NIPES

Journal of Science and Technology Research

Journal homepage: www.nipesjournals.org.ng

Bottom-Hole Pressure Prediction from Wellhead Data Using Developed Machine Learning Models

Otamere Blessing[a], Igbinere Sunday Agbons[b]
blessing.otamere@uniben.edu[a] and agbons.igbinere@uniben.edu[b]
Department of Petroleum Engineering, University of Benin, Benin City, P.M.B 1154, Edo State, Nigeria

Article Info

***Keywords*:**
Bottom-hole Pressure, Random Forest Regression, Gradient Boosting Regression, Modeling, Mean Absolute Error, Mean Squared Error, Reservoir, Petroleum Industry, Downhole gauge, Normalization.

Received 04 August 2021
Revised 24 August 2021
Accepted 29 August 2021
Available online 04 September 2021

https://doi.org/10.37933/nipes/3.3.2021.24

https://nipesjournals.org.ng

Abstract

The accurate prediction or measurement of bottom-hole pressures in oil and gas reservoirs cannot be over-emphasized in the Petroleum Industry. Mechanistic, numerical and analytical models have been developed and deployed to determine bottom-hole pressure. Some of these models developed have failed to predict bottom-hole pressures to an acceptable accuracy. However, the down-hole gauges measure the bottom-hole pressures to an acceptable accuracy, but, are expensive to use and maintain. This study focused on developing models using random forest regression and gradient boosting regression to predict bottom-hole pressures in oil and gas reservoirs accurately. The input data used was collected from the Volve Field (Jurassic sandstone reservoir) and filtered and correlated successively. The data was normalized using Python programming to prepare the data sets for input into the model. The results showed that the random forest regression model has an accuracy of 97.80% while the gradient boosting regression model has an accuracy of 95.83%. The average magnitudes of the errors are 0.0067 and 0.01266 for random forest and gradient boosting regression models respectively. The developed models predicted the bottom-hole pressures for the reservoir with an acceptable degree of accuracy and error magnitude. The Random Forest Regression and the Gradient Boosting Regression models were seen to be economical and accurate in solving the problem of predicting bottom-hole pressures in oil and gas reservoirs.

1.0 Introduction

In the upstream sector of oil and gas industry, the accurate determination of bottom-hole pressures facilitates the determination of the productivity of wells and adequate management of the well production system can be achieved. With the increased use of permanent down-hole gauges, measuring bottom-hole pressure (BHP) gets faster. However, these down-hole gauges require continuous maintenance and calibration which are very costly to carry out. Also, by engaging in well intervening operations to measure BHP is an expensive task, associated with production risk and interruption. For these reasons, the motivation of the prediction of BHP has been argued. The bottom-hole pressure (BHP) is the pressure acting on the walls of the hole. In large diameters, this pressure has limited impacts on the wellbore, but in the case of smaller diameters, it can generate down-hole problem such as total circulating loss [1]. When the wall is static, the bottom-hole pressure can equal to the hydrostatic pressure generated by the column of the drilling fluids. During the circulation, the bottom-hole pressure equals

to the sum of hydrostatic pressure and frictions generated through the circulating system [2]. There is nothing more important in petroleum engineering than a definite knowledge of the pressure at the bottom of an oil well at any existing operating condition and the relationship between this pressure and the pressure within the producing formation. Knowledge of bottom-hole pressures is fundamental in determining the most efficient methods of recovery and the most efficient lifting procedure, yet there is less information about these pressures than about any other part of the general problem of producing oil [1].

Models have been developed to determine BHP directly from surface readings using multiphase correlations or mechanistic methods [3]. Technological advancements have contributed to the accurate predictions of BHP data because machine learning algorithms have been developed to better determine BHP data from surface measurements [3]. Multi-phase correlations analysis is the key to determining bottom-hole pressure from wellhead data because the pressure gradient of the multiphase flow pattern is obtained for a particular length of tubing, but, the prediction is not a single estimation of pressure gradient; the flow pattern has to be actively considered [3]. There are several multiphase correlations, mechanistic models and machine learning algorithms that have been used to predict bottom-hole pressure in multiphase flow [4, 5, 6, 7, and 8]. However, their general applicability is questionable. Correlations that address only a specific class of problems exist and these types of correlation usually perform better than those which attempt to solve a variety of problems. Usually, the higher the number of variables in the model the lesser the reliability and general applicability of the correlations and with advances in drilling and completions operations, complex completion design and various wellbore trajectories which result in different pipe configuration and changing inclination, the multiphase correlation is affected. Modern advancements in pressure prediction methods revealed that most of the mechanistic models produced with lesser accuracy and adjustments are still needed. Machine learning models such as artificial neural network (ANN) and regression techniques such as linear regression have been of great help in this area with an acceptable degree of accuracy.

