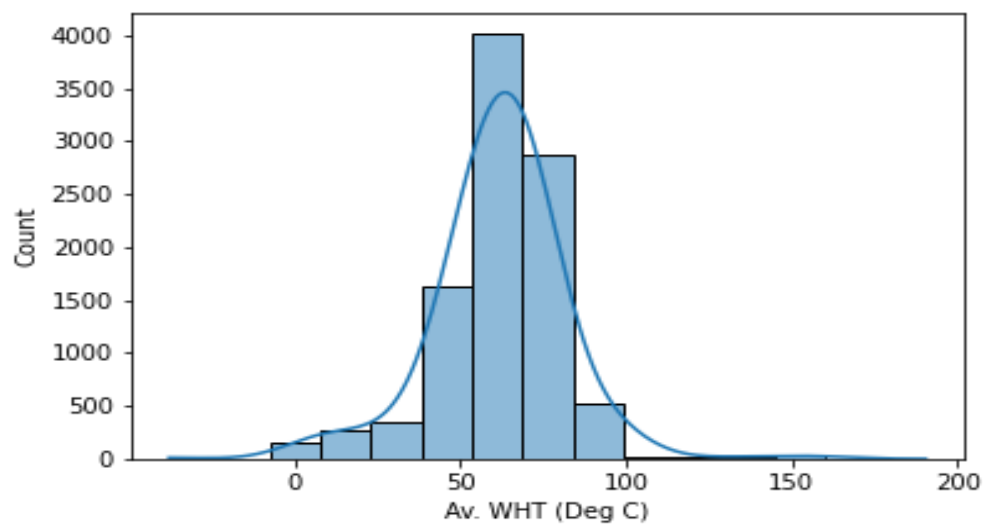
(b)
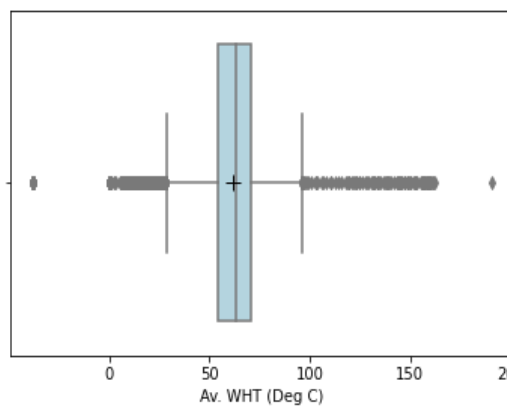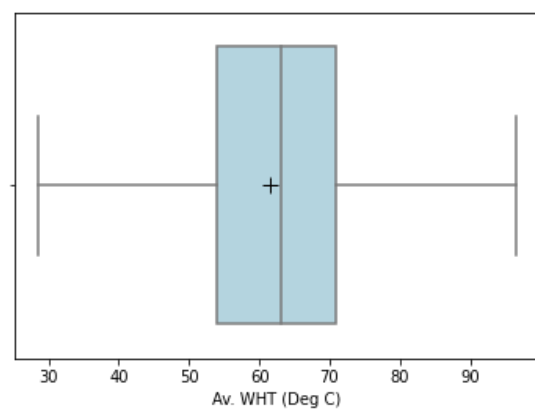
(c)   **(d)**

Figure A.17: (a) Kernel Density Estimation plot for Av. WHT (Deg C) (b) Histogram for Av. WHT (Deg C) (c) Boxplot for Av. WHT (Deg C) (d) Boxplot without outliers for Av. WHT (Deg C)

Av. WHP (bar)
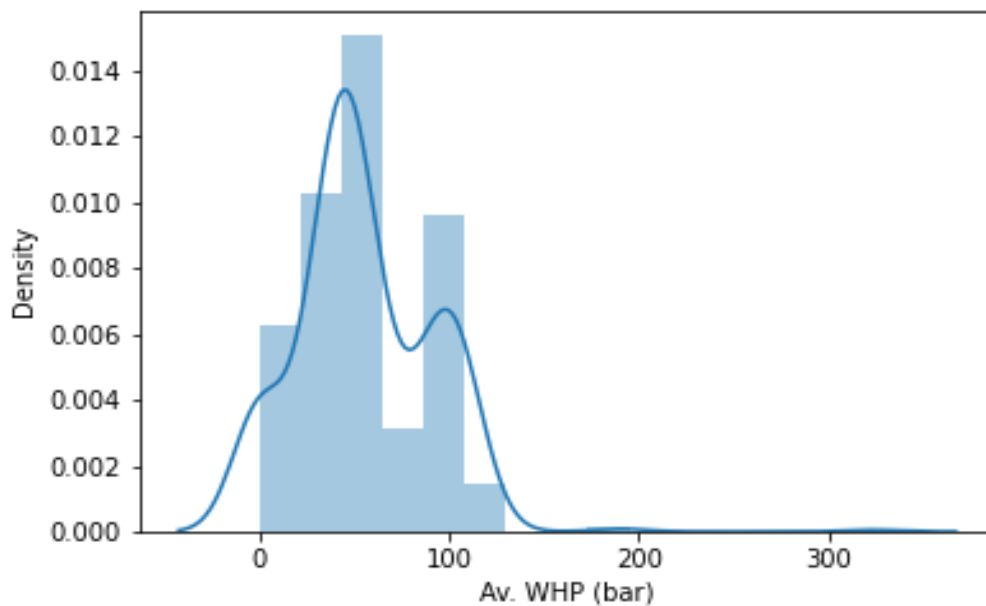
(a)
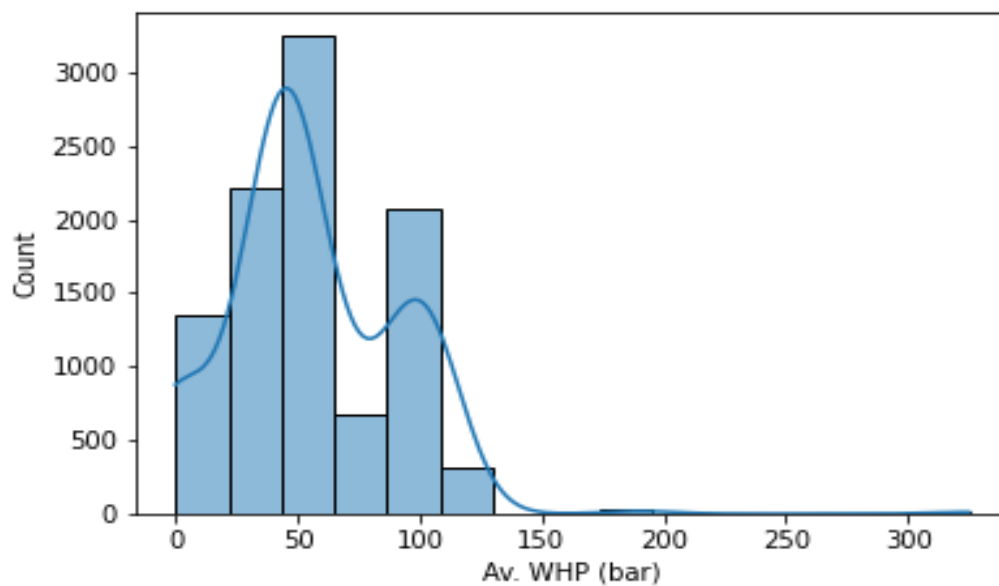
(b)
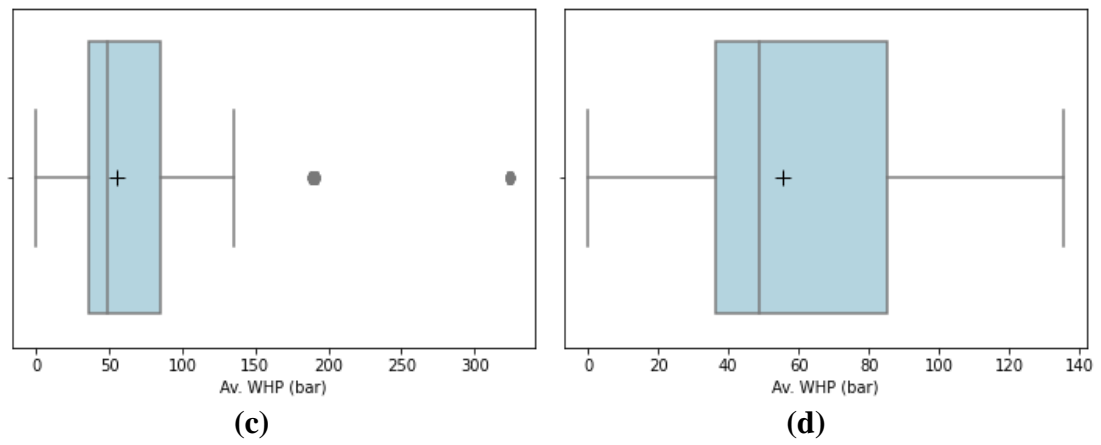
Figure A.18: (a) Kernel Density Estimation plot for Av. WHP (bar) (b) Histogram for Av. WHP (bar) (c) Boxplot for Av. WHP (bar) (d) Boxplot without outliers for Av. WHP (bar)

