Table of Contents

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 key elements to the growth of the economy. There has 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 [3]. Additionally, the oil and gas industry has also provided jobs to thousands of individuals [2].

There are many oil and gas reserves in different corners of the world. In Indonesia in particular, the Energy Ministry has recorded that 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, demand for oil and gas in Indonesia is rising [5]. Furthermore, 50% of Indonesia’s energy is derived from oil [5]. This results in Indonesia importing nearly 350,000 BPD and 50,000 barrels of fuels per day from other countries [5].

Oil and gas have many uses and have a strong impact on a country’s economy. Therefore, oil and gas industries often make use of dashboard software applications in order to help them manage it. There are numerous dashboard software applications available such Lynx and INTviewer. These applications are similar yet they also have their differences. Lynx offers petroleum data services, geophysical and Geographical Information System (GIS) services [6]. 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 $60,000 per year [9]. These types of applications can benefit oil and gas industries greatly, however these software applications are expensive. Therefore, the goal of this project is to develop a web-based dashboard application with similar features intended for the oil and gas industries in Indonesia at a lower cost.

1.2 Scope

In this project the author’s responsibility was to create a predictive model capable of predicting oil and gas production. The author had to collect and scrape useful data in order to make a dataset. This dataset would then be processed and cleaned in order to be used to train the machine learning model. After training the model, the author will connect the AI to the website created by the other members of the author’s team by using RestAPI.

Student	Role
Elizabeth Chan	Design UI/UX of the GIS Developing a visualization of information-bearing heat mapped map using GIS Testing (e.g. unit test & integration test)
Kotrakona Harinatha Sreeya Reddy	Collecting and processing data Using the data collected to develop predictive models Visualizing data through diagrams such as charts as well as doing data analytics
Vicky Vanessa	Designing UI/UX of the frontend of the website application Visualizing the data of oil, gas, and water Testing

Table 1.1 : Scope of Student Activities

1.3 Objective

This section will describe the author's aims as well as the vision and mission of the author and the author's team.

1.3.1 Aim

The main aim of this 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. This website application aims to obtain latent information from oil and gas production data which will be used to build a predictive model. This predictive model will help the oil and gas industry by showing areas that are more likely to contain more oil and gas.

1.3.2 Vision and Mission

The vision of the author’s team is to increase the use of local service in the oil and gas industries to gain more profits and indirectly increase the national income. Another vision of the author and the author’s team is to make this website known internationally. The mission of the author and the team is to develop a high quality yet affordable web-based application. This application will consist of essential features and exclude features that are unnecessary, which would lower the cost for the buyer. In addition to this, this application will also help engineers comprehend complex data and gain better insights on how the data can be used

1.4 Structure

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

1.4.1 Chapter 1

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

1.4.2 Chapter 2

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

1.4.3 Chapter 3

Chapter 3 will detail the problem even further and describe the data and metrics used by the author to create the predictive model

1.4.4 Chapter 4

Chapter 4 focus 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

CHAPTER 2

THEORETICAL FOUNDATION

This chapter will cover theories regarding oil and gas production and provide insights on which factors impact oil and gas production. Furthermore, it will delve into Machine Learning techniques, Deep Learning and Prediction Models.

2.1 Mining Oil and Gas

As has been mentioned previously, oil and gas are resources with high demand as it can be utilized in a variety of ways. Therefore, there is a great necessity to mine these resources. Oil and gas are mined in an underground reservoir, which is an oil bearing rock or formation. A rotary drill using a drilling fluid is used to drill into the reservoir [20]. This drill needs to have proper rheology as it needs to form a low permeability filter cake on the wellbore surface and also contain specific characteristics to avoid chemo-mechanical mishaps [20].

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 volume of oil and dissolved gas at a certain temperature and pressure which is needed to manufacture one barrel of oil [21]. B_o is either greater than or equal to unity [23].

The equation for oil formation volume factor is :

$$B_o = \frac{(V_o)pT}{(V_o)_{sv}} . \qquad (2.1)$$

In Equation 2.1, *Bo* is the oil volume factor, *(Vo)* is the volume of oil, *(Vo)sc* is volume of oil measured under standard conditions, *p* is the pressure at the reservoir whereas *T* is the temperature at the reservoir.

