Unsupervised machine learning techniques to detect oil and gas presence in hydrocarbon reservoirs.

Harish Dhinakaran

H00365604

A Dissertation submitted in fulfilment of the requirements for the degree of

MSc in Artificial Intelligence

Student Declaration of Authorship

Course code and name:	F21MP – Masters Project and Dissertation
Type of assessment:	**Individual**
Coursework Title:	Masters Dissertation
Student Name:	Harish Dhinakaran
Student ID Number:	H00365604

Declaration of authorship. By signing this form:

- **I declare** that the work I have submitted for individual assessment OR the work I have contributed to a group assessment, is entirely my own. I have NOT taken the ideas, writings or inventions of another person and used these as if they were my own. My submission or my contribution to a group submission is expressed in my own words. Any uses made within this work of the ideas, writings or inventions of others, or of any existing sources of information (books, journals, websites, etc.) are properly acknowledged and listed in the references and/or acknowledgements section.
- I confirm that I have read, understood and followed the University's Regulations on plagiarism as published on the University's website, and that I am aware of the penalties that I will face should I not adhere to the University Regulations.
- I confirm that I have read, understood and avoided the different types of plagiarism explained in the University guidance on Academic Integrity and Plagiarism

Student Signature *(type your name):* *Harish Dhinakaran*

Date: *15/08/2022*

Abstract

Well logging is an important part of the oil and gas industry which has been performed for many years. It gives engineers the ability to determine hydrocarbon formations along the reservoirs. This study addresses the issue of manual well log interpretation in hydrocarbon exploration being a tedious process. With the advancement of machine learning algorithms over the years, it is possible to use these to our advantage to help simplify the process of manually interpreting well logs or act as an aid in determining hydrocarbon bearing regions. In this study different unsupervised machine learning methods are used to identify potential hydrocarbon extraction points based on the obtained manual interpretation from the well log data and are compared in terms of which method proves to be the most effective. Dimensionality reduction is a useful tool to reduce high dimensional data into smaller dimensions. In this study PCA was used as a dimensionality reduction technique and was tested along with unsupervised machine learning methods such as K-Means clustering and SOM to find useful patterns of data in the well logs. They proved to be effective in clustering regions of hydrocarbons, after comparison with the manually interpreted regions. This study also discussed the influence of certain attributes such as gamma ray on K-Mean's ability to cluster data and how high variance attributes influenced the performance of PCA.

Acknowledgements

I would like to acknowledge and express my sincere gratitude to my research supervisor **Dr. Smitha Kumar** *for her continuous guidance, support, and patience. I would also like to thank my family for supporting and encouraging me.*

Contents

List of Tables

List of Figures

Abbreviations

DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DPHI	Density Porosity Hydrogen Index
HI	Hydrogen Index
NPHI	Neutron Porosity Hydrogen Index
PCA	Principal Component Analysis
RHOB	Bulk Density(ρ_b)
SOM	Self Organizing Maps
t-SNE	t-Stochastic Neighbour Embedding
WCSS	Within-Cluster-Sum-of-Squares

1. Introduction

Hydrocarbon exploration is an important process in the Oil and Gas industry [1]. In this study, the aim is to contribute further toward the development of machine learning in the oil and gas industry using a novel approach. It is often a time-consuming process to manually interpret the data from well logs to identify profitable regions of oil and gas reservoirs if any [2]. By identifying clusters and patterns of data in said well logs we can investigate the underlying hydrocarbon accumulation. The main approach for conducting this research was inspired by lithological classifications of well log data into discrete rock facies [3]. Lithology is the study of the general physical characteristics of rock formation in a specific area. The application aspect of this study can prove effective in reducing the amount of time and effort spent by geologists and engineers in interpreting well logs. In oil and gas exploration, well log data have been used for two main reasons. To identify hydrocarbon bearing zones and to weigh out the cost to profit ratio of the detected hydrocarbon accumulation before the start of the drilling process. Unsupervised machine learning methods such as clustering are frequently used to find similar patterns in data or groups in unlabelled data. The goal of this study is to implement different unsupervised learning methods and dimensionality reduction techniques to the well log data and investigate whether the methods prove effective in grouping data into clusters of hydrocarbon-bearing zones. The different clustering methods used are also compared.

1.1. Aim and Objectives

This study aims to successfully identify regions of hydrocarbon accumulations on a well log data using the means of unsupervised machine learning.

Furthermore, the specific objectives are as follows: -

- Manual interpretation of the well log dataset to identify hydrocarbon bearing regions.
- Reducing a high dimensional well log data into a bidimensional dataset.
- Experimenting with different clustering methods on the dataset to identify hydrocarbon bearing regions and comparing how effective the different methods are.

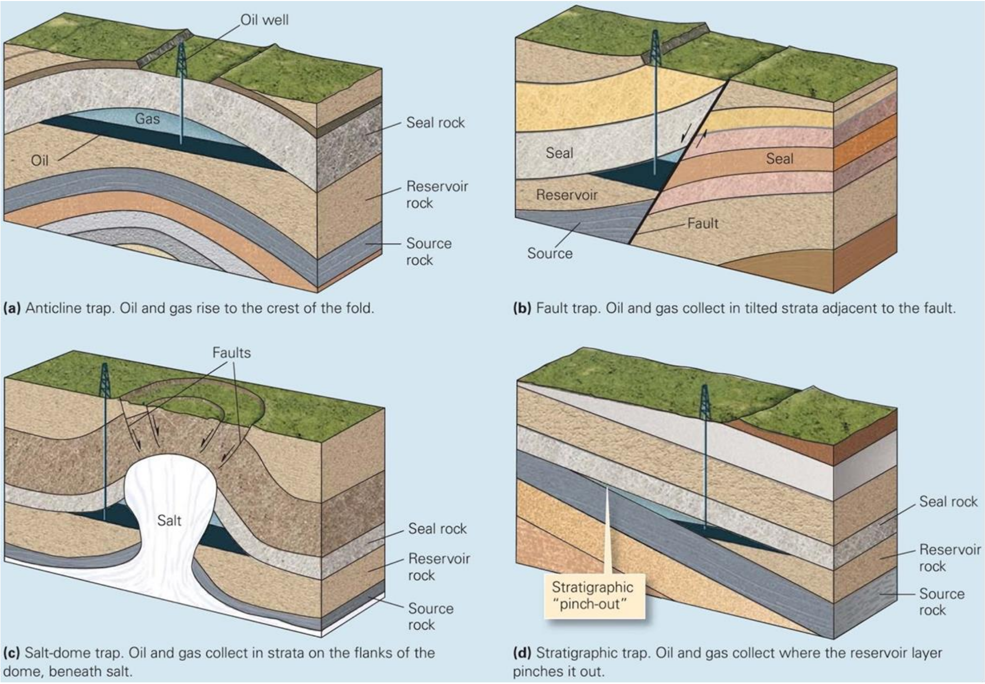
2. Background Study and Literature

In this section a detailed understanding is provided about hydrocarbon exploration and the terminologies associated with wireline logging methods of oil and gas exploration data. Unsupervised machine learning and dimensionality reduction methods are also discussed.

2.1 Hydrocarbon exploration

Hydrocarbon exploration (oil and gas) is the exploration done by petroleum engineers and geophysicists to detect regions of hydrocarbon accumulation, mainly petroleum and natural gas, on the earth's surface. Traditional hydrocarbon exploration and evaluation mainly focus on traps, i.e., areas with favourable conditions of source, reservoir, caprock, trap, migration, and preservation. At present, it is unclear the exact formation mechanisms of these kind of lithologic-stratigraphic reservoirs [1].

Exploration Geophysics is a subset of geophysics that uses physical methods on the earth's surface to measure the physical qualities of the subsurface as well as anomalies in those properties, such as seismic, gravitational, magnetic, electrical, and electromagnetic approaches. It is used to infer the presence of economically valuable geological deposits such as fossil fuels and other hydrocarbons, as well as to provide information on the geological position of the deposit [4]



(a) Anticline trap. Oil and gas rise to the crest of the fold.

(b) Fault trap. Oil and gas collect in tilted strata adjacent to the fault.

(c) Salt-dome trap. Oil and gas collect in strata on the flanks of the dome, beneath salt.

(d) Stratigraphic trap. Oil and gas collect where the reservoir layer pinches it out.

Figure 1 - Different types of naturally occurring hydrocarbon traps [5]

The five main elements of a petroleum system are [2]:-

- Source rock - Petroleum is composed of hydrocarbons that was formed over millions of years due to the decaying of plant and animal remains. As bacteria break down organic plants and animal material, the layers of sediment settled on top. Over time the layers of sediment are transformed into sandstone, limestone, and other sedimentary rock due to heat and pressure, while the organic matter was transformed into petroleum.
- Reservoir Rock – Reservoir rocks are permeable rocks where hydrocarbons accumulate. The porous nature of these rocks allows for the hydrocarbons to be extracted. The volume of hydrocarbons that a reservoir can store depends on the porosity i.e larger the porosity more hydrocarbons it can store.

- Seal or Cap Rock – Seal or cap rock is present above the reservoir rock. Unlike reservoir rocks, these rocks have a low permeability that prevents the migration of oil and gas. If such a barrier wasn't placed, the oil and gas would float to the surface and form oil seeps.
- Trap – A trap can be defined as the overall configuration or system of the hydrocarbon regions. The different types of naturally occurring hydrogen traps are shown in figure 1.
- Migration – Migration is the process by which hydrocarbons move from the source rock to the trap.

2.2 Well log interpretation

Well logging refers to the logging of continuously obtained measurements within wellbores from one region of depth to another in periodic intervals of distance [6]. The curves on the resulting log are used for analysis and indicate the presence of hydrocarbons if any and to what extent [7]. Some of the useful logs that help us in identifying the presence of hydrocarbons are resistivity, gamma ray, density porosity and neutron porosity logs [8].

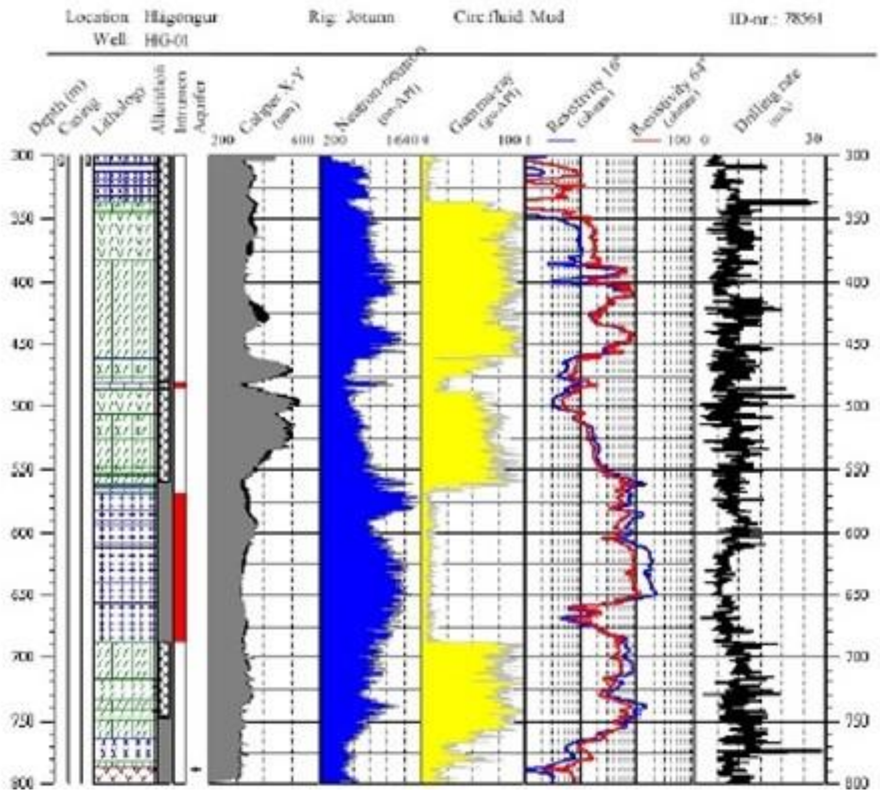


Figure 2 - Example of a geophysical well log [9]

2.2.1 Resistivity log

Resistivity logging is a well logging technique that involves measuring the electrical resistivity of the rock or sediment in a borehole to characterize it. Resistivity is a fundamental material attribute that describes how strongly a substance resists electric current flow. To reduce the resistance of the contact lines, resistivity is measured using four electrical probes in these logs [6].

2.2.2 Density porosity log

A density logging tool is used to calculate the electron density which is proportional to the formation's bulk density. A contact tool that emits gamma rays from a source is known as a density recording instrument. The gamma rays that are emitted collide with the formation electrons and disperse. The number of returning gamma rays is counted by a detector set at a specified distance from the tool source. The number of gamma rays that return is a measure of the formation's bulk density which is denoted as ρ_b or RHOB [6].

