International Journal of Intelligent Networks 4 (2023) 29–37

Contents lists available at ScienceDirect

International Journal of Intelligent Networks

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

Unsupervised machine learning technique for classifying production zones in unconventional reservoirs

Karrar A. Abbas [a,*], Amir Gharavi [b], Noor A. Hindi [a], Mohamed Hassan [a], Hala Y. Alhosin [a], Jebraeel Gholinezhad [a], Hesam Ghoochaninejad [a], Hossein Barati [a], James Buick [a], Paria Yousefi [a], Reham Alasmar [a], Salam Al-Saegh [a]

[a] University of Portsmouth, Faculty of Technology, School of Energy and Electronic Engineering, Lion Gate Building, Lion Terrace, Portsmouth, Hampshire, PO1 3HF, United Kingdom
[b] University College London, United Kingdom

ARTICLE INFO

Keywords:
Machine learning
Sweet spots
Unsupervised classification
Supervised classification
Unconventional reservoirs
Clustering analysis
Support vector machine

ABSTRACT

Significant amounts of information are rapidly increasing in bulk as a consequence of the rapid development of unconventional tight reservoirs. The geomechanical and petrophysical characteristics of the wellbore rocks influence the sweet and non-sweet areas of tight unconventional reservoirs. Using standard approaches, such as data from cores and commercial software, it is difficult and costly to locate productive zones. Furthermore, it is difficult to apply these techniques to wells that do not have cores. This study presents a less costly way for the systematic and objective detection of productive and non-productive zones via well-log data using clustering unsupervised and supervised machine learning algorithms. The method of cluster analysis has been used in order to classify the productive and non-productive reservoir rock groups in the tight reservoir. This was accomplished by assessing the variability of the reservoir characteristics data that are forecasted by looking at the dimensions of the well logs. The Support vector machine as a supervised machine learning algorithm is then used to evaluate the classification accuracy of the unsupervised algorithms based on the clustering labels. The application made use of approximately ten different variables of rock characteristics including zonal depth, effective porosity, permeability, shale volume, water saturation, total organic carbon, young's modulus, Poisson's ratio, brittleness index, and pore size. The findings show that both clustering techniques identified the sweet areas with high accuracy and were less time-consuming.

1. Introduction

Unconventional tight reservoirs are becoming primary hydrocarbon resources owing to the ongoing advancements in exploration theory, the steady rise in the world's demand for oil and gas, the continuous fall in conventional oil and gas output, and the effective use of sophisticated horizontal well drilling and multi-stage hydraulic fracturing methods [1]. Organic quality (OQ), rock quality (RQ), and mechanical quality (MQ) are the three determinants that determine the viability of unconventional resources [2]. The drilling of horizontal wells and the selection of perforation clusters both benefit from mapping sweet spots, which may lead to the maximum production and recovery of unconventional resources if done correctly. In the past, geoscientists have traditionally found sweet spots via the examination of well logs [3]. The use of artificial intelligence and machine learning is one of the most intriguing technologies that has lately entered the area of unconventional reservoirs. Unconventional reservoirs are reservoirs that are not normally found in their natural state. Within the realm of artificial intelligence lies the subfield known as "machine learning," in which intelligence may be produced even without the use of exact programming [4]. The procedures that are utilized to evaluate sweet spots or productive zones in complicated reservoirs may be greatly improved by the use of machine learning algorithms [5]. As can be seen in Fig. 1, machine learning may be broken down into one of two primary categories: either supervised or unsupervised [6].

The process of extracting useful information from data is referred to as analytics. This process utilizes a variety of methods, tools, and processes. The term "techniques" encompasses a wide range of ideas,

* Corresponding author.
E-mail addresses: up958633@myport.ac.uk (K.A. Abbas), a.gharavi@ucl.ac.uk (A. Gharavi).

https://doi.org/10.1016/j.ijin.2022.11.007
Received 11 September 2022; Received in revised form 31 October 2022; Accepted 21 November 2022
Available online 11 December 2022

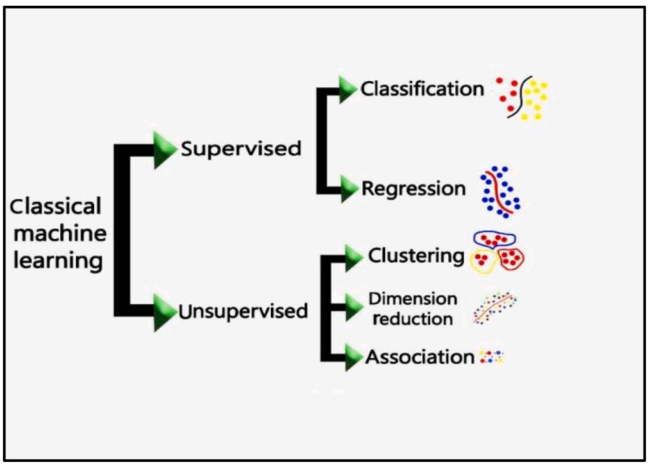

Fig. 1. Machine learning methods [6].

including "artificial intelligence," "machine learning," and "deep learning" algorithms. Machine learning (ML) is a branch of artificial intelligence (AI) that includes a wide range of data processing methods such as classification, regression, and clustering. Two major categories of machine learning are supervised and unsupervised approaches [7]. The most essential components of unsupervised machine learning are the input variables and the anticipated values [8]. In the oil and gas industry, machine learning systems that make use of wireline logs are increasingly being employed to address the geoscientific issues that are faced in the exploration, development, and production of oil and gas [9]. Data driven modelling tools are a potential and alternative strategy to assess the productive zones in unconventional formations. These techniques include statistical data analysis, data clustering and classification, and machine learning algorithms [10]. The extensive use of machine learning algorithms within the oil and gas industry has resulted in an increase in the number of obstacles that must be overcome in order to successfully extract petroleum from the subsurface. Machine learning techniques have been proposed as the best solution for strong learning capability and computational efficiency [11]. Low porosity, significant heterogeneity, and diagenesis all have an impact on the tight sandstone reservoirs. Determination of reservoir and completion qualities of the unconventional reservoirs has emerged as one of the important challenges to be tackled in oil and gas exploration and production since tight reservoirs often exhibit ultra-low petrophysical parameters [12,13]. The sweet spot in the reservoir is the most important part of tight oil and gas exploration, which is directly connected to the process of selecting exploration sites and determining the amount of tight oil and gas resources. In addition, it is necessary in order to ensure the successful development of tight oil and gas. Different scales may be used to locate sweet spots in hydrocarbon reserves. Seismic interpretation methods, such as the discovery of stratigraphic or structural traps, may be used to locate sweet spots at the field scale [14]. Sweet spot identification along the well involves locating areas with appropriate petrophysical properties connected to in-place hydrocarbon volume and formation conductivity, such as formation thickness, rock porosity, fluid saturation, rock permeability, rock relative permeability, and rock wettability. Sweet spot is often influenced by more complicated and unconventional elements in complex formations like tight gas reservoirs and organic-rich mud rocks. Based on seismic characteristics associated with the existence of natural cracks, Glaser et al. (2013) [15] documented the detection of reservoir size sweet spots in shale formations. Evaluation of the original in-place hydrocarbon volume along the well sweet-spot involves assessing formation lithological features and depositional characteristics such total organic content (TOC), maturity, hydrocarbon saturation, and porosity [16].

