MACHINE LEARNING WORKFLOWS FOR DETECTION OF HIGH WATER CUT UNCONVENTIONAL WELLS USING PETROPHYSICAL LOGS

A Thesis

by

JONATHAN DOMINIC FOSTER

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of
MASTER OF SCIENCE

Chair of Committee,	Ryan Ewing
Co-Chair of Committee,	Siddharth Misra
Committee Member,	Hongbin Zhan
Head of Department,	Julie Newman

August 2021

Major Subject: Geophysics

ABSTRACT

For many years, high water-cuts in the Delaware basin have been the source of much frustration for oil and gas operators producing in the region. The goal of this thesis was to construct automated workflows which are capable of predicting very early on in a horizontal well's lifetime whether or not it will produce a substantially higher amount of water compared to hydrocarbon. With the intent to accomplish this goal, two different data-driven workflows have been developed. Each workflow focused on the differentiation of high water-producing wells (HWPs) and low water producing wells (LWPs) using machine learning (ML) algorithms. Both data-driven workflows use well log data, which provide information about the rock properties surrounding a given wellbore. The first data-driven workflow extracted out summary features from the well logs with respect to depth intervals below the kick-off point of a given wellbore, which is the point which a wellbore begins to transition from vertical to lateral. Using features extracted from well log data from 20 horizontal wells from the Delaware basin, supervised ML algorithms were trained to differentiate and predict which wells would be HWPs and LWPs. Logistic regression proved to be the most accurate supervised ML algorithm for the first proposed workflow. This workflow produced promising median F1 and Mathew's correlation coefficient (MCC) scores of 0.96 and 0.92, respectively, for 100 cross-validation training iterations. The second data-driven workflow used unsupervised ML algorithms to assign a predicted lithology to every sample for 500 ft of well log data for 17 wells from the Delaware basin. This resulted in 5 unique lithologies which were found when all 17 wells were combined together. Using these predicted lithologies as a guide, features were extracted for all 17 wells and then used to train supervised machine learning algorithms to differentiate the two well classes: HWP and LWP. Using 100 cross-validation training iterations, three supervised algorithms

proved very comparable: K-Nearest neighbors, logistic regression, and support vector machine. Each of these supervised algorithms produced a median MCC score of 0.90. The geologic meaning of the most informative features from both workflows were also interpreted.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a thesis committee consisting of Professor Siddharth Misra of the Petroleum Engineering Department and Professors Ryan Ewing and Hongbin Zhan of the Geology and Geophysics Department.

The data used in this thesis was generously provided by Mark Nibbelink of Enverus.

The students of Dr. Siddharth Misra's research group provided technical support across many stages of the thesis project.

Funding Sources

Graduate study was supported by the Berg-Hughes Center for Petroleum and Sedimentary Systems and the Crisman Institute for Petroleum Research.

TABLE OF CONTENTS

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

1.1 Background

With the increase in popularity of the usage of horizontal drilling and hydraulic fracturing in combination, the practice of producing hydrocarbons from unconventional formations greatly expanded within the U.S. These unconventional formations earned their name primarily due to the fact that they were uneconomical targets until the early 2000s or so. These formations were not common targets for hydrocarbon production due to the nature of their tight, low-permeability rock matrices. This of course changed as horizontal drilling and hydraulic fracturing became a more economical prospect for operators. One of the most oil-prolific regions of onshore United States is the Delaware basin, which is the western-most region of the greater Permian located in northwest Texas and southeast New Mexico. Although the Delaware basin is a major oil-producing region for the US, the primary fluid produced from hydrocarbon wells in the Delaware basin is actually water. One study found that a quarter of a set of 10,000 shale-oil wells in the Permian basin are producing at least 70% formation water (Male, 2019). Produced water forecasts are estimating as high as 30 million barrels/day from target formations and wells with a water-to-oil ratio (WOR) of 10:1 is not unheard of (Duman, 2019). The percentage of water produced from a hydrocarbon-producing well is referred to as a water-cut. The brine produced from these wells is not readily useable for drinking or irrigation, thus must be managed or processed in some manner.

1.1.1 Motivation

There are many concerns which revolve around the transportation, reuse, and storing of the formation water produced from these high water-cut wells in the Delaware basin. There are estimations which predict unit cost for produced water will rise to over US$5.00/bbl (Duman, 2019). With the cost of operating these wells rising at their current rates, 20% of undrilled Permian barrels can become non-commercial. The reuse of produced water has the potential to offset some of the cost of water management by saving money on hydraulic fracking jobs, but in especially high water-cut ratio wells operators are unable to cheapy reinject all produced volumes (Duman, 2019).

One common method of reuse of formation water in the Delaware or Permian basin is through saltwater disposal (SWD) wells (Scanlon et al., 2017). Saltwater disposal wells are wells which are used to inject saltwater into the subsurface. The SWD solution raises concerns for many who are worried about the environmental impacts that may follow the reinjection of produced formation water into shallower, conventional geologic formations (Scanlon et al., 2017). These shallower formations are now being reported to be experiencing overpressure by operators in the Delaware. Another approach which is being considered is to drill deeper injection wells into the Ellenberger formation. Injection of disposal water into deeper formations can induce seismicity in the area (Scanlon et al., 2017). Aquifer contamination is another primary concern resulting from produced formation water. The produced formation water from shale-oil plays has been tested to contain Benzine and other BTEX compounds at above-safe levels of drinking and irrigation water quality (Khan et al., 2016).

1.1.2 Objective

For the purposes of mitigating the above-mentioned hazards from produced formation water, two novel data-driven workflows have been created to predict and classify which wells will produce an excessive amount of water. The goal of developing these two workflows was to provide quick, automated processes for determination of relative water production very early on in a horizontal well's lifespan. The incorporation of machine learning techniques allows for rapid testing to determine if a well will produce a large quantity of water compared to its target hydrocarbon fluid. Both of the proposed workflows use a combination of petrophysical logs, well trajectory data, and production data for each well to train machine learning (ML) algorithms in order to predict well classification. With knowledge of whether or not a well will produce an excessive amount of water as soon as the well is logged, operators can make an informed decision on whether or not production from a well is going to be more harmful or more beneficial overall. Along with the proposed workflows, the geological features which appear to be significant when differentiating high water-cut wells and low water-cut wells were also determined.

1.2 Machine Learning in Oil and Gas

Machine learning methods are becoming increasingly popular in the oil and gas industry (Miah et al., 2020; Mohamed et al., 2015; Hajizadeh, 2019). The workflows proposed in this thesis draw inspiration from studies which have shown success in the utilization of machine learning algorithms predicting properties of geologic formations or production for wells. One study which utilizes methods that overlap greatly with the methods proposed in this thesis is Guevara et al. (2017). In Guevara's approach, well log data from vertical wells is used to interpolate features in the horizontal wellbore where production occurs and

predict cumulative production itself. To perform these tasks, they used a relatively large suite of well logs and extracted features relative to formations in the vertical wellbore (Guevara et al., 2017).

1.3 Well Log Suite

During the development of the proposed workflows, the controls placed on the well log suite has varied over time. The sample count of available wells will increase as controls on the required well log suite are decreased. However, the increase in sample count has resulted in less accurate relative water production. The well log suite which has resulted in the best predictability of relative water production is as follows: gamma ray (GR), neutron porosity (NPHI), density porosity (DPHI), deep resistivity (ILD), and shallow resistivity (ILS). This combination of well logs has proven to be the most effective for differentiation of HWPs and LWPs after a series of testing different data sets with varying sample count and well log suites. This suite limits the sample count quite strictly, with the data set only consisting of 20 horizontal wells in the Delaware basin. Of these 20 wells, ten of them are HWPs and ten of them are LWPs. The final well log suite is the classic "triple combo" of gamma ray, porosity, and resistivity measurements. The triple combo has been a staple feature in petrophysical interpretation for many decades, due to its versatility in interpretable data. These 5 well logs provide insight into lithology, fluid saturation, and pore space in the surrounding rock body. An example of the well log data is displayed in *Fig. 1* below.

