Final Report

Daniel Addokwei Tetteh

June 20th, 2022

Development of a Data-driven Model for Estimation of Reservoir Properties (Shale Volume, Porosity, And Fluid Saturations) Using Well Logs: A Case Study of the Volve Field; Norwegian North Sea

1. Introduction

Well logging is the discrete or continuous recording of measurements within a wellbore with a well logging tool or a probe (Hearst and Nelson, 1985). The logging tools are usually placed at the end of a wireline tool and lowered into a petroleum well to measure rock and fluid properties. Standard well logs in the oil and gas industry include resistivity, density, sonic, gamma-ray, caliper, and neutron porosity logs. Well logs provide essential information about geologic formations, which are used in the analysis of geophysical properties of wells, evaluation of formation rock characteristics, estimation of hydrocarbons initially in place, investigation of reservoir pressures, and estimation of in-situ petrophysical, geomechanical and geochemical reservoir properties which is essential for reservoir modeling and production forecasting amongst others (Mondol, 2015).

Despite the numerous advantages of using well logs, well logging tools do not directly measure these properties. On the contrary, these properties are obtained by processing, interpreting, and calibrating the well logs and thus are associated with several uncertainties that may affect the overall outcome (Moore et al., 2011). Furthermore, petroleum engineers and geophysicists use statistical and empirical methods together with well logs to build three-dimensional reservoir models for reserve estimation and improvement of production. However, the process is usually costly, cumbersome, and associated with human errors (Wang et al., 2021).

Data-driven models have recently been a better alternative in reservoir property estimation and characterization using well logs (Fajana et al., 2018; Saproetti et al., 2018). Machine learning has tremendous potential in predicting reservoir properties better than conventional methods using the large volumes of well log data presently available in the oil and gas industry. It is a less cumbersome approach and can be done at a minimal cost.

2. Objectives

1. The primary objective of this project is to develop a machine learning data-driven model to estimate reservoir properties, including shale volume, porosity, and fluid saturation,

based on a standard set of well logs, including gamma-ray, bulk density, neutron porosity, resistivity, and sonic.

2. A secondary objective is to cluster well logs from the training data into various electrofacies (i.e., clusters showing rock intervals with specific characteristics). These electrofacies would be depth-matched and compared with lithological plots obtained from the original database as an extra step in validating the performance of the clustering algorithm.

2.1.Learning Objectives

This project would promote a thorough exploration of the Exploratory Data Analysis (EDA), Imputation techniques, and outlier detection and removal techniques. Also, feature engineering, model development, and model performance improvement techniques like cross-validation and data stratification will be investigated. Overall deployment of data-driven models as a cost-effective solution to pertinent petroleum engineering problems is investigated.

3. Data Source and Acquisition

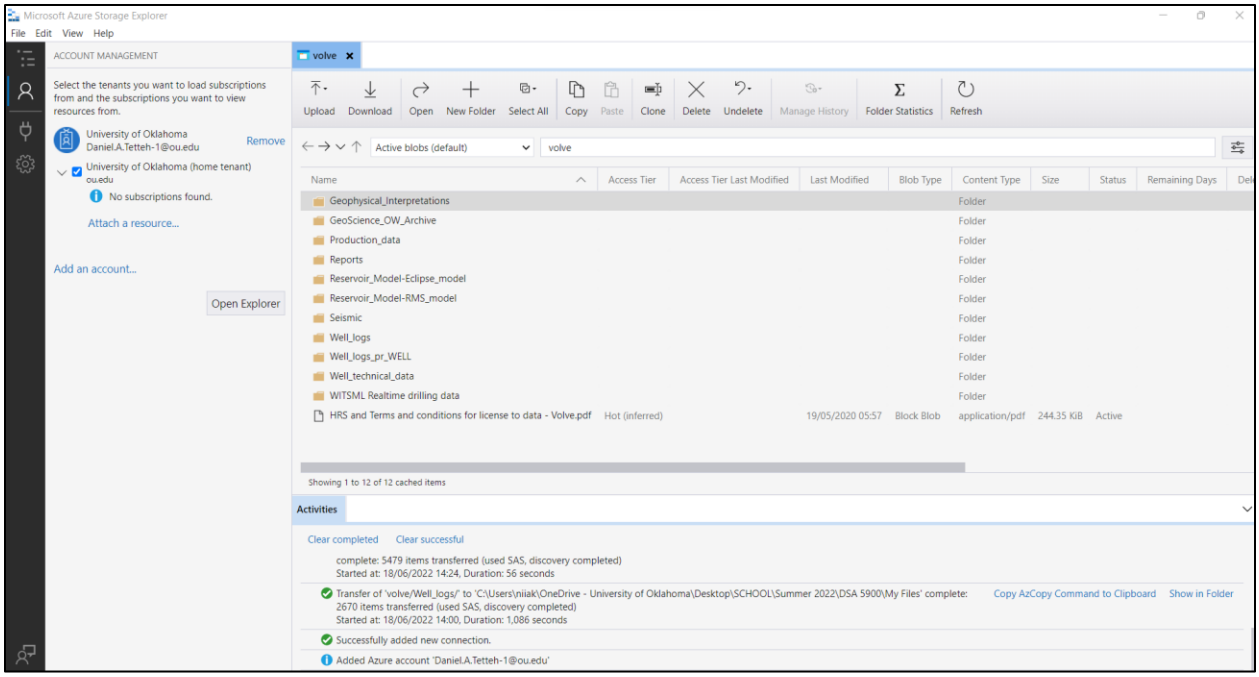
Data employed in this work is from the Volve Field owned by Equinor.

Field Description: The Volve is a Norwegian offshore petroleum field discovered in 1993. The field is located about 5 km from the north Sleipner Ost field, where the water depth is about 85m deep. The Hugin sandstone formation was the main producer from the field, with a permeability of about 2913 millidarcy (Deepak). The field has a porosity of about 0.226 and showed a peak oil production of 56,000 barrels per day with an overall production of about 63 million barrels (Sen et al., 2019). The data was made available for public consumption in 2018.

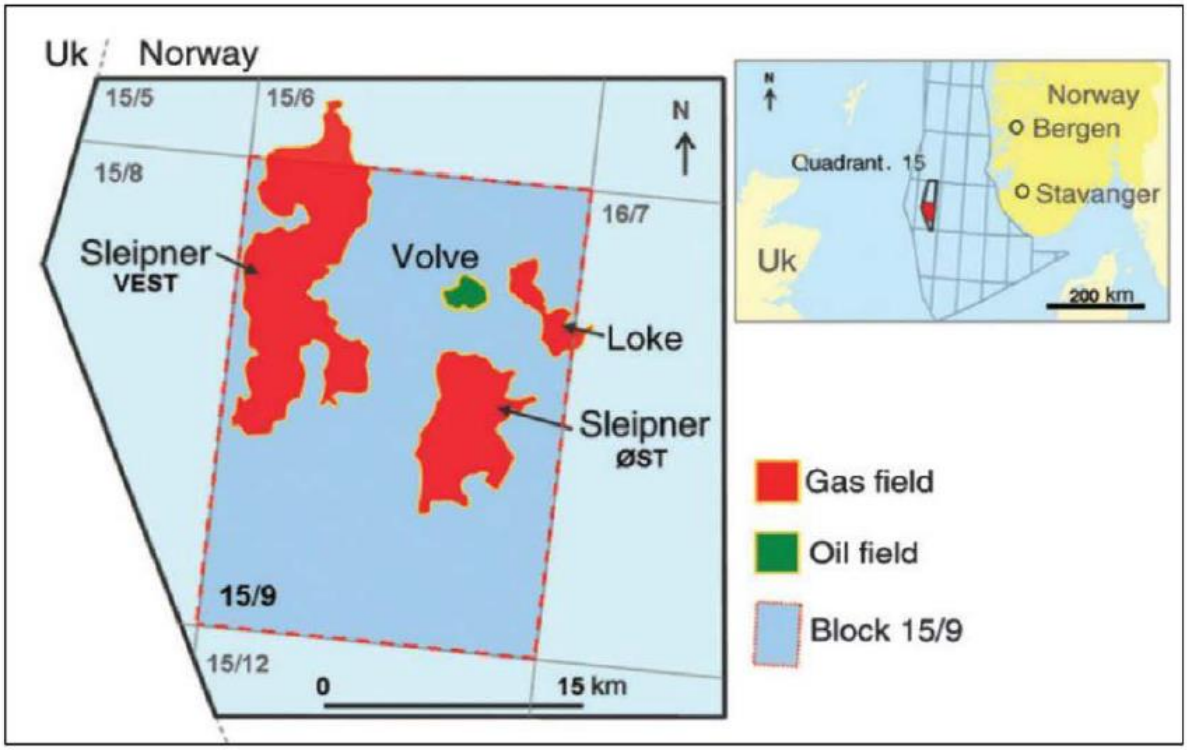
Data Acquisition: Microsoft Azure Storage Explorer was used to acquire data from the Equinor official website (https://www.equinor.com/energy/volve-data-sharing.

Available data include, Well log, reservoir modeling, seismic, production, and geophysical data, amongst others. For this work, data of interest include well logs, lithology and reservoir properties, field maps, and geological maps.

Key Challenges: Key challenges faced in the data acquisition process were extracting pertinent information from cumbersome oil field data formats and converting them into usable formats in a python coding environment. Oil and gas datasets usually have numerous inconsistencies and errors, making data mining very slow and unreliable. A .csv format form of the same data was obtained from the Society of Petrophysics and Well Log Analysts (SPWLA) PDDA (Petrophysical Data-Driven Analytics) division (https://github.com/pddasig/Machine-Learning-Competition-2021). The data has about 10 different wells with well log data and target petrophysical data to be predicted. i.e., the volume of shale (VSH), porosity (PHIF), and water saturation (SW). The training and test data sets obtained were distinct from each other. The data obtained from SPWLA was cross-checked and validated with the data acquired from the official equinor website.



Volve Field Data on Microsoft Azure Storage Explorer



Volve Field Location in the North Sea (Ravasi et al., 2015).

4. Workflow

The workflow for this project is described as follows.

1. Description of data source and acquisition
2. Explanation of fundamental concepts in estimating petrophysical properties from well logs.
3. **Exploratory Data Analysis**: Data description, univariate analysis, bivariate analysis, handling of missing values and outliers, and final data preparation for modeling. The process would be done for both the training and test datasets.
4. **Unsupervised Learning**: Well logs from the training datasets would be clustered into electrofacies using unsupervised classification algorithms; K-Means and Agglomerative Clustering. The clusters would be depth-matched and compared to lithological plots from the original dataset to validate the clustering algorithm.
5. **Supervised Learning**: Four main regression algorithms, Linear Regression, Support Vector Machine, K-Nearest Neighbours, and Random Forest models, would be explored in developing a single best model for predicting the target features. Optimization techniques like feature engineering and boosting algorithms would be employed to optimize the model's performance where applicable.
6. **Train-test split and Model Validation**: The training dataset would be divided into a training dataset and a validation data set. Standard methods like data stratification and cross-validation validation would be employed to avoid data imbalance and over-fitting or under-fitting the regression models. The performance of the models would be determined using R^2 and adjusted R^2 values and root-mean-square-error values on training and validation datasets.
7. Finally, the best model would be used to predict the test dataset's target variables. As an extra step to validate the model performance, the well logs of the test dataset would be clustered into elctrofacies and depth-matched in a similar manner to the training dataset. Also, the predicted values (i.e., Volume of Shale, Porosity, and Water saturation) would be plotted against depth and compared to their respective values for a given formation depth as seen in lithology plots from the original volve dataset from Equinor.

5. Fundamentals of Petrophysical Analysis

Before describing the exploratory data analysis (EDA) process, this section explains some primary petrophysical analysis concepts are explained showing meaningful relationships between well logs and the petrophysical properties to be predicted.

5.1.Estimation of Formation Water Saturation

The famous Archie’s equation, as shown below, is used to determine the saturation of water (Sw) in an uninvaded zone in a formation. The equation shows the dependency of water saturation on formation resistivity, obtained from resistivity logs.

$$S_w^{\,n} = \frac{R_w}{(\Phi^m \times R_t)}$$

where:

- S_w = water saturation of the uninvaded zone
- n = saturation exponent, which varies from 1.8 to 4.0 but normally is 2.0
- R_w = formation water resistivity at formation temperature
- Φ = porosity
- m = cementation exponent, which varies from 1.7 to 3.0 but normally is 2.0
- R_t = true resistivity of the formation, corrected for invasion, borehole, thin bed, and other effects

The Archie's Equation ((Archie, 1952); AAPG, 2022)

5.2. Porosity Estimation from Well Logs

Formation bulk density, neutron, and sonic logs (which usually depict shear and compressional wave travel times) are used in estimating formation porosity. The associated equations are shown below;

$$\Phi = \frac{\rho_{ma} - \rho_b}{\rho_{ma} - \rho_f}$$

where:

- Φ = porosity
- ρ_{ma} = matrix density (see table below)
- ρ_b = formation bulk density (log value)
- ρ_f = density of the fluid saturating the rock immediately surrounding the borehole—usually mud filtrate (use 1.0 for freshwater and 1.1 for saltwater mud)

Use the lithology matrix densities to determine porosity and average P_e to determine lithology listed in the table below.

Lithology	Density, g/cc	Average pe
Sandstone	2.65	1.8
Limestone	2.71	4.8
Dolomite	2.876	3.0
Anhydrite	2.977	5.05
Salt	2.032	4.6

Porosity estimation from formation density (AAPG, 2021; Alberty, 1992)

A standard early method for estimating porosity from neutron logs was proposed by (Brown and Bower; 1958) using neutron counts.

$$\log \phi = -mN_d + K$$

ϕ = porosity
N_d = neutron count
m = slope of best-fit line
K = a constant

Porosity estimation from neutron logs (Brown and Bower; 1958)

Also, in modern petroleum engineering, neutron and formation bulk density porosities are combined to estimate the true porosity of a formation.

Calculate porosity using the equation

$$\Phi = \left(\frac{\Phi_N^2 + \Phi_D^2}{2} \right)^{1/2}$$

where Φ is percent porosity, Φ_N is neutron percent porosity, and Φ_D is density percent porosity.

Formation porosity estimation equation from neutron and bulk density porosities (AAPG, 2021)

$$\frac{1}{v} = \frac{\phi}{v_f} + \frac{(1-\phi)}{v_{ma}}, \quad \text{.....(1)}$$

where

- ϕ = fractional porosity of the rock
- v = velocity of the formation (ft/sec)
- v_f = velocity of interstital fluids (ft/sec)
- v_{ma} = velocity of the rock matrix (ft/sec)

In terms of transit time (Δt):

$$\Delta t = \phi \Delta t_f + (1-\phi)\Delta t_{ma}, \quad \text{.....(2)}$$

or

$$\phi = \frac{\Delta t - \Delta t_{ma}}{\Delta t_f - \Delta t_{ma}}, \quad \text{.....(3)}$$

where

- Δt = acoustic transit time (µsec/ft)
- Δt_f = acoustic transit time of interstitial fluids (µsec/ft)
- Δt_{ma} = acoustic transit time of the rock matrix (µsec/ft)

Porosity estimation from acoustic logs (AAPG, 2021; Tenchov, 2016)

5.3 Estimation of Volume of Shale from GR logs

$$I_{GR} = \frac{GR_{log} - GR_{min}}{GR_{max} - GR_{min}} \qquad V_{Sh} = \frac{I_{GR}}{3 - 2\, I_{GR}}$$

The volume of shale estimation from GR logs (Moradi et al., 2016)

IGR = gamma ray index
GRmin = gamma ray response for the cleanest formation
GRmax = gamma ray response in shale layer
GRlog = gamma ray log value in zone of interest
Vsh= volume of shale

It must be mentioned that these estimations are not final. They are combined with outcomes from geological and reservoir models, laboratory core analysis, and tests to determine the appropriate reservoir petrophysical properties.

6. Data Exploration

The training data set has 17 features, as shown below;

- WELLNUM - Well, number
- DEPTH - Depth, unit in feet
- DTC - Compressional Travel-time, unit in nanosecond per foot
- DTS - Shear Travel-time, unit in microseconds per foot
- BS - Bit size, unit in inch
- CAL - Caliper, unit in Inc.
- DEN - Density, unit in Gram per cubic centimeter
- DENC - Corrected density, unit in Gram per cubic centimeter
- GR - Gamma Ray, unit in API
- NEU - Neutron, unit in dec
- PEF - Photo-electric Factor, unit in barns/e
- RDEP - Deep Resistivity, unit in Ohm.m
- RMED - Medium Resistivity, unit in Ohm.m
- ROP - Rate of penetration, unit in meters per hour
- PHIF - Porosity, a unit equals to the percentage of pore space in a unit volume of rock.
- SW - Water saturation
- VSH - Shale Volume

The depth unit was changed to meters to correlate the log and lithological plots from model development to that obtained from the field data. The original training dataset had 318967 rows with 17 columns. The statistics of the “raw” training data are shown in Table 1.

Table 1: Statistics of rawdata for training

	WELLNUM	DEPTH	DTC	DTS	BS	CALI	DEN	DENC	GR
count	318967.000000	318967.000000	318967.000000	318967.000000	318967.000000	318967.000000	318967.000000	318967.000000	318967.000000
mean	4.081012	2133.045195	-7791.051225	-8605.136709	-958.267420	-6982.041578	-7000.762197	-7205.371313	-43.026647
std	2.462805	1157.611118	4168.057493	3491.479107	2969.159669	4592.475975	4582.288708	4486.572798	990.102434
min	0.000000	102.156797	-9999.000000	-9999.000000	-9999.000000	-9999.000000	-9999.000000	-9999.000000	-9999.000000
25%	2.000000	1145.399959	-9999.000000	-9999.000000	8.500000	-9999.000000	-9999.000000	-9999.000000	23.330000
50%	4.000000	2104.700016	-9999.000000	-9999.000000	17.500000	-9999.000000	-9999.000000	-9999.000000	55.508400
75%	6.000000	3063.998854	-9999.000000	-9999.000000	26.000000	8.556900	2.246500	0.030300	78.978000
max	8.000000	4770.601431	181.813900	388.839700	36.000000	20.330400	3.089600	0.334158	1124.440000

	NEU	PEF	RDEP	RMED	ROP	PHIF	SW	VSH
count	318967.000000	318967.000000	318967.000000	318967.000000	318967.000000	318967.000000	318967.000000	318967.000000
mean	-7007.119486	-7221.969013	-949.836141	-804.023955	-989.366887	-8515.777016	-8515.693410	-8585.158486
std	4578.760758	4480.143550	2944.074718	4177.156020	3028.429856	3554.017382	3554.217716	3484.038870
min	-9999.000000	-9999.000000	-9999.000000	-9999.000000	-9999.000000	-9999.000000	-9999.000000	-9999.000000
25%	-9999.000000	-9999.000000	0.648200	0.669800	15.197600	-9999.000000	-9999.000000	-9999.000000
50%	-9999.000000	-9999.000000	1.102400	1.146000	24.969900	-9999.000000	-9999.000000	-9999.000000
75%	0.085100	0.058400	2.020650	2.154900	30.633700	-9999.000000	-9999.000000	-9999.000000
max	1.463474	13.840700	80266.800000	97543.400000	208.633000	0.403294	1.000000	3.654300

A significantly large number of the individual column data values was “-999,” which is quite unexpected for the features in context. A log plot and box plot were thus plotted to visualize the logs and target variables, as shown in fig 1 and 2, respectively. From the log plot, we can infer that the ‘-999’ values are possible outliers as no significant deflections are seen from the surface to a depth of about 2500m for all logs beside the gamma ray (GR) log. From domain knowledge of petroleum well logging, well logging tools are usually deployed at depths of interest (usually the pay zone). The Hugin sandstone, the primary producing formation in the Volve field, is found at a depth of about 2500m from the surface across all wells (Sen et al., 2019). The basic boxplot also showed many outliers.