In complicated reservoirs, factors such as matrix permeability, which takes into account the contribution of micro cracks and kerogen, fluid viscosity, and overpressure are employed to measure formation conductivity. However, there may be only a minor association between hydrocarbons already present and the permeability of virgin formations when looking at eventual output [17]. The most trustworthy way for linking the physical and operational factors to hydrocarbon output is to use field data. The use of field data, on the other hand, is dependent on the accurate assessment of reservoir petrophysical parameters, which may be both time-consuming and costly. In addition, it may be difficult to accurately quantify these petrophysical characteristics in complex reservoirs such as shaly sands and organic-rich mud rocks because of the intrinsic complexity of the rock samples. This is owing to the fact that such reservoirs have a high degree of heterogeneity [18]. The application of machine learning algorithms offers a quick practical technique that is also reliable for analyzing formation parameters. Developing proxy models that can correlate reservoir petrophysical properties and operational parameters to the ultimate economic hydrocarbon recovery from complex reservoirs is possible with the assistance of such empirical models, which can also assist in evaluating the effectiveness of the fracture treatment.

The oil industry has made extensive use of machine learning in a variety of contexts and context-specific applications. However, with the advent of machine learning, the oil industry transitioned to new and improved algorithms and programs that are more likely to solve difficult and complex problems in a timely manner while producing results that are as optimized. Machine learning enables big oil and gas businesses to get timely and reliable data to support their business decisions [19]. Machine learning can be used to make categories (clustering) and predictions (classification or regression) about the future, which enables us to use it to anticipate the occurrence of certain events and attempt to take steps to mitigate or lessen their impact. Since there are so many datasets related to oil and gas exploration and production, machine learning plays another crucial function in the petroleum business. Machine learning can assist in selecting the most crucial information from these datasets. This might indicate the next stage of return on investment (ROI). Taking direct, quick, and accurate judgments and actions is one of the key functions played by machine learning in the business [20]. The oil industry has been using artificial intelligence for decades in a variety of ways and applications. However, with the advent of machine learning, the industry transitioned to new and improved algorithms and programmes that have the tendency to solve difficult and complex operations in a timely manner while producing results that are as optimized as they can be. The main oil and gas businesses can acquire quick and accurate data to support their business operations when they make use of machine learning. In this section, we will explore the areas in which machine learning has been used by the oil and gas sector throughout the course of time [21].

In this investigation, we have used both unsupervised machine learning techniques as well as supervised machine learning algorithms for the purpose of forecasting production and non-production zones. The wireline logs of the tight oil reservoir are the training data that are used for the learning process. In the unsupervised method, we used machine learning techniques called k-means and hierarchical clustering. These techniques attempt to divide the dataset into a certain number of unique clusters that have been predefined, with each data point only belonging to a single one of those groups. For the supervised learning technique, the support vector machine algorithm was used, and the intended output consisted of cluster labels based on the environment of deposition from core data that was added to the training dataset. The findings from each methodology are reviewed in relation to the efficacy of each strategy in predicting zone distribution with a greater degree of precision.

The following will serve as the format for the article:

- In terms of the methodology, the most crucial step is data preprocessing, which consisted of transforming the raw data into a scaled version with a distribution that ranged between 0 and 1, with

0 representing the mean and 1 representing the standard deviation. Furthermore, supervised and unsupervised methods are used to cluster and evaluate the available dataset.

- The results section gives two examples of the use of k-means and hierarchical algorithms to determine the optimal number of clusters, followed by the use of a support vector machine to evaluate the classification accuracy that yields the best outcome. In the results and discussion section, findings are interpreted.
- Analyzing dataset
- In the section devoted to the conclusion, the methodologies and their results were compared, and the ones that proved to be the most effective were highlighted. Additionally, the benefits and advantages of each method on its own were discussed, as well as the algorithms and methodologies that are utilized in machine learning.

2. Methodology

Python-based numpy, pandas, and sklearn libraries were used in order to implement the machine-learning methods that were utilized in the research project. These are several tools that have been created specifically for the purposes of data analytics and machine learning. Both the unsupervised technique, which includes grouping data based on similarities and distance and the supervised approach, in which we have goal data, were used in our efforts to generate classifications based on learning from the data that was made accessible (clusters). At each and every point in the project's development, data cleaning, pre-processing, and visualization are absolutely necessary for the project to be a success in machine learning [22].

Clustering is a set of basic unsupervised learning methods that helps to group the data into meaningful groups that indicate an underlying pattern. These approaches help to organize the data in a way that is more easily understood. The data are ordered into classes that have a high similarity within each class but a low similarity across classes [22]. Ten formation parameters of an unconventional tight oil reservoir for the Nene Marine Field were used to demonstrate their influence in determining which zones were productive and which were not, as well as the overall effectiveness of machine learning methods in enhancing the workflow of data-driven operations [23]. The range and statistical qualities of the features that were chosen to be used for this investigation are shown in Table 1.

2.1. Data preprocessing

The process of normalizing data is often required by a significant number of machine learning algorithms. In order to maximize the effectiveness of the supervised and unsupervised machine learning models', the standardization approach is used to normalize both the inputs and the target variables. Standard scaling normalizes feature values by first subtracting the value of the mean, so that standardized values always have a mean of zero and then dividing by the standard deviation, hence that the resulting distribution has a standard deviation of one. It is possible to write the equation for the standard scaling as follows [24]:

Table 1
Statistics properties of input variables.

