Evaluating feature selection techniques using gradient boosting regression model for reservoir characterisation.  
  
Daniel Asante Otcherea, b,∗, Tarek Omar Arbi Ganat a, Jude Ojeroc, Mohamed Takid  
a Department of Petroleum Engineering, Universiti Teknologi PETRONAS, 32610, Seri Iskandar, Perak Darul Ridzuan, Malaysia

b Department of Petroleum Engineering, Curtin University, CDT 250, Miri Sarawak, Malaysia

Abstract

Feature Selection, a critical data preprocessing step in machine learning, is an effective way in removing irrelevant variables thus reducing the dimensionality of input features. In this paper, we propose a custom ensemble feature selection algorithm based on lasso regularisation and random forest to handle high dimensional wireline log data. Statistical comparison based on prediction errors and computational efficiency with eight existing feature selection techniques was done using Volve well log data sets.

Analysis of the results shows that the proposed custom ensemble is very effective and efficient in selecting the relevant well logs for predicting permeability, porosity and water saturation. This proves that the xxx, xxx and xxx can be used in identifying the best input features for permeability, porosity and water saturation predictions respectively.

This paper presents an ensemble feature selection method to identify the most informative spectral features for practical applications in plant phenotyping. To rank spectral features, six feature selection methods were used as the base for the ensemble: correlation-based feature selection, ReliefF, sequential feature selection, support vector machine-recursive feature elimination (SVM-RFE), LASSO logistic regression, and random forest. The best results were achieved by the ensemble of ReliefF, SVM-RFE, and random forest, which drastically reduced the dimension of the hyperspectral data set from 215 to 15 features, while improving the accuracy in classifying the salt-treated vegetation pixels from the control pixels by 8.5%. The result of salt tolerance assessment of the four wheat lines using the derived multispectral data set was similar to that of the hyperspectral data set.

Keyword: **Feature Selection Techniques; Dimensionality Reduction Techniques; Artificial Intelligence; Ensemble Machine Learning; Decision Tree Algorithm**

# Introduction

Reservoir characterisation based on petrophysical properties is of key interest in the industry as economic decisions as to the viability of a reservoir highly depends on its accurate predictions. For this reason, companies spend huge amounts of money and time and gathering data to aid in improved estimation of these properties. With the abundance of well log data, reservoir characterisation using supervised machine learning techniques has gained a lot of popularity. Supervised machine learning has chalked up huge successes over the years in the petroleum industry but any increase in accuracy and time efficiency will have enormous contributions. Supervised machine learning techniques, however, achieves better accuracy and time efficiency with relevant input variables. In the case of reservoir characterisation, not all well logs are relevant and incorporating them as input variables leads to a situation called Garbage In Garbage Out (GIGO) (Rose and Fischer, 2011). To aid in mitigating the effect of GIGO, input variables needs to be reduced to eliminate the variables that does not have any correlation to the target or dependent variable. There are various techniques used in reducing the input variables and selecting only relevant logs for prediction. This is termed as dimensionality reduction techniques which is based on the correlation amongst high dimensional data to extract only the meaningful low-dimensional sets of data without losing relevant information (De Silva and Tenenbaum, 2003). This step is one of the most important preprocessing steps in the application of machine learning algorithms (Blessie and Karthikeyan, 2012; Kou et al., 2020).

Before dimensionality reduction technique is applied to input variables, an in depth understanding and knowledge of how the input logs relate to the target is relevant. Studies that involves a large variety of data points, like reservoir characterisation, offers an excess of data that requires statistical techniques to evaluate the correlation amongst data leading to a less subjective process [Telnæs, N., Dahl, B]. Statistical analysis of these input logs are also key in selecting relevant logs that have strong association with the target. Feature selection of relevant logs tends to have a significant impact on the performance of supervised machine learning models as it reduces dimensionality and improves the computational time efficiency of supervised machine learning models (Laib and Kanevski, 2019; Wei et al., 2020). This helps prevent multicollinearity and overfitting that results from highly dimensional and complex data, improves readability and interpretability of the data (Moghimi et al., 2018; Wei et al., 2020). Currently, the use of unsupervised machine learning techniques has been used to establish correlation amongst multivariable datasets for dimensionality reduction. These techniques offers some form of accuracy in feature selection but does not work best with highly complex and nonlinear datasets as each suffers from their inherent limitations. Correlation, however, does not imply causation as different datasets has a significant impact on the types of correlation methods to use. There are therefore no systematic rules or methodologies in the application of correlation techniques to determine the strength of association.

Regrettably, several applications of statistical methods used in feature selection tends to fail in selecting relevant logs for the prediction of several reservoir petrophysical properties. The main aims of this study are to;

1. Understand the relationship amongst key wireline logs relevant for accurate reservoir characterization,
2. Illustrate the performances of the different feature selection techniques used in analyzing high dimensional data
3. Identify the best feature selection techniques that attempts to capture inherent geological knowledge and improve reservoir characterisation accuracy.