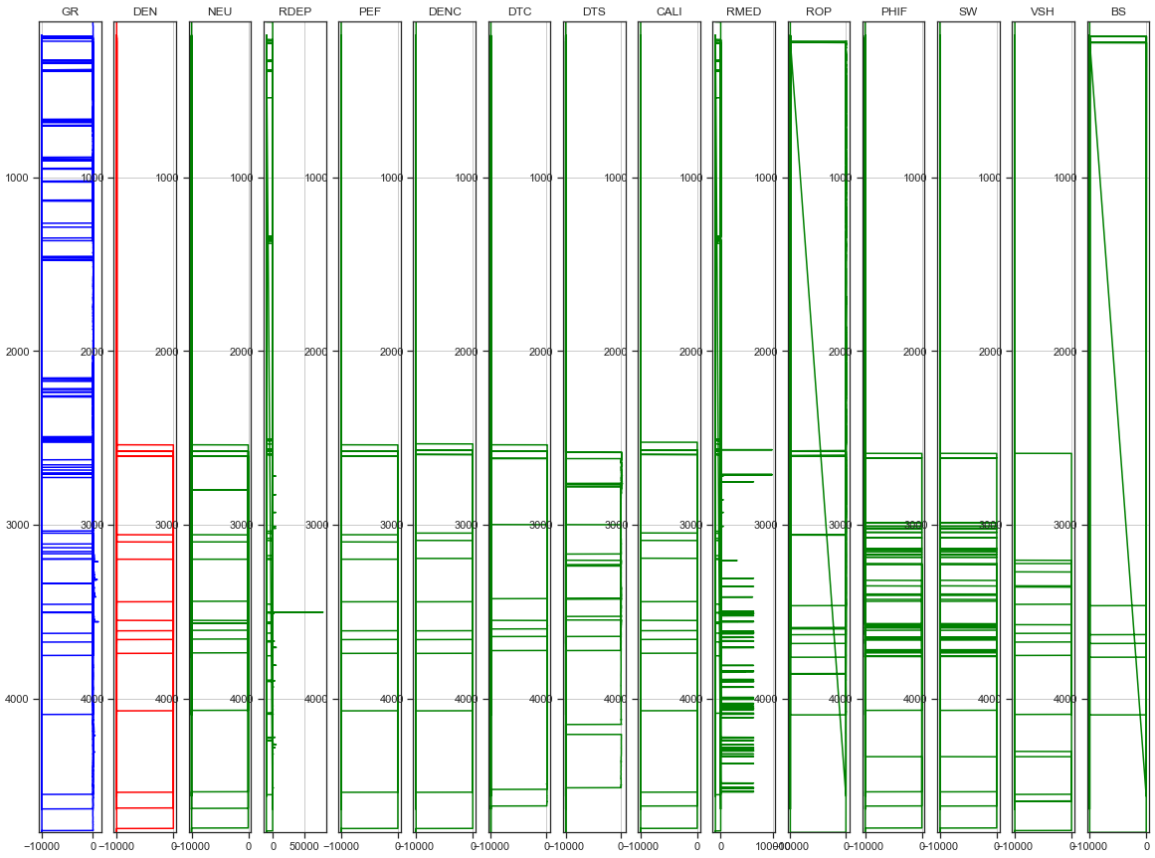


Fig 1. log plot of rawdata

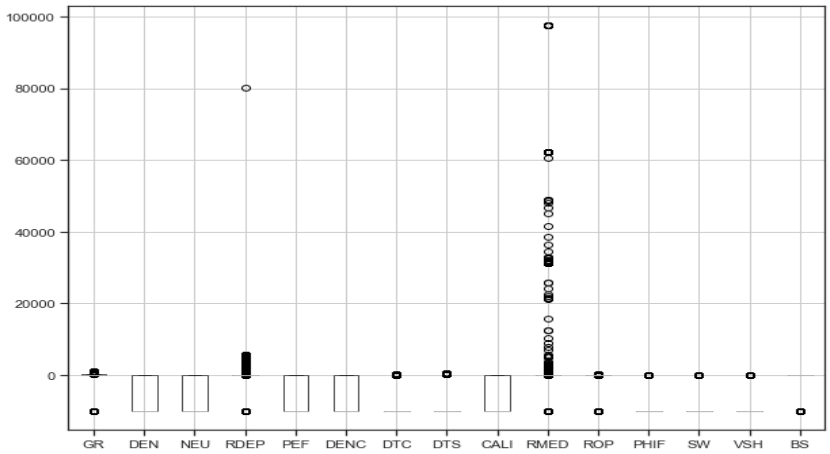


Fig 2. Boxplot of rawdata

6.1.Univariate Analysis

The following inferences were made from the univariate analysis of the raw data set.

1. Well number and well depth are uniformly and normally distributed, respectively, which is expected. This is because 'well number' is nominal. Also, we expect more data to be collected at a depth of interest, around 5000 to 10000ft.
2. Bit size is also nominal as it depends on the hole size being drilled.
3. ROP, DTS, and GR logs are right-skewed, while PEF showed some bimodality.
4. DTC, DTS, DEN, DENC, NEU, PEF, and PHIF have a more significant proportion of their distribution with average values of -999. This is unusual using domain knowledge as compression travel time, shear travel time, density, and neutron porosity logs do not usually provide such high negative values in oil and gas operations.

   Assumption: We would assume that this data was obtained from depths where the logging tools were not deployed, which are not of interest to the study.

5. RDEP and RMED values have averages of -948 and -804, respectively which is unusual for a conventional oil field like the Volve field. Resistivity logs are usually presented on logarithmic scales from 0.2 to 2000 ohms (https://glossary.oilfield.slb.com/en/terms/r/resistivity_log).

6.2.Missing Data

Before determining the missing data, all "-999" values were replaced with NAN values. Again, this was an assumption based on domain knowledge that logging tools are only deployed at depths of interest. Thus, no values (represented by -999 in this case) were recorded for depths from the surface to about 2400m. The statistics of the features after replacing '-999' values with NAN are shown in Table 2.

6.3.Data Cleaning

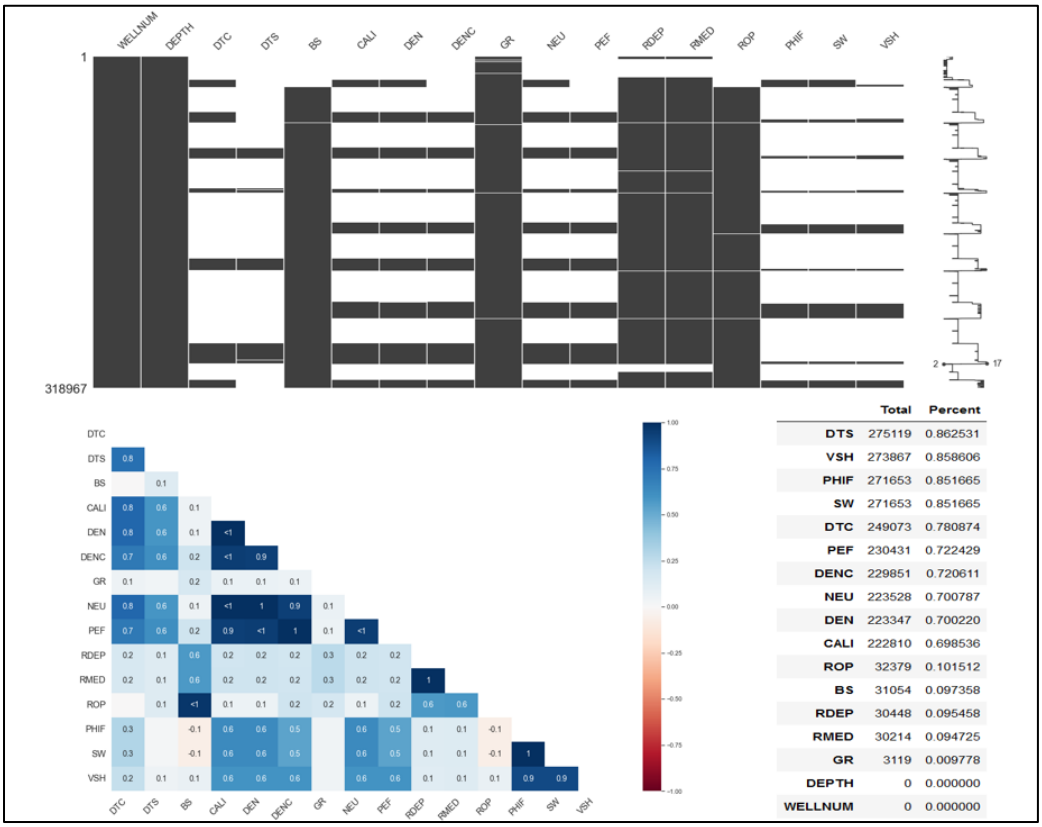
First, all missing values in the target variables, PHIF, SW, and VSH, were removed. This was done because the originality of the data in these features cannot be changed. The new dataset has a shape of 42309 rows and 17 columns which is good enough to build a model. Also, the statistics of this new dataset showed that they are within depths of interest. A minimum and a maximum depth of 2588.97m to 4744.80m, respectively, with an average of 3732.46m. The data still had missing values, as shown in Fig 4, although it reduced significantly. This dataset was referred to as rawdata1.

Table 2: rawdata statistics after replacing '-999' with NAN values.

	WELLNUM	DEPTH	DTC	DTS	BS	CALI	DEN	DENC	GR
count	318967.000000	318967.000000	69894.000000	43848.000000	287913.000000	96157.000000	95620.000000	89116.000000	315848.000000
mean	4.081012	2133.045195	77.155278	140.490795	16.856695	8.697590	2.452806	0.050244	55.288621
std	2.462805	1157.611118	15.387921	36.085217	7.071820	0.384869	0.156333	0.020949	38.602323
min	0.000000	102.156797	1.025100	74.822400	8.500000	6.000000	1.626600	-0.982700	0.148800
25%	2.000000	1145.399959	66.363775	119.019775	8.500000	8.578100	2.311100	0.043900	24.329750
50%	4.000000	2104.700016	72.396000	130.534100	17.500000	8.625000	2.506800	0.053100	55.983000
75%	6.000000	3063.998854	85.584000	144.340900	26.000000	8.687500	2.576100	0.060600	79.240600
max	8.000000	4770.601431	181.813900	388.839700	36.000000	20.330400	3.089600	0.334158	1124.440000

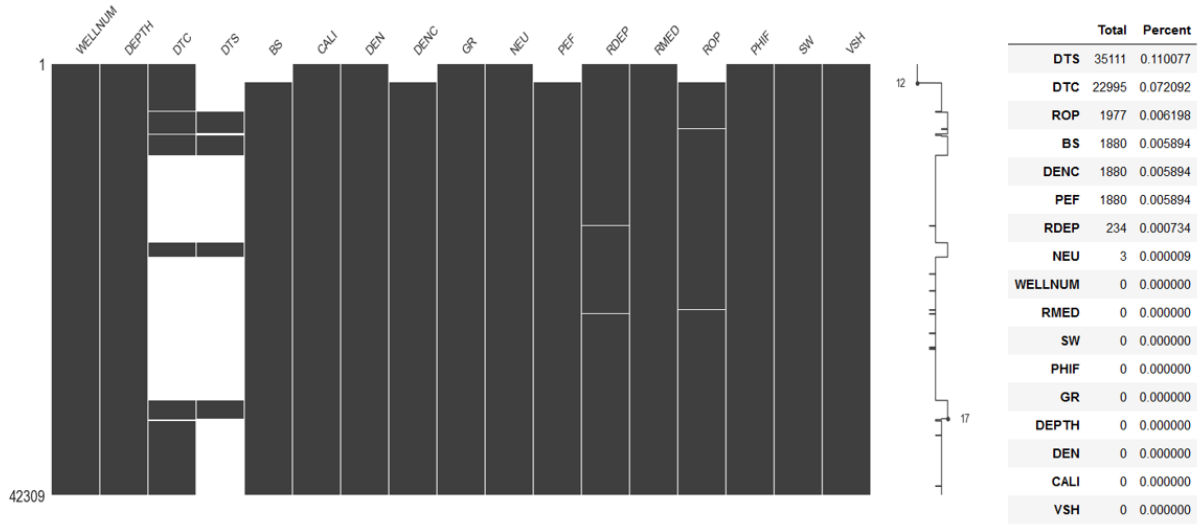
	NEU	PEF	RDEP	RMED	ROP	PHIF	SW	VSH
count	95439.000000	88536.000000	288519.000000	288753.000000	286588.000000	47314.000000	47314.000000	45100.000000
mean	0.173838	5.757870	5.140624	158.102867	28.550509	0.137371	0.701001	0.307895
std	0.095899	2.533365	168.911711	3082.521918	18.161868	0.085265	0.350021	0.254315
min	-0.003400	-0.023200	0.065000	0.064900	0.000000	0.000000	0.013000	-0.248000
25%	0.106300	4.891800	0.743400	0.769700	19.451200	0.060000	0.336000	0.118375
50%	0.156900	6.143150	1.208590	1.255200	25.916100	0.133000	0.933496	0.246509
75%	0.218700	7.737325	2.183900	2.348500	32.220200	0.216000	1.000000	0.390100
max	1.463474	13.840700	80266.800000	97543.400000	208.633000	0.403294	1.000000	3.654300

The missing data in the data afterward are summarized in Fig 3. The white spaces in the first plot represent the missing values. The percentage of missing values in the individual features and the fraction of the missing values between any two features are also depicted.



	Total	Percent
DTS	275119	0.862531
VSH	273867	0.858606
PHIF	271653	0.851665
SW	271653	0.851665
DTC	249073	0.780874
PEF	230431	0.722429
DENC	229851	0.720611
NEU	223528	0.700787
DEN	223347	0.700220
CALI	222810	0.698536
ROP	32379	0.101512
BS	31054	0.097358
RDEP	30448	0.095458
RMED	30214	0.094725
GR	3119	0.009778
DEPTH	0	0.000000
WELLNUM	0	0.000000

Fig 3: Missing value plots in rawdata



	Total	Percent
DTS	35111	0.110077
DTC	22995	0.072092
ROP	1977	0.006198
BS	1880	0.005894
DENC	1880	0.005894
PEF	1880	0.005894
RDEP	234	0.000734
NEU	3	0.000009
WELLNUM	0	0.000000
RMED	0	0.000000
SW	0	0.000000
PHIF	0	0.000000
GR	0	0.000000
DEPTH	0	0.000000
DEN	0	0.000000
CALI	0	0.000000
VSH	0	0.000000

Fig 4: Missing Values in rawdata1

6.4.Handling Missing Data

The missing values were handled using a data imputation technique employed by Akinnikawe et al. (Akinnikawe et al., 2018)

Imputation Technique: Petroleum well logs and petrophysical data are measured on different scales and are highly variable. This is due to the significant differences in mineralogical constituents of different rock formations encountered during well logging. The use of conventional machine learning imputation techniques like imputing mean or any other measure of central tendency would significantly affect the integrity of the data and model accuracy. In this work, we predicted the missing data using rows within rawdata1 without missing values with a K-Nearest Neighbors (KNN) linear regression model. The method was first used by (Akinnikawe et al., 2018) in generating synthetic Photoelectric and Unconfined compressive strength logs from other wireline logs. The missing data for DTS, DTC, PEF, DENC, and ROP were predicted and imputed as they were significantly large. Multicollinearity between the individual predictors was determined using the variance inflation factor (VIF) method, and a threshold value was determined based on the range of values of VIF. Generally, all the well logs were highly correlated hence very high VIF values. The defined threshold value was thus, 100 which is higher than the conventional value of 10. All other missing values were removed. The final dataset after imputation and cleaning was called ‘rawdata_new’ with 42072 rows and 15 columns. Log plots, distribution plots, and boxplots were done for the new data set rawdata_new.