Regression analysis is a form of predictive modeling technique which investigates the relationship between a dependent (target) and independent (predictor) variables. This technique is used for forecasting, time series modeling and finding the causal effect relationship between the variables [9]. For example, relationship between rash driving and number of road accidents by a driver is best studied through regression. Regression analysis is an important tool for modeling and analyzing data. Here, we fit a curve or line to the data points, in such a manner that the differences between the distances of data points from the curve or line is minimized [10]. Gradient Boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees [11]. Also, gradient boosting machines or simply, GBMs, the learning procedure consecutively fits new models to provide a more accurate estimate of the response variable. The principle idea behind this algorithm is to construct the new base-learners to be maximally correlated with the negative gradient of the loss function, associated with the whole ensemble [11]. The objective of any supervised learning algorithm is to define a loss function and minimize it. Equation 1 can be used to develop Gradient Boosting algorithm by determining the Mean Squared Error (MSE) as loss which is defined as;

$$\text{Loss} = \text{MSE} = \sum_{i=1}^{n}\left(y_i - y_i^p\right)^2 \quad (1)$$

Where,
$y_i = ith\ Target\ Value,\ \ y_i^p = ith\ Prediction,\ \ L(y_i y_i^p) = Loss\ Function$

The target is to have minimum loss function (MSE). By using gradient descent and updating our predictions based on a learning rate, we can find the minimum value of MSE. Boosting model is an ensemble technique in which the predictors are not made independently, but sequentially. This technique employs the logic in which the subsequent predictors learn from the mistakes of the previous predictors. Therefore, the observations have an unequal probability of appearing in subsequent models and ones with the highest error appear most. So the observations are not chosen based on the bootstrap process, but based on the error. The predictors can be chosen from a range of models like decision trees, regressors, classifiers etc. Because new predictors are learning from mistakes committed by previous predictors, it takes less time or iterations to get close to actual predictions. But we have to choose the stopping criteria carefully or it could lead to over fitting on training data. Gradient boosting is an example of boosting algorithm.

Random forest is a machine learning technique which uses ensemble learning method for classification and regression. Random forest is a bagging technique and not a boosting technique. The trees in random forests are run in parallel. There is no interaction between these trees while building the trees. It operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. A random forest is a meta-estimator (i.e. it combines the result of multiple predictions) which aggregates many decision trees, with some helpful modifications [12]. The data used in this study were obtained from the *Volve Field*, located in the central part of the North Sea, Norway. The reservoir contains oil in a combined stratigraphic and structural trap in Jurassic sandstones in the Hugin Formation.

2.0 Methodology

The method adopted to achieve the set objectives of this study involved data collection from reservoirs located in *Volve Field*. These data were normalized for adequate application in the Python programming language because, from its raw state, it falls short of what is required to achieve the desired modeling and prediction of bottom-hole pressures. The Anaconda software which has the Python programming language was used to carry out the coding on the Jupyter environment. Python is a high level programming language which emphasizes code readability. The Python was used to develop Random Forest Regression and Gradient Boosting Regression models for the prediction of bottom-hole pressure (BHP). The Microsoft Excel tool was also used for data filtering and preparation before using in the Python Programming. The number of datasets used for the models are same. It consists of 3522 rows and 11 columns after normalization of the data sets for use. The features of the columns of datasets that were used are as follows; well depth, average down-hole pressure, average down-hole temperature, average annulus pressure, average choke size, average well-head pressure, average well-head temperature, pressure differential in chokes, bore gas volume, bore oil volume and bore water volume.