2.2.3 Neutron porosity log

The hydrogen index in a reservoir is closely connected to porosity, and neutron porosity measurement uses a neutron source to quantify it. It is denoted as Neutron Porosity Hydrogen Index (NPHI) [6].

2.2.4 Gamma ray log

Gamma ray logging is used to characterize the rock in a borehole or drill hole by monitoring naturally occurring gamma radiation emitted by them. It is measured in American Petroleum Institute units (API units) [10].

2.3 Critical Analysis and evidence of similarity to lithological classifications.

The main approach for conducting this study was inspired by lithological classifications of well log data into discrete rock facies such as sandstone, shale, limestone etc. Lithology is the study of the general physical characteristics of rock formation in a specific area [11]. Several studies have been conducted in classifying the well log data into the respective lithological facies using clustering techniques which are homogeneous to the method that will be applied in this study. These studies often inadvertently gave insight into the underlying accumulation of hydrocarbon in those regions of rock formation.

In well log interpretation sufficient evidence shows that the neutron-density crossover followed by a low gamma ray index, suggests the presence of hydrocarbons in that region of depth as discussed by Ijasan *et al.* [12].

Similarly, the research conducted by Ishwar *et al.* [8] in petrophysical well log analysis for hydrocarbon exploration suggests the presence of hydrocarbon in regions of low gamma ray, high resistivity and crossover between neutron porosity and bulk density as shown in the well log image below

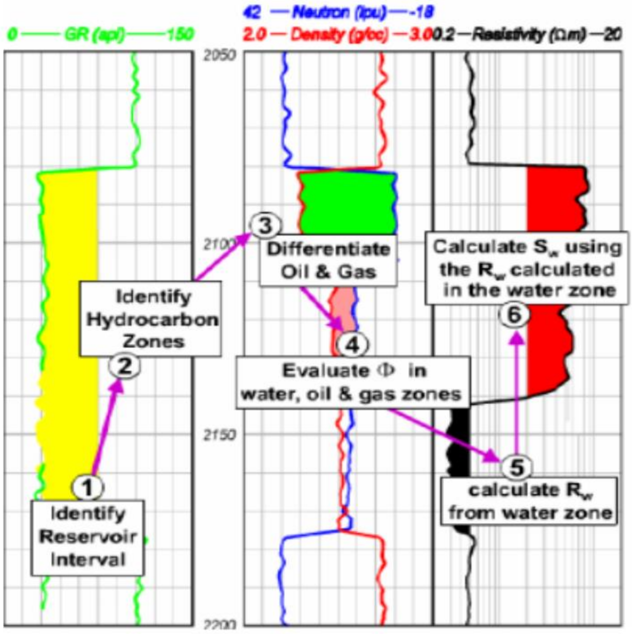


Figure 3 - Well log interpretation [13]

In the study conducted by Mahmoud *et al.* [14] it is suggested that lithological classification can be performed where the lithological contents of this rock unit were analysed using the cross plots of petrophysical parameters including shale volume, porosity, and hydrocarbon saturation. He also took advantage of the neutron-density cross-plots, and litho-saturation cross plots of the studied wells, using the well log data.

In the research conducted by Hossain *et al.* [15] lithology interpretation, they have performed K-means clustering and dimensionality reduction using Principal Component Analysis (PCA). The study was successful in producing bidimensional components from an initial dataset which was then used to perform the clustering analysis.

The hypothesis being introduced is that due to the nature of similarity in studies conducted between the lithological classification of hydrocarbon reservoirs and well log interpretation of hydrocarbon reservoirs, we can implement a novel approach and apply the same unsupervised learning techniques to identify useful clusters of data, in this case with neutron-density crossover, low gamma ray and high resistivity. Instead of classifying that data as facies of rock based on the well log data, they will be grouped into clusters of hydrocarbon bearing zones (Both oil and gas).

2.4 Dataset

The dataset that will be used for this project is a well log from the formations in the Norwegian North Sea spanning 86 wells. This dataset is publicly available and was compiled by EXPLOCROWD. This dataset has been used in the FORCE 2020 machine learning competition [16] to predict the lithofacies using machine learning models, however, in this study, it will be used to extract information about the underlying hydrocarbon accumulation in the region that these wells span.

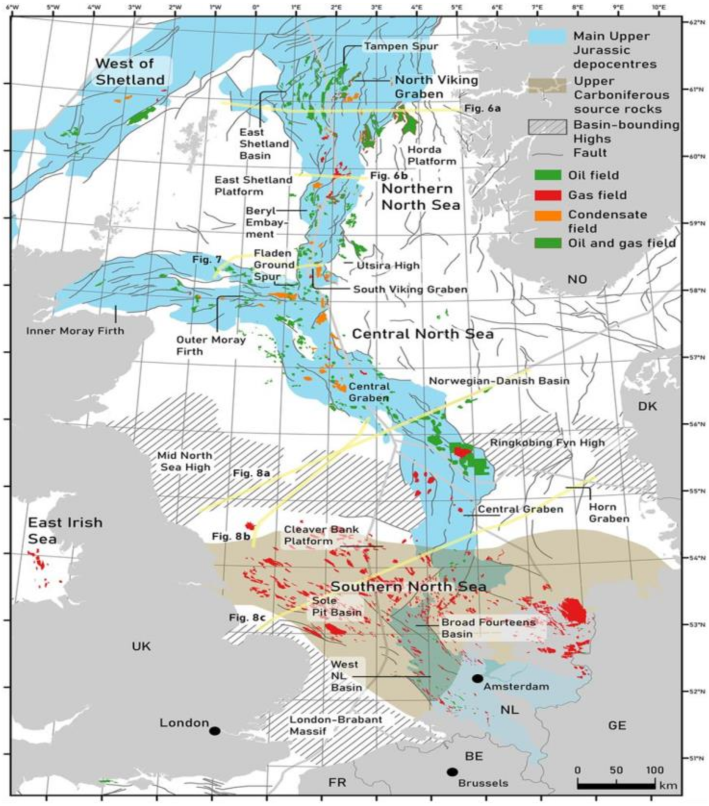


Figure 4 - Location map of the North Sea Basin [17]

These wells span the South and North Viking graben and penetrate a highly variable geology from the late Permian evaporites in the south and the deeply buried Brent delta facies in the North. The reason for the selection of this dataset is due to the rich formations of hydrocarbon reservoirs that have been accumulated over a long period ranging from the late Permian periods to the early cretaceous periods as shown in the figure below. (Green circle suggests the presence of oil, red valve suggests the presence of gas)

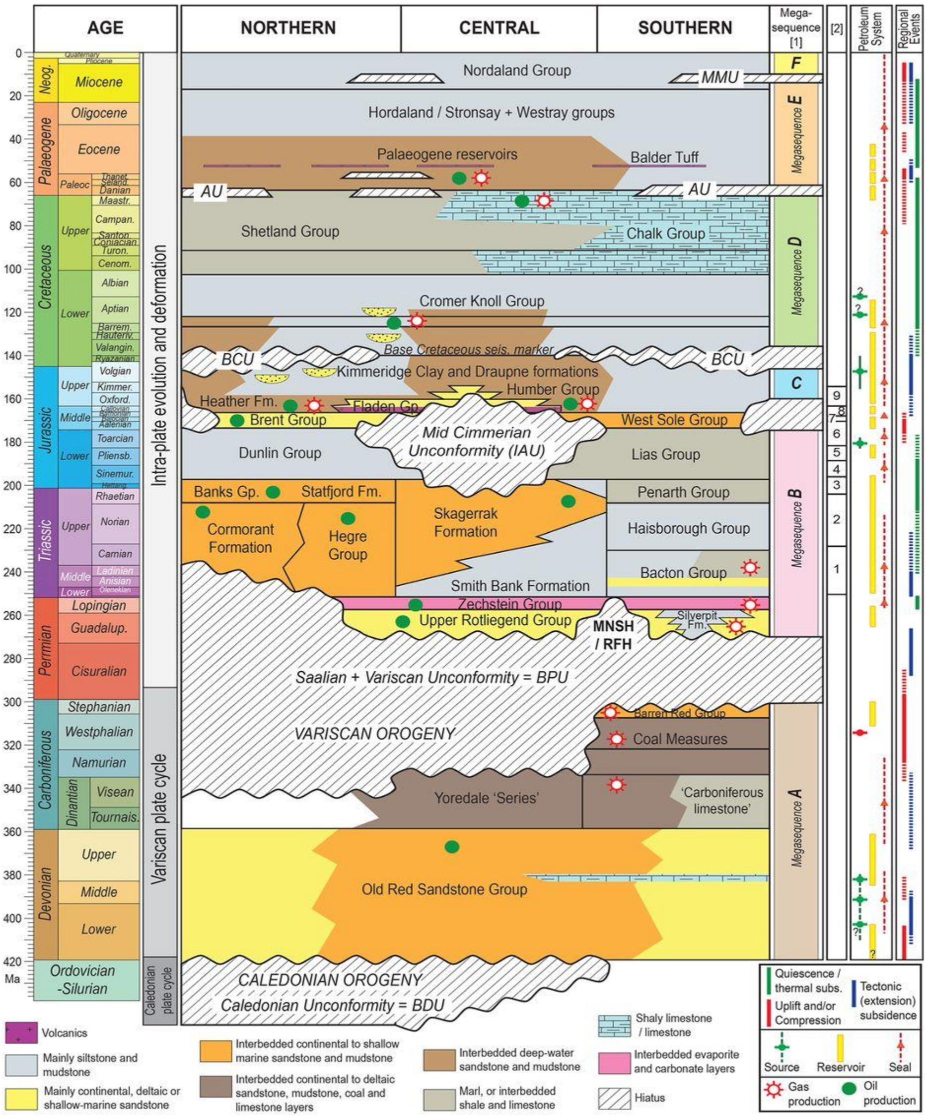
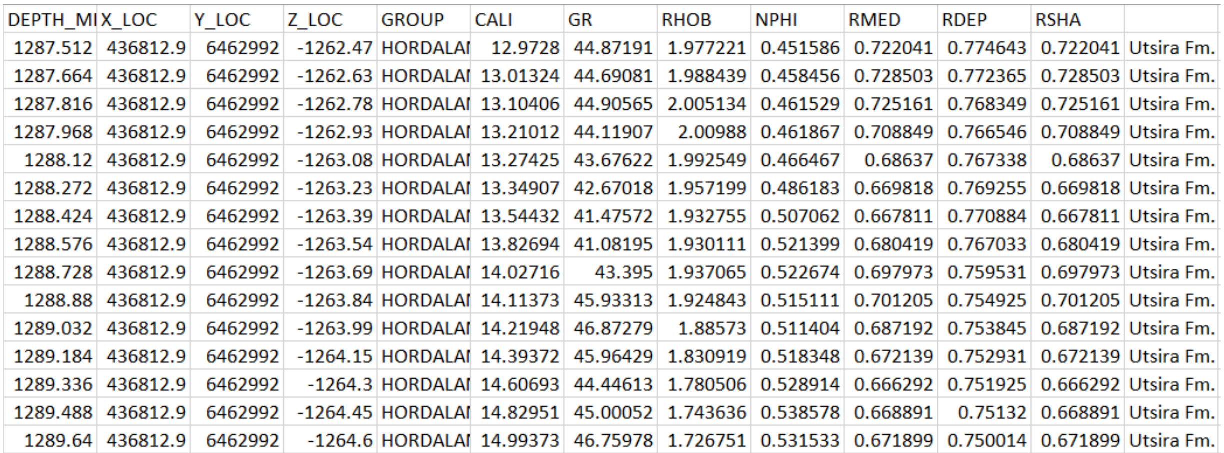


Figure 5 - Synthetic Stratigraphy for the North Sea Basin [17]

This dataset is extensive in size and contains 29 attributes in total, which is divided into 27 numerical attributes and 2 text attributes. The sample size is large and contains 1,048,574 samples. It contains the log readings of 86 wells over a range of groups and formations, each well contains readings taken at equal intervals of depth. Some notable features are depth, location, mud weight, density, porosity, gamma ray and resistivity.