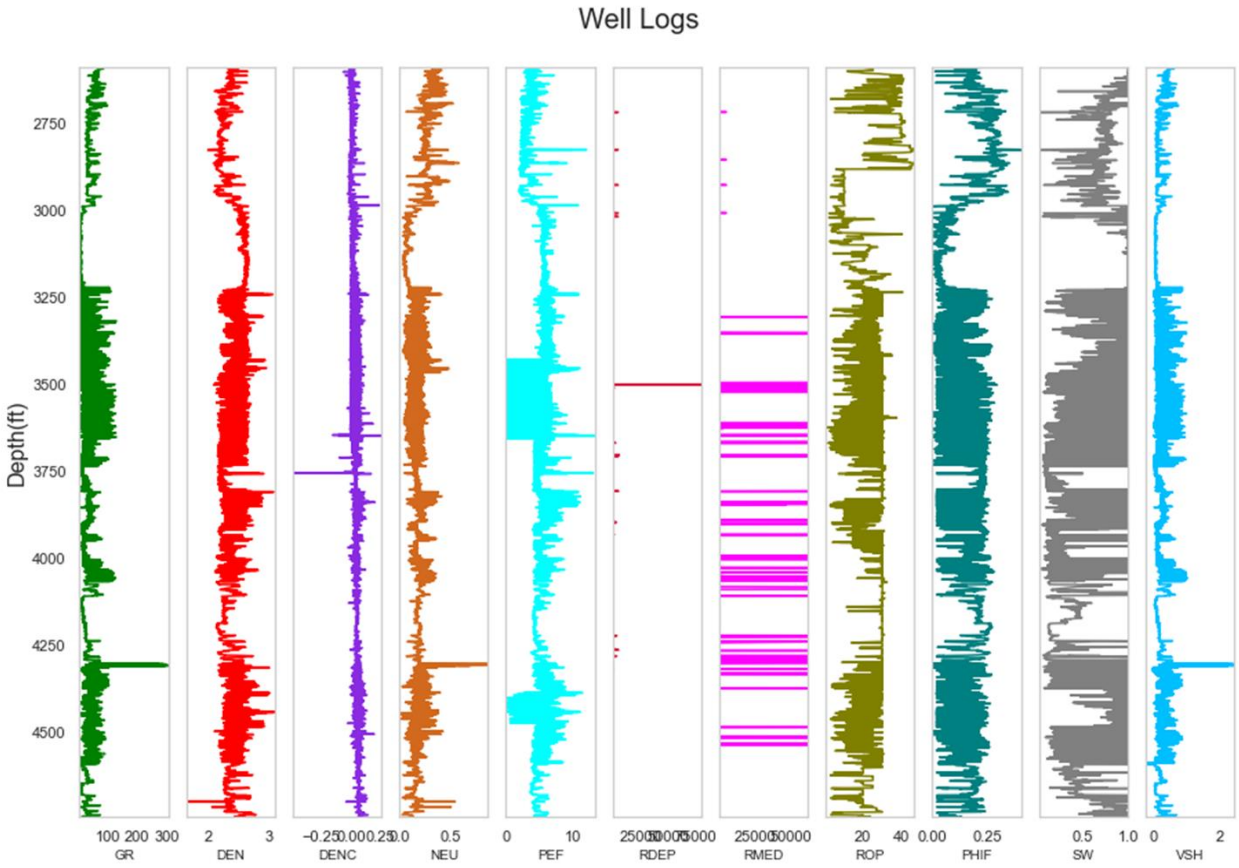


Fig 5. Log plot for rawdata_new

6.5.Handling Outliers

Outliers within the dataset were detected using the isolation forest algorithm, as shown in Fig 7. below. The boxplots also confirmed the presence of outliers. The z-score transform was used to handle the outliers, where all data that fell beyond a threshold value of 3 for all the features were removed. The new data set was named “rawd_zscore.” Other outlier handling techniques will be explored based on the model's performance. Significant number of outliers within the individual features were removed. The pair plot and boxplot below in figs 9 and 10 show how the data looked after outlier removal.

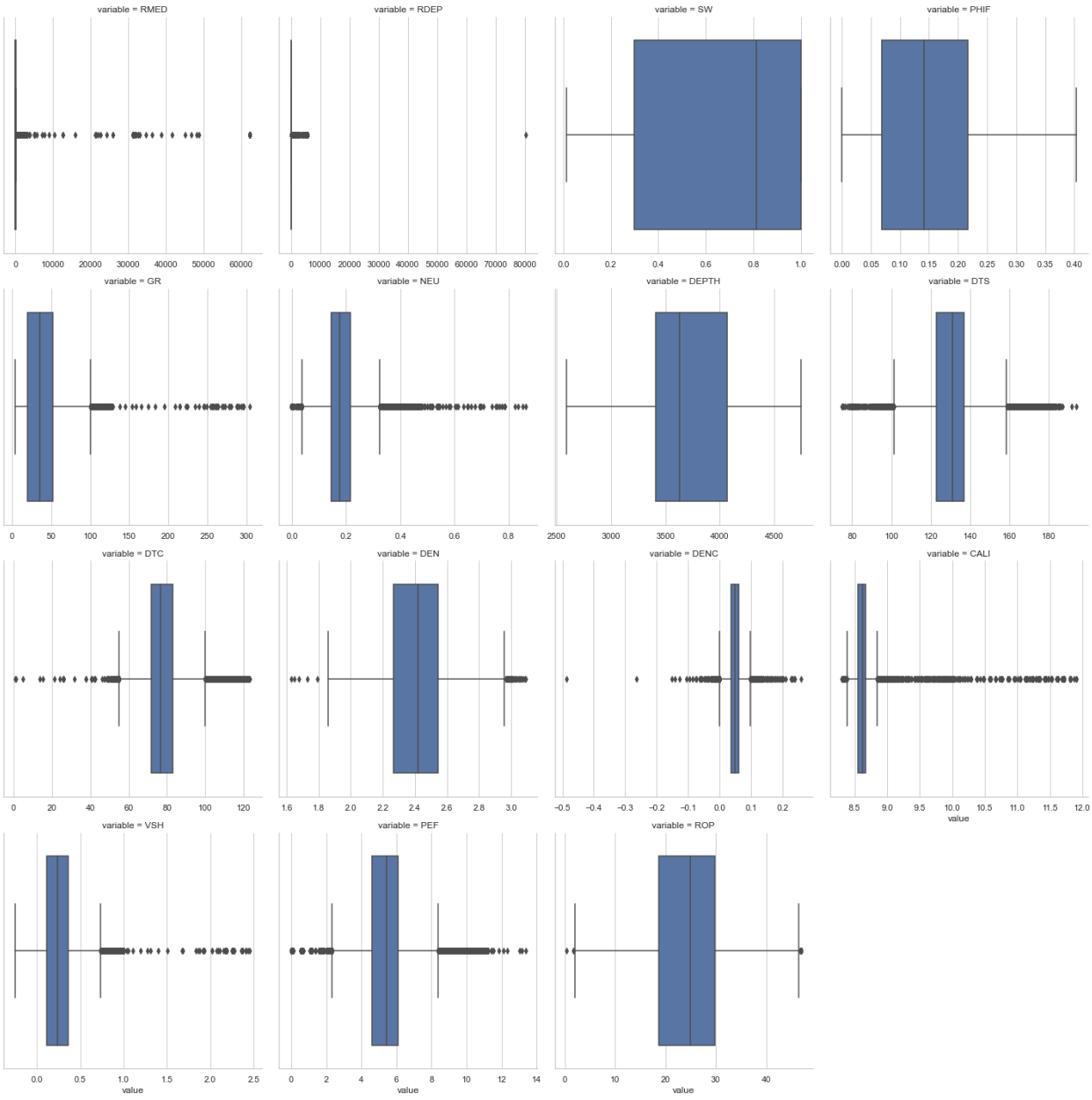


Fig 6. Boxplot for rawdata_new

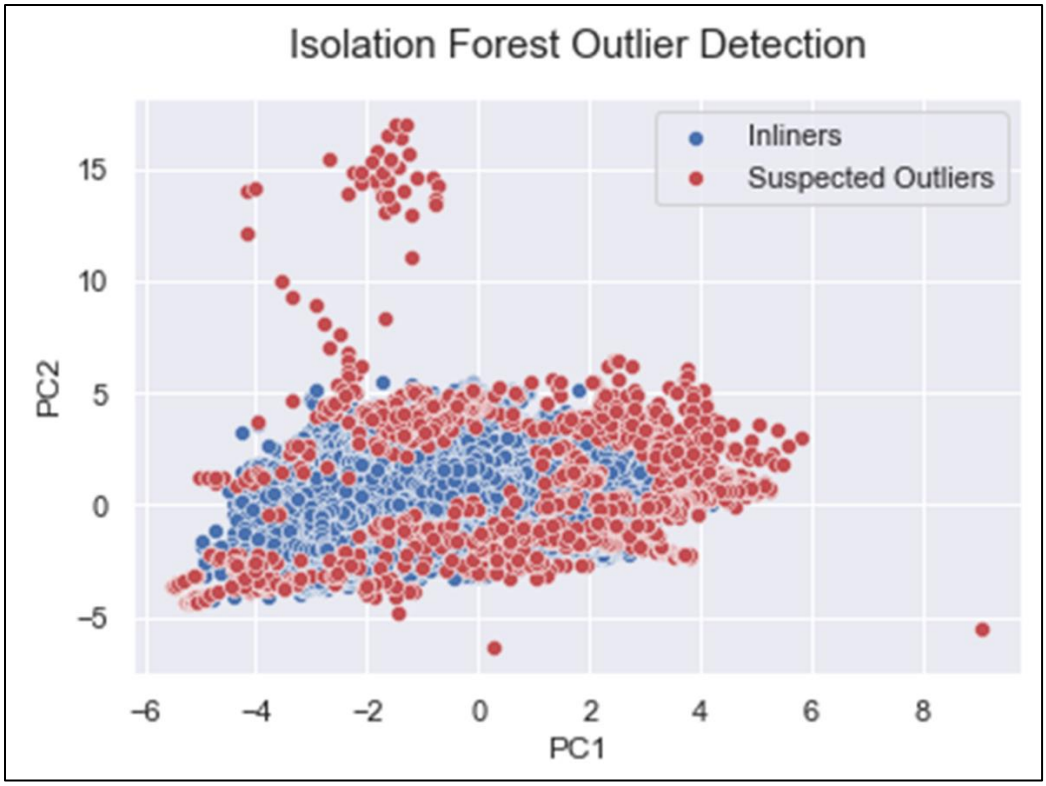


Fig 7: Outlier detection in rawdata_new

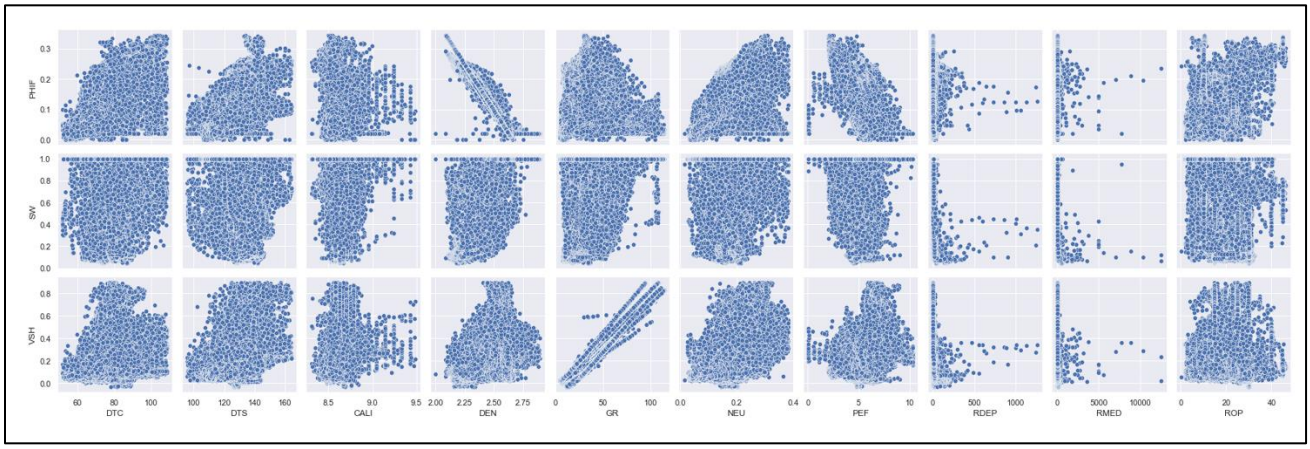


Fig 9: Pairplot after outlier removal

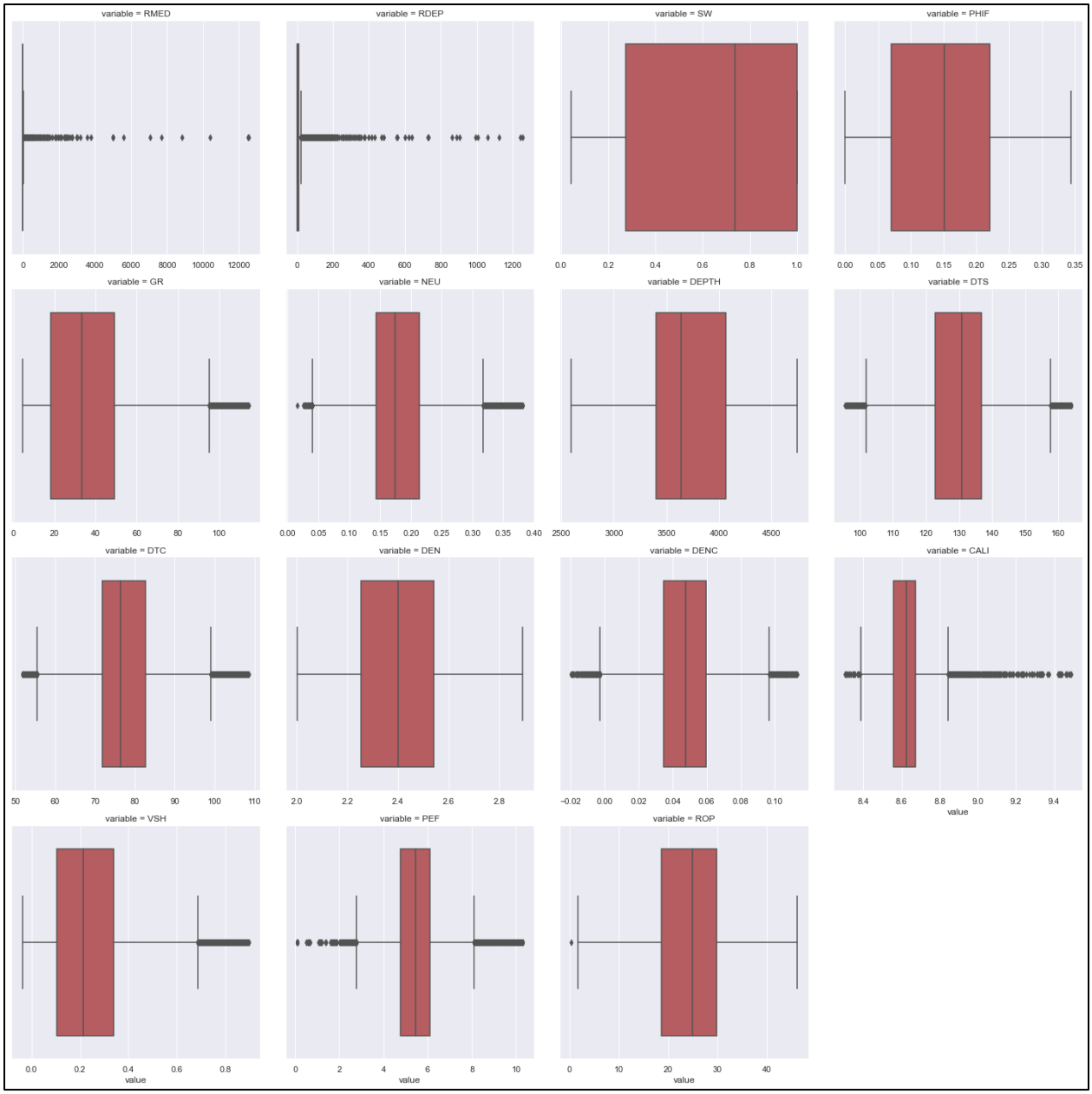


Fig 10: Boxplot after outlier removal

7. Machine Learning Methodology

Both Supervised and unsupervised learning techniques will be employed in this work. Two different data sets were employed in this process. Rawdata_new (without missing values but with outliers) and rawd_zscore (without missing values and outliers treated using a z-score threshold value of 3). This was done to evaluate the effects of outliers on the model performances.

7.1.Unsupervised Learning

K-means and agglomerative clustering were explored in clustering well logs into electrofacies (zonation). The clusters were compared with lithological maps from the acquired dataset to ascertain distinct formations and validate the performance of the clustering algorithm.

7.2.Supervised Learning

Regression models would be developed using the following machine learning algorithms; Linear Regression, K-Nearest Neighbors, and Random Forest. Gradient boosting and other optimization techniques would be employed where necessary. A train-test split would be used to divide the training data set for model training and validation. Stratification would be ensured in splitting to avoid bias in data sampling for training and validation. Grid-search cross-validation would be conducted to prevent over-fitting models and ensure optimal model performance. The target variables would be predicted using the test data.