2.1 Random Forest Regression Algorithm

The following sequences were adopted to run the datasets in the random forest regression algorithm;

a) Initialize the random forest regression model.
b) Fit the training data set of both the input variables and the output variables to the random forest regression model.
c) Train the model with the training set data using the random forest regression model.
d) Evaluate the model by fitting the test data set to the model for the bottom hole prediction process.
e) Compute the accuracy of the prediction model, the mean absolute error (MAE) and the mean squared error (MSE).

$$Accuracy = \frac{Input\ of\ Test\ Data\ Set}{Output\ of\ Test\ Data\ Set} * 100 \quad (2)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - y_{avg}| \quad (3)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(|y_i - y_{avg}|\right)^2 \quad (4)$$

2.2 Gradient Boosting Regression Algorithm

The following sequences were adopted to run the datasets in the gradient boosting regression algorithm;

a) Initialize the gradient boosting regression model.
b) Fit the training data set of both the top features/input variables and the target/output variables to the gradient boosting regression model.
c) Train the model with the training set data using the gradient boosting regression model.
d) Evaluate the model by fitting the test data set to the model for the bottom hole prediction process.
e) Compute the accuracy of the prediction model, the mean absolute error (MAE) and the mean squared error (MSE).

$$Accuracy = \frac{Input\ of\ Test\ Data\ Set}{Output\ of\ Test\ Data\ Set} * 100 \quad (5)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - y_{avg}| \quad (6)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(|y_i - y_{avg}|\right)^2 \quad (7)$$

3.0 Results and Discussions

The results presented in Table 1 showed the full datasets collected from reservoirs in *Volve Field* transformed in a descriptive statistical summary generated in excel, which include the central tendency, dispersion, percentiles, and standard deviation as shown in the Figure 1 to Figure 16 These statistical summary data gives the general statistical description of all the available data which is presented in the appendix and appeared too cumbersome or bulky to appreciate in the usage for the models developed and to also appreciate the extent of the normalization of these data. A closer look at the Table 1 from the first column which captured the well depth, counting from the serial number 3522, you have the mean and standard deviation of all the data in that column to be 5978.644 and 768.173 respectively. While the first and second quantiles of 25%

and 50% are 5351 and 5599, the minimum and maximum values on that column are 5351 and 7078. The minimum values is seen to be zeros (0) in the average down-hole pressure and temperature, which are also recorded in average choke size, average wellhead pressure and temperature, as well as the average annulus pressure. As earlier stated Table 1 summarizes what is presented in Appendix A and also shows the statistical behavior of the datasets.

Table 1 Descriptive statistical summary

	Well Depth	Average Down-hole Pressure	Average Down-hole Temp.	Average Annulus pressure	Average Choke Size	Average Wellhead Pressure	Average Wellhead Temp.	Differential Pressure in Choke
Count	3522	3516	3516	3509	3493	3522	3522	3522
Mean	5978.644	251.264	103.403	18.084	60.922	48.681	77.288	20.113
Std	768.173	22.400	5.2998	5.978	35.680	22.754	15.813	20.982
Min	5351	0	0	0	0	0	0	0.013
25%	5351	238.329	100.018	13.964	24.882	31.092	74.800	2.773
50%	5599	253.1395	105.585	19.662	62.6692	39.976	81.675	10.9215
75%	7078	265.652	106.376	22.217	100	64.5572	87.42	32.5502
Max	7078	334.656	107.508	30.02	100	120.889	92.071	111.525

3.2 Data Filtering and Normalization

The data were distributed in a way that does not encourage the application of Python programing language algorithm and it is important to understand how the variables in these data are distributed. The univariate distribution of each feature is plotted as a histogram and fitted with a kernel density estimate (KDE) as shown in Figure 1 to 16. These normalizations revealed how the data are either skewed to a particular direction or abnormal in distribution. KDE were adopted to represent these data in density probability functions and it is referred to as the non-parametric way to estimate the probability density function of the variables, most of which in this case have distributions which are not normal, as some are bimodal and others skewed to the left. Hence, there is a need for the data to be normalized using data transformation, to enable the model to perform better. The Figures 1 to 16 showed the probability plots of the datasets and the normalized distribution of the datasets. These data make up the buildup data for the random forest regression and the gradient boosting regression models.