DEPTH_MI	X_LOC	Y_LOC	Z_LOC	GROUP	CALI	GR	RHOB	NPHI	RMED	RDEP	RSHA	
1287.512	436812.9	6462992	-1262.47	HORDALAI	12.9728	44.87191	1.977221	0.451586	0.722041	0.774643	0.722041	Utsira Fm.
1287.664	436812.9	6462992	-1262.63	HORDALAI	13.01324	44.69081	1.988439	0.458456	0.728503	0.772365	0.728503	Utsira Fm.
1287.816	436812.9	6462992	-1262.78	HORDALAI	13.10406	44.90565	2.005134	0.461529	0.725161	0.768349	0.725161	Utsira Fm.
1287.968	436812.9	6462992	-1262.93	HORDALAI	13.21012	44.11907	2.00988	0.461867	0.708849	0.766546	0.708849	Utsira Fm.
1288.12	436812.9	6462992	-1263.08	HORDALAI	13.27425	43.67622	1.992549	0.466467	0.68637	0.767338	0.68637	Utsira Fm.
1288.272	436812.9	6462992	-1263.23	HORDALAI	13.34907	42.67018	1.957199	0.486183	0.669818	0.769255	0.669818	Utsira Fm.
1288.424	436812.9	6462992	-1263.39	HORDALAI	13.54432	41.47572	1.932755	0.507062	0.667811	0.770884	0.667811	Utsira Fm.
1288.576	436812.9	6462992	-1263.54	HORDALAI	13.82694	41.08195	1.930111	0.521399	0.680419	0.767033	0.680419	Utsira Fm.
1288.728	436812.9	6462992	-1263.69	HORDALAI	14.02716	43.395	1.937065	0.522674	0.697973	0.759531	0.697973	Utsira Fm.
1288.88	436812.9	6462992	-1263.84	HORDALAI	14.11373	45.93313	1.924843	0.515111	0.701205	0.754925	0.701205	Utsira Fm.
1289.032	436812.9	6462992	-1263.99	HORDALAI	14.21948	46.87279	1.88573	0.511404	0.687192	0.753845	0.687192	Utsira Fm.
1289.184	436812.9	6462992	-1264.15	HORDALAI	14.39372	45.96429	1.830919	0.518348	0.672139	0.752931	0.672139	Utsira Fm.
1289.336	436812.9	6462992	-1264.3	HORDALAI	14.60693	44.44613	1.780506	0.528914	0.666292	0.751925	0.666292	Utsira Fm.
1289.488	436812.9	6462992	-1264.45	HORDALAI	14.82951	45.00052	1.743636	0.538578	0.668891	0.75132	0.668891	Utsira Fm.
1289.64	436812.9	6462992	-1264.6	HORDALAI	14.99373	46.75978	1.726751	0.531533	0.671899	0.750014	0.671899	Utsira Fm.

Table 1 - Sample Dataset

2.5 Clustering Analysis

Clustering is an unsupervised machine learning technique that helps us to identify patterns or similar objects of interest within a specified multidimensional dataset. There are no prior input-output relationships i.e., the data is untagged or unlabelled unlike in supervised machine learning [18]. By using clustering methods, we can segregate our data into insightful homogeneous groups, in this case, we aim to investigate similarities in data that follow a specific pattern which is explained in the critical analysis section.

2.5.1 SOM

In exploratory study of high-dimensional data, SOM or Self-Organizing Maps can be used to highlight relationships between data components.

SOM can be used to find groupings of reservoir and fluid variables in the data that have similar trends [19]. In a reduced space 2D lattice, an unsupervised learning method groups qualities with similar relationships. The input data is assigned to neurons on a lattice based on their similarity. Heat maps can be used to visualize the distances between data points on a 2D lattice. We can produce a heat map for each attribute using SOM lattice data projection, which illustrates mutual links between the patterns in the data, allowing geological attributes to be classified depending on the trends they exhibit [20].

2.5.2 K-Means

K-Means is a clustering algorithm based on the similarity of data that is in proximity in multidimensional space [20]. K-means randomly chooses a centroid (cluster centre) for each cluster from the dataset at the first iteration. Then, it places every point in the cluster with the closest centroid until no point is left un-clustered. In the next iteration, a new centroid is calculated for each cluster, which represents the mean of the points in this cluster. All the points are then assigned to the cluster with the closest new centroid. This process continues until no point changes its cluster which means the algorithm has converged [21]. K-means clustering requires us to find the optimal number of clusters as the name suggests.

2.5.2.1 The Elbow Method

The Elbow method involves plotting the WCSS values of the dataset against the number of clusters in range and provides a point which resembles the elbow of a human arm. This point gives information on the best potential number of clusters ‘k’. The squared distance between each cluster member and its centroid is added together to obtain the WCSS values [22]. An example of the elbow method is shown in figure 6.

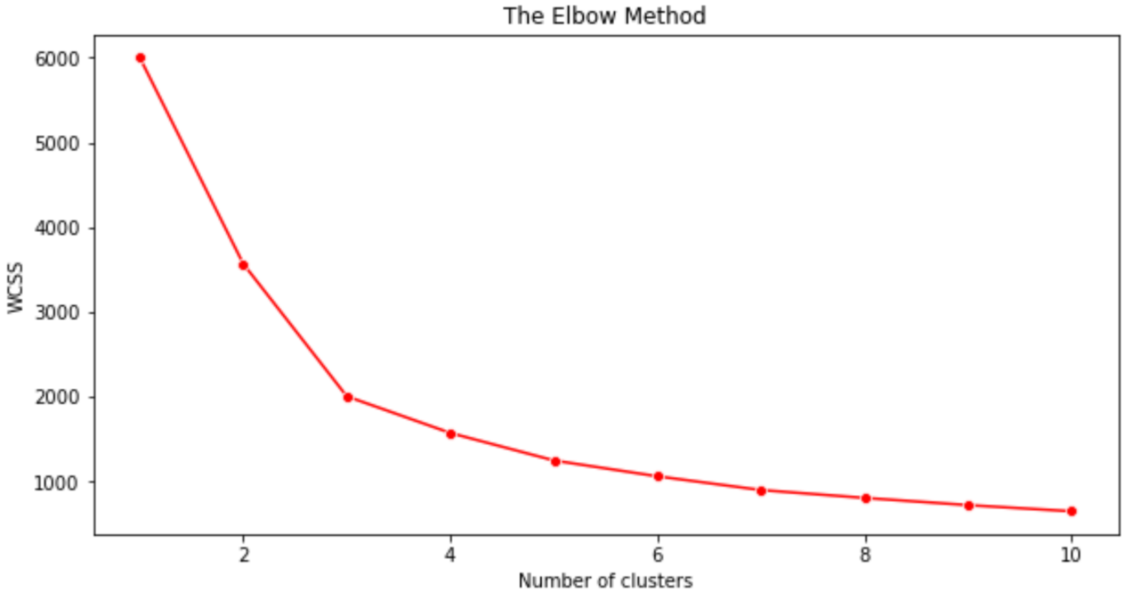


Figure 6 - Example of the elbow method

In this example the optimal number of clusters would be 3.

2.5.3 DBSCAN

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a clustering algorithm that is based on two parameters; the first one is minPts which is the minimum number of neighbours to form a cluster and the second one is eps, which is the maximum distance between two points in a neighbourhood. Based on these two parameters, DBSCAN classifies points into three types; Core points which have minPts neighbours with distance less than or equal to eps. Border points are points that lie within an eps radius from a core point but have less than minPts neighbours in their eps radius and outliers which are the remainder or not clustered [23].

2.6 Dimensionality Reduction Techniques

Clustering algorithms struggle in high dimensional data due to the *Curse of Dimensionality,* which is the tendency for machine learning algorithms to work well in low dimensions but break down as the dimension increases [24]. For clustering algorithms that rely on distance metrics, the points of data seem to be equidistant as the number of dimensions increases [25]. A method that can be used to circumvent this curse of dimension is dimensionality reduction. Visualization holds tremendous importance in this study and is important that high dimensional data is viewed appropriately in a 2D space. As explained in the critical analysis section, clusters need to be identified in a specific pattern. Identifying this pattern will be disrupted if the attributes contributing to this pattern in data are not present during dimensionality reduction. Therefore, the selection of dimensionality reduction technique needs to account for the preservation of the local structure.

2.6.1 PCA

Principal Component Analysis (PCA) is a dimensionality reduction technique that analyses the different observations and extracts the important information from the dataset to illustrate the pattern of similarity between the observations and the variables as points on maps, and to express it as a collection of new orthogonal variables called principal components [26]. For example, if we have a large multivariate dataset with n dimensions and we want to reduce it to m dimensions where m < n, the idea is that it is possible to implement PCA to reduce the dimensions from n to m while retaining the variance in the original dataset up to a considerable extent [27].

2.6.2 t-SNE

To solve the issue of clustering more than 3 dimensions t- Stochastic neighbour embedding is very effective It is impossible to view for example a 4D scatter plot on a 2D screen. Therefore t-SNE can be used to view the same information on a 2D plane. The t-SNE algorithm simulates the probability distribution of each point's neighbours. The term "neighbours" refers to the group of points closest to each other. This is described as a Gaussian distribution in the original, high-dimensional space. This is modelled as a t-distribution in the 2-dimensional output space. The capacity to preserve local structure, which is critical in locating valuable clusters of data from the well log, is t-SNE's advantage. This indicates that in a 2D scatter plot, points that are near to one another in a high-dimensional data set will tend to be close to one another [28].

3. Methodology

3.1. Overview

Two different unsupervised learning techniques are used, namely K-MEANS [29] and SOM [30]. These methods have proved to be effective when implemented on lithological datasets as explained in the critical analysis section. The four main attributes that were used for the clustering analysis are Gamma Ray, Bulk Density, Neutron Porosity and Resistivity as mentioned in the critical analysis section. Another attribute that holds importance is the depth of the reservoir, as it is crucial to have the data points ordered by depth to plot the well logs. Principal Component Analysis (PCA) was used to reduce the 4-dimensional data into their respective 2-dimensional components. Therefore, there is a total of 2 clustering methods and 1 dimensionality reducing technique that will be implemented.

3.2. Pre-processing of Dataset

For the pre-processing of the well log dataset, the following steps will be implemented: -

- Removing columns that do not contribute to the clustering analysis in any way. Since this dataset was primarily used for lithological classification, the non-essential attributes can be removed keeping only those that are crucial for identifying hydrocarbon-bearing zones. All textual attributes can also be dropped.

- Removing samples which contain missing values, this is crucial to perform dimensionality reduction. Dimensionality reductions cannot be performed with the existence of NA values.
- Splitting the dataset for each individual well. For instance, if there are 100 unique wells present in the dataset, that would mean splitting the original well log dataset into 100 sub datasets.
- It is imperative that all the individual datasets are ordered by depth, here depth can be considered an interval data, where each sample is differentiated by periodic depth levels. It is very essential to do this, and the samples cannot be randomized/shuffled due to the reason that well logs can be comprehended only when they are ordered by depth.

.

3.3. Manual well log interpretation

As the dataset is untagged, measuring the performance metrics and validation of clustering methods will be insufficient in determining the efficiency of the model. Hence it is necessary to interpret the well log manually and identify the potential hydrocarbon-bearing zones for each well, before performing the clustering analysis. Another reason to consider manual well log interpretation is to help identify and remove wells that have no potential for hydrocarbons at all.

3.4. Software and Hardware Specifications

For this study, Jupyter Notebook was used to run the simulations. Primarily the python libraries such as pandas dataframe, matplotlib, seaborn, numpy and scikitlearn were used. In addition to these the minisom library [31] was used to implement a minimalistic version of the SOM.

The hardware specifications are as follows: -

- The Central Processing Unit (CPU) is Intel(R) Core (TM) i7-7500U CPU @ 2.70GHz -2.90 GHz.
- 8GB Memory.
- 2GB Nvidia Geforce 940mx Graphics Processing Unit (GPU).

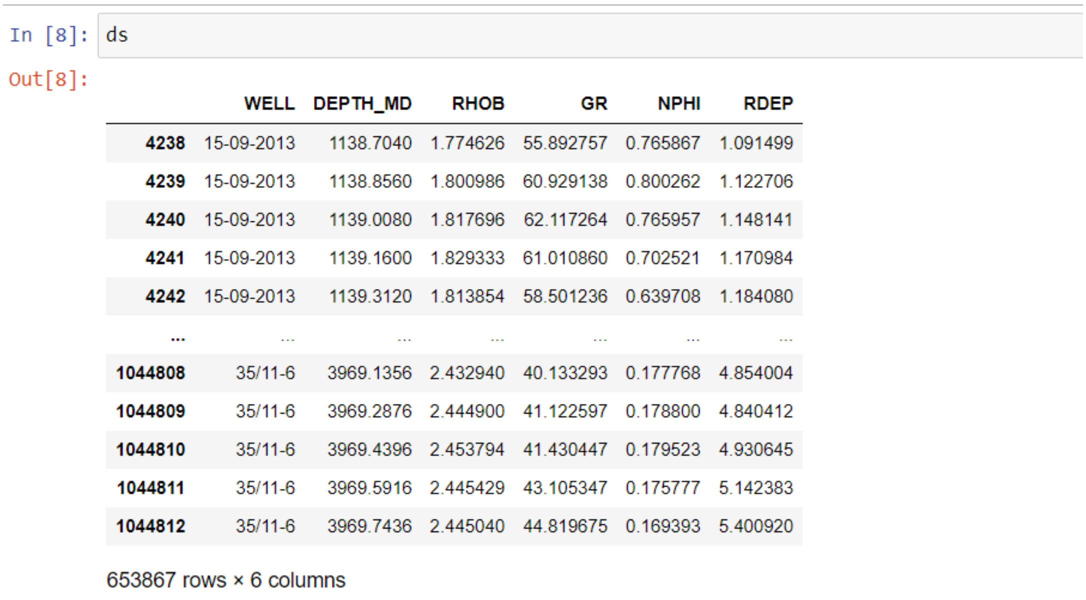
3.5. Evaluation and validation of results

The results obtained from manual well log interpretation and through clustering analysis were compared and tabulated. The goal is to be able to evaluate how effective the clustering methods are in identifying hydrocarbon bearing regions that were previously identified from the manual interpretation of the well log dataset.