VI.   Feature statistics in Volve and Kyle Master dataset after forward filling

Hours Online
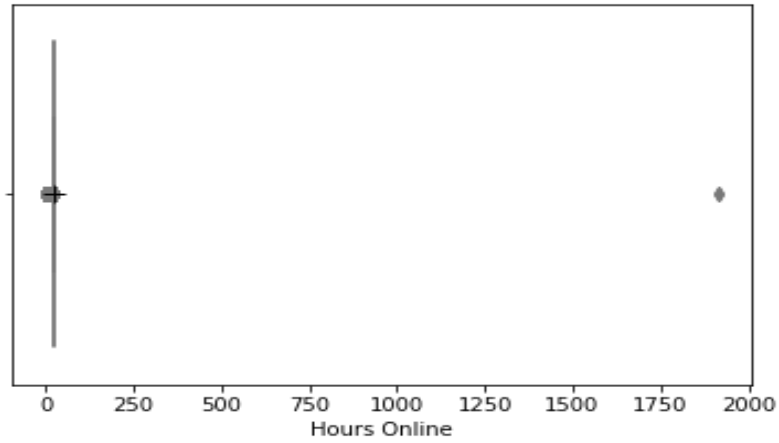
(a)
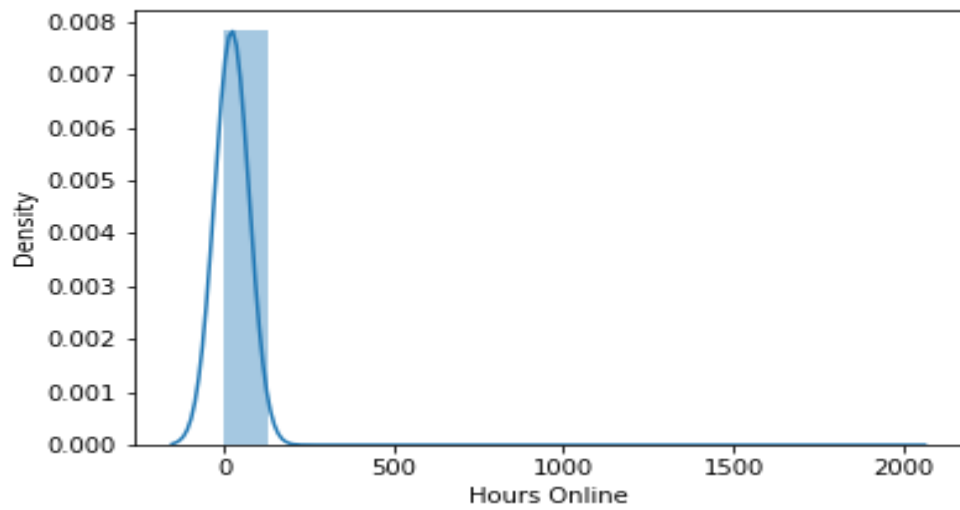
(b)
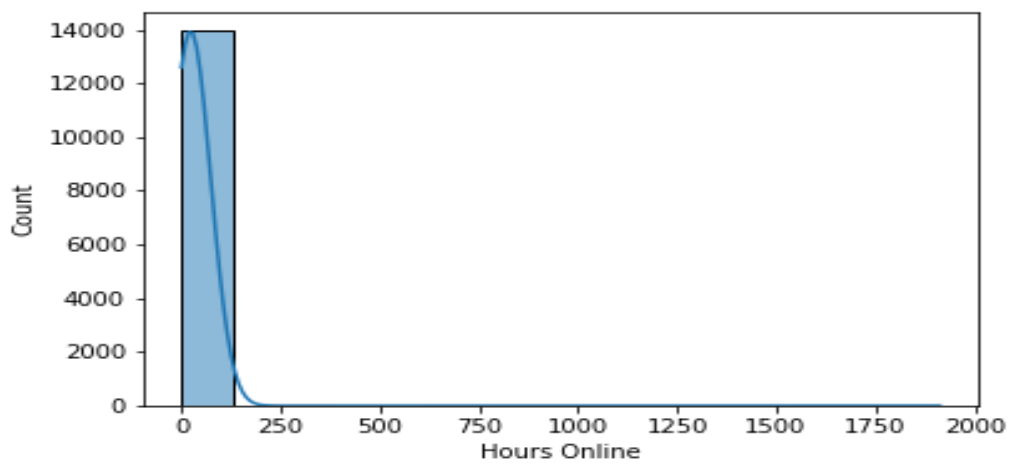
(c)