Parameters	Range	Mean	Standard Deviation
Depth	2244–2697.6	2463.213	117.036
Porosity (φ)	0.016–1.0	0.479	0.211
Shale Volume (V_{sh})	0.001–1.0	0.387	0.185
Water saturation (S_w)	0.014–1.0	0.312	0.198
Permeability (k)	0.0004–1.0	0.135	0.172
Total Organic Carbon (TOC)	0.107–1.0	0.522	0.188
Young's Modulus	0.099–1.0	0.494	0.203
Poisson Ratio	0.046–1.0	0.455	0.160
Brittleness Indicator	0.032–1.0	0.404	0.185
Pore Size	0.003–1.0	0.212	0.212

$$z_i = \frac{x_i - \mu}{\sigma} \tag{1}$$

$$\mu = \frac{1}{n}\sum_{i=1}^{n}(x_i) \tag{2}$$

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2} \tag{3}$$

Where z_i illustrates the standard score of the ith sample, x_i denotes the ith sample, μ represents the mean value of the samples. In addition, σ signifies the standard deviation of the samples, and n indicates the number of samples in the dataset.

2.2. Unsupervised machine learning

Unsupervised machine learning works with a dataset that has not been labeled or a dataset of unknown structure. The using methods of unsupervised learning that will be able to investigate the structure of the dataset in order to derive relevant information without the direction of a known outcome (target) variable. The purpose of clustering is to discover a natural grouping in the data in such a way that the items included within the same cluster are more comparable to one another than they are to those contained within separate clusters [25].

2.2.1. Finding similarities using distances

The methods for clustering data begin with the presumption that the information being analyzed contains subsets that are comparable to or identical with one another. One method for determining the degree of resemblance between two things is to calculate the distance between them using a variety of metrics [26].

2.2.1.1. Euclidean distance. The euclidean distance is the radial distance between two samples or instances in the dataset [26]. Calculating the euclidean distance between two observations x_1 and x_2 can be accomplished by the following equation:

$$D(x_1, x_2) = \sqrt{\sum (x_{i1} - x_{i2})^2} \tag{4}$$

Where x_{i1} represents the value of the ith feature for the first observation and x_{i2} shows the value of ith feature for second observation.

2.2.2. K-means clustering

K-means is an unsupervised approach for machine learning that is used for the division of clusters. It aims to accomplish the objective of forming k clusters and allocating data points to them in such a manner that there is high similarity within each cluster (internal cohesion), and low similarity between clusters (external separation). The K-means clustering method uses the following stages to complete its process as can be shown in Fig. 2 [26]:

- Step one: initially, choose a number of clusters (which can be optimized later).
- Step two: select randomly the cluster centroids (number of centroids = number of clusters).
- Step three: assign each data point to the nearest cluster centroid by calculating the distance between each data point and centroid. The commonly used distance calculation for k-means clustering is euclidean distance, a scale value that measures the distance between two data points.
- Step four: update cluster centroid position. A Centroid is computed as the average of data points in a cluster.

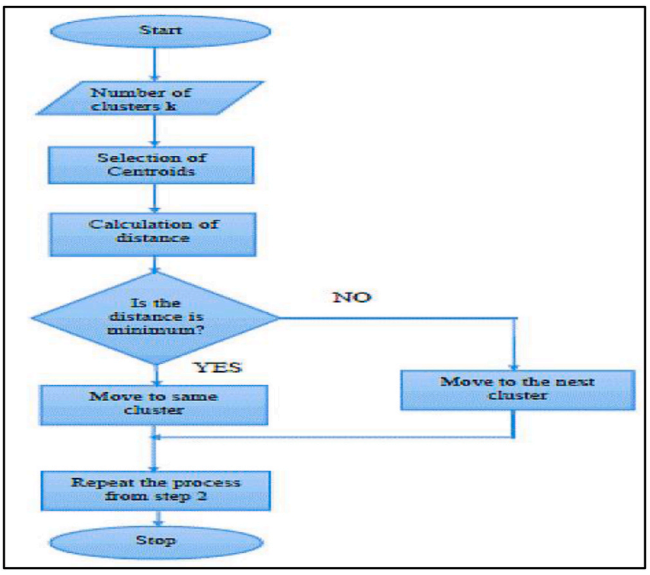

Fig. 2. Flowchart of k-means clustering [20].

- Step five: repeat steps three and four until all data points are closest to the cluster centroid and no data point will switch cluster once this happen, the algorithm will stop working.

2.2.3. Hierarchical clustering

The hierarchical clustering technique is known to be one of the most frequent and commonly utilized methods for the process of clustering in the field of machine learning. When using the agglomerative clustering method, one begins at the cluster leaf and works their way upward until one reaches the cluster root. The workflow of the hierarchical clustering method can be seen in Fig. 3. The following stages are used to build clusters via the use of hierarchical clustering [26]:

- Step one: begin with each data point located in its own individual cluster.
- Step two: find the data points that are the closest to one another (using a measurement that is suitable for distance), then combine those points into a cluster.
- Step three: return to step 2 and continue doing so until all of the data points are combined into a single cluster.

It is essential to calculate the degree of similarity between two clusters before deciding whether to combine or split them. The degree of similarity between two clusters may be calculated using a few different methods, including the following [27]:

- Single-Linkage
- Complete Linkage
- Group Average
- Distance Between Centroids
- Ward's Method

In the course of the current investigation, complete linkage between two clusters was used in order to determine the degree of similarity that existed between them. Complete linkage may be defined as the distance between two clusters C_1 and C_2 that equal to the longest distance that can be found between the points P_i and P_j, in which P_i belongs to cluster C_1 and P_j belongs to cluster C_2 [26].

$$D(C_1,C_2)=\max_{\substack{P_i\in C1\\P_j\in C2}}\left(d\left(P_i,P_j\right)\right) \tag{5}$$

2.3. Supervised machine learning

Supervised machine learning techniques are used to accomplish aiming to learn the link between parameters and their output. Indeed, It accomplishes this by developing a function that relates the inputs to the outcome. The model must be fitted/trained using features and a target dataset, which are both necessary in supervised learning. Afterward, the succeeding classifier is utilized to forecast an unidentified attributes dataset that has no matching target data. Regression algorithms and classification algorithms are two types of supervised algorithms. The most significant distinction between regression and classification is that the goal dataset in regression is a collection of continuous variables, while in classification the target dataset is a collection of categorical variables or set of discrete variables. A typical supervised learning method can handle equal regression and classification tasks; however, the procedure is normally developed to withstand one instance and then changed to tackle the other situation. A support vector algorithm is an example of this kind of method [27].

2.3.1. Support vector machine

The support vector machine is often used for solving issues involving regression and classification. In order to categories the data points, the fundamental idea behind support vector machines is to create a hyperplane or a series of hyperplanes in a high-dimensional feature space. In order to translate the input vectors into a space with a high dimension, kernel functions like linear, polynomial, and radial basis function (RBF) are used as mapping tools. The RBF algorithm is the one that is employed as the kernel function in Support vector classifier (SVC). In order to prevent the issues that arise from overfitting, there are two crucial parameters namely, the penalty parameter (C) and the kernel parameter (γ) that need to be selected with care. The greater the value of the penalty parameter that is applied, the greater the amount of mistake that is punished [27]. In this study, the penalty parameter and the kernel parameter are fine-tuned to enhance the performance of SVC. The values of C and γ that provide the best results are 215.44 and 0.001 correspondingly these values are obtained by using Grid Search Cross Validation technique [27]. The workflow of the support vector machine method can be seen in Fig. 4.

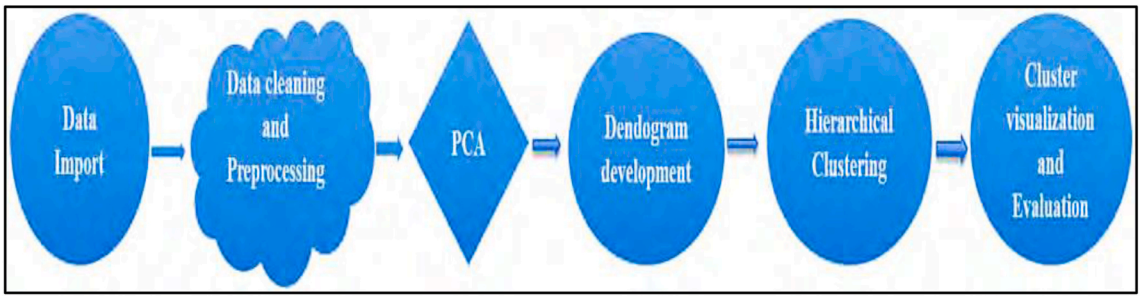

Fig. 3. Workflow for Hierarchical clustering [21].

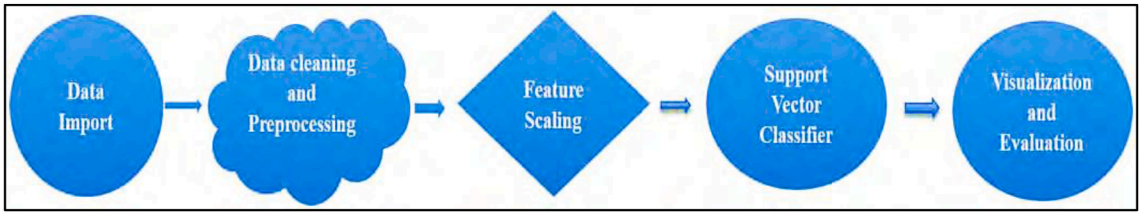

Fig. 4. Workflow for support vector machine (SVM) [21].

3. Results and discussion

3.1. K-means results

It is possible that the k-means algorithm's widespread use is due to the fact that it is not only incredibly simple to develop but also highly effective in terms of its use of computer resources when compared to other clustering methods. In the field of prototype-based clustering, the k-means method is considered to be a representative example. Prototype-based clustering implies that each cluster is represented by a pattern, which might be the centroid (average) of comparable points with continuous characteristics or the medoid (the most typical or often occurring point) with categorical features. One of the limitations of the k-means clustering method is that it requires us to choose the number of clusters, k, in advance, despite the fact that this algorithm is particularly effective at locating groups of data that have a spherical form. A poor clustering performance might be the consequence of an incorrect decision for the value of k [24].

3.1.1. Finding optimal number of K-means clusters

3.1.1.1. Elbow curve method. The cluster's variance will be maximum if it is assumed that all of the data points belong to only one cluster. The overall variance across all clusters will start to decrease as the number of clusters rises. However, if it supposes that each data point is a cluster by itself, the total variance will be zero. As a result, the Elbow curve approach takes the quantity of clusters into account when calculating the variance. The ideal cluster size is determined such that increasing the number of clusters does not dramatically alter the variance. The approach relies on the idea of reducing the inertia, often known as the within clustering sum of squares (WCSS) that occur inside a cluster. Inertia can be expressed by Ref. [25]:

$$\mathrm{WCSS} = \sum_{i=1}^{n} \min_{\mu_j \in C} \left(\left\| X_i - \mu_j \right\|^2 \right) \tag{6}$$

Where X_i represents each data point in the dataset and μ_j shows the mean of points of each cluster in the dataset (centroid of the cluster) while n is the total number of the records in the available dataset. The elbow diagram, which may be seen as a plot of WCSS versus the number of clusters, is shown in Fig. 5. Based on this graph, it can be deduced that the elbow is created with a K value of somewhere near 2. After K = 2, the WCSS begins a gradual decline this is due to the samples being nearer their designated centroids., and a reduction in variance begins as the number of clusters grows.

3.1.1.2. Silhouette analysis. The silhouette coefficient, which is calculated using the silhouette analysis, is used to provide an evaluation of the cluster splitting process's overall quality. The following is the formula that must be used to calculate the silhouette coefficient:

$$S_{(i)} = \frac{b_{(i)} - a_{(i)}}{\mathrm{Max}\left(a_{(i)}, b_{(i)}\right)} \tag{7}$$

The silhouette coefficient for a particular data point is denoted by the symbol $S_{(i)}$.The $a_{(i)}$ represents the average distance that separates this particular data point from all of the other data points that are included within the same cluster. The $b_{(i)}$ is the average distance that separates this particular data point from all of the other data points that are located in the nearest cluster. $S_{(i)}$ may take on any value between −1 and 1 [20].

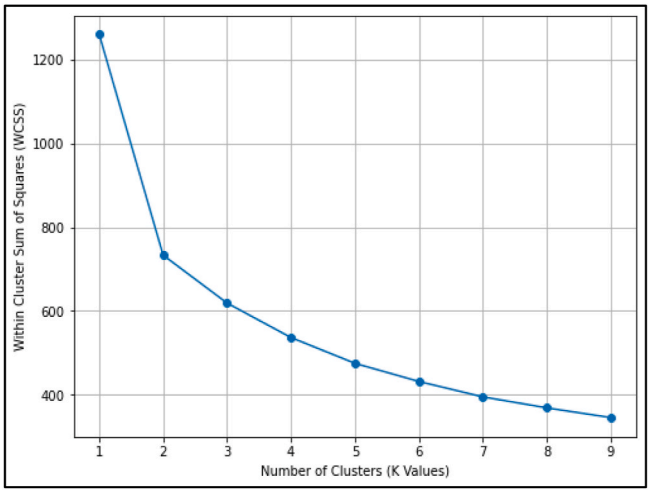

Fig. 5. Elbow diagram for k means clustering.