Once this done, the optimal input features will be used to train a machine learning model to predict different petrophysical properties of reservoir to test the robustness of the model using the differently selected features. The key contribution of this research is to identify key wireline logs that can improve the prediction accuracy and minimize the prediction errors of machine learning models for several reservoir characterisation.

# Brief Background of Dimensionality Reduction Techniques

There are two main type of dimensionality reduction techniques namely feature selection and dimensionality reduction. Although both feature selection and dimensionality reduction performs the same functions, the main differences are that feature selection only identifies relevant features but keeps the variables in its original form whereas dimensionality reduction transforms data by combining variables (Bolon-Candeo 2019). The three main classes of feature selection techniques to be used in this study are filter, wrapper and embedded methods. These will be compared to other feature selection methods. With the data being continuous, methods that works well with continuous variables will be applied in this study.

## Feature Selection

### Filter method

This method only filters out relevant features with high correlation to the target using correlation matrixes like Pearson's r, Spearman's rho (rs) and Kendall's Tau (τ). The filter model selects relevant subsets of the input variables by measuring their general characteristics based on distance, dependency and consistency (Blessie and Karthikeyan, 2012). The Pearson correlation is the most popular filter method that uses parametric test to measure the linear relationship between two variables. It works best with continuous data that has a normal distribution function. The Spearman and Kendall correlation methods are better suited for non-normally distributed data and uses non-parametric tests (Rock, 1987). Spearman correlation tends to measure the degree of correlation between two variables whereas Kendall correlation, an extension of Spearman, measures the strength of dependence between two features (Akoglu, 2018). There has been various suggestion that Kendall correlation gives a more accurate generalization in data as compared to Spearman features (Akoglu, 2018). For filter methods, the relationship between variables are measured between 1 (direct relationship) and -1 (inverse relationship) with 0 indicating of no correlation.

### Wrapper method

The wrapper method is an iterative process that employs a machine learning model, evaluates the performance of the model with different subsets of the input variables and selects the high-ranking variables that results in the best performing model. Although it is a computationally expensive method and prone to overfit (Wei et al., 2020) (Rao et al, 2019), it tends to outperform the filter method (Rao et al, 2019). There are various types of wrapper methods but only three will be used in this study.

1. Backward feature elimination (BFE): This an iterative process that begins with all the input variables then removes the worst performing ones. The performance metric used to evaluate the performance of the variables is p-value. The highest performing variables are then selected for model training. This is done until the predetermined number of features are all exhausted.
2. Recursive feature elimination (RFE): This method performs a rigorous search for the relevant features by iteratively creating models and evaluating their performances. The worst performing variables are recursively removed using the remaining variables to create the models. The features are then ranked using an accuracy metric according to their order of elimination (Ebrahimy and Azadbakht, 2019). It reports its evaluation using hierarchical ranking, true for relevant variables and false for irrelevant variables.

### Embedded method

Embedded methods are iterative processes that perform feature selection as part of the model training (Rao et al, 2019). The result is a model that carefully emphasizes only the most relevant features that contributes the most to the model training process. The most used embedded methods are the regularisation methods that penalizes features based on a given coefficient threshold. There are various regularisation methods, but this study applies the Linear Regression with L1 regularization (Lasso Regularization) (Boyd and Vandenberghe, 2004) for feature selection. For lasso regularisation, if the input data is irrelevant, its coefficient gets penalized, made 0 and removed leaving only the relevant features. Therefore, regularization (penalization) methods introduces supplementary constrictions into optimising predictive algorithms allowing the selection of fewer coefficients nudging (Cite).

### Intrinsic method

Decision tree ensembles, in this study the Random Forest, are useful in the automatic selection of relevant features. The Random Forest method is one of the most commonly used algorithms that results in good model performances, robust against overfitting and offers easy interpretability (Ebrahimy and Azadbakht, 2019). The ease of interpretability is because it estimates the importance of each input feature on the decision tree. Aside the many advantages this method provides, it tends to assign equal importance to correlated features. This leads to an overall reduction in importance for the features built on the same tree as compared to uncorrelated features. The model also tends to be biased towards features with high cardinality.

### Univariate feature selection method

The univariate feature selection method works by assessing the strength of the relationship of each input data to the target based on univariate statistical tests. The SelectKBest combines p-values and F-test scores to evaluate the performance of the parameters and selects the relevant features. The F-test adds an additional estimate of the amount of linear dependency between two random variables. This makes it statistically robust than the wrapper method which selects features only based on the p-value. This is empirically stated as:

# Background of Supervised Machine Learning Models Used

## Gradient Boosting Regressor (GBR)

The Gradient Boosting Regressor (GBR) is another ensemble model that is an iterative collection of sequentially arranged tree models so as the next model learns from the error of the former model. This is a machine learning model that makes predictions using "boosting" of the ensemble of weak prediction models often decision trees to form a stronger model (Rao et al, 2019). A GBR with M number of trees can be stated as;

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

Where hm is a weak learner that performs badly individually, γm is a scaling factor adding the contribution of a tree to the model. GBR uses the gradient descent loss function to minimize errors by updating the initial estimation with the new estimation. Thus, a final model is created with the combination of all preliminary estimations with the suitable weights. The GBR model implemented in this study is from the GradientBoostingRegressor() method provided in scikit-learn (Pedregosa et al., 2011).