4. Implementation

4.1. Pre-Processing

The original dataset consists of 27 columns and a sample size of 1,048,574 which includes 86 unique wells. These are the dimensions of the dataset before rows containing incomplete values were removed/dropped and the essential attributes were selected. There was no feature selection performed in this scenario as the goal of this study is to use four attributes GR (Gamma Ray), RDEP (Resistivity), RHOB (Bulk Density) and NPHI (Neutron Porosity) to identify the hydrocarbon bearing regions. In addition to these 4 attributes, we require the depth values to visualize the well log plots and the well name to separate the well log dataset into unique Pandas data frames. Therefore, all the other columns can be considered non-essential to this research study and are hence removed along with rows containing null values. This results in a dataset of 6 attributes and a sample size of 653,866 which is approximately 38% less than the original dataset. This new dataset contains 85 unique wells.



```
In [8]: ds
Out[8]:
               WELL   DEPTH_MD      RHOB         GR      NPHI      RDEP
   4238  15-09-2013  1138.7040  1.774626  55.892757  0.765867  1.091499
   4239  15-09-2013  1138.8560  1.800986  60.929138  0.800262  1.122706
   4240  15-09-2013  1139.0080  1.817696  62.117264  0.765957  1.148141
   4241  15-09-2013  1139.1600  1.829333  61.010860  0.702521  1.170984
   4242  15-09-2013  1139.3120  1.813854  58.501236  0.639708  1.184080
    ...         ...        ...       ...        ...       ...       ...
1044808     35/11-6  3969.1356  2.432940  40.133293  0.177768  4.854004
1044809     35/11-6  3969.2876  2.444900  41.122597  0.178800  4.840412
1044810     35/11-6  3969.4396  2.453794  41.430447  0.179523  4.930645
1044811     35/11-6  3969.5916  2.445429  43.105347  0.175777  5.142383
1044812     35/11-6  3969.7436  2.445040  44.819675  0.169393  5.400920

653867 rows × 6 columns
```

Figure 7 - Initial Dataset

Although each well is assigned a separate Pandas data frame, the original dataset is still retained as this will undergo dimensionality reduction and then be clustered. There is no scaling, normalizing or transformations performed on this dataset to preserve the originality of the data. It would be impractical to do so because the resulting datapoints cannot be plotted as a well log. Using the pairplot() function from seaborn library the 4 main attributes from the initial dataset can be viewed as shown in figure 8.

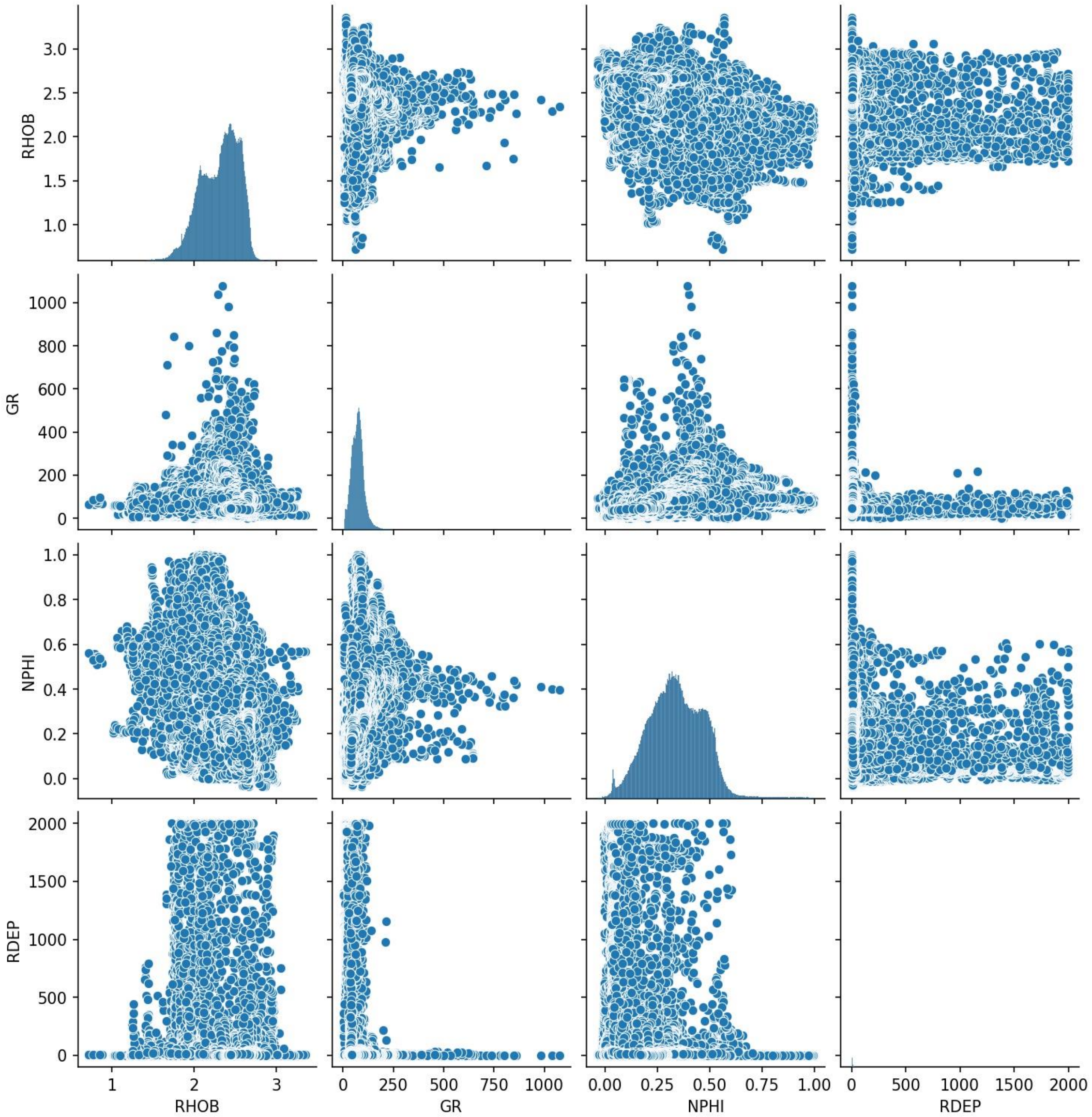
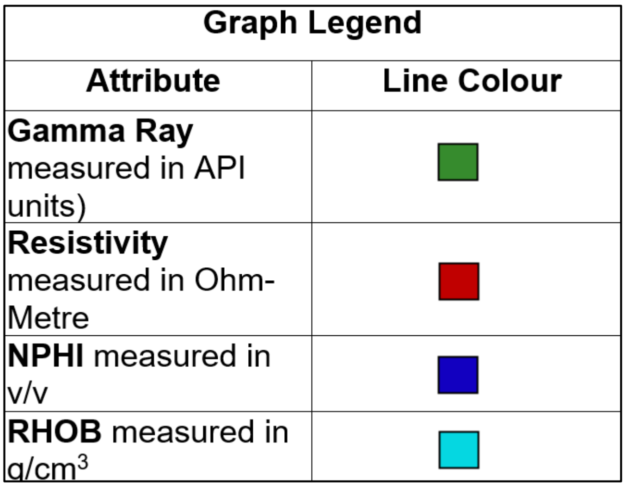


Figure 8 - Initial Data Exploration

4.2. Interpretation of the Well Logs

To get the best results from this experiment certain criterions were placed for identifying the hydrocarbon bearing regions to ensure the most accurate evaluation in the later stages.

- As mentioned in section 2.3, a hydrocarbon bearing (both oil and gas) is identified by a low Gamma ray, high Resistivity and crossover between Neutron porosity (NPHI) and Bulk density (RHOB).
- An example of a well log plot with clear indication of hydrocarbon bearing regions is shown in figure 11.
- An example of a well log plot which was discarded is shown in figure 10.
- The legend for the graph is given in figure 9.



Graph Legend

Attribute	Line Colour
Gamma Ray measured in API units)	🟩
Resistivity measured in Ohm-Metre	🟥
NPHI measured in v/v	🟦
RHOB measured in g/cm^3	🟦

Figure 9 - Graph Legend 1

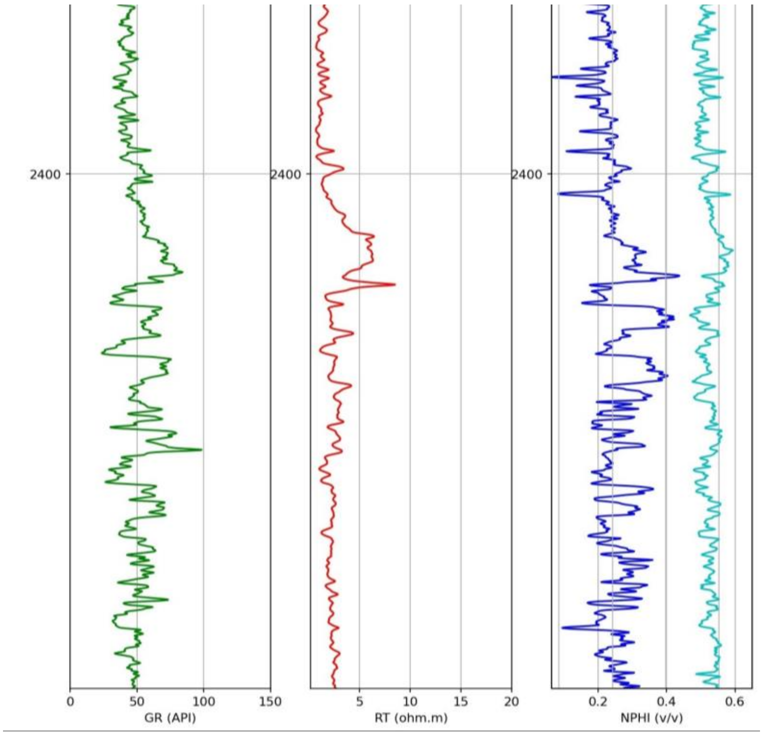


Figure 10 – Well log plot with no indication of the presence of hydrocarbons

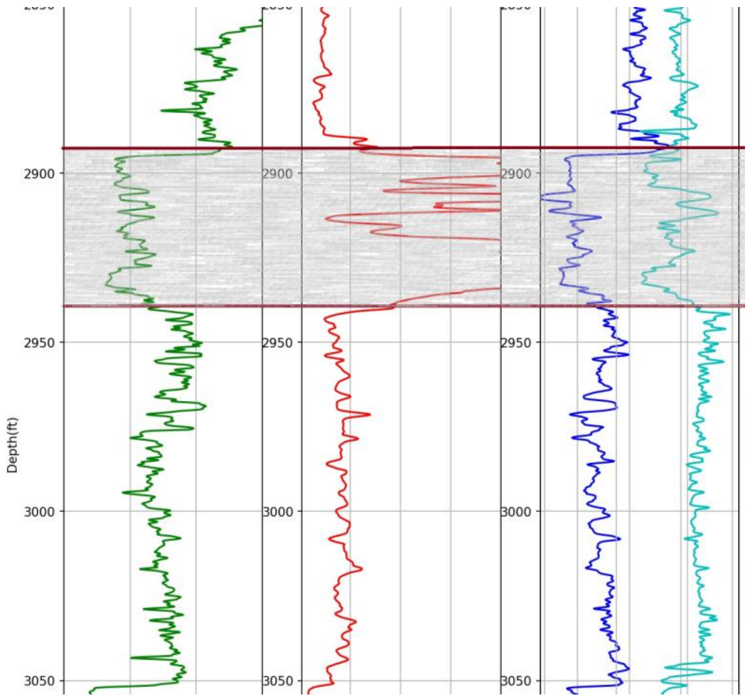
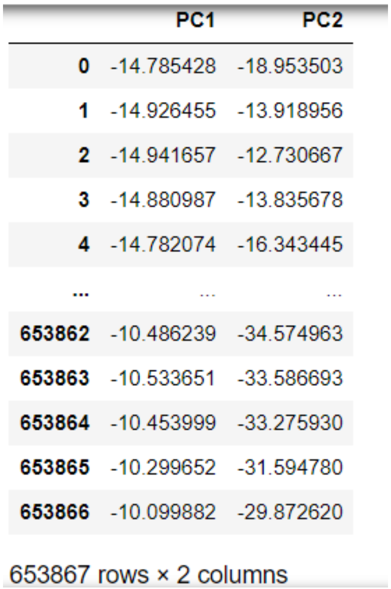


Figure 11 - Well log plot with clear indication of the presence of hydrocarbons

After manually interpreting **85** different well logs, it was found that **51** of them either did not have visible hydrocarbon bearing regions or did not contain any hydrocarbon bearing regions at all. This leaves **34** datasets with visible hydrocarbon zones. These zones were marked and will be used later after the clustering analysis is performed.