3.2.1 Normalization of Well Depth

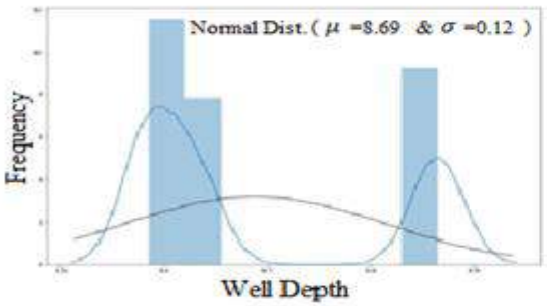

Figure 1: Normalized Data for Well Depth

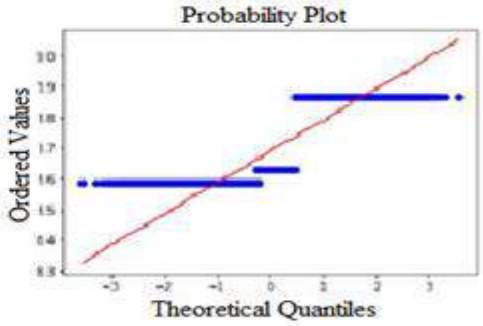

Figure 2: Probability Distribution of Well Depth

3.2.2 Normalization of Down-hole Pressure

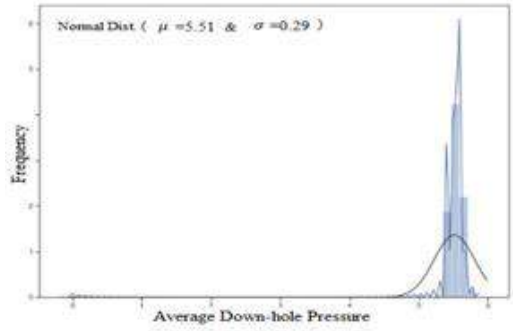

Figure 3: Normalized Data for Down-hole Pressure

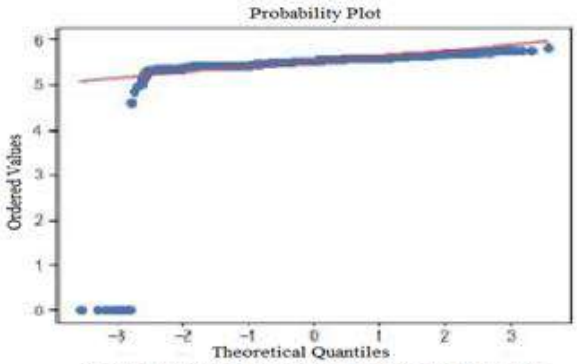

Figure 4: Probabilit Distribution of Down-hole Pressure

3.2.3 Normalization of Down-hole Temperature

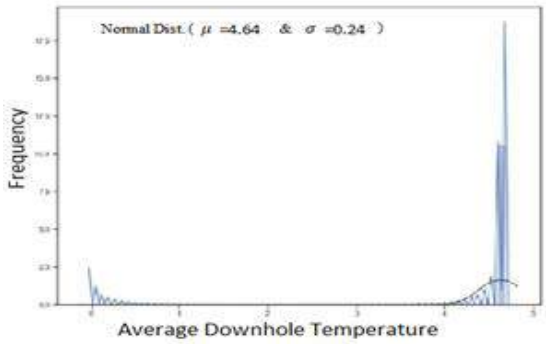

Figure 5: Normalized Data of Downhole Temperature

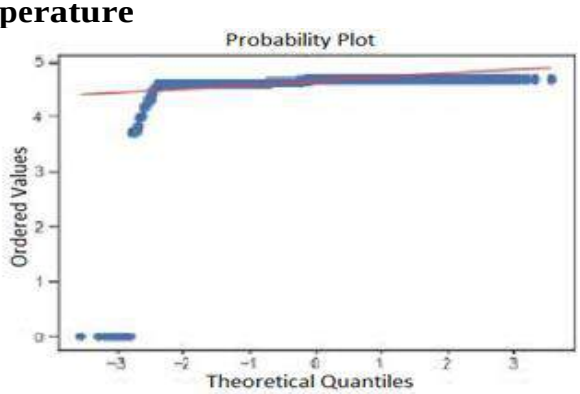

Figure 6: Probability Distribution for Downhole Temperature

3.2.4 Normalization of Annulus Pressure

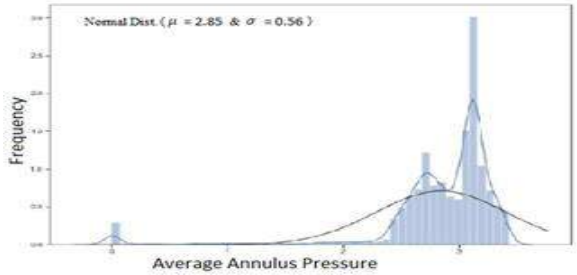

Figure 7: Normalized Data of Annulus Pressure

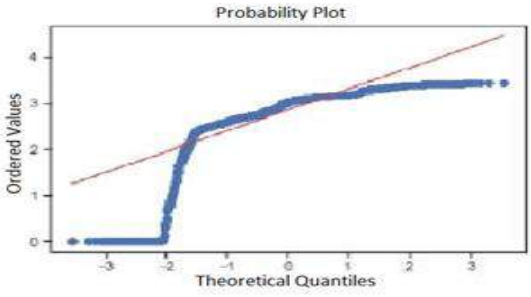

Figure 8: Probability Distribution for Annulus Pressure

3.2.5 Normalization of Choke Size

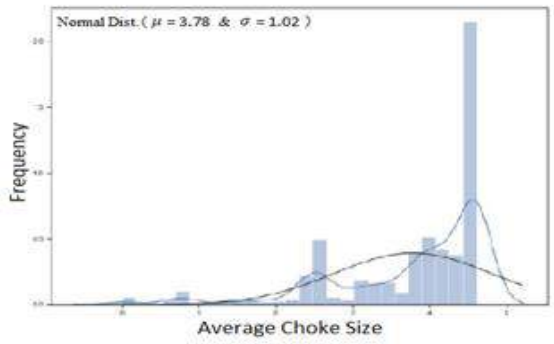

Figure 9: Normalized Data of Choke Size

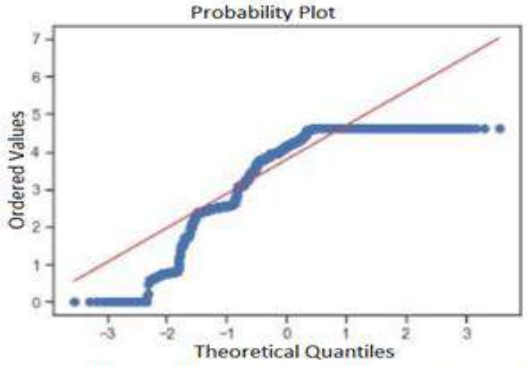

Figure 10: Probability Distribution for Choke Size

3.2.6 Normalization of Wellhead Pressure

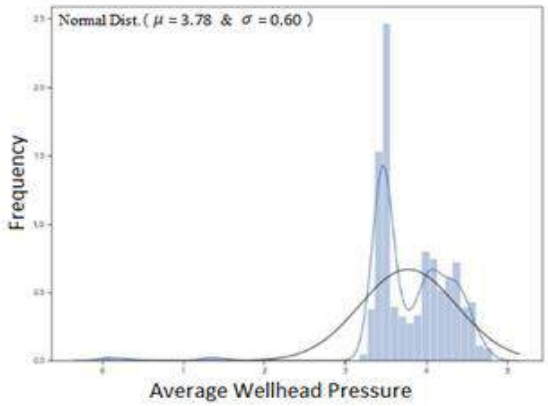

Figure 11: Normalized Data of Wellhead Pressure

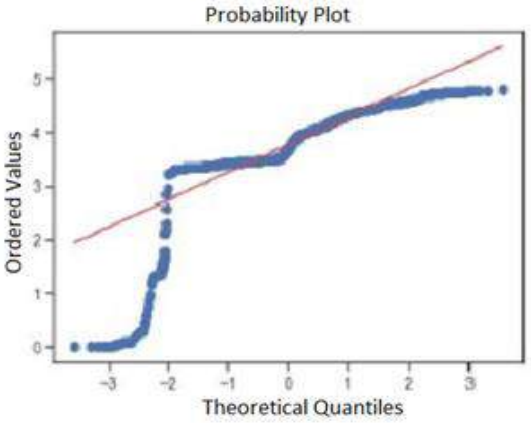

Figure 12: Probability Distribution for Wellhead Pressure

