AN ARTIFICIAL INTELLIGENCE APPROACH FOR PREDICTION OF WATER PRODUCTION IN OIL AND GAS WELLS.

WRITTEN BY

OKORO BLESSING C

20151013363

DEPARTMENT OF PETROLEUM ENGINEERING

SCHOOL OF ENGINEERING AND ENGINEERING TECHNOLOGY

FEDERAL UNIVERSITY OF TECHNOLOGY OWERRI IMO STATE

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE AWARD OF A BACHERLOR OF ENGINEERING (B.ENG) DEGREE IN PETROLEUM ENGINEERING

MARCH, 2021

CERTIFICATION

This is to certify that this project work on **"AN ARTIFICIAL INTELLIGENT APPROACH FOR PREDICTION OF WATER PRODUCTION. THE VOLVE FIELD AS CASE STUDY"** was written by **Okoro Blessing Chimezie,** with registration number **20151013363** under the supervision of Engr. Dr. N.C IZUWA. It is my original work except for those areas referenced and that no part of this work has been presented elsewhere for an award or any degree or certificate.

……. ……

ENGR. DR. N.C IZUWA Date

(*Project Supervisor*)

……. …….

ENGR. DR. K.C IGWILO Date

(*Head of Department*)

……. …….

……. Date

(*External Examiner*)

DEDICATION

This Project work is dedicated to God Almighty for his immense grace, provision, and enablement throughout my undergraduate studies. Also, this work is dedicated to my loving parents Mr. and Mrs. Chima Okoro, who consistently and undoubtedly supported me throughout my undergraduate studies. They provided the much-needed encouragement, financial support to see to the completion of my undergraduate studies here in this great citadel of learning.

ACKNOWLEDGEMENT

I, in a special way acknowledge my project supervisor, Engr. Dr. N.C IZUWA for his unparalleled support, encouragement, mentorship which in turn inspired the success of this project work. My profound Gratitude goes to my H.O.D, Engr. Dr. K.C IGWILO for his contributions to this work with the jurisdiction as the Head of Department Petroleum Engineering. My Gratitude also goes to my course adviser Engr. Dr. NWACHUKWU ANGELA and all the staffs of Petroleum Engineering Department who guided and supported me in my course of study.

My appreciation goes to Mr. YAYA ADEBISI and all the staffs of Wireline and Well Logging Department, B.G Technical for their knowledgeable contributions in Production Engineering and petroleum Engineering to in General which contributed to the success of my project work. Special thanks go to the various Oil Exploration and Production company who provided the production data used in this project work.

I also sincerely appreciate my friend, Obasi Emmanuel who enlightened me on Artificial Intelligence, sacrificed his time and made sure every single objective of this work was accomplished.

I thank the rest of my friends and well-wishers who have in one way or the other contributed to my welfare. I pray that the Almighty God will reward each and every one of you abundantly.

ABSTRACT

Production of water is unavoidable in the lifespan of a well because most fields experience decline in oil production with increasing water production with time especially fields with strong water drive. This is usually controlled by understanding the behavior of the reservoir to deduce if water production is in excess or will be in excess in future. Therefore, searching for a good model for determination of water production is usually needed.

This project presents a Machine Learning Model for predicting water production in oil and gas fields. In developing the model in this work, 7504 production data set was gotten from the Volve field in the Norwegian Sea. had the coefficient of determination (R^2) of 0.73298 and Mean Square Error of 0.0406 which indicates that the model fits into the data set and it is reliable.

The model was optimized to avoid overfitting by reducing the variance so that it will not perform poorly when used on a different production set. Model validation was carried out using the cross validation by splitting the data into training and testing set. The coefficient of determination (R^2) was improved to 0.73652 and Mean Square error of 0.0407, which is good but not so high, the data set was used on other models for better predictive capacity and the Random Forest Model gave a better coefficient of determination (R^2) of 0.92 and RMSE of 0.11, this indicates a good reliability of the proposed model and it can be used for water prediction in oil and gas reservoirs.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE: INTRODUCTION

1.1 BACKGROUND OF STUDY

Reservoir management faces remarkable challenges in optimizing profitability while satisfying a number of constraints (physical, financial, geopolitical and human). To optimize profitability, engineers have traditionally used mathematical models, field data and knowledge to make decisions about the best operating scenario. (Luigi .S. et al, 2004). Water production is expected to increase with the life of a reservoir, this produced water is present with the hydrocarbons in the reservoir and it's brought to the surface alongside with the hydrocarbons, it could either be from aquifer or from water flooding process. (Echufu-Agbo.O., 2010). At a certain point in the well`s life, actual performance may not satisfy expected values, nor will it meet physical and economic constraints, since production of oil and gas from reservoir is usually accompanied by water/brine (produced water) as the well is being depleted. At this stage, a remediation action or workover would be performed if analysis predicts additional and significant economic value creation by so doing. Water production affects optimization of hydrocarbons since the brine reduces the amount of oil and gas that will be produce during that period. Most times, water production is done at the cost of gas recovery and in serious cases, the water influx becomes too much causing the gas production to choke off. Some traditional method used to reduce the effect of water influx in oil and gas well in order not to produce mixture of gas and water include installation of downhole separators, conventional treatment such as recompletion or cement squeeze etc. (K. J. Waro et al, 2000). Water production can be controlled effectively by understanding the water production mechanism. Water production is one of the factors that affect production optimization in wells thereby reducing the economic lifetime of the well. The process of producing, handling and disposal of water is not usually cost effective, and they are also time consuming when

compared to generating machine learning models for water production in the well. As optimization algorithms and reservoir simulation techniques continue to develop and computing power continues to increase, upstream oil and gas facilities previously thought not to be candidates for advanced control or optimization are being given new consideration.

The important feature of machine learning is its ability to learn from previous experiences making it easy to control a process. And machine learning algorithms may be as important to the Oil and Gas industry as the Internet has become to the society because by analyzing historical data, the algorithm can know the relationship of different parameters and their effect to production. More accurate analyses may lead to more confident decision making. And better decisions can mean greater production efficiencies, cost reduction, and reduced risks. Upstream is no stranger to Artificial Intelligence. Oil and Gas companies use thousands of sensors installed in sub surface wells and surface facilities to provide continuous data-collecting, real-time monitoring, and environmental conditions. (Abdelkader .B. et al, 2014). To support the real-time decision making, Oil and Gas companies need tools that integrate and synthesize diverse data sources into a unified whole. Having a machine learning algorithm that has the ability to predict water production rate based on adjusting the control parameters is an important tool. This machine learning model gives the water production rate landscape at its peak and valley which is high and low production, this enables the variables to be controlled and to also know how much to adjust them for optimum hydrocarbon production. Machine learning is of great interest to production and operation work, it can predict future performance based on historical results, or to identify sub-par production zones, it can shift assets to more productive areas, improve oil recovery rate by integrating and analyzing seismic, drilling and production data to provide self-service business intelligence to reservoir engineers.

This project proposes an approach using Machine learning model to predict water production by analyzing different data sets at different points in the well to determine the best model for the prediction. Machine learning algorithms can learn the key information patterns within multidimensional information domain. So, engineering efforts can be reduced.

Finally, the performance of this model will be compared with other models.

1.2 STATEMENT OF PROBLEM

In petroleum fields, one of the most important aspect of production is finding a way to optimize production in a cost- effective manner and to save time. Production and handling of produced water usually define the hydrocarbon reserve and the economic lifetime of the wells in the field therefore predicting and evaluating this water production is important. The work of an operator is to devise optimal operating strategies to achieve these goals. This project will develop a prediction model that will ease and automate the decision making of field operators.

1.3 AIM OF STUDY

To predict water production using an artificial intelligent approach. The volve field as CASE STUDY.

1.4 OBJECTIVES OF STUDY

In order to achieve the above set goals, the following objectives will be considered:

1. Studying the water production problems in the Niger Delta and other regions.
2. Developing a model that effectively predict water production.
3. Optimizing the model for better performance and reliability.

4. Prediction of water production using different Machine Learning models.

1.5 SIGNIFICANCE OF STUDY

During the life- time of a producing well, production may gradually start to decline due to increase in water cut as the well is being depleted and other factors that affect production. This often result to the use of different well intervention method which is usually expensive and takes time. With Machine learning algorithm, one can be able to discover new information and identify patterns that will enable improvement of production by developing a model for the prediction of water production in a well.

1.6 SCOPE OF STUDY

The focus on this project was to apply a Machine learning approach for prediction of water production. Different machine learning algorithms will be deployed on production data from Volve Field. The best is selected and the developed model will be consequently fine-tuned through hyper parameterization of controlled variables. These variables include; flowrate, gauge sizes, oil, water and gas production, GOR, BHFP, choke sizes, wellhead temperature and pressure, tubing sizes.

CHAPTER TWO: LITERATURE REVIEW

2.1 WATER PRODUCTION IN OIL AND GAS WELLS

Petroleum production comprises of two separate but intimately connected general systems which are (1) the reservoir, which is a porous and impermeable storage with unique flow characteristics, (2) and the artificial structures, which is made up of the bottomhole, the well and wellhead assemblies, as well as the surface gathering, separation, and storage facilities.

Production engineering is a part of petroleum engineering that ensures that optimum well design and equipment are selected in order to maximize production in a cost-effective manner.

During the whole life span of oil and gas wells, oil gas and water are produced with water being the biggest by product during the production and this is usually unavoidable. (Tyler F Hussey et al, 2017). Unwanted fluids such as produced water are controlled by understanding the behavior of the reservoir and can be deduced if the recent oil, gas or water production are in excessive or will be excessive in future. (Andrea .Q. et al, 2020). Most oil fields under water drive or natural aquifer produces water along with oil in due time, the current world daily water production in oil wells is about 3 BWPD per barrel of oil though some wells produce higher quantity. Good water production is observed at a rate less than the water/oil economic limit, but bad water production is seen in wells that have no sufficient oil for handling the cost of water disposal. If a well is producing with 80% water cut, the cost of lifting the water will be less than the cost of handling though some wells with 90% water cut may produce enough hydrocarbon to be economically productive.

Water cut increases as the reservoir is being depleted during primary production. This is usually seen in usually seen in reservoirs that have natural water drive where the aquifer is in pressure communication with the reservoir, hence as hydrocarbon is being produced the water from the aquifer fills the space left

behind increasing the water saturation of the rock. As the pressure in the reservoir declines over time, the quantity of water produced will rise until the cost of handling becomes more than the hydrocarbon being produced. (Bailey .B. et al, 2000).

2.2 CAUSES OF WATER PRODUCTION IN OIL AND GAS WELLS

Successful water control is usually predicted by knowing the source and position of the intruding water and it can be seen in production logs, production history and direct measurement. Some causes of water production in oil and gas wells are;

- Coning: Water coning in production well is as a result of pressure gradients created by the production of fluids around the wellbore. It also occurs when the oil water contact is near perforations having high vertical permeability. (Onwukwe S.I, 2015).
- Edge water drive from poor areal sweep, this problem is usually seen in reservoirs having different areal permeability.
- Fractures or faults between the injector and the producer wells or production of water from fractures that cuts through a deeper water zone.
- Gravity Segregated Layer: This occurs when water in thick high permeability reservoirs moves downward from the aquifer into the permeable formation and sweeps only the lower part of the reservoir.
- Flows channeling from poor primary cementing that does not completely isolate the water bearing zone from the pay zone.
- Leaks from casing, packers and tubing.
- Moving oil-water contact: The transition zone is the contact between oil and water, instability in the zone can cause water encroaching into the reservoir leading to water production.

