International Journal of Intelligent Networks 3 (2022) 87–101

Contents lists available at ScienceDirect

International Journal of Intelligent Networks

journal homepage: www.keaipublishing.com/en/journals/international-journal-of-intelligent-networks

Application of machine learning techniques for identifying productive zones in unconventional reservoir

Amir Gharavi *, Mohamed Hassan, Jebraeel Gholinezhad, Hesam Ghoochaninejad, Hossein Barati, James Buick, Karrar A. Abbas

University of Southampton, School of Chemical Engineering and Chemistry, Southampton, University Road, Highfield Campus, Southampton, SO17 1BJ, United Kingdom

ARTICLE INFO

Keywords:
Machine learning
Quick analyser
Exploratory data analysis
Feature importance
Hyperparameter tuning
Feature engineering
Unconventional resources

ABSTRACT

Unconventional reservoirs are the productive zones in other words the rock quality and the mechanical properties of the rocks this process is devastating if humans or people try to search for the best reservoirs. So we can use machine learning (ML) algorithms to help us find and search easily and fast for the best reservoirs with less human interaction as possible. The objectives of this paper is to use machine learning (ML) techniques to predict and classify the reservoirs based on the properties of each reservoirs and choose the best reservoir. In this paper we have made a comparison between the different types of machine learning algorithm and described how we get the best and worst result for each one, the comparison we made gave us that the AdaBoost algorithm gave the worst performance measured in the accuracy while the random forest (RF) algorithm gave the best performance, this paper aim to make improvement of the process of searching for productive zones using ML algorithms.

1. Introduction

Identifying the productive area or "sweet spot" in unconventional resources has been a real challenge over the past decade because there are huge number of parameter to consider when searching for the best area. We need a new technique to make the life easier and the business more profitable, but before we discuss how we solved this problem in this paper let's understand the idea first. In unconventional resources rely on three factors: Organic Quality (OQ), Rock Quality (RQ), and Mechanical Quality (MQ) [1]. Mapping sweet spot benefits the horizontal well drilling and the selection of perforation clusters that can result in the highest production and recovery in unconventional resources. Traditionally, geoscientists determine sweet spots from the interpretation of well logs [2]. One of the most exciting technologies that have recently entered the field of unconventional reservoirs is the application of Artificial Intelligence and Machine Learning. Machine Learning (ML) is a field within Artificial Intelligence (AI) where intelligence is induced regardless of precise programming [3]. ML algorithms can significantly improve workflows used for evaluating sweet spots or productive zones in complex reservoirs [4]. ML is mainly divided into two categories, namely, supervised and unsupervised as shown in Fig. 1 [5].

In this article we talk about machine learning algorithms that significantly improve the workflow used to evaluate and identify productive areas in complex reservoirs where a set of dominant machine learning classification algorithms such as logistic regression, decision tree, random forest, K-nearest neighbours, and boost (AdaBoost) are introduced and Gradient Boosting) to automatically identify productive areas.

The article format will be as follows:

- Talking about the methodology, the most important of which is the rapid analyser technology that determines the ideal tank interval and targets production areas, which is much faster and simpler than traditional 2D and 3D sweet spot modelling.
- Results and their discussion which includes QA validation and comparison of Quick Analyser results with traditional logging results then comparison of Quick Analyser (QA) results with MDT Log, 3D Petrel sweet spot mapping and Geology sweet spot mapping.
- Data Analysis
- Machine Learning (Logistic Regression, Decision Tree Technique, Nearest Neighbours (KNN), Boosting, Random Forest, Comparison of All Techniques).
- Conclusion: In the conclusion, the techniques and their results were compared, and the best ones were shown, the benefits and

* Corresponding author.
E-mail address: amir.gharavi@port.ac.uk (A. Gharavi).

https://doi.org/10.1016/j.ijin.2022.08.001
Received 2 August 2021; Received in revised form 6 July 2022; Accepted 1 August 2022
Available online 11 August 2022

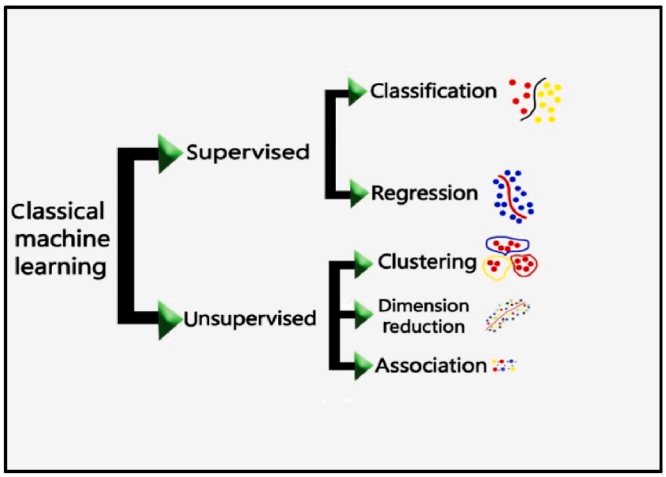

Fig. 1. Machine learning techniques.

advantages of each one separately, and what are the algorithms and methodologies used from machine learning.

Machine Learning (ML) algorithms can significantly improve workflows used for evaluating sweet spots or productive zones in complex reservoirs [4]. ML is mainly divided into two categories, namely, supervised and unsupervised as shown in Fig. 1 [5]. Supervised learning can be used for two sets of problems or data, classification and regression. In classification problems the outcome variable or response variable (Y) takes discrete values. The primary objective of a classification model is to predict the probability of an observation belonging to a class, known as class probability [6]. In this paper application of Machine Learning (ML) for classification in an unconventional oil field (tight oil reservoir) were investigated in conjunction with a novel method, Quick Analyser (QA), for identifying productive zones.

Supervised learning can be used for two sets of problems or data, classification and regression. In classification problems the outcome variable or response variable (Y) takes discrete values. The primary objective of a classification model is to predict the probability of an observation belonging to a class, known as class probability [6]. In this paper application of Machine Learning (ML) for classification in an unconventional oil field (tight oil reservoir) were investigated in conjunction with a novel method, Quick Analyser (QA), for identifying productive zones.

The oil industry has used artificial intelligence for decades in many ways and applications, but with the machine learning era, the industry moved to new and better algorithms and programs that tend to solve hard and complex operations quickly and with optimized result as possible. With the use of machine learning the oil and gas major companies can get fast and accurate data to support their business actions, we will discuss where the ML has been used over time by the oil and gas industry [4].

In order to start with, we use ML to predict future results so we can use ML to predict the happening of some events and try to take actions to prevent them or minimize their effects. Another important role of ML inside the oil and gas industry that there is so many data to handle data of marketing, fields, workers, etc. ML can help with this data to choose what is really important. This can define the next step of ROI (Return on investment). One of the important roles of ML in the industry is to take direct, fast and correct decisions and actions [7].

2. Methodology

2.1. Quick Analyser (QA) technique

Reservoir characterization using subsurface or sweet spot modelling during exploration and appraisal is a time-consuming process which heavily relies on 2D and 3D modelling. The Quick Analyser method identifies the ideal reservoir interval and targets the productive zones. This method is much quicker and simpler than traditional 2D and 3D sweet spot modelling and provides initial information about the subsurface without using any traditional simulation and with robust results for field development.

QA technique, which is a data-driven approach, integrates all subsurface parameters such as geomechanics, petrophysics, geochemistry and drilling to identify the productive zone. This method can optimize field development by reducing technical and commercial risk and establish a repeatable and more profitable recovery strategy and it is applicable for any type of reservoirs (conventional or unconventional).

Applying the QA method can provide:

- Initial and vital information about the subsurface quickly and with high accuracy
- More complete understanding of the reservoir on using real-time data through data acquisition
- Cost effective field planning and optimal well placement (identifying the location of productive zones benefits the horizontal well drilling and completion zones)

The QA method uses an optimum value for each parameter to characterize and evaluate the reservoir utilising real time data. In general, to apply the Quick Analyser technique the following steps are taken:

1. Export all available data from logs for petrophysics analysis
2. Export all calculated data for geomechanics analysis
3. Export all calculated data for geochemistry analysis
4. Using cut-off values or optimum values for all above parameters
5. Integrate all of these values to detect productive zones
6. Carry out validation by conducting sensitivity analysis via machine learning

For example, applying the QA method can identify and evaluate a resource potential with depth, thickness, porosity, permeability, water saturation, oil saturation, total organic carbon, and use them to identify the productive zone. Table 1 through 5 demonstrate classification and characterization of geological subsurface factors using QA technique. This is based on classifying and characterising each parameter according to their optimum (cut-off) values.

Table 6 demonstrate the output of the QA method. The results column shows that if all the conditions for the optimum values for each parameter are met (Equation (1)) it will show 1(productive zone) otherwise 0 (not productive zone).

$$\mathrm{IF}(\mathrm{AND}(\mathrm{K} <= 3, \varphi < 3,\ \ V_{Sh} = 1,\ \ S_o = 1,\ S_w = 1), "1", "0") \quad \text{Equation 1}$$

Where K represents the permeability, φ is the porosity, V_{Sh} depicts the shale volume, S_o and S_w show the oil and water saturations respectively.

Table 7 presents an illustration of the most significant inferential statistical characteristics of dependent and independent variables, including the total number of observations, a five-number summary, the mean, and the standard deviation.

Table 1
Permeability classification of QA method.

Permeability range	Classification	Permeability and color key
0.01>	Very low	4
0.01-1	Low	3
1-10	Mid	2
10-100	Good	1
100 <	Extremely Good	0

Table 2
Porosity classification of QA method.

Porosity range	Classification	Porosity and color key
0.1>	Very low	3
0.1-0.2	Mild	2
0.3<	Good	1

Table 3
Shale volume classification of QA method.