3.2.7 Normalization of Wellhead Temperature

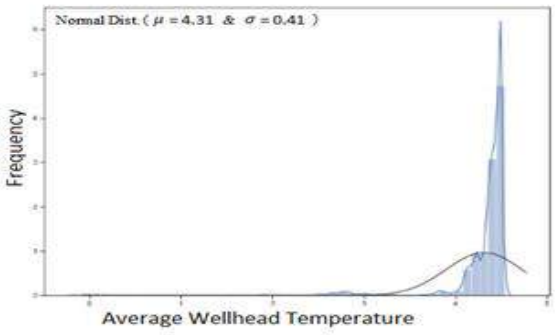

Figure 13: Normalized Data of Wellhead Temperature

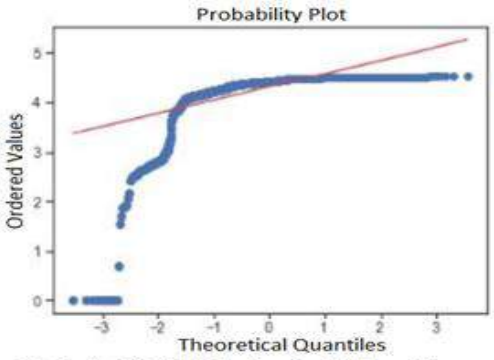

Figure 14: Probability Distribution for Wellhead Temperature

3.2.8 Normalization of Differential Pressure in Chokes

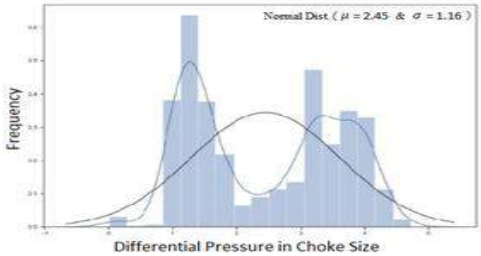

Figure 15: Normalized Data of Differential Pressure in Choke Size

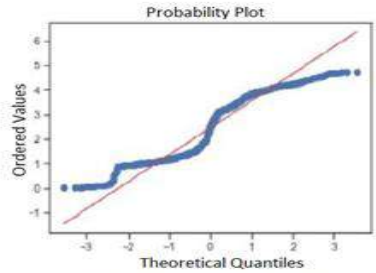

Figure 16: Probability Distribution for Differential Pressure in Choke Size

3.3 The Reservoir Data Relationship

Heat map visualization technique helped in the analysis of relationship of variables as shown in Figure 17, it showed the colour ranges from -1 to +1. For better understanding of the relationship between the variables, with a high observation on the target variable, it is important that the correlation heat map be visualized. In this research, it helped to understand the relationship of all features with the average down-hole pressure, as there are negative and positive correlations with the target variable. These negative and positive correlations indicate an increase to increase relationship, for example, the average down-hole pressure relationship with the average choke size pressure is 0.11, an increase in the average down-hole pressure resulted in an increase in the average choke size pressure. While an increase to a decrease relationship between the average down-hole pressure and average annulus pressure is -0.19, meaning they both have an inverse relationship. A closer look at the heat map showed all the variables relationship either in the increasing order or decreasing manner. Perfectly correlated variables gave a unity value indicating an over-fitting data while the non-correlated variable gave zero value, the other positive or negative values as seen in Figure 17 were the areas of interest applied in the building of the random forest regression and gradient boosting regression models. The axis AB in the vertical and horizontal direction is the well depth, similarly, BC is the average down-hole pressure, CD is the average down-hole temperature, DE is the average annulus pressure, EF is the average choke size pressure, FG is the average wellhead pressure, GH is the average wellhead temperature and HI is the differential pressure in choke.

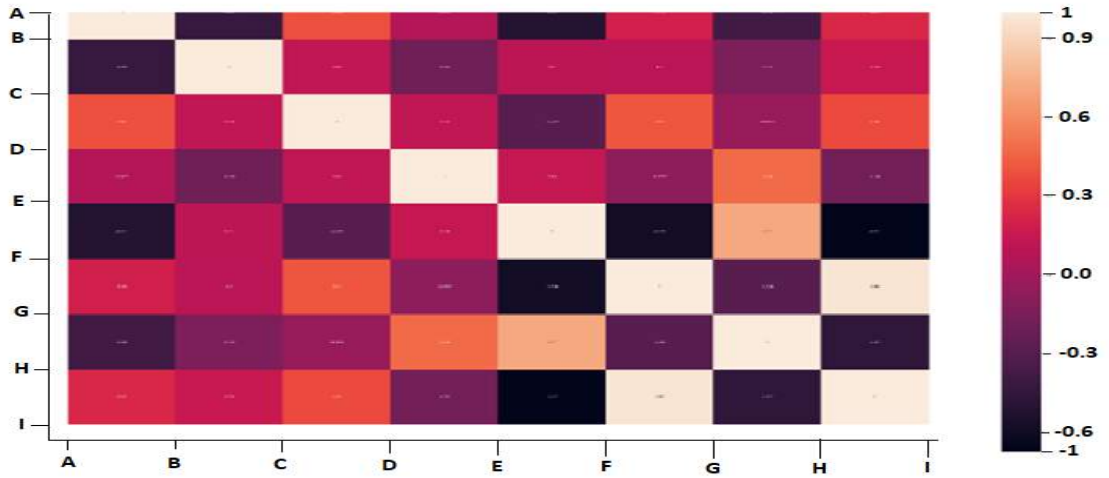

Figure 17: The Heat Map Variable Relationship of the Dataset Features

3.4 Random Forest Regression and Gradient Boosting Regression Results

The models predicted future values of the average down-hole pressure and the results were compared with the real value of the average down-hole pressure.