# Criteria for Model Evaluation

In model regression analysis, the difference between the actual data points and the best fit line produced by the algorithm is the model’s error. With multiple data points, the error of the model will be determined using the following criteria;

1. Mean Absolute Error (MAE): This represents the mean of the absolute value of the errors that indicates the deviation from true probability. This is mathematically expressed as;

|  |  |  |
| --- | --- | --- |
|  |  | (2) |

1. Execution time: This is the time that it takes the model to learn and make predictions from the input logs.
2. Root Mean Squared Error (RMSE): RMSE is also a popular performance evaluation metric for models because it is interpretable as the standard deviation of the prediction errors. This is written as;

|  |  |  |
| --- | --- | --- |
|  |  | (3) |

# Materials and Methodology

The core objective of this research is to characterise sandstone reservoirs using wireline logs in the Volve field. The selected variables for reservoir characterisation are as follows; CALI, DRHO, DT, GR, NPHI, PEF, RACEHM, RACELM, RD, RHOB, ROP and RT whereas the targeted output are horizontal permeability (KLOGH), porosity (PHIF) and water saturation (Sw). For this study, one well with 2793 data points was used to train and test all the feature selection techniques. Before using machine learning algorithms, a systematic workflow was developed to help identify appropriate variables for each target output to help prevent the issue of “GIGO”. For this process to be done, the input data needs to be understood as table 1 and 2 summarizes the descriptive statistics of all the input and output variables respectively. The covariance matrix was calculated to quantify the degree of correlation amongst features (figure 1). This analysis helps to determine the data distribution and their expected behaviour (Akande et al., 2015).

Table 1 Descriptive statistics of all input data

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | CALI | | DRHO | | DT | GR | | | NPHI | PEF | | RACEHM | | RACELM | | RD | RHOB | | | ROP | RT |
| count | | 2793 | 2793 | 2793 | | | 2793 | 2793 | | | 2793 | 2793 | 2793 | | 2793 | | | 2793 | 2793 | | 2793 | |
| mean | | 8.63 | 0.06 | 75.60 | | | 62.54 | 0.16 | | | 6.06 | 1.48 | 1.42 | | 1.53 | | | 2.52 | 20.01 | | 1.65 | |
| std | | 0.03 | 0.01 | 4.46 | | | 25.89 | 0.04 | | | 0.43 | 0.54 | 0.46 | | 0.58 | | | 0.08 | 3.17 | | 0.96 | |
| min | | 8.55 | 0.01 | 60.22 | | | 17.73 | 0.06 | | | 4.62 | 0.42 | 0.45 | | 0.41 | | | 2.29 | 10.00 | | 0.37 | |
| 0.25 | | 8.63 | 0.05 | 72.58 | | | 44.20 | 0.13 | | | 5.77 | 1.10 | 1.06 | | 1.14 | | | 2.47 | 18.49 | | 1.14 | |
| 0.50 | | 8.63 | 0.06 | 74.97 | | | 52.48 | 0.15 | | | 6.05 | 1.52 | 1.46 | | 1.59 | | | 2.53 | 19.98 | | 1.63 | |
| 0.75 | | 8.65 | 0.06 | 77.93 | | | 78.75 | 0.19 | | | 6.36 | 1.85 | 1.77 | | 1.90 | | | 2.59 | 22.56 | | 2.01 | |
| max | | 8.73 | 0.12 | 90.98 | | | 129.28 | 0.28 | | | 8.36 | 4.70 | 2.41 | | 5.70 | | | 2.75 | 25.40 | | 19.62 | |

Table 2 Descriptive statistics of output data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | KLOGH | PHIF | SW | VSH |
| Count | 2793.00 | 2793.00 | 2793.00 | 2793.00 |
| mean | 5.11 | 0.08 | 0.99 | 0.40 |
| std | 48.90 | 0.05 | 0.06 | 0.22 |
| min | 0.00 | 0.02 | 0.33 | 0.03 |
| 0.25 | 0.00 | 0.04 | 1.00 | 0.25 |
| 0.50 | 0.04 | 0.08 | 1.00 | 0.32 |
| 0.75 | 0.39 | 0.11 | 1.00 | 0.53 |
| max | 1332.59 | 0.21 | 1.00 | 1.00 |