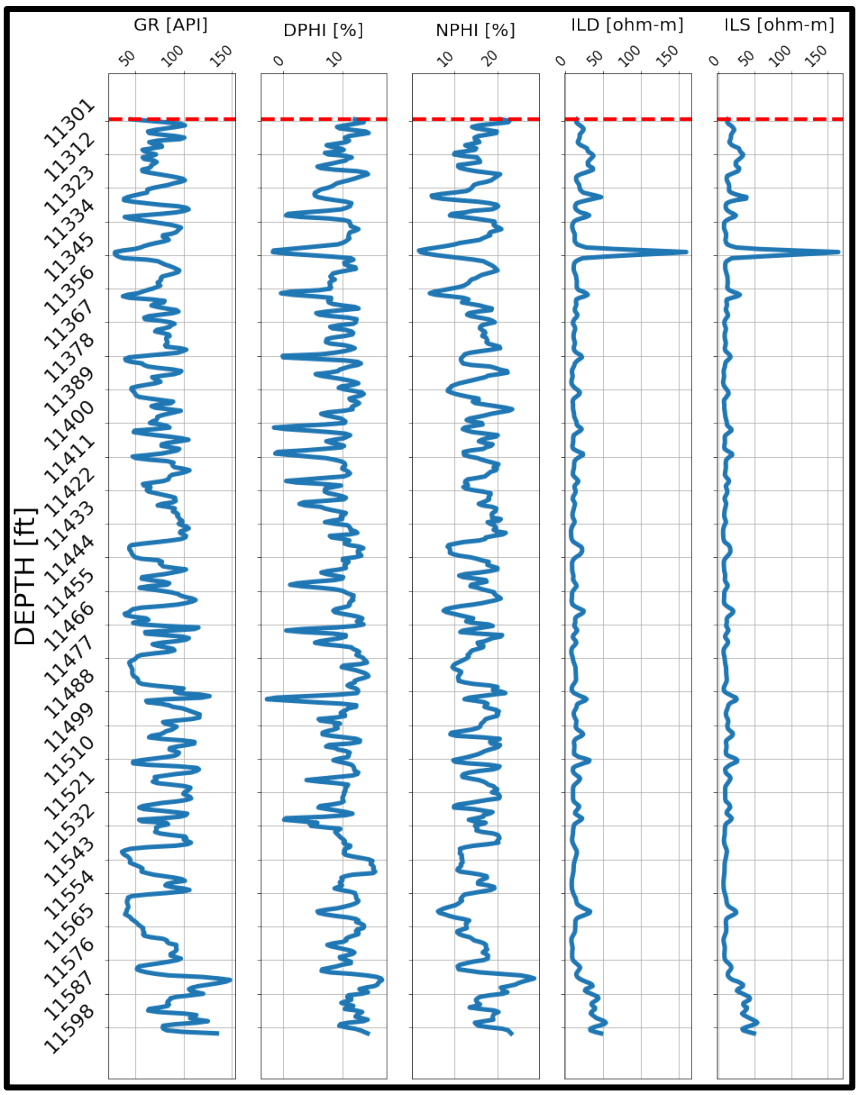

Fig. 1: Example display of well log data used in this thesis.

1.4 Production Data

It is important to note that the production data used in this study are calculated values provided by Enverus. The governing body for oil and gas production in the state of Texas, the Texas Railroad Commission (TXRRC), does not require operators to report the number of barrels of water produced from hydrocarbon wells. The data used in this study consists of synthesized water volumes which are calculated on the lease level by averaging well

tests. Once well tests are averaged, a water-cut percentage is allocated to each individual based on its well test. Although values used in this analysis are not directly reported water production data, Enverus has ensured that this method of produced water approximation is highly accurate and has been rigorously tested on a random well-by-well spot-check basis. Information on this topic was acquired through personal email exchange with Enverus support.

1.5 Target Labels

In order to assign a class to each well, we must examine the production data provided for each well. The key variable which these workflows are intended to predict is the relative water production ratio (WPR) for each well. As mentioned in the previous section, the values provided by Enverus are not exact reported values. To account for this discrepancy, the wells are discretized into two categories: high water producer (HWP) and low water producer (LWP). The category is calculated using the cumulative water production for each well and the total produced fluid for each well. The WPR is calculated using eq. 1:

$$WPR = \frac{Produced\ Water}{Total\ Produced\ Fluid} \quad \text{(eq. 1)}$$

The WPR is calculated at four different time intervals, all concentrated around the beginning of a well's production. These four time intervals are: 2 months, 6 months, 1 year, and then 2 years of first production. Once a WPR has been calculated for each of these 4-time intervals, they are all averaged together to become one average WPR. Wells which produce a WPR ≥ 0.70 are labeled as a high-water producer. Low water producers are wells whose WPR is < 0.50. The data sets used in the proposed workflows only utilize these two categories of wells. This is done intentionally to provide some distance in

feature-space between the two classes, so that the ML algorithms are primarily differentiating the two extreme cases. The labels of HWP and LWP operate as the target variables. These target variables can be considered the dependent variables for the data sets which will be generated in both workflows. The features, or the independent variables which describe the rock bodies in the subsurface, will be used to train machine learning algorithms to predict which well class a given well will belong to.

1.6 Spatial Distribution of Wells

The primary area of interest for this thesis is the Delaware basin, where high water-cut wells are a severe issue. The spatial distribution of the Delaware basin data set is shown *Fig. 2*. In the Delaware data set, there are 20 horizontal wells in total. 10 of these 20 are HWPs and the remaining 10 are LWPs. These wells are targeting either the Bone Springs sands or Wolfcamp shale, which are the primary targets for hydrocarbon production in the region. From these wells and the wells of the proof-of-concept data sets, well log data has been gathered and will be used to extract features for ML algorithm training. The spatial distributions for the two proof-of-concept data sets are also illustrated in *Fig. 2*, these two data sets are taken from the Fort Worth basin and Gulf Coast region.

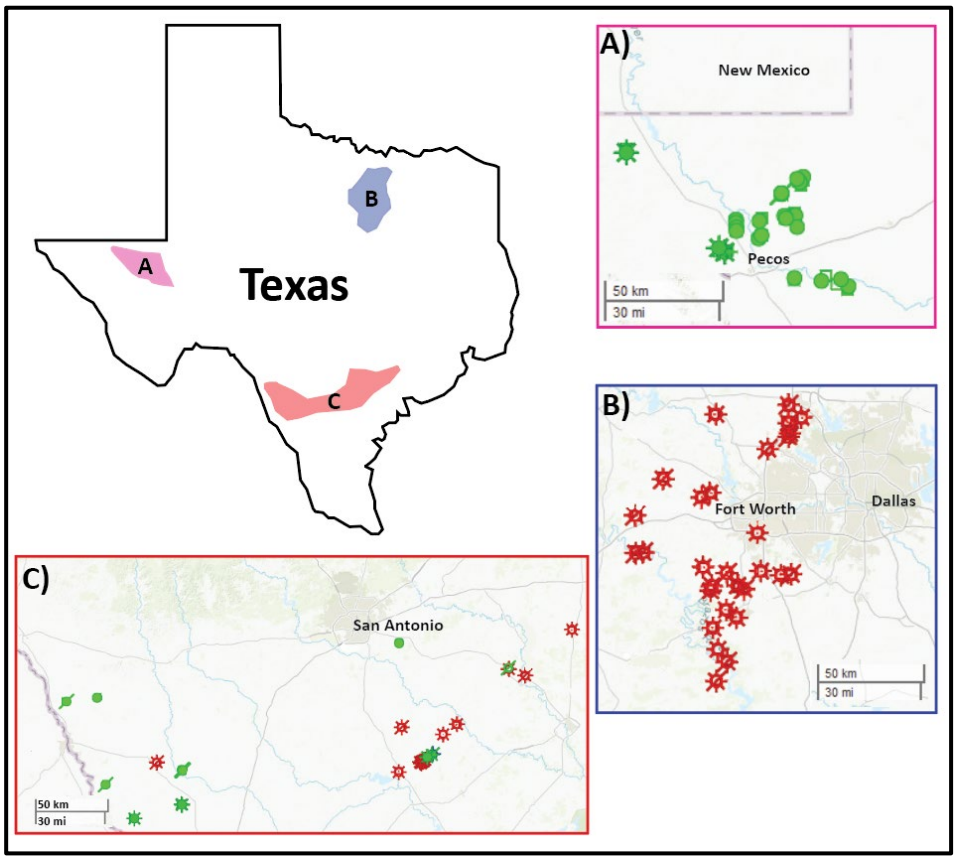

Fig. 2: Spatial distributions of all three data sets utilized – (A) Delaware Basin data set; (B) Fort Worth data set; and (C) Gulf Coast Region data set.

CHAPTER II

HIGH WATER CUT PREDICTION USING DEPTH-BASED FEATURE EXTRACTION

2.1 Interval of Interest

In our first approach, the data sets utilized consisted of well log data taken directly below the kick-off point (KOP) of the wellbore. In this analysis, the KOP is defined as the point at which the vertical wellbore began to transition into the horizontal wellbore. This data was used to train supervised ML algorithms to classify and predict relative water production. The generalized workflow is illustrated in *Fig. 3*. To begin this process, the KOP was calculated for each well. KOP calculation was performed algorithmically using trajectory data which is available for each horizontal well. From the trajectory data, the inclination of the drill bit at each depth increment is provided and used to determine when the horizontal drilling has begun. This method allowed us to generate an accurate calculation of where the KOP is located in measured depth (MD). The KOP acted as an anchor point in the well logs. This is a key component, due to the fact that well logs are provided in MD as well.