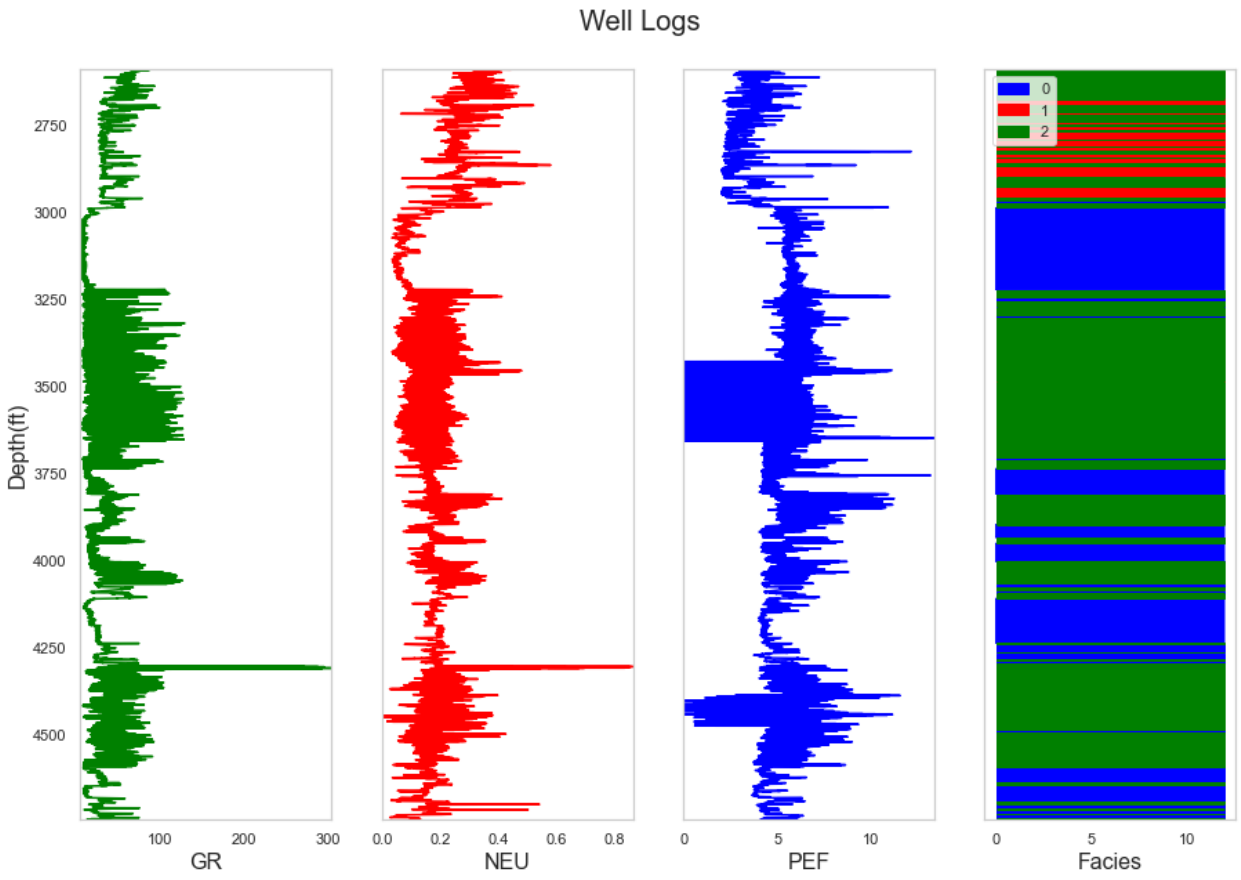
7.3.Result Validation

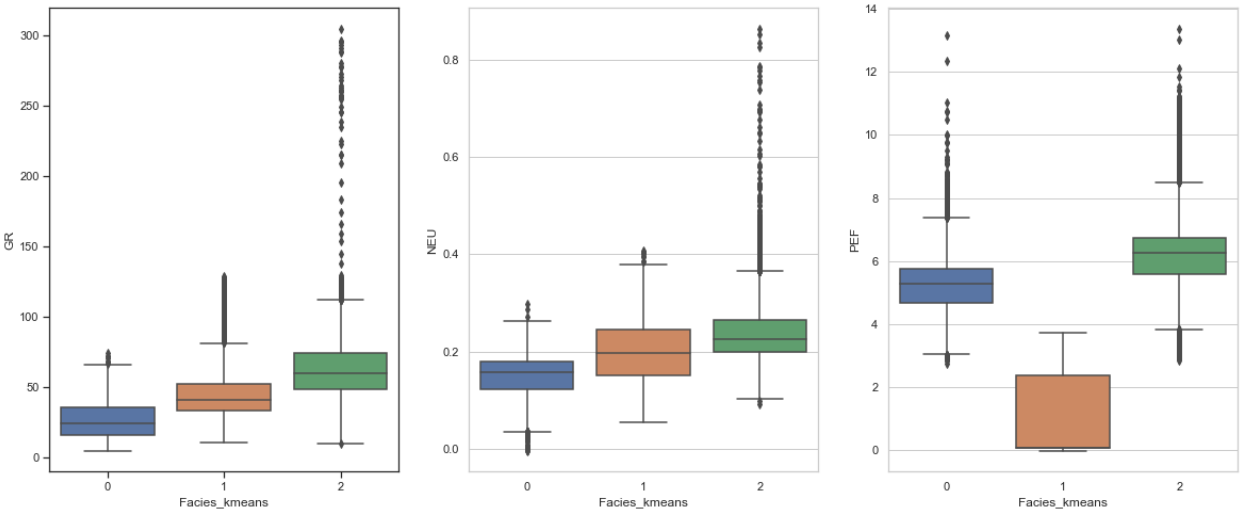
As an extra step to validate the performance of the best model above, the predicted target variables would be represented on a log plot and compared with log plots obtained from the equinor data set. The plots will be compared on a depth basis.

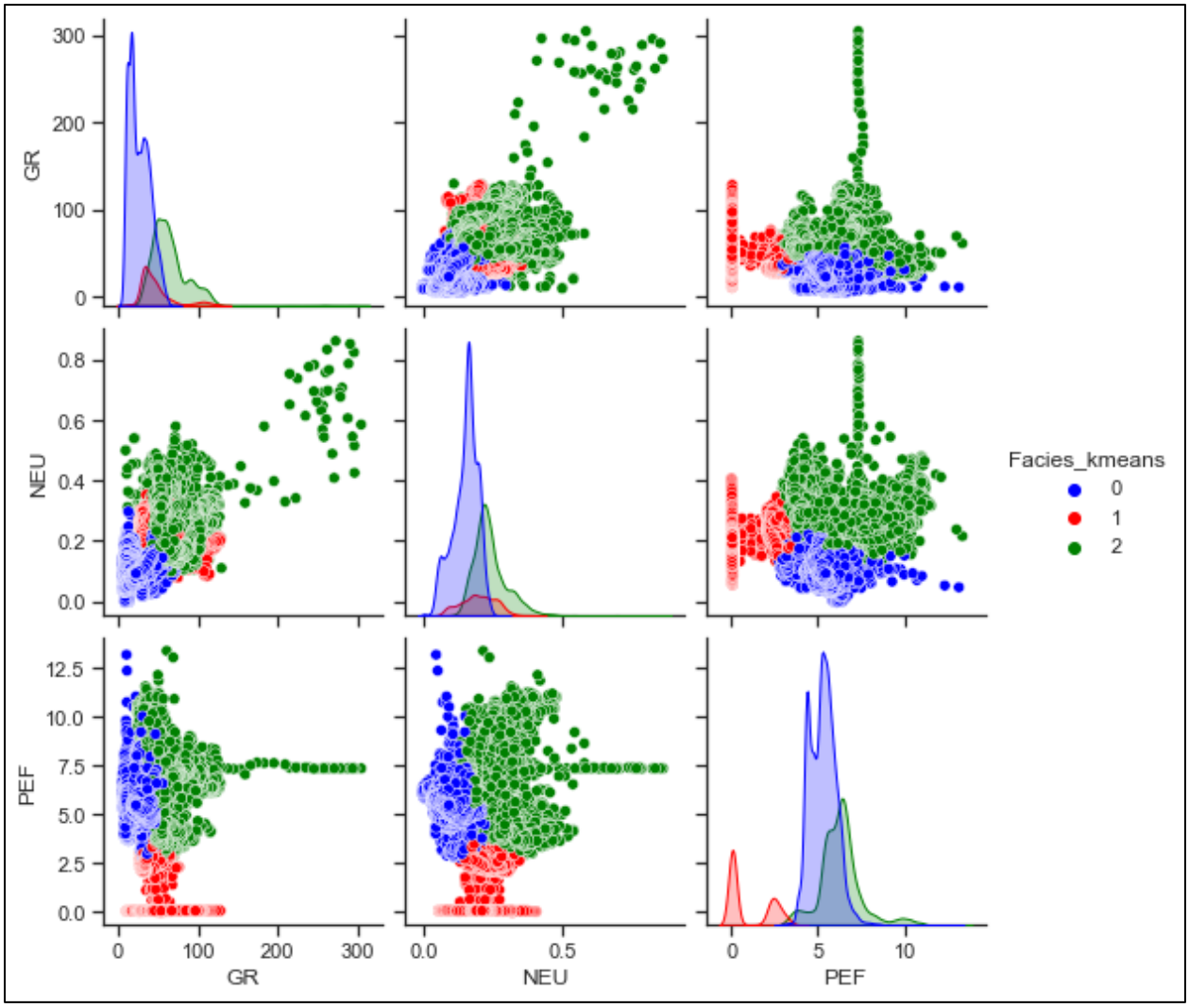
8. Results and Analysis

Unsupervised Clustering

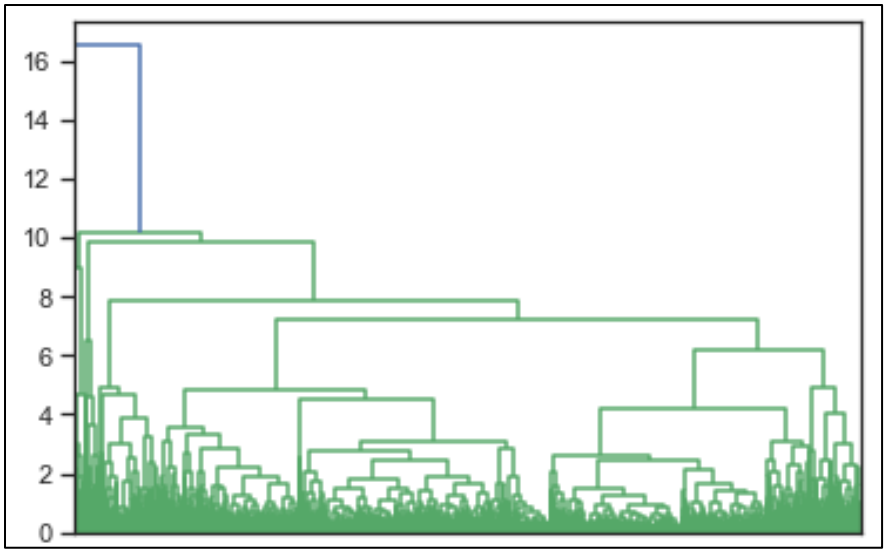
KMeans Clustering: GR, NEU, PEF, RDEP.RMED and DENC logs were selected for clustering. The data was scaled using the standard scaler method. A simple KMeans clustering algorithm was deployed on the scaled data and an “elbow plot” was made based on the “sum of square within” to determine the optimal number of clusters. This was further corroborated with a silhouette analysis plot. Based on the results of the silhouette analysis and elbow plots, 5 clusters were selected. Results however showed indistinct number of clusters. A boxplot was made with 5 clusters and significant overlap for datasets within the various clusters was observed. The clusters were also indistinct in the pair plots for the selected logs. The results from log and box plots for five clusters can be seen in the appendix. It was also observed that RMED, RDEP and DENC significantly affected the performance of the clusters from log plots. A new dataset with GR, NEU and PEF was created and used for clustering. An optimal number of clusters of 3 was determined and a more distinct clustering was observed based on plots.

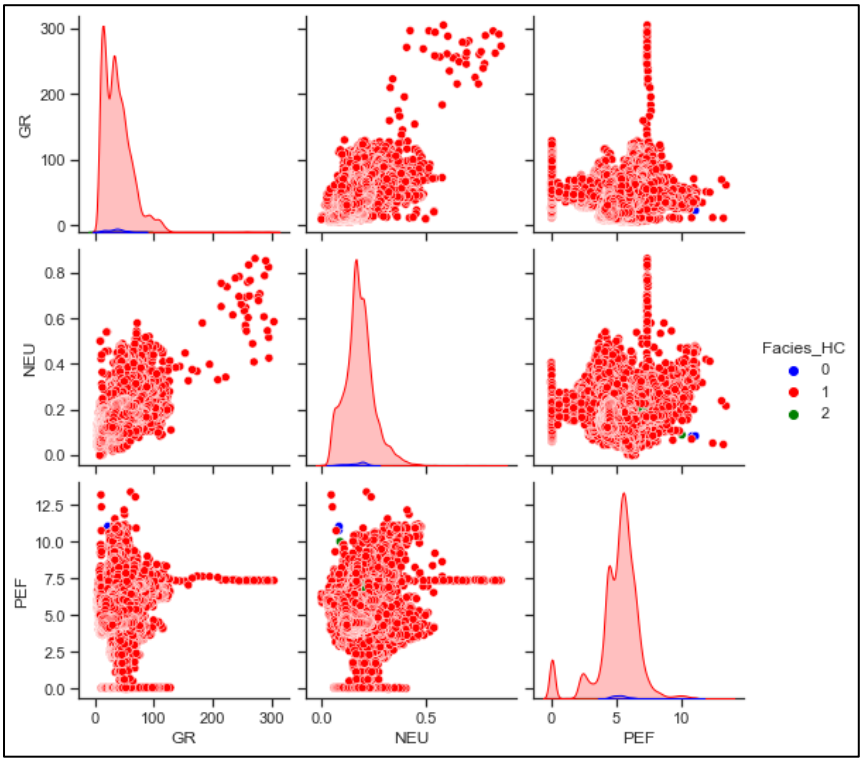

Well Logs
Depth(ft)
2750
3000
3250
3500
3750
4000
4250
4500
100
200
300
GR
0.0
0.2
0.4
0.6
0.8
NEU
0
5
10
PEF
0
1
2
0
5
10
Facies


GR
Facies_kmeans
NEU
Facies_kmeans
PEF
Facies_kmeans



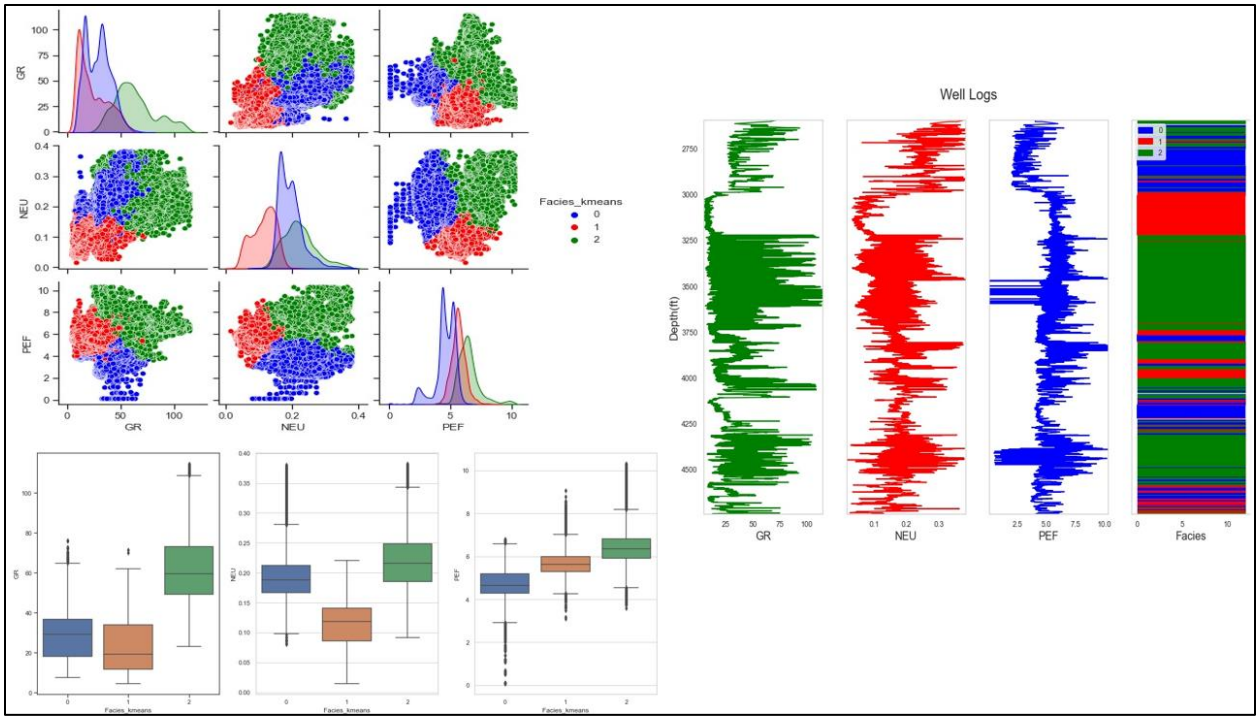
Hierarchical Clustering: The process for KMeans clustering was repeated for hierarchical clustering using the agglomerative clustering algorithm. The optimal number of clusters of 3 was used. Results showed that agglomerative clustering did do not perform well in clustering the data. This may be due to the high correlation within the individual features. The dendrogram and pair plots from the agglomerative clustering are shown in Figs F and G.

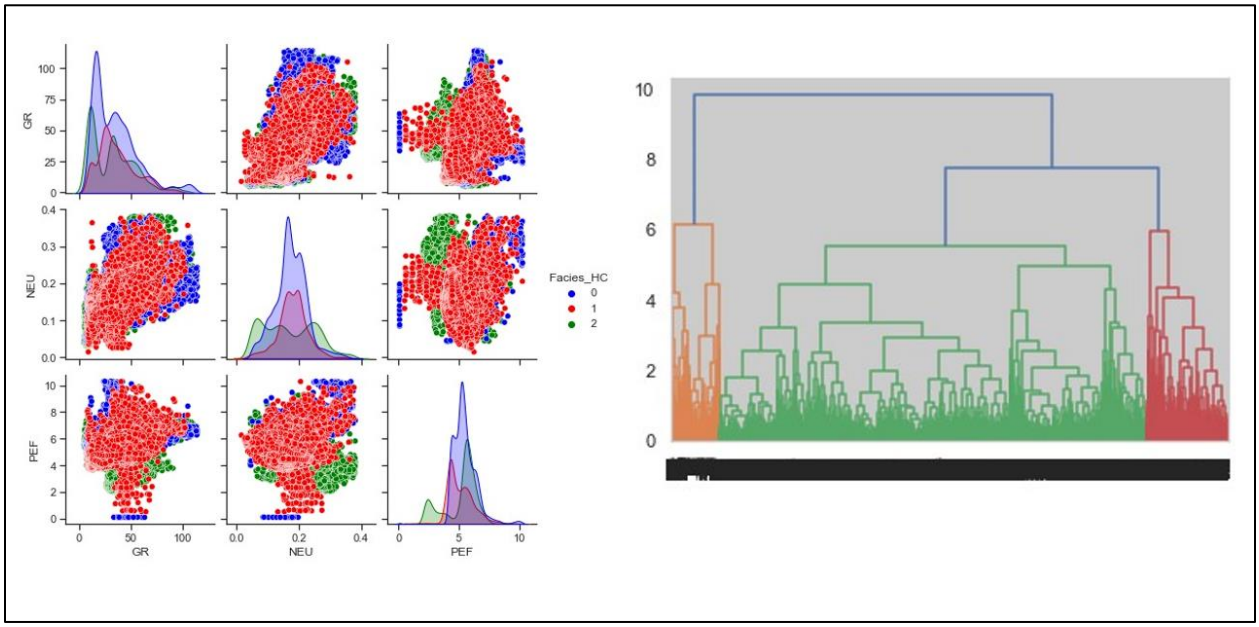




Effects of Outliers

To ascertain the effects of outliers the rawd_zscore dataset was used for KMeans and Hierarchical clustering following the same procedure with the same number of clusters. The results showed more distinct clustering in both instances, especially in hierarchical clustering for this dataset as shown below.