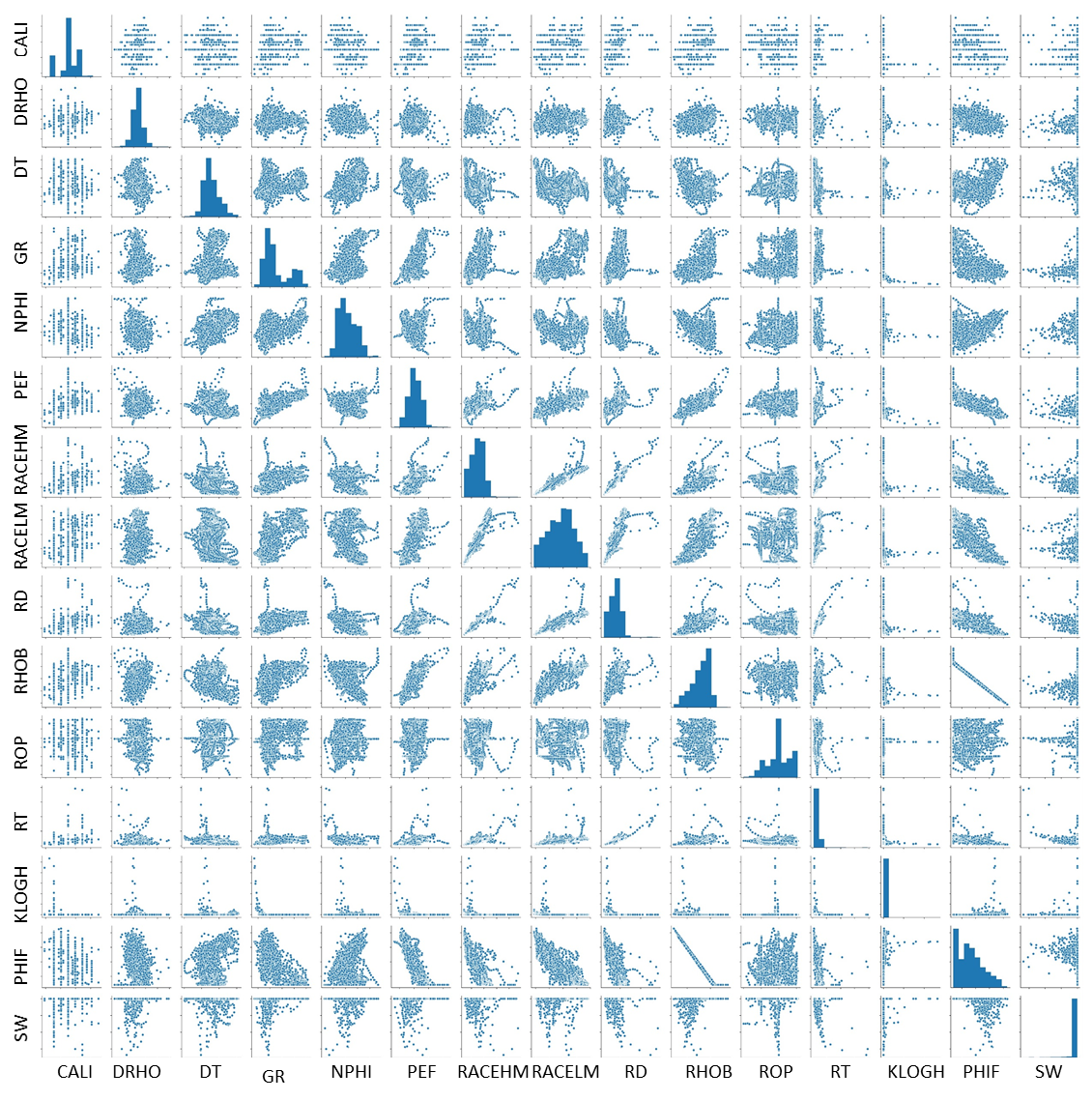


Figure 1 Correlation matrix of input and output variables.

## Application of feature selection techniques

By using Jupyter notebook python programming language, all input variables were used to predict porosity, permeability and water saturation using the Gradient Boosting Regressor (GBR) machine learning algorithm. The input variables were then correlated to the response variable using the Pearson, Spearman and Kendall correlations to show their relationships and covariance. Although Pearson correlation is most commonly used due to its preference for indicating associations between linearly related variables, the latter two correlations are much preferred when data is not normally distributed and shows some non-linear associations within wide ranges (Akoglu, 2018). The three correlations with values closer to +1 showing positive relationship, -1 indicating inverse relations and 0 showing no relationship. Features showing a relationship above a threshold value of 0.5 were selected after which feature-feature correlation was done to remove features with correlation above 0.8. This is an easy way in detecting multicollinearity. The resulting features were used to make further predictions using GBR.

The wrapper methods were also applied by selecting the input and output variables. An iterative process was applied to select the relevant features that fell below the P-value cut-off of 0.05 for BFE. P-value indicates the probability value that the correlation between variables is statistically significant (Akoglu, 2018). A significant threshold level of 0.05 shows 95% confidence that the correlation between the variables is significant. Conventionally;

1. p-value < 0.001: shows strong evidence of significant correlation.
2. p-value < 0.05: shows moderate evidence that the correlation is significant.
3. p-value < 0.1: indicates weak evidence that the correlation is significant.
4. p-value > 0.1: shows no evidence of significant correlation.

For the BFE technique, the Ordinary Least Squares (OLS) model which is used for performing linear regression was employed in the iteration process. Random feature selection was used to choose features for RFE by splitting input data into train and test data. Based on the accuracy metric used to rank feature according to their importance, a variable rank of 1 indicates it is most important. It also provides its support with True being relevant and False being irrelevant features. An iterative loop was subsequently applied to find the optimum number of features that will achieve the highest accuracy for all reservoir properties.

The embedded method was also applied for feature selection using Lasso regularisation. A default penalisation (c) of 3 was used bearing in mind that increasing c increases the number of features removed whiles a low c tends not to remove irrelevant features. This iterative method using the Lasso best score eliminates irrelevant features after penalising their coefficients. The Random Forest (RF) intrinsic method was used to select relevant features for model training using ten estimators, one minimum sample leaf and two minimum sample splits. The MSE error criterion was used. The SelectKBest was defined using the f\_regression function, then the train and test features (70:30) were transformed to select relevant features for the target outputs. All input variables were assigned a score with the top 5 scoring features selected for permeability. For porosity, all the features that scored above 0 were selected whiles the top 8 scoring features for water saturation were chosen for model training.

## Application of machine learning model

The GBR model was used to estimate permeability, porosity and water saturation from all the input logs. After applying all of the feature selection techniques, the relevant logs selected by each method were subsequently used to predict each target. The holdout cross validation technique of 70:30 training and testing was used. Bear in mind that the main idea is not the model accuracy but the prediction error. Analysing the errors helps in understanding the biasness and variance of the model. The error rate is usually referred to as the bias with the most influencing parameter being the selection of the input data. The variance of a model refers to the deterioration of the performance accuracy of the model on test data compared to the training data. A proper understanding of this is determined by the feature selection techniques employed in this study to select relevant logs for reservoir characterisation. The model parameters used are Talk of the model settings. The results of all the feature selection techniques in terms of error and computational time were then compared to the baseline model that used all the input parameters. Due to the stochastic nature of the GBR model or differences in numerical accuracy, the permeability and water saturation results may vary. Hence, they were run ten times with the average of the outcome selected for use in this study.

# Results and Discussion

The various techniques applied for feature selection of relevant logs to predict permeability, porosity and water saturation needs to agree with the domain knowledge. Usually machine learning algorithms are applied to reservoir characterisation without proper domain analysis of the target. Researchers tend to subjectively decide the logs to use as input features. After, they apply various algorithms that helps reduce dimensionality of input data with some techniques selecting different combinations of input logs to predict the same output for different wells. Understanding the geological relationship between the input wireline logs and the output helps in knowing the best techniques to apply in machine learning. Selecting too few or many input features causes problems for machine learning models. Selecting variables with high feature-feature correlations leads to the occurrence of multicollinearity whereas subjectively omitting a relevant input variable results in biasness. There is therefore the need for a comprehensive evaluation of feature selection techniques that can provide a suitable balance in too many or few relevant feature selections for machine learning models. A summary of the uses of each input wireline log in relation to the targets are detailed in table 3.

Table 3 Uses of wireline logs used in this study

|  |  |
| --- | --- |
| Wireline log | Uses |
| Gamma Ray (GR) | 1. Lithology interpretation  2. Calculation of shale volume  3. Calculation of permeability and porosity |
| Caliper (CALI) | 1. Detection of permeable zones |
| Resistivity logs [shallow (RACELM), medium (RACEHM), deep (RD), true (RT)] | 1. Lithology interpretation  2. Detection of hydrocarbon bearing zones  3. Calculation of water saturation |
| Bulk Density (RHOB) / Bulk Density Corrected (DRHO) | 1. Interpretation of lithology  2. Detection of hydrocarbon bearing zones  3. Calculation of porosity |
| Neutron Porosity (NPHI) | 1. Detection of hydrocarbon bearing zones  2. Calculation of porosity and permeability |
| Sonic (DT) | 1. Calculation of porosity and permeability |
| Photoelectric (PEF) | 1. Determination of minerals for lithology interpretation |
| Rate of Penetration (ROP) | 1. Lithology interpretation  2. Calculation of porosity and permeability |

Since some logs have a direct relationship to the output variables, our feature selection techniques should be able to select them as relevant for their respective outputs. Some logs also exhibit non-linear and/or complex relationships with the target hence it is expected that the feature selection techniques will be robust enough to identify the degree of the relationship.

Table 4 below shows how each technique performed in the selection of relevant features for permeability, porosity and water saturation predictions. Unsurprisingly, some techniques could not select any logs whiles others were not able to reduce the input variables. With the filter method, Pearson correlation statistically did not find any of the input variables to have a moderate to strong linear correlation. Pearson’s usually requires normally distributed variables for its parametric test (Akoglu, 2018) as such cannot handle the highly skewed permeability and water saturation distribution. Although the Pearson correlation is the most popular technique amongst researchers, subsurface data are not always normally distributed due to heterogeneity and involves multivariate analysis hence it will not always be appropriate. However, five of the input logs were selected as relevant for porosity prediction. This shows that porosity distribution in the reservoir has some correlation with some wireline logs although its empirical calculation is dependent on density and neutron logs. Kendall’s and Spearman’s correlations tend to handle non-normal data distributions better hence were able to select five and six relevant features each for permeability and porosity respectively. However, no input logs were selected for water saturation indicating how complex the relationship between the wireline logs and this output is. Water saturation has been deemed to be the most challenging petrophysical property to be measured (Kadhim 2014, Gholano, 2016). Interestingly, none of these techniques selected the neutron log as having a strong to moderate correlation to porosity hence its input in empirical calculation of porosity seems to affect its accuracy. The RHOB, GR, RT and PEF logs, however, were the most selected relevant logs. The only limitation with the filter method is that the strength of the relationship between the input and output cannot be determined. Usually, the significance of the input to the output is mistakenly reported instead of the strength of the relationship.

The BFE method was able to identify the features whose removal resulted in the most minimum change in the model performance hence those features were dropped for all the outputs. This method selected eight and eleven relevant features for permeability and water saturation predictions respectively. The model, however, selected all input features for porosity prediction indicating its unsuitability for use in dimensionality reduction. The RFE also demonstrated its inability to reduce input variables for prediction of the three outputs as the method selected all twelve of the input logs as relevant. With the two wrapper methods selecting all the input logs as relevant for porosity prediction, it confirms that porosity itself is relevant in the calculation of permeability and water saturation. Therefore, logs that are used to estimate water saturation and permeability will invariably exhibit some form of relationship with porosity. However, the main aim of feature selection is only to select the most relevant logs hence regardless of the kind of relationship that exists between wireline logs and porosity, some will be more relevant than others. The inability of the wrapper methods to identify these logs makes it unsuitable for use in porosity prediction.

The Lasso model based on its best score eliminated all irrelevant features after penalising their coefficients. Only GR was selected as relevant for permeability prediction and although it agrees with the domain knowledge and may probably be the best log, it was expected that some other logs like CALI or DT would have been selected as well. Only five logs were deemed relevant for porosity prediction. Although the selected logs exhibits a relationship with porosity, the most relevant log based on linear relationship, RHOB, was not selected hence making this technique questionable. For water saturation, most of the relevant features selected by this method agreed with the domain knowledge.

The Random Forest technique, with its renowned potential in feature selection, selected 8 and 9 relevant logs for permeability and water saturation respectively. It was expected that high feature-feature correlation will be assigned equal importance and also give preference to features with unique values. The technique, however, greatly reduced the dimensionality for porosity prediction to two features. The SelectKBest technique complimenting its P-value score with F-test scores selected 5, 7 and 8 relevant logs for permeability, porosity and water saturation predictions respectively. It is expected to perform better than the filter method due to the added scoring metric of F-test which estimates the degree of linear dependency between random variables.

Table 4 Selection of features by the feature selection techniques for reservoir characterisation

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Feature selection technique | Permeability | | Porosity | | Water saturation | |
| Number of features | Comments | Number of features | Comments | Number of features | Comments |
| Baseline | 12 | Baseline model uses all features | 12 | Baseline model uses all features | 12 | Baseline model uses all features |
| Pearson | 0 | Could not select any logs | 5 | GR, RACELM, RT, PEF, RHOB | 0 | Could not select any logs |
| Kendall | 5 | GR, RACELM, RD, PEF, RHOB | 5 | RACELM, RD, RT, PEF, RHOB | 0 | Could not select any logs |
| Spearman | 6 | GR, RACELM, RD, PEF, RHOB, RT | 6 | GR, RACELM, RD, PEF, RHOB, RT | 0 | Could not select any logs |
| BFE | 8 | CALI, DRHO, DT, RD, RACEHM, ROP, RHOB, PEF | 12 | Model selected all features | 11 | CALI, DRHO, DT, NPHI, GR, ROP, RACEHM, RD, RT, PEF, RHOB |
| RFE | 12 | Model selected all features | 12 | Model selected all features | 12 | Model selected all features |
| Lasso | 1 | GR | 5 | GR, DT, PEF, RD, RACELM | 7 | ROP, RT, PEF, RD, RACEHM, DT, RHOB |
| Random Forest | 8 | GR, RACELM, DRHO, RT, RD, PEF, RHOB, DT | 2 | GR, RHOB | 9 | DRHO, DT, NPHI, ROP, RACEHM, RD, RT, PEF, RHOB |
| SelectKBest | 5 | CALI, GR, RACELM, PEF, RHOB | 7 | RACEHM, GR, RACELM, PEF, RHOB, RD, RPCELM | 8 | CALI, DRHO, GR, ROP, RT, PEF, RHOB, RACELM |

Accurate permeability measurements are required along the entire length of reservoirs as their distributions and variations are critical to the development of completion strategies, dynamic-flow calculations amongst others. Permeability distribution within reservoirs is affected by several parameters like volume of shale, porosity amongst others. Hence logs that relate to these parameters are expected to be identified in conjunction with permeability detection logs. The empirical calculation or permeability involves a multivariate regression analysis between porosity log and shale volume. The measurement of porosity highly influences the estimation of hydrocarbon volumes. This petrophysical property is critical in the development of reservoirs hence most wireline logs run in a well carries vital information about it. The calculation of porosity depends highly on the accurate interpretation of lithology. Total porosity (PHIF), used as a target in this study, is empirically calculated from wireline logs using RHOB which is then corrected to the overburden. NPHI is usually integrated in its calculation due to its ability to correct against variations caused by mud filtrate invasion but surprisingly none of the techniques selected it as relevant. The feature selection techniques that came close to selecting logs with relationships to permeability and porosity are the Random Forest and SelectKBest methods. The estimation of water saturation has been stated as the most difficult petrophysical property to measure because there are various approaches used in its calculation which often leads to widely different results. The accurate estimation of water saturation also equates to hydrocarbons originally in place hence it is extremely critical to the life of a field and development strategies to be employed. From wireline logs, the resistivity logs have been used in its calculation whiles relating it to porosity hence the feature selection techniques are expected to choose these relevant logs. Based on this, the Lasso, Random Forest and SelectKBest techniques were able to select most of the relevant logs. Further analysis on how these techniques will perform when used to train a supervised machine learning model is discussed below.

## Model prediction using selected features

The feature selection techniques used in this study exhibited diversity in terms of the selected relevant features. This led to the model performing differently with some techniques resulting in better model performance. The main aim of this study is not the R-Squared results of the input features but to determine the most robust feature selection technique that will result in low errors and high computational efficiency. In reservoir characterisation, the minimum increase in accuracy hence reduction in error highly influences decision making that can impact the life of a field.

In terms of permeability prediction (figure 2), the Random Forest technique outperformed the baseline and other techniques by scoring a MAE of 1.9576 and RMSE of 12.6428. The SelectKBest technique resulted in the second lowest error estimation albeit with a more efficient computational time due to a smaller number of input features. The superiority of the Random Forest technique can be attributed to the selection of DT, RT and RD logs which the SelectKBest technique failed to select as a relevant log. The low errors recorded by both Random Forest and SelectKBest techniques are because the relevant logs selected are lithology and porosity based logs. This confirms the dependency of permeability on lithology and porosity. The Pearson correlation technique could not identify any relevant logs for predictions, however, Spearman and Kendall techniques recorded lower errors and higher computational efficiency over the baseline. The BFE technique had the worst performance for permeability predictions scoring 5.3491 and 34.8034 for MAE and RMSE respectively. This is because the technique eliminated the GR log which has a strong relationship to lithology, adding to the increased error measurement. This reason is affirmed by the lasso technique that selected only GR log and recorded lower error measurements than the BFE method.

Figure 2 Evaluation of permeability prediction using all feature selection techniques

For porosity, the best performing technique in terms of error measurement was the baseline, RFE and BFE techniques because all the input features were selected but this results in higher execution time (figure 3). The purpose of feature selection is to eliminate some features to reduce dimensionality as well hence, techniques that selects some features and exhibits error measurements close to the baseline are more appropriate. With this in mind, SelectKBest method scored 0.0025 MAE and 0.0036 RMSE recording an execution time of 0.42 seconds. All the filter methods performed comparably well and even better than the SelectKBest in terms of execution time due to the few features selected. The Lasso technique, with similar number of features as the Pearson and Kendall method recorded the worst error measurement as expected. This is because, RHOB which has the most significant relationship with porosity was not selected as a relevant log. This clearly confirms that although the aim is to reduce the number of input features, domain analysis of the target and input variables is critical to machine learning. The most efficient feature selection technique for porosity prediction is the Random Forest that selected only two features hence recorded an execution time of 0.17 seconds. The technique however, recorded a MAE of 0.0033 and RMSE of 0.0043.

Figure 3 Evaluation of porosity prediction using all feature selection techniques

In predicting water saturation, the filter methods could not identify any relevant logs hence were not used in the GBR model (figure 4). With the difficulty in predicting water saturation well known, the lasso regularisation technique resulted in the least MAE and RMSE of 0.0034 and 0.0117 respectively. The SelectKBest method, in comparison to the lasso technique, exhibited the most efficient computational time albeit having more features, an RMSE of 0.01225 and a MAE of 0.0038. The time efficiency of the SelectKBest can be attributed to less noise in the types of logs selected. The wrapper methods resulted in the worst performance. Random Forest technique recorded a MAE of 0.0035 which is less than that of the SelectKBest technique. However, its RMSE was 0.00032 higher than that of SelectKBest. This may be attributed to the variance of the frequency distribution of error magnitudes of the selected logs. The performance of the Random Forest technique as compared to the lasso technique is due to the types of logs selected. The addition of DRHO and NPHI logs increased the error predictions indicating its unsuitability for water saturation prediction. Also, the sensitivity of decision trees to certain data distributions as shown in the crossplots in figure 1 can be attributed to its performance.

Figure 4 Evaluation of water saturation prediction using all feature selection techniques

# Conclusions

In this study, eight feature selection techniques for data dimensionality reduction have been presented. The GBR model was used to train the reduced datasets and compared each technique through MAE, RMSE and execution time. It is acknowledged that dimensionality reduction is not just to make the execution of models faster but to improve model performance as well. Based on the evaluation criteria used, the Random Forest feature selection technique proved to be the most effective for permeability prediction. The SelectKBest technique, however, was the best in removing unimportant features whiles scoring comparative minimum errors for permeability prediction. For porosity prediction, the most interesting results identified was that all logs used as input features gave better performance. However, the best technique to reduce the input features and still give comparable model performance is the SelectKBest technique. The prediction of water saturation also showed that Lasso regularisation is the best technique in terms of dimensionality reduction and low prediction errors.

The results confirm that combining these techniques with a supervised machine learning model can lead to a satisfactory dimensionality reduction and improved model performance with respect to the baseline features. Although there is no straightforward approach in dimensionality reduction, this study offers a possible approach in improving prediction accuracies, reducing associated risks and uncertainties in the domain of reservoir characterisation. As the overall decrease in reservoir properties prediction errors will in turn increase the efficiency of exploration, production and management activities. Based on the success of this study, efforts will be made to combine the best feature selection techniques in developing a custom ensemble that has not been implemented in reservoir characterisation.

# Acknowledgement

The authors express their sincere appreciation to University Teknologi Petronas and the Centre of Research in Enhanced Oil recovery for financially supporting this work through YUTP grant (015LCO-105).

# Computer code and data availability

This study is executed via Jupyter Notebooks hosted at repository link. The Volve well data used is hosted at xxx.

# References

Akande, K.O., Olatunji, S.O., Owolabi, T.O., AbdulRaheem, A.A., 2015. Comparative analysis of feature selection-based machine learning techniques in reservoir characterization. Soc. Pet. Eng. - SPE Saudi Arab. Sect. Annu. Tech. Symp. Exhib. 1–12. https://doi.org/10.2118/178006-ms

Akoglu, H., 2018. User’s guide to correlation coefficients. Turkish J. Emerg. Med. 18, 91–93. https://doi.org/10.1016/j.tjem.2018.08.001

Blessie, E.C., Karthikeyan, E., 2012. Sigmis: A feature selection algorithm using correlation based method. J. Algorithms Comput. Technol. 6, 385–394. https://doi.org/10.1260/1748-3018.6.3.385

Boyd, S., Vandenberghe, L., 2004. Convex Optimization, Convex Optimization. Cambridge University Press. https://doi.org/10.1017/cbo9780511804441

De Silva, V., Tenenbaum, J.B., 2003. Global versus local methods in nonlinear dimensionality reduction, in: Advances in Neural Information Processing Systems.

Ebrahimy, H., Azadbakht, M., 2019. Downscaling MODIS land surface temperature over a heterogeneous area: An investigation of machine learning techniques, feature selection, and impacts of mixed pixels. Comput. Geosci. 124, 93–102. https://doi.org/10.1016/j.cageo.2019.01.004

Kou, G., Yang, P., Peng, Y., Xiao, F., Chen, Y., Alsaadi, F.E., 2020. Evaluation of feature selection methods for text classification with small datasets using multiple criteria decision-making methods. Appl. Soft Comput. J. 86, 105836. https://doi.org/10.1016/j.asoc.2019.105836

Laib, M., Kanevski, M., 2019. A new algorithm for redundancy minimisation in geo-environmental data. Comput. Geosci. 133, 104328. https://doi.org/10.1016/j.cageo.2019.104328

Moghimi, A., Yang, C., Marchetto, P.M., 2018. Ensemble Feature Selection for Plant Phenotyping: A Journey from Hyperspectral to Multispectral Imaging. IEEE Access 6, 56870–56884. https://doi.org/10.1109/ACCESS.2018.2872801

Pedregosa, F., Michel, V., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D., Pedregosa, F., Varoquaux, G., Gramfort, A., Thirion, B., Grisel, O., Dubourg, V., Passos, A., Brucher, M., Perrot andÉdouardand, M., Duchesnay, A., Duchesnay EDOUARDDUCHESNAY, Fré., 2011. Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot. J. Mach. Learn. Res. 12, 2825–2830.

Rock, N.M.S., 1987. CORANK: A Fortran-77 Program to Calculate and Test Matrices of Pearson, Spearman, and Kendall Correlation Coefficients with Pairwise Treatment of Missing Values. Comput. Geosci. 13, 659–662.

Rose, L.T., Fischer, K.W., 2011. Garbage In, Garbage Out: Having Useful Data Is Everything. Measurement 9, 222–226. https://doi.org/10.1080/15366367.2011.632338

Wei, G., Zhao, J., Feng, Y., He, A., Yu, J., 2020. A novel hybrid feature selection method based on dynamic feature importance. Appl. Soft Comput. J. 93, 106337. https://doi.org/10.1016/j.asoc.2020.106337