4.3. Dimensionality Reduction

The Principal Component Analysis (PCA) was performed on the dataset to reduce the dimensions of each datapoint's four main attributes namely GR, RDEP, NPHI and RHOB to their reduced two-dimensional components namely PC1 and PC2 as shown in figure 12.



	PC1	PC2
0	-14.785428	-18.953503
1	-14.926455	-13.918956
2	-14.941657	-12.730667
3	-14.880987	-13.835678
4	-14.782074	-16.343445
...	...	...
653862	-10.486239	-34.574963
653863	-10.533651	-33.586693
653864	-10.453999	-33.275930
653865	-10.299652	-31.594780
653866	-10.099882	-29.872620

653867 rows × 2 columns

Figure 12 - Principal Components of Dataset

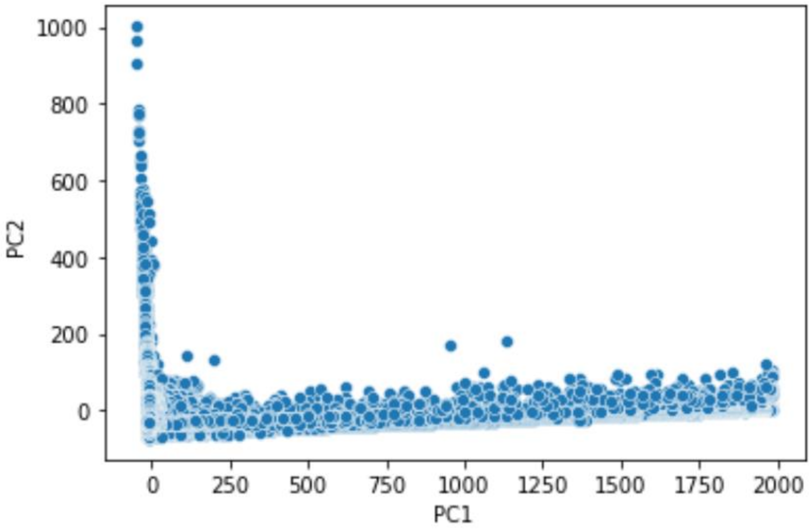


Figure 13 - Scatter plot of Principal Components

It was important that the dimensionality reduction and clustering analysis were performed on the entire initial dataset to avoid overfitting and obtain more generalized clusters of data [32]. This would ensure the efficiency of the clustering on new unseen data in a more practical scenario. The scatter plot of the newly obtained 2D points is shown in figure 13. The index of the original dataset was reset along with the indexes of the Principal Components dataset and were concatenated, therefore the 4D data points were linked with their respective 2D points.

4.4. Implementation of K-Means

To perform the K-Means clustering algorithm an appropriate number of clusters to be formed had to be chosen. To fulfil this requirement the elbow method was used. The elbow method was performed on the dataset as shown in figure 14 containing the Principal Components for a range of 0-20 clusters and 5 clusters was found to be the potential best.

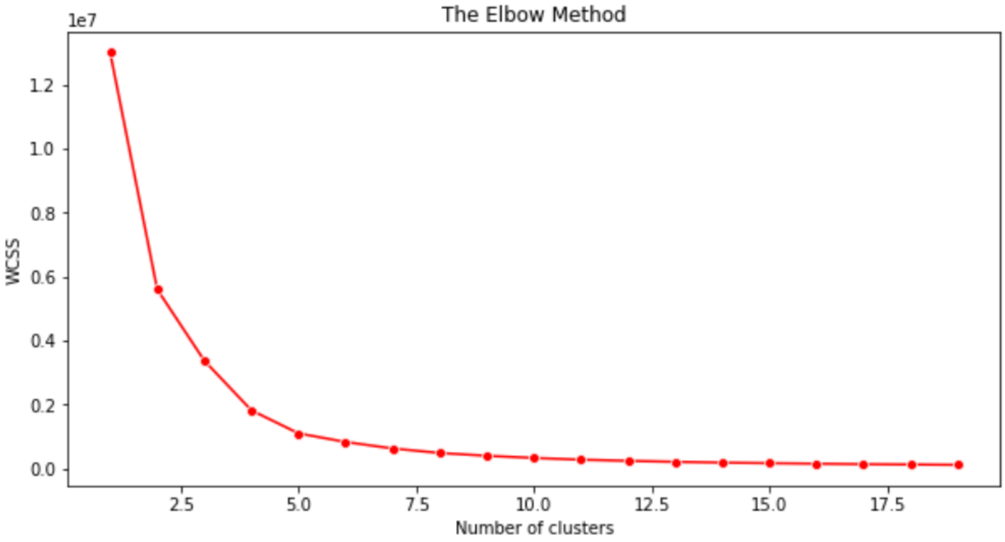


Figure 14 - Elbow method for K-Means Clustering

The K-Means clustering analysis was then performed on the dataset containing the principal components. The scatter plot is shown in figure 15.

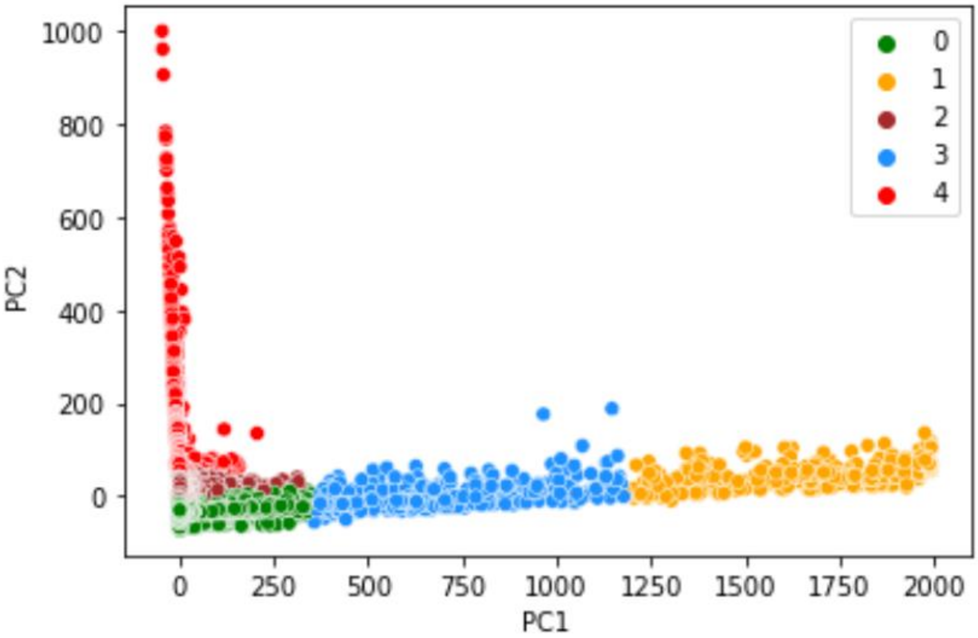


Figure 15 - Visualizing K-Means clusters formed on the bidimensional dataset

The cluster labels obtained in the form of a 1-D array were then added to the original dataset as a new column.

4.5. Implementation of SOM

The main idea behind Self Organizing Maps is that all the samples that are mapped into a specific neuron will belong to that specific cluster [33]. The SOM had been implemented using the minisom library [31]. The minisom package provided a minimalistic implementation of the SOM. The shape of the SOM was adjusted in such a way the resulting number of clusters is the same as K-Means i.e., 5 clusters. This was done so that the two different unsupervised learning techniques can be compared effectively. Due to the large size of the dataset, the SOM model was trained over 50,000 iterations with a Gaussian neighbourhood function.

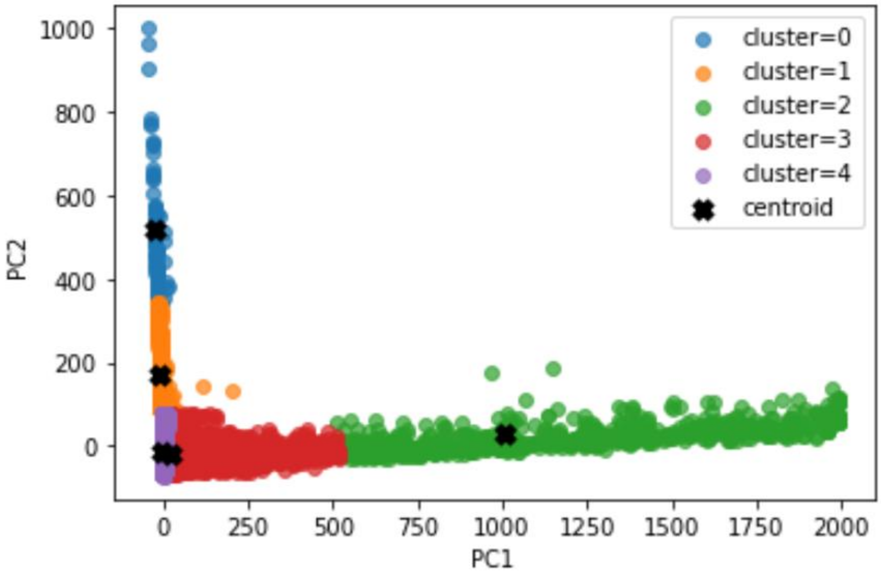
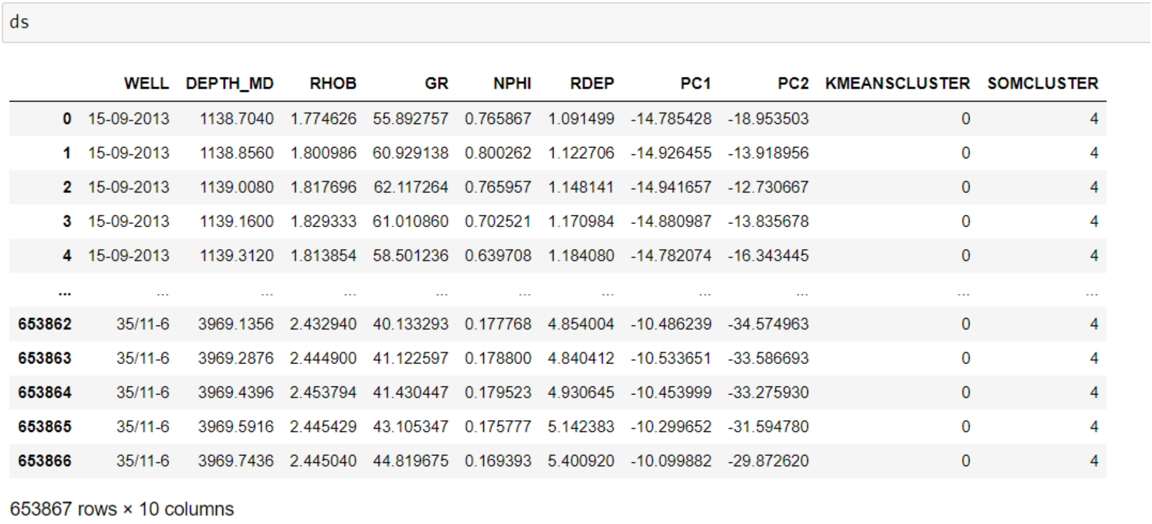


Figure 16 - Visualizing SOM Clusters formed on the bidimensional dataset

To identify each cluster accurately the bidimensional coordinates of the neuron in the SOM were transformed into a mono dimensional index. Like K-Means, these labels of clusters were then added to the original dataset as a new column.

The resulting dataset contains the original 4-D datapoints, their bidimensional coordinates and the clusters they belong to respectively as shown in figure 17. This dataset was not shuffled/rearranged at any point of the implementation phase; therefore, it contains the well wise datapoints in increasing order of depth. This will be used to plot the well wise log graphs and to visualize which cluster each data point falls into. The aim is to find datapoints belonging to a hydrocarbon region assigned to a specific cluster for K-Means and SOM. The hydrocarbon bearing regions that are found from the resulting 34 wells after manual well log interpretation is visualized along with a discrete line graph formed by the K-Means clusters and SOM clusters.



```
ds
```

	WELL	DEPTH_MD	RHOB	GR	NPHI	RDEP	PC1	PC2	KMEANSCLUSTER	SOMCLUSTER
0	15-09-2013	1138.7040	1.774626	55.892757	0.765867	1.091499	-14.785428	-18.953503	0	4
1	15-09-2013	1138.8560	1.800986	60.929138	0.800262	1.122706	-14.926455	-13.918956	0	4
2	15-09-2013	1139.0080	1.817696	62.117264	0.765957	1.148141	-14.941657	-12.730667	0	4
3	15-09-2013	1139.1600	1.829333	61.010860	0.702521	1.170984	-14.880987	-13.835678	0	4
4	15-09-2013	1139.3120	1.813854	58.501236	0.639708	1.184080	-14.782074	-16.343445	0	4
...	...	...	...	...	...	...	...	...	...	...
653862	35/11-6	3969.1356	2.432940	40.133293	0.177768	4.854004	-10.486239	-34.574963	0	4
653863	35/11-6	3969.2876	2.444900	41.122597	0.178800	4.840412	-10.533651	-33.586693	0	4
653864	35/11-6	3969.4396	2.453794	41.430447	0.179523	4.930645	-10.453999	-33.275930	0	4
653865	35/11-6	3969.5916	2.445429	43.105347	0.175777	5.142383	-10.299652	-31.594780	0	4
653866	35/11-6	3969.7436	2.445040	44.819675	0.169393	5.400920	-10.099882	-29.872620	0	4

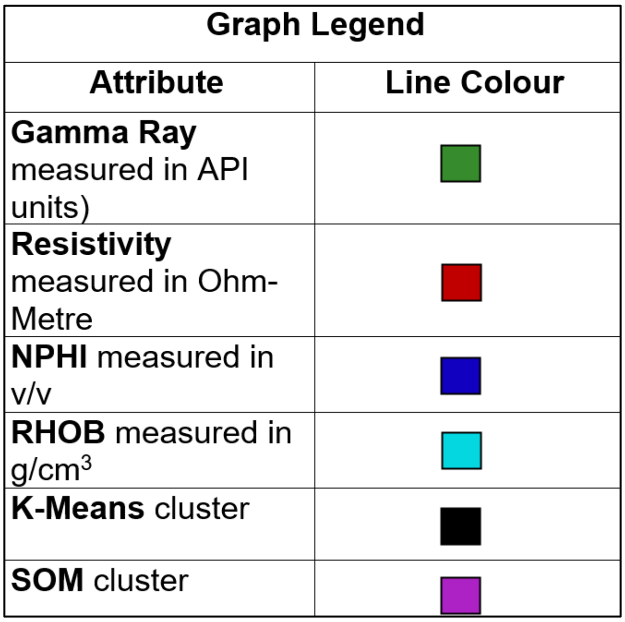
653867 rows × 10 columns

Figure 17 - Final Obtained Dataset

5. Results and Inferences

For the purpose of this experiment and visualization, 3 best wells out of 34 are selected and displayed below. The regions of hydrocarbon are highlighted.

The legend for the following wells is as given below: -



Graph Legend	
Attribute	**Line Colour**
Gamma Ray measured in API units)	🟩
Resistivity measured in Ohm-Metre	🟥
NPHI measured in v/v	🟦
RHOB measured in g/cm^3	🟦
K-Means cluster	⬛
SOM cluster	🟪

Figure 18 - Graph Legend 2

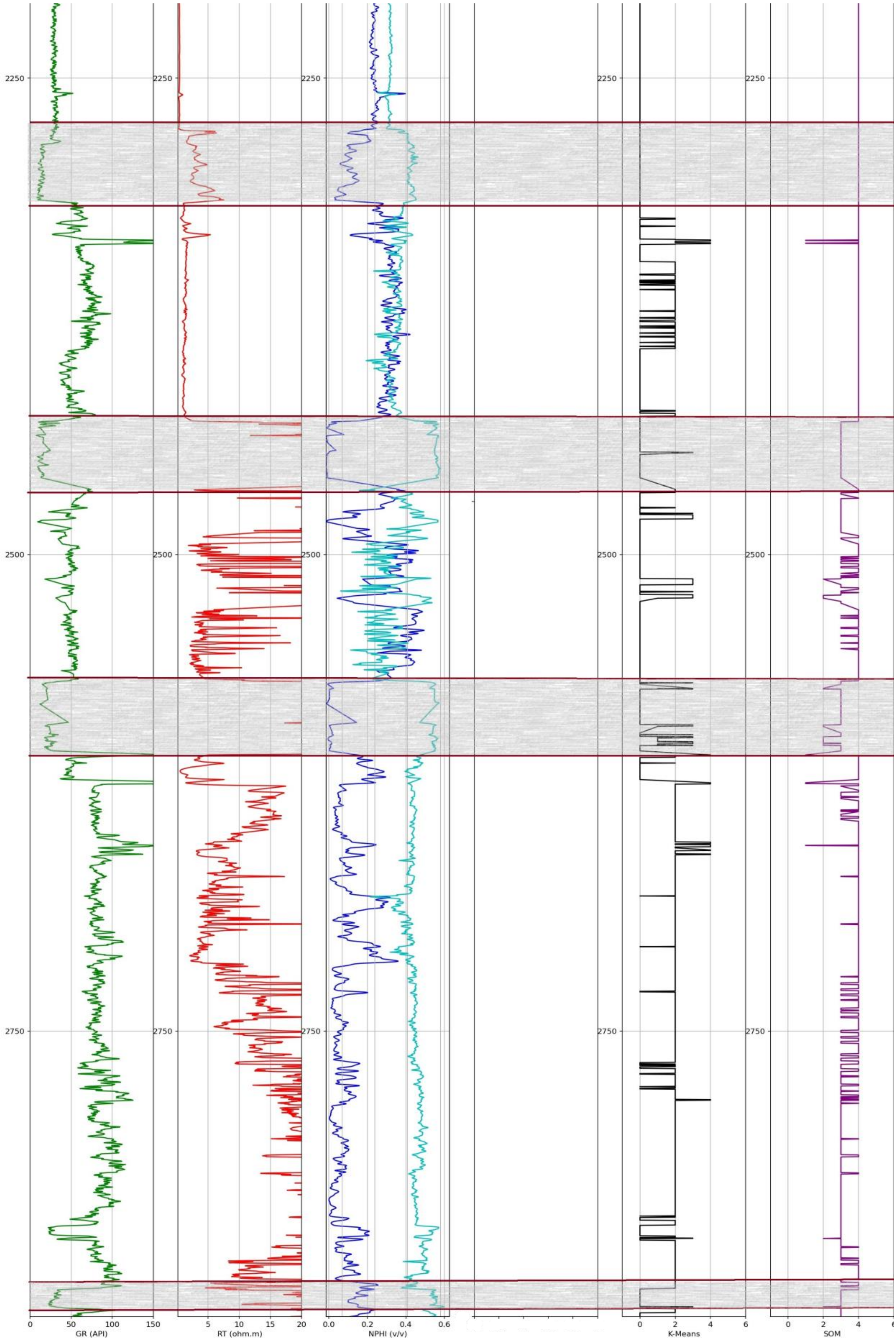


Figure 19 - Well '16-04-2001' Log Plot

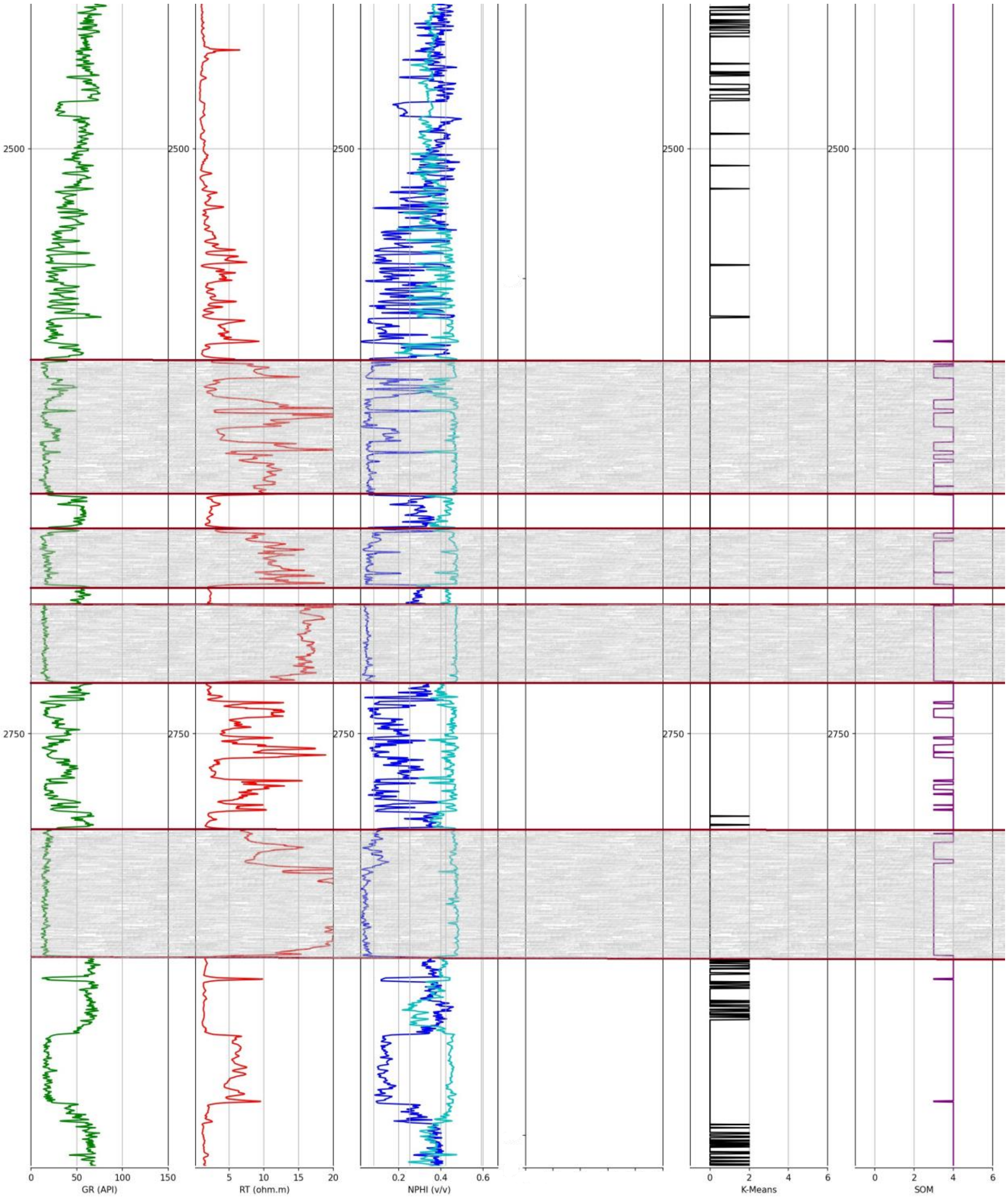


Figure 20 – Well '25-02-2007' Log Plot

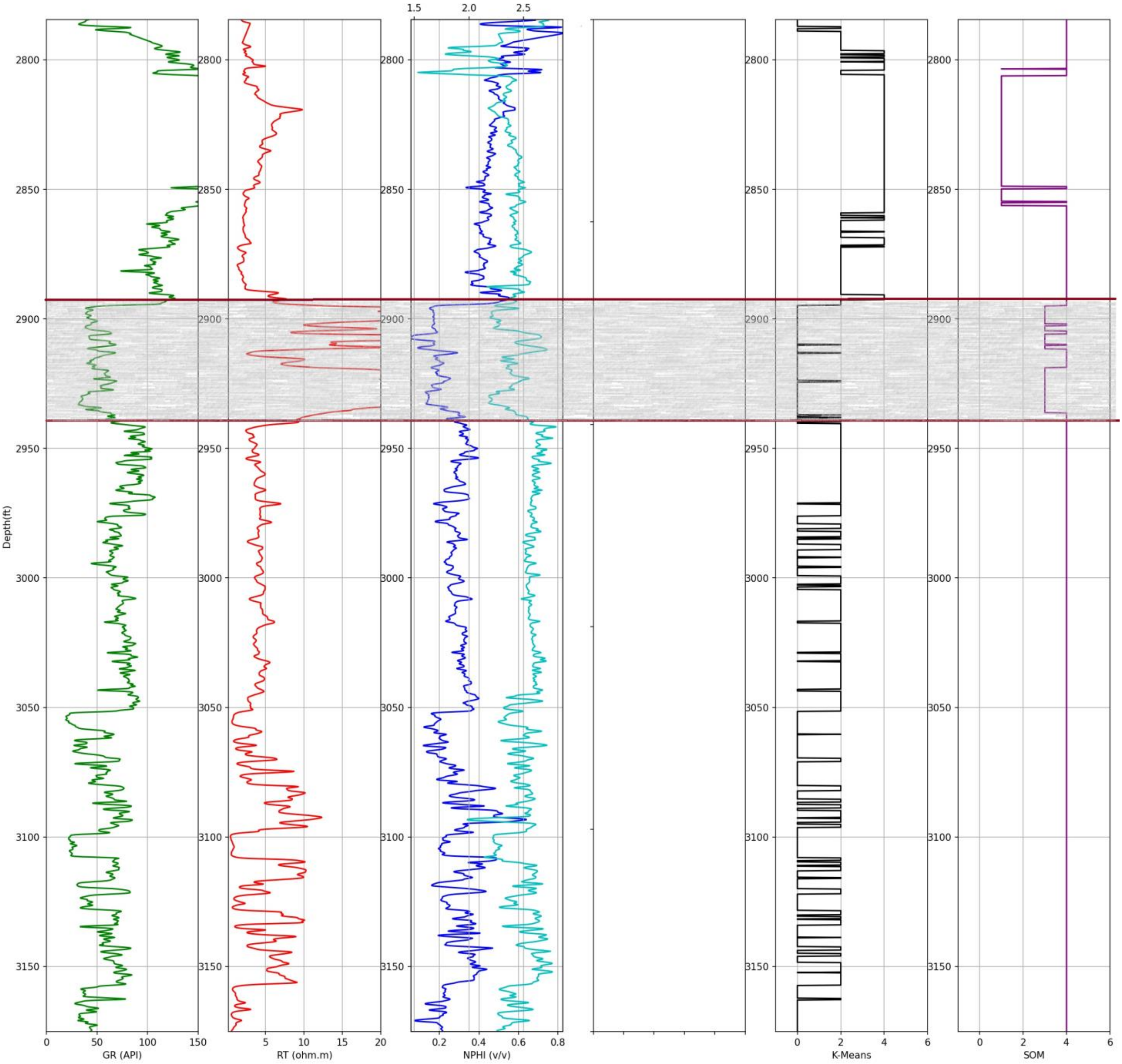


Figure 21 - Well '25-05-2004' Log Plot

In figure 19,20 and 21, it is perceived that the K-Means clusters the data points belonging to a hydrocarbon region into cluster 0 and the SOM clusters the same data points into cluster 3. The total number of identified hydrocarbon regions, the number of clusters formed by K-Means and SOM were noted and tabulated in table 2.

S.No.	Well Name	No. of hydrocarbon regions identified from manual interpretation of well log	No. of clusters identified by KMEANS	No. of clusters identified by SOM
1.	Ds1	6	1	6
2.	Ds4	4	3	1
3.	Ds6	3	3	3
4.	Ds9	2	1	0
5.	Ds11	4	0	3
6.	Ds12	4	4	2
7.	Ds13	4	2	4
8.	Ds14	4	1	3
9.	Ds20	2	2	0
10.	Ds22	1	0	0
11.	Ds25	5	2	4
12.	Ds26	3	3	3
13.	Ds28	2	2	2
14.	Ds29	1	1	1
15.	Ds31	2	1	0
16.	Ds32	1	1	1
17.	Ds38	2	2	2
18.	Ds40	2	1	1
19.	Ds41	1	1	1
20.	Ds50	4	4	0
21.	Ds55	1	1	1
22.	Ds57	3	3	3
23.	Ds62	3	1	1
24.	Ds64	2	2	2
25.	Ds65	1	1	1
26.	Ds66	4	4	4
27.	Ds67	1	1	1
28.	Ds75	1	1	1
29.	Ds77	1	1	1
30.	Ds78	3	3	3
31.	Ds81	1	0	1
32.	Ds83	1	0	1
33.	Ds84	1	0	1
34.	Ds85	1	1	1
	Total:	81	54	59

Table 2 - Tabulation of results from manual well log interpretation and clustering analysis

The K-Means clustering algorithm was able to cluster **54** out of **81** hydrocarbon regions while the SOM algorithm was able to successfully cluster **59** out of **81** points. Although the SOM algorithm seems to have only a slight edge over K-Means it is far more superior as the K-Means algorithm seemed to be heavily influenced by the Gamma Ray values.

5.1. Influence of Gamma Ray on K-Means Clustering

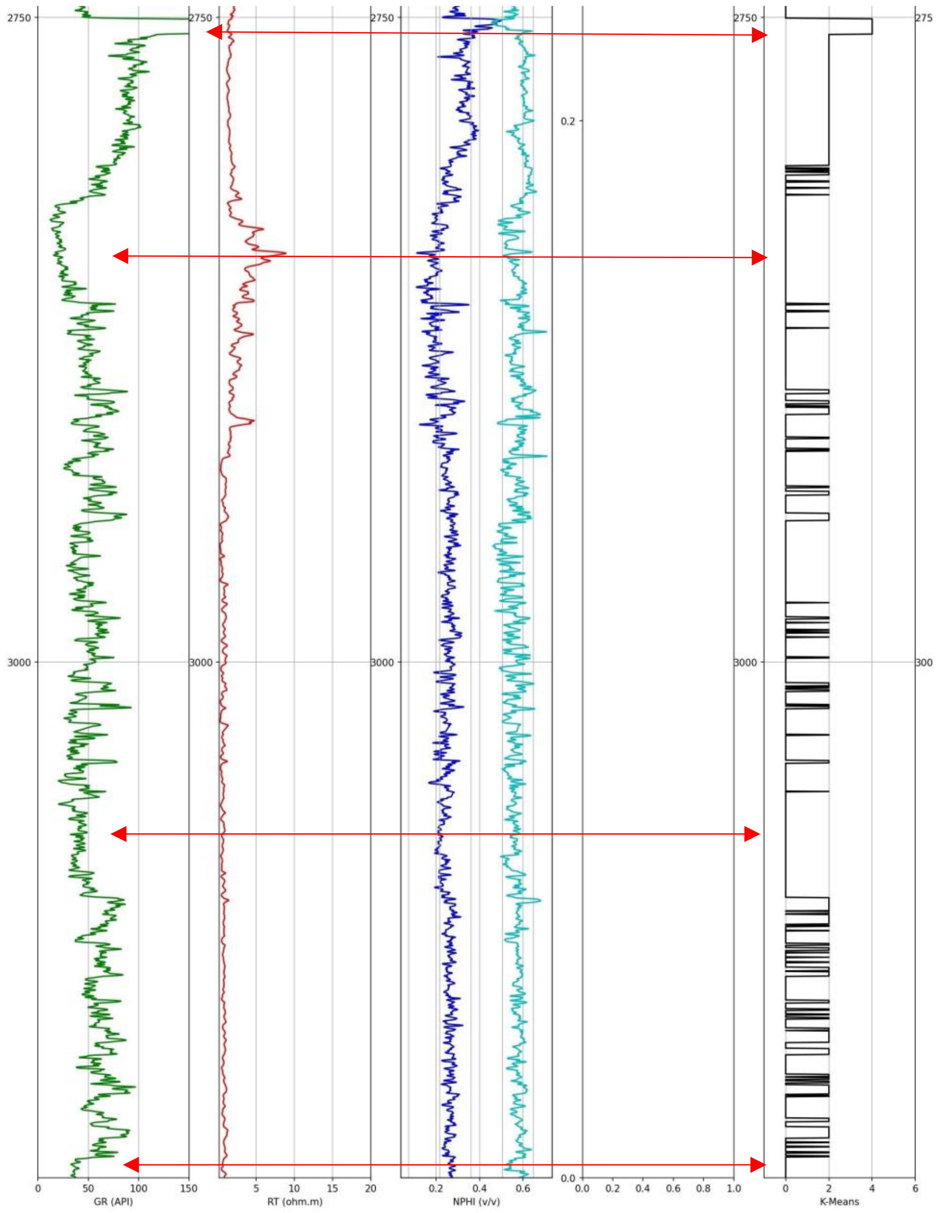


Figure 22 - Influence of Gamma Ray on K-Means clustering

In figure 12 the Gamma ray plot (in green) and the K-Means discrete line graph (in black) are almost identical. The K-Means is essentially a discretised plot of the continuous Gamma ray values. It was inferred that the K-means algorithm set its centroids based on the value of gamma ray of the data points. Lower values of gamma ray tended to fall into cluster 0 and 2, whereas higher values tended to fall into cluster 4. It was also found that due to this influence the K-Means triggered a lot of false positives. A low gamma ray value does not necessarily mean that the region contains hydrocarbon as the clusters need to also account for high resistivity and RHOB-NPHI crossover. Therefore, SOM was better at clustering these data points, as it accounted for all the patterns identified in the dataset.

In conclusion the combined indication of K-Means and SOM is a good method to identify the hydrocarbon regions with a low false positive rate.

5.2. Domination of PCA components by high variance attributes

In the initial exploratory analysis of the original dataset a pairwise plot of the 4 main attributes were examined. The resistivity vs gamma ray plot of the original dataset was compared with the principal components scatter plot and they were found to be identical as shown in Figure 23 and 24.

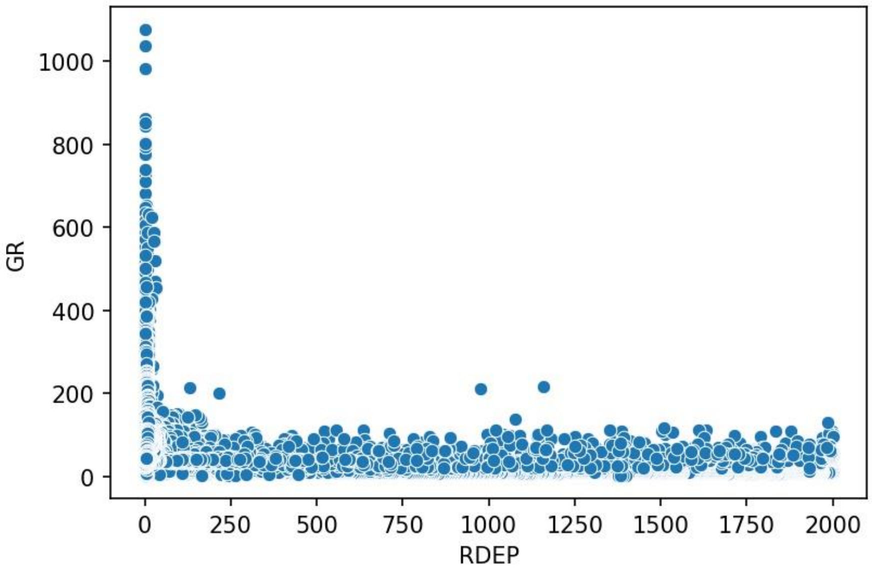


Figure 23 – Resistivity Vs Gamma Ray plot of the original dataset

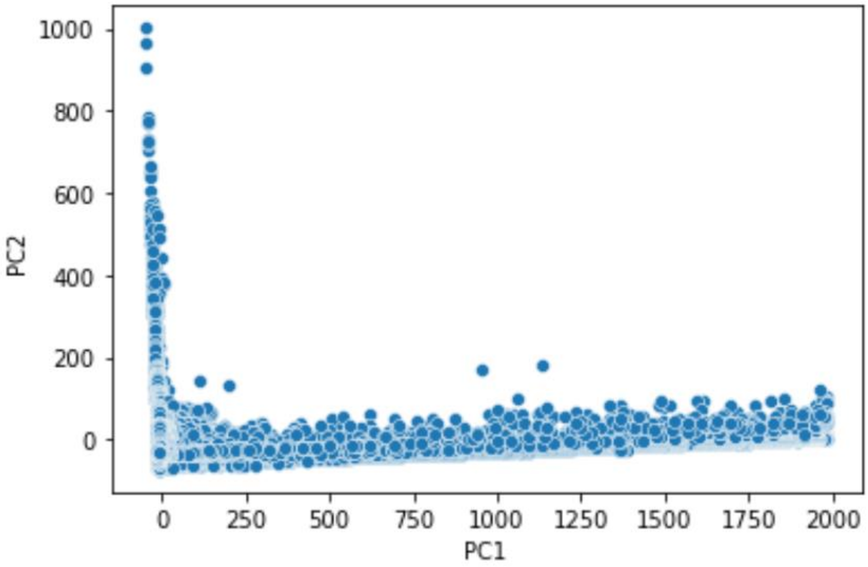
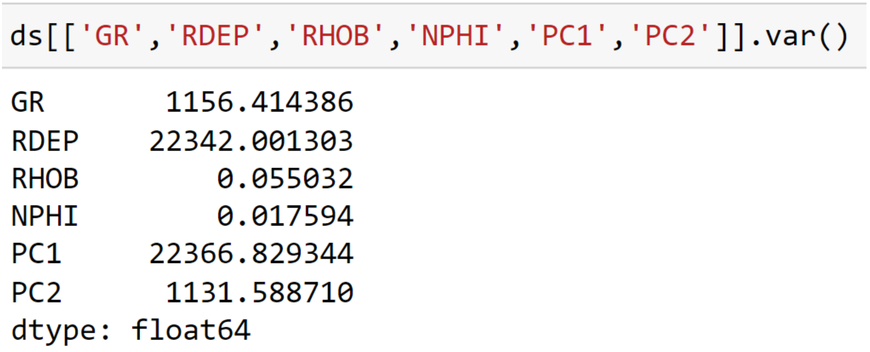


Figure 24 – PC1 Vs PC2 scatter plot from the bidimensional dataset

This suggests that the PCA algorithm computed the best correlating features out of the lot as it is dominated by the variables displaying high variance [34]. To confirm this the variance of the different numerical columns present in the original dataset were tested using the var() function.



```
ds[['GR','RDEP','RHOB','NPHI','PC1','PC2']].var()
```

```
GR         1156.414386
RDEP      22342.001303
RHOB          0.055032
NPHI          0.017594
PC1       22366.829344
PC2        1131.588710
dtype: float64
```

Figure 25 – Variance of attributes

As shown in figure 25 due to the high variance of Gamma ray and Resistivity as compared to the low variance of Neutron porosity index and Bulk Density index, the principal components are dominated by them. This also explains the influence of Gamma ray on the K-Means algorithm. In this specific scenario the use of feature selection instead of dimensionality reduction would have in theory yielded similar results for the clustering analysis.