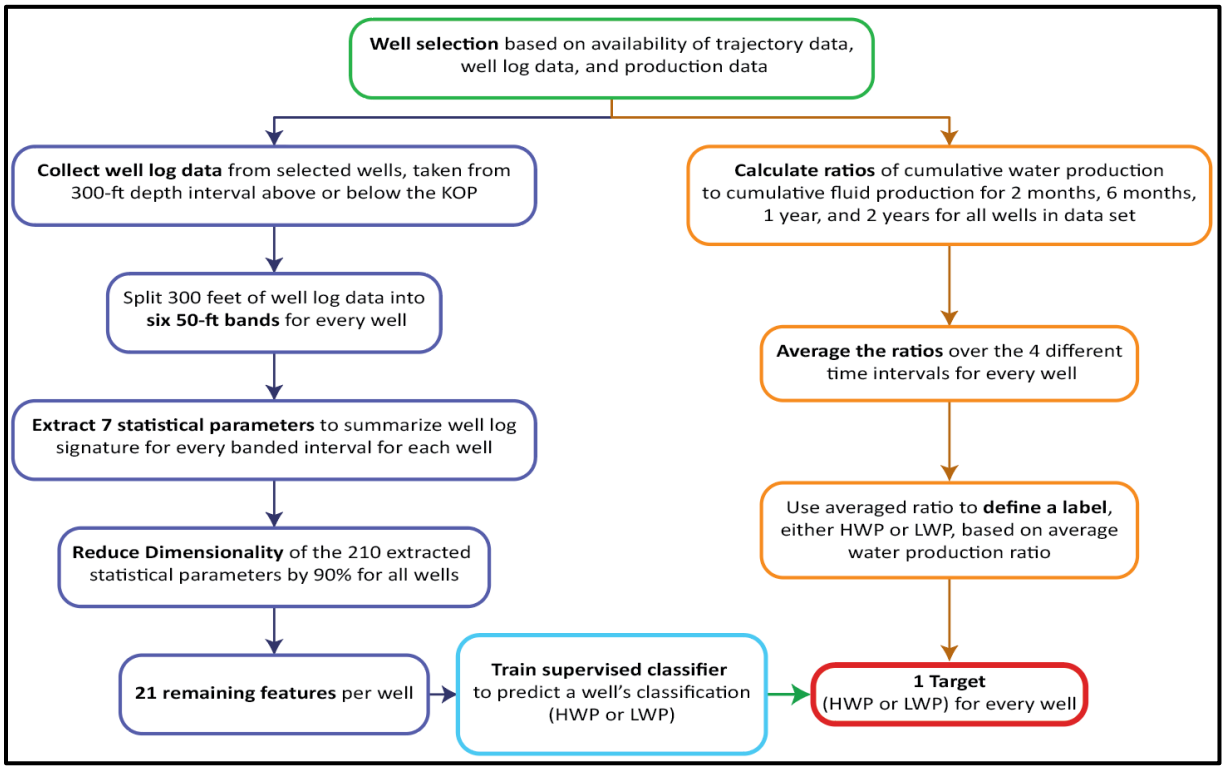

Fig. 3: Illustration of the data-driven workflow for chapter 2 where we train supervised classifiers to identify and predict excess water-producing wells using statistical parameters taken from 300 ft of well log data below the KOP.

Once the KOP was calculated, the well log data had to be truncated to only include depth intervals which could prove useful to ML algorithms later in the process. Many regions of the wellbores have been tested with the methods described in this thesis. Of these many, one region far outperformed the rest and that region was the depth intervals near the KOP of the wellbore. The region which was utilized in the first proposed data-driven workflow consists of 300 ft directly below the KOP. This provided us with 600 samples for each of the 20 wells in the data set, as well logs are sampled at every ½ foot.

2.2 Feature Extraction

After the well logs have been truncated to only include the KOP and 300 ft of logged data below the KOP, feature extraction was performed. A key component that this analysis wanted to preserve in the feature set is the relative depth of the wellbore. Keeping this information in our extracted numerical data allowed us to make inferences about the influence that the petrophysical differences with respect to relative depth had on our two well classes.

First, the 300 ft of well log data was divided into six different bands, or depth intervals. This process splits the well log data into six 50 ft bands. After the well log interval is split into 6 bands, seven statistical summary parameters are extracted for each well log, from each band. Given that we have 5 well logs, six band intervals, and 7 summary parameters, this results in a total of 210 features extracted to train ML algorithms. These seven parameters were: mean, median, kurtosis, skewness, root-mean square (RMS), entropy, and variance.

2.3 Feature Reduction

Not all 210 generated features were useful to ML algorithm training, there were redundant features in the data set. Thus, it was necessary to remove less useful features and preserve the most informative features for training supervised ML models. To separate algorithms from models, it is important to define them moving forward. An algorithm is essentially the foundation which a model will be generated by training it on well data to predict our two well classes. To begin the feature reduction process, an analysis of variance (ANOVA) F-test was applied to all the features with respect to the target label. The ANOVA test is very standard in the statistical realm. The F-test generates two linear regression models,

one built with randomly selected constants attached to a feature and another with one constant attached to a feature. If both of these generated models produce similar results, then the null hypothesis is accepted and the feature has been determined to carry no statistical significance with relation to the target (Osogba et al., 2020). The result of the F-test was two values for every feature, the p-value and F-value. To sum their utility briefly, a significantly low p-value (≤ 0.05, generally) indicates strong statistical significance and a F-value ≥ 1.0 also indicates statistical significance. With these typical values in mind, cut-off thresholds were applied to the features. The threshold values used for the final data set were p-values ≤ 0.08 and F-values ≥ 1.0. These threshold values were produced by adjusting threshold values and testing the data sets by training the supervised algorithms, then evaluating performance with various thresholds.

The second reduction method utilized in this workflow was focused on collinearity between features. This further eliminated redundancies in the feature-set by removing features which are essentially sharing information. For the purpose of reducing collinearity, a Pearson's correlation coefficient (PCC) (Guyon and Elisseeff, 2003) was calculated for each feature with respect to every other feature remaining in the data set. After a PCC is generated for all features, their relative correlation to one another is displayed in a correlation matrix as shown in *Fig. 4*. Due to the nature of our features, it was necessary to remove them based on a hierarchical scheme. The features are presented in the following form: "Well log", "statistical parameter", "band interval which parameter was taken (numbered from 0 – 5)." Since the geology will not vary substantially from band to band, it is expected that many features between two bands will have similar values. This can be considered an "inter-band" relationship. Redundancy with relation to our geologic

features which are described by the well logs are primarily a feature of "intra-band" relationships, or high collinearity within the same band interval. In general, the mean of any well log + band combination was preserved over all others, followed by skewness. A PCC value ≥ 0.80 was considered to be sufficient redundancy among features. The final result of both reduction methods was a data set of 21 features per well, for a total of 90% feature reduction.

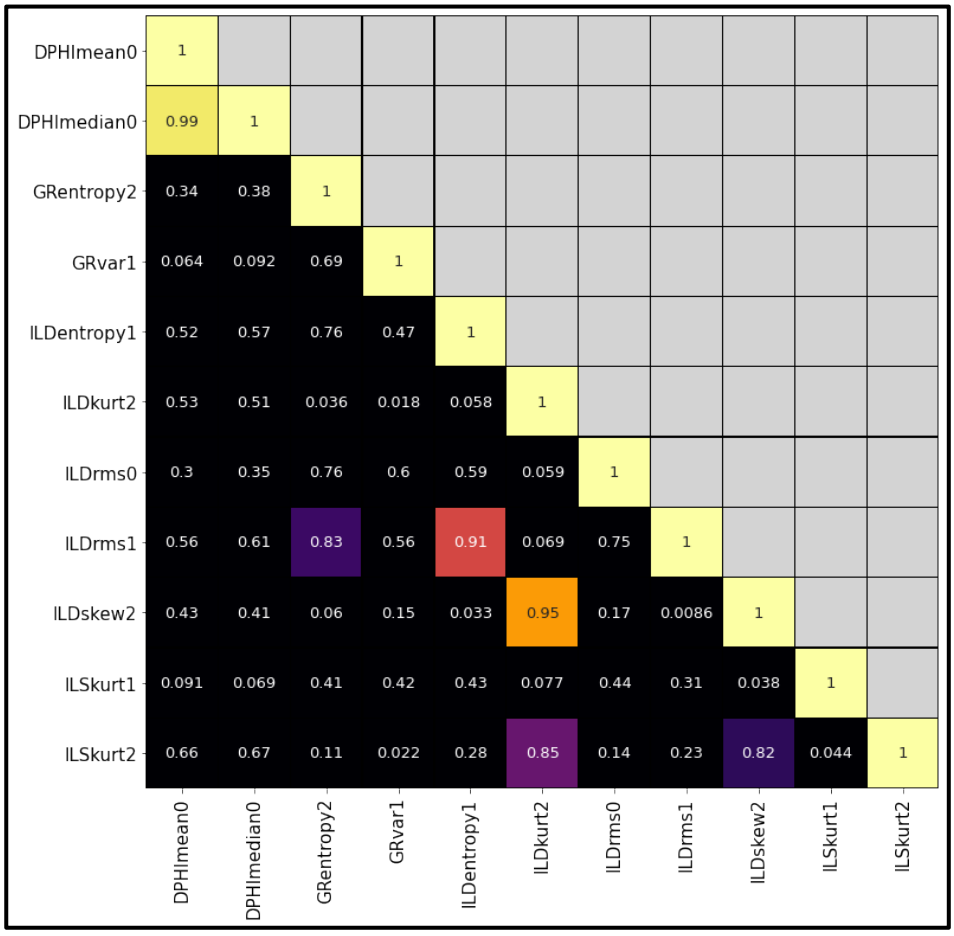

Fig. 4: Heatmap of the Pearson correlation coefficients for all feature pairs. Feature pairs with Pearson correlation coefficient (PCC) greater than 0.80 are shown with bright colors, while the black-colored boxes are low PCC values.

