


Prediction of Oil Production through Linear Regression Model and Big Data Tools

Rehab Alharbi¹, Nojood Alageel², Maryam Alsayil³, Rahaf Alharbi⁴, and A'aeshah Alhakamy⁵ 

Faculty of Computers and Information Technology, Master of Artificial Intelligence at University of Tabuk, Saudi Arabia^{1,2,3,4,5}
Industrial Innovation & Robotics Center (IIRC), and Faculty of Computers and Information Technology,
Department of Computer Science at University of Tabuk, Saudi Arabia³

;

Abstract—Fossil fuels, including oil, are the most important sources of energy. They are commonly used in various forms of commercial and industrial consumption. Producing oil is a complex task that requires special management and planning. This can result in a serious problem if the oil well is not operated properly. Oil engineers must have the necessary knowledge about the well's status to perform their duties properly. This study proposes a linear regression method to predicate the oil production value. It takes into account various independent variables, such as the pressure, downhole temperature, and pressure tubing. The proposed method can accurately reach a very close prediction of the actual production value by achieving very interesting results at the end of this study.

Keywords—Big data; machine learning; oil production; regression model; features; prediction; PySpark

I. INTRODUCTION

The current rapid evolution of science and technology causes more progress in the oil field development, especially with the increasing demand for petroleum resources worldwide. For instance, the prediction of crude oil production represents the capacity of the oil field in the future, and it also directly impacts the future planning for this field [1].

In the petroleum industry, oil production and its future prediction have always been the center of interest to many scientists and researchers. This is because oil plays an impactful and effective role in energy production. Naturally, oil is produced within oil wells. In fact, many factors influence the production of the well including internal geological factors as well as geological location, time, and production equipment which are considered external factors. These influencing factors make it hard for the prediction of oil production to be precise and accurate, especially when using traditional methods like curve analysis and mathematical modeling [2]. Nonetheless, the need to accurately predict oil production is a necessity. As a matter of fact, the efficient accurate prediction can help save manpower, the consumed materials, and the resources required for the extraction process in addition to enhancing the economical aspect of the oilfields [3].

Oil production has an influential role in improving the economy. It is also a part of the political aspects of developing and established countries. Fossil fuel is considered a common industrial energy source due to the fact that it's versatile, easily transported, accessible, and expensive [4], [5]. Moreover, the

oil serves various uses in industry, agriculture, energy transformation as well as commercial and residential services as shown in Fig. 1.

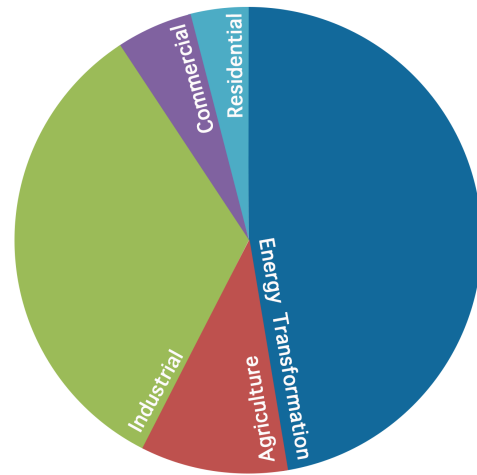


Fig. 1. Uses of fossil fuels and oil in multiple sectors: energy transformation, Agriculture, industrial, commercial, and residential

When discussing oil properties, it's important to mention that oil is found underground in porous rocks where rock strata are stacked above them creating pressure. This pressure is compensated by the pressure of the fluid (oil, gas, or water) which is in the reservoir. When drilling takes place, the fluid rises to the level of the surface because of the reservoir pressure. In addition, this lost fluid pressure is compensated by a displacement of the deeper fluid from underground to fill its place [6]. Consequently, the reservoir pressure decreases. It keeps decreasing gradually until it is no longer possible for the fluid to flow to the surface. In this case, a pump can be used to control the flow of fluid to the surface by controlling the difference between the pressure in the well and the reservoir pressure. Furthermore, the pump mode can be adjusted to have control over the bottom-hole pressure which can be a mode for setting the well's operations in the sense.

It is important to be able to predict the oil production rate or volume since it provides the engineers with useful information. As such, they can, in turn, evaluate and opti-

mize reservoir management-related issues. Unfortunately, it is extremely difficult to perform highly accurate predictions without taking into consideration the subsurface conditions [7]. On the other hand, computing technology and data analytic can help resolve this issue of oil production forecasting while maintaining the complexity of the relationships between the inputs and outputs [8].

Prediction is the process of producing estimations based on observations. Upon analyzing the observable occurrences, it is then possible to project the data to make future predictions for a similar type of occurrence. Often these observable data are too large, and no deductions or predictions can be done by simply observing them. Manual analysis and interpretation are both tiresome and time-consuming in addition to the great error margin [9]. Prediction techniques vary in their type and content as can be seen in Fig. 2, yet some of them are more advanced than others.

In oil and petroleum exploration and development, artificial intelligence (AI) has gained a remarkable spot as it can be the solution for the issues that arise in this field. Recent studies have shown that using data-driven models ensures better prediction results than experience-based prediction models [10]. It is important to mention that AI is involved in the prediction of oil and gas properties, optimization of well layout plans, and prediction of reservoir physical properties and that of oil production among others [11]. Fig. 3 illustrates the various interaction models that can be used for predictions including data-driven methods such as regression, classification, and dimensional reduction.