Figure A.19: (a) Boxplot for Hours Online (b) Kernel Density Estimation Plot for Hours Online (c) Histogram for Hours Online after forward filling
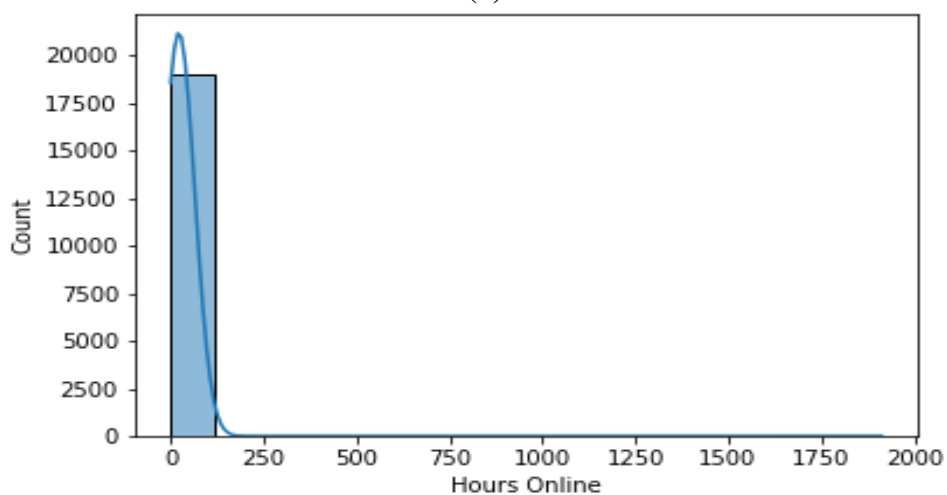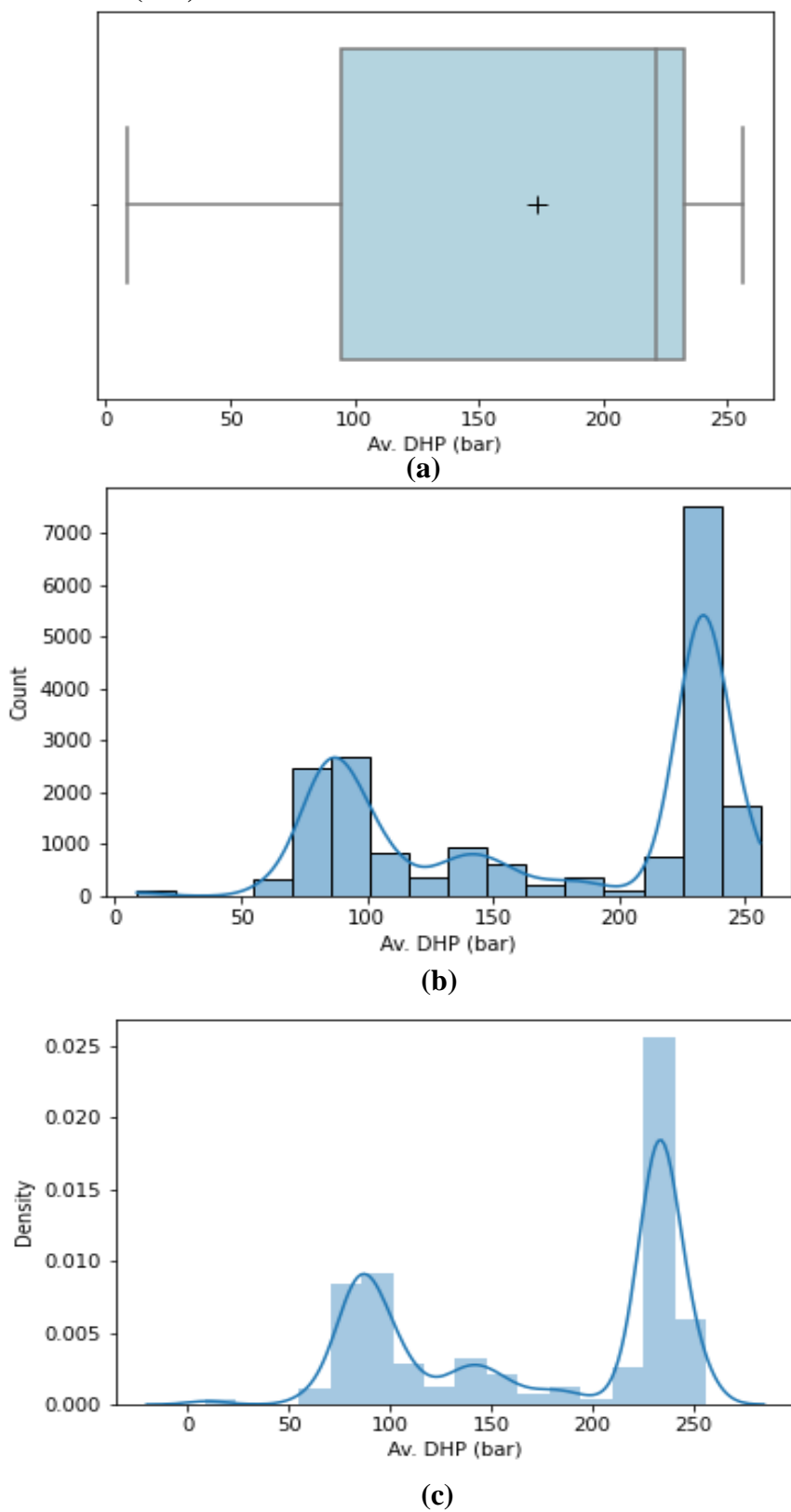
Av. DHP (bar)
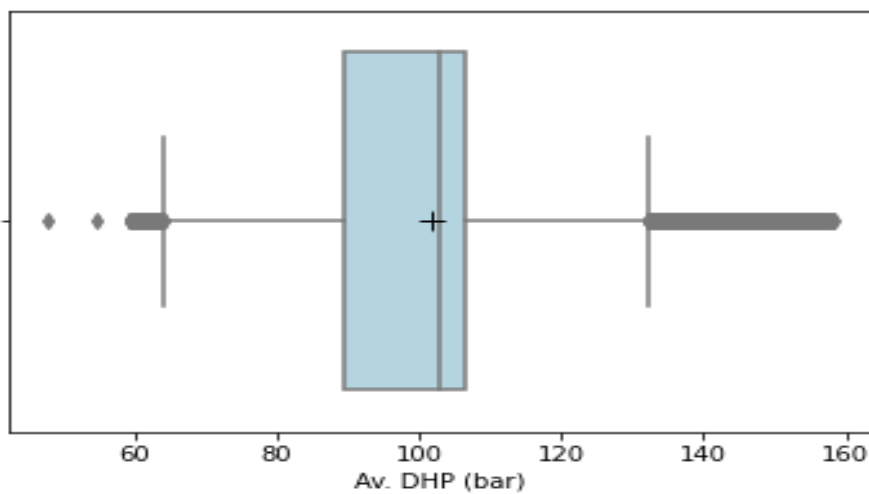
(a)
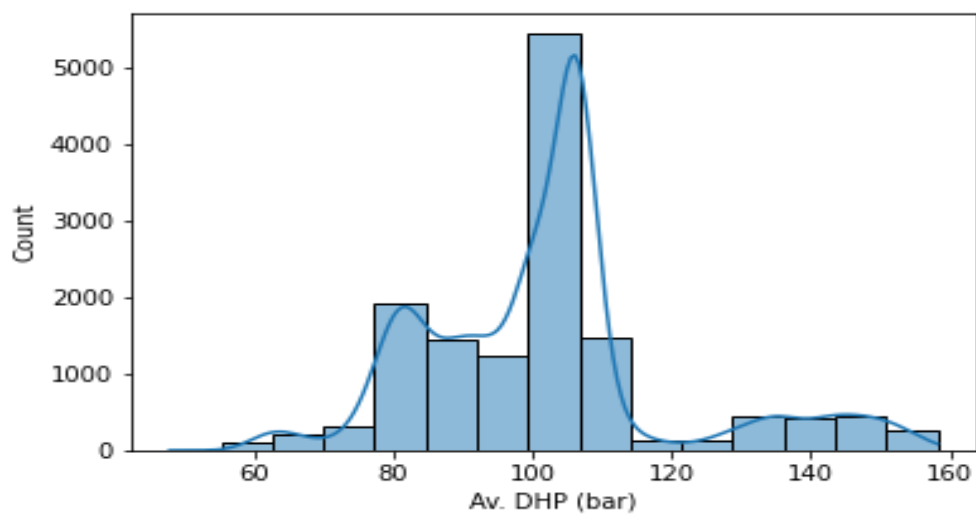
(b)
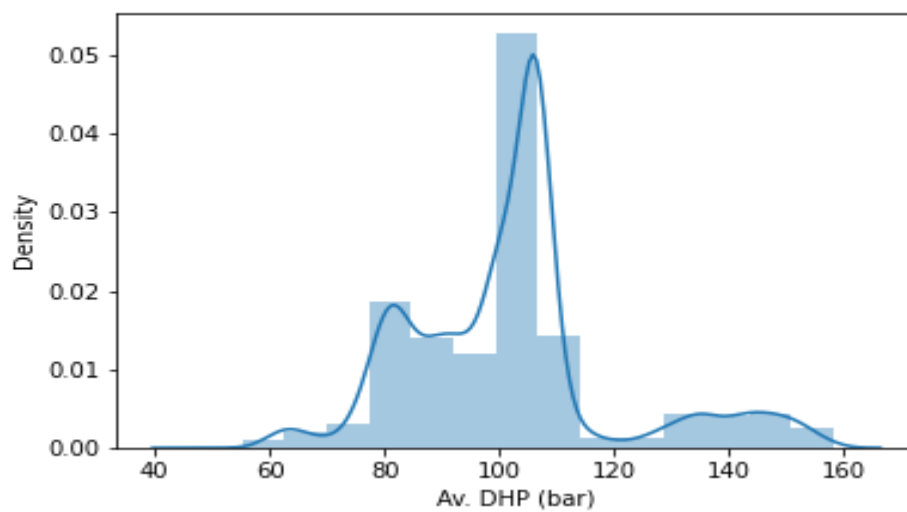
(c)

Figure A.20: (a) Boxplot for Av. DHP (bar) (b) Histogram for Av. DHP (bar) (c) Kernel Density
Estimation Plot for Av. DHP (bar) after forward filling

Av. DHT (Deg C)

(a)

(b)

(c)
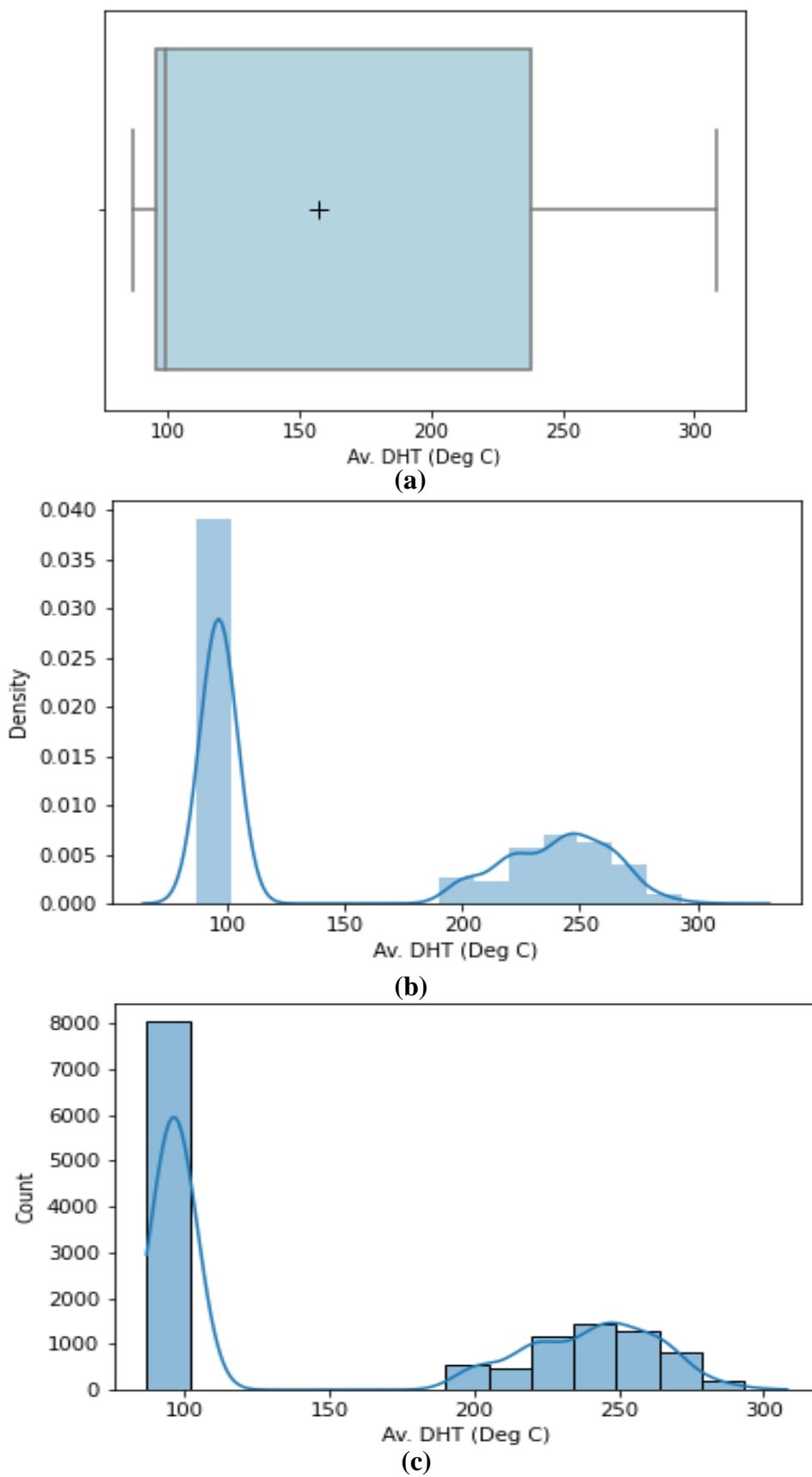
Figure A.21: (a) Boxplot for Av. DHT (Deg C) (b)) Kernel Density Estimation Plot for Av. DHT
(Deg C) (c) Histogram for Av. DHT (Deg C) after forward filling

Oil (m3)
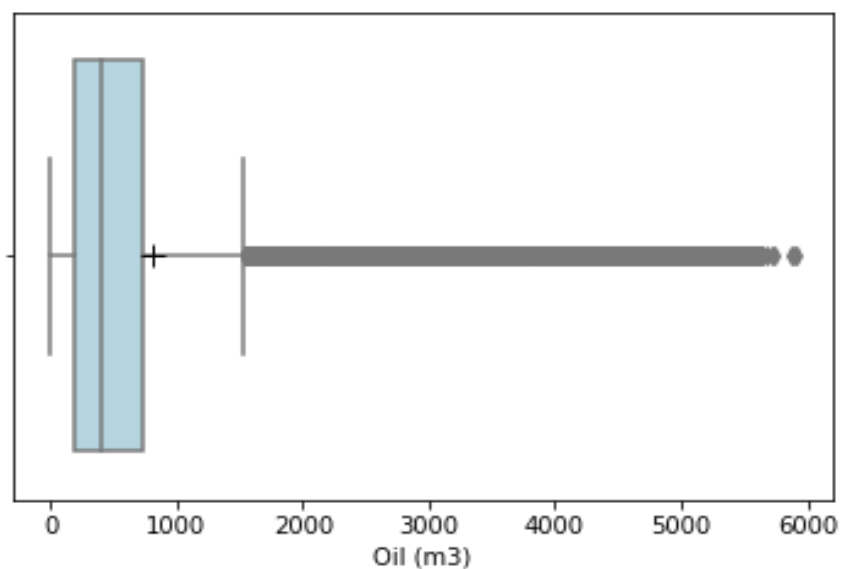
(a)
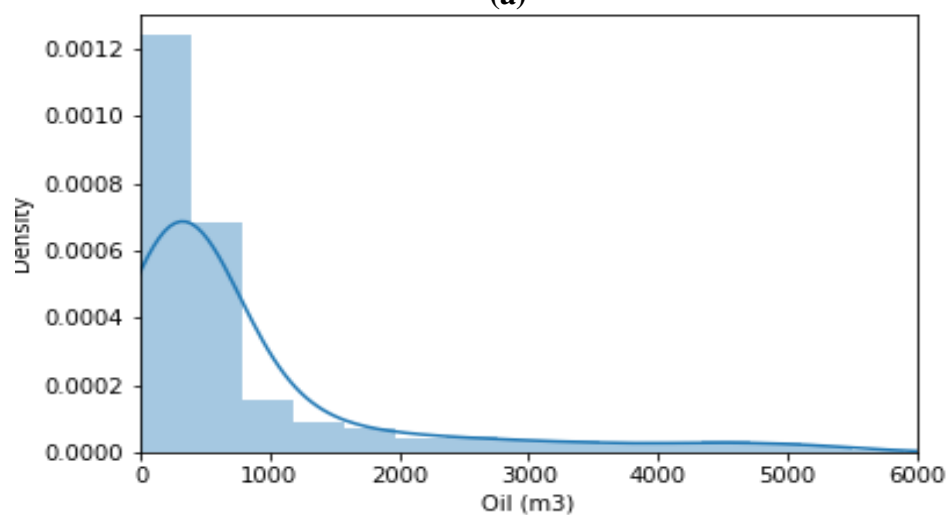
(b)
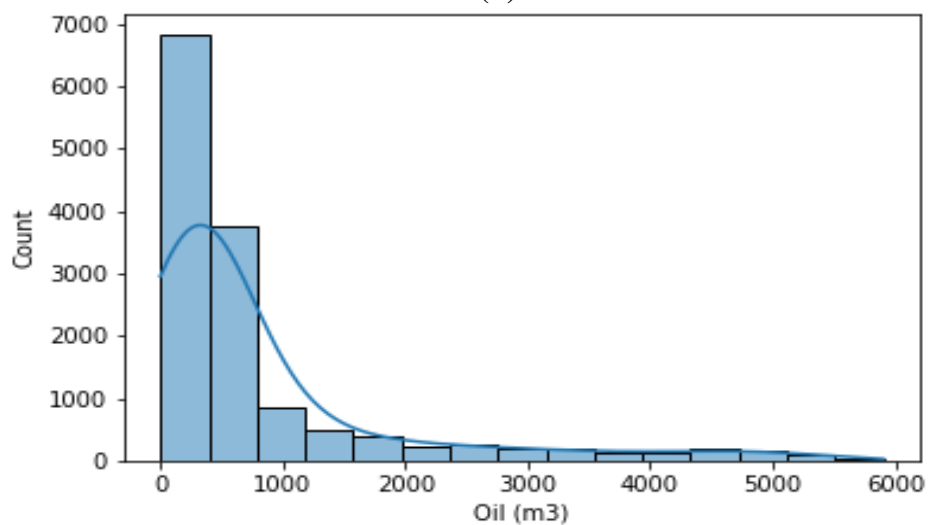
(c)

Figure A.22: (a) Boxplot for Oil (m3) (b)) Kernel Density Estimation Plot for Oil (m3) (c)
Histogram for Oil (m3) after forward filling

Gas (m3)

(a)

(b)