Shale volume	Classification	Shale Color key
0.5>	High	2
0.5<	Low	1

Table 4
Oil saturation classification of QA method.

Oil Saturation	Classification	Oil Sat color code
0.5 <	High	2
0.5 >	Low	1

Table 5
Water saturation classification of QA method.

Water Saturation	Classification	Water Sat color key
0.6 <	High	**2**
0.6 >	Low	**1**

Table 6
Input and results of QA for Nene Marine fiel well number NNM-1.

Depth (m)	Permeability	Porosity	Shale	So	Sw	Results
2420.57	2	2	2	1	2	0
2420.72	2	2	2	1	2	0
2420.87	2	2	1	1	2	0
2421.03	2	2	1	1	2	0
2421.18	2	2	1	1	2	0
2421.33	2	2	1	1	2	0
2421.48	2	2	1	1	2	0
2421.64	2	2	1	1	2	0
2421.79	2	2	1	2	1	1
2421.94	2	2	2	2	1	0
2422.09	2	2	1	2	1	1
2422.25	2	2	1	2	1	1
2422.40	2	2	1	2	1	1
2422.55	2	2	1	2	1	1
2422.70	2	2	1	2	1	1
2422.86	2	2	1	2	1	1
2423.01	2	2	1	2	1	1
2423.16	2	2	2	1	2	0
2423.31	2	2	2	1	2	0
2423.46	2	2	2	1	2	0
2423.62	2	2	2	1	2	0

Boxplot is another standardized way of displaying the distribution of data based on a five-number summary ("minimum", first quartile (Q1), median, third quartile (Q3), and "maximum"). It tells about the outliers and what their values are. It can also indicate if the data is symmetrical, how tightly the data is grouped, and if and how the data is skewed. Fig. 2 shows the boxplot for porosity compared to the target zones 0 (non-productive zone) and 1 (productive zones).

2.2. Experiment and environment

Our work has been conducted in python environment using python language version 3.8.2, this paper used a self-gathered dataset from the research field that contain the needed data which is reading about the land rocks and some other metric.

3. Results & discussion

3.1. QA validation

In order to validate Quick Analyzer (QA) results (our dataset) the following steps were taken:

A. Compare Quick Analyser results with traditional logging results using data obtained from an oil field located in Middle East, Oman, (case study 1)
B. Compare Quick Analyser (QA) results with MDT log and 3D Petrel and Geology sweet spot mapping, (case study 2)
C. Compare with unsupervised methods, MDT log and poresize distribution (case study 3)

3.1.1. Tight oil reservoir, Middle East, (Case study 1)

In order to demonstrate the efficiency of the proposed Quick Analyser method, a case study from an Oman oilfield was undertaken. Fig. 3 shows the results for this case study (A). The results clearly match with SLB Techlog results and the following conclusion can be drawn.

- In Fig. 3, there are 8 tracks which represent depth, calliper, gamma ray, permeability, porosity, water saturation, shale volume and QA results, respectively.
- Red colour in the last column in Fig. 3 represents the productive zones detected by QA (last track) and the blue colour shows the non-productive zones.
- There have been two perforations in the intervals of 4800–4825 m and 4895–4905 m (highlighted).
- The QA clearly detected the areas of higher porosity and permeability and lower shale volume and water saturation in the mentioned intervals.
- As it shown in Fig. 3, QA detected a thin non-productive zone (blue colour) in the perforated depth of 4815. By investigating the well performance reports, it turned out that this depth was cemented after perforation. This could have been avoided by using the QA, prior to the perforation process.

3.1.2. Nene Marine field, tight oil reservoir, (Case study 2)

The MDT Modular Formation Dynamics Tester log (mD/cp) provides fast and accurate pressure measurements and high-quality fluid sampling. It can also measure permeability anisotropy, reservoir heterogeneity and reservoir quality. The MDT log (mD/-cp) is able to acquire most of the data requirements needed for accurate and timely decision-making (see Fig. 4). Therefore, the following conclusion can be drawn:

- Comparing QA results (last column) with sweet spot mapping and MDT log from a west African tight oil field (NNM1), shows that upper Djeno A formation matches with sweet spot mapping, and higher MDT reading (green colours).
- In addition, in Middle Djeno formation, QA correctly detected non-productive zone (brown colour).
- In other regions, QA also shows a strong correlation with sweet spot mapping and MDT readings.

3.1.3. Quick Analyser Vs Machine Leaning (unsupervised) & Poresize distribution (Case study 3)

Fig. 5 illustrates the comparison of several unsupervised methods, comprising of Fuzzy C-means clustering (FCM), K-means, Hierarchial, Geolog software results, and QA results (Cutoffs), with MDT data, and poresize distirbution. Green regions show productive zones while red area shows non-productive zones. Similar to case study A and B, QA

Table 7
Inferential statistical analysis.

	DEPT	PHIE_2014	VSH_2014	SWE_2014	K_FZI_2014	TOC
Count	140.000000	140.000000	140.000000	140.000000	140.000000	140.000000
Mean	2463.213929	0.476207	0.354419	0.310618	0.135067	0.518374
Std.	117.036635	0.215008	0.217709	0.199981	0.172214	0.193306
Min	2244.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	2374.225000	0.337436	0.227080	0.157053	0.015827	0.381720
50%	2458.450000	0.450444	0.336371	0.271866	0.038556	0.508769
75%	2570.425000	0.647758	0.508057	0.448310	0.233173	0.646662
Max	2697.600000	1.000000	1.000000	1.000000	1.000000	1.000000
	Youngsmodulus	Poisson'sRatio	BrittlenessIndicator	poresize	Super RT	
Count	140.000000	140.000000	140.000000	140.000000	140.000000	
Mean	0.491025	0.451739	0.401229	0.212168	0.214286	
Std.	0.207723	0.164976	0.188157	0.213502	0.411799	
Min	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	0.317761	0.362376	0.291125	0.044520	0.000000	
50%	0.505679	0.464613	0.391189	0.099123	0.000000	
75%	0.667133	0.532225	0.493612	0.375561	0.000000	
Max	1.000000	1.000000	1.000000	1.000000	1.000000	

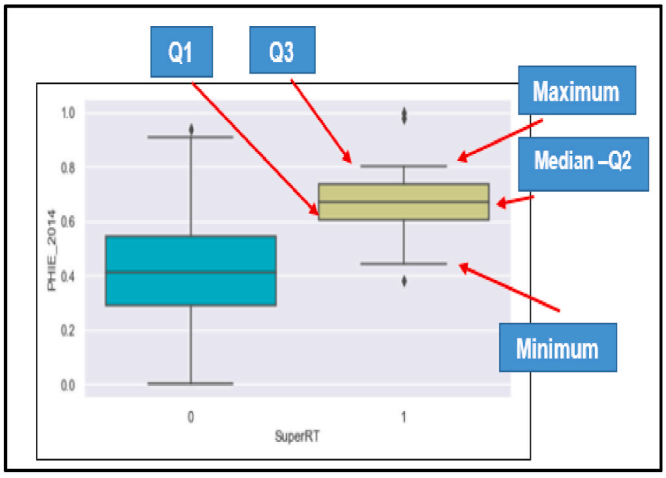

Fig. 2. Porosity boxplot.

shows high correlation with unsupervised methods and poresize distribution as well as MDT reading in order to detect productive zone from non prodcuvte region. Even in some depths, QA shows better and more accurate results.

3.2. Data analysis

3.2.1. Exploratory data analysis (EDA)

EDA was the first stage to conducting Machine Leaning (ML). For Data scientists and stakeholders, EDA is one of the vital stages that provides certain insights and statistical measures. It is used to describe and present the key features and to perform variable selection [8]. The dataset is based on geological parameters and characteristics of an unconventional basin (tight oil sandstone) based on the Quick Analyser (QA) method. The dataset contains 140 records and 11 columns. It has several features related to the potential zones (productive zones) and whether the zones are productive zones or not. The objective, in this case, is to predict which zones are productive zones or potentials and to determine which parameters play important roles in order to detect these zones from unproductive zone (feature importance). Table 8 shows the first 10 columns as the independent variables, where the response variable Y (shown in column "Super RT") is equal to 1 for a productive zone and 0 otherwise).

EDA Analysis of the full 140 data points showed that there are around 78% observation of unproductive zone and 21% observation of productive zone.

Fig. 6 provide heatmap and correlation plot of all the varibles. The correlation plot is used for measuring the strength and direction of the linear relationship between two continuous random variables.

1. The correlation value lies between −1.0 and 1.0. The sign indicates whether it is positive or negative correlation.
2. −1.0 indicates a perfect negative correlation, whereas +1.0 indicates perfect positive correlation

Fig. 6 shows that porosity (PHIE_2014), permeability (K_FZI_2014) and TOC are positively correlated with the targets zone (Super RT) whereas mechanical properties such as Brittleindex and Young's modulus, and poison's ratio as well as shale volume are not so strongly correlated. More Details about data analysis (EDA) results are shown in [9] Appendix 1.

3.3. Machine learning

The following ML classification techniques such as Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbours and Boosting (AdaBoost, Gradient Boosting) were investigated to determine productive zones(target variable) which obtained by QA method in the previous section. Before the model was build, the dataset was split X:Y to create a training and test dataset (with 70:30 ratio). The model was then built using the training set and tested using the test set.

3.3.1. Logistic regression

Logistic Regression is one of the supervised machine learning algorithms used for classification. In logistic regression, the dependent variable is categorical [10]. Logistic regression is a statistical model in which the response variable takes a discrete value and the explanatory variables can be either continuous or discrete. If the outcome variable takes only two values, then the model is called binary logistic regression model. The outcomes are called positive (usually coded as Y = 1) and negative (usually coded as Y = 0). Then the probability that a record belongs to a positive class, P(Y = 1), using the binary logistic model is given by sigmoid function which converts the input into range 0 and 1.