AI incorporates computational power and human intelligence in order to achieve reliable smart systems that can resolve extremely nonlinear and highly complex problems. Thus, AI paves the way for computers to make data-based decisions [12]. Machine learning (ML), on the other hand, is included within AI. It delivers statistical tools for big data analysis and exploration. ML is actually further subdivided into supervised, unsupervised, and reinforced learning.

Even though AI has been involved in E&P practices before, the recent expansion of digital techniques in the oil field made it increasingly involved in advanced predictive and prescriptive analytics [13].

Moreover, AI provides great potential in solving issues in almost all areas of the oil industry including prediction, classification, and clustering. The data-driven systems can locate and define the relationship between oil production and other data acquired from the field. This is achieved through using sensors and ML models. In reservoir engineering, combining ML techniques with data analytics has various benefits [14]. The studies show that this combination can help predict the bottom-hole pressure, optimize water flooding, and forecast hydrocarbon production.

ML algorithms include Support Vector Machine, K-means, decision tree, Apriori, neural networks, and Naive Bayes algorithm [11]. Linear regression is one of the basic regression techniques for forecasting, predicting, or estimating. In the past years, specifically in 1800, Gauss created a method called the Least Square Method for the purpose of suiting an equation according to linear parameters. A linear regression illustrates the existing relationship between a random factor with a

dependent factor, whose prediction or estimation is the center of focus in this study. In both research and statistical studies, linear regression is still frequently used [15]. Linear Regression as can be seen from Fig. 2 is one of the qualitative prediction methods.

Notably, the researchers of this study are highly motivated by the increasing importance of oil and fossil fuels production. Hence, the main objective of the study is to develop a system that can analyze the fed data and transform them into valuable inputs that can be used for oil production prediction in the future. The proposed system is based on Linear Regression which achieves high levels of accuracy. For a better understanding of the presented work, the developed model is evaluated to answer the following research questions:

(R.Q.1) *What factors directly affect the oil production value?*

(R.Q.2) *Can a Linear Regression model effectively predict the production value of oil based on different factors?*

This study is composed of three parts. The first part presents some related works to the prediction of oil production. These studies use different methodologies for the purpose of analyzing the data in the oil fields by presenting suitable models. Then, the methodology is the second part of this study which is explained in detail. This includes the levels of the structure of the dataset, implementations, importing the data, exploratory data analysis, data processing, data scaling, feature selection, and applying machine learning algorithms. The last part is the results part where the researchers explain that the linear regression method suggested by them has proven to be highly successful by having results that are really close to the actual results.

It is important to mention that this study has a high significance in the field of fossil fuels, especially in oil production. As such, it attempts to decrease and reduce the problems that may occur in the prediction process of oil production. The researchers highlight that this study with its simple method, linear regression, can achieve high and real results in comparison to other methods with much more complex features. Such a method can be flexible, practical, and effective for explaining and illustrating the process of successful oil predictions. It can add more knowledge and information for future research for the purpose of enhancing and saving the most significant source of energy for the upcoming generations.

II. RELATED WORK

When considering the topic of oil production, many studies come to light especially those that attempt to accurately predict the rate or volume of oil production. ML models are quite popular in this case. Nonetheless, many studies prefer to use linear regression specifically due to its many advantages.

Emeke and Bello addressed the issue of predicting the oil production volume in an area within Nigeria through linear regression techniques [16]. In their study, the authors have made use of MATLAB in analyzing the data acquired from the oil field from Kwale, Niger Delta. In addition, multivariate linear regression has been chosen as a model for its capability of housing multiple variables. For regression, the input variables are the metered volume, the basic sediment and

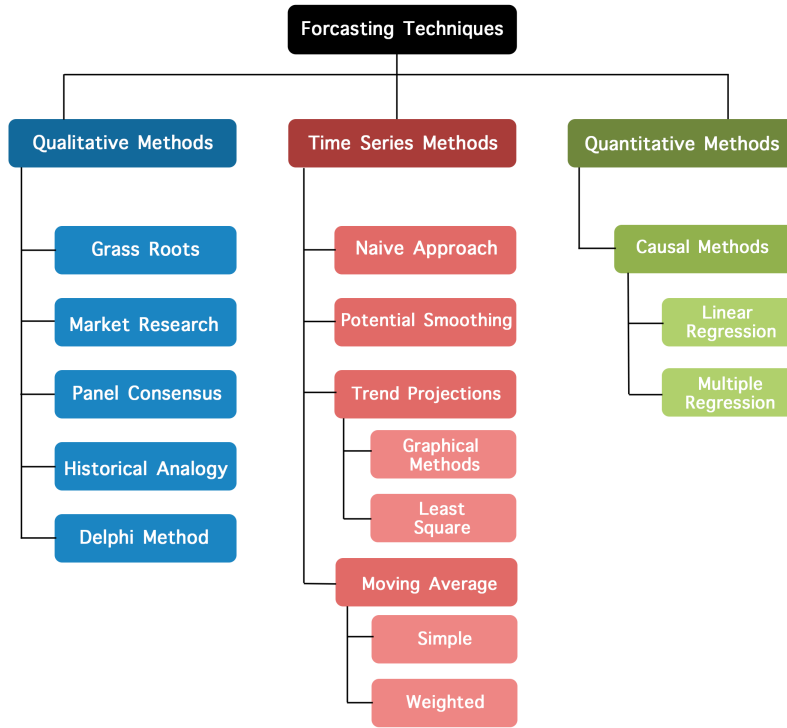


Fig. 2. Prediction traditional methods including qualitative, time series and quantitative methods

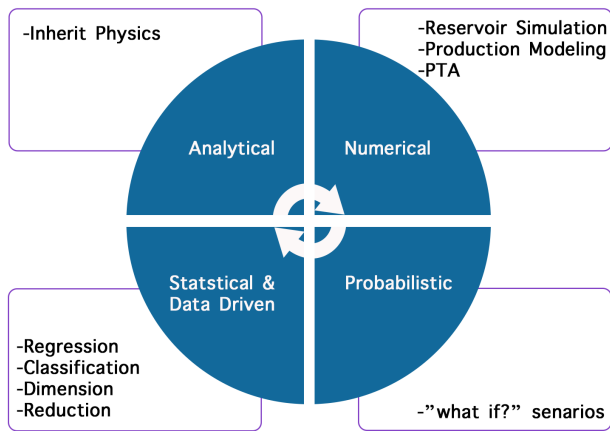


Fig. 3. Prediction tools including data-driven methods such as regression models

water, the volume correction factor, the metered factor, gross standard volume, API @60, and temperature among others. These data are collected over the course of several months, and the entire dataset is split into 70% for training the LR model and 30% for testing it. After that, a couple of models are considered (with different input variables) in contrast to the new generalized model which has been obtained by coefficient comparison and inspection of the lowest root mean square error value RMSE. One of the distinctive features of the new model is the exclusion of some factors like water percentage and basic sediment. As a result, the RMSE value for the new model is

much lower than the old one. Furthermore, the residual values are calculated to see the differences between the predicted values and the actual values, and the new generalized model, which are low, indicates a close prediction. In addition, the R-squared value is 0.9889.

In another study, Mgbemena and Chinwuko aim to predict the crude oil production value in Nigeria through implementing multiple linear regression [17]. In this study, the authors attempt to produce a daily, weekly, monthly, and yearly forecast. To do so, the authors initially analyzed many forecasting models before creating the LR model. The past historical data for the oil field are collected from the years 2010 to 2016. For linear regression, the least-squares equation is used. Overall, the performances of six models are compared. Among the graphical forecasting, 3-year simple moving average, 6-year simple moving average, 8-year simple moving average, least squares LR and the exponential smoothing with 0.1 or 0.6 as weighing factors, the Least Squares Linear Regression model have achieved the lowest mean squared error MSE value. This means that the lowest mean method makes a good prediction with minimal error. After the comparison, the authors developed their own multiple linear regression models to predict the daily oil production values. This model takes into consideration the number of natural wells which are flowing, the number of gas-lifted weak wells, and the amount unbiased compressors which are running. As a result of comparing the developed system's predicted value and the actual value, the system misses the direct hit point by 43 barrels per day, which is a very satisfactory result.

In their study, Kumar and his colleagues aim to develop

TABLE I. COMPARISON OF LITERATURE REVIEW STUDIES ACCORDING TO YEAR, AIM, AND APPROACH

Paper	Year	Aim	Approach
[16]	2019	To create a system that can accurately predict oil production in the field in Nigeria	Multivariate linear regression model has been developed taking into consideration the following factors: metered volume, gross standard volume, metered factor, and volume correction factor.
[17]	2020	To develop a model that can perform a precise forecast for oil production in Niger Delta Region	A comparison of many models has been performed, and the multiple variables regression model has been finally developed achieving very good results.
[18]	2019	To predict the oil and gas production rates based on Machine Learning	RapidMiner tool is used in addition to linear regression SET ROLE operator and APPLY MODEL operator.

a forecasting model based on machine learning to predict gas and oil production [18]. To execute the study, data has been acquired from sensors and processed, then the CRISP (Cross Industry Standard Process) model has been applied to the processed data. This involves understanding the data, preparing data, modeling, and evaluating. After that, the data is imported to Rapid Miner where correlation analysis is performed to check for existing relationships between them, the variables, and the dependent variable. Next, linear regression is applied based on the identified variables from the correlation analysis. Finally, data visualization techniques are used to visualize and discuss the data. Ultimately, six variables are chosen for input to achieve oil production and output prediction, including average wellhead temperature, average choke size, average gas lift rate, etc. Upon experimentation and testing, bottom-hole pressure has been found to be an important variable in predicting the average oil production rate. Table I presents a comparison between the literature review studies according to the year, aim, and approach.

III. METHODOLOGY

A. Dataset

The Norwegian oil and gas company Equinor has announced starting in June 2018, all of the data from the Norwegian continental shelf are available for research and study. This has given academic institutions, students, and researchers worldwide permission to use crucial scientific information in line with the Equinor Open Data License without the requirement for further written consent.

The Volve production data has been released in the form of an excel file which is made up of two (02) sheets, namely Daily Production Data and Monthly Production Data. Oil production estimation is one of the major duties of production engineers, depending on several operational parameters that include tubing differential pressure, bottom hole pressure, and wellhead pressure. This estimation process can happen as an effective way of applying empirical correlations or nodal analysis. Furthermore, physics is generally included in most of the literature work since it involves a lot of assumptions.

The researchers of this study are inspired to apply a completely different approach based on Data-Driven applications, such that physics is being taught to the computer based on data

only. This approach ensures that no complexity or assumptions are being made and that the results are purely generated from the fed data. Both linear and polynomial regression models are going to be used to predict the oil production from the date of the Volve field daily production. In the Norwegian North Sea, a hydrocarbon reservoir termed Volve has been operational for a couple of years between 2005 and 2016. Volve has a 54% recovery which is considered good. Additionally, the Volve dataset is considered the top open-source dataset for exploration and production in terms of completion.

B. Implementation

Apache Spark is a framework for data processing that can swiftly process operations on extremely large data sets and distribute processing operations over several computers, either individually or in conjunction with other distributed computing technologies. It is a super-quick machine learning and large data analytics engine. The Apache Spark community has published a tool, PySpark, to enable Python with Spark. Python programmers may work with RDDs by using PySpark.

1) *Spark MLlib*: Spark MLlib is used in Apache Spark to execute machine learning. Popular tools and algorithms may be found in MLlib. MLlib is a scalable Machine Learning library in Spark that offers both high-quality algorithms and high performance. Clustering, classification, regression, collaborative filtering, and pattern mining are examples of machine learning algorithms. Additionally, MLlib contains lower-level machine learning primitives like the general gradient descent optimization technique.

The main Machine Learning API of Spark is spark.ml. For creating ML pipelines, the library Spark.ml provides a higher-level API built on top of DataFrames. Below are some utilities for Spark MLlib:

- Pipelines
- Featurization
- ML Algorithms
- Persistence
- Utilities

2) *ML algorithms*: The core of MLlib is its ML algorithms. These include well-known learning techniques including collaborative filtering, clustering, regression, and classification. The goal of MLlib is to standardize APIs to make it simpler to integrate several algorithms into a single pipeline or workflow. The Pipelines API is one of the core ideas, and the scikit-learn project serves as an inspiration for the pipeline concept.

3) *Transformer*: An algorithm known as a Transformer may change one DataFrame into another DataFrame. Basically, a Transformer performs the transform () function, which adds one or more columns to one DataFrame to change it into another. For example: A feature transformer may take a DataFrame, read one column (for example, text), map it into another column (for example, feature vectors), and then produce a new DataFrame with the mapped column attached.

A learning model may use a DataFrame as an input, read the column containing feature vectors, predict the label for every feature vector, and then produce a new DataFrame with the predicted labels attached as a column.

4) *Estimator*: An algorithm which is known as an Estimator may be fitted to a DataFrame to create a Transformer. In terms of technical implementation, an Estimator uses the function fit(), which takes a DataFrame and outputs a Model, a Transformer. For instance, using fit() trains a Logistic Regression Model, which is a Model and subsequently a Transformer, from a learning method like Logistic Regression, which is an Estimator. Both Estimator.fit() and Transformer.transform() are stateless functions. Alternative notions may enable stateful algorithms in the future. Each Transformer or Estimator instance has a unique ID which may be used to provide parameters (discussed below).

5) *Featurization*: Selection, dimensionality reduction, feature extraction, and transformation are all parts of featurization. Feature extraction is the process of extracting features from raw data. Furthermore, the term "feature transformation" refers to the resizing, updating, or changing of features. In feature selection, a small subset of essential characteristics is chosen from a large pool of features.

6) *Pipelines*: A pipeline specifies an ML process by connecting several transformers and estimators. Additionally, it offers resources for building, assessing, and fine-tuning ML Pipelines. It is typical in machine learning to run a series of algorithms to analyze and learn from data. Such a process is represented by a pipeline in MLlib, which consists of a series of pipeline stages (Transformers and Estimators) that must be executed in a certain order. In this part, the researchers of this study utilize this simple method as an example. As such, the pipeline example below does the data preparation in the following order:

- Use the String Indexer method to determine the index of the category columns.
- Use OneHot encoding to present the category columns.
- Use the String indexer to locate the column of the output variable "label"
- VectorAssembler is used to assemble category and numerical columns. A transformer called VectorAssembler creates a single vector column from a provided list of columns.

C. Importing Data

The first step is to import the data, including three essential libraries, namely Spark which is used to import product data and to develop DataFrame, in addition to the Seaborn and Matplotlib libraries.

There are several parameters considered in the uploading of the dataset. It contains various data such as the good type being water injection or oil production, as the average downhole pressure, average drill pipe tubing, date of production, oil volume, average downhole temperature, gas volume, average choke size, average well-head temperature, average annulus pressure, average well-head pressure, water volume, and type of flow (production or injection). In addition, the data on oil production is documented based on the daily collection.