2.4 Data Processing

2.4.1 Scaling Features

Once feature reduction was complete, the data set was nearly prepared for building our supervised ML models. The final step before training was to scale and transform the features. This was a necessary step to ensure that the features are relatively Gaussian in distribution, as most ML models assume Gaussian distribution as part of their theory. To achieve more Gaussian distributions, two methods were utilized to scale and transform the data: A Z Score transform and Yeo-Johnson transformation. The Z Score transform shifts a feature's distribution with respect to the mean, μ, and the standard deviation, S. It was performed with *eq. 2*, where X represents a given sample from the feature space (Jain et al., 2005).

$$Z = \frac{X-\mu}{S} \qquad \text{(eq. 2)}$$

After reshaping the features with a Z Score transform, a Yeo-Johnson transformation was applied to shift the curve further. These two methods in combination generated a more bell-like curve which the ML algorithms tended to perform better on. The Yeo-Johnson transform uses the following set of equations (*eq. 3*), where λ can be any real number (Yeo and Johnson, 2000):

$$\psi(\lambda, x) = \begin{cases} \{(x+1)^{\lambda} - 1\}/\lambda & (x \geq 0, \lambda \neq 0), \\ \log(x+1) & (x \geq 0, \lambda = 0), \\ -\{(-x+1)^{2-\lambda} - 1\}/(2-\lambda) & (x < 0, \lambda \neq 2), \\ -\log(-x+1) & (x < 0, \lambda = 2). \end{cases} \qquad \text{(eq. 3)}$$

A key note in this process to take into account is that due to the small sample size which is utilized in this analysis requires that training data and the testing data be scaled

simultaneously. This is key to the training, so as to ensure that all the data are in the same relative distribution after scaling.

2.4.2 Cross-Validation

In order to properly determine the efficacy of this workflow cross-validation was necessary. Cross-validation is a technique which allows a model to be trained and tested on various iterations of the data set, to allow for a balanced evaluation of the model. Cross-validation splits the samples of a data set into *k* number of folds, for which then *k – 1* folds are used for training (Stone, 1974). The remaining fold, which is set aside from the training data, was used to test and evaluate the trained model. The cross-validation process allows the model to mix-and-match training and testing samples in multiple randomly generated training and testing schemes as it shuffles around the samples. This process allows the model's efficacy to be determined with a relatively wide spectrum of training data.

2.5 Supervised Algorithms

2.5.1 K-Nearest Neighbors

In the process of developing this workflow, a small collection of supervised methods was tested. Of the tested models, three outperformed the others by a substantial margin. The first of these three models was K-nearest neighbors (KNN). The KNN algorithm calculates the Euclidean distance between samples in feature space. Samples which illustrate similar characteristic in their features can be considered "neighbors," since they will be relatively close to one another in terms of Euclidean distance. When neighbors in a neighborhood are mostly of one class and distant from a neighborhood of the other class, a boundary is

drawn to represent the neighborhood. This is directed by defining how many neighbors, *n*, within a neighborhood is to be considered a "majority vote."

2.5.2 Support Vector Machine Classifier

The second supervised method which performed well was the support vector machine (SVM). The support vector machine is a learning machine for two-group classification challenges. The assumptions are: input vectors are mapped non-linearly to a feature space with very high dimensionality. Within the highly dimensional feature space, a linear decision surface is constructed to separate the two classes (Cortes and Vapnik, 1995). This algorithm can be extended to construct decision planes or hyperplanes, but within the confines of this workflow a linear decision surface with a dimensionality of 1 performs the best.

2.5.3 Logistic Regression

The third algorithm which performed well on the data sets tested was the logistic regression. The logistic regression is a modified version of the linear regression, with the distinction being that the output variable of the logistic regression is binary (Hosmer et al., 1995). Logistic regression algorithms use a Sigmoid function (*eq. 4*) to transform the variables of a data set.

$$Sigmoid(x) = \frac{1}{1-e^{-x}} \quad \text{(eq. 4)}$$

This regression model assigns weights to all features within a given data set. The weights of these features are passed through the sigmoid function to determine if the sample will fall on a value of 1 or 0. The logistic regression finds a boundary between samples of

different classes by determining the maximum log-likelihood distribution which best represents the data (Wu et al., 2019).

2.5.4 Hyper-Parameter Tuning

Each of these supervised methods have attributes which alter their learning habits. These attributes are defined upon training the algorithms and are called "hyper-parameters." Hyper-parameters vary between each algorithm and optimizing these parameters is a key component to an applied data science workflow. One of the most popular methods of hyper-parameter tuning is the grid search method (Bergstra and Bengio, 2012). With grid search, the process is to provide a range of values for any set of hyper-parameters to be used in a set of trials. The trials are run with each pair or combination of hyper-parameters and the set of hyper-parameters which has the best performance is preserved for training the model.

2.6 Evaluation of Supervised Models

The efficacy of these supervised models was quantified. Commonly, classification problems are simplified into a confusion matrix (*Fig. 5*) (Visa et al., 2011). The confusion matrix is a 2 x 2 array which helps to illustrate a classifier's ability to correctly predict a class for a sample. A confusion matrix has two axes: the predicted axis and the true axis. Both of these axes have binary possibilities, a false and a true class. This limits the possibilities of a prediction into four categories: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The true quadrants of the confusion matrix represent the correct predictions from a classifier, while the false quadrants represent incorrect predictions.

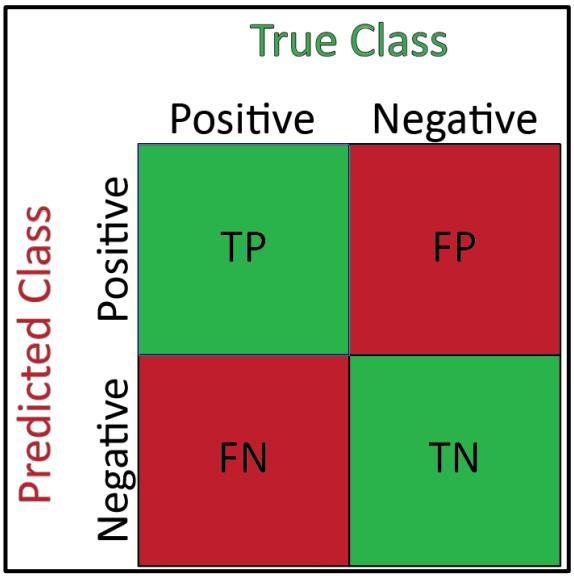

Fig. 5: Illustration of a typical confusion matrix. TP represents true positive; TN is true negative; FP is false positive; and FN is false negative.

Quantifying results from a confusion matrix has been done in a number of different ways. The methods used in this analysis are: F1 score and Matthew's correlation coefficient (MCC). F1 score is a common metric for evaluating classifiers. F1 scores range from 0.0 to 1.0, with 1.0 being a perfect score. The F1 score is calculated by using the following eq. 5:

$$F_1 = \frac{TP}{TP+\frac{1}{2}(FP+FN)} \quad \text{(eq. 5)}$$

Although the F1 score (Chicco and Jurman, 2020) is reliable for balanced data sets, it has a weakness of biasing towards the true positive predictions. This issue is supplemented by bringing in the Matthew's correlation coefficient (MCC). The MCC has the same goal as the F1 score, but uses a different equation (*eq.* 6) and thus different penalties. MCC is much more punishing for incorrect predictions in comparison to F1 score, due to the fact more weight is provided to the false classes. The MCC score ranges from -1.0 to 1.0, with

a score of -1.0 being all incorrect predictions and 1.0 being all correct predictions (Chicco and Jurman, 2020).