3.3.1.1. Logistic regression results. To understand how many observations the model has classified correctly and how many has not, a cut-off probability of 0.5 (default) initially was used. The actual column in Table 9 depicts the actual label of the SuperRT in the test dataset, while predicted column depicts what the model has predicted by taking 0.5 as cut-ff probability value. For obesrvation10, the model predicateds very

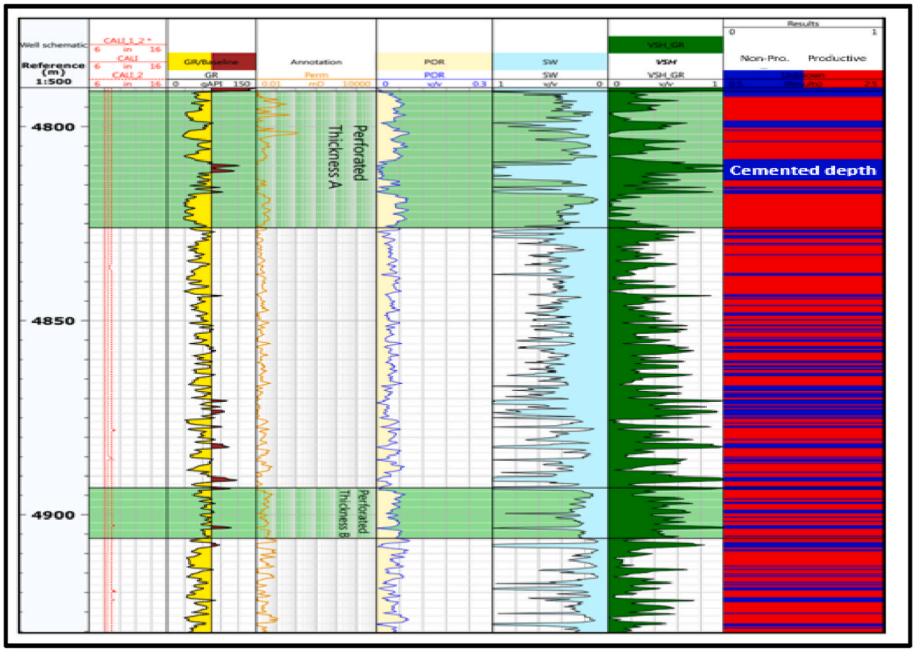

Fig. 3. Results of proposed QA method on an oilfield from the Middle East.

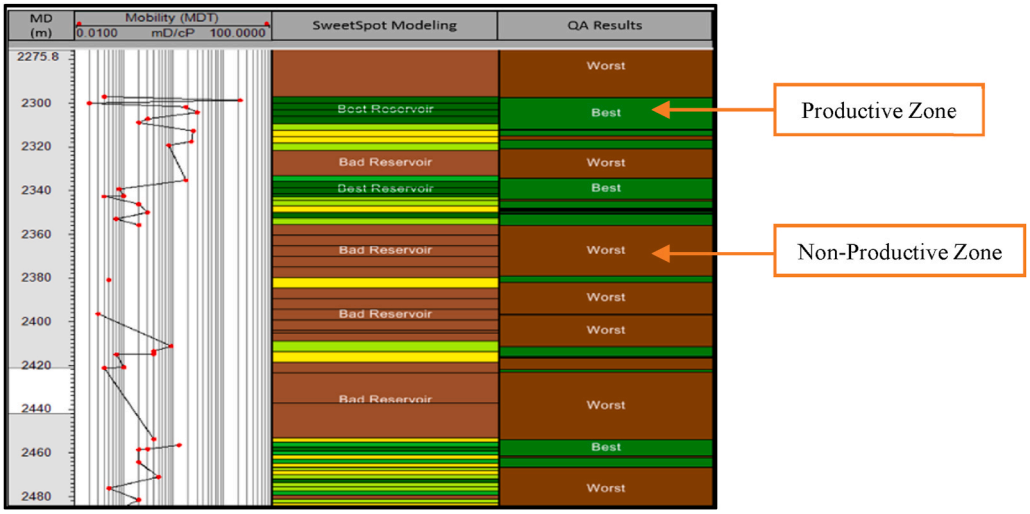

Fig. 4. Comparison of MDT reading, sweet spot mapping and QA method for well number NNM1.

low probability (0.2) of being a non-prodcutive zone whereas it is actually a prodcutive zone. The model has wrongly classified this one. Similarly, for observation of 104, the model predicted high probability (0.7) of being a prodcutive zone whereas it is actually a non-prodcutive zone. The model has wrongly classified this one either. However, the model predicts high probability (0.9) of being a productive zone for observation 0, which is actually a productive zone. The model correctly predicted the class in this case.

3.3.1.2. Confusion matrix & measuring accuracies. Fig. 7 and Table 10 illustrate confusion matrix results and linear regression (LR) reports. In Fig. 7, the columns represent the predicted label (class), while the rows represent the actual label (class) [6]. For example, out of 11 (i.e., 5 + 6) productive zones (Good Zone), only 6 have been classified correctly as productive zones (Poor Zone) and rest 5 have been classified as unproductive zones when the cut-off probability is 0.5.

Table 10 gives a detailed report of precision, recall, and F-score for each class. Recall for positive cases (Y = 1) are only 55%, which suggests some of the cases were predicted as negatives.

The model is very good at identifying the unproductive zone (Y = 0) with F1 score of 91%, but not very good at identifying productive zones (Y = 1) with recall score of 55%. Overall accuracy based on logistic regression is 86%. This is the result for cut-off probability of 0.5%. Fig. 8 depicts the distribution of predicted probability values for productive and non-productive zones to understand how well the model can distinguished non-productive zones from productive zones. The larger the overlap between predicted probabilities for different classes, the higher the misclassification will be.

3.3.1.3. ROC and AUC curves. ROC and AUC are two important measures of model performance for classification problems at various threshold settings. The higher ROC curve, the better the model. As it can be seen, the ROC curve for the logistic model was 89% is shown in Fig. 9, indicating a good model.

Fig. 5. Comparison of all used methods for sweet spot detection by mobility and poresize distribution.

3.3.1.4. Finding optimal classification cut-off. The overall accuracy, sensitivity and specificity will depend on the choosen cut-off probability. This was investigated using the Youden index [6], and the cost-based approach [6]. Both methods gave an optimal cut-off of 0.23. Applying this gave the improved results, shown in Fig. 10.

Fig. 10 and Table 11 show new confusion matrix with optimal cut-off using Yodens' index and cost function (P = 0.23). With cut-off probability of 0.23, the model is able to classify the prodcutive zones better and F1-score and Recall for prodcutive zones (Y = 1) has improved to 0.77 and 0.91 from 76% to 55% respectively.

3.3.2. Decision tree technique

Decision tree is one of the most important and powerful predictive analytics methods applied for generating business rules. It use tree-like structure to predict the value of an outcome variable. The algorithm use the complete data and start with the root data then splits the notes into multiple branches. In this technique, the data is divided into subsets in order to create more uniform branches (children nodes), [6]. Decision tree results shows the following in Fig. 11.

The following points can be concluded:

- At the top node, there are 98 observations of which 79 are unproductive zones and 19 are productive zones. The corresponding Gini index is 0.313.
- TOC is the most important feature for splitting productive and unproductive zones in the dataset when compared to other features and hence, chosen as the top splitting criteria.
- The first rule (TOC< 0.5) means if the zones has TOC values below 0.7 or above
- This rule has split the dataset into two subsets represented by the second level nodes. On the left node, there are 79 samples (i.e., TOC below 0.7) and on the right node, there are 19 samples (i.e. having TOC above 0.7).
- The nodes represented by dark shades depict unproductive zones, while the nodes represented by light shades are productive zones.

One of the rules can be interpreted as: If the zone does have TOC values above 0.7 and volume of shale below 0.4 and brittle index is below 0.218, it then there is high probability of being a productive zone. There are 17 records in the dataset that satisfy these conditions and 2 of them have unproductive zones. Another rule: If TOC values is below 0.6, shale volume is above 0.2, and brittle index is below 0.067, then there is high probability of being a poor zone. There are 78 samples in the datasets and 74 records of which in the dataset that satisfy these conditions and 4 of them have productive zones. Parameters such as TOC, Volume of shale, SWE (water saturation) and brittle index are most important parameters to identify good zones or productive zones.

3.3.2.1. Decision tree accuracy. Fig. 12 provides ROC AUC score of the decision tree model. As Shown in Fig. 12, DT model has AUC score of 69% and is lower than the LR model.

3.3.2.2. DT confusion matrix. Fig. 13 and Table 12 provide confusion matrix and classification report of the decision tree model. As it can be seen recall for the positive cases (Y = 1, productive zones) is 0.55, suggesting some misclassification.

3.3.2.3. Finding optimal criteria and max depth. Figs. 14 and 15 and Table 13 provide ROC curve, confusion matrix and classification report of the tuned decision tree model based on GridSearchCV results.

Compared to the previous DT model (default model), tune DT model accuracy has been increased to 0.79 from 0.76.

3.3.3. K-Nearest Neighbours (KNN)

KNN algorithm compares distance of the data points to determine the similarities between data points [5]. Different approaches exist to calculate the distance; however, Euclidean distance (length of the line segment between the two points) is the usual method to calculate the distance [6]. As the distance between two points decreases, the similarity increases. KNN classifies data points based on their similarities in labels, (see Fig. 16).

3.3.3.1. KNN accuracy. As Shown in Fig. 17, KNN has AUC score of 91% and is better than two previous models LR and DT with 89% and 73% respectively.