D. Exploratory Data Analysis

It is critical to perform exploratory data analysis in any machine learning approach, even though it is also time-consuming. This step is essential since upon encountering the data, it must be understood before continuing with the other steps. There are many processes included in the exploration of data, such as discovering patterns, checking assumptions, and spotting anomalies through visualizations and statistical summaries. Fig. 4 illustrates the distribution of acquired data of the seven wells.

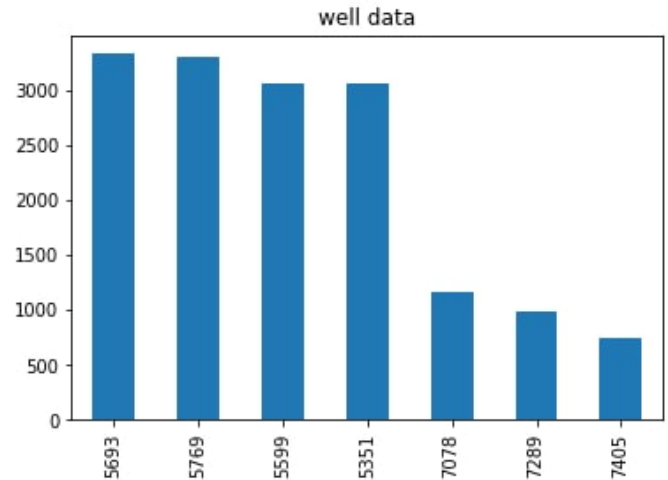


Fig. 4. Distribution of data acquired from seven wells

Fig. 5 shows the ECDF plots of each of the wells. From these plots, it can be observed that 40% of the data points relative to well number 4 have zero bore oil production. For well number 5, the observations show that 20% of their data points have zero bore oil production. These indicate that there is no flow in these wells since zero bore oil production is different from not available NA values. Thus, when the seven wells are considered in this study. Well number 4 and well number 5 data should be eliminated, as they can't be used for training. On the other hand, well number 1 shows empty values in the ECDF plot, and well number 2 has some missing values. Hence, for these two wells, the ECDF of water injection volume is plotted to check if they are injectors.

Fig. 6 shows the water injection plots for wells number 1 and number 2. By analyzing the graph, it becomes clear that well number 1 is an injector well. On the other hand, data of well number 2 indicates that it is an injector in addition to being a producer.

Generally, wells can be either injectors or producers. When fluids are poured into the underground to be placed in porous geological formations such as limestone, deep sandstone, or a shallow soil layer, then the well has termed an injection well. Additionally, the injected fluid can be either water, salt water "brine", wastewater, or chemically treated water. On the other hand, a well specified for extracting oil or gas is called a production well. Fig. 7 shows the difference in well data between the two well types.

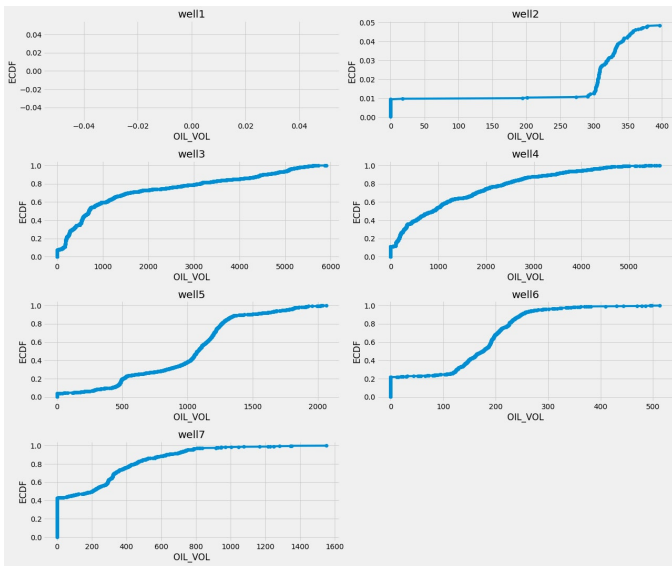


Fig. 5. ECDF plots for each of the seven wells

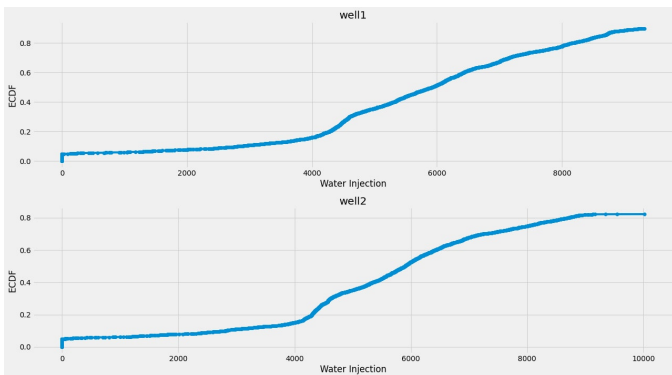


Fig. 6. ECDF plots for well number 1 and 2 referring to water injection data

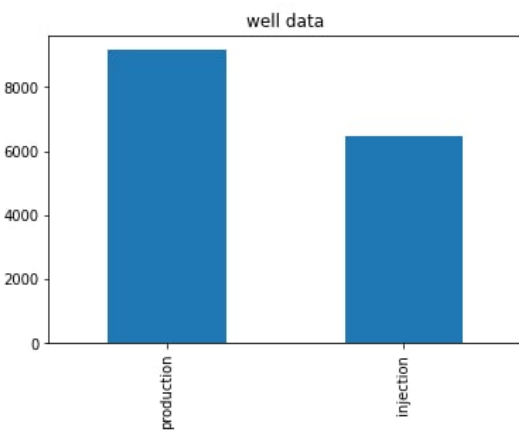


Fig. 7. Well data for a production well in comparison to an injection well

E. Data Pre-Processing

Data preparation is the next step after data analysis. Pre-processing the data involves dealing with string formats inserted in a numerical column, negative values in permeability columns,