6. Limitations of this study

Dimensionality reduction techniques such as t-SNE were explored but the hardware was unable to produce the bidimensional components as the process surpassed its memory limit. Implementing the t-SNE algorithm on the individual well log datasets would not have been a viable solution to this as clusters of data needed to be formed on the entire dataset at once to find meaningful patterns. DBSCAN was also explored as a method for clustering data, but also required more memory than the hardware was able to provide.

7.Professional, Legal and Ethical issues

7.1. Professional and Legal issues

All the work presented in this document that is not my own has been appropriately referenced. This study did not involve any sensitive datasets. The dataset that was used for this study is publicly available and was obtained from the FORCE 2020 machine learning competition [16]. The original well log data comes from the Norwegian government provided by a NOLD 2.0 license and compiled by EXPLOCROWD [35]. Therefore, this study does not present any legal issues. The software used i.e., Jupyter notebook is a 100% open-source software, free for all to use and released under the liberal terms of the modified BSD license [36].

7.2. Ethical issues

This study did not involve any user roles or interaction with human subjects; therefore, it is not in violation of any ethics code.

8. Conclusion and Future work

This study aimed to use unsupervised machine learning to identify regions of hydrocarbons in a well log dataset. Initially, manual well log interpretation was performed on the well logs to correctly identify hydrocarbon bearing zones, later, upon the successful implementation of K-Means and SOM on the bidimensional dataset obtained from PCA, clusters were formed which provided information on the presence of hydrocarbon in the well log with SOM performing slightly better than K-Means. The results of this experiment were tabulated, and graphical visualizations were provided. Additionally, the original dataset was also explored to understand the influence of the selected attributes on the performance of the clustering analysis and dimensionality reduction technique. It was inferred that the Gamma Ray values influenced the way in which K-Means formed its clusters on the well log data. This was attributed to the fact that the high variance of Gamma ray and Resistivity values dominated the components obtained from Principal Component Analysis. The Gamma Ray influence on K-means leads to a high false positive rate as a region of low Gamma Ray does not necessarily mean an underlying presence of hydrocarbon, therefore SOM was found to be a preferable method. Based on the comparison of manually interpreted regions and those identified by SOM, it was found that SOM performed well but not perfectly. In conclusion the combined indications obtained from K-Means and SOM prove to be useful in detecting the presence of hydrocarbon in the well log data.

The t-SNE and DBSCAN algorithms could not be performed, but these experiments show promise for future work, along with other machine learning techniques such as Gaussian Mixture Models and Hidden Markov Models. Supervised deep learning methods and Fuzzy Inference Systems can also be performed on labelled well log datasets.

References

[1] W. HU, J. BAO, and B. HU, “Trend and progress in global oil and gas exploration,” *Petroleum Exploration and Development*, vol. 40, no. 4, pp. 439–443, Aug. 2013, doi: 10.1016/S1876-3804(13)60055-5.

[2] J. Craig and F. Quagliaroli, “The oil & gas upstream cycle: Exploration activity,” *EPJ Web of Conferences*, vol. 246, p. 00008, 2020, doi: 10.1051/epjconf/202024600008.

[3] T. S. Bressan, M. Kehl de Souza, T. J. Girelli, and F. C. Junior, “Evaluation of machine learning methods for lithology classification using geophysical data,” *Computers & Geosciences*, vol. 139, p. 104475, Jun. 2020, doi: 10.1016/J.CAGEO.2020.104475.

[4] “Basic exploration geophysics (Book) | OSTI.GOV.” https://www.osti.gov/biblio/6982729 (accessed Apr. 20, 2022).

[5] “How Do We Actually Find Oil?. Geological Modeling for Dummies | by Erik Engheim | Medium.” https://erik-engheim.medium.com/how-do-we-actually-find-oil-4d0e58d67004 (accessed Aug. 15, 2022).

[6] O. Serra, “Fundamentals of well-log interpretation, 2. The interpretation of logging data.,” *Fundamentals of well-log interpretation, 2. The interpretation of logging data.*, 1986, Accessed: Apr. 18, 2022. [Online]. Available: https://books.google.com/books/about/Fundamentals_of_Well_log_Interpretation.html?id=o3jxAAAAMAAJ

[7] R. Baker, “Oil and Natural Gas: Offshore Operations,” *Encyclopedia of Energy*, pp. 581–594, Mar. 2004, doi: 10.1016/B0-12-176480-X/00258-8.

[8] N. B. Ishwar and A. Bhardwaj, “Petrophysical Well Log Analysis for Hydrocarbon exploration in parts of Assam Arakan Basin, India.”

[9] B. Steingrímsson, “GEOPHYSICAL WELL LOGGING: GEOLOGICAL WIRELINE LOGS AND FRACTURE IMAGING”.

[10] G. Asquith, D. Krygowski, S. Henderson, and N. Hurley, “Basic well log analysis,” *Basic well log analysis*, 2004, doi: 10.1306/MTH16823.

[11] I. M. Mohamed, S. Mohamed, I. Mazher, and P. Chester, “Formation Lithology Classification: Insights into Machine Learning Methods,” *Proceedings - SPE Annual Technical Conference and Exhibition*, vol. 2019-September, Sep. 2019, doi: 10.2118/196096-MS.

[12] O. Ijasan, C. Torres-Verdín, and W. E. Preeg, “Interpretation of porosity and fluid constituents from well logs using an interactive neutron-density matrix scale,” *http://www.seg.org/interpretation*, vol. 1, no. 2, pp. T143–T155, Oct. 2013, doi: 10.1190/INT-2013-0072.1.

[13] L. A. Lubis, D. P. Ghosh, and M. Hermana, “Elastic and Electrical Properties Evaluation of Low Resistivity Pays in Malay Basin Clastics Reservoirs,” *IOP Conference Series: Earth and Environmental Science*, vol. 38, no. 1, Aug. 2016, doi: 10.1088/1755-1315/38/1/012004.

[14] M. Mahmoud, M. Ghorab, T. Shazly, A. Shibl, and A. A. Abuhagaza, “Reservoir characterization utilizing the well logging analysis of Abu Madi Formation, Nile Delta, Egypt,” *Egyptian Journal of Petroleum*, vol. 26, no. 3, pp. 649–659, 2017, doi: 10.1016/j.ejpe.2016.11.003.

[15] T. M. Hossain, J. Watada, I. A. Aziz, and M. Hermana, “Machine learning in electrofacies classification and subsurface lithology interpretation: A rough set theory approach,” *Applied Sciences (Switzerland)*, vol. 10, no. 17, Sep. 2020, doi: 10.3390/app10175940.

[16] P. Bormann, P. Aursand, F. Dilib, S. Manral, and P. Dischington, “FORCE 2020 Well well log and lithofacies dataset for machine learning competition,” Dec. 2020, doi: 10.5281/ZENODO.4351156.

[17] S. Patruno, H. Kombrink, and S. G. Archer, “Cross-border stratigraphy of the Northern, Central and Southern North Sea: a comparative tectono-stratigraphic megasequence synthesis,” *Geological Society Special Publication*, vol. 494, no. 1, pp. 13–83, 2022, doi: 10.1144/SP494-2020-228/ASSET/E97D76AD-4AEB-4F03-9822-4EB5F2D9D797/ASSETS/IMAGES/LARGE/02_GSLSPECPUB2020-228F32.JPG.

[18] T. S. Madhulatha, “An Overview on Clustering Methods,” *IOSR Journal of Engineering*, vol. 02, no. 04, pp. 719–725, May 2012, doi: 10.48550/arxiv.1205.1117.

[19] T. Kohonen, “The Self-Organizing Map,” *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990, doi: 10.1109/5.58325.

[20] R. E. Brackenridge, V. Demyanov, O. Vashutin, and R. Nigmatullin, “Improving Subsurface Characterisation with ‘Big Data’ Mining and Machine Learning,” *Energies (Basel)*, vol. 15, no. 3, Feb. 2022, doi: 10.3390/en15031070.

[21] J. A. Hartigan and M. A. Wong, “Algorithm AS 136: A K-Means Clustering Algorithm,” 1979.

[22] H. Humaira and R. Rasyidah, “Determining The Appropiate Cluster Number Using Elbow Method for K-Means Algorithm,” Mar. 2020, doi: 10.4108/EAI.24-1-2018.2292388.

[23] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise,” 1996. [Online]. Available: www.aaai.org

[24] “Comparison of Subspace Projection Method with Traditional Clustering Algorithms for Clustering Electricity Consumption Data | Semantic Scholar.” https://www.semanticscholar.org/paper/Comparison-of-Subspace-Projection-Method-with-for-Piao-Park/bf1dce45a409e2fdaca65dda5ca6558d9a61f751#citing-papers (accessed Apr. 20, 2022).

[25] D. Digitalcommons@usu, S. Andreas, F. San, A. Fault, and J. Bryan, “Clustering and Classifying Geophysical Rock Properties of the San Andreas Fault,” 2020, Accessed: Apr. 12, 2022. [Online]. Available: https://digitalcommons.usu.edu/phys_capstoneproject/85

[26] H. Abdi and L. J. Williams, “Principal component analysis,” *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, Jul. 2010, doi: 10.1002/WICS.101.

[27] I. Jolliffe, “Principal Component Analysis,” *Encyclopedia of Statistics in Behavioral Science*, Oct. 2005, doi: 10.1002/0470013192.BSA501.

[28] L. van der Maaten and G. Hinton, “Visualizing Data using t-SNE,” *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[29] H. Yang, H. Pan, H. Ma, A. A. Konaté, J. Yao, and B. Guo, “Performance of the synergetic wavelet transform and modified K-means clustering in lithology classification using nuclear log,” *Journal of Petroleum Science and Engineering*, vol. 144, pp. 1–9, Aug. 2016, doi: 10.1016/J.PETROL.2016.02.031.

[30] C. F. Chun, W. W. Kok, H. Eren, and R. Charlebois, “Lithology classification using self-organizing map,” *IEEE International Conference on Neural Networks - Conference Proceedings*, vol. 1, pp. 526–531, 1995, doi: 10.1109/ICNN.1995.488233.

[31] “GitHub - JustGlowing/minisom: MiniSom is a minimalistic implementation of the Self Organizing Maps.” https://github.com/JustGlowing/minisom (accessed Aug. 15, 2022).

[32] S. Bubeck and U. von Luxburg, “Overfitting of clustering and how to avoid it,” 2007.

[33] J. A. F. Costa, “Clustering and visualizing SOM results,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6283 LNCS, pp. 334–343, 2010, doi: 10.1007/978-3-642-15381-5_41/COVER.

[34] I. Jolliffe, “Principal Component Analysis,” *Encyclopedia of Statistics in Behavioral Science*, Oct. 2005, doi: 10.1002/0470013192.BSA501.

[35] “EXPLOCROWD.” https://www.explocrowd.com/ (accessed Apr. 11, 2022).

[36] “The 3-Clause BSD License | Open Source Initiative.” https://opensource.org/licenses/BSD-3-Clause (accessed Apr. 18, 2022).