3.3.3.2. KNN confusion matrix. Fig. 18 and Table 14 provide confusion matrix and classification report of KNN model. As shown in Table 14 the recall of positive cases has improved from 0.45 (DT model) to 0.73 in KNN model. The above model accuracy is obtained by considering default number of neighbours (k = 5). The Recall score has been increased significantly comparing with LR and DT models. Resulting in prediction improvement in positive cases (Y = 1).

3.3.3.3. KNN hyperparameters tuning. Hyperparameter tuning is a technique used in ML and DL in order to find optimal value for hyperparameters. GridSearchCV method can be used in ML algorithms such as DT, KNN, LR, RF, etc. in order to find these optimum values. GridSearch

Table 8
Dependent and independent variables of available data.

Index	Dependent Variables (Features)										Independent Variable (Target)
	DEPT	PHIE_2014	VSH_2014	SWE_2014	K_FZI_2014	TOC	Youngsmodulus	Poisson'sRatio	BrittlenssIndicator	Poresize	SuperRT
0	2410.5	0.800479	0.000000	0.114840	0.312939	0.950979	0.444156	0.388466	0.304002	0.418947	1
1	2389.5	0.694545	0.136057	0.025256	0.213306	0.808909	0.410370	0.341185	0.254701	0.330534	1
2	2353.0	0.755884	0.000000	0.319136	0.287191	0.621279	0.221638	0.493735	0.344649	0.406034	0
3	2625.3	0.340280	0.477093	0.534075	0.010663	0.413167	0.769919	0.475907	0.481169	0.031940	0
4	2415.5	0.535026	0.165163	0.160427	0.152076	0.644643	0.457917	0.309380	0.236411	0.283460	1
5	2414.8	0.790825	0.000000	0.322094	0.527241	0.604112	0.257674	0.481917	0.344157	0.679391	0
6	2481.6	0.251964	0.753323	0.339771	0.009685	0.254303	0.471627	0.545112	0.475948	0.033121	0
7	2533.9	0.333640	0.511666	0.191938	0.010048	0.405470	0.669914	0.684009	0.749114	0.030593	0
8	2309.0	0.730730	0.044108	1.000000	0.432428	0.838833	0.248832	0.343794	0.225494	0.602269	1
9	2344.1	0.495219	0.599340	0.184913	0.177391	0.550711	0.525521	0.375107	0.322749	0.356990	0

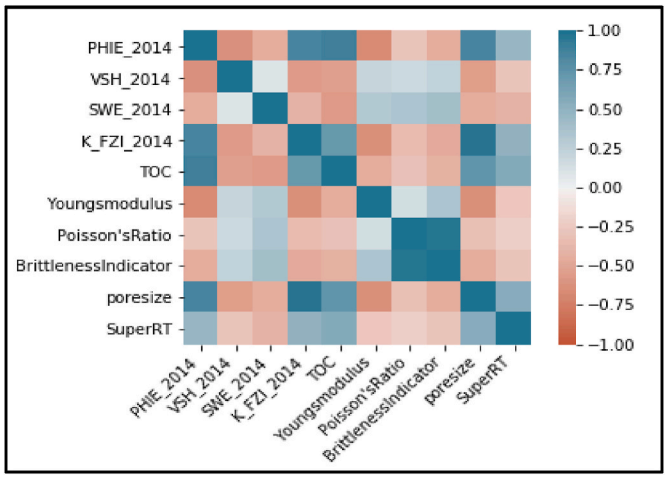

Fig. 6. Correlation and heat map plot of each varibles.

Table 9
Actual and predicted outcome with predicted probability.

Index	actual	predicted_prob	predicted	Comments
26	0	0.146128	0	correct classification
19	1	0.540402	1	
104	0	0.700420	1	Misclassification
138	0	0.097014	0	correct classification
42	1	0.699048	1	
0	1	0.952959	1	
110	0	0.029897	0	
18	0	0.115295	0	
10	1	0.256822	0	Misclassification
81	1	0.686216	1	correct classification

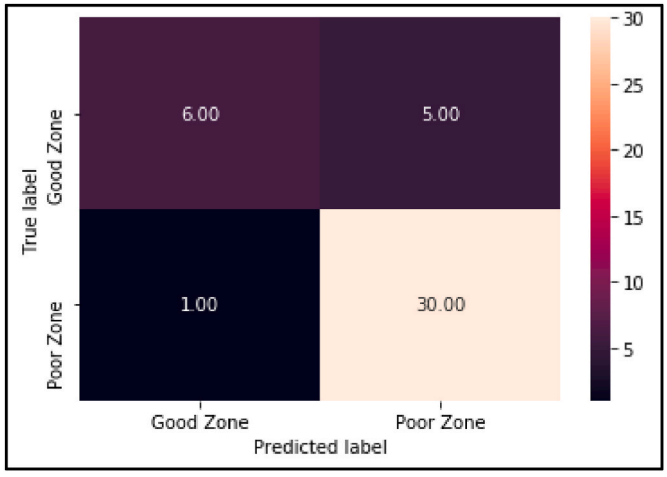

Fig. 7. Logistic regression confuction matix.

Table 10
Logistic regression calssification Report.

	Precision	Recall	f1-score	Support
0	0.86	0.97	0.91	31
1	0.86	0.55	0.67	11
Accuracy			0.86	42
Macro avg	0.86	0.76	0.79	42
Weighted avg	0.86	0.86	0.85	42

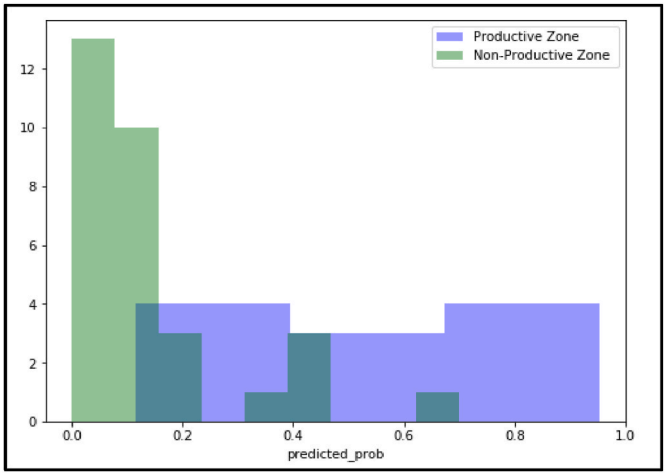

Fig. 8. Distribution of predicted probability values by the model for both productive and non-productive zones.

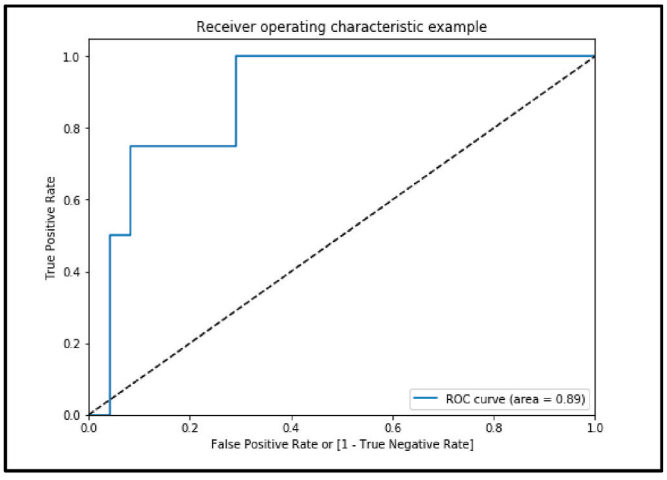

Fig. 9. ROC curve for logistic regression method.

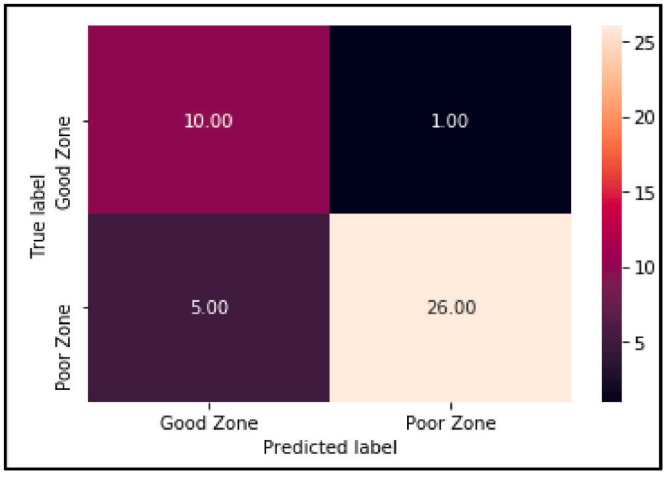

Fig. 10. Confusion matrix with optimal cut-off using Yodens' index.

suggests the best combination of parameters is K (neighbours) of 7, Canberra distance, and the corresponding roc_auc score is 89. Finally, a new model was built based on optimal hyperparameter tuning. The results in Fig. 19 and Table 15 show improvement for increasing F1 score

Table 11
Logistic regression calssification report based on cut-off value of 0.23.

	Precision	Recall	f1-score	Support
0	0.96	0.84	0.90	31
1	0.67	0.91	0.77	11
Accuracy			0.86	42
Macro avg	0.81	0.87	0.83	42
Weighted avg	0.89	0.86	0.86	42

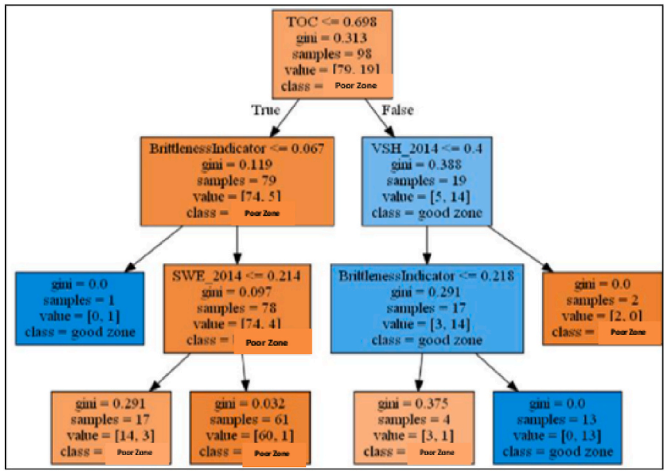

Fig. 11. Designed decision tree.