Table 2: Real Values of the Bottom-hole Pressure and Predicted Bottom-hole Pressure

S/N	Real Values (Psi)	Random Forest Regression (Psi)	Gradient Boosting Regression (Psi)
1	273.94693	273.8574452	274.509041
2	316.01419	315.9951645	316.003403
3	317.55131	317.7651927	317.485427
4	277.79768	277.8411459	277.822714
5	273.59437	273.4923454	273.278095
6	272.29849	272.1972860	272.139310

7	271.56095	271.4266188	272.145176
8	271.07159	271.4393995	271.794827
9	270.69052	270.6281025	271.242867
10	270.24503	270.2176104	270.047943
11	269.99314	269.9569647	270.442691

Table 3: Accuracy, MAE and MSE of the Models

In Percent values	Random Forest regression	Gradient Boosting regression
Accuracy	97.79909589	95.83176645
Mean Absolute Error (MAE)	0.00671815	0.012661687
Mean Squared Error (MSE)	0.001122985	0.002126792

The results show that the models gave an accuracy of above 95percent which is a high degree of accuracy for use. It shows that any of the models can be used for bottom-hole pressure prediction. From the results compared in Table 3, the Random Forest Regression model gives the highest accuracy of bottom-hole pressure prediction with least Mean Absolute and Mean Squared Error

4.0 Conclusion

Based on the results obtained from the procedure outlined in the methodology section of this work the following conclusions can be drawn:

a. The datasets affecting the bottom-hole pressure were filtered and successfully correlated as shown on the heat map.
b. A Random Forest Regression model and Gradient Boosting Regression model were successfully developed.
c. The bottom-hole pressures of the *Volve field* were successfully predicted using the Random Forest Regression model and Gradient Boosting Regression model and produced a high degree of accuracy of 95 % and above. Indicating that both models can be effectively deployed in predicting bottom-hole pressure having input data similar to that applied in this study.
d. The Random Forest regression model was seen to be the model that performed better than the Gradient Boosting Regression model with a higher accuracy 97.8% for the bottom-hole prediction for the reservoir and having a 0.67% and 0.11% for mean absolute error and mean squared error respectively.

5.0 References

[1] Choh, S.-J., Milliken, K.L. and McBride, E.F. (2003) A tutorial for sandstone petrology: architecture and development of an interactive program for teaching highly visual material. Comput. Geosci. 29, pp 1127–1135.
[2] Bjørlykke, K., Jahren, J., 2010. Sandstones and sandstone reservoirs. In: Petroleum Geoscience. Springer, Berlin, Heidelberg, pp. 113–140.
[3] Akinsete, O. and Adesiji B. A. (2019) Bottom-Hole Estimation from Wellhead Data Using Artificial Neural Network. SPE-198762-MS presented at the Nigeria Annual International Conference and Exhibition, Lagos Nigeria 5-7 August.
[4] Baxendall, P.B. and Thomas, R. (1961), "The Calculation of Pressure Gradients in High rate Flowing wells", Journal of Petroleum Technology, SPE-2-PA, 13 (10) pp. 1023-1028

[5] Fancher, G. H., and Brown, K. E. (1963). Prediction of Pressure gradients for Multiphase Flow in Tubing, Society of Petroleum Engineering Journal paper presented at the 37th Annual Fall Meeting of SPE, October 7-10
[6] Hagedon, A. R and Brown, K. E. (1965). Experimental Study of Pressure Gradients occurring during continuous two-phase flow in small diameter vertical conduits. Journal of Petroleum Technology, p.475-484.
[7] Duns, H., And Ros. N. C. J. (1963). Vertical flow of Gas and Liquid Mixtures in wells, Proc. of the Sixth World Congress,Vol.10, Section 2, Paper 22.PD6. 451465.
[8] Beggs, H. D., and Brill, J. P. (1973). A Study of Two-Phase Flow in Inclined Pipes, Journal of Petroleum Technology, Vol. 25, No. 5, pp. 607-617.
[9] Nguyen, B.T.T., Jones, S.J., Goulty, N.R., Middleton, A.J., Grant, N., Ferguson, A., Bowen, L., (2013). The Role of Fluid Pressure and Diagenetic Cements for Porosity Preservation. AAPG Bulletin Vol. 97 (8) pp. 1273-1302
[10] Bjørlykke, K., (2014). Relationships Between Depositional Environments, Burial History and Rock Properties. Some Principal Aspects of Diagenetic Process in Sedimentary Basins. Sedimentary Geology, Elsevier, Vol. 301 Issue 1-2, pp. 1-14
[11] Natekin, A. and Knoll, A. (2013). Gradient Boosting Machiness, A Tutorial. Frontiers in Neurorobotics, Volume 7, Article 21, pp 1- 21. doi:10.3389/fnbot.2013.00021 www.frontiersin.org
[12] Liaw, A. and Wiener, M. (2002). Classification and Regression by Random Forest. Vol. 2 (3), pp. 18 – 22. www.researchgate.net/publication/228451484

Appendix

Table A1: Input Datasets

S/N	Well Depth	Average Downhole Pressure	Average Downhole Temp	Average Annulus pressure	Average Choke Size	Average Wellhead Pressure	Average Wellhead Temp	Differential Pressure in Choke size	Bore Oil Volume	Bore Gas Volume	Bore Water Volume
1	7078	273.947	105.551	21.55	2.5408	94.565	55.959	66.404	190	29,120	0
2	7078	316.014	102.196	0	0.0036	0.849	18.786	0.449	0	0	0
3	7078	317.551	101.74	0	0	0.819	16.437	0.348	0	0	0
4	7078	277.798	104.933	1.653	6.11618	96.496	41.019	68.441	590	88,733	0
5	7078	273.594	105.44	17.309	9.95129	96.201	52.455	67.944	1,066	161,227	0
6	7078	272.298	105.538	24.75	9.75875	95.512	55.184	67.258	1,060	160,270	0
7	7078	271.561	105.585	28.259	9.88211	95.042	57.214	66.797	1,070	160,951	0
8	7078	271.072	105.614	22.087	9.78896	94.743	58.377	66.511	1,070	160,232	0
9	7078	270.691	105.64	22.075	9.78433	94.516	58.627	66.283	1,062	159,484	0
10	7078	270.245	105.667	23.186	9.91491	94.298	57.726	66.063	1,074	162,197	0
11	7078	269.993	105.695	24.284	9.8808	94.06	58.426	65.84	1,070	161,999	0
12	7078	269.958	105.718	25.182	9.88084	93.969	58.197	65.426	1,054	160,095	0
13	7078	269.873	105.742	26.022	9.84737	93.86	58.751	65.156	1,039	159,050	0
14	7078	269.561	105.776	26.623	9.86483	93.555	58.551	65.221	1,051	160,116	0
15	7078	269.412	105.803	27.543	9.7961	93.228	58.849	64.991	1,045	158,507	0
16	7078	269.362	105.824	27.969	9.83735	93.079	57.457	64.819	1,045	158,972	0
17	7078	267.613	105.886	23.733	9.91057	91.851	58.585	63.583	1,076	162,410	0
18	7078	266.957	105.925	21.241	10.09222	91.445	57.518	63.179	1,089	165,357	0
19	7078	268.022	105.907	20.317	9.78648	91.937	56.503	63.668	1,042	158,398	0
20	7078	266.027	105.982	23.008	9.86604	90.752	60.084	62.49	1,047	158,557	59

21	7078	266.56	105.978	23.037	9.81311	91.007	59.501	62.765	1,024	156,092	59
22	7078	266.43	105.995	23.65	10.10168	90.856	59.891	62.611	1,050	160,103	60
3218	5351	298.614	98.06	0.061	1.2028	0.071	0	0.126	0	0	0
3219	5351	298.925	98.04	0.062	1.11962	0.073	0	0.154	0	0	0
3220	5351	299.549	98.018	0.058	1.28649	0.599	0	0.337	0	0	0
3221	5351	302.878	90.783	0.136	1.19928	16.28	0	15.997	0	0	0
3222	5351	334.656	78.804	0	1.18873	72.136	0	71.816	0	0	0