- If $S_{(i)}$ is equal to 1: it indicates that the data point in question is relatively near to other points that belong to the same cluster while being far apart from points that belong to the neighbor cluster.
- If $S_{(i)}$ equals 0: it shows that the data point in question is located very close to the edge of its cluster.
- If $S_{(i)}$ is equal to −1: it illustrates this particular data point has been placed in the incorrect cluster.

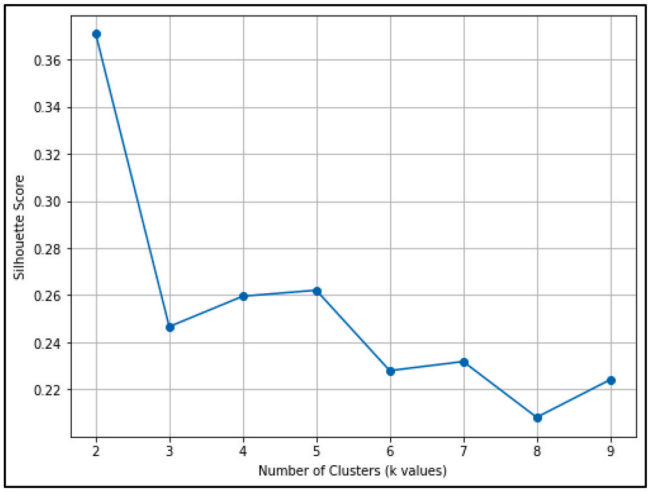

Fig. 6. Silhouette analysis for k means clustering.

Fig. 6 displays the final silhouette coefficient, whereas Fig. 7 displays the final silhouette diagram. The silhouette score is calculated by averaging the silhouette coefficient of all of the data points. The silhouette coefficients show how successfully the clustering is done, with greater values indicating better clustering. Given that the silhouette coefficient is the greatest of the measured quantities, and that k = 2 corresponds to the hypothesis that there are exactly two clusters, the value of the coefficient is about (0.3707).

3.1.1.3. Davies bouldin index analysis. The davies bouldin index (DBI) is a statistic used to evaluate various clustering techniques. Its most common use is checking whether or not a certain number of clusters was correctly divided using the k-means clustering algorithm. It's possible to figure out the Davies-Bouldin Index by doing the following:

$$D_{(i,j)} = \frac{d_{(i)} + d_{(j)}}{d_{(i,j)}} \tag{8}$$

The davies bouldin index for a certain set of clusters is denoted by the notation $D_{(i,j)}$ (e.g., clusters i and j). The $d_{(i)}$ and $d_{(j)}$ represent the average distance between each point and its corresponding cluster's centroid with i and j being the clusters in question. The distance between the centers of clusters i and j is denoted by the notation $d_{(i,j)}$ [20].

Fig. 8 shows the findings of a study into the average maximum Davies Bouldin index for each cluster. Cluster 2 is the lowest-valued cluster, with a value of (1.0608). This is the lowest total value in the graph, and it clearly shows that there are just two clusters in the available dataset.

The output variable that is shown in Table 2 includes labels that may be consulted in order to ascertain which cluster observations fall into. It can be noted that the first three observations are classed as belonging to cluster 1 whereas the fourth, seventh, and eighth records are categorized as belonging to cluster 0 respectively and the remaining observations will be labeled based on two clusters 0 and 1.

3.1.2. SVM classifier accuracy based on K-means clustering

The confusion matrix and the classification report based on the support vector machine classifier were applied in order to illustrate the effective performance of the clustering method. Fig. 9 and Table 3 both provide visual representations of this interpretation. By identifying the number of observations that correlate to the true positive and false negative, the confusion matrix displays a relatively high level of performance of the support vector machine that relies on k means clustering results. To be more exact, 19 of the 20 records point to an accurate categorization, whereas one of the records provide evidence of incorrect classification. On the other hand, the classification report demonstrates

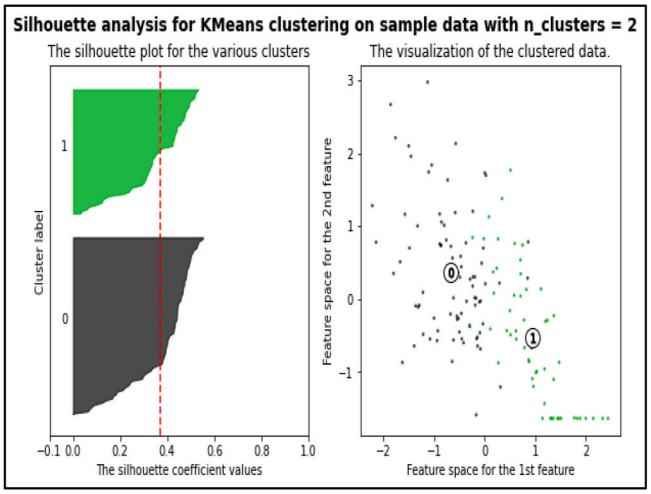

Fig. 7. Silhouette diagram for k means clustering.

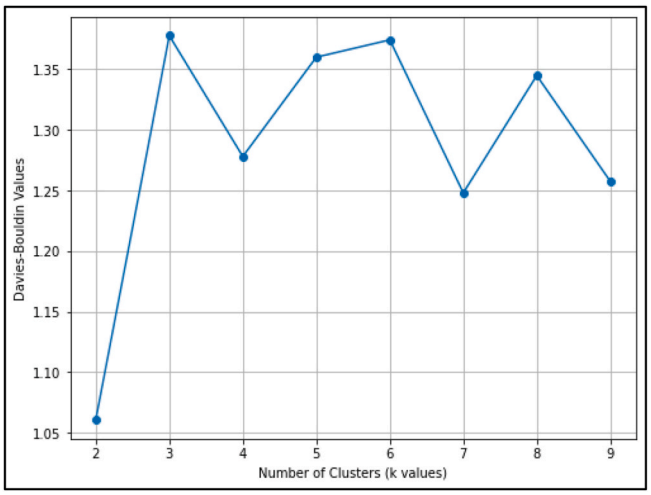

Fig. 8. Davies bouldin index analysis for k means clustering.

the excellent performance of the model in identifying the productive zone (class 1). More precisely, the recall and the precision of class 1 refer to approximately 95% and 100% respectively, and the overall accuracy of the model is 98%, with 2% of the error occurring during the course of implementation.