2.3 EFFECT OF PRODUCED WATER

The water that moves up to the surface with the hydrocarbon is called produced water, it is sometimes called brine or formation water. Produced water contain some chemical characteristics of the formation and hydrocarbon since it has been in contact with them. Produced water contains; chemical additives from drilling, oil due to contact with hydrocarbon formation, salt from saline formation, naturally occurring radioactive materials and other inorganic/organic compounds. (Earl Hangstrom, 2016). These constituent of produced water leads to;

- Corrosion in the wellbore and pipes.
- Gas hydrate formation which could plug the pipelines.
- Hydrocarbon solid deposition.
- Costly disposal of the water or reinjection.

Due to these reasons excessive water production in oil and gas wells is not beneficial.

2.5 WATER PRODUCTION ANALYSIS

It is important to understand the fundamentals of fluid flow across the production system to be able to predict the performance of individual wells and the water production in wells and reservoirs. The production system carries reservoir fluids from the reservoir to the surface. The basic components of the production system includes the reservoir, well-bore, surface equipment etc (Larry .W.L et al, 2007).

2.5.1 Reservoir

Reservoir rocks are rocks (usually sedimentary) that are capable of storing fluids inside their pores, so that the fluids either water, oil, or gas can be accumulated. Reservoir is one of the elements of petroleum system that can accumulate hydrocarbons, the rock must have good porosity and permeability to accumulate

and drain oil in economical quantities. The reservoir is made up of different interconnected geological flow units. The shape and position of the reservoir to the aquifer can have a great effect on the water production in the well, reservoirs having high permeability streaks, fractured formation or coning in relation with a limited aquifer may have early water production or encroachment from edge wells.

2.5.2 The Well and Well Completion

Hydrocarbon wells are drilled with the larger bore hole sections at the top of the well. Each section is cased to the surface and cemented until last section of the casing which is the production casing is cemented, it is usually run with a production tubing to isolate the annulus between the outside of the tubing and the inside of the casing. When poor primary cementing is done and it does not completely isolate the water bearing zone or the surface water, water flows from this zone into the production tubing. Excessive water production of water occurs in structurally low wells especially reservoirs with water drive where water is the displacing mechanism. (Micheal Economides et al, 2013).

2.5.3 Formation permeability around the well

The value of the formation permeability in the reservoir is different from the permeability close to the wellbore in most producing wells, it might be as a result of formation damage/skin damage from well completion method, and it leads to alteration of the permeability around the well which may lead to excess water production since the higher permeability layer controls the fluid flow.

2.5.4 Water Flooding Operation

Waterflooding is a method of secondary recovery where water is injected into the reservoir to move the hydrocarbon into the producing, the process of adding water and moving the hydrocarbons increases the reservoir pressure and causes increase in hydrocarbon production till it reaches it economic limit or water

break through occurs, when this happens there will be excess water production and less hydrocarbon production. (SPRI Team, 2018).

2.5.5 Wellhead pressure and choke performance

The choke is usually placed on the wellhead to help maintain the wellhead pressure and hence control the bottomhole flowing pressure and the production rate. They are used for limiting production rates and control flowrates to prevent water or gas coning. The choke performance has a relationship with water production in oil and gas wells, increase in the size of the choke increases water production. (Mohammed S. Al-Jawaad *et al*, 2006).

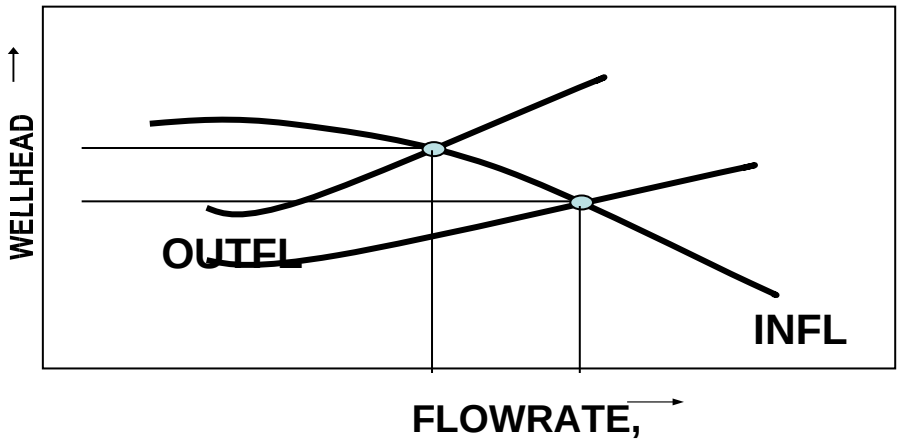

Fig 2.1: Effect of wellhead pressure on production rate

2.6 EXISTING MODELS FOR WATER PRODUCTION

Many existing models have been developed for the prediction of water cut and water production. Some of these existing models will be presented and discussed in this work.

Warren's Model

In 1998 warren developed on model for prediction of water production. The model was developed by making a linear plot of cumulative fractional flow against the cumulative oil production. The model is expressed as.

$$\frac{Np}{Wp + Np} \propto Np$$

Lawal's Model

In 2007, based on the Arp's production model, Lawal et al developed a model for the prediction of water cut. The model assumes that the water production rate declines exponentially therefore fitting an exponential curve to the plot of water production versus cumulative production rate and can be used for prediction. The model is expressed as.

$$f_{wt} = 1 - f_{oi}\, e^{-at}$$

Where f_{wt} is the water cut at a production time t, f_{oi} is the initial oil cut and a is a constant.

Dhafer Al Shehri's Model

In 2019 Dhafer Al Shehri et al modified Purvis and Lawal's Model after discovering from plots that the trends deviates from linear and exponential decline at the later part of production data and get distorted without a visible trend. The model is expressed below.

$$f_{wm} = \frac{1}{1 + a * Q_O * e^{\left(\frac{Q_o}{BHP}\right)}}$$

Where f_{wm} is the modelled water cut, Q_o is the oil production rate, BHP is the bottom hole pressure and the constant a is determined from the production history of the data, it allows for flexibility of achieving an exponential, hyperbolic or harmonic increase of water by changing the value of the constant to match the required trend.

2.7 ARTIFICIAL INTELLIGENCE (AI)

Artificial Intelligence is the method of combining human intelligence and computing power in producing relevant and intelligent solution to complex problem. Artificial Intelligence uses algorithms to understand problems and provide solutions. It provides a new way of thinking and tackling a problem and also helps to utilize complicated data sets and identify parameters that drive production which is one of the challenges in petroleum industry.

In oil and gas industry, Artificial Intelligence systems are trained using large volume of raw production data, this helps to improve production data for generating analytics by enabling automatic pattern recognition and classification. They can be used to estimate well production performance based on the geologic properties and completion design at a well location. The combination of the experiences from drillers, engineers and geologist can provide a strategic plan for developing wells and optimization to maximize revenue and recovery.

Production decline curves and history matching were the traditional methods used to estimate the estimated ultimate recovery, this depends heavily on the experiences of the scientists and engineers, but Artificial Intelligence helps to identify parameter relationships by extracting production data. Most times, the

information provided by Artificial Intelligence proves the assumptions of reservoir engineers and geologist.

Prediction models and estimations can be created using Artificial Intelligence systems, through this oil and gas industries can analyze production procedures which will help identify areas of inefficiency and develop different method of maximizing production. (Naveen Joshi, 2019).

2.8 MACHINE LEARNING

The type of Artificial Intelligence accesses incoming data for anomalies which could cause problem ahead in monitored equipment is called Machine Learning. Machine learning is a type of AI where computers are programmed algorithmically to learn from new data and experience to improve at a task when evaluated using a metric, instead of being told about the outcome. It is the ability of an algorithm to learn from data and improve in accuracy over a period. It is used to find data patterns for making decisions and prediction, as it processes more data the algorithm becomes more accurate. It is very useful when the data sets are very large to manually analyze (Matthew Taylor, 2021). This is illustrated with the simple example of a problem in which the aim is to identify handwritten digits, and the metric is the percentage of digits classified using the database of human-labelled images of handwritten digits serving as experience.

The “learning” part uses big amounts for the computer to be able to understand underlying patterns.

When building a machine learning model, the data needs to be processed in a way that it will be adequate for the computational stability of the machine learning algorithm. A necessary step that should be taken in data preprocessing involves rescaling the input features and output variable (for regression problems) to make their ranges consistent before feeding them as an input to the machine learning algorithm.

Advancement in computational power and highly efficient machines, has brought a significant improvement in machine learning algorithms efficiency over the past few years. (Y.N Pandey, 2020).

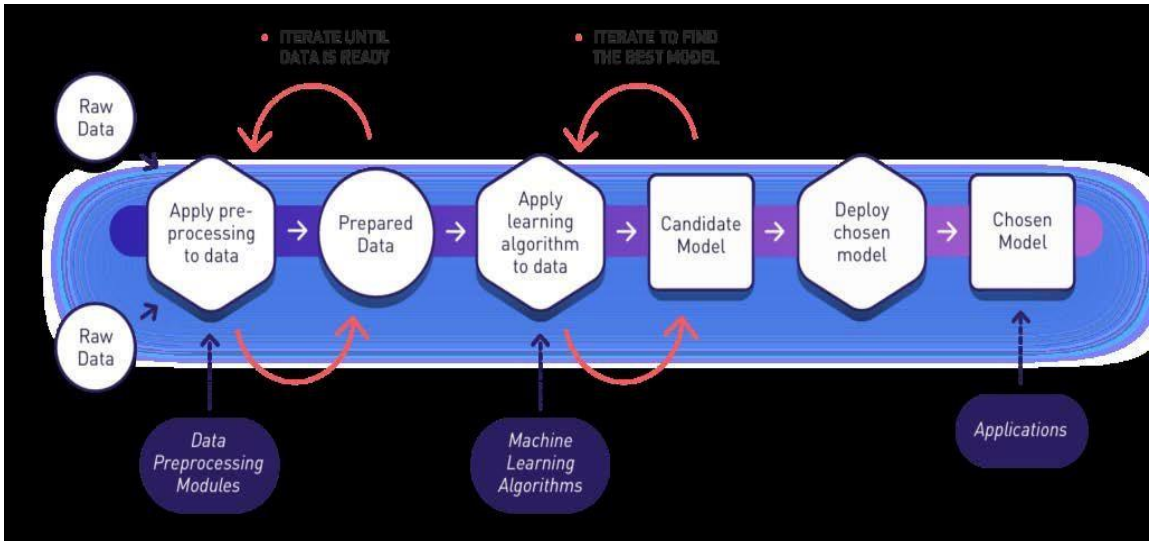

Fig 2.2: A typical workflow for a machine learning project (Shengnan Chen, 2019).

1. Integrating raw data sets from different data sources into a target data base.
2. Cleaning target data by removing noise, duplicated, and inconsistent data.
3. Transforming data into appropriate forms by dimension reduction or normalization.
4. Applying various Machine Learning algorithm to data sets and selecting the best candidate model.
5. Using the Machine Learning model in making decisions.