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad \text{(eq. 6)}$$

2.7 Proof-of-Concept Data Sets

It was also deemed appropriate to extend this workflow to other data sets with similar data. This allowed us to determine if the success, or lack thereof, of this workflow could be extended to other geologic regions outside of the Delaware basin. For purposes of validating this data-driven workflow, two data sets outside of the Delaware basin were also constructed. One of these validation data sets are from the Fort Worth basin and the other from the Gulf Coast region. It should be noted that the wells within these two validation data sets were not restricted by target formation, as the Delaware data set is, but they possessed sufficient well log, production, and trajectory data to be processed through this workflow. The Fort Worth data set has 29 wells, 11 HWPs and 18 LWPs. The Gulf Coast data set has 24 wells, 9 HWPs and 15 LWPs. The distribution of these wells geographically is illustrated in *Fig. 1*.

2.8 Determination of Important Features

Once the best performing algorithm for each data set was determined, the next step in this analysis was to determine which features were providing the most information to the generated models. This was determined using a method known as permutation testing (Pesarin and Salmaso, 2010). Permutation testing is a method which every feature in the data set is isolated and removed from the training data, one at a time. For every training iteration, the model is tested to determine how much prediction accuracy has changed

based on the chosen evaluation metric. This process provided insight into how much information each feature individually provided to a given model.

Permutation testing was performed on the 10 best performing models, which were determined through model evaluation metrics (F1 score and MCC score). The cycle of cross-validation was performed 100 times. For each iteration, the F1 score, MCC score, and the model generated was stored. Using the scores as a guide, the 10 best performing models from the best performing algorithm for the data set were gathered for permutation testing. The features were automatically presented in a ranked format for each model, although most of the features were not provided a rank nor any weight by the permutation process. The features are ranked from 1 to *n* ranked features, where 1 is the highest ranked. In order to take into account in the frequency which a feature appeared to be informative across multiple models, the number of models which a feature appeared significant in, f_{Ri}, was used in the following equation (*eq. 7*):

$$R_f = \frac{\sum_i^n R_i}{f_{R_i} * 100} \quad \text{(eq. 7)}$$

Where R_f, represents the final rank for a feature and R_i represents each rank for a feature for model *i*. From *eq. 7*, the feature with the lowest R_f value was the most informative feature for the set of 10 best performing models. With this ranking system, it was determined which well log, statistical parameter, or depth interval was the most informative when differentiating HWPs and LWPs.

2.9 Results and Interpretations

2.9.1 Delaware Basin Supervised Results

Out of the three supervised models utilized in this study the logistic regression consistently performed better than SVC and KNN, although only marginally better than SVC. After 100 training and testing cross-validation iterations, the logistic regression algorithm generated a bimodal distribution of F1 and MCC scores with very low variance (*Fig. 6*). The overwhelming majority of iterations stacked up on the median for both scoring metrics of 0.96 and 0.92 for F1 and MCC, respectively.

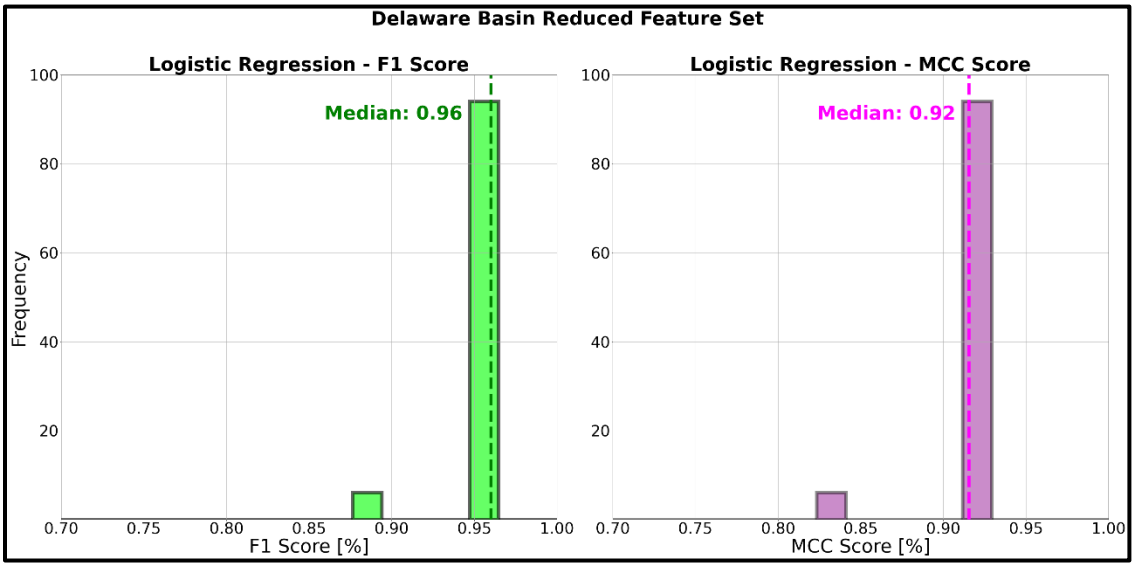

Fig. 6: Prediction performances of logistic regression over 100 training iterations after applying dimensionality reduction on the Delaware Basin data set. The median F1 score, represented by the green dashed line, was 0.96. The median MCC score, represented by the dashed purple line, was 0.92.

The variance coefficients for these distributions are significantly less than 1. This provides insight into how well the logistic regression models are performing on different permutations of the data set as it splits and shuffles the samples during cross-validation procedures. The F1 and MCC scores for the SVC models are 0.95 and 0.91, respectively. The distribution of the SVC metrics over 100 training iterations has more variance, but not

enough to raise the variance coefficients for either score above 1.0. The median F1 and MCC scores for the KNN algorithm are 0.91 and 0.84, respectively. This is a noticeable drop in both scores and the variance is significantly larger than both SVC and logistic regression.

2.9.2 Delaware Basin Most Informative Features

Using the methods described in section *2.8* of this thesis, the most informative features for the best performing models were ascertained for the Delaware basin data set. For each data set in this analysis, we sought to determine the top 10 features for each the best performing model per data set. Logistic regression, for example, is the best performing model on the Delaware data set, so the most informative features for this algorithm are the focus for this data set. The results of feature ranking methods are displayed in *Table 1*. Readers will notice that there are only 7 ranked features in *Table 1*. This is due to the fact that the best performing models of the logistic regression algorithm only seem to utilize 7 features to predict well classification. The feature names are constructed as follows: "Well log mnemonic (GR, for example)," "statistical parameter," and "band from which features were extracted." The bands were numbered from 0 – 5, where 0 was the closest to the KOP and 5 was the farthest from the KOP. Upon observing the results from feature ranking, one will notice a few key points. The first being that the majority of the most important features were extracted from the closest bands to the KOP, and most of them from the 0^{th} band. Secondly, resistivity and porosity seem to play a key role in differentiating the two well classes. Resistivity and porosity logs consist of 6 of the 7 top ranked features for the logistic regression models.

Table 1: The most informative features to the best performing logistic regression models for the Delaware basin data set is presented. These features are ranked from most informative to least informative and are numbered 1 – 7, respectively. The feature names are constructed as follows: "Well log mnemonic (GR, for example)," "statistical parameter," and "band from which features are extracted." Terms like "rms", "var", and "kurt" are all shortened versions of statistical parameters "root mean square", "variance", and "kurtosis."

Delaware Basin Logistic Regression Ranked Features	
FEATURE	**RANK**
ILDvar0	1
ILSkurt2	2
NPHIvar1	3
ILDrms0	4
GRvar1	5
NPHImean0	6
DPHImean0	7

2.9.3 Interpretation of Delaware Basin Most Informative Features

The most important feature for the logistic regression models when differentiating HWPs and LWPs is the variance of deep resistivity (ILD) nearest to the kick-off point (KOP). This indicates that there is a distinction between HWPs and LWPs when involving formation resistivity near the KOP. If the formation resistivity is changing rapidly throughout this 50 ft interval, it may be indicative of thin intervals of highly water-saturated rock which could be conduits or containers of formation water. Four out of seven top ranked features are within the 0^{th} band and the other 3 are in the 1^{st} and 2^{nd}. This conveys that the location of the KOP is of critical importance to separating the two well classes.

2.10 Proof-of-Concept Data Set Results

To ensure that this workflow could be universally applied to basins outside of the Delaware, additional data was gathered from the Fort Worth basin and Gulf Coast region. The methods described for the Delaware basin worked comparably well for our proof-of-concept data sets. The results for the Fort Worth basin are shown in *Fig. 7*. The logistic regression algorithm was also the best performing algorithm for the Fort Worth data set. The median F1 score and median MCC score were 0.89 and 0.82, respectively. This was a noticeable decrease in overall accuracy, compared to the Delaware data set, but these scores are indicative that the logistic regression models are still predicting accurately most of the time.