empty columns, null values, and theoretically incorrect values.

From analyzing the data using a box plot, it is evident that the data is highly skewed depending on the well, and missing data need to be fixed. The forward filling is the technique to be performed to fill in the missing data here. However, usually, in the case of missing data, the mean value is used, yet this cannot be the case in the data of this study since it is skewed. For this reason, the null value is filled with a value just above it.

F. Data Scaling

In order to prepare the data for machine learning approaches, data scaling or normalization should be done. The scaling process is meant to alter the numeric values within the columns into a common scale while keeping the differences in the range of values and maintaining the information. The data is transformed in a way that the features are within a specific range [0, 1].

$$x' = (x - x_{min}) / (x_{max} - x_{min}) \quad (1)$$

G. Feature Selection

Feature selection is also called feature engineering since the input features of the data are examined and evaluated for their importance and impact on the results. In this way, the features, that contribute to getting the precise predicted value and are correlated to the anticipated output, are chosen either automatically or manually. The importance of this step is explained when the accuracy of the model decreases when using irrelevant features is noticed. However, using highly correlated features is also tricky since it might lead to reducing the accuracy given that the input data becomes homogenous and lacking variability. It might also lead to data leakage which increases the accuracy beyond acceptance. For this reason, correlation heat maps must be generated. The correlation heat maps are generated to determine the highly correlated features. Fig. 8 shows the variation of oil production volume with respect to downhole pressure and temperature.

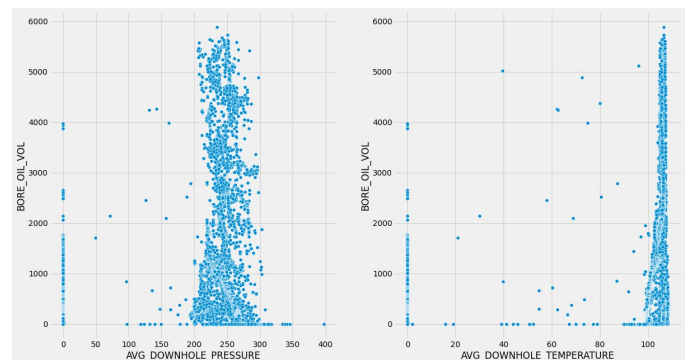


Fig. 8. Variation of oil volume with respect to downhole pressure and temperature

H. Applying Machine Learning Algorithm

Linear regression is one of the many existing regression methods, which are considered among the most important

machine learning and statistical techniques. In regression, relationships among the different variables are explored. For instance, applying regression in a company allows exploring how the salary of its employees changes according to their experience, education, their role in the company, etc. In this case, the data of each employee is a representation of one observation, and in this regression a problem, the dependent variable is the salary, whereas the independent variables are the role, education level, years of experience, etc.

In a regression analysis, a chosen phenomenon is considered with a group of observations, where each observation involves at least two features or variables. Then, an attempt of finding the relationship of dependency between the various features or variables occurs based on assumptions and further testing and analysis. Thus, regression not only finds how a phenomenon is influenced by a feature but also how the different features are related to each other. The outputs of regression are often denoted by y , whereas the inputs are denoted by x , such that the independent features can be put in a vector $x = (x_1, \dots, x_r)$ with r being the number of inputs.

Forecasting is one of the other uses of regression, where output can be predicted based on a set of predictors. Regression is used in many different fields, including economics, computer science, and the social sciences. The importance of regression increases with the increase of data volume and practical value of data.

Outputs, responses, and dependent variables are the terms used for the dependent features, whereas inputs, regressors, and predictors are the terms used for the independent features. In regression problems, there are often one continuous dependent variable and many continuous, categorical, or discrete independent variables.

Linear Regression is a simple supervised machine learning model. In this problem, the aim is to create a linear relationship between the dependent variable "Oil Production", and the other training independent variables. The linear equation will assign a coefficient to each training feature, and an intercept is added to the equation as well. Input instance –feature vector: $x = (x_0, x_1, \dots, x_n)$, and the predicted Output: $y = w_0x_0 + w_1x_1 + \dots + w_nx_n + b$, therefore, for parameters to estimate (PtoE) is represented as follow:

$$PtoE = \begin{cases} \hat{w} & := \hat{w}_0, \dots, \hat{w}_n: \text{feature } \frac{\text{weights}}{\text{model}} \text{ coefficients} \\ \hat{b} & : \text{constant bias } \frac{\text{term}}{\text{intercept}} \end{cases} \quad (2)$$

The variation of actual responses $y_i, i = 1, \dots, n$ occurs partly due to the dependence of the predictors x_i . However, there's also an additional inherent variance in the output.