2.8.1 Machine Learning Categories

Machine learning algorithms can be divided into three broad categories;

- **Supervised Learning**

  In supervised learning, various tags, markers and description are placed on the data and the response variable is known, the mapping between the input features and output variables is provided by a machine learning algorithm. The regression and classification problems are the common categories of supervised learning which are dictated by the type of output variable. But if the output variable is continuous, it falls into the regression category.

  While in classification problems, the output variable usually contains multiple classes or labels, In supervised learning, the model training process continues with making improvements by evaluating errors until a certain level of accuracy is achieved.

- **Unsupervised Learning**

  In unsupervised learning, there is no present information to the data attached to the data added to the system and the relationships are gotten from the data provided to the algorithm. Some of the algorithms in this unsupervised learning can show hidden structures and relationships between input features. Some examples of unsupervised learning include dimensionality reduction algorithms, clustering, and associative rule learning.

✦ Reinforcement Learning

In reinforcement learning, the algorithm is made in such a way that for every sequence of decision made by the algorithm, there will be a reward or penalty associated with it, this helps the algorithm to learn the decisions it has to make to achieve a set objective. The reinforcement learning is usually seen as semi-supervised but the algorithm uses a trial and error method in finding solutions in an uncertain and complex environment by either being penalized or rewarded for their actions. (Y.N Pandey, 2020).

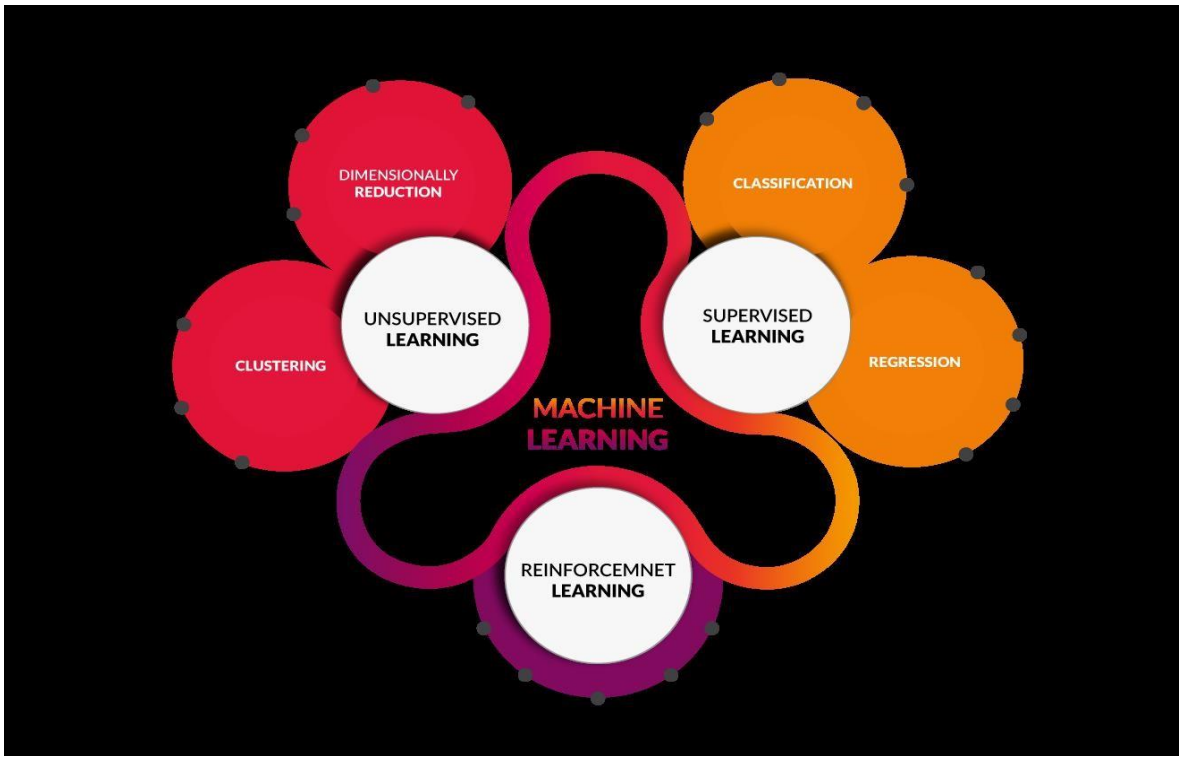

Fig 2.3: Machine learning categories (Shengnan Chen, 2019).

2.8 REVIEW OF EXISTING LITERATURES

T. Cross et al, 2020 used data from the North Dakota Industrial Commission to train a model for predicting water production at a 30-day increments after initial production out of 720 days. It gave them a highly accurate way of predicting water production and how to analyze the choice of the impact of water cut by operators.

Andrea Quintero et al, 2020 optimized the design of a conformance treatment with the use of reservoir simulator and also evaluated the efficiency of the conformance solution. After the conformance treatment, the water production was reduced by almost 65%.

In 2019, Dalal Al-Subaiei et al worked on the production data in the North Kuwait Imtegrated Digital oil field by preparing a well model for every well in NK asset considering the PVT parameters, completions and surface co-ordinate. They designed a smart work flow and displaced the oil gain opportunities from the workflow run.

Azad Almasov et al (2020) used a machine learning model which could be obtained with either least squares support vector regression (LS-SVR) or Guassian process regression (GPR) to accurately approximate the NPV during optimization. The two ML base methods prove to be quite efficient in production optimization than using a stochastic gradient computed for a high fidelity compositional simulator directly in a gradient ascent algorithm.

Guofan Luo et al (2018) integrated the geological properties and well completion strategies for 2061 horizontal wells in Bakken field. They made use of data driven technique to gain insight about the production performance of the well,

they made a predictive model with deep learning using the important parameters providing a good relationship between the input parameters and the normalized production.

In 2019 Christoph Kandziora and Siemens AG applied artificial intelligence to optimize oil and gas production by incorporating AI into a machine learning base predictive maintenance model. They used it to identify previously unknown ESP operating anomalies and other multiple kind of anomalies, therefore providing valuable decision support.

In 2019, Abdulaziz Al-Qasim et al applied nodal analysis to eleven different existing wells to optimize their production flow rate. They achieved the optimization by changing the tubing and flowline sizes, minimizing the skin factor and controlling the water cut. They ran different parametric scenarios on different chokes and pipeline for each of the eleven wells.

C.M.F Galas, 2003 proposed an approach to oil reservoirs where he emphasized on matching the cumulative produced fluid as a function of time. He varied the parameters that control each of the mechanism to produce the history matching making sure that the changes were consistent with the uncertainty in the parameters.

In 2019, Dhafer Al Shehri et al analyzed the post breakthrough performance of water with the use of empirical models. They developed different model to capture the trend of the redefined water cut plotted against the flow rate, they validated the model with an existing field and it was proven to be a good tool for water production estimation.

CHAPTER THREE: MATERIALS AND METHODS

3.1 MATERIALS

3.1.1 TOOLS USED IN DEVELOPING THE MACHINE LEARNING MODEL

In developing the machine learning model, we used python and jupyter notebook as the integrated development environment which enabled the python to communicate directly with the code. Jupyter notebooks basically provides an interactive means for developing python-based data science applications. It helps to describe the analysis process step by step and can be used for data transformation and cleaning, data visualization, machine learning and much more.

3.1.2 DATA AQUISITON

In developing a machine learning model, it is necessary to choose the right content and sources of data for the model. The volve field that covers a wide range of production data was stimulated in this project. The volve field is located in the Norwegian North Sea, the field consists of seven distinct wells. To develop the model, 15,634 production data points were gotten from this field with parameters which include tubing size, choke size, well head temperature and pressure, oil production rate, gas production rate, water production rate and water influx volume.

3.2 METHODS USED IN DEVELOPING THE MACHINE LEARNING MODEL

3.2.1 EPLORATIVE DATA ANALYSIS

When working with data sets, it is necessary to understand the data so that lots of useful information can be gotten from it. Explorative data analysis is carried out

on data to discover anomalies, investigate the data and check assumptions using graphical and statistical representation.

The following steps were carries out during the explorative data analysis;

- The necessary libraries (pandas, numpy, matplotlib, seaborn) used for data analysis were imported.
- The data set was loaded and read.
- After loading, the basic parameters of the data set were explored using summary method by examining the head (first few rows) of the pandas data frame to gain an idea of the content.
- The statistical summary of the data set was gotten and checked for null values. The missing value was eliminated to get equal number of points in the column.

Out[9]:

	ON_STREAM_HRS	AVG_DOWNHOLE_PRESSURE	AVG_DOWNHOLE_TEMPERATURE	AVG_DP_TUBING	AVG_ANNULUS_PRESS	AVG_CHOKE_SIZE_P
count	15349.000000	8980.000000	8980.000000	8980.000000	7890.000000	8919.000000
mean	19.994093	181.803869	77.162969	154.028787	14.856100	55.168533
std	8.369978	109.712363	45.657948	76.752373	8.406822	36.692924
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	24.000000	0.000000	0.000000	83.665361	10.841437	18.952989
50%	24.000000	232.896939	103.186689	175.588861	16.308598	52.096877
75%	24.000000	255.401455	106.276591	204.319964	21.306125	99.924288
max	25.000000	397.588550	108.502178	345.906770	30.019828	100.000000

Fig 3.1: statistical summary of data set

- Data Visualization: This is the graphical representation of data with the use of visualization tools to be able to identify the trend and patterns in the data set. The visualization tools used were; bar charts, histogram, kernel density plot and the scatter plots.

  The bar chart: This was used on the categorical data set to show the corresponding values to each data set (frequency of the variable). Fig 3.2 below shows the wellbore code of the five different wells with the number of data entries recorded from each of the wells.

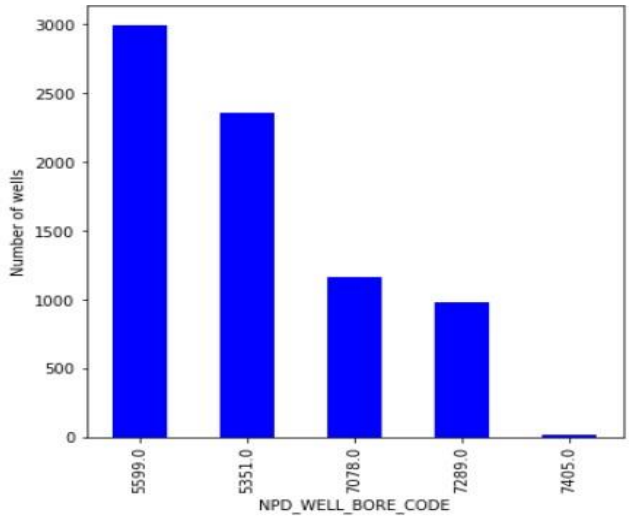

Fig 3.2: bar chart showing the different wells and their data entries.

The histogram: This was used to check the distribution and the range the numerical data falls into. It shows a visual representation of the numerical data by displaying the number of data points within a range of value. In fig 3.3, it is observed that the average choke size had its highest entries between 90 to 100.

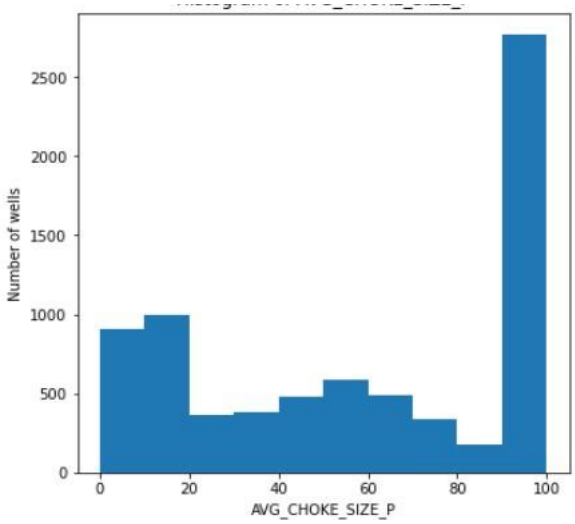

Fig 3.3: data visualization of choke size and the range of data entries

The kernel density plot: This was used to check if the data set were normally distributed. The figure below shows that the average downhole pressure is close to normal distribution but for the machine learning model

to work well it must be normally distributed and therefore will be analyzed during the data preparation.

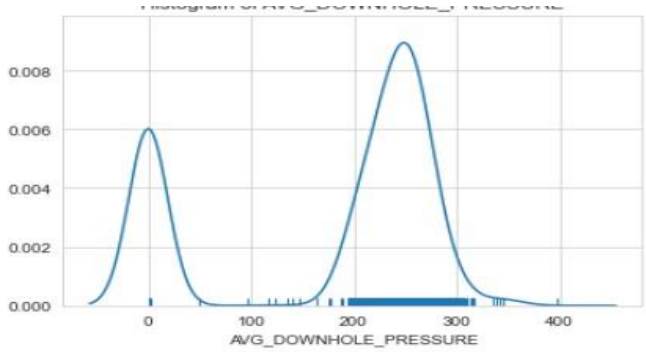

Fig 3.4: kernel density plot of avg downhole pressure

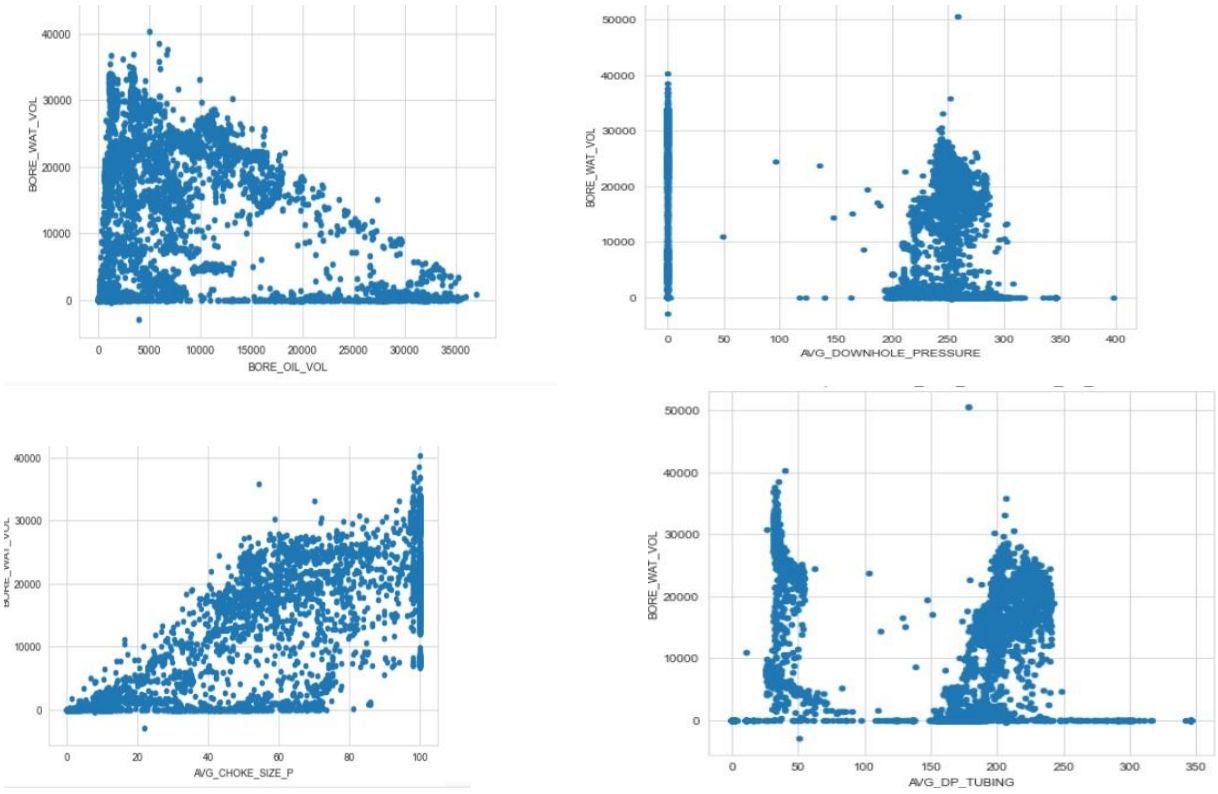

Fig 3.5: scatter plot of water production and other variables

The scatter plots: This is used to show the correlation relationship between variables. In developing our machine learning model, the scatter plot was used to show the relationship between water production and other variables.

Some scatter plots below shows that some variables such as oil volume and choke size show a linear relationship with the water production while some variables do not reflect such and therefore will not be used in our analysis.

3.2.2 DATA PREPARATION FOR ANALYSIS

The development of our machine learning model can't be done with raw features. These features will be selected and transformed to form new features to create a predictive model. The process of extracting these features from the raw data and transforming them is known as feature engineering. If good feature engineering is not done it will be difficult to get good insight from your data because good features can make poor machine learning models function well.

Approach to the feature engineering model includes.

Aggregating Categories: This is done on categorical variables with too many unique categories to reduce the number so that the predictive power of the machine learning model will not be limited. The data set in this work was investigated to ensure that each category has sufficient samples, the five different wells were renamed to well A, B, C, D and E and it was observed that the well will be negatively biased towards well D and E since they had insufficient input entry. Well E was aggregated with well B and well D with C making well A, B and C to have compactible values and sufficient data for regression. From the box plot below, the water production range of these three wells categories is distinctive and will be useful in predicting water production from wells in the volve field.

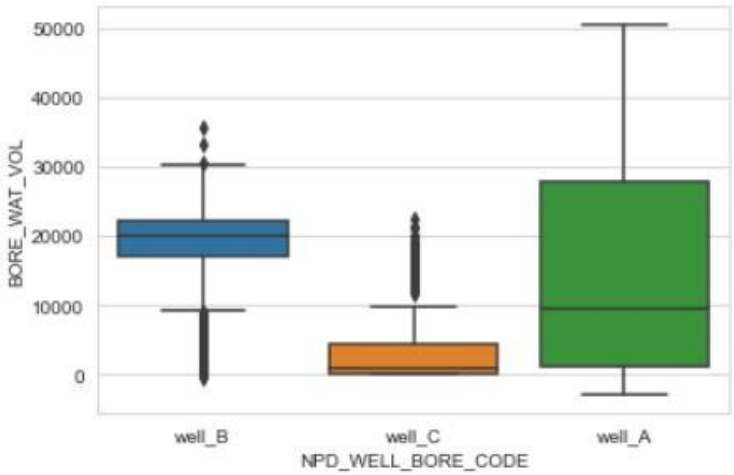

Fig 3.6: Box plot showing the three wells.

Transforming Numerical Variables: Transformation of values are applied to improve the performance of the machine learning model. This is done to make the relationship between variables more linear and make distributions closer to normal or at least more symmetric. From the data analysis, the distribution of water production is both skewed to the right and multimodal, a log transformation was carried out by taking the logarithm of the entries and dropping the entries with zero values to a more distinct and normal distribution. The square root power was also taken, and it gave the same result.

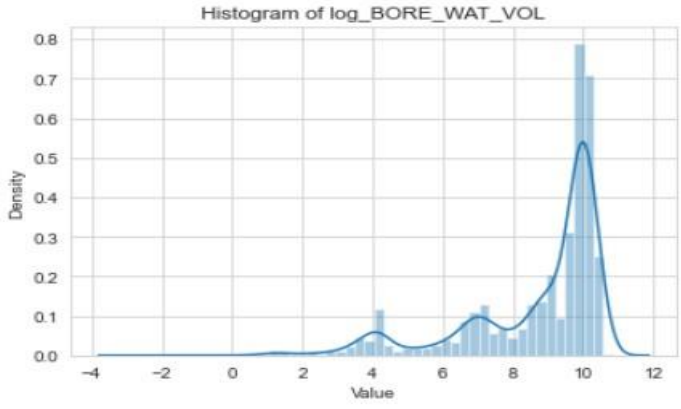

Fig 3.7: Logarithm transformation of bore water volume.

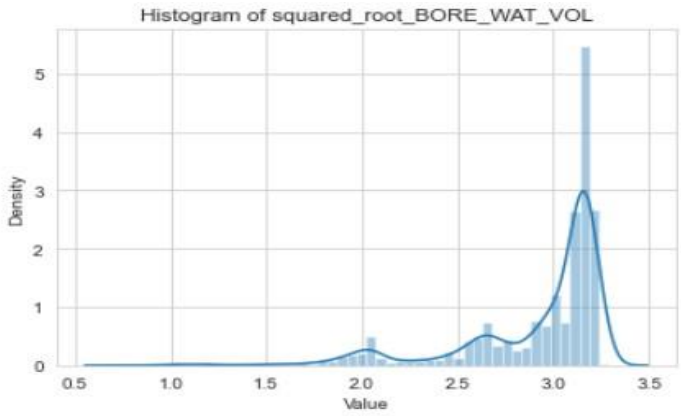

Fig 3.8: Square root transformation of bore water volume

3.2.3 DATA WRANGLING

Data wrangling are the different procedures used to transform raw data set in order to be ready for use depending on the work being carried out. In this work, the columns in the data set was checked and the columns that are not needed for the modelling was eliminated. From the scatter plot in the data visualization the columns that had no correlation with the bore water volume, the well type and the flow type was also dropped since they were the same for all entries. After the data wrangling, the prepared data for analysis was gotten having 12 columns and 6591 rows.

3.2.4 APPLYING LINEAR REGRESSION MODEL

Linear regression models the relationship between two variables in the data set by generating a linear equation to the observed data. It is used to predict the value of a dependent variable based on the value of an independent variable. Steps carried out in applying linear regression to model the water production includes;

Peparation of the model matrix: This involves the preparation of the matrix needed for training and testing. The machine learning model can not be applied on non numeric variables therefore the catergorical variables were converted to a

set of dummy variables ranging from 0 to 2 for well A, B and C respectively, three columns were created for the dummy variables so that they will have the same weight.

Procedures carried out in preparing the model matrix;

- The numeric features were added to the data set to complete the model matrix. The model matrix had 11 features; 3 dummy variables and 8 mumeric features.
- The model matrix constructed was splitted for training and testing with the training having the larger data.
- The numeric features were rescaled so that they can have similar range of values. This prevents features from having an undue influence on the model training because they have larger numeric values. They were scaled within the range of -1 to 1.

Constructing the Linear Regression Model: The linear regression model was computed using the prepared data that was split into training and test subset. With the dummy variables created there were 11 features, therefore the model required 11 coefficient. Since it contained dummy variables there was no specified intercept. The equation for the multiple regression is;

$$\hat{y} = f(\vec{x}) = \vec{\beta} \cdot \vec{x} + b = \beta 1x1 + \beta 2x2 + \cdots + \beta nxn + b$$

where; $\hat{y}$ are the predicted values or scores, $\vec{x}$ is the vector of feature values with components $\{x1, x2, \cdots, xn$, $\vec{\beta}$ is vector of model coefficients with components $\{\beta 1, \beta 2, \cdots, \beta n$, b is the intercept term, if there is one. You can think of the linear regression function $f(\vec{x})$ as the dot product between the beta vector $\vec{\beta}$ and the feature vector $\vec{x}$, plus the intercept term b. Since no intercept will be fit, the intercept value or bias will be accommodated in the coefficients of the dummy variables for the categorical features. The model is fit using the fit method with the numpy array of features and the label. The sklearn import linear model was

used to compute the least squares linear model and a linear regression model object was created with the Linear Regression method.

The residuals (difference between the predicted value and actual value) were analyzed using histogram and scatter plot to present a better approach to decide if the regression model is a good fit. Since the residual for a good regression model are random and normally distributed.

The features that were transformed during feature engineering to get a good range for the variables were untransformed using the squared and exponential method to get the real values used in the model.

3.2.5 OPTIMIZATION OF THE LINEAR REGRESSION MODEL

Regularization and Bias-Variance trade-off

Regularization is done to modify the model to perform well with new data set outside the training data set. When a model is overfit that is, it learned the training data set too well, it will perform poorly when new data sets are introduced.

Regularization methods are tools used to prevent overfitting of machine learning models. This method reduces the variance (measurement of deviation in the response variable while estimating it over a different training sample of data set) in the model, therefore introducing bias (measurement of the deviation or error from the real value of function) since the stronger the regularization the lower the variance and the greater the bias. Hence, when applying regularization, the bias-variance trade-off will be dealt with.

In this work, the regularization was done with the L2 Regularization and the L1 Regularization.

L2 Regularization

It is also called the Ridge Regression. It used the principle of sum of square error of the model coefficient to reduce the variance in the model to prevent overfitting. L2 Regularization can drive some coefficient towards zero, usually not zero.

L1 Regularization

This is also known as the Lasso method. The L1 Regularization was used to limit the sum of the absolute values of the model coefficient therefore reducing the variance in the model. L1 Regularization can drive model coefficient to zero.

3.2.6 SELECTION AND VALIDATION OF THE MODEL

The aim of model selection is to get the model that performs best for the model that performs best for the problem that is to be solved. Finding the optimum value of the regularized parameter for the L2 and L1 Regularization model that was created is an example of model selection.

The model selection was used to.

- Select optimal model hyperparameters used to determine the model characteristics.
- Select the features used for the model.
- Compare the different model types.

Cross Validation

When new independent data cannot be obtained to validate a model, the data set can be split into training set for developing the model and testing set for evaluating the predictive ability as an alternative. This is called cross validation.

In this work, the data set were split into 10 folds roughly equal in size. For each part that was used as training sample, the remaining nine were used as testing sample to predict the model. 10 folds were fitted for the 11 candidates totaling 110 fits. The statistical summary for the 10 folds for each of the 11 candidates were gotten. The R-Squared was plotted against the test and train score for the 10 folds to get the optimum value of number of features for building the final model.

3.2.7 USING THE PREPARED DATA ON DIFFERENT MACHINE LEARNING MODELS

The prepared data was used on Support vector regression, Decision tree regression and the Random Forest regression to know if the Adjusted R-Squared, R-Squared and the Root mean square error can be improved.

Decision Tree Regressor

Decision tree is a predictive modelling approach in machine learning that uses a decision tree as the predictive model. It represented the independent variables in the prepared data (features) as branches and the dependent variable (the label) as the leaves in developing the model.

Random Forest Regressor

The random forest regressor was used to correct for overfitting to the training data set in the decision trees for a better predictive ability.

CHAPTER FOUR: RESULT AND INTERPRETATION

4.1 RESULT PRESENTATION

The linear regression equation gotten from the linear regression model becomes.

y= 2.88666811A1 + 2.90943831A2 + 3.00821597A3 + 0.01563032A4 + 0.01042541A5 – 0.05834426A6 – 0.03392682A7 – 0.31664735A8 – 0.12218957A9 – 0.1325319A10 + 0.04959461A11.

where y is the predicted value of water production and A1, A2, A3…...A11 are Well A, Well B, Well C, Time, Annulus Pressure, Choke size, Wellhead Pressure, Wellhead Temperature, DP Choke size, Oil volume and Gas volume respectively.

It was evaluated to know how good it performs and how it approximates the relationship based on the mean square error, root mean square error, mean absolute error, median absolute error and R-squared.

Mean Square Error	0.040696920547311194
Root Mean Square Error	0.20173477773381365
Mean Absolute Error	0.12196762870172971
Median Absolute Error	0.0703554289999444
R-Squared	0.7365842976053831
Adjusted R-Squared	0.7329776931608321

Table 4.1: Result from the linear regression model.

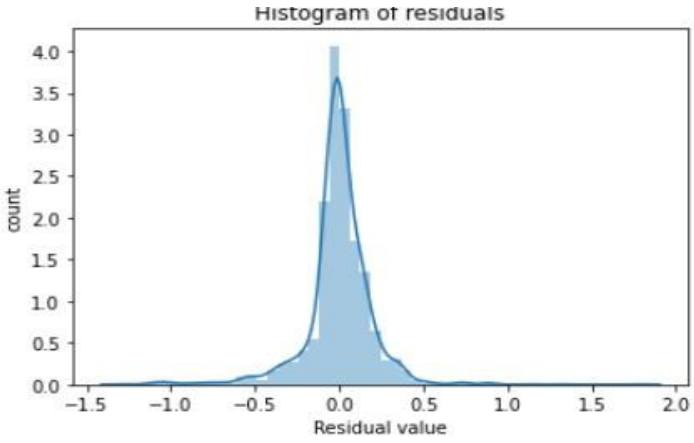

Fig 4.1: Histogram of residual value.

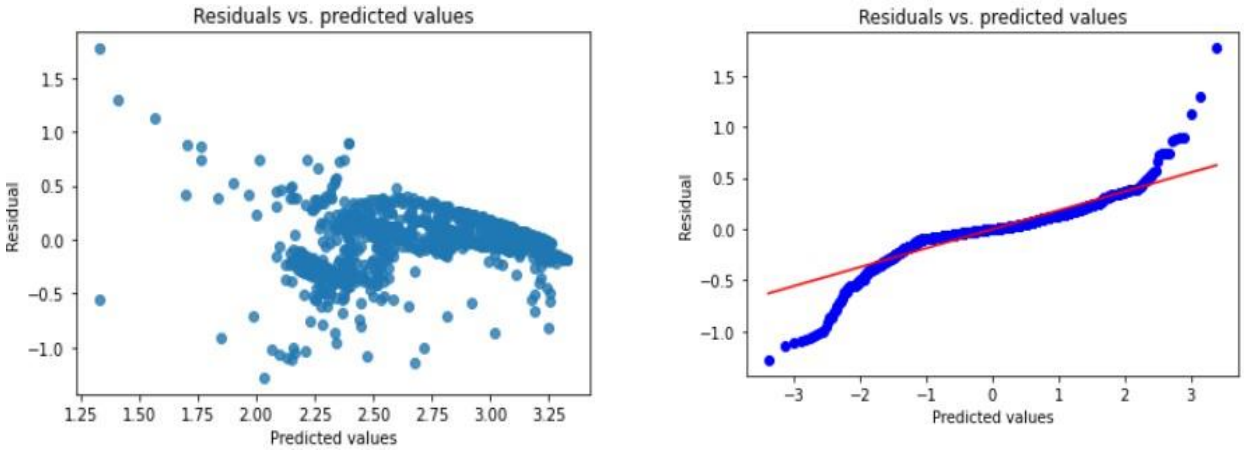

Fig 4.2: Residual versus predicted values

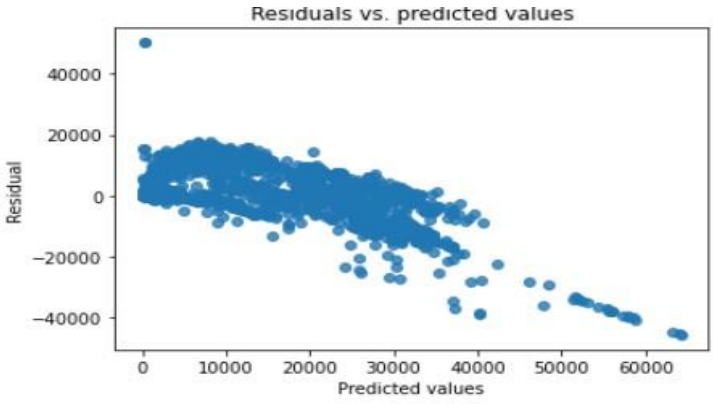

Fig 4.3: Residual vs predicted values after it was untransformed.

Evaluating the model based on the best L2 Regularization gave the following result.

Mean Square Error	0.04070615992168422
Root Mean Square Error	0.20175767623980065
Mean Absolute Error	0.12202485841650257
Median Absolute Error	0.07046772519040378
R-Squared	0.7365244946459103

Table 4.2: Result from L2 Regularization.

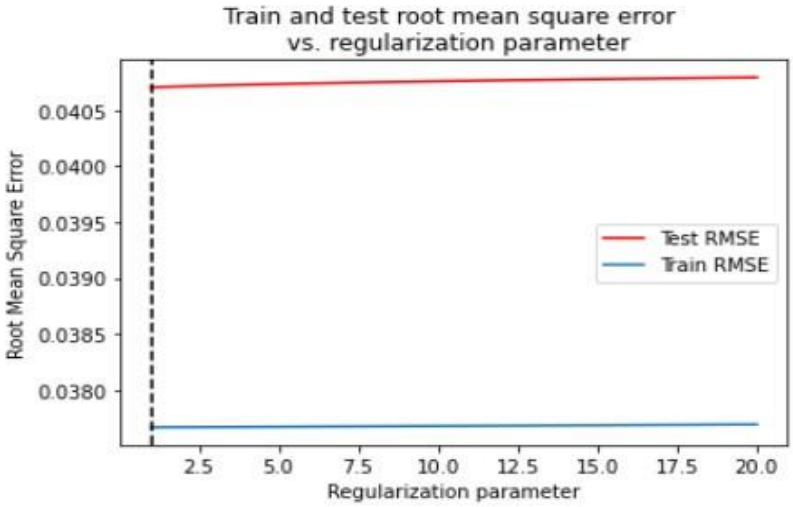

Fig 4.4: Train and test root mean square error vs regularization parameter.

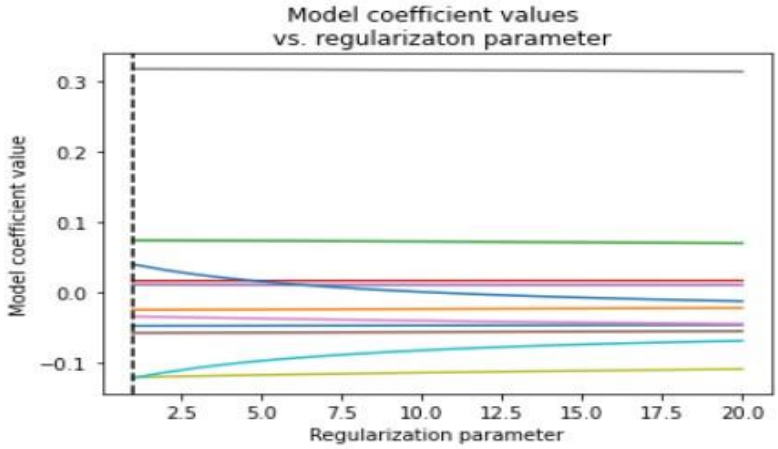

Fig 4.5: Model coefficient values vs regularization parameter.

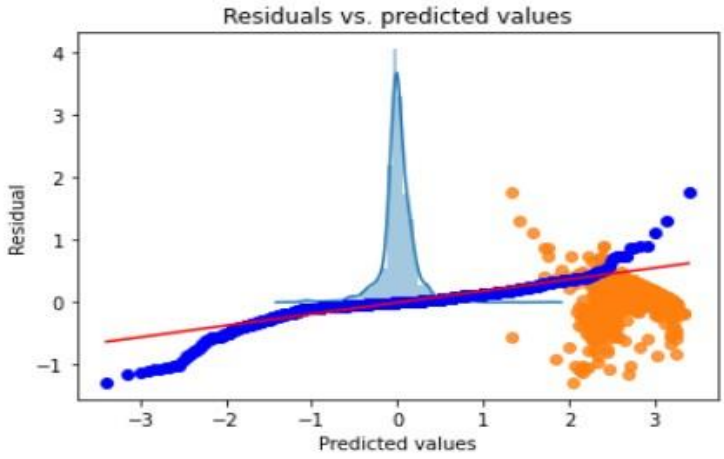

Fig 4.6: Residual vs predicted values for L2 Regularization.

Evaluating the model based on the best L1 Regularization gave the following result.

Mean Square Error	0.04073779682919529
Root Mean Square Error	0.2018360642432251
Mean Absolute Error	0.12220180648432995
Median Absolute Error	0.07120483928888155
R-Squared	0.7365244946459103

Table 4.3: Result from L1 Regularization.

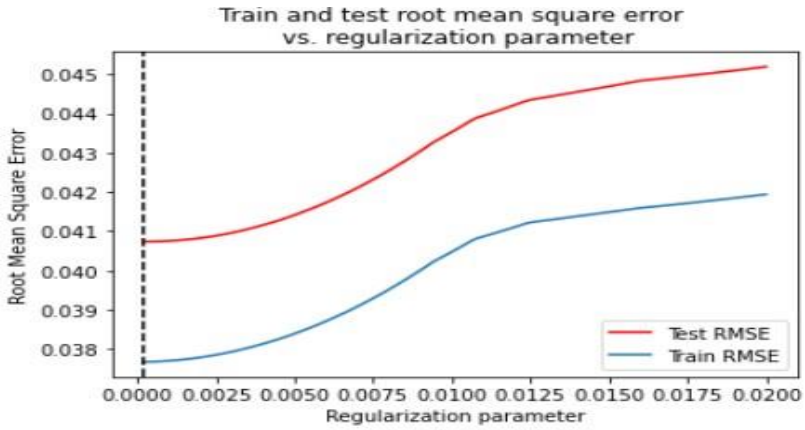

Fig 4.7: Train and test root mean square error vs regularization parameter.

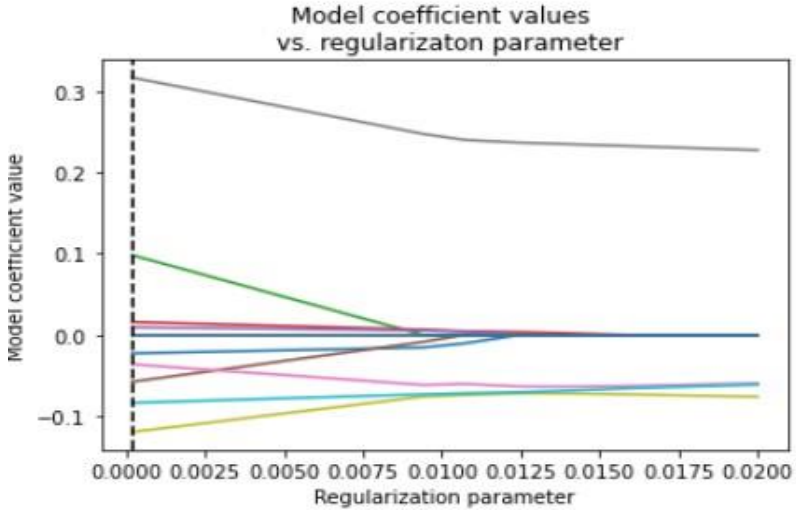

Fig 4.8: Model coefficient values vs regularization parameter.

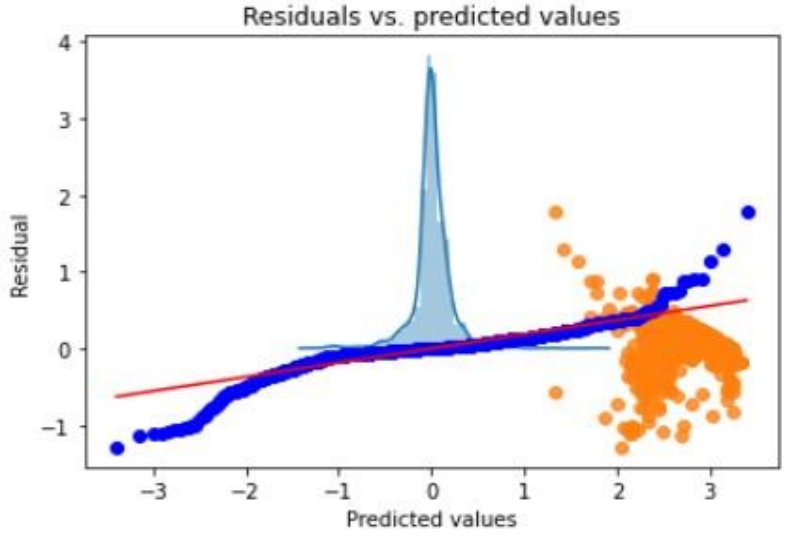

Fig 4.9: Residual vs predicted values for L1 Regularization.

	mean_fit_time	std_fit_time	mean_score_time	std_score_time
0	0.008046	0.005835	0.000278	0.000093
1	0.004351	0.000142	0.000263	0.000089
2	0.004504	0.000728	0.000263	0.000053
3	0.004038	0.000526	0.000248	0.000045
4	0.003174	0.000126	0.000211	0.000013
5	0.002979	0.000497	0.000207	0.000010
6	0.002349	0.000029	0.000199	0.000006
7	0.002060	0.000127	0.000221	0.000021
8	0.001722	0.000094	0.000281	0.000152
9	0.001234	0.000052	0.000222	0.000032
10	0.000735	0.000049	0.000211	0.000006

ect	params	split0_test_score	split1_test_score	split2_test_score	split3_test_score	...	split2_train_score	split3_train_score
1	{'n_features_to_select': 1}	0.495863	0.451901	0.489543	0.480248	...	0.479721	0.481031
2	{'n_features_to_select': 2}	0.669080	0.738825	0.724442	0.642832	...	0.695225	0.703906
3	{'n_features_to_select': 3}	0.695087	0.763243	0.753136	0.648286	...	0.713096	0.724310
4	{'n_features_to_select': 4}	0.734634	0.746253	0.768357	0.684568	...	0.742150	0.751021
5	{'n_features_to_select': 5}	0.733528	0.745209	0.767898	0.684141	...	0.742271	0.751128
6	{'n_features_to_select': 6}	0.728417	0.755364	0.779078	0.673816	...	0.746473	0.757533
7	{'n_features_to_select': 7}	0.729482	0.755822	0.778683	0.673958	...	0.746732	0.757730
8	{'n_features_to_select': 8}	0.729485	0.755730	0.779293	0.674423	...	0.746873	0.757899
9	{'n_features_to_select': 9}	0.734256	0.755738	0.779294	0.674418	...	0.746873	0.757900
10	{'n_features_to_select': 10}	0.732882	0.760081	0.784010	0.674222	...	0.748158	0.759775
11	{'n_features_to_select': 11}	0.732748	0.761329	0.783763	0.675458	...	0.748737	0.760193

split4_train_score	split5_train_score	split6_train_score	split7_train_score	split8_train_score	split9_train_score	mean_train_score	std_train_score
0.480980	0.490070	0.473428	0.474806	0.481008	0.485151	0.481040	0.004619
0.695586	0.518043	0.695278	0.694664	0.697569	0.707075	0.680184	0.054216
0.713969	0.723282	0.713461	0.712984	0.717088	0.727714	0.717725	0.005404
0.742053	0.749019	0.741167	0.741473	0.744642	0.750907	0.745321	0.003628
0.742183	0.753747	0.741297	0.745776	0.749854	0.756097	0.747338	0.004870
0.746771	0.753909	0.746991	0.746022	0.750085	0.756313	0.750557	0.004030
0.747017	0.753909	0.747203	0.746282	0.750085	0.756313	0.750703	0.003968
0.747243	0.753984	0.747471	0.746282	0.750169	0.756410	0.750832	0.003951
0.747243	0.753985	0.747471	0.747797	0.751826	0.758838	0.751536	0.004240
0.749169	0.756463	0.748835	0.747798	0.752131	0.758853	0.752652	0.004253
0.749693	0.756846	0.749599	0.748294	0.752132	0.759414	0.753065	0.004199

Fig 4.10: Statistical summary of the cross validation.

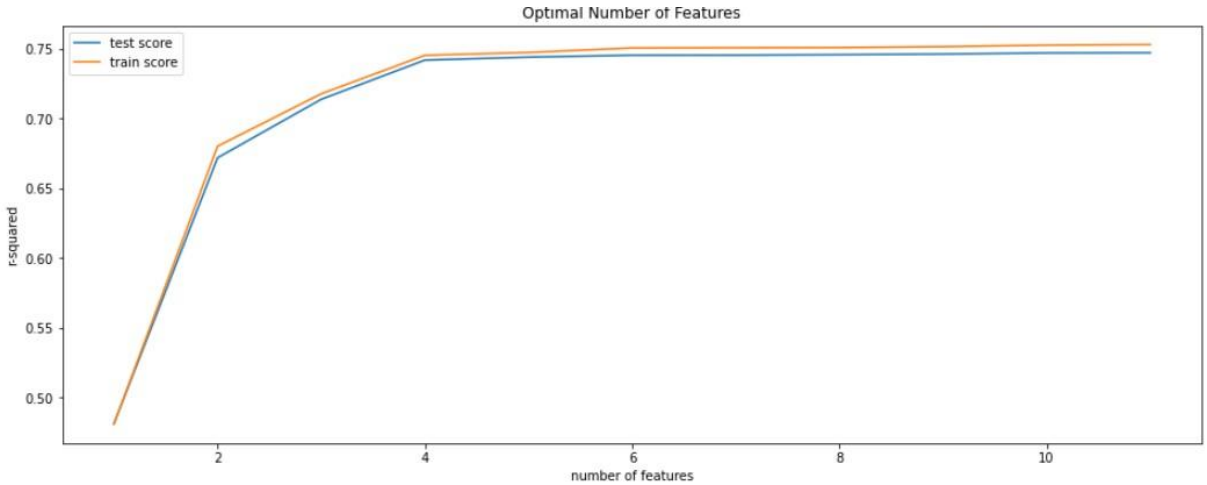

Fig 4.11: Plot of R-Squared vs number of features.

MODELS	ADJUSTED R^2	R-SQUARED	RMSE
Random Forest Regression	0.91	0.92	0.11
Support Vector Regression	0.90	0.90	0.12
Decision Tree Regression	0.86	0.86	0.14
Linear Regression	0.74	0.74	0.20

Table 4.4: Different models and their coefficient of determination.

Checking the accuracy of the various models used for predicting the water production for the first fifty values in the test set.

```
print(y_test[:50])
```

```
[3.18304818 2.65459681 3.10250933 2.21380156 2.04655759 3.10268924
 2.67412645 2.52558913 2.04397806 3.15463101 2.52077055 3.1278518
 3.20410513 3.21938562 3.14356766 1.95726973 3.07871326 2.58149533
 2.38945707 2.90841809 2.98442676 3.1802852  3.22065607 3.21027351
 2.98084944 3.20651649 3.20511914 3.00281703 3.13626721 3.16885324
 3.13364475 2.66723294 3.15073707 2.84711659 2.66438014 3.13419321
 2.67859656 2.15669643 3.22347372 2.63533602 2.99669655 3.15196713
 2.89144908 2.98179273 2.30854019 3.00482229 3.16646464 2.03702787
 3.12159218 3.21050295]
```

Fig 4.12: actual values of water production

```
In [129]: print(y_pred1[:50])
```

```
[3.07848326 2.48369699 3.21046796 1.8349685  2.28850624 3.2150767
 2.72027932 2.41772497 2.31959425 3.23232578 2.4219721  2.86052312
 3.08732027 3.24670309 3.02728945 2.30336635 3.09052525 2.53916835
 2.39684127 2.57389364 3.0451662  3.18386622 3.22071007 3.14993942
 2.9895031  3.16706637 3.1929461  3.09702416 3.31215492 3.08346174
 3.21586526 2.75249454 3.23191641 2.52732119 2.58758085 3.21681614
 2.73960297 2.54764529 3.21724771 2.72394077 3.0145141  3.17315842
 2.92566745 3.00390584 2.50216355 3.09530174 3.04325285 2.21326267
 3.00760969 3.17678487]
```

Fig 4.13: predicted values of water production using linear regression model.

```
print(ypred_regressor[:50])
```

```
[3.10039692 2.62896203 3.08538808 2.40855792 2.0995914  3.08089602
 2.77839816 2.46902482 2.14910845 3.17311206 2.52030211 3.10278973
 3.12251479 3.14878568 3.1808052  1.83588404 3.05471497 2.57138326
 2.38042694 2.77763326 3.07484249 3.17491157 3.14578407 3.16585657
 3.02088675 3.1623654  3.14550652 3.04450357 3.07406741 3.10211268
 3.11815872 2.73458124 3.08517226 2.71315239 2.7147139  3.14671674
 2.72803642 2.55362531 3.16909547 2.66836453 3.06997065 3.17826822
 2.9878594  3.06417551 2.42740094 3.05106046 3.10942233 1.98604272
 3.20875742 3.17249852]
```

Fig 4.14: predicted values of water production using support vector regression model.

```
print(ypred_rfr[:50])
```

```
[3.1757291  2.57703683 3.10304326 2.02502519 2.02670332 3.12643824
 2.7182003  2.592608   2.04671115 3.15956128 2.55177941 3.12795019
 3.15133273 3.22259275 3.13588376 1.9548603  3.07767749 2.59331205
 2.21616294 2.90811624 2.98401189 3.18015992 3.22364957 3.2141099
 2.99089011 3.20685509 3.20266314 2.99907107 3.13705307 3.16877235
 3.1327605  2.675478   3.11342171 2.81801431 2.66730782 3.14252979
 2.67489229 2.55679133 3.22827543 2.62662832 3.04289372 3.15037265
 2.94515588 2.98306882 2.40419019 3.00070227 3.1689483  2.04289737
 3.15215071 3.21299629]
```

Fig 4.15 predicted values of water production using random forest model.

4.2 RESULT DISCUSSION

The model developed the coefficient of the linear regression equation above with the intercept of zero.

Mean Square Error

Mean square error of approximately 0.041 was gotten from the model, this shows the average of the square of errors that is the average squared difference between the predicted value and the actual value. The mean square error is

always positive or greater than zero. While zero is a perfect predictor, a value close to zero will represent a good quality predictor.

Root Mean Square Error

The square root of the average of the squared difference between the predicted value and actual value from the model is approximately 0.202.

Mean Absolute Error

The arithmetic average of the actual value and the value predicted by the model I approximated 0.122. A mean absolute error of zero shows the model is a perfect predictor.

Median Absolute Error

The measure of the of the median of the absolute deviation from the median of the distribution in the model is approximately 0.070.

R-Squared (R^2)

This is also called the coefficient of determination; it portrays how good the model fits the dataset and shows how closely the predicted values are to the actual values. The R-Squared gotten from the linear regression model is approximately 0.737. The value of R-Square lies between 0 and 1, where 0 indicates that the that model fits perfectly to the data set provided.

Adjusted R-Squared

The Adjusted R-Squared from the model is approximately 0.733. It is a modified R-Squared, it shows the amount of the variance by only the independent variable that affect the dependent variable.

The histogram in **Fig 4.1** shows that the residuals are in small range and there are some noticeable skews in the distribution and the cross plot in **Fig 4.2** shows that they were generally random.

From **Fig 4.3** above, the scatter plot of residual versus predicted values after transformation showed that the residual decreases as the predicted value increases.

L2 Regularization

The train and test root mean square error plotted against regularization parameter in **Fig 4.4** and the model coefficient value plotted against regularization parameter in **Fig 4.5** shows the optimum value gotten at (1, 0.04070615992168422).

The Residual plotted against the predicted values in **Fig 4.6** compared the error metrics achieved to those of the un regularized model. The error metrics for the regularized model were better which indicates that the regularized model generalize better than the unregularized model. The residuals are closer to normally distributed than the unregularized model.

L1 Regularization

The train and test root mean square error plotted against regularization parameter in **Fig 4.7** and the model coefficient value plotted against regularization parameter in **Fig 4.8** shows the optimum value gotten at (0.0002, 0.04073779682919529).

Comparing the performance metrics of L1 Regularized model in Fig 4.9 to the metrics of the un regularized model and the L2 Regularized model shows that the metrics are in between the two previous model and the residual are closer to the unregularized model.

Fig 4.10 shows the statistical summary for the 10 folds for each of the 11.

Fig 4.11 shows the plot of R-Squared versus the test score and train score (number of features). From the figure the optimum value for the number of features gotten for building the final model is 0.7365244946459103.

Comparison of different models with the linear regression model in table 4.4 shows that the random forest regression had the highest value of R-Square which was 0.92 and the lowest value of Root mean square error which was 0.11 compared to the Decision tree, linear regression, and support vector regression.

CHAPTER FIVE: CONCLUSION AND RECOMMENDATION

5.1 CONCLUSION

Accurate determination of water production has been a major challenge in the oil and gas development and management. This project shows that the predicted values for the water production with the linear regression model can be used for the prediction of water production in oil and gas wells instead of the approximated and complex empirical correlations.

The linear regression model that was developed for water prediction using the volve production data that was gotten from the Norwegian Sea had the coefficient of determination (R^2) of 0.73298 and Mean Square Error of 0.0406 which indicates that the model fits into the data set and it is reliable.

The model was optimized to avoid overfitting by reducing the variance so that it will not perform poorly when used on a different production set. Model validation was carried out using the cross validation by splitting the data into training and testing set. The coefficient of determination (R^2) was improved to 0.73652 which is good but not so high, the data set was used on other models for better predictive capacity and the Random Forest Model gave a better coefficient of determination (R^2) of 0.92 and RMSE of 0.11, this indicates a good reliability of the proposed model.

Hence, the Random Forest Model can be used to predict water production in different oil and gas wells.

5.2 RECOMMENDATION

The linear regression model has some shortcomings which includes the fact that the linear regression model is limited to linear relationships, its sensitive to outliers and looks at the mean of the dependent variables. The linear regression

model also assumes that the errors must be normally distributed. Therefore, in a way to improve the predictive values the Random Forest model or the support vector regression should be used for the prediction of water production since they gave a higher coefficient of determination from the result above.

REFERENCES

Echufu-Agbo Ogbene Alexis. (2010). Diagnostic plot for analysis of water production and reservoir performance.

Bailey .B, Crabtree .M, Tyrie .J, Elphick .J, Kuchuk .F, Romano .C and Roodhart .L. (2000). The challenge of water control. *Oilfield Review.*

T. Cross, K. Sathaye, K. Darnell, D. Niederbut and K.Crifasi. (2020). Predicting water production in the willston basin using a machine learning model. *URTeC: 2756.*

G. Zamonsky, P.E Lancentre and A. E. Lacentre. (2005). Towards better correlations for water production prediction using sensitivity analysis and numerical simulation models. *SPE 94457.*

K. J Wawro, Talisman Energy, F. R. Wassmuth, and J. E. Smith. (2000). Reducing water production in a naturally fractured gas well using sequential gel/gas slug injection. *SPE 59746.*

SPRI TEAM. (2018). Secondary recovery method: Water flooding. *Sierra Pine Resources International Blog.*

Dhafer A. S (2019). Comprehensive evaluation of water breakthrough with a novel method to estimate water production. *SPE-198643-MS.*

Swain .K, Silpakom .D, Chutarat .S, Tanarat .K, Patharud .P, Tawam .W, Ikenna .C. and Khamis.A. (2020). Collaborative analytical technique to

improve production allocation in a mature oil field producing high water cut wells. *OTC-30433-MS.*

C. M. F Galas. (2003). The art of history matching – Modelling water production under primary recovery. *Canada International Petroleum Conference.*

Yisa .A., Emeka .O., Augusta .E., Ernest .B., Onyebuchi .O. and Ubong .U. (2020). Modelling the scaling tendency of produced water. A case study of Niger delta brown field. *SPE-203715-MS.*

Chris Capenter. (2020). Capacitance – Resistance Model used for integrated detection of water production. *Journal of Petroluem Technology*.

Andrea .Q., Eduardo .D., and Alex .O. (2020). Successful control of high-water production with thixotropic conformance technology in a horizontal well; Diagnosing the water production mechanism. *SPE-199828-MS.*

Boyun Guo, William .C. Lynos and Ali Ghalambor. (2007). Petroleum production engineering. A computer – Assisted Approach. *Elsevier Inc.*

Earl Hangstrom, Christopher Lyles, Mala Pattanayek, Bridgette Deshields and Mark Berkman. (2016). Produced water – Emerging challenges, risks and opportunities. *Environmental claims journal.*

Miguel Armenta. (2003). Mechanisms and control of water inflow towards wells in gas reservoirs with bottom water drive. *LSU Doctoral Dissertations.*

Naveen Joshi. (2019). Leveraging Artificial Intelligence in the oil and gas industry. *Allern Technology*.

Yogendra .N. P., Ayush .R., Sribharath .K., Srimoyee .B. and Luigi .S. (2020). Machine learning in the oil and gas industry. Including geosciences, reservoir engineering and production engineering with python.

Shengnanchen. (2019). Application of machine learning model to predict well productivity in Montney and Duvernay. *University of Calgary*.

Matthew Taylor. (2021). Machine learning in the Oil and Gas Industry. *New engineer*.

Micheal Economides, Daniel .A.H., Christine .E. and Ding .Z. (2013). Petroleum production system. *Person Education Inc.*

Mohammed Al-Jawad and Dhifaf Sadeq. (2006). Well performance analysis based on flow calculations and IPR. *Journal of Engineering*.

Larry .W. Lake and Joe Dunn Clegg. (2007). Petroleum engineering handbook volume iv, production operations engineering. *Society of Petroleum Engineering.*

Luigi .S., Joao .O., Turiassu .A. and Alvaro .E. (2004). Real time petroleum optimization. *Rio Oil and Gas Expo and Conference*.

S .I. Onwukwe. (2015). Techniques of controlling water coning in oil reservoirs. *ADR Journals.*

APPENDIX

import modules for the project work

In [2]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

load volve production dataset

In [2]:
```
volve=pd.read_excel(r'C:\Users\DELL\Documents\VolveFieldProductionData.xlsx')
```

In [3]:
```
volve.head(5)
```

Out[3]:

	DATEPRD	WELL_BORE_CODE	NPD_WELL_BORE_CODE	NPD_WELL_BORE_NAME	NPD_FIELD_CODE	NPD_FIELD_NAME	NPD_FACILITY_CODE	NPD_
0	2014-04-07	NO 15/9-F-1 C	7405.0	15/9-F-1 C	3420717.0	VOLVE	369304.0	M
1	2014-04-08	NO 15/9-F-1 C	7405.0	15/9-F-1 C	3420717.0	VOLVE	369304.0	M
2	2014-04-09	NO 15/9-F-1 C	7405.0	15/9-F-1 C	3420717.0	VOLVE	369304.0	M
3	2014-04-10	NO 15/9-F-1 C	7405.0	15/9-F-1 C	3420717.0	VOLVE	369304.0	M
4	2014-04-11	NO 15/9-F-1 C	7405.0	15/9-F-1 C	3420717.0	VOLVE	369304.0	M

5 rows × 24 columns

In [15]:
```
def count_unique(volve, cols):
    for col in cols:
        print('\n' + 'For column ' + col)
        print(volve[col].value_counts())

cols=['NPD_WELL_BORE_CODE','NPD_FIELD_CODE','NPD_FACILITY_CODE' ]
count_unique(volve, cols)
```

```
For column NPD_WELL_BORE_CODE
5599.0    2993
5351.0    2360
7078.0    1159
7289.0     978
7405.0      14
Name: NPD_WELL_BORE_CODE, dtype: int64

For column NPD_FIELD_CODE
3420717.0    7504
Name: NPD_FIELD_CODE, dtype: int64

For column NPD_FACILITY_CODE
369304.0    7504
Name: NPD_FACILITY_CODE, dtype: int64
```

In [13]:
```
volve.dropna(axis=0,inplace=True)
```

In [14]:
```
volve.count()
```

Out[14]:
```
DATEPRD                     7504
WELL_BORE_CODE              7504
NPD_WELL_BORE_CODE          7504
NPD_WELL_BORE_NAME          7504
NPD_FIELD_CODE              7504
NPD_FIELD_NAME              7504
NPD_FACILITY_CODE           7504
NPD_FACILITY_NAME           7504
ON_STREAM_HRS               7504
AVG_DOWNHOLE_PRESSURE       7504
AVG_DOWNHOLE_TEMPERATURE    7504
AVG_DP_TUBING               7504
AVG_ANNULUS_PRESS           7504
AVG_CHOKE_SIZE_P            7504
AVG_CHOKE_UOM               7504
AVG_WHP_P                   7504
AVG_WHT_P                   7504
DP_CHOKE_SIZE               7504
BORE_OIL_VOL                7504
BORE_GAS_VOL                7504
BORE_WAT_VOL                7504
BORE_WI_VOL                 7504
FLOW_KIND                   7504
WELL_TYPE                   7504
dtype: int64
```

from the above codes wriiten missing values have been droped and all have equal number of datapoints

```
def plot_histogram(volve, cols, bins = 10):
    import pandas as pd
    import numpy as np
    for col in cols:
        fig = plt.figure(figsize=(6,6)) # define plot area
        ax = fig.gca() # define axis
        volve[col].plot.hist(ax = ax, bins = bins) # Use the plot.hist method on subset of the data frame
        ax.set_title('Histogram of ' + col) # Give the plot a main title
        ax.set_xlabel(col) # Set text for the x axis
        ax.set_ylabel('Number of wells')# Set text for y axis
        plt.show()
    for col in cols:
        volve[col]=pd.to_numeric(volve[col])
    return volve

p= list(volve.columns)
num_cols=p[1::]
plot_cols=['NPD_WELL_BORE_CODE','NPD_FIELD_CODE','NPD_FACILITY_CODE','ON_STREAM_HRS',
 'AVG_DOWNHOLE_PRESSURE',
 'AVG_DOWNHOLE_TEMPERATURE',
 'AVG_DP_TUBING',
 'AVG_ANNULUS_PRESS',
 'AVG_CHOKE_SIZE_P' ]
plot_colss=['NPD_WELL_BORE_CODE','NPD_FIELD_CODE','NPD_FACILITY_CODE','FLOW_KIND', 'WELL_TYPE' ]
plot_histogram(volve,plot_cols)
```

```
def plot_scatter(volve, cols, col_y = 'BORE_WAT_VOL'):
    for col in cols:
        fig = plt.figure(figsize=(7,6)) # define plot area
        ax = fig.gca() # define axis
        volve.plot.scatter(x = col, y = col_y, ax = ax)
        ax.set_title('Scatter plot of ' + col_y + ' vs. ' + col) # Give the plot a main title
        ax.set_xlabel(col) # Set text for the x axis
        ax.set_ylabel(col_y)# Set text for y axis
        plt.show()

prodVariable=['ON_STREAM_HRS', 'AVG_DOWNHOLE_PRESSURE',
       'AVG_DOWNHOLE_TEMPERATURE', 'AVG_DP_TUBING', 'AVG_ANNULUS_PRESS',
       'AVG_CHOKE_SIZE_P', 'AVG_CHOKE_UOM', 'AVG_WHP_P', 'AVG_WHT_P',
       'DP_CHOKE_SIZE', 'BORE_OIL_VOL', 'BORE_GAS_VOL']
plot_scatter(volve, prodVariable)
```

Add the numeric features

To complete the model matrix, execute the code in the cell below to concatenate the three numeric features.

```
]: Features = np.concatenate([Features, np.array(volvePrepared1[[ 'ON_STREAM_HRS', 'AVG_ANNULUS_PRESS',
       'AVG_CHOKE_SIZE_P', 'AVG_WHP_P', 'AVG_WHT_P', 'DP_CHOKE_SIZE',
       'BORE_OIL_VOL', 'BORE_GAS_VOL']])], axis = 1)
   Features[:2,:]
```

```
:[12]: array([[0.00000000e+00, 0.00000000e+00, 1.00000000e+00, 2.40000000e+01,
        2.30083271e+01, 9.86603563e+00, 9.07523181e+01, 6.00839548e+01,
        6.24895608e+01, 6.58579799e+03, 5.59938159e+06],
       [0.00000000e+00, 0.00000000e+00, 1.00000000e+00, 2.40000000e+01,
        2.30368419e+01, 9.81311063e+00, 9.10065954e+01, 5.95012948e+01,
        6.27646710e+01, 6.44226475e+03, 5.51234674e+06]])
```

```
]: Features.shape
```

```
:[13]: (6591, 11)
```

Applying Linear Regression to model Water production

```
[3]: import pandas as pd
     from sklearn import preprocessing
     import sklearn.model_selection as ms
     from sklearn import linear_model
     import sklearn.metrics as sklm
     import numpy as np
     import numpy.random as nr
     import matplotlib.pyplot as plt
     import seaborn as sns
     import scipy.stats as ss
     import math

     %matplotlib inline
```

```
[66]: volvePrepared=pd.read_csv(r'C:\Users\DELL\Documents\VolvePrepared.csv')
```

```
[6]: volvePrepared.columns
```

Evaluate the model

we use the test dataset to evaluate the performance of the regression model. As a first step, we execute the code in the cell below to compute and display various performance metrics and examine the results.

```
y_test[np.isnan(y_test)] = np.median(y_test[~np.isnan(y_test)])
```

```
def print_metrics(y_true, y_predicted, n_parameters):
    ## First compute R^2 and the adjusted R^2
    r2 = sklm.r2_score(y_true, y_predicted)
    r2_adj = r2 - (n_parameters - 1)/(y_true.shape[0] - n_parameters) * (1 - r2)

    ## Print the usual metrics and the R^2 values
    print('Mean Square Error      = ' + str(sklm.mean_squared_error(y_true, y_predicted)))
    print('Root Mean Square Error = ' + str(math.sqrt(sklm.mean_squared_error(y_true, y_predicted))))
    print('Mean Absolute Error    = ' + str(sklm.mean_absolute_error(y_true, y_predicted)))
    print('Median Absolute Error  = ' + str(sklm.median_absolute_error(y_true, y_predicted)))
    print('R^2                    = ' + str(r2))
    print('Adjusted R^2           = ' + str(r2_adj))

y_score = lin_mod.predict(x_test)
print_metrics(y_test, y_score, 28)
```

Cross Validation in sklearn

1. Using K-Fold CV

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import re

import sklearn
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import MinMaxScaler
from sklearn.preprocessing import PolynomialFeatures
from sklearn.preprocessing import scale
from sklearn.feature_selection import RFE
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import cross_val_score
from sklearn.model_selection import KFold
from sklearn.model_selection import GridSearchCV
from sklearn.pipeline import make_pipeline
```

4]:

```
# step-1: create a cross-validation scheme
folds = KFold(n_splits = 10, shuffle = True, random_state = 100)

# step-2: specify range of hyperparameters to tune
hyper_params = [{'n_features_to_select': list(range(1, 12))}]


# step-3: perform grid search
# 3.1 specify model
lm = lin_mod_l2
lm.fit(x_train, y_train)
rfe = RFE(lm)

# 3.2 call GridSearchCV()
model_cv = GridSearchCV(estimator = rfe,
                        param_grid = hyper_params,
                        scoring= 'r2',
                        cv = folds,
                        verbose = 1,
                        return_train_score=True)

# fit the model
model_cv.fit(x_train, y_train)
```

```
Fitting 10 folds for each of 11 candidates, totalling 110 fits
```

```
# pip install lazypredict
```

```
import lazypredict
```

```
from lazypredict.Supervised import LazyRegressor
reg = LazyRegressor(verbose = 0, ignore_warnings = False, custom_metric = None)
models, predictions = reg.fit(x_train, x_test, y_train, y_test)
print(models)
```

```
100%|██████████| 42/42 [00:20<00:00,  2.09it/s]
```