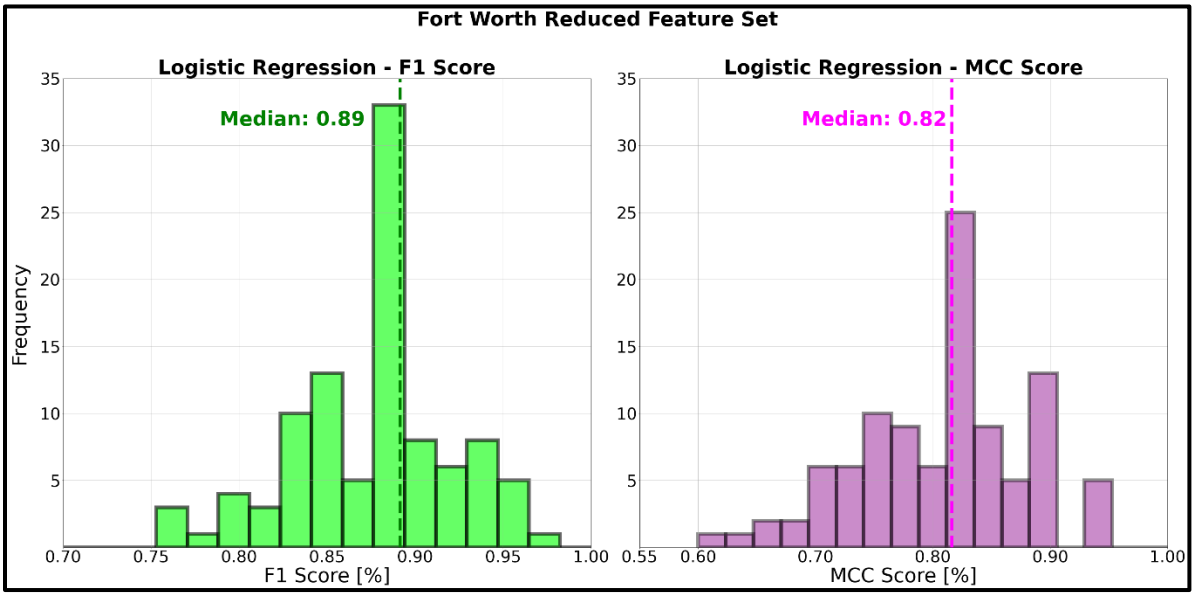

Fig. 7: Prediction performances of logistic regression over 100 training iterations after applying dimensionality reduction on the Fort Worth data set. The median F1 score, represented by the green dashed line, was 0.89. The median MCC score, represented by the dashed purple line, was 0.82.

The results for the Gulf Coast region are slightly improved from the Fort Worth data set. The logistic regression, again, proves to be the most accurate algorithm for separating the two classes of wells. As shown in *Fig. 8*, the median scores for both metrics are back up to the low 0.90's. The Gulf Coast data set still performs slightly less favorable than the Delaware basin data set by relative variance, but is still relatively accurate for most cross-validation iterations.

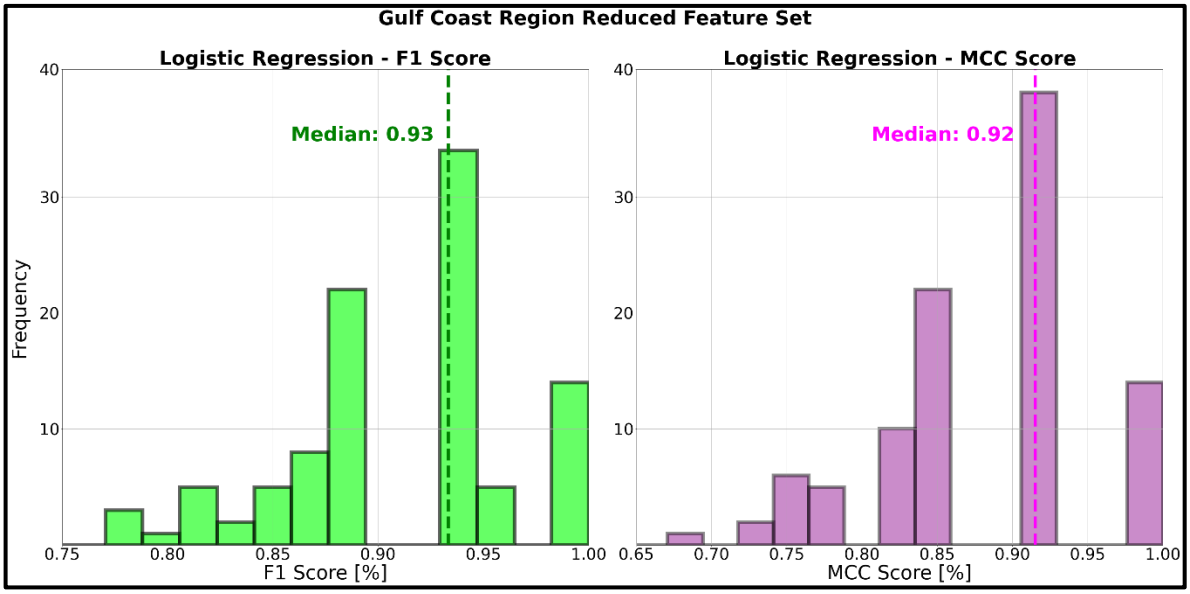

Fig. 8: Prediction performances of logistic regression over 100 training iterations after applying dimensionality reduction on the Gulf Coast Region data set. The median F1 score, represented by the green dashed line, was 0.93. The median MCC score, represented by the dashed purple line, was 0.92.

With the median scores for both of these regions as high as they are, it is reasonable to assume that this workflow can be applied to different regions. It should be noted that the features which contribute the most to the logarithmic models vary from basin to basin. Therefore, it is critical to train the models with respect to wells with similar geological data to new wells which relative water production is to be predicted for. Doing so will ensure

that models are trained to predict high water-cut wells for the particular geologic region the user is operating in.

It is also critical to note that the proof-of-concept data sets have not been subject to further scrutinization based on target formation. In the Delaware basin data set, the more restrictions placed on target formation and well production type increase the algorithm's ability to correctly predict well class based on these methods. In all likelihood, both data sets would display increased predictability based on the methods proposed if the samples were made more homogeneous.

2.11 Chapter II Conclusions

A data-driven workflow which utilizes features extracted from petrophysical logs based on relative depth to the kick-off point (KOP) of a given wellbore has been developed to predict high water-cut wells. These features can be extracted from various regions of the vertical and horizontal portions of the wellbore; however, the most effective region is directly below the KOP in the transitionary region between vertical and horizontal wellbores. 210 statistical summary parameters in total are extracted as features in this workflow. These 210 features are then reduced, by 90% for the Delaware basin data set, and then used to train supervised ML algorithms to predict which well class a well belongs to: HWP or LWP. Using 100 cross-validation iterations, which were trained and evaluated using F1 and MCC scores, generate median F1 and MCC scores of 0.96 and 0.92, respectively. These scores signify that this workflow can reliably predict, given proper training data, whether or not a brand new well in the same geologic region will produce a high or low amount of water relative to target hydrocarbon.

This workflow was further evaluated using two proof-of-concept data sets which are entirely separate from the basin of interest, the Delaware basin. The data set from the Fort Worth basin generates a histogram of F1 and MCC scores with median values of 0.89 and 0.82, respectively. The Gulf Coast data set generates a histogram of F1 and MCC scores with median values of 0.93 and 0.92, respectively. The results from the proof-of-concept data sets show that this workflow can likely be extended to other regions, which may experience similar problems of high water-cuts, and reliably predict high water-cut wells.

From the top 10 best performing models, generated from training a logistic regression algorithm, the most informative features to best performing models were determined using permutation testing. The resistivity logs (ILD and ILS) and the porosity logs (NPHI and DPHI) seemed to provide the most informative information to ML models to separate these two well classes. Given that the resistivity curves provide insight into the fluid saturation of a rock and the porosity curves provide insight into potential fluid storage, it makes logical sense that these features would be important when separating HWPs and LWPs.

CHAPTER III

HIGH WATER CUT PREDICTION USING UNSUPERVISED LITHOLOGY-BASED FEATURE EXTRACTION

3.1 Interval of Interest

The second data-driven workflow developed in this incorporates unsupervised ML methods. For this workflow, the interval of interest is expanded. Still using the KOP as the anchoring point, the 300 ft below is now accompanied by 200 ft above the KOP as well. The unsupervised methods are used to create pseudo-lithology types in each of these wells. With the additional 200 ft, the Delaware basin data set is reduced by 3 wells for a total of 17 for workflow validation. This increase in sample count provides unsupervised methods with 1000 sample points per well, however. The generalized workflow for the unsupervised approach is presented in *Fig. 9*.