From equation 2.1, it can be inferred that temperature and pressure are important 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 [21]. Afterwards, the oil will again expand as the pressure increases [21]. Often the effect of the temperature and pressure change when the oil reaches the surface is minimal and will cancel out one another [21].

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 certain temperature and pressure which is needed to manufacture one standard volume of gas [22]. This relationship can be expressed as :

$$B_g = \frac{V_{p,T}}{V_{sc}} . \qquad (2.2)$$

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 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 [24]. Variable z can be expressed as :

$$z = \frac{V_{actual}}{V_{ideal}} \tag{2.4}$$

where V_{actual} is the actual volume of n-moles of gas at a certain temperature and pressure, and V_{ideal} is the ideal volume of n-moles of gas at the same temperature and pressure [24]. 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.

$$B_g = \frac{zTP_{sc}}{T_{sc}P}. \tag{2.5}$$

In Equation 2.5, *Bg* 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 z-factor at standard conditions (1.0) [24]. 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 :

$$B_g = 0.0283\frac{zT}{P}. \tag{2.6}$$

2.2 Machine Learning

Machine Learning 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* [10]. It is a branch of Artificial Intelligence (AI) that centres on using the data and algorithms to echo the way humans act and learn [11]. Machine learning helps uncover insights, make classifications and predictions from the data given in order to aid users [11]. Machine learning depends on a dataset, which is a collection of data which will be regarded as one unit by the machine [12]. 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. In the context of the author's problem, a high quality dataset would mean a dataset which provides oil production value, gas production value and data regarding sensors such as pressure and temperature.

Machine learning learns in three different ways, namely supervised, unsupervised and reinforced learning.

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 [13]. The input variables are fed into the machine learning model and after the training phase, it will apply what it learned onto unknown data [14]. This type of machine learning is one of the most common methods and usually used for classification and regression problems. There are several types of supervised machine learning models, namely Naive Bayes, Linear Regression, Support Vector Machine (SVM), KNN and others [14].

Unsupervised learning is another part of machine learning and AI, however it is not as widely used compared to supervised learning. Unlike supervised learning, unsupervised learning makes use of unlabelled and unclassified training data. This type of learning aims to obtain meaningful information from these data [16]. Unsupervised learning is often used for clustering problems such as gene sequence analysis, item recognition, market research and many more [16]. There are several types of unsupervised learning models, such as hierarchical clustering, apriori algorithm, k-means clustering, and many more [17].

Reinforcement machine learning is a complex part of AI that is gaining more popularity. Reinforcement learning is the process of training a machine in order to make a sequence of decisions during a certain scenario [19]. The core of reinforcement learning is making decisions sequentially, therefore the output is reliant on the state of the current input and the following input would then be reliant on the output of the previous input [19]. Reinforcement learning is commonly used for games, industrial automation, robotics, traffic control systems and many more [19]. Algorithms for reinforcement learning include Q-Learning, State Action Reward State Action (SARSA), and Deep Q Neural Networks [19].

In terms of this project, the most suitable type of machine learning would be Supervised Learning as there are datasets available which contain production values of oil and gas, as well as values of sensors such as temperature and pressure.

2.3 Handling Missing Data

2.3.1 Mechanism of Missingness

Missing data in a dataset could prove problematic as it could affect the model's ability to perform well. There are three possible mechanisms for missing data in a dataset, these mechanisms are :

1) *MCAR (Missing Completely at Random) :* 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 it would not greatly impact the model parameters [25].
2) *MAR (Missing at Random) :* In MAR, the missing values are also random such as in MCAR, however there are possibilities of the data in question to be dependent on other values in the dataset. In this situation, the missing values are also considered negligible as the model can include covariates which represents the missing values [25].
3) *MNAR (Missing Not at Random) :* The missing values are strongly dependent on the other values in the dataset, both missing and present. MNAR is the most serious reason for mechanism for missing data as it cannot be ignored and could affect the model's performance [25]. In these cases, it is recommended to validate the data gathering process [25].

In order to counteract the effects of the missing values on the dataset, there are several actions that can be done. These actions include, *Listwise Deletion, Pairwise Deletion, Central Value Imputation,* and *Regression Imputation*.