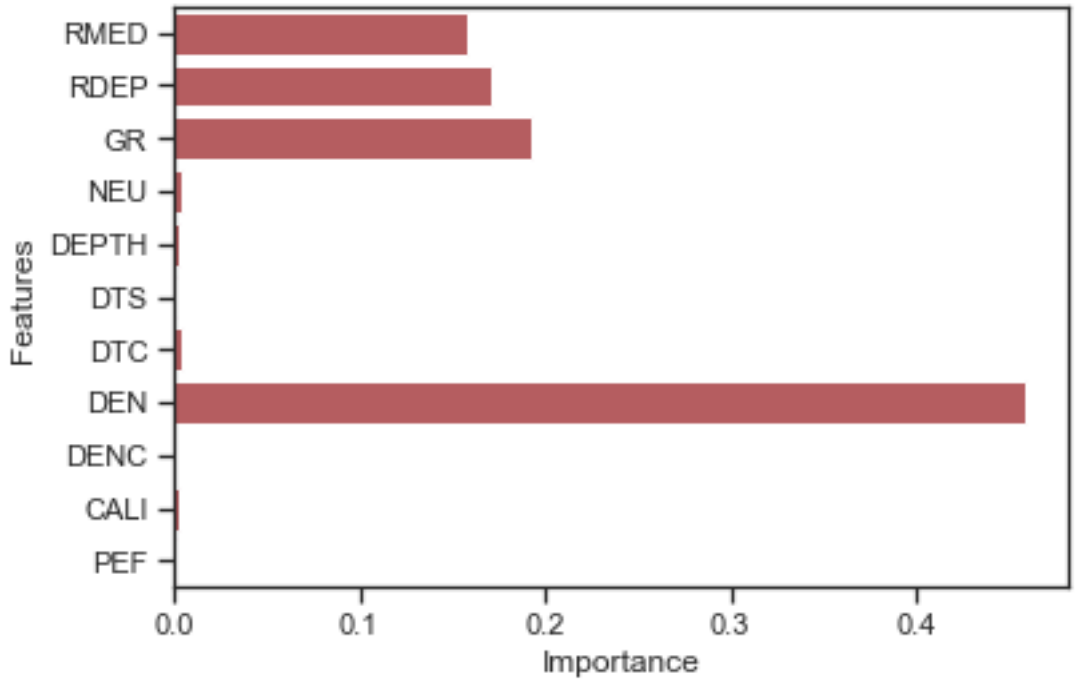
Supervised Learning

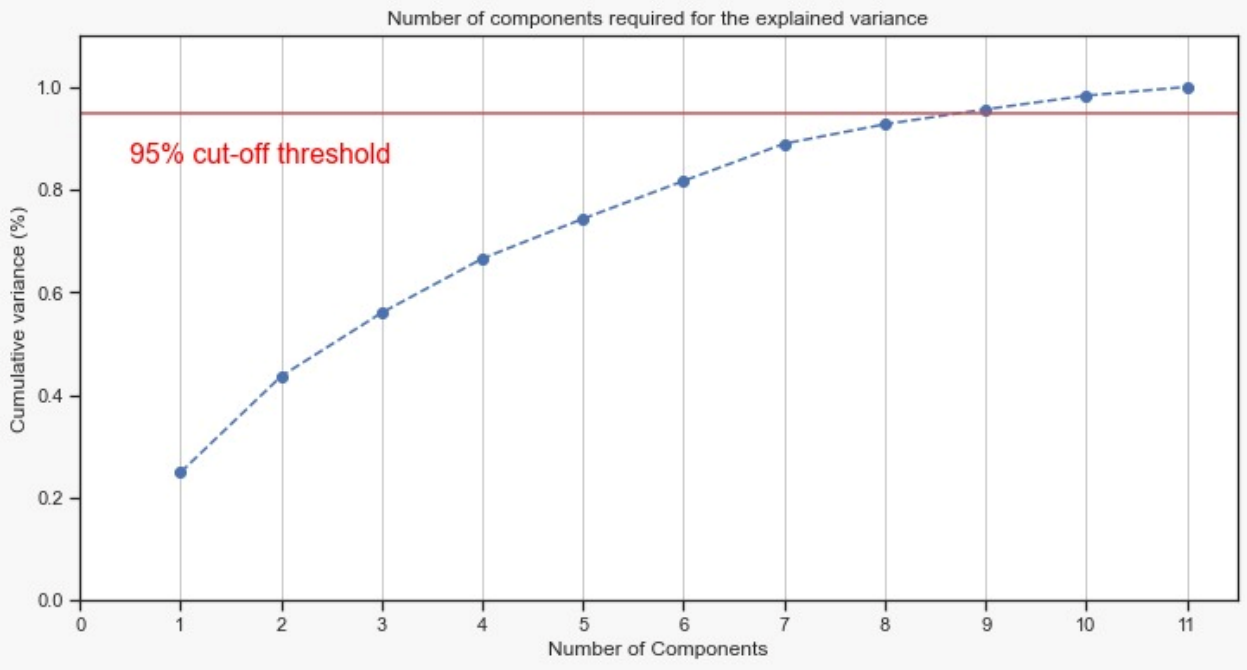
As discussed earlier, regression models were built with five (5) different machine learning algorithms; Linear Regression, K Nearest Neighbors (KNN), Support Vector Machine (SVM), Random Forest (RF) and Gradient Boost Regression (GBR). The data was divided into a training and a validation dataset in the ratio 85:15 respectively. A 5-fold cross-validation was carried out

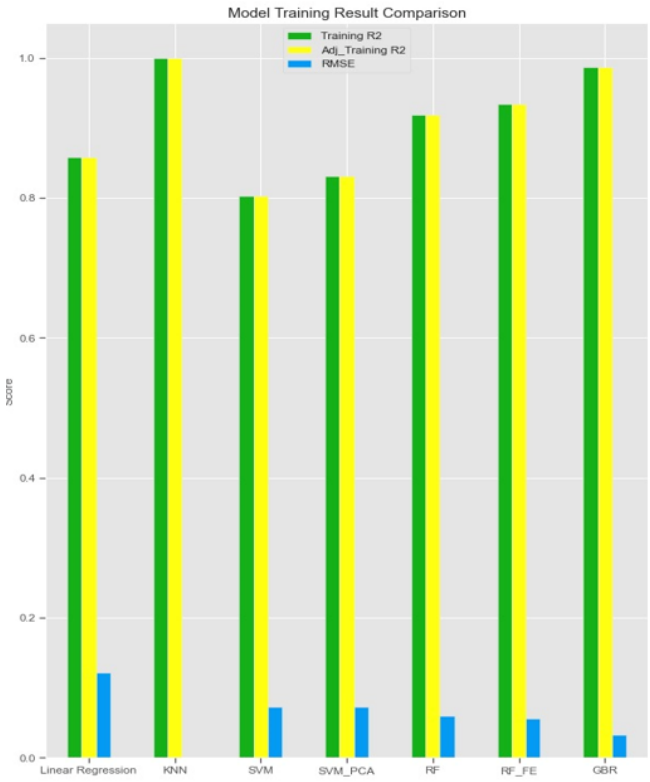
for all machine learning algorithms with hyper-parameter variation except linear regression. This was done to prevent overfitting of the training dataset. The regression algorithms were ranked based on their R-squared, adjusted R-squared and root-mean-squared-error (RMSE) values. The initial results (as shown in Fig C) showed KNN and GBR as the best models based on their RMSE and R-squared values.

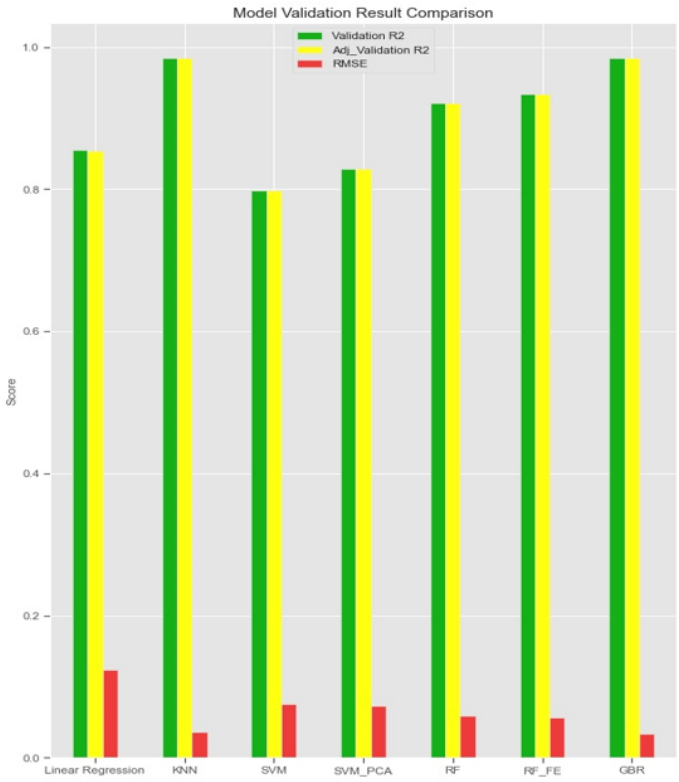
Feature Engineering: Principal component analysis was conducted on the data set separately for Support Vector Machine. A cumulative explained variance vs number of components plot was created to determine the optimal number of components for a threshold cumulative explained variance of 95%. An optimal number of components of 9 was selected. A new SVM model was developed with 9 eigenvectors (principal components) and the metrics measured.

In order to improve upon the performance of the random forest model, a feature importance plot was created, and the important features were selected as seen in Fig B. DTS, DENC AND PEF were dropped from the original features. An improvement was seen in the performance of the random forest model. The results are summarized in Fig C. below. Graphical representations of the performance of each of the model can be found in the appendix.





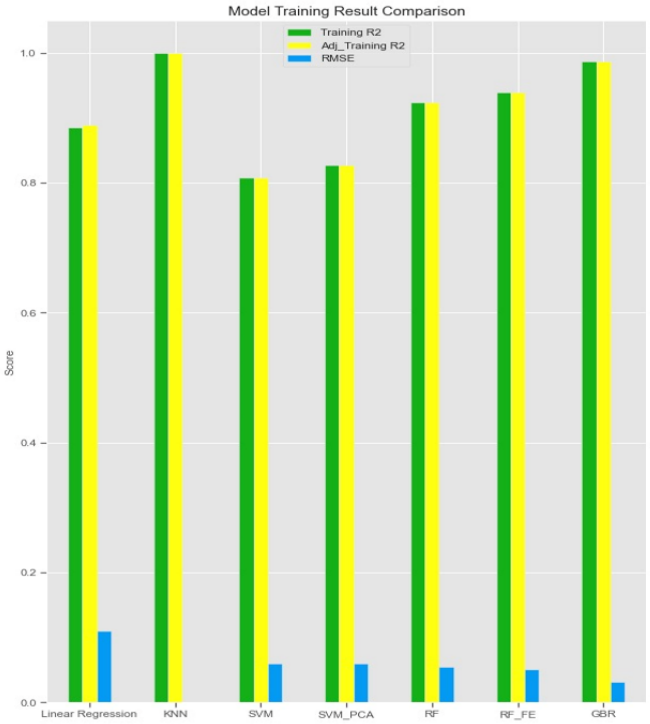


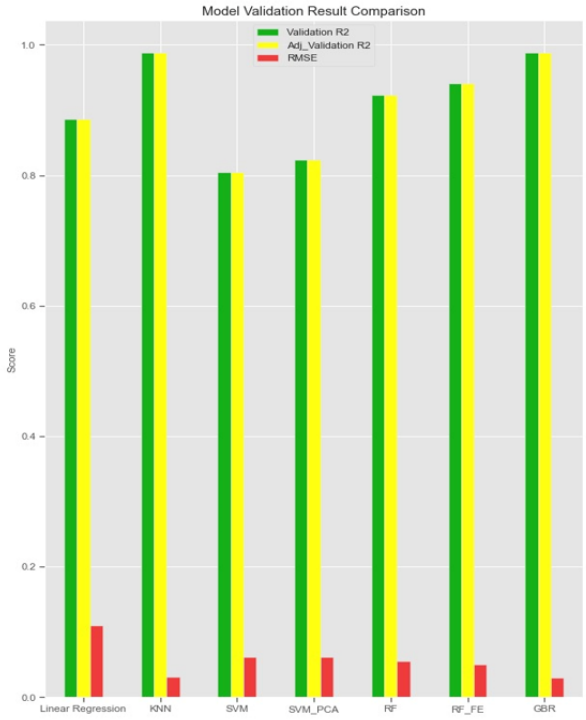


Summary of results of model results for rawdata_new

Effects of Outliers

Similar to the unsupervised clustering, the process was repeated for the rawd_zcore dataset to observe the effects of outliers. The results are shown in Fig F.





Rawdata_new

	Training R2	Adj_Training R2	RMSE
Linear Regression	0.857800	0.857700	0.12100
KNN	1.000000	1.000000	0.00000
SVM	0.802970	0.802920	0.07266
SVM_PCA	0.830440	0.830400	0.07226
RF	0.918480	0.918557	0.05917
RF_FE	0.933810	0.933790	0.05523
GBR	0.986178	0.986173	0.03195

	Validation R2	Adj_Validation R2	RMSE
Linear Regression	0.854500	0.854300	0.12303
KNN	0.984650	0.984620	0.03648
SVM	0.797604	0.797580	0.07463
SVM_PCA	0.828050	0.828020	0.07296
RF	0.920620	0.920520	0.05823
RF_FE	0.933785	0.933670	0.05610
GBR	0.984655	0.984634	0.03300

Rawd_zscore

	Training R2	Adj_Training R2	RMSE
Linear Regression	0.88563	0.888559	0.11024
KNN	1.00000	1.000000	0.00000
SVM	0.80752	0.807500	0.05982
SVM_PCA	0.82680	0.826760	0.05980
RF	0.92371	0.923690	0.05507
RF_FE	0.93897	0.938940	0.05021
GBR	0.98655	0.986530	0.03132

	Validation R2	Adj_Validation R2	RMSE
Linear Regression	0.886330	0.886160	0.11033
KNN	0.987140	0.987120	0.03110
SVM	0.805080	0.805050	0.06153
SVM_PCA	0.823940	0.823900	0.06122
RF	0.923060	0.922950	0.05518
RF_FE	0.940718	0.940599	0.04942
GBR	0.988040	0.988039	0.02939

From the results shown, an overall improvement can be seen in the model performance when outliers were removed as seen in the case of the rawd_zscored dataset. It was more prominent in linear regression, support vector machine and random forest. Also, the result showed a slight overfitting by KNN and SVM models for each dataset while RF slightly underfitted the datasets.

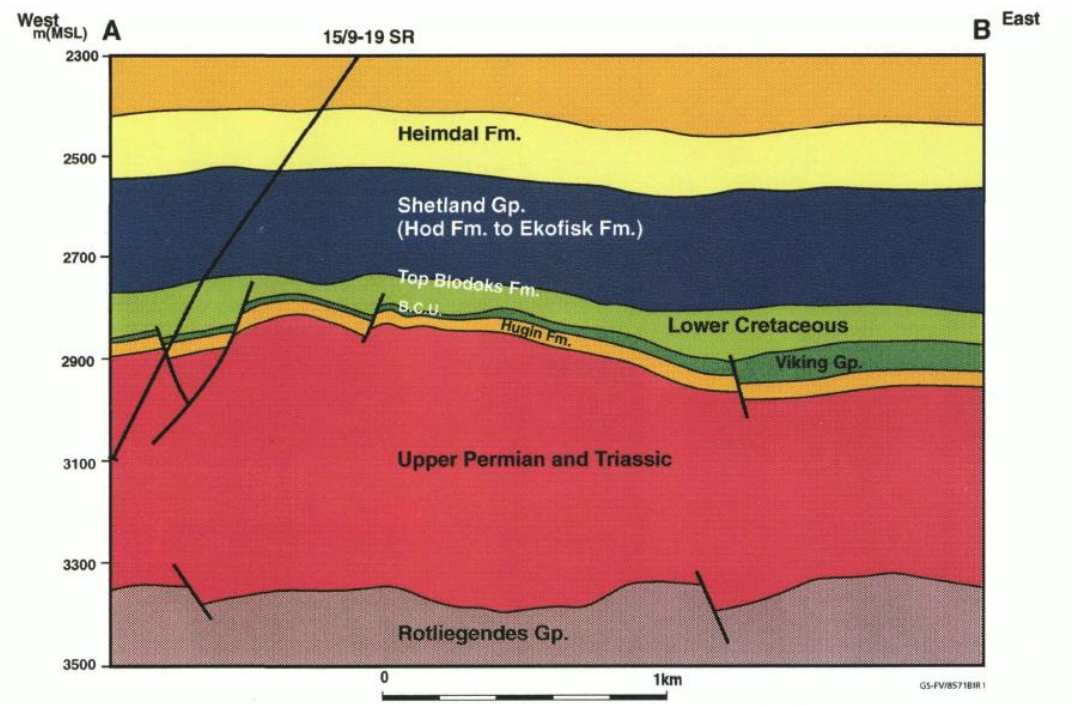
RESULTS VALIDATION

Unsupervised Clustering

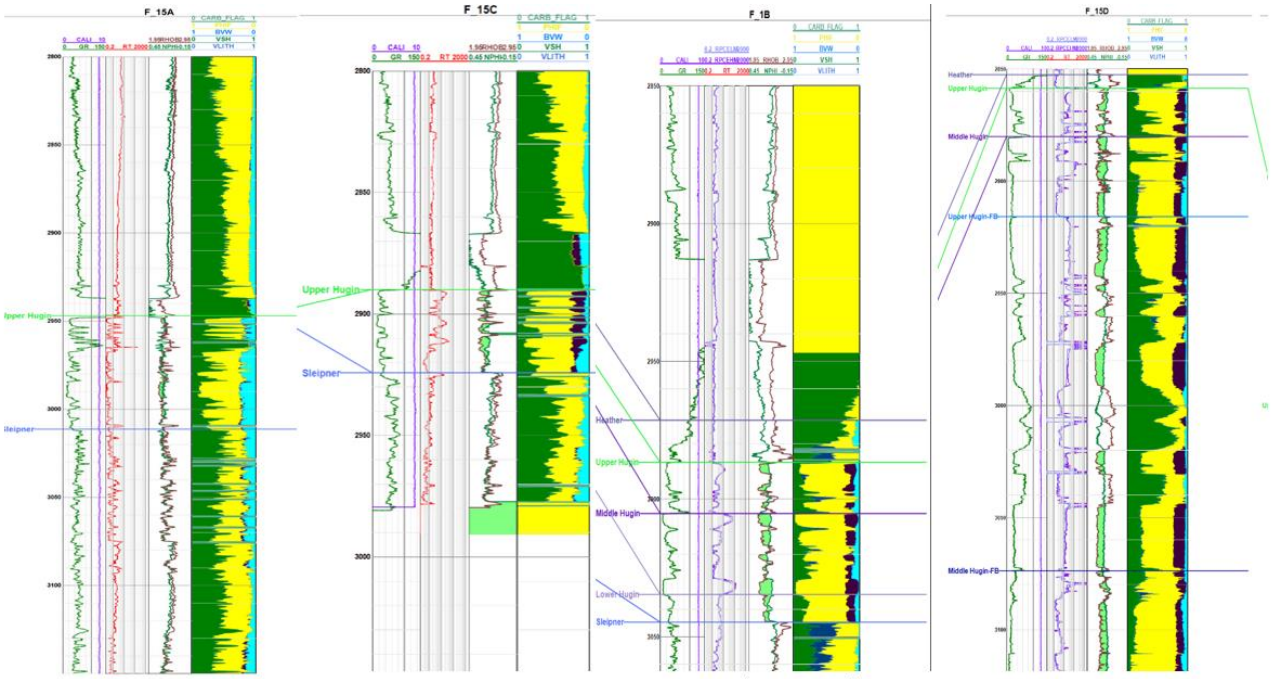
In order to validate the performance of the clustering algorithm, the following approach was employed.

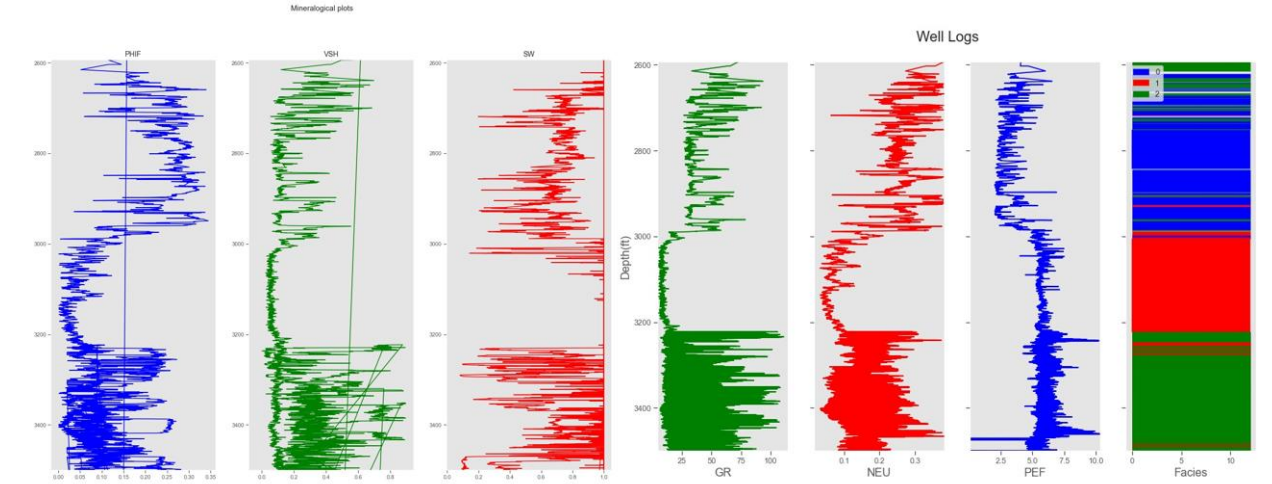
1. The clusters were juxtaposed against the corresponding well logs and explained based on domain knowledge.
2. Depth-matched cross-sectional lithological maps from the Volve Field report was correlated with the clusters to distinguish between the electrofacies in the clusters. This was supported with well log plots and mineralogical plots.

The structural cross-section of the Volve field along well 15/9-19 is shown in Fig D below revealing the formation types at various depths which are predominantly sandstones interspersed with shales (Le Huy et al., 2019; Discover Volve, 2019). To validate the clusters formed by the Kmeans algorithm, clusters were compared to lithological cross-sections for selected wells on the volve field for different directions (i.e. East-West and NorthEast to SouthWest) from the field. These comparisons were made in conjunction mineralogical plots from the data set and estimations were made on the performance of the model.



Structural cross-section along the 15/9-19 SR well path. See Figure 3.12 for location.





Mineralogical plots and KMeans clusters from rawd_zscore data

From the cross-sectional lithological plots, VLITH represents the volume of the specific lithology, every other alias is the same as that presented in the data dictionary in the introduction. The lithological plots shows that the Hugin formation viewed from different cross-sections spans from a depth of about 2880m to 3000m. Right above the Upper Hugin formation is a shaly zone as seen by high shale volume (in green) in almost all the cross-sections. Shaly zones are often associated with high gamma ray readings, low porosities and higher photoelectric factor values compared to sandstones (Doveton, 2004). From the mineralogical plots, depths from about 2600 to about 2800m have high volume of shale (VSH) readings and low porosity values. Water saturation is also very high in this region, and this is characteristic of most shales as they tend to absorb more

water. From the log readings, gamma ray is relatively high in this region. A similar formation is encountered at depths from 3200m to 3400m which has higher photoelectric factor readings. Depths from about 2800m to 3000m has porosities, lower volume of shale lower gamma ray readings, indicative of a sandstone. This is similar to the lithological cross-sections and the top of the Hugin formation which is the main producing sandstone can be found here. The cross-sections show top of Hugin formation to be within this depth range. Cluster 0 can thus be estimated as another sandstone reservoir. The formation at depths from 3000 to 3200m also has similar properties as that seen for cluster 1 but with lower porosity readings. This can also be estimated as a different type of sandstone with lower porosity. The structural cross-section in Fig Q shows that formation at the mentioned depth (3000 to 3200m) is an upper Permian and Triassic sandstone.

Supervised Learning

The performance of the regression models was validated with the validation set using the R^2, adjusted R^2 and RMSE values. The statistics of the predicted values by the best model was compared to that of the training data as an extra step of ensuring that predicted values are within range and represent the same formation.

Table: Statistics of predicted target variables vs training target variables (A) for rawdata_new (B) for rawdata_zscore

(A)

	VSH	PHIF	SW
count	42072.000000	42072.000000	42072.000000
mean	0.275740	0.141716	0.669053
std	0.207018	0.081228	0.354409
min	-0.248000	0.000000	0.013000
25%	0.113200	0.068200	0.300000
50%	0.233000	0.141000	0.813400
75%	0.360625	0.217000	1.000000
max	2.460104	0.403294	1.000000

Statistics of target variables in training data

	VSH_pred_GBR	PHIF_pred_GBR	SW_pred_GBR
count	11189.000000	11189.000000	11189.000000
mean	0.262836	0.136589	0.652931
std	0.192266	0.075282	0.333168
min	0.043442	-0.007586	0.023589
25%	0.136606	0.069408	0.464067
50%	0.198609	0.125974	0.659574
75%	0.342482	0.209208	0.976171
max	2.269940	0.296290	1.216506

(B)

	VSH	PHIF	SW
count	36526.000000	36526.000000	36526.000000
mean	0.253626	0.145851	0.642839
std	0.184674	0.081524	0.357788
min	-0.035820	0.000000	0.043000
25%	0.106000	0.070000	0.275600
50%	0.213300	0.151000	0.737000
75%	0.338175	0.221200	1.000000
max	0.896700	0.344653	1.000000

Statistics of target variables in training data

	VSH_pred_GBR	PHIF_pred_GBR	SW_pred_GBR
count	11189.000000	11189.000000	11189.000000
mean	0.254251	0.142174	0.589021
std	0.150618	0.072917	0.317989
min	0.071524	0.011632	0.077917
25%	0.140485	0.077290	0.290462
50%	0.202924	0.132139	0.619337
75%	0.333919	0.213039	0.900616
max	0.886239	0.289527	1.268572

Statistics of predicted target variable by GBR model

9. Deliverables

In this project, an attempt was made to develop a data-driven model to predict three major petrophysical properties based on well logs obtained from the Volve Field. Below are deliverables based on outcomes and results.

1. A depth-matched plot of clusters (electrofacies) developed with the KMeans unsupervised clustering algorithm using well logs. These clusters may assist paraphysicists and formation evaluation engineers in determining formation lithologies at various depths.
2. A data-driven model developed with a gradient boost regressor algorithm that predicts the petrophysical properties with a substantial level of accuracy and minimal errors based on R-squared and RMSE values.

This model hopes to reduce the cost and cumbersomeness involved in estimation of petrophysical properties in oil and gas formation evaluation and reservoir engineering. The model can be further explored, and its shortcomings improved to be made deplorable in oil and gas business settings.

10. SELF-ASSESSMENT

This was an overall challenging project as it demanded thorough understanding of both Data Science and Analytics and Petroleum engineering concepts. Its technical nature required knowledge in Geology, Petrophysics and reservoir engineering.

Furthermore, almost all key concepts in Data Science and Analytics (DSA) were employed as it involved both unsupervised classification and supervised regression analysis. DSA concepts employed included univariate and bivariate analysis, missing value imputation techniques and outlier handling processes. Furthermore, much learning on cluster analysis was achieved with the use of Sklearn packages for KMeans and Hierarchical clustering, box plots, silhouette analysis, and violin plots was achieved.

In development of the regression model, feature engineering was significantly employed. Here the concept of explained variance in principal component analysis was thoroughly understood and employed. The concept of the variance inflation factor in correlation analysis between predictors was also well understood and used. Selection of important features to improve model performance was also experimented as seen in the random forest model. Major challenges experienced include extracting all relevant data from data sources available and interpreting supplementary data to corroborate thorough understanding of the model performance. Also some challenge was experienced in validating clustering results with cross-sectional lithological plots from field report as this was done in different direction. Unfortunately, the data used in the modelling had no directional component.

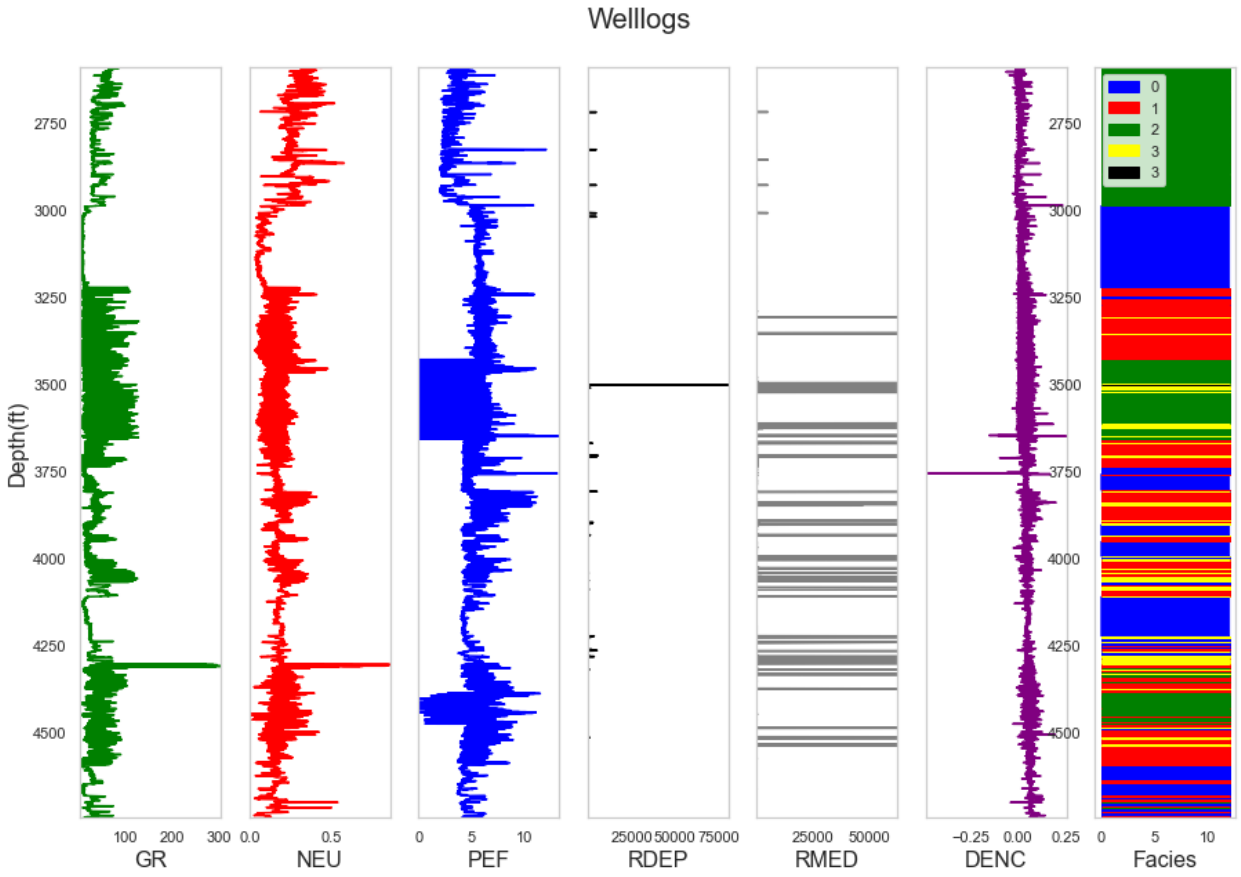
The project was a research project and was supervised by Dr. Matt Beattie. Supervisor details are shown below;

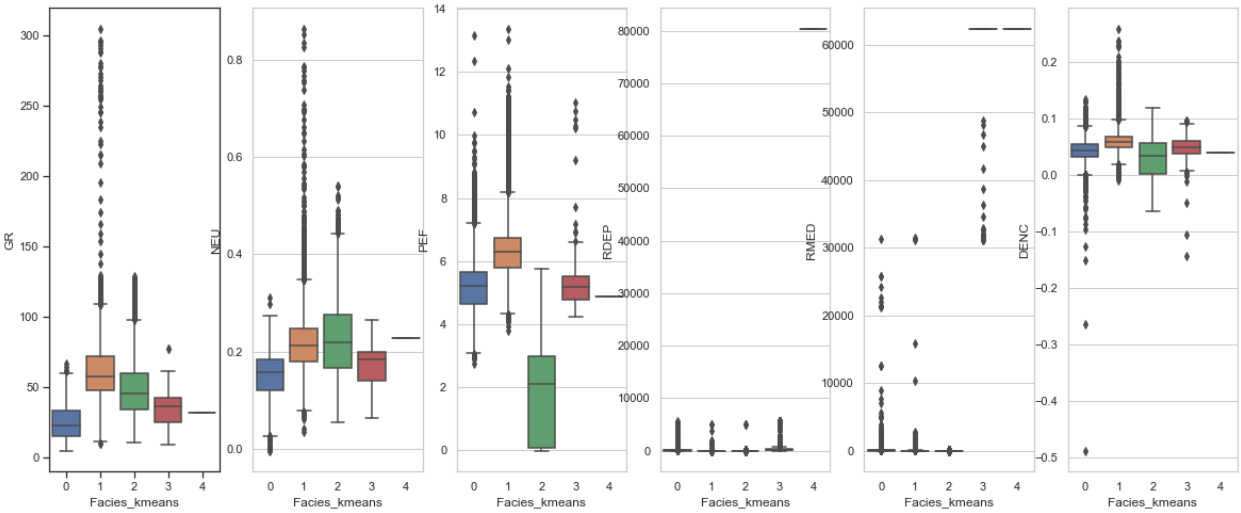
Matthew Beattie
Adjunct Professor
Data Science and Analytics.
The University of Oklahoma
mjbeattie@ou.edu

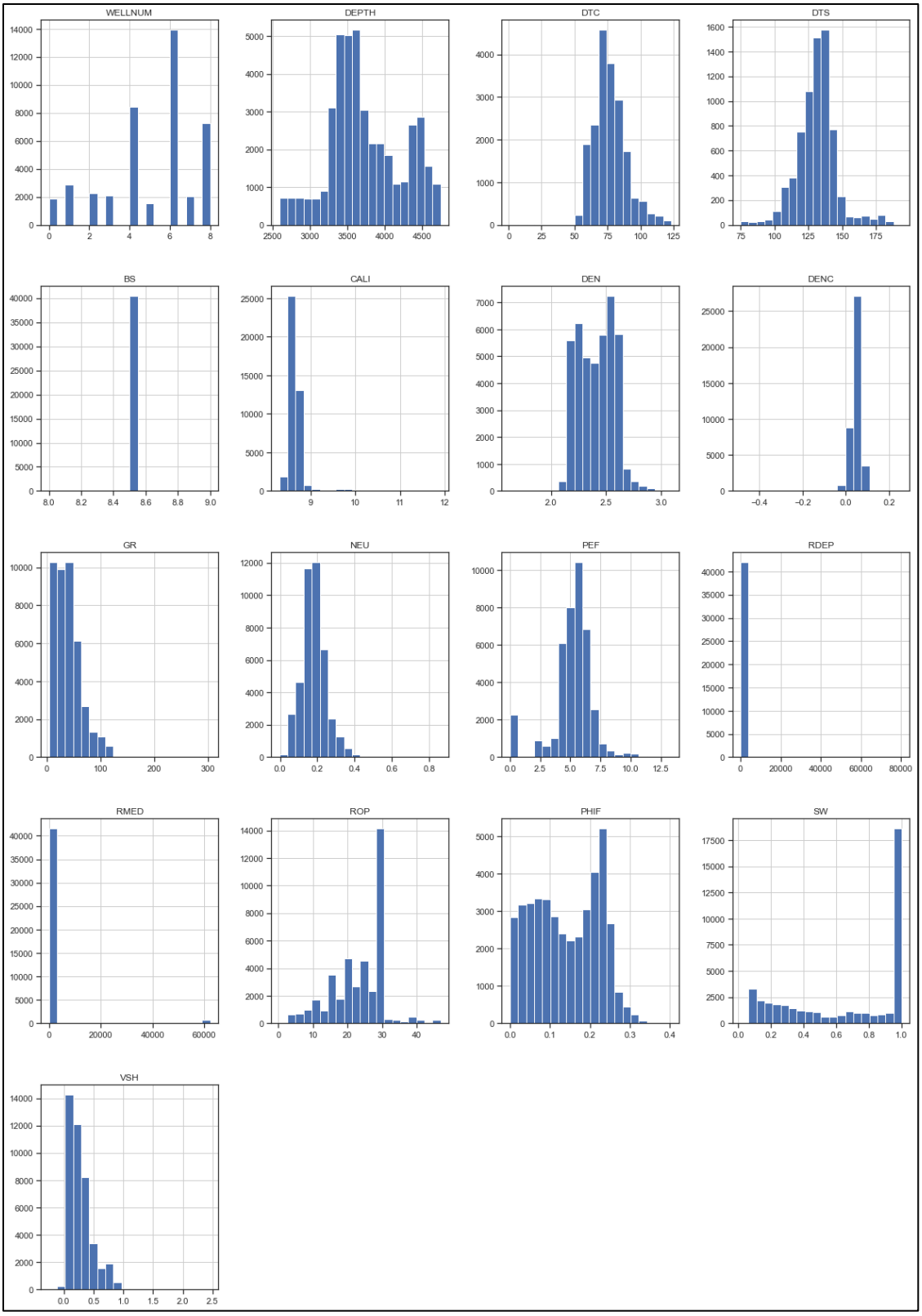
11. References

1. Archie, G. E. (1952). Classification of carbonate reservoir rocks and petrophysical considerations. Aapg Bulletin, 36(2), 278-298.
2. Alberty, M. (1992). Standard interpretation: Part 4. wireline methods.
3. Brown, A. A., & Bowers, B. (1958). Porosity determinations from neutron logs. The Petroleum Engineer, 30.
4. Tenchov, G. G. (2016). Porosity evaluation from acoustic log using the theory of mixtures.
5. Moradi, S., Moeini, M., Al-Askari, M. K. G., & Mahvelati, E. H. (2016, October). Determination of shale volume and distribution patterns and effective porosity from well log data based on cross-plot approach for a shaly carbonate gas reservoir. In IOP Conference Series: Earth and Environmental Science (Vol. 44, No. 4, p. 042002). IOP Publishing.
6. Hearst, Joseph R., and Philip H. Nelson. "Well logging for physical properties." (1985).
7. Mondol, Nazmul Haque. "Well logging: Principles, applications, and uncertainties." In Petroleum Geoscience, pp. 385-425. Springer, Berlin, Heidelberg, 2015.
8. Moore, William R., Y. Zee Ma, Jim Urdea, and Tom Bratton. "Uncertainty analysis in well-log and petrophysical interpretations." (2011): 17-28.
9. Wang, Jun, Junxing Cao, Jiachun You, Ming Cheng, and Peng Zhou. "A method for well log data generation based on a spatio-temporal neural network." Journal of Geophysics and Engineering 18, no. 5 (2021): 700-711.
10. Equinor Volve Data set; https://www.equinor.com/energy/volve-data-sharing
11. American Association of Petroleum Geologists.(2022). Density-Neutron Log Porosity. https://wiki.aapg.org/Density-neutron_log_porosity
12. Sen, S., & Ganguli, S. S. (2019, April). Estimation of pore pressure and fracture gradient in Volve field, Norwegian North Sea. In SPE oil and gas india conference and exhibition. OnePetro.
13. Gupta, I., Tran, N., Devegowda, D., Jayaram, V., Rai, C., Sondergeld, C., & Karami, H. (2020). Looking ahead of the bit using surface drilling and petrophysical data: Machine-learning-based real-time geosteering in Volve field. SPE Journal, 25(02), 990-1006.
14. Le Huy, K. P. (2019). Source-Rock, Reservoir and Hydrocarbon Mapping Using Far-Offset, Elastic Impedance and Extended Elastic Impedance Seismic Volumes, Volve Field, Offshore Norway, North Sea. *Bulletin of Earth Sciences of Thailand*, *11*(2), 1-12.
15. Discover Volve. 2019. https://discovervolve.com/2019/07/18/the-seismic-interpretation-and-the-formations/
16. Doveton, J. H., & Merriam, D. F. (2004). Borehole petrophysical chemostratigraphy of Pennsylvanian black shales in the Kansas subsurface. *Chemical geology*, *206*(3-4), 249-258.

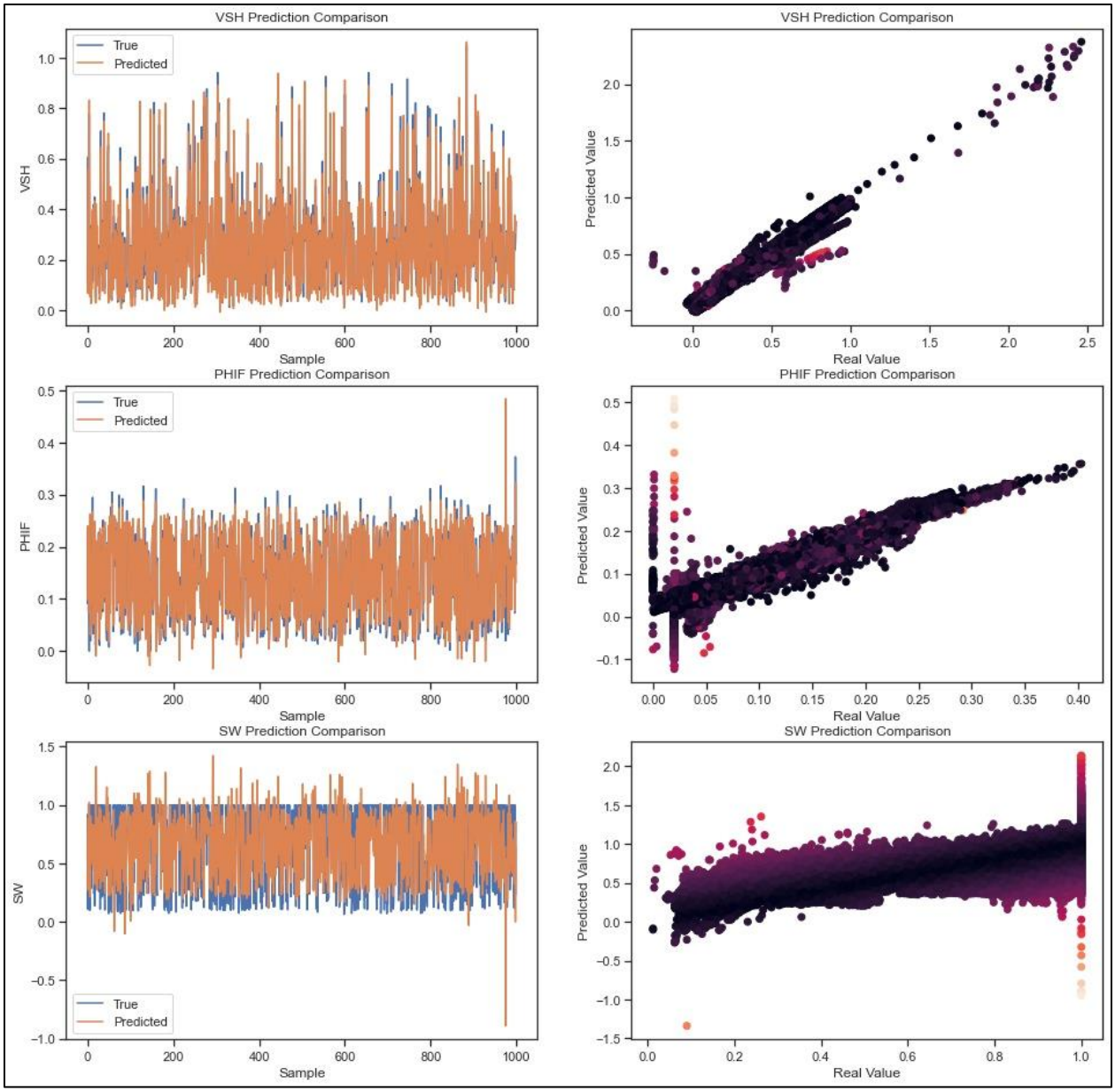
APPENDIX


Welllogs
Depth(ft)
GR
NEU
PEF
RDEP
RMED
DENC
Facies
0
1
2
3
3

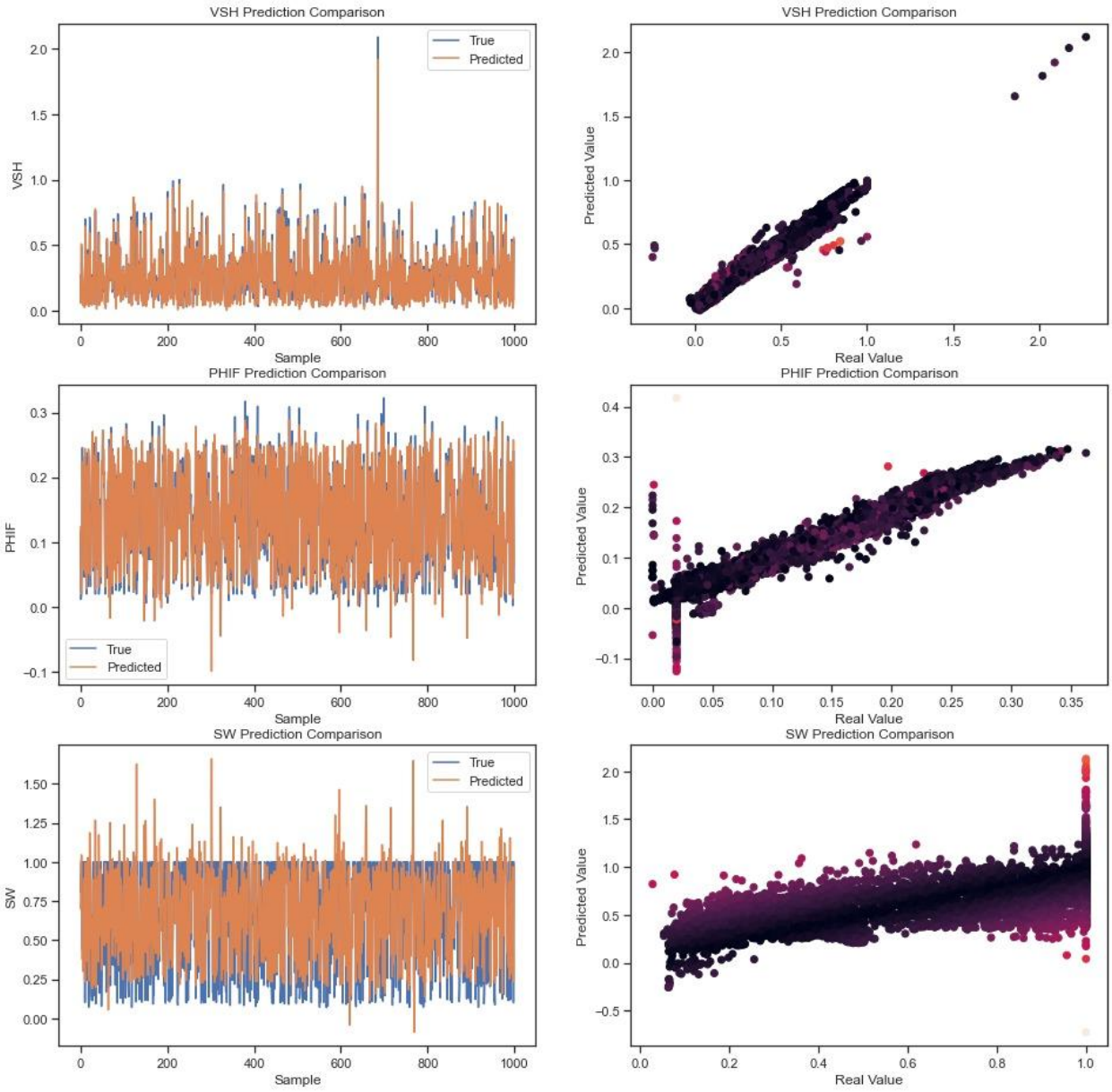

GR
NEU
PEF
RDEP
RMED
DENC
Facies_kmeans



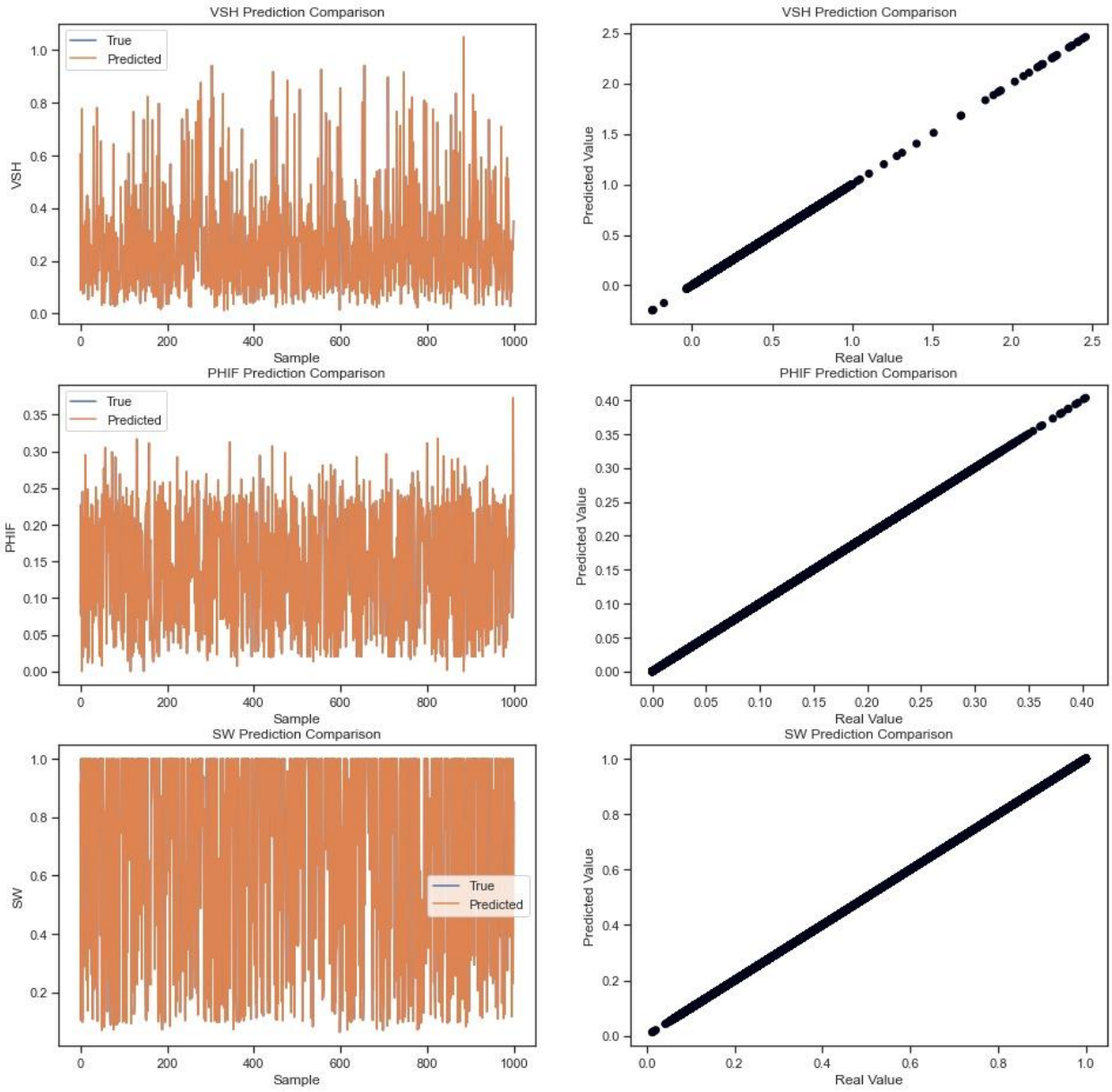
Histogram of features in rawdata1



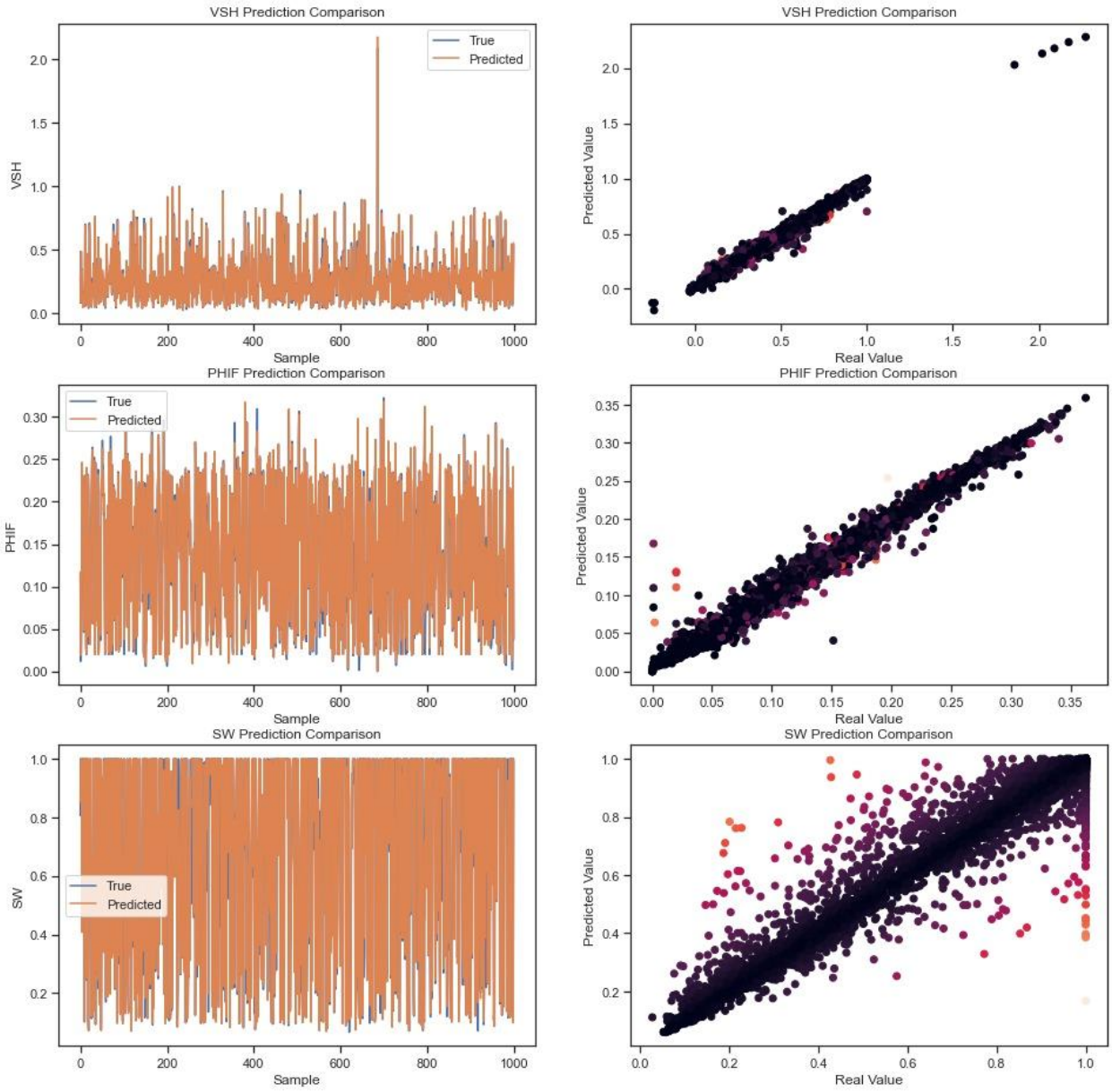
Linear regression training results for rawdata_new



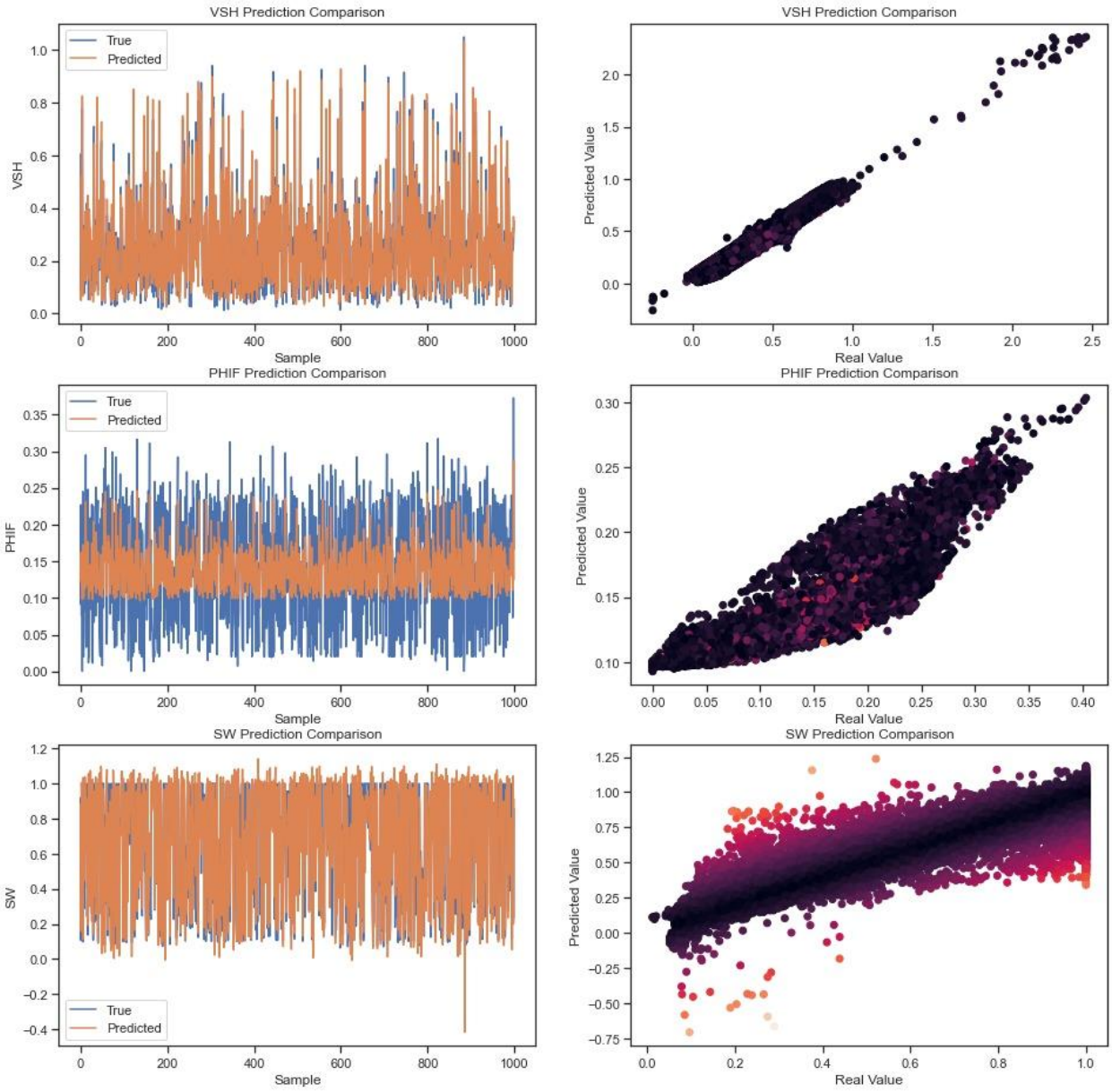
Linear regression validation results for rawdata_new



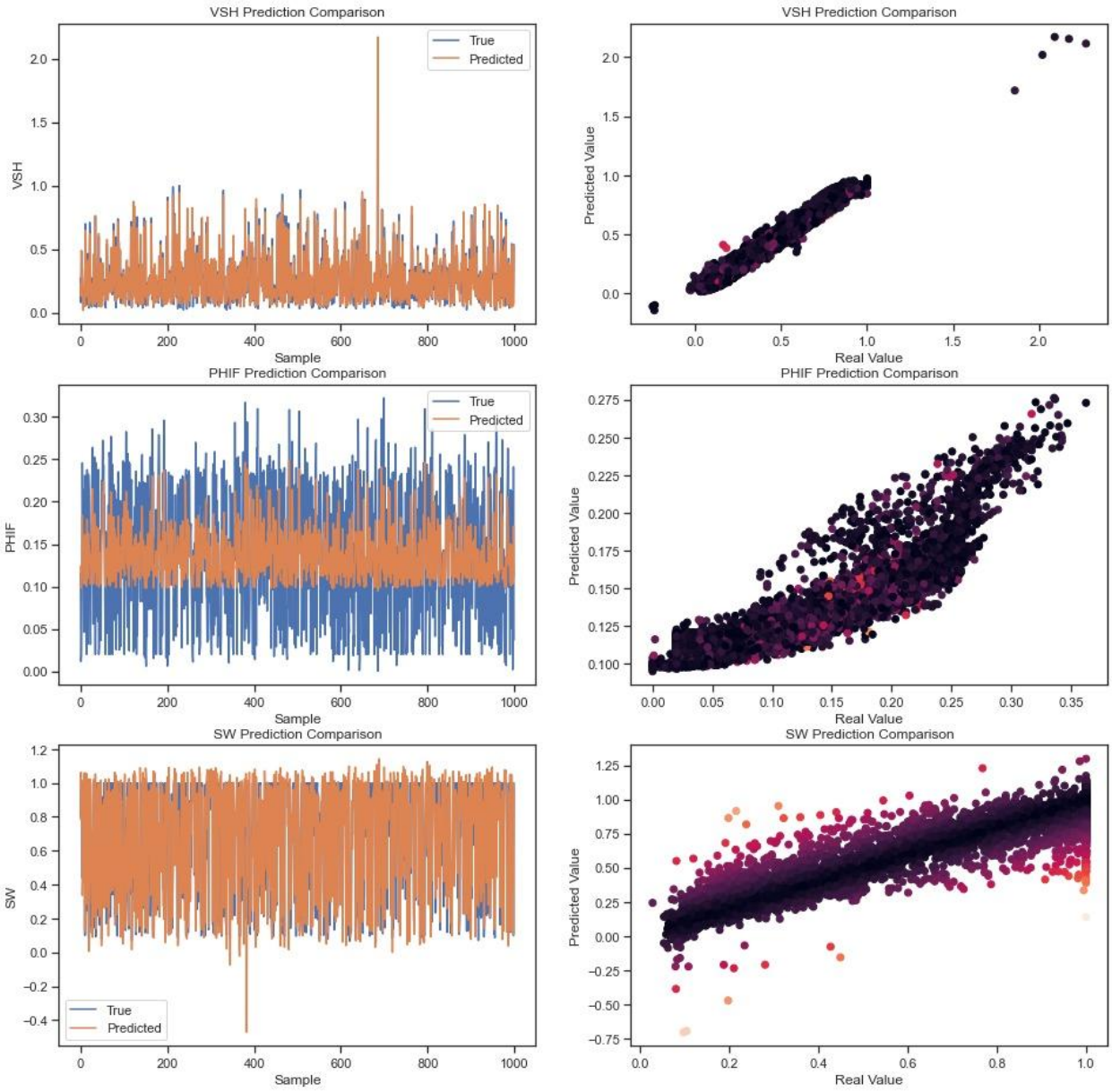
KNN Training results for rawdata_new



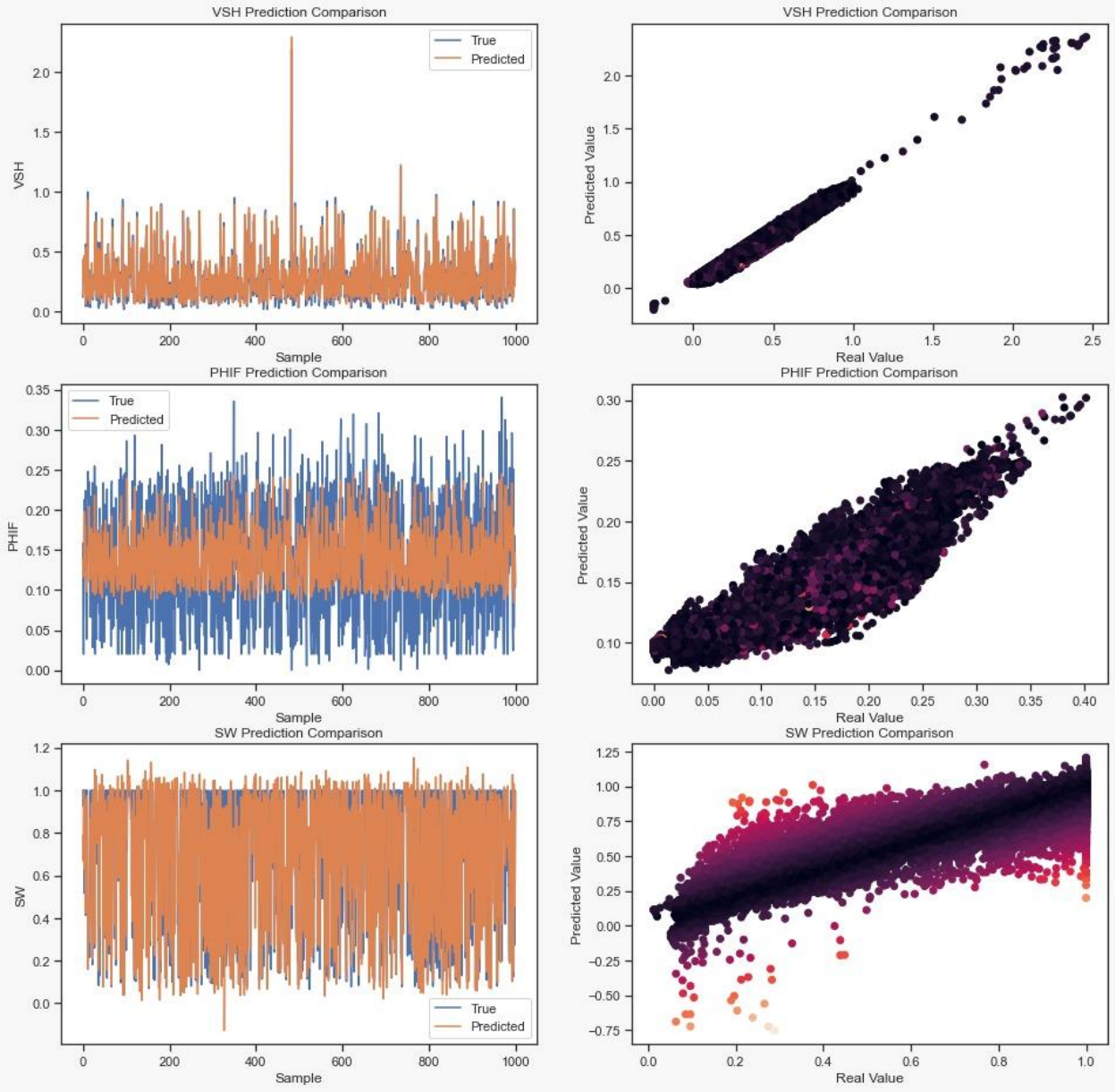
KNN validation results for rawdata_new



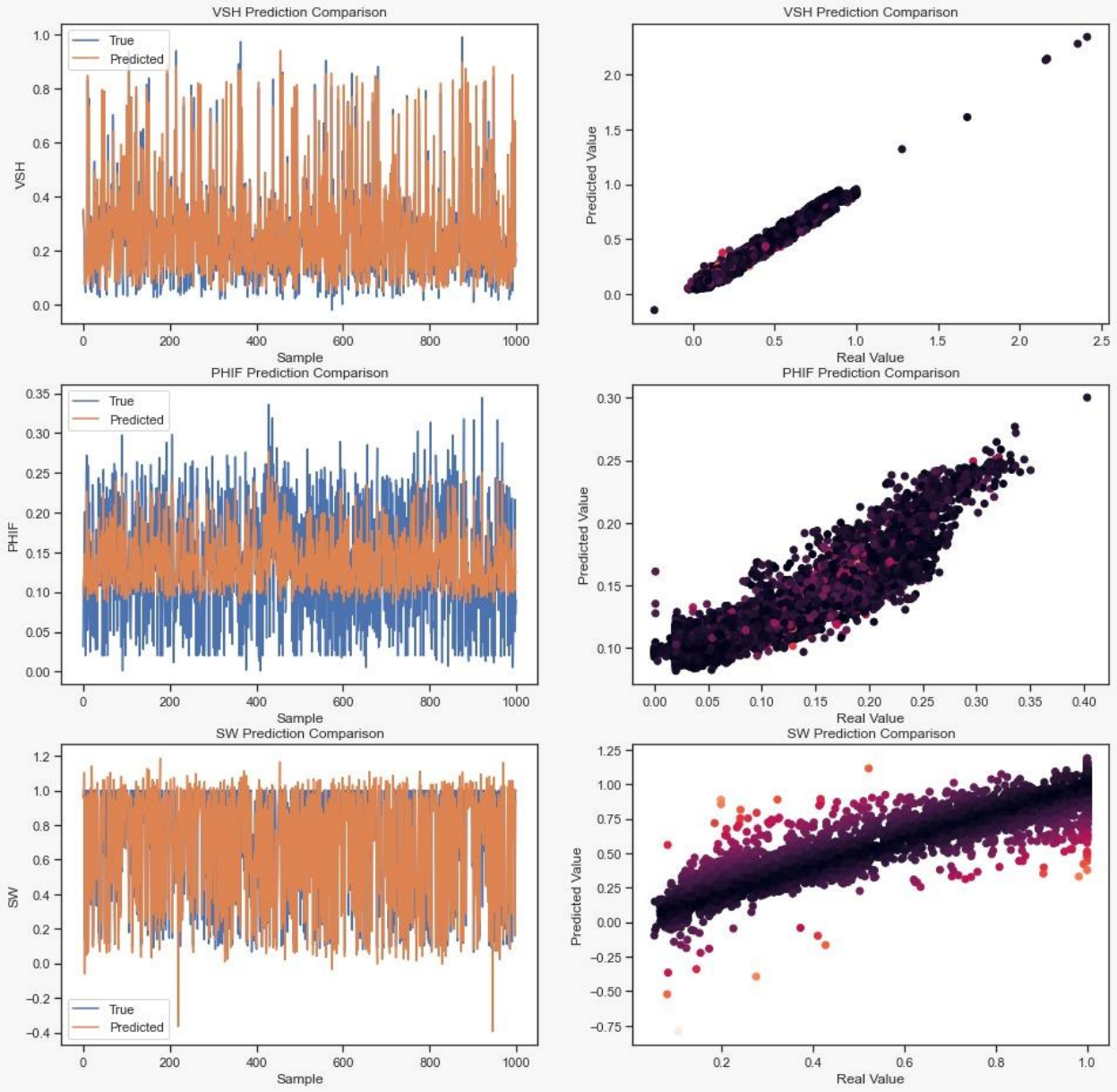
Support Vector machine training results for rawdata_new



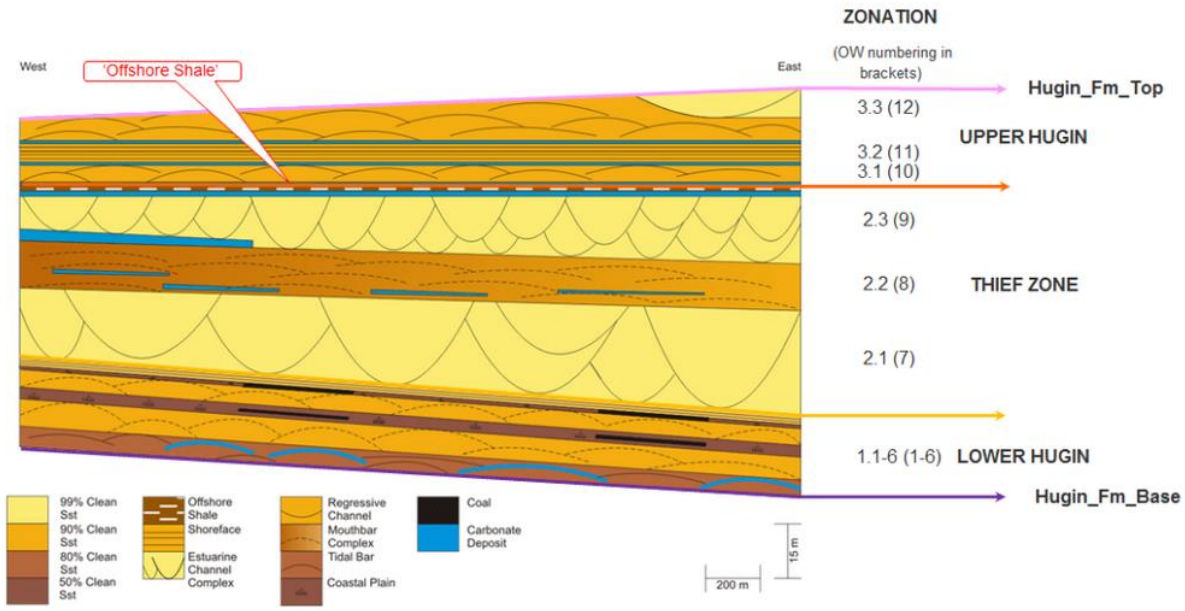
Support vector machine validation results for rawdata_new



Support Vector Machine with PCA training results for rawdata_new



Support vector machine with PCA validation result for rawdata_new



Hugin reservoir stratigraphy