(c)
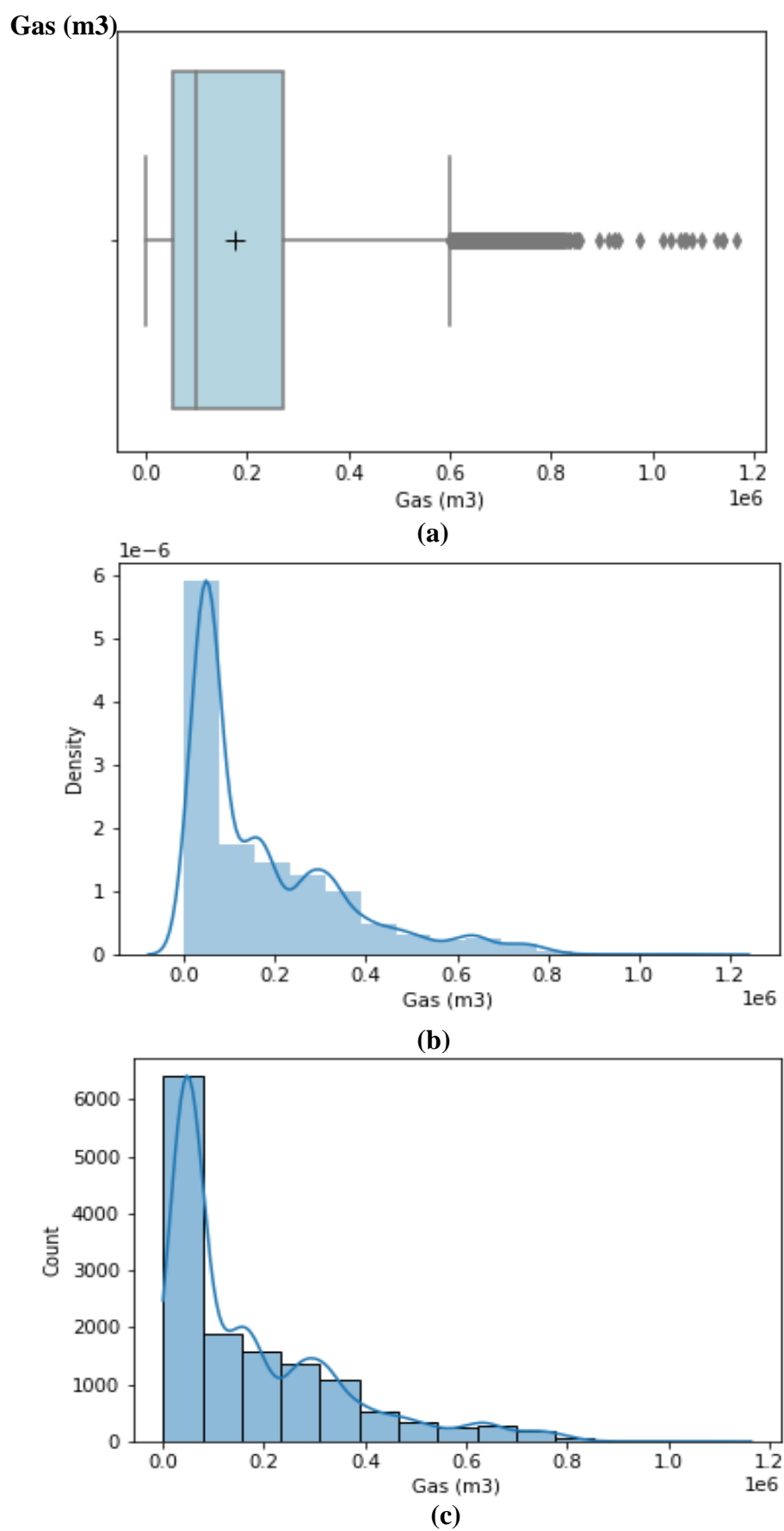
Figure A.23: (a) Boxplot for Gas (m3) (b)) Kernel Density Estimation Plot for Gas (m3) (c) Histogram for Gas (m3) after forward filling
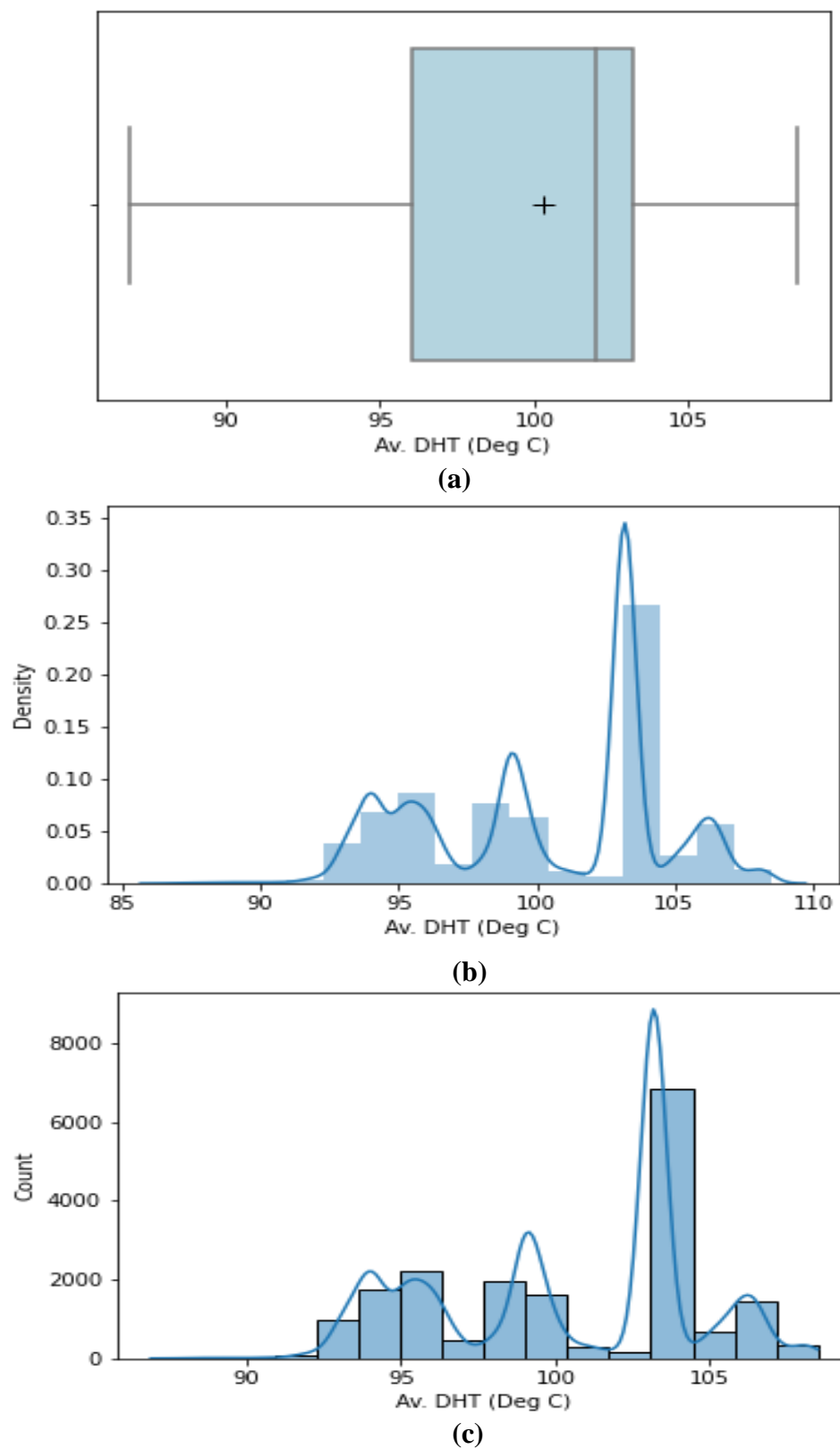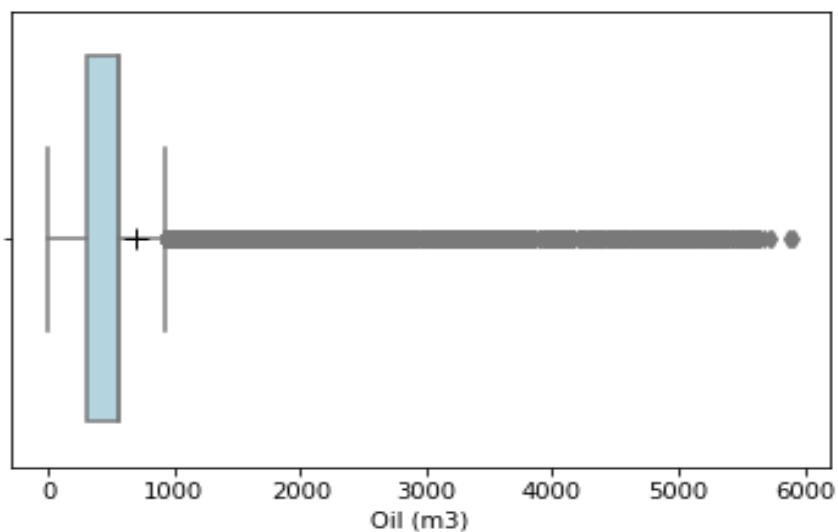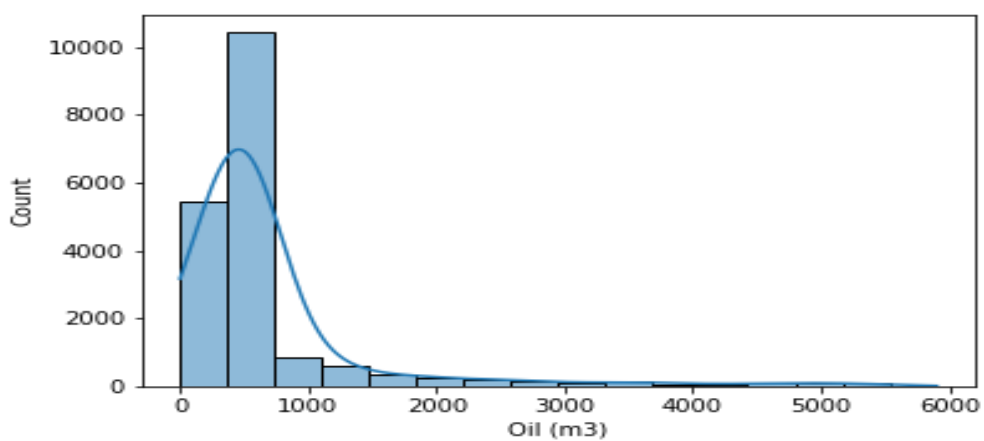
Av. WHT (Deg C)

(a)

(b)

(c)
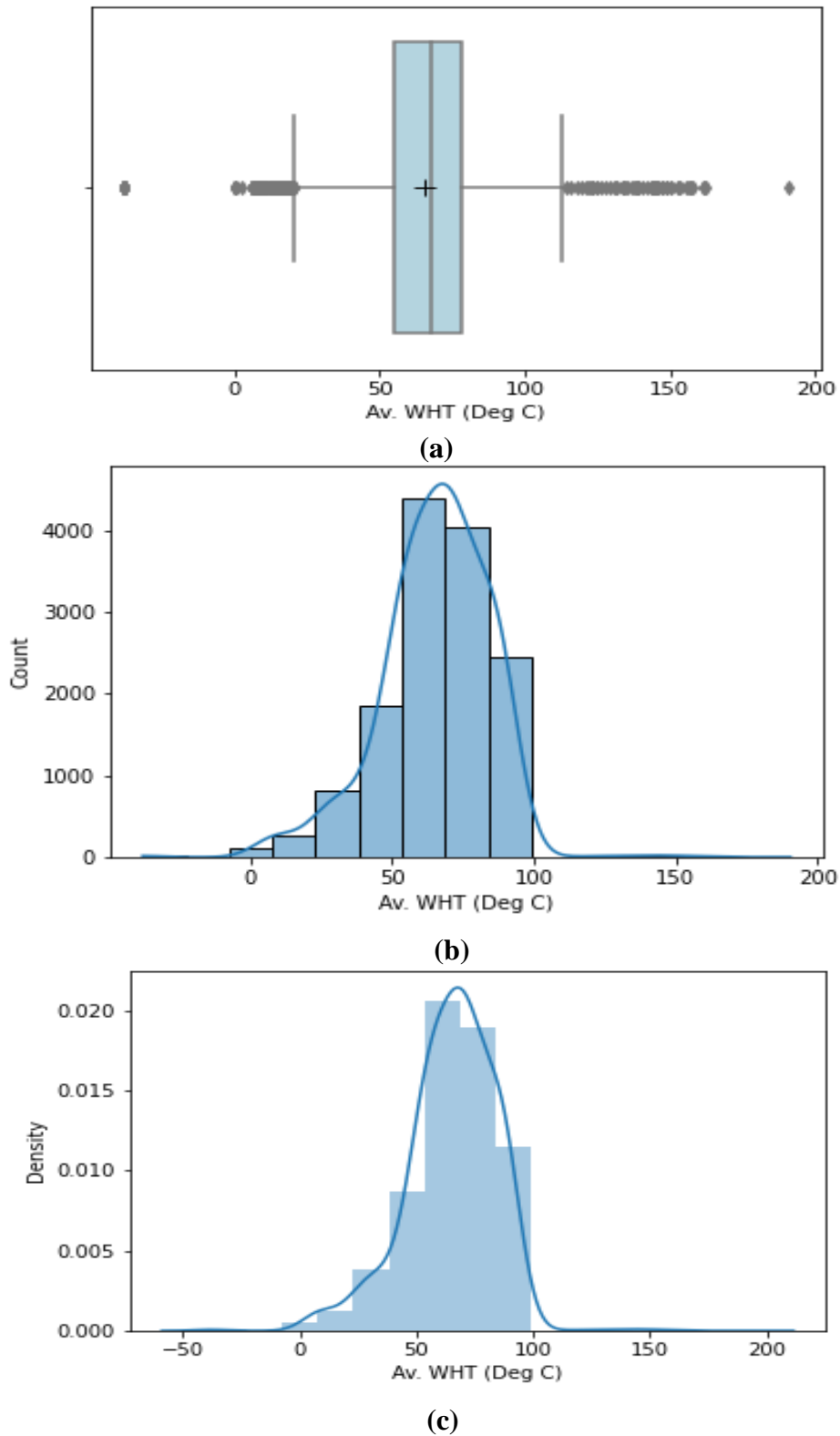
Figure A.24: (a) Boxplot for Av. WHT (Deg C) (b)) Histogram for Av. WHT (Deg C) (c) Kernel Density Estimation Plot for Av. WHT (Deg C) after forward filling
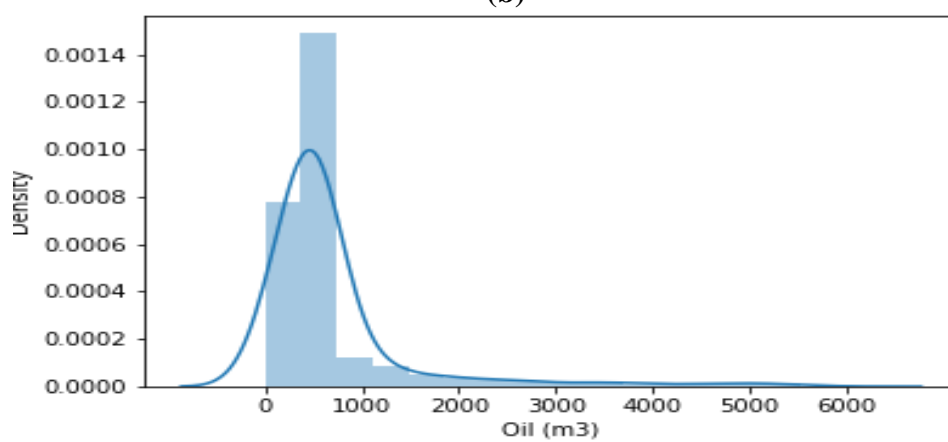
Av. WHP (bar)
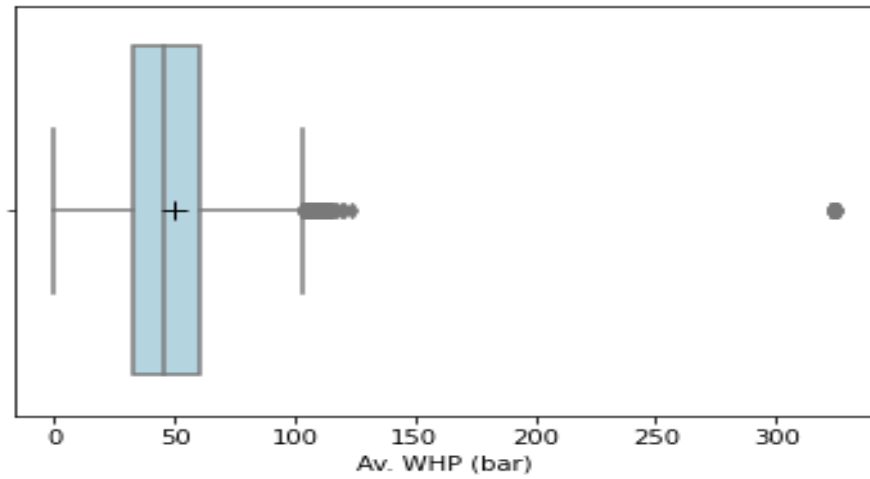
(a)
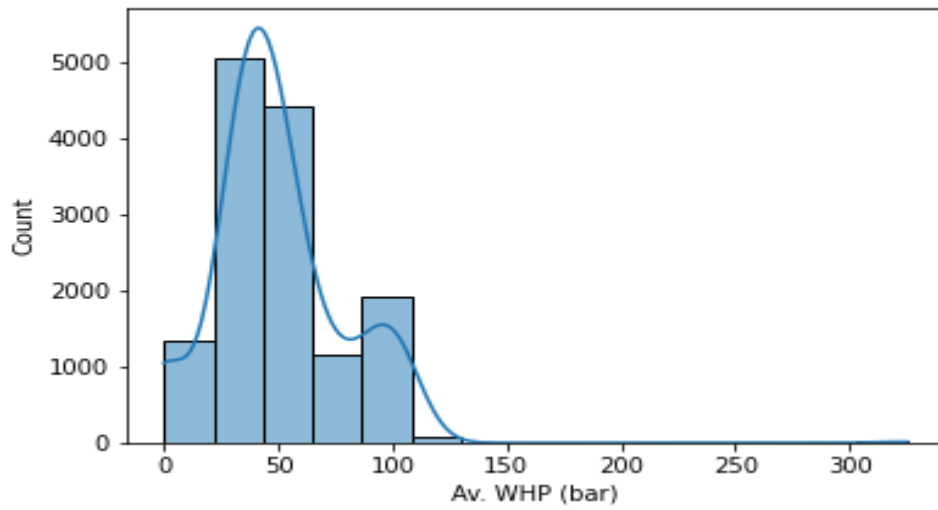
(b)
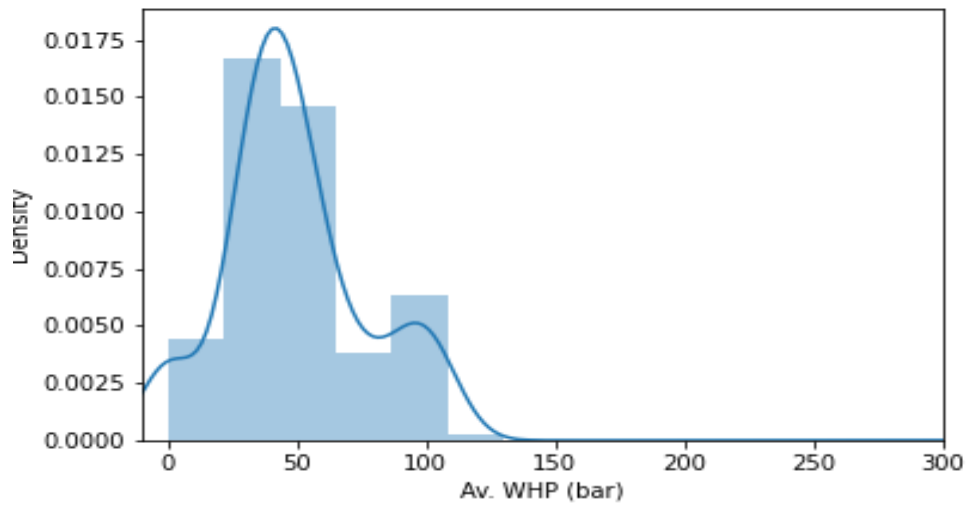
(c)

Figure A.25:(a) Boxplot for Av. WHP (bar) (b)) Histogram for Av. WHP (bar) (c) Kernel
Density Estimation Plot for Av. WHP (bar) after forward filling

VII. **Feature statistics in Volve and Kyle Master dataset after median imputation**

Hours Online
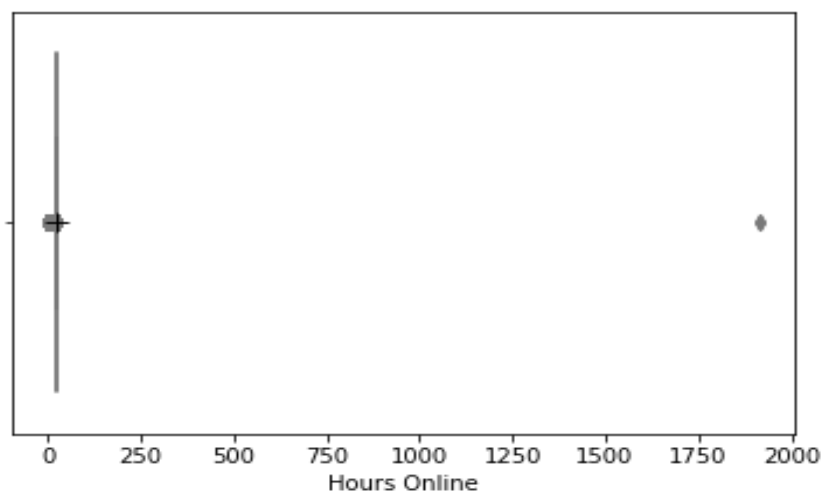
(a)

(b)
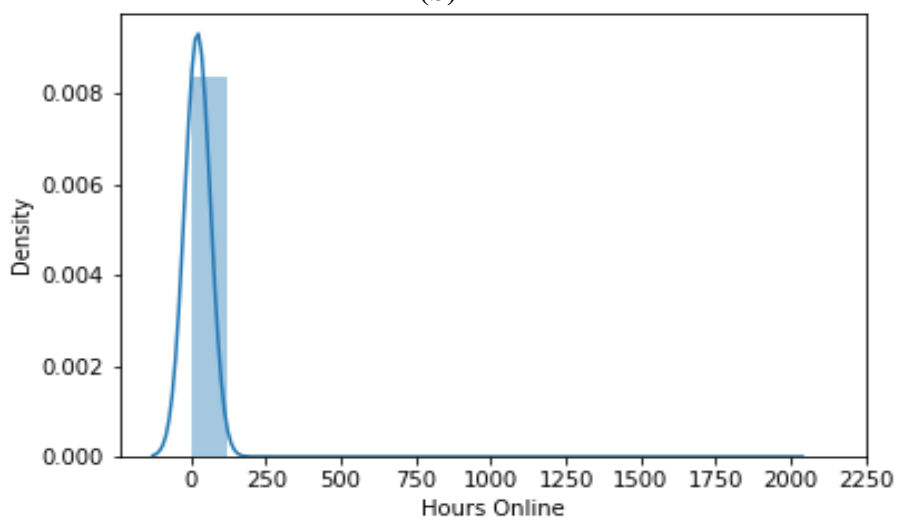
(c)

Figure A.26: (a) Boxplot for Hours Online (b)) Histogram for Hours Online (c) Kernel Density Estimation Plot for Hours Online after median imputation

Av. DHP (bar)

(a)

(b)

(c)

Figure A.27: (a) Boxplot for Av. DHP (bar) (b) Histogram for Av. DHP (bar) (c) Kernel
Density Estimation Plot for Av. DHP (bar) after median imputation

Av. DHT (Deg C)

(a)

(b)

(c)

Figure A.28: (a) Boxplot for Av. DHT (Deg C) (b) Histogram for Av. DHT (Deg C) (c) Kernel Density Estimation Plot for Av. DHT (Deg C) after median imputation
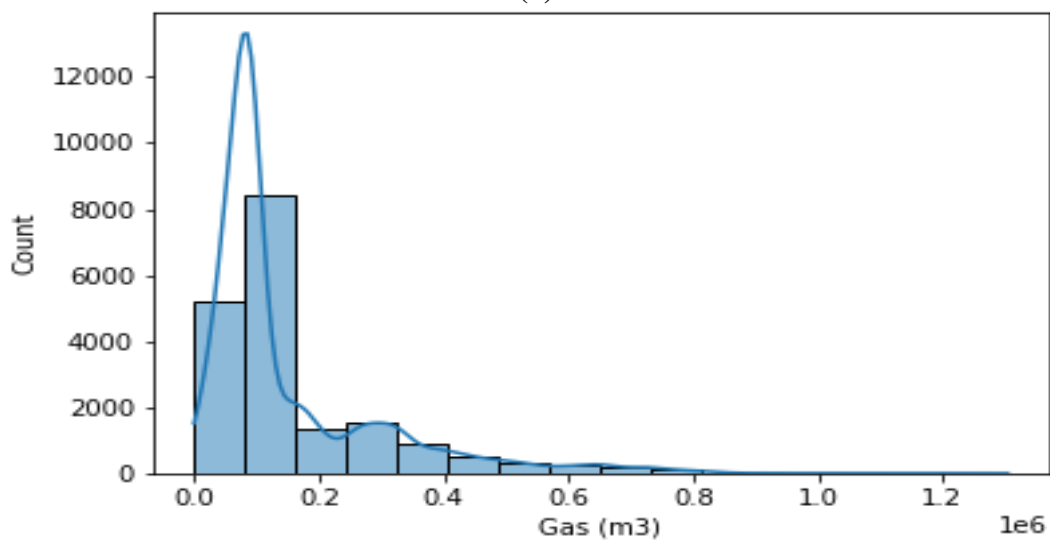
Oil (m3)

(a)

(b)

(c)

Figure A.29: (a) Boxplot for Oil (m3) (b) Histogram for Oil (m3) (c) Kernel Density
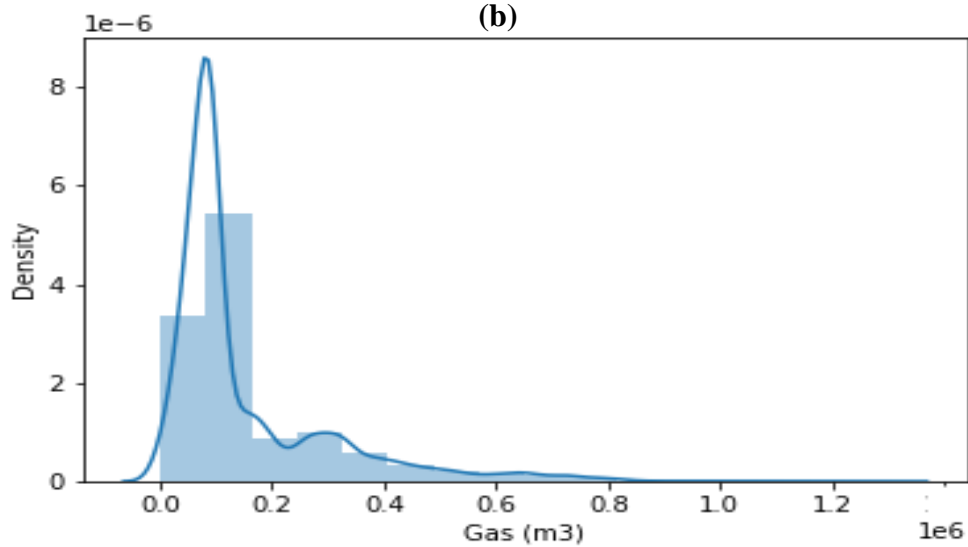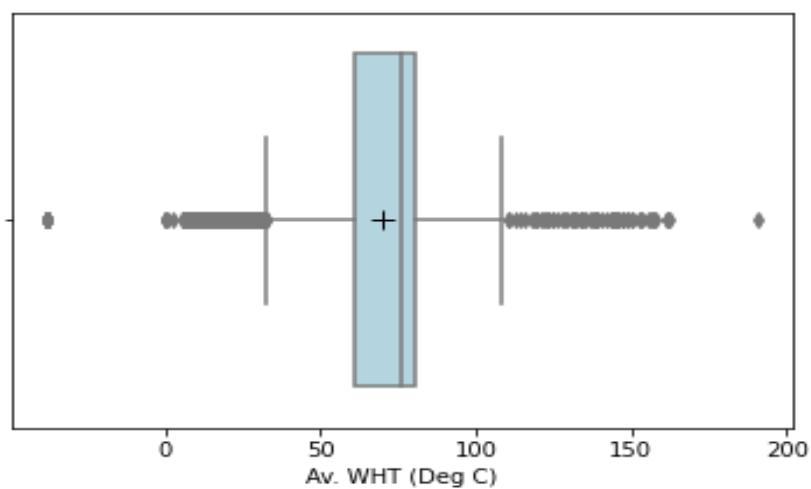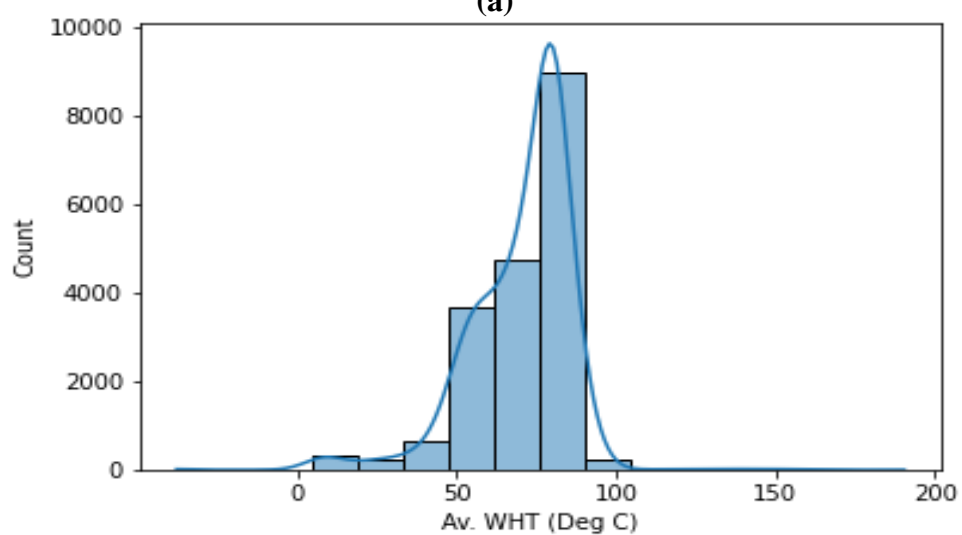Estimation Plot for Oil (m3) after median imputation

Gas (m3)
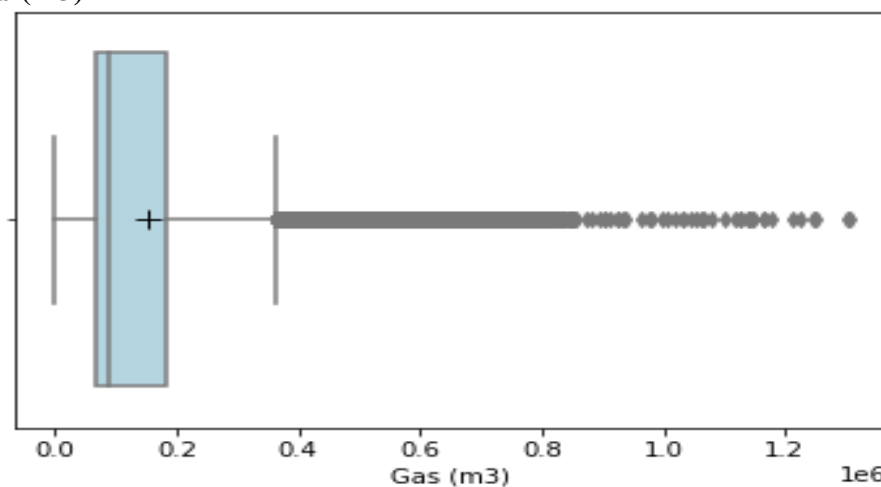
(a)

(b)

(c)

Figure A.30: (a) Boxplot for Gas (m3) (b) Histogram for Gas (m3) (c) Kernel Density
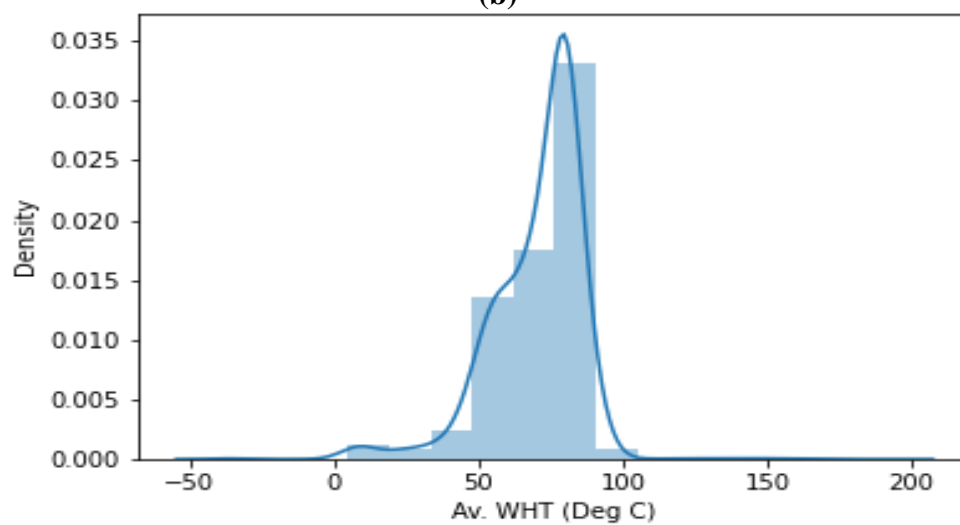Estimation Plot for Gas (m3) after median imputation

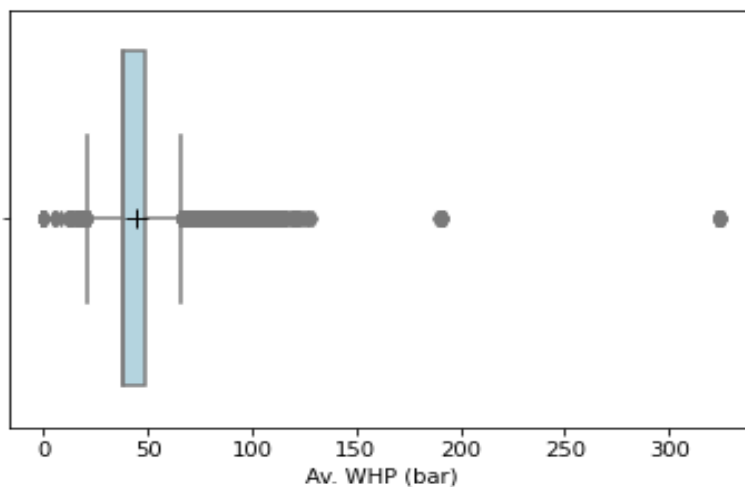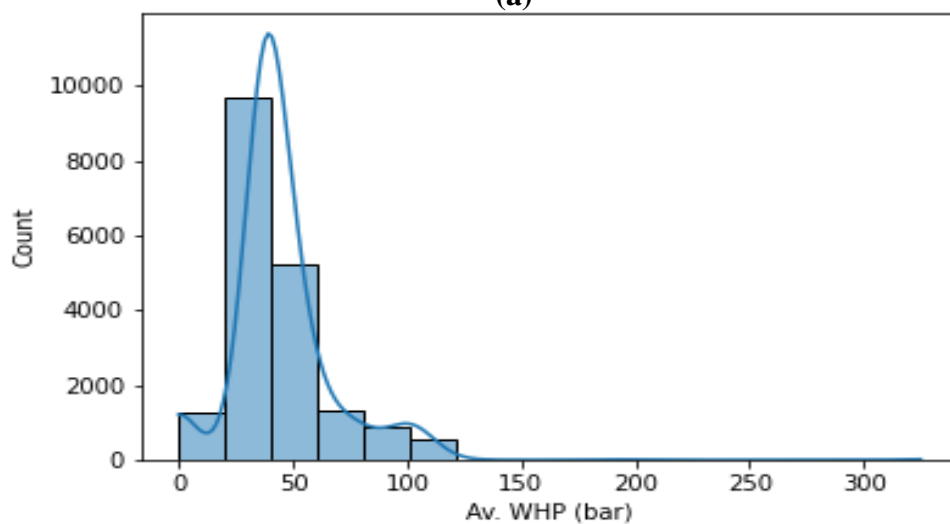**Av. WHT (Deg C)**

(a)

(b)

(c)

Figure A.31: (a) Boxplot for Av. WHT (Deg C) (b) Histogram for Av. WHT (Deg C) (c) Kernel Density Estimation Plot for Av. WHT (Deg C) after median imputation
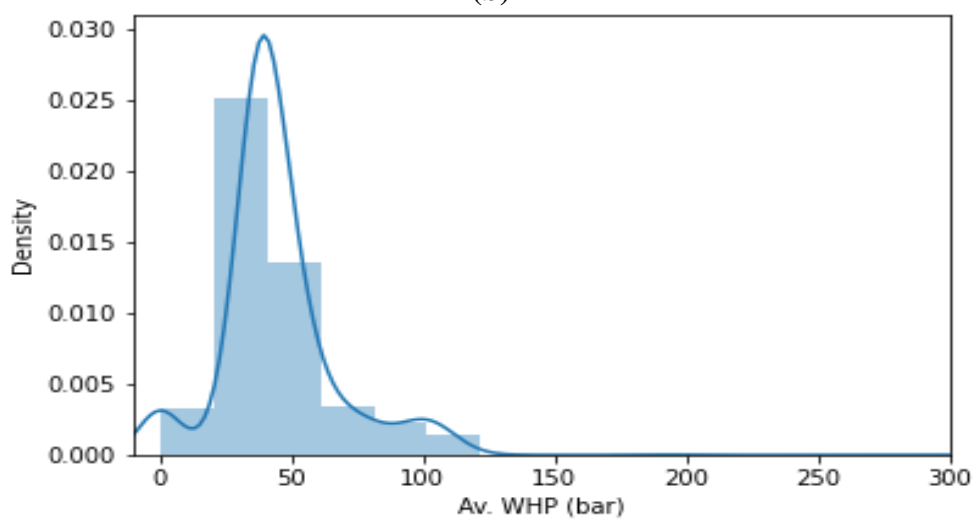
Av. WHP (bar)

(a)

(b)

(c)

Figure A.32: (a) Boxplot for Av. WHP (bar) (b) Histogram for Av. WHP (bar) (c) Kernel Density Estimation Plot for Av. WHP (bar) after median imputation

Appendix B

VDR Website Application

The party involved in this development consist of:

Product Owner:	**Ardimas Andi Purwita**
Client:	**Widodo Nugroho - PT Geodwipa Teknika Nusantara (GTN)**
Developers:	**Chan Elizabeth**
	Kotrakona Harinatha Sreeya Reddy
	Vicky Vanessa

Problem Statement:

The client wishes to have a more cost-effective application as an alternative since the application they were using is not cost-effective enough for the client's company.

Scope of work of the project:

The developers are to create a website application frontend prototype for the VDR Website Application to Visualize Production Data for the Oil and Gas Industry. The aims of a fully functional prototype are:

- visualizing oil and gas data,
- being more cost-effective, and
- a predictive model capable of predicting oil and gas production.

The prototype will consist of the features that the client demand which are:

1. Visualization of oil and gas data

   Building a viewer's page which allows the user to choose which file they choose to visualize. Visualization is available in 2D and 3D images, i.e., well logs and seismic data. The file they choose is the file that the user uploaded in the file management, which will be described later.

2. Map application along with the showcase feature

   An application which allows the user to see and upload the location of oil and gas reserves. It shows the data the client has for the reserves for the showcase, e.g., tabular data, snapshots of the location.

3. Prediction model to predict the oil and gas production

   Using pressure and temperature sensor data to predict oil and gas production values so that the user can focus on wells that contain more oil and gas.

4. File management to store the user's files

   The user can store their files and store them into folders. These files can and will be used in the other features.

Approved by

Ardimas Andi Purwita

Product Owner

Widodo Nugroho

Client

Figure B.1: Proof for Scope of Work