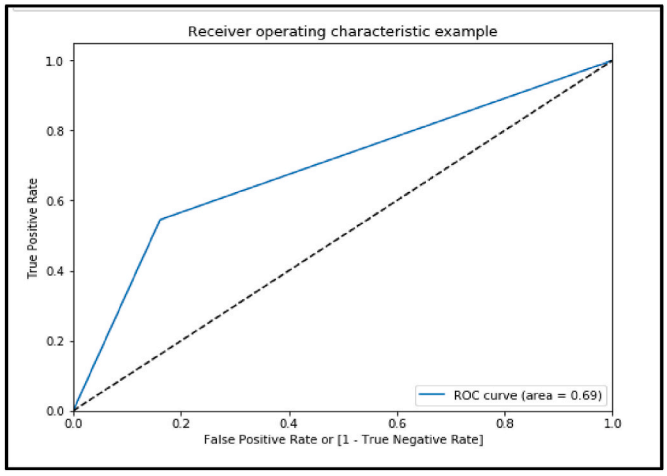

Fig. 12. ROC curve for dection tree model.

and recall of class 1 compared with previous KNN model (default) using k = 5. As it can be seen this model has higher percentage of true positive compare to previous KNN model. In business context, the objective is to build a model that will have a high number of true positive.

3.3.4. Random forest

Random forest is one of the most popular ensemble techniques used in the industry due to its performance and scalability. A random forest is an ensemble of decision trees (classification and regression tree), where each decision tree is built from bootstrap samples (sampling with replacement) and randomly selected subset of features without replacement. The decision trees are normally grown deep (without pruning) [6].

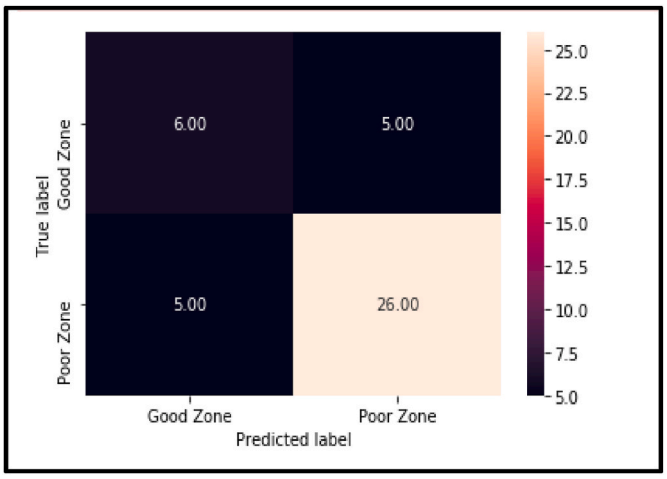

Fig. 13. DT model confusion matrix.

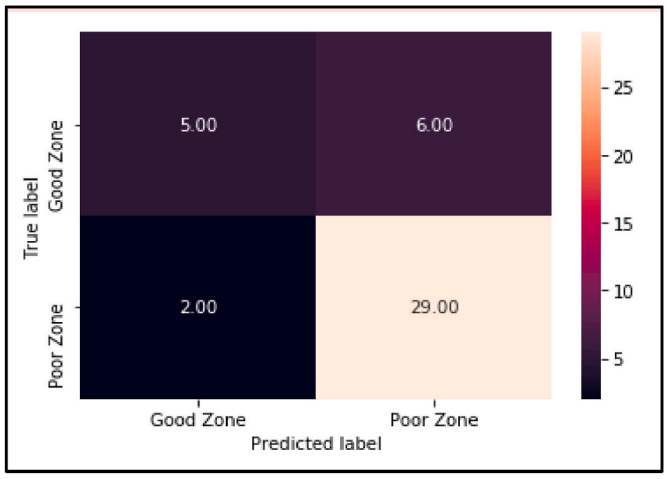

Fig. 15. Tuned DT model, confusion matrix.

Table 12
DT calssification reort based on default values.

	Precision	Recall	f1-score	Support
0	0.84	0.84	0.84	31
1	0.55	0.55	0.55	11
Accuracy			0.76	42
Macro avg	0.69	0.69	0.69	42
Weighted avg	0.76	0.76	0.76	42

Table 13
Tuned DT model classification report of based on GridSearchCV results.

	Precision	Recall	f1-score	Support
0	0.82	0.90	0.86	31
1	0.62	0.45	0.53	11
Accuracy			0.79	42
Macro avg	0.72	0.68	0.69	42
Weighted avg	0.77	0.79	0.77	42

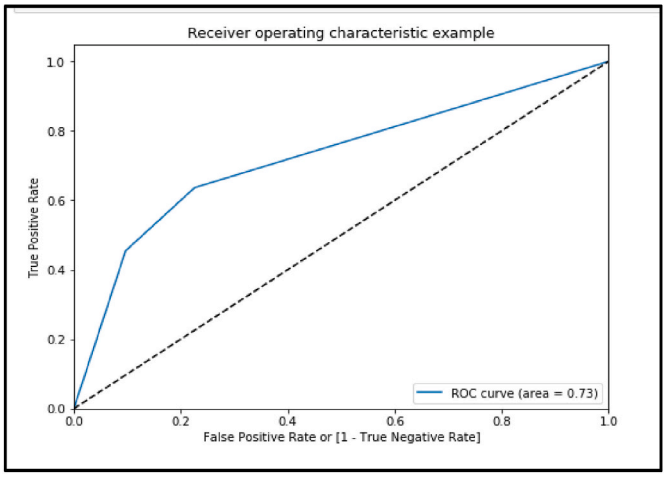

Fig. 14. DT ROC curve for tunned.

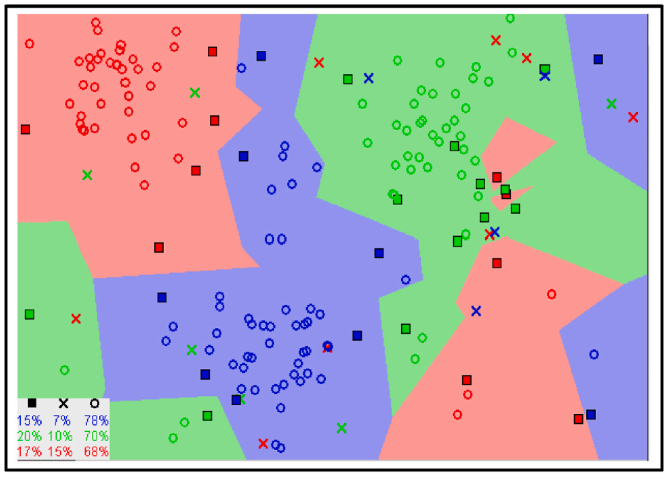

Fig. 16. Similar data points typically exist close to each other (Wikipedia).

3.3.4.1. Random forest accuracy. The plot of ROC curve is shown in Fig. 20. The corresponding AUC value for the RF model is 91%.

AUC for the random forest model is 90% (Fig. 20) and better compared to the DT and LR models. However, the accuracy still can be improved by using grid search by fine-tuning the hyperparameters. Fig. 21 and Table 16 show confusion matrix and classification report for the RF model. The model detects 6 out of 11 productive zones with recall 55%. The overall accuracy of the model is 76%.

3.3.4.2. Optimal parameters tuning. Similar to the previous models, GridSearchCV technique was conducted to find the optimal values for hyperparameters. Therefore, parameters such as max_depth, n_estimators and max_features of 15, sqrt, 10 respectively were calculated as the best model. AUC score has reached 0.92 with the following optimal values for the hyperparameters (see Fig. 22).

3.3.4.3. Random forest confusion matrix. Fig. 23 and Table 17 provide confusion matrix and classification of random tuned forest model.

As shown in Table 17 the precision, recall and F1 score for positive

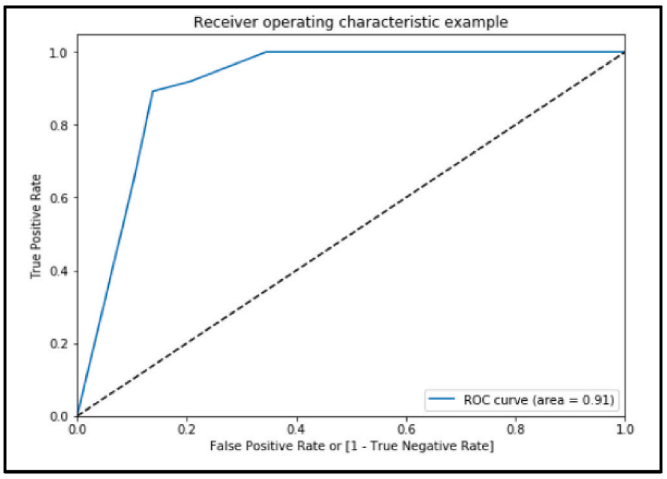

Fig. 17. ROC AUC curve for KNN model.