The coefficient of determination R^2 resembles the degree of variation of y depending on x , through the particular regression model. If R^2 has a large value, it means that the model is highly capable of explaining the variation of the dependent variable according to the independent variables, thus it is a better fit. The value $R^2 = 1$ corresponds to $SSR = 0$.

The whole workflow of the represented system can be summed up in Fig. 9 Variables such as the downhole temperature, pressure, DP tubing, annulus press, and choke size are

considered as an input for our linear regression model, which can develop an output of predicted oil production volume based on these features.

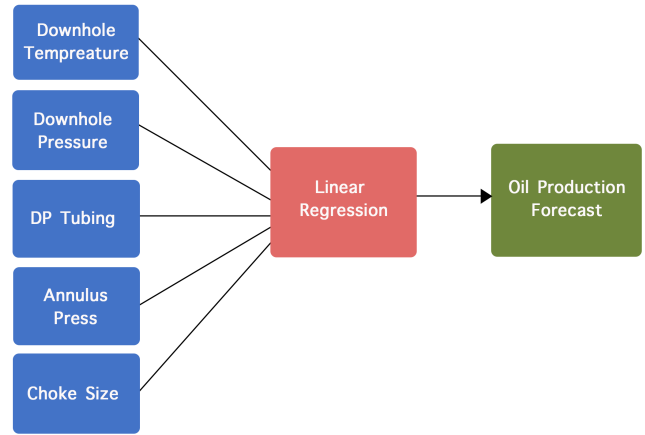


Fig. 9. Flow chart of oil production prediction system based on linear regression model

IV. RESULTS

Scatterplot and Matplotlib libraries are used to show how the model is predicting the oil production in comparison to the actual production volume (Fig. 10).

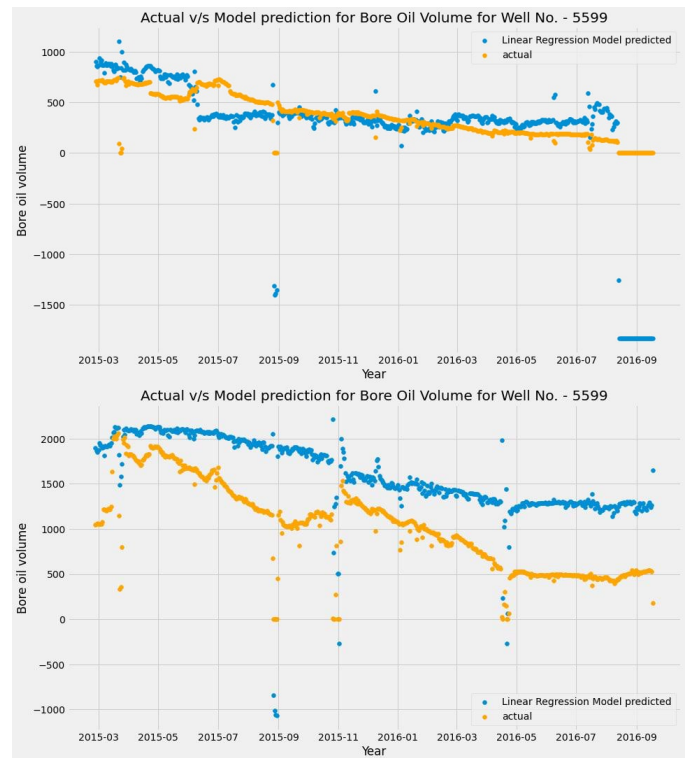


Fig. 10. Scatterplot and Matplotlib libraries using to show how the model is predicting the oil production in comparison to the actual production volume

$$RMSE = \frac{\sum_{i=1}^n (y_i - \hat{y})^2}{n} \quad (3)$$

RMSE	MSE
0.30292596179787934	0.2861434684281355

Mean Squared Error represents the average of the squared difference between the original and predicted values in the data set. It measures the variance of the residuals.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y})^2}{n}} \quad (4)$$

Root Mean Squared Error (RMSE) is the square root of Mean Squared error. It measures the standard of residuals.

Based on the above two results, it obvious that the model which is used in this study is correct and accurate because the mean absolute error that has been received after the analyses is 0 (MSE is 0.2 & RMSE is 0.3). This adds more advantages to the method used by the researchers of this study.

The results of this study can be explained by comparing the results of other studies presented in the literature review part. In study [16], the features used in the model are mostly related to volumes such as standard volume and metered volume. However, the model of this study, linear regression, takes into consideration much broader features such as temperature, pressure, and choke size that explains the system more. Moreover, in study [17], eight models have been used to make oil predictions which are very complex. Whereas, in this study, only a linear regression model has been used due to its simplicity and accuracy which have achieved really interesting results. Hence, a good prediction has been obtained in this study through only one mode. Furthermore, the researchers in this study haven't focused on each feature individually and which one is more important, but have favored getting a general model that takes all the features into account contrary to what Kumar and his colleagues have done in study [18].