3.2. Hierarchical clustering results

An alternate method to prototype-based clustering is hierarchical clustering. Hierarchical clustering techniques have the benefit of allowing us to display dendrogram (visualizations of a binary hierarchical clustering), which might also aid in the understanding of the findings by generating useful categories. We do not have to define the number of clusters up front, which is another helpful benefit of this hierarchical method. Agglomerative and divisive hierarchical clustering are the two basic methods used in this process. In divisive hierarchical clustering, we begin with a single cluster that includes all of our samples and then divide it into successively smaller clusters, reducing the size of each cluster until it contains a single sample. The opposing strategy, agglomerative clustering, will be the main topic of this paper. Each sample is first clustered individually, and then we combine the closest pairings of clusters until only one cluster is left [24].

3.2.1. Finding optimal number of clusters

3.2.1.1. Dendrogram graph. Dendrogram graph is used in hierarchical clustering to describe the given data as a cluster tree. Each group relates to two or more successor groups. The groupings are then structured in the form of a tree and layered inside one another, which should ultimately result in a sensible categorization system. A visual representation of the data included in the whole set is provided by the process where clusters at one level combine with clusters at the level above using a degree of similarity [26].

Fig. 10 shows the dendrogram graph, which may be seen as a plot of the euclidean distance between each data point in the dataset on the y-axis versus the number of data points on the x-axis. The dendrogram vertical lines indicate how far apart certain groups are from one another. Based on this figure, it can be deduced that the threshold distance for the vertical line is around 10. The number of clusters will be equal to the number of vertical lines that are crossed by the line that was constructed using the threshold. The findings indicate since the red line intersects 2 vertical lines, as a result the number of clusters is 2 clusters.

Fig. 11 shows how a heat map combined with a hierarchical clustering dendrogram may be used in practice to graphically represent the range of values present in a given sample matrix. This heat map shows

Table 2
Independent variables and k-means identified clusters.

Index	Dependent Variables (Features)									Independent Variable by K-Means Clustering
	Porosity	V_{sh}	S_w	Permeability	TOC	Youngs Modulus	Poisson's Ratio	Brittlenss Indicator	Pore Size	Labels
0	1.513601	−1.633796	−0.982498	1.036566	2.245967	−0.226440	−0.384903	−0.518592	0.971988	1
1	1.019131	−1.006603	−1.432071	0.455948	1.508376	−0.389675	−0.672526	−0.781551	0.556391	1
2	1.305443	−1.633796	0.042748	0.886517	0.534255	−1.301511	0.255474	−0.301786	0.911287	1
3	−0.634463	0.565501	1.121402	−0.724972	−0.546211	1.347443	0.147021	0.426387	−0.847184	0
4	0.274548	−0.872430	−0.753724	0.099126	0.655550	−0.159958	−0.866005	−0.879109	0.335116	1
5	1.468537	−1.633796	0.057589	2.285428	0.445124	−1.127410	0.183582	−0.304410	2.196231	1
6	−1.046692	1.838864	0.146301	−0.730672	−1.370985	−0.093718	0.568016	0.398538	−0.841632	0
7	−0.665457	0.724875	−0.595590	−0.728558	−0.586172	0.864283	1.412958	1.855551	−0.853515	0
8	1.188035	−1.430469	3.459615	1.732896	1.663732	−1.170127	−0.656656	−0.937337	1.833714	1
9	0.088741	1.129033	−0.630843	0.246650	0.167883	0.166663	−0.466168	−0.418599	0.680754	1

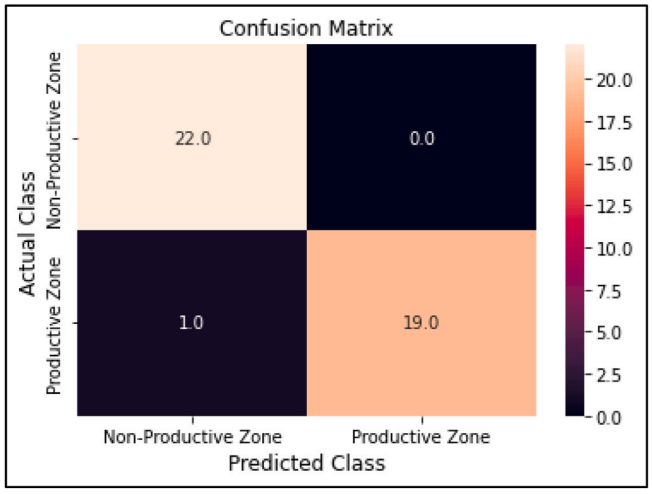

Fig. 9. Support vector machine confusion matrix based on K-means clustering.

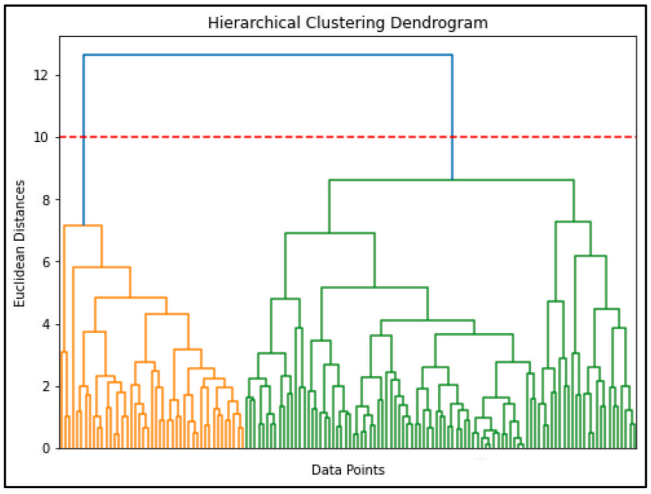

Fig. 10. Dendrogram illustration graph.

Table 3
Support vector machine calssification report based on K-means clustering.

	Precision	Recall	F1-Score	Support
0	0.96	1	0.98	22
1	1.00	0.95	0.97	20
Accuracy			0.98	42
Macro avg	0.98	0.97	0.98	42
Weighted avg	0.98	0.98	0.98	42

how the dendrogram's clustering of the samples is represented in the order of the rows. A simple dendrogram and a heat map with color-coded values for each sample and feature provide us a fantastic picture of the data set. The degree to which rows are similar or dissimilar to one another, as well as the node to which each row belongs, may be seen in the row dendrogram.

3.2.1.2. Silhouette coefficient. The silhouette score is used to assess the quality of clusters generated using clustering algorithms in terms of how well samples are grouped with other samples that are similar to each other. The silhouette scores are shown in Fig. 12 in relation to the number of clusters with the greatest score being about (0.3652) which places it towards the clustered number two (k = 2). This provides a stronger signal that the quality of the grouping improves in direct proportion to the silhouette score.

The predicted output that is displayed in Table 4 has labels that may be referred to in order to determine which cluster observations belong to

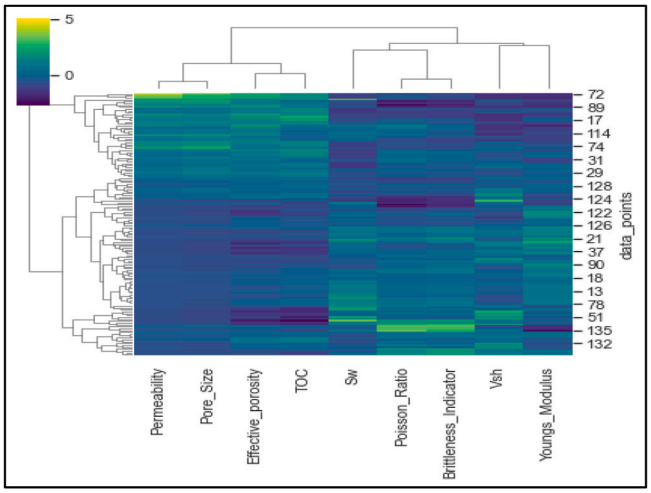

Fig. 11. Dendrogram heat map illustration.

which group. It is important to note that the first three observations have been labeled as belonging to cluster 1, while the fourth and fifth recordings have been labeled as belonging to cluster 0. The following records will be grouped based on the two clusters 0 and 1, respectively.

3.2.2. SVM classifier accuracy based on hierarchical clustering

In order to demonstrate the efficient operation of the hierarchical

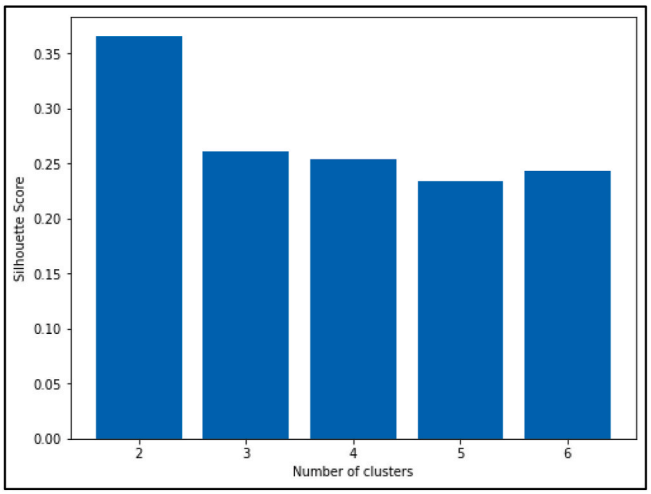

Fig. 12. Silhouette coefficient for Hierarchical clustering.

clustering approach, the confusion matrix as well as a classification report that was generated using a support vector machine classifier were both used. Fig. 13 and Table 5 both provide visual representations of this interpretation. A relatively high degree of performance of the support vector machine is shown by the confusion matrix, which is dependent on the results of hierarchical clustering. This is accomplished by determining the number of observations that correspond to the true positive and the false negative. To be more specific, 15 of the 15 records lead to a valid categorization, whilst none of the records give proof of any inaccurate classification. In contrast, the report on categorization reveals that the model performs very well when it comes to locating the yielding zone (class 1). The recall and the precision of class 1 correspond to 100% and 94% respectively, while the total accuracy of the model is 98%, with 2% of the mistake happening during the process of implementation.

4. Conclusions

In this article, an intelligent method for predicting sweet spots and non-sweet spots in tight unconventional reservoirs using unsupervised and supervised machine learning algorithms are proposed. The results that were produced by using k-means and hierarchical clustering to create productive and non-productive zones give a method of constructing zonal prediction that is strictly driven by data and has less bias when compared to the manual method that is dependent on human assessment. The supervised method support vector machine offers a way to anticipate the accuracy of the clusters to a very large degree in core and wireline logging of the well. Although the technique above produced a very accurate forecast with the basic data that was available, more data will result in a higher accuracy score and a more accurate prediction. In addition to this, the advantage of the grid search cross-validation tool was centered on hyperparameter tweaking. This is done in order to gain better parameters, which ultimately leads to improved support vector machine outcomes. The findings obtained from both the unsupervised and the supervised approaches help to reduce the amount of bias that is present in the zonal identification modeling. We came to the conclusion that predicting sweet spots based on available core and wireline logging data and distributing for uncored depth intervals is typically more reliable than traditional manual zonal prediction. This is despite the fact that both methods present a more efficient way of predicting productive zones in comparison to the traditional method of manual zonal prediction. On the other hand, the zonal prediction may be performed using a classification approach that is based on clustering even if core data are not available. Because a more accurate predictive model may be developed with increased amounts of data to train the machine with, it is essential to keep in mind that huge amounts of data that are organized concisely are an essential component

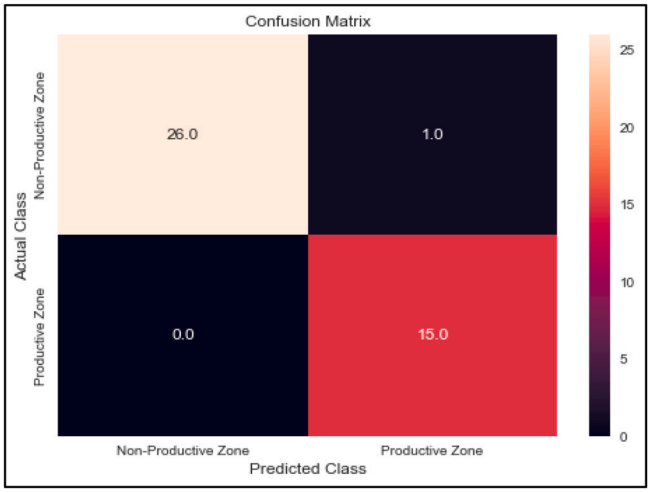

Fig. 13. Support vector machine confusion matrix based on hierarchical clustering.

Table 5
Support vector machine calssification report based on hierarchical clustering.

	Precision	Recall	F1-Score	Support
0	1.00	0.96	0.98	27
1	0.94	1.00	0.97	15
Accuracy			0.98	42
Macro avg	0.97	0.98	0.97	42
Weighted avg	0.98	0.98	0.98	42

Table 4
Independent variables and hierarchical predicted clusters.

Index	Dependent Variables (Features)									Independent Variable by Hierarchical Clustering
	Porosity	V_{sh}	S_w	Permeability	TOC	Youngs Modulus	Poisson's Ratio	Brittlenss Indicator	Pore Size	Labels
0	1.513601	−1.633796	−0.982498	1.036566	2.245967	−0.226440	−0.384903	−0.518592	0.971988	1
1	1.019131	−1.006603	−1.432071	0.455948	1.508376	−0.389675	−0.672526	−0.781551	0.556391	1
2	1.305443	−1.633796	0.042748	0.886517	0.534255	−1.301511	0.255474	−0.301786	0.911287	1
3	−0.634463	0.565501	1.121402	−0.724972	−0.546211	1.347443	0.147021	0.426387	−0.847184	0
4	0.274548	−0.872430	−0.753724	0.099126	0.655550	−0.159958	−0.866005	−0.879109	0.335116	0
5	1.468537	−1.633796	0.057589	2.285428	0.445124	−1.127410	0.183582	−0.304410	2.196231	1
6	−1.046692	1.838864	0.146301	−0.730672	−1.370985	−0.093718	0.568016	0.398538	−0.841632	0
7	−0.665457	0.724875	−0.595590	−0.728558	−0.586172	0.864283	1.412958	1.855551	−0.853515	0
8	1.188035	−1.430469	3.459615	1.732896	1.663732	−1.170127	−0.656656	−0.937337	1.833714	1
9	0.088741	1.129033	−0.630843	0.246650	0.167883	0.166663	−0.466168	−0.418599	0.680754	0

of any machine learning algorithm.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] R.A. Kerr, Energy. Natural gas from shale bursts onto the scene, Science (New York, NY) 328 (5986) (2010) 1624–1626.
[2] J.B. Aldrich, J.P. Seidle, Sweet spot identification and optimization in unconventional reservoirs, Mt. Geol. 52 (3) (2018) 5–12.
[3] J. Tang, B. Fan, L. Xiao, S. Tian, F. Zhang, L. Zhang, D. Weitz, A new ensemble machine-learning framework for searching sweet spots in shale reservoirs, SPE J. 26 (2021) 482–497, 01.
[4] A. Casian, M. Zannato, Predicting Teamfight Tactics Results with Machine Learning Techniques, 2020.
[5] S. Tandon, Integrating machine learning in identifying sweet spots in unconventional formations, in: SPE Western Regional Meeting, OnePetro, 2019, April.
[6] S. Aghabozorgi, A.S. Shirkhorshidi, T.Y. Wah, Time-series clustering–a decade review, Inf. Syst. 53 (2015) 16–38.
[7] B. Hall, Facies classification using machine learning, Lead. Edge 35 (10) (2016) 906–909.
[8] U. Ashraf, H. Zhang, A. Anees, H. Nasir Mangi, M. Ali, Z. Ullah, X. Zhang, Application of unconventional seismic attributes and unsupervised machine learning for the identification of fault and fracture network, Appl. Sci. 10 (11) (2020) 3864.
[9] P.P. Mandal, R. Rezaee, Facies classification with different machine learning algorithm–An efficient artificial intelligence technique for improved classification, ASEG Extended Abstracts 2019 (1) (2019) 1–6.
[10] M. Shaheen, M. Shahbaz, A. Guergachi, Data mining applications in hydrocarbon exploration, Artif. Intell. Rev. 35 (1) (2011) 1–18.
[11] B. Shelley, S. Stephenson, The use of artificial neural networks in completion stimulation and design, Comput. Geosci. 26 (8) (2000) 941–951.
[12] J. Lai, G. Wang, Y. Ran, Z. Zhou, Y. Cui, Impact of diagenesis on the reservoir quality of tight oil sandstones: the case of Upper Triassic Yanchang Formation Chang 7 oil layers in Ordos Basin, China, J. Petrol. Sci. Eng. 145 (2016) 54–65.
[13] D. Zheng, X. Pang, F. Jiang, T. Liu, X. Shao, Y. Huyan, Characteristics and controlling factors of tight sandstone gas reservoirs in the Upper Paleozoic strata of Linxing area in the Ordos Basin, China, J. Nat. Gas Sci. Eng. 75 (2020), 103135.
[14] M.R. Giles, S.H. Tennant, Sweet spots: what are they, where are they, how are they created and are they important anyway?, in: SPE/EAGE European Unconventional *Resources Conference And Exhibition*, vol. 2014 European Association of Geoscientists & Engineers, 2014, February, pp. 1–6, 1.
[15] K.S. Glaser, C.K. Miller, G.M. Johnson, B. Toelle, R.L. Kleinberg, P. Miller, W. D. Pennington, Seeking the sweet spot: reservoir and completion quality in organic shales, Oilfield Rev. 25 (4) (2013) 16–29.
[16] S. Tandon, Z. Heidari, Effect of internal magnetic-field gradients on nuclear-magnetic-resonance measurements and nuclear-magnetic-resonance-based pore-network characterization, SPE Reservoir Eval. Eng. 21 (2018) 609–625, 03.
[17] R.S. Pilcher, J. McDonough Ciosek, K. McArthur, J.C. Hohman, P. Schmitz, Ranking production potential based on key geological drivers-Bakken case study, in: International Petroleum Technology *Conference*, OnePetro, 2011, November.
[18] S. Tandon, Z. Heidari, H. Daigle, Pore-scale evaluation of nuclear magnetic resonance measurements in organic-rich mud rocks using numerical modeling, in: SPE/AAPG/SEG Unconventional *Resources Technology Conference*, OnePetro, 2017, July.
[19] S. Tandon, Integrating machine learning in identifying sweet spots in unconventional formations, in: SPE Western Regional Meeting, OnePetro, 2019, April.
[20] H. Saleh, The Machine Learning Workshop, Packt Publishing, 2020.
[21] D.O. Fadokun, I.B. Oshilike, M.O. Onyekonwu, Supervised and unsupervised machine learning approach in facies prediction, in: SPE Nigeria Annual International *Conference And Exhibition*, OnePetro, 2020, August.
[22] A. Gharavi, M. Hassan, J. Gholinezhad, H. Ghoochaninejad, H. Barati, J. Buick, K. A. Abbas, Application of machine learning techniques for identifying productive zones in unconventional reservoir, International Journal of Intelligent Networks 3 (2022) 87–101.
[23] S. Raschka, V. Mirjalili, Machine Learning and Deep Learning with Python, Scikit-Learn and TensorFlow, vol. 189, Packt Publishing, UK, 2017, p. 195.
[24] M. Swamynathan, Mastering machine learning with python in six steps: a practical implementation guide to predictive data analytics using python, Apress (2019).
[25] A.A. Patel, Hands-on Unsupervised Learning Using Python: How to Build Applied Machine Learning Solutions from Unlabeled Data, O'Reilly Media, 2019.
[26] A. Sircar, K. Yadav, K. Rayavarapu, N. Bist, H. Oza, Application of Machine Learning and Artificial Intelligence in Oil and Gas Industry, Petroleum Research, 2021.
[27] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.