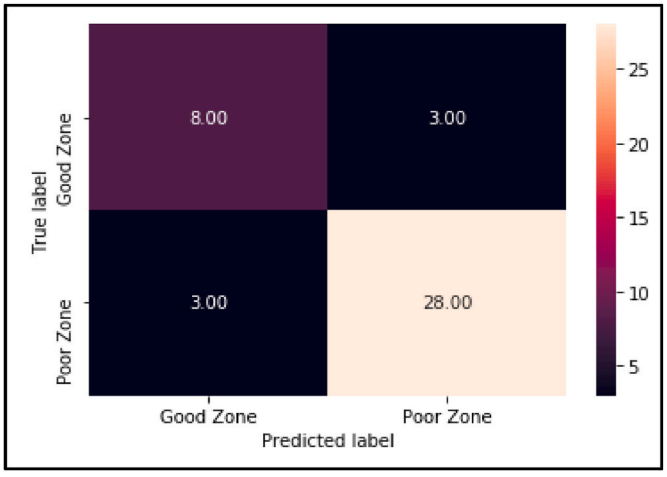

Fig. 18. KKN confusion matrix results using default values.

Table 14
KNN model classification report using default values.

	Precision	Recall	f1-score	Support
0	0.90	0.90	0.90	31
1	0.73	0.73	0.73	11
Accuracy			0.86	42
Macro avg	0.82	0.82	0.82	42
Weighted avg	0.86	0.86	0.86	42

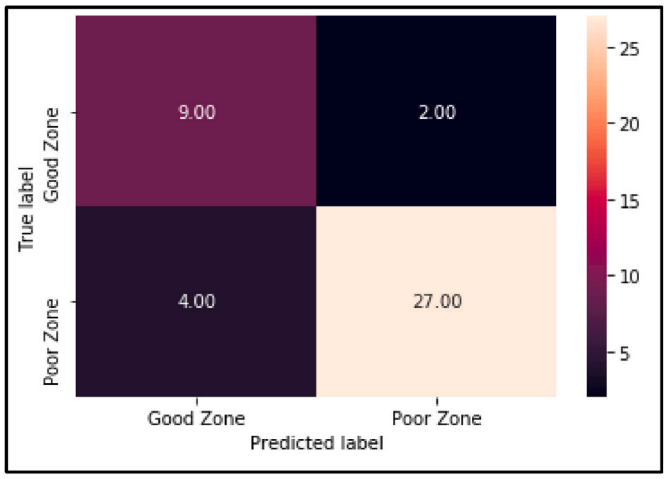

Fig. 19. Tuned KKN confusion matrix results.

Table 15
Tuned KNN model classification report based on GridSearchCV results.

	Precision	Recall	f1-score	Support
0	0.93	0.87	0.90	31
1	0.69	0.82	0.75	11
Accuracy			0.86	42
Macro avg	0.81	0.84	0.83	42
Weighted avg	0.87	0.86	0.86	42

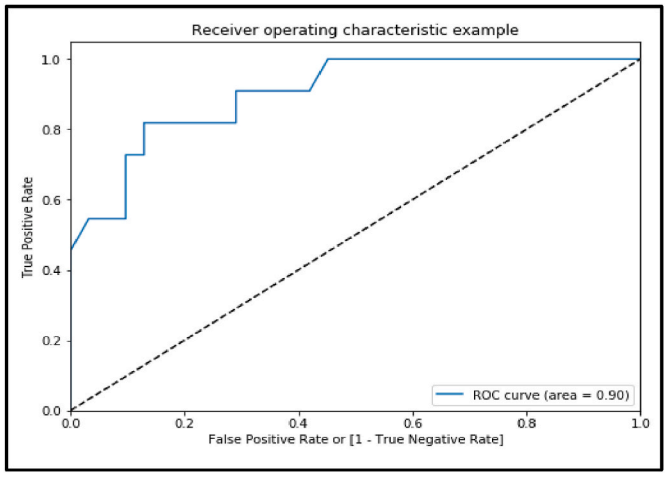

Fig. 20. ROC AUC curve for random forest.

cases are 0.78 and 0.64. 0.70 respectively which are better than what was obtained by two previous models, RF model (default model) and DT model namely. The overall accuracy of this model also improved from 76% to 86%.

3.3.4.4. Finding important features. Random forest algorithm reports feature importance by considering feature usage over all the trees in the forest.). Fig. 24 shows the tuned random forest model feature importance. The top 5 features are TOC, Poresize, K_FZI, SWE and Brittleindex.

Fig. 25 represents the cumulative sum of features importance can show the amount of variance explained by the top five features.

The top five features provide nearly 80% of the information in the data with respect to the outcome variable. This technique can also be used for feature selection. Random forest being a black box model, cannot be interpreted. But it can be used to select a subset of features using feature importance criteria and build simpler models for interpretation.

3.3.5. Boosting

Boosting is another popular ensemble technique which combines multiple weak classifiers into a single strong classifier, boosting is done by create and training model with any ML algorithm and after that create and train another model to correct the mistakes of the first model and increase the accuracy of prediction. A weak classifier is one which is slightly better than random guessing. That is, the error is less than 50%. Any classification algorithm can be used for boosting and is called the

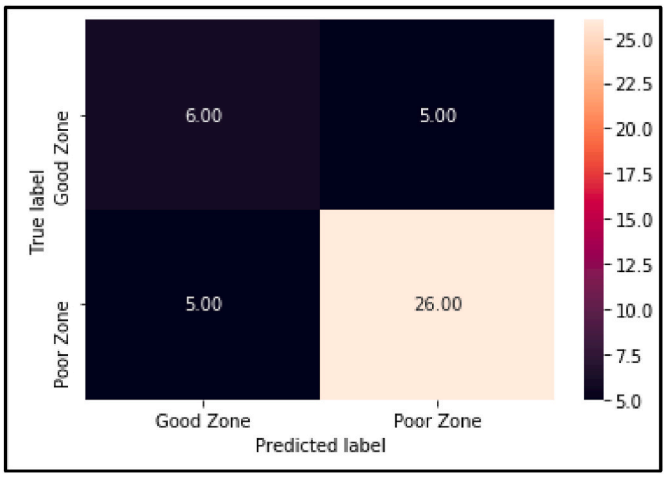

Fig. 21. Random forest confuction matix.

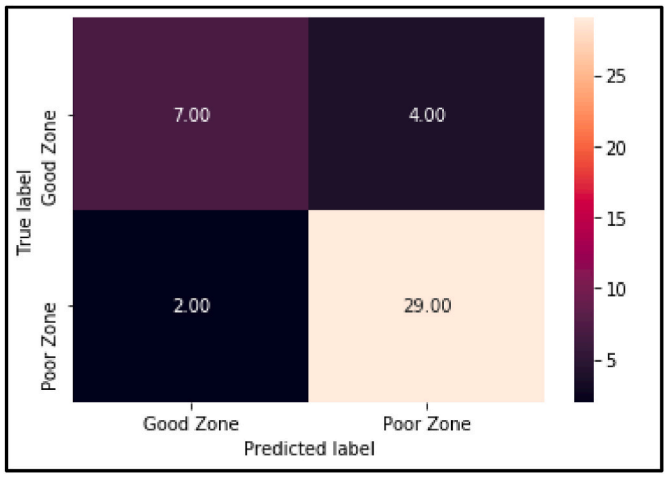

Fig. 23. Confusion matrix of tuned random forest model.

Table 16
Radom forest calssification report.

	Precision	Recall	f1-score	Support
0	0.84	0.84	0.84	31
1	0.55	0.55	0.55	11
Accuracy			0.76	42
Macro avg	0.69	0.69	0.69	42
Weighted avg	0.76	0.76	0.76	42

Table 17
Classification report of tuned random forest model.

	Precision	Recall	f1-score	Support
0	0.88	0.94	0.91	31
1	0.78	0.64	0.70	11
Accuracy			0.86	42
Macro avg	0.83	0.79	0.80	42
Weighted avg	0.85	0.86	0.85	42

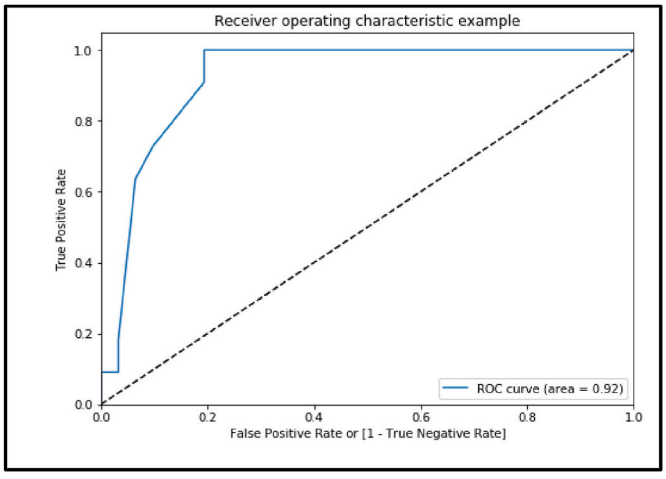

Fig. 22. Improved random forest ROC curve.

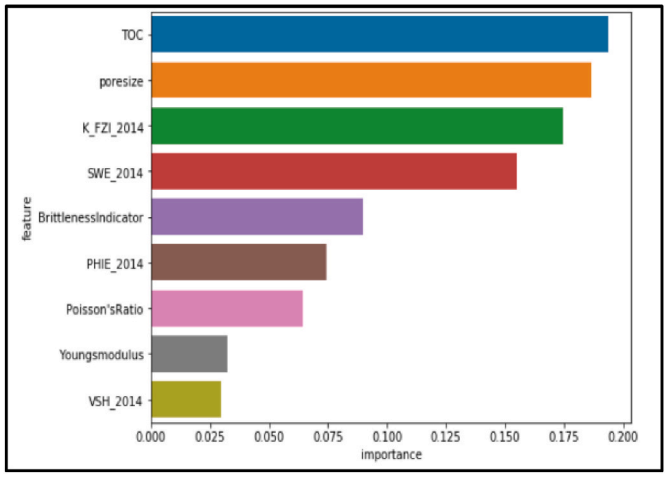

Fig. 24. Feature importance results (tuned random forest model).

base classifier [6]. Boosting builds multiple classifiers in a sequential manner as opposed to bagging, which can build classifiers in parallel. Boosting builds initial classifier by giving equal weights to each sample and then focuses on correctly classifying misclassified examples in subsequent classifiers.

Two most widely used boosting algorithms are:

- AdaBoost
- Gradient Boosting

3.3.5.1. AdaBoost. In AdaBoost, each record in the training dataset will receive a weight, which indicates the possibility of using that record for training. For the first classifier, AdaBoost will use an equal weight for all of the examples (random sampling). Afterward, the weight of the misclassified records will be increased, in order to increase the possibility of their selection. This way, the next classifier learns to classify them more efficiently [6]. Similar to the KNN model the AdaBoost model has an AUC score of 0.91% (see Fig. 26). Fig. 27 and Table 18 provide confusion matrix and classification of AdaBoost model.

As it can be seen form Table 18, the F1 score and recall for positive

	feature	importance	cumsum
4	TOC	0.193844	19.384390
8	poresize	0.186512	38.035638
3	K_FZI_2014	0.174460	55.481684
2	SWE_2014	0.154922	70.973900
7	BrittlenessIndicator	0.089686	79.942504
0	PHIE_2014	0.074142	87.356747
6	Poisson'sRatio	0.064332	93.789991
5	Youngsmodulus	0.032500	97.039981
1	VSH_2014	0.029600	100.000000

Fig. 25. Cumulative sum of features importance.

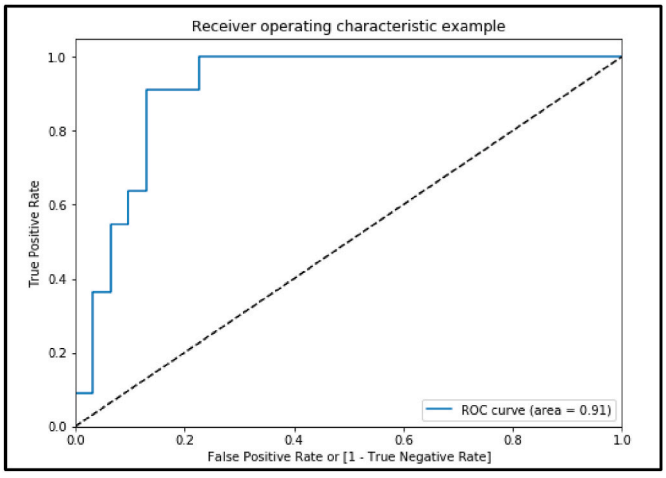

Fig. 26. AdaBoost ROC curve.

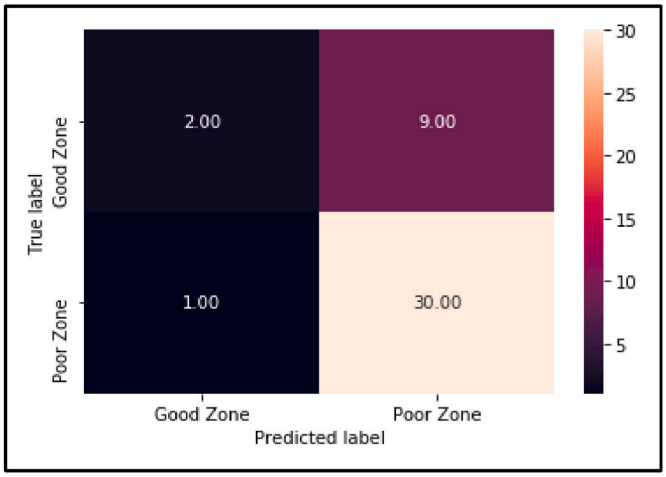

Fig. 27. Confusion matrix of the AdaBoost model.

Table 18
Classification report of AdaBoost model.

	Precision	Recall	f1-score	Support
0	0.77	0.97	0.86	31
1	0.67	0.18	0.29	11
Accuracy			0.76	42
Macro avg	0.72	0.57	0.57	42
Weighted avg	0.74	0.76	0.71	42

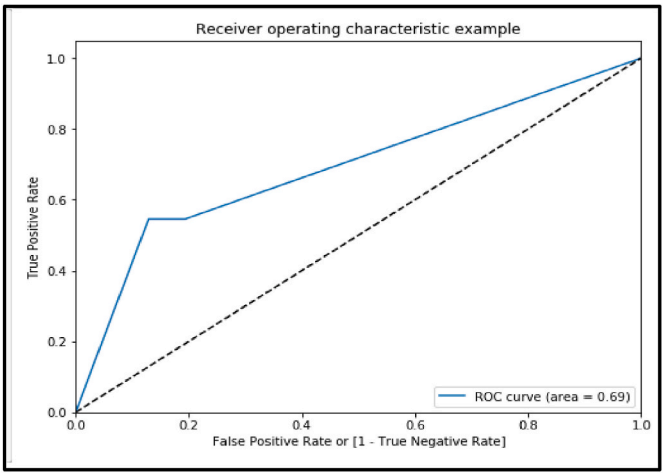

Fig. 28. Gradient Boosting ROC AUC curve.

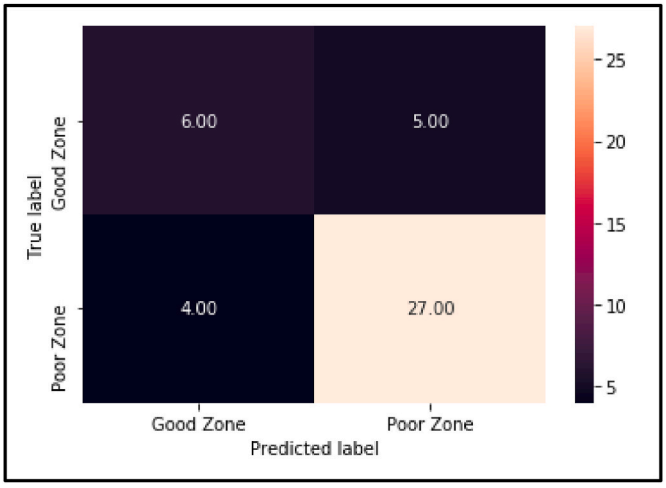

Fig. 29. Gradient Boosting model confusion matrix.

Table 19
Classification report of Gradient Boosting model.

	Precision	Recall	f1-score	Support
0	0.84	0.87	0.86	31
1	0.60	0.55	0.57	11
Accuracy			0.79	42
Macro avg	0.72	0.71	0.71	42
Weighted avg	0.78	0.79	0.78	42

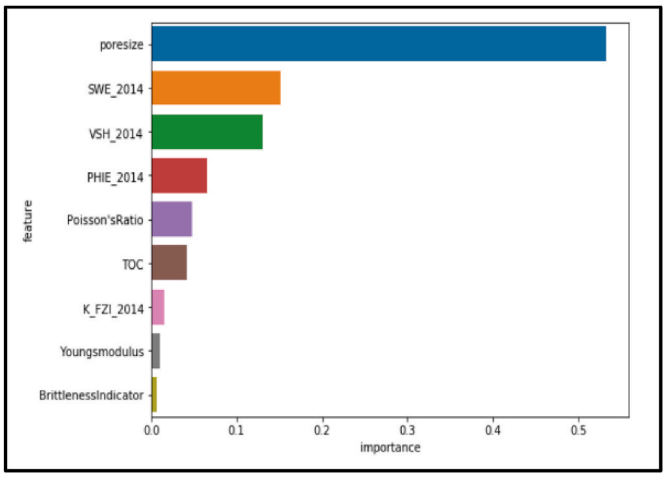

Fig. 30. Feature sorted by their importance values in the gradient boosting model.

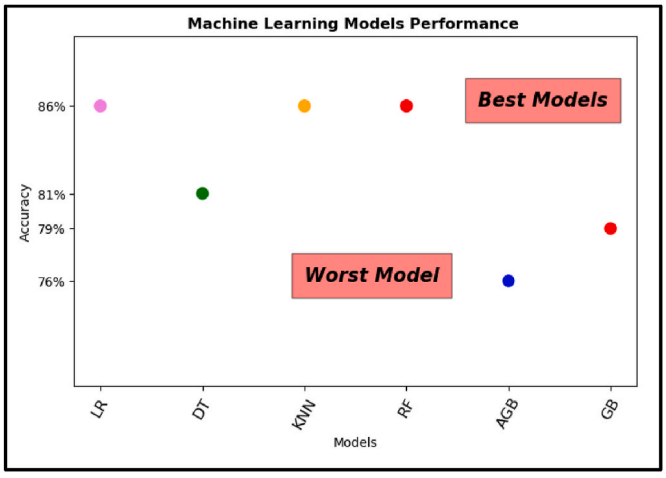

Fig. 31. Comparison of the used methods.

cases are 0.29 and 0.18 respectively, which are far lower than the previous models. The model detects 11 out of 42, productive zones with recall reaching 18%, which is the lowest among all models.

3.3.5.2. Gradient boosting. In Gradient Boosting, the main focus is on the residuals from previous classifiers and a model will be fit to residuals. This technique leverages the residuals patterns and improves the model with weak classifiers. When the model recognizes that the residuals are out of patterns, it will stop the residuals modelling. The base classifier in Gradient Boosting is Decision Tree [6,11]. The ROC curve of gradient boosting is shown in Fig. 28. The corresponding AUC is 0.69. Fig. 29 and Table 19 illustrate Gradient Boosting confusion matrix results and classification report.

The model detects 11 out of 42 productive zone cases with recall reaching 55%. Comparing to AdaBoost model Gradient boosting model showed a better performance in detecting the positive cases (Y = 1), with Recall and F1 score of 55% and 57% compared to 18% and 28% respectively. The overall accuracy of Gradient boosting model is 79%. Like Random Forest algorithm, the boosting algorithm also provides feature importance based on how each feature has contributed to the accuracy of the model. Gradient boosting also selected the poresize, SWE and VSH and PHE as well as TOC as top features (see Fig. 30) which have maximum information about whether a zone is productive or not.

3.4. Comparison of all techniques

Comparing the performance of all methods are shown in Fig. 31 Random Forest (RF) and KNN models showed the highest accuracy (∼ 86%), while boosting algorithms, AdaBoost and Gradient Boosting models, showed the lowest accuracy 76% and 79% respectfully among other sensitives.

4. Conclusion

This research has made a complete study and analysis about the classification of reservoir by understanding their characterization and extracting features from them, this features will be used in training and testing the machine learning models, most important characterization we used in our work are subsurface or sweet spot modelling during exploration and appraisal characterizing the reservoir is a time-consuming process and chore and heavily relies on 2D and 3D modelling. However, Quick Analyser (QA) method provides results, which are robust and quick baseline for field development that reduces technical and commercial risk and establishes repeatable, more profitable recovery without any traditional 2D and 3D modelling. Different machine learning algorithms were used in this study in order to detect sweet spots as a supervised learning problem and used tools and methodology from machine learning to build data-driven sweet spot classifiers. In this research we used a different type of machine learning models and for each model we calculated the performance in order to compare between the models. The results we found by running work is as follow the LR, KNN and Random Forest show a fair degree of promise, scoring the highest 86%, boosting algorithms AdaBoost and Gradient Boosting models exhibited the lowest accuracy of 76% and 79% respectfully among all sensitivities, as result from this research is to make the process of searching for the best areas a robust process and with the highest accuracy as possible.

Declaration of competing interest

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of this article.

Appendix 1

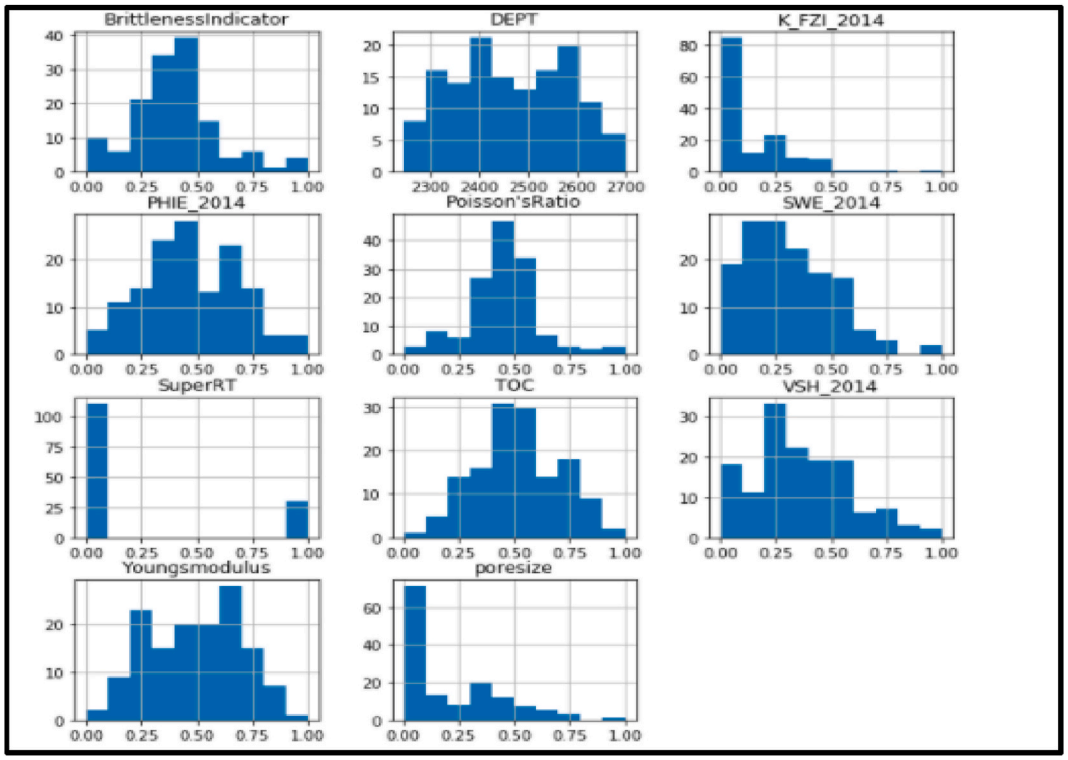

Fig. 1. Histogram of the input varibles distribution.

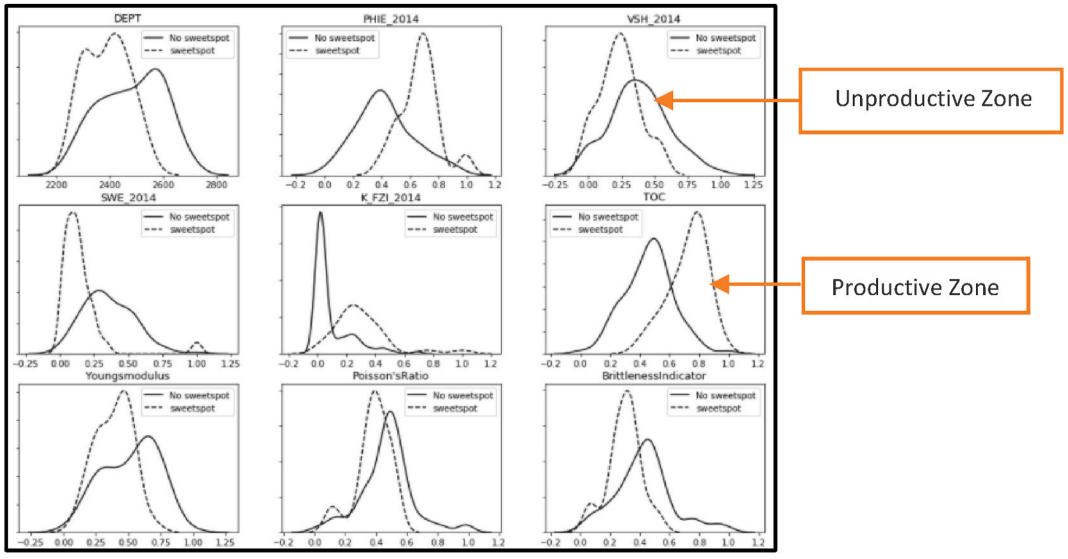

Fig. 2. Density distribution of each varibles in regards to the target variable.

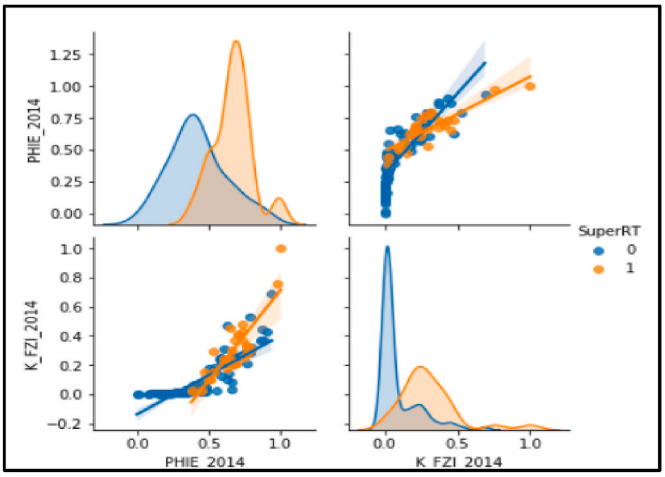

Fig. 3. Relationship between porosity and permeability (k_FZI).

References

[1] Jeffrey B. Aldrich1, John P. Seidle, Sweet Spot" Identification and Optimization in Unconventional Reservoirs, 2018.

[2] Jizhou Tang, Bo Fan, Lizhi Xiao, Shouceng Tian, Liyuan Zhang, David Weitz, A. John, Paulson, A New Ensemble, 2021.

[3] J. Moolayil, J. Moolayil, S. John, Learn Keras for Deep Neural Networks, Apress, Birmingham, 2019, pp. 33–35.

[4] S. Tandon, Integrating machine learning in identifying sweet spots in unconventional formations, in: SPE Western Regional Meeting, OnePetro, 2019, April.

[5] S. Aghabozorgi, A.S. Shirkhorshidi, T.Y. Wah, Time-series clustering–A decade review, Inf. Syst. 53 (2015) 16–38, 2015.

[6] Manaranjan Pradhan, U Dinesh Kumar, Machine Learning Using Python, Kindle Edition, 2018, p. 259.

[7] S. Wang, S. Chen, Insights to fracture stimulation design in unconventional reservoirs based on machine learning modelling, J. Petrol. Sci. Eng. 174 (2019) 682–695.

[8] ashish. Srimal, Why EDA is necessary for machine learning? Medium website (2019) [Online]. [5 March 2020]. Available from: https://medium.com/@srimalashish/why-eda-is-necessary-for-machine-learning.

[9] Will Koehrsen, Histograms and density plots in Python [Online]. [23 March 2018]. Available from: https://towardsdatascience.com/histograms-and-density-plots-in-python-f6bda88f5ac0.

[10] Saishruthi Swaminathan, Logistic regression detailed overview. Towards data science website [Online].10 march 2020. Available from: https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc, 2018.

[11] D. Han, J. Jung, S. Kwon, Comparative study on supervised learning models for productivity forecasting of shale reservoirs based on a data-driven approach, Appl. Sci. 10 (4) (2020) 1267.