V. CONCLUSION

Today, fossil fuels, such as oil, are the most important sources of energy. They are commonly used in various industrial and commercial sectors. However, their production is complex and requires special management and planning.

It is crucial that oil engineers are aware of the status of their wells and are able to perform their duties properly. In this study, they proposed a linear regression method to estimate the oil production value. This method can be used to analyze the various independent variables that affect the oil production process.

The proposed method would be able to accurately predict the oil production value. It can also achieve interesting results by analyzing the data collected during the study.

REFERENCES

[1] H. Hu, Y. Pu, and X. Guan, "Oil field crude oil production level prediction method based on ahp-pso-bp," in *2020 IEEE 8th International Conference on Information, Communication and Networks (ICIN)*, 2020, pp. 214–218.

[2] W. Liu, W. D. Liu, and J. Gu, "Forecasting oil production using ensemble empirical model decomposition based long short-term memory neural network," *Journal of Petroleum Science and Engineering*, vol. 189, p. 107013, 2020. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S092041052030108X>

[3] S. Pan, J. Wang, and W. Zhou, "Prediction on production of oil well with attention-cnn-lstm," *Journal of Physics: Conference Series*, vol. 2030, no. 1, p. 012038, sep 2021. [Online]. Available: <https://dx.doi.org/10.1088/1742-6596/2030/1/012038>

[4] K. I. Wong and P. K. Wong, "Optimal calibration of variable biofuel blend dual-injection engines using sparse bayesian extreme learning machine and metaheuristic optimization," *Energy Conversion and Management*, vol. 148, pp. 1170–1178, 2017. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0196890417306088>

[5] Y. Xing, Z. Zheng, Y. Sun, and M. Agha Alikhani, "A review on machine learning application in biodiesel production studies," *International Journal of Chemical Engineering*, vol. 2021, 2021. [Online]. Available: <https://doi.org/10.1155/2021/2154258>

[6] A. Davtyan, A. Rodin, I. Muchnik, and A. Romashkin, "Oil production forecast models based on sliding window regression," *Journal of Petroleum Science and Engineering*, vol. 195, p. 107916, 2020. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0920410520309712>

[7] X. Ma and Z. Liu, "Predicting the oil production using the novel multivariate nonlinear model based on arps decline model and kernel method," *Neural Computing and Applications*, vol. 29, no. 2, pp. 579–591, 2018. [Online]. Available: <https://doi.org/10.1007/s00521-016-2721-x>

[8] S. D. Mohaghegh, "Subsurface analytics: Contribution of artificial intelligence and machine learning to reservoir engineering, reservoir modeling, and reservoir management," 2020.

[9] M. Rajesh *et al.*, "Price prediction for pre-owned cars using ensemble machine learning techniques," *Recent Trends in Intensive Computing*, vol. 39, p. 178, 2021.

[10] I. Makhotin, D. Koroteev, and E. Burnaev, "Gradient boosting to boost the efficiency of hydraulic fracturing," *Journal of Petroleum Exploration and Production Technology*, vol. 9, no. 3, pp. 1919–1925, 2019.

[11] S. Qin, J. Liu, X. Yang, Y. Li, L. Zhang, and Z. Liu, "Predicting heavy oil production by hybrid data-driven intelligent models," *Mathematical Problems in Engineering*, vol. 2021, 2021.

[12] Z. Tariq, M. S. Aljawad, A. Hasan, M. Murtaza, E. Mohammed, A. El-Husseiny, S. A. Alarifi, M. Mahmoud, and A. Abdurraheem, "A systematic review of data science and machine learning applications to the oil and gas industry," *Journal of Petroleum Exploration and Production Technology*, vol. 11, no. 12, pp. 4339–4374, 2021.

[13] M. Maucec and S. Garni, "Application of automated machine learning for multi-variate prediction of well production," in *SPE Middle East Oil and Gas Show and Conference*. OnePetro, 2019.

[14] C. S. W. Ng, A. Jahanbani Ghahfarokhi, and M. Nait Amar, "Application of nature-inspired algorithms and artificial neural network in water-flooding well control optimization," *Journal of Petroleum Exploration and Production Technology*, vol. 11, no. 7, pp. 3103–3127, 2021.

[15] F. Salim and N. A. Abu, "Used car price estimation: Moving from linear regression towards a new s-curve model," *International Journal of Business and Society*, vol. 22, no. 3, pp. 1174–1187, 2021.

[16] K. B. C. Emeke, "A novel model developed for forecasting oilfield production using multivariate linear regression method," *Journal of Science and Technology Research*, vol. 29, no. 2, pp. 579–591, 2019.

[17] C. Mgbemena and E. a. Chinwuko, "Forecast of crude oil production output in an oil field in the niger delta region of nigeria," *International Journal of Industrial Engineering & Production Research*, vol. 31, no. 1, 2020. [Online]. Available: <http://ijiepr.iust.ac.ir/article-1-891-en.html>

[18] A. K. K., R. Ramasree, and M. Faisal, "Performing predictive analysis using machine learning on the information retrieved from production data of oil & gas upstream segment," in *2019 International Conference on Communication and Signal Processing (ICCSPP)*, 2019, pp. 0385–0391.