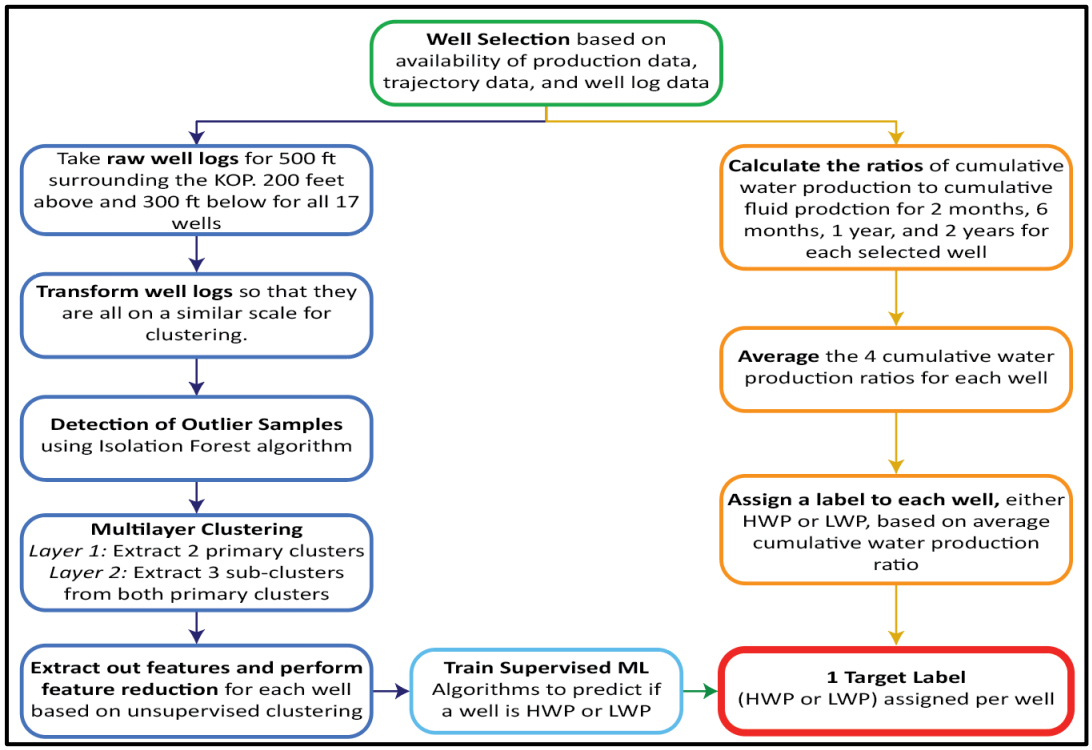

Fig. 9: Illustration of the generalized workflow for Chapter 3 methods. The well log data is clustered into 6 unique pseudo-lithologies using unsupervised ML algorithms, where features are extracted out for each cluster and each well log. These features are then used to train supervised models to predict well class.

3.2 Multi-Layer Clustering

3.2.1 Preprocessing Data

The clustering process began with all the wells being treated as one continuous well. This is crucial to the application of unsupervised models, as any variation in the well logs could cause a change in predicted lithology label for samples. Treating all the well log data as if they were all one well ensured that the lithology labels were consistent across all wells. If the unsupervised models were to be ran on the wells individually, cluster *A* in well 1 might not be the same cluster *A* as in well 2 in terms of petrophysical properties.

Once wells are all in one bucket, the well logs are scaled and transformed. Well logs such as resistivity tend to function on a logarithmic scale compared to other well logs. For this reason, it is necessary to apply a log base 10 transform to bring it more into scale with the other well logs. This transformation is applied to both shallow and deep resistivities. In order to supplement the unsupervised models, a handful of features were created from existing well logs. These created features are relationships such as gamma ray divided by neutron porosity and average porosity. After these new features are created, all features in the data set are scaled using a Z score transform and Yeo-Johnson transform as described in section *2.4.1*.

After scaling, the process of removing outlier samples began. Outlier samples will cause shifts in the distributions of unsupervised clusters, due to potentially noisy data. To remove this noisy data, an isolation forest algorithm is utilized to detect outliers. Operating under the assumption that outliers in a sample pool are "few and different" from normal samples, the isolation forest algorithm builds a series of trees. Because outliers possess different attributes compared to normal samples, outliers tend to be isolated from the rest of the samples near the root of the tree as opposed to deeper in the tree (Liu et al., 2008). As for this workflow, a contamination percentage of 8% is assumed. The contamination percentage is the percentage of data which is presumed to be outliers.

3.2.2 K-Means Clustering

This workflow utilizes two unsupervised clustering methods. The first of these methods is K-Means. A K-Means algorithm applies a process of partitioning a population of *N*-dimensions in to *k* sets (MacQueen, 1967). K-Means determines *k* initial cluster centers,

where *k* is defined by the user, and then each cluster center is refined to be the mean of constituent samples within similar clusters (Wagstaff et al., 2001).

3.2.3 Spectral Clustering

The second unsupervised model used in this workflow is spectral clustering. Spectral clustering constructs an affinity matrix, *A*, from a given set of points. That affinity matrix is then used to define a diagonal matrix, *D*, which is then transformed to *L* using *equation 8*.

$$L = D^{-\frac{1}{2}} * A * D^{-\frac{1}{2}} \quad \text{(eq. 8)}$$

The largest eigenvectors of *L* are then used to reshape the matrices further to cluster the samples, using K-means or another algorithm which attempts to minimize distortion (Ng et al., 2001).

3.2.4 Cluster Validation

3.2.4.1 Agreement

K-Means and spectral clustering are used simultaneously to cluster the data. They were ran simultaneously as a means to validate the clusters they are generating. For example, if two separate clustering algorithms are defining the same relative boundaries in feature space to separate *X* number of clusters, this provides solid evidence that these clusters are signal and not noise. This is illustrated in *Fig. 10*, which shows the distributions of a feature space which was clustered by both algorithms individually. Given that produced results which are greatly similar, this gives us good indication that these boundaries are likely separating two different lithology types in the well logs.

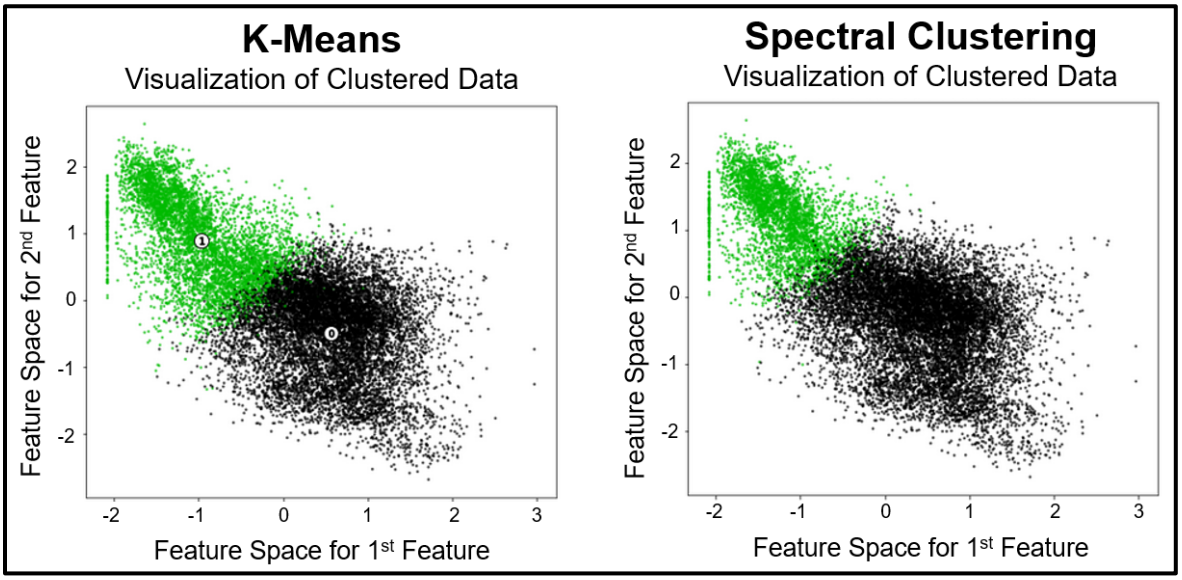

Fig. 10: Graphical representation of two methods of unsupervised clustering generating roughly the same boundaries between two clusters.

3.2.4.2 Silhouette Score

The second validation metric being used to evaluate the unsupervised methods is the silhouette score. Once data samples are clustered and assigned a label, the silhouette score for the samples can be determined. The silhouette score is algorithmic process which is commonly used to validate clustering performance. This algorithm calculates the distance between samples within the same cluster, say cluster A, and compares it to the distance of samples in cluster A to samples in cluster B (Rousseeuw, 1987). In short, this generates a silhouette score for every sample based on how small the intra-cluster distance is and how large the inter-cluster distance is. This silhouette score is then averaged based on each sample in the selection, which is defined when unsupervised models are run. An averaged silhouette score for a cluster can range from -1.0 to 1.0, with a perfect score being a 1.0. An illustration of the silhouette score for the used in this study is in *Fig. 11*.