2.3.2 Listwise Deletion

Listwise deletion, also known as row deletion, is the process of removing all data that contains a missing value [26]. This strategy is one of the most commonly used method for dealing with missing values. Furthermore, many statistical software packages utilizes this strategy for data analysis [26]. However, this procedure is only effective if the dataset follow the MCAR mechanism. If the dataset does not follow the MCAR mechanism, then the listwise deletion would result in biased parameters [26]. This could cause the model to perform poorly during its training. In addition to this, using listwise deletion is not preferable if the dataset size is small or the percentage of missing data is large [27]. This is because dropping the missing values in this situation would greatly reduce the chances of the model to learn meaningful observations.

2.3.3 Pairwise Deletion

The concept of pairwise deletion is derived from listwise deletion, however, it attempts to reduce the amount of data loss that occurs due to listwise deletion [26]. Pairwise deletion does this by only deleting the missing value and leaving the other variables in the row intact [27]. This mechanism is often preferred over listwise deletion as it allows more data to be used as it does not delete the entire row. However, pairwise deletion only yields optimal results if the missing data mechanism is MCAR. If the missing data mechanism is MAR or MNAR, it could result in biased estimates and parameters in the model [26].

2.3.4 Central Value Imputation

Central value imputation is the process of filling in the missing data in the dataset with their central tendencies [27]. These central tendencies could either be the mean, median or mode. The mode is normally used to fill in the missing data for categorical variables whilst the mean and median are often used to fill in for numerical variables [26]. 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 [28]. Additionally, filling the missing values with the mean could reduce variance in the dataset [27].

2.3.5 Regression Imputation

Regression imputation is process of using present variables to make an prediction which would be used to fill in the missing values [26]. One of the advantages of regression imputation is that it saves a lot of the dataset in contrast to other methods such as listwise deletion [26]. Furthermore, this method also does not greatly alter the shape and standard deviation of the data distribution [26]. However, this method could result in overfitting and also weakens variance [26].

2.4 Outliers

Outliers can be defined as *an data in a dataset that strays from the other data* [35]. It is necessary to detect these outliers as it could skew the model's training which would reduce the accuracy of the model [34]. Outliers removal is usually one of the earliest steps in a machine learning problem [34]. There are several methods that can be utilized in order to identify these outliers. One of those methods is by using the *Tukey's*

Method. The Tukey's Method is based on statistics where data is expected to follow a distribution model such as normal distribution [36]. A data is considered an outlier it deviates from the model [36]. 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) [36]. 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 r^{th} 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 [36]. The Tukey's Method involves calculating the Interquartile Range (IQR) between the lower quartile and the upper quartile in a boxplot [36]. 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 falls between the upper limit and lower limit [36].

2.5 Correlations

Correlation is known as a statistical measure which describes how one feature is related to another feature [39]. 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 [39]:

1) *Positive Correlation :* A positive correlation denotes when the value of one feature increases, the value of the other feature increases as well [39]. In a graph format, a strong positive correlation would have a positive gradient as shown in Figure 2.1.
2) *Negative Correlation :* A negative correlation indicates that when the value of one feature increases, the value of the other feature would decrease [39]. A negative correlation would have a negative gradient as shown in Figure 2.2.
3) *No Correlation* : No correlation indicates that the features being assessed are not related, therefore, a change in one feature would not impact the other feature [39]. In a graph format, features with no correlation would look like Figure 2.3.

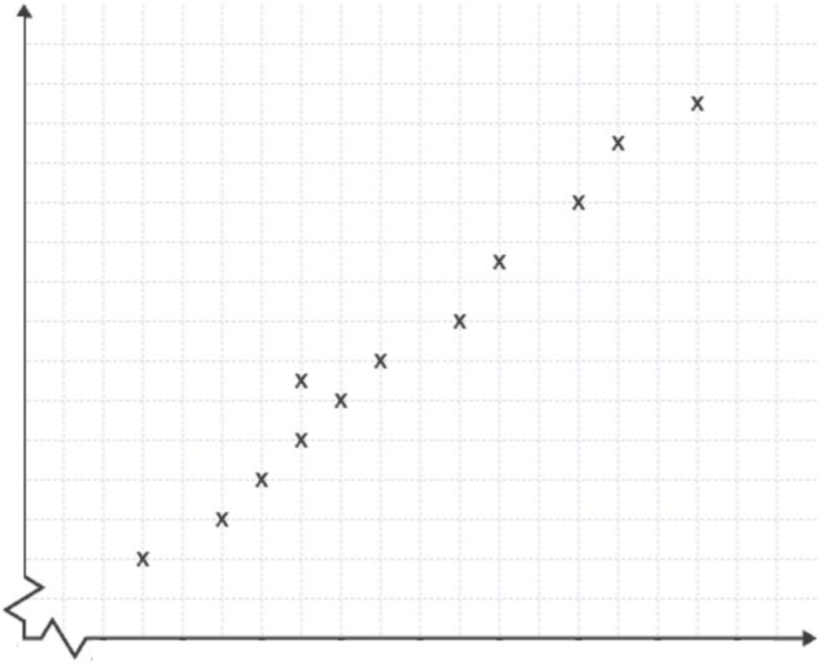

Figure 2.1 : Positive Correlation [40]

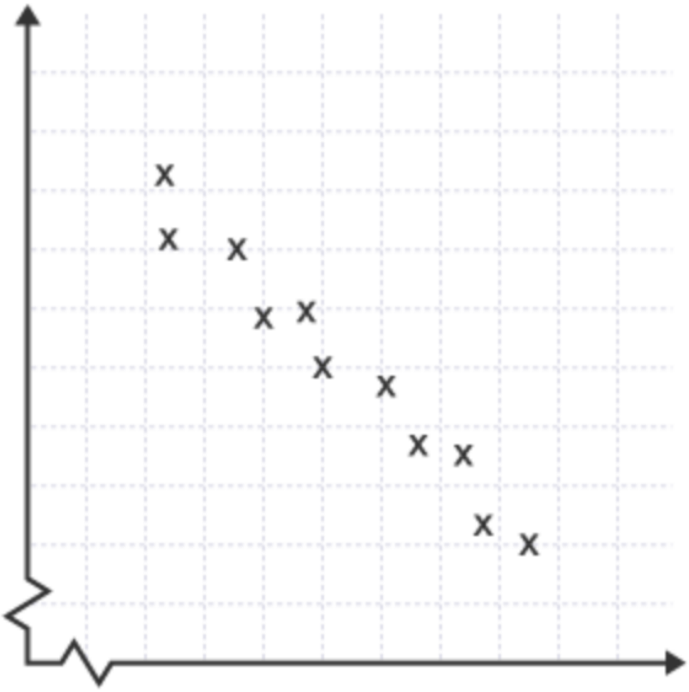

Figure 2.2 : Negative Correlation [40]

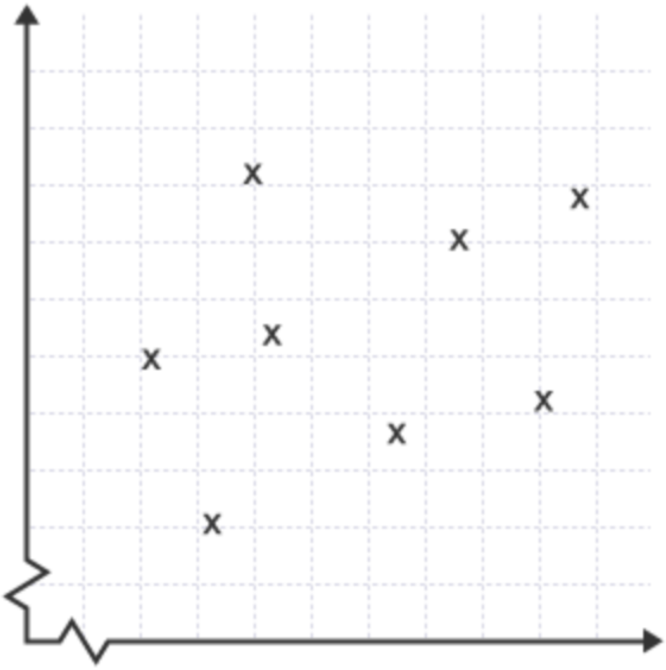

Figure 2.3 : No Correlation [40]

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

2.5.1 Pearson Correlation

In Pearson correlation, the features being compared gets assigned a value between -1 and 1 [41]. 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 express that if one feature is present than the other feature will unquestionably be present as well [41]. 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 [41]. 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 which behaves differently [41]. 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 [41].

The equation for Pearson correlation is :

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

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.

2.5.2 Spearman Correlation

The Spearman correlation is a method that measures the strength and direction of the relationship between two features in a dataset [42]. 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 [42]. However, the relationship between the features does not have to be linear [42]. The correlation values in Spearman correlation follows 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 [42]. If the correlation value is 0, then a change in a variable would not affect the other variable [42]. On the other hand, if the correlation value is 1, then as one variable increases, the other variable would increase as well [42].

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

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \qquad (2.12)$$

where ρ is the spearman correlation, d_i is the difference between the features, and *n* is the total number of values [43]. Equation 2.12 can only be used is there are no duplicates in the dataset. If duplicates exists in the dataset, then the second equation will be used. The second equation is :

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 (y_i - \bar{y})^2}} \qquad (2.13)$$

where ρ is the spearman correlation, *x* is the value of feature x, x̄ is the mean of feature x, y is the value of feature y and ȳ is the mean of feature y [43].

2.6 Feature Selection

Feature Selection is the process of cutting down the input variables which will be fed into the models [37]. This is useful as it gets rids of the noise in the dataset so that the model can focus on the useful information [37]. 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 [38]. 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 it could reduce the model's accuracy [38]. It would lead to lack of variation in the data or even result in data leakage which would make the model perform unrealistically good [38].

2.7 Models

2.8 Software Development Life Cycle

Software Development Life Cycle (SDLC) is the process which is made up of steps that a certain software can follow in order to develop in a proper manner [30]. This would make it more likely for the project to be accomplished on time whilst ensuring the quality of the product is suitable for the user [29]. The activities for a specific SDLC can be labelled as [31] :

1) *Understanding the case*
2) *Deciding solution scheme*
3) *Coding based on the solution decided*
4) *Testing*

However, these activities are quite broad, therefore it can be broken down even further to better illustrate SDLC process [31]. The phases of SDLC are :

1) *Requirements Analysis :* This is the first phase of SDLC. In this phase, the business requirements of the project are gathered. The project managers and stakeholders will discuss to define the requirements of the software. These requirements could include answering questions such as "who will use the software" or "how will the system be used" [31]. After the discussion, a Software Requirement Specification (SRS) document will be created which will contain the results of the discussion [31].
2) *Design :* The main objective of this phase is to turn the requirements specified in the first phase into an architecture [31]. In this phase, the hardware and

system requirements are specified so that the architecture of the software can be defined [31]. Additionally, this phase is where testers are required to define what needs to be tested and how it should be tested [31].

3) *Development :* In this phase, the results of the design phase is converted into a system that meets the user requirements. This phase is also referred to as the coding phase. All the developers and engineers play an active role in this phase, and they are required to follow the requirement guidelines defined beforehand [32]. It is the longest yet most crucial phase in the entire SDLC process. Additionally, the process of the development phase will be recorded in a document entitled Source Code Document (SCD) [32].
4) *Testing :* The next phase is the testing phase where the software developed in the previous phase will be tested. There is usually a specific team whose purpose is solely for testing the software, their job is to conduct series of tests on the software [32]. The testing team will document any errors they encounter and send this report to the development team so that the developers can attempt to remove the errors [31]. The testing phase is one of the most essential phase as it decides whether the software is eligible to be released to the users [32].
5) *Deployment and Maintenance :* In this phase, the software has passed the testing phase and is bug-free, therefore it is now deployed and useable by the client [32]. Additionally, in this phase, there are possibilities that the software needs to be updated due to technological advancements. Therefore, the developers need to maintain the software to ensure that its performance will not decline [30].

Over the years, the SDLC model has been adapted into different kinds of models. These models include the *Waterfall Model, V-shaped Model, Incremental Model, Agile*

Methodology and many more [31]. The agile methodology in particular is known for constant iterations for software testing and development [30]. In this methodology, it is common for the development phase and the testing phase to occur concurrently [31]. The Agile methodology contains twelve core principles, which are [33] :

1) *Customer satisfaction*
2) *Adaptive to changing requirements*
3) *Regular software delivery, the faster the better*
4) *Productive collaboration between developers and stakeholders*
5) *Support developers by supplying ideal work environment and believe that they will accomplish the project*
6) *Direct face to face communication for team discussion*
7) *Assess progress by checking on working software*
8) *Encourage maintainable development*
9) *Constant focus on technical quality and design*
10) *Simplicity is vital*
11) *Working units that can organize themselves will provide the ideal output (design, software architecture, requirements)*
12) *Occasional reflection so that the team can improve*

The agile methodology also consists of framework entitled SCRUM. SCRUM is an agile development methodology that is based on an iterative as well as an incremental process [31]. One of the main features of SCRUM is that it focuses more on feedback, revisions, and frequent customer engagement rather than documenting procedures and predicting a plan of action for accomplishing the project [31]. In SCRUM, there are three main roles which are Product Owner (PO), Scrum Master (SM), and Scrum Team

(ST). There is a lack of guidelines or descriptions for how the project should be accomplished in SCRUM, most of the decision-making is left to the team doing the project as the team knows best [31]. There are three constants in SCRUM which are Product Backlog, Sprint Backlog, and Sprint Goal [31]. Product Backlog is the list of things that need to be done by the PO, Sprint Backlog is the list of things selected by the ST that needs to be done in the current sprint cycle, whereas Sprint Goal is the endgame of the current sprint [31]. SCRUM Methodology is beneficial for complicated projects and this methodology greatly helps the project progress in an efficient manner [31].

2.9 RestAPI

CHAPTER 3

SYSTEM DESIGN

CHAPTER 4

SOLUTION DESIGN

4.1 Data Preprocessing

REFERENCES

[1] M. S. Vassiliou, *Historical dictionary of the petroleum industry*. Rowman & Littlefield, 2018.

[2] World Petroleum 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.

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

[4]R. Ranggasari, "Oil and gas reserves potential in eastern Indonesia reaches 9.8bn barrels," *Tempo*, 07-Dec-2021. [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] I. Investments, "Crude oil," *Indonesia*. [Online]. Available: https://www.indonesia-investments.com/business/commodities/crude-oil/item267.

[6] "Lynx Information Systems," About Lynx Information Systems. [Online]. Available: http://www.lynxinfo.co.uk/about.html.

[7] Lynx information systems. License Pricing - Lynx Information Systems. (n.d.). Retrieved from http://www.lynxinfo.co.uk/download-pricing.html.

[8] "Intviewer - Fast Geoscience Visualization, Analysis & QC," INT, 02-Aug-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. Geoscience Analysis and QC, Simplied. (n.d.). Retrieved from https://www.int.com/products/intviewer/.

[10] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electronic Markets*, vol. 31, no. 3, pp. 685–695, 2021.

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

[12] I. Sydorenko, "What is a dataset in machine learning," *High quality data annotation for Machine Learning*, 05-Apr-2021. [Online]. Available: https://labelyourdata.com/articles/what-is-dataset-in-machine-learning.

[13] Y. Gong, W. Xu, P. Cunningham, M. Cord, and S. Delany, "Supervised Learning," in *Machine learning techniques for multimedia*, Springer US, 2008, pp. 21–49.

[14] IBM Cloud Education, "What is supervised learning?," *IBM*, 2020. [Online]. Available: https://www.ibm.com/cloud/learn/supervised-learning.

[15] A. R. van Loon, A. M. Jeevan, A. V. Murugesan, A. A. Lakshminarayanan, and A. A. Strife, "Machine learning explained: Understanding supervised, unsupervised, and reinforcement learning,"

Big Data Made Simple, 04-Feb-2019. [Online]. Available: https://bigdata-madesimple.com/machine-learning-explained-understanding-supervised-unsupervised-and-reinforcement-learning/.

[16] K. El Bouchefry and R. S. de Souza, "Learning in big data: Introduction to machine learning," *Knowledge Discovery in Big Data from Astronomy and Earth Observation*, pp. 225–249, 2020.

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

[18] G. Sharabok, "What is reinforcement learning?," *Medium*, 17-Sep-2020. [Online]. Available: https://towardsdatascience.com/what-is-reinforcement-learning-99f9615918e3. [Accessed: 27-Feb-2022].

[19] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. Cambridge, MA: The MIT Press, 2020.

[20] Maitland, G. C. (2000). Oil and gas production. *Current Opinion in Colloid & Interface Science*, *5*(5–6), 301–311.

[21] El-Banbi, A., Alzahabi, A., & El-Maraghi, A. (2018). Black Oils. In *PVT Property Correlations* (pp. 147–182). Elsevier. https://doi.org/10.1016/B978-0-12-812572-4.00007-2

[22] El-Banbi, A., Alzahabi, A., & El-Maraghi, A. (2018). Dry Gases. In *PVT Property Correlations* (pp. 29–63). Elsevier. https://doi.org/10.1016/B978-0-12-812572-4.00003-5

[23] Mokhatab, S., Poe, W. A., & Mak, J. Y. (2019). Natural Gas Fundamentals. In *Handbook of Natural Gas Transmission and Processing* (pp. 1–35). Elsevier. https://doi.org/10.1016/B978-0-12-815817-3.00001-0

[24] Ahmed, T. (2010). Reservoir-Fluid Properties. In *Reservoir Engineering Handbook* (pp. 29–135). Elsevier. https://doi.org/10.1016/B978-1-85617-803-7.50010-9

[25] Rosenthal, S. (2017). Data Imputation. In *The International Encyclopedia of Communication Research Methods* (pp. 1–12). Wiley. https://doi.org/10.1002/9781118901731.iecrm0058

[26] Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, *64*(5), 402. https://doi.org/10.4097/kjae.2013.64.5.402

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

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

[29] A. Mishra and D. Dubey, "A Comparative Study of Different Software Development Life Cycle Models in Different Scenarios," *International Journal of Advance Research in Computer Science and Management Studies*, vol. 1, no. 5, pp. 1–6, 2013.

[30] G. Gurung, R. Shah, and D. P. Jaiswal, "Software development life cycle models-A comparative study," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, pp. 30–37, 2020.

[31] J. Shah, "A Comparative Study of Software Development Life Cycle Models," *Institute of Technology, Nirma University*.

[32] P. Pedamkar, "What is SDLC: Different phases and models of SDLC," *EDUCBA*, 31-Jan-2022. [Online]. Available: https://www.educba.com/what-is-sdlc/. [Accessed: 05-Mar-2022].

[33] Agile Manifesto, *Principles behind the Agile Manifesto*. [Online]. Available: https://agilemanifesto.org/principles.html. [Accessed: 05-Mar-2022].

[34] Fernández, Á., Bella, J., & Dorronsoro, J. R. (2022). Supervised outlier detection for classification and regression. *Neurocomputing*, *486*, 77–92. https://doi.org/10.1016/j.neucom.2022.02.047

[35] Ben-Gal, I. (n.d.). Outlier Detection. In *Data Mining and Knowledge Discovery Handbook* (pp. 131–146). Springer-Verlag. https://doi.org/10.1007/0-387-25465-X_7

[36] Salgado, C. M., Azevedo, C., Proença, H., & Vieira, S. M. (2016). Noise Versus Outliers. In *Secondary Analysis of Electronic Health Records* (pp. 163–183). Springer International Publishing. https://doi.org/10.1007/978-3-319-43742-2_14

[37] V. Kumar and S. Minz, "Feature Selection: A literature Review", *Smart Computing Review*, vol. 4, pp. 1 - 19, 2014. Available: https://faculty.cc.gatech.edu/~hic/CS7616/Papers/Kumar-Minz-2014.pdf. [Accessed 6 March 2022].

[38] K. Menon, “Feature selection in machine learning [2021 edition] - simplilearn,” Simplilearn.com, 16-Sep-2021. [Online]. Available: https://www.simplilearn.com/tutorials/machine-learning-tutorial/feature-selection-in-machine-learning#what_is_feature_selection.

[39] F. Malik, “Understanding value of correlations in data science projects,” Medium, 10-Jun-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%20with%20each%20other. [Accessed: 06-Mar-2022].

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

[41] Nettleton, D. (2014). Selection of Variables and Factor Derivation. In *Commercial Data Mining* (pp. 79–104). Elsevier. https://doi.org/10.1016/B978-0-12-416602-8.00006-6

[42] “Spearman correlation coefficient: Definition, formula and calculation with example,” *QuestionPro*, 15-Jan-2020. [Online]. Available: https://www.questionpro.com/blog/spearmans-rank-coefficient-of-correlation/.

[43] Swapnilbobe, “Spearman's correlation,” *Medium*, 13-Apr-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.