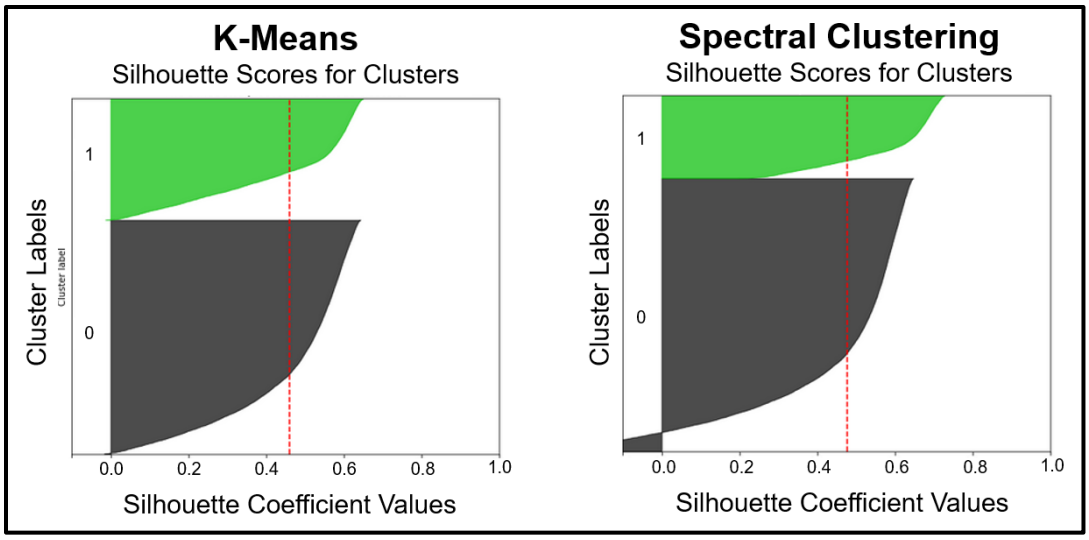

Fig. 11: Illustration of silhouette scores generated for both algorithms (K-Means and Spectral clustering). The dashed red line on each graph denotes the average silhouette score for both plots.

3.3 Clustering Results

The average silhouette score for a 2-cluster setup, which is the best apparent setup, was 0.46 for K-Means and 0.48 for spectral clustering. Both of the validation metrics producing relatively positive results for a 2-cluster setup indicates that this is likely the most optimal cluster selection for the large all-wells-in-one setup. All samples are labeled by both K-Means and spectral clustering. Each label is paired with its counterpart label from the other algorithm. This keeps consistency between both labels and acts as another bottle-neck for the unsupervised methods. This splits the data into two primary clusters, with one cluster small enough to consider noise from the first layer of clusters. The two primary clusters are labeled cluster A and cluster B. It is worth noting that the depth index for each sample and what well it belongs to had been preserved through this entire process. This allowed us to reinject samples back into the correct order and the correct well after clustering.

Cluster A contained 10,308 samples and cluster B contained 4063 samples after the first layer of unsupervised clustering. With this many samples per cluster, and reasonable geologic expectations, it was decided that these two primary clusters could be further reduced. The two clusters were then separated from one another for further, individual clustering. These two clusters were clustered into smaller sub-clusters using the same methods as described for layer one, aside from scaling the data as this was preserved into layer 2. Using the same validation scheme for layer 2 clustering, it was determined that cluster A should be subdivided further into 3 clusters and cluster B should be subdivided into 2. Again, the best performing unsupervised methods by silhouette score were K-Means and spectral clustering. The average silhouette score for both clusters A and B subdivision were approximately 0.48. Cluster A ended up being split into 3 sub-clusters, while cluster B was divided into two sub-clusters. The final 5 clusters are named "A0, A1, A2, B0, and B1" The number of samples per cluster is presented in the following *Fig. 12.*

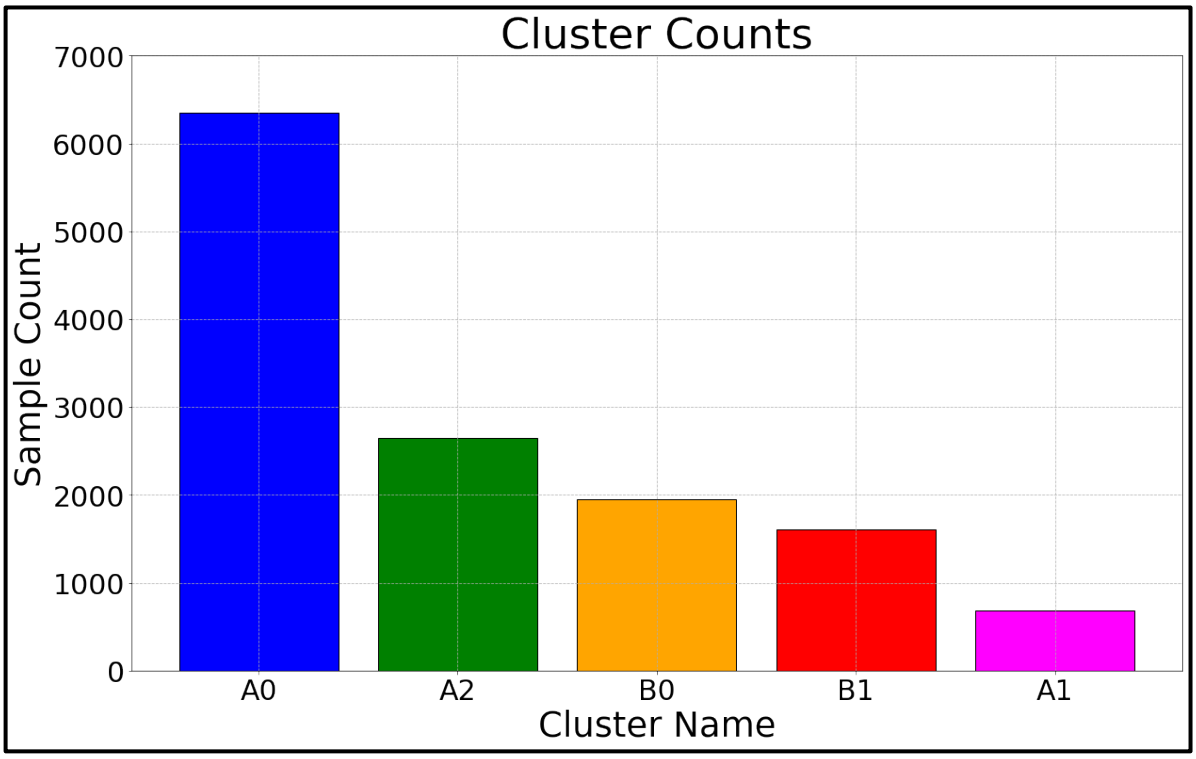

Fig. 12: Bar graph of the sample count within each of the sub-clusters generated through the unsupervised methods described in *Chapter 3*.

3.4 Feature Extraction and Selection

These clusters can be considered predicted lithologies, determined automatically by the unsupervised clustering algorithms. The samples, along with their respective cluster label, are returned to the individual wells. The scaled well logs were used in feature extraction, with respect to each cluster. The well logs were scaled as one continuous data set, so there was not an issue with varying scales between wells during feature extraction. Each feature extracted was done with respect to each of the 5 clusters. The first feature extracted from the clustered data is a binary type feature which answers: is *X* cluster in this well? This feature gave us an idea of the importance the presence of a particular rock type is for differentiating HWPs and LWPs. The next feature extracted counts the number of samples

which belong to each cluster for a given well. Similar to the first feature, the cluster count feature provided us with an idea on the frequency a predicted unsupervised lithology could be used to differentiate HWPs and LWPs. Along a similar line of reasoning, a third feature which was extracted checked if there are 30 samples or more for a given cluster in a well.

If there were 30 samples of a given cluster in a well, similar statistical parameters to chapter 2 were extracted per each of the 5 well log for that cluster. However, the interquartile range (IQR) was implemented in statistical parameter extraction for the chapter 3 workflow. Two variants of IQR were extracted. One IQR used the range between the 75^{th} percentile and the 25^{th} percentile, while the other variant calculates the range between 95^{th} percentile and the 2^{nd} percentile.

Due to the fact the depth data was preserved for each well, it was also possible to extract features based on the depth. The outlier removal process destroyed some samples of data from various wells. Therefore, some depth samples have been lost from well to well. With this in mind, the thickness of depth intervals must be iterable across all wells in the data set. To accomplish this, the minimum depth is subtracted from the maximum depth for every well and the difference is divided into 5 separate depth intervals. From the newly defined depth intervals, the sample count for each cluster was calculated for every interval and every well. The total feature extraction process for chapter 3's data-driven workflow resulted in a total of 336 features for training supervised models.

3.5 Training Supervised Algorithms

3.5.1 Feature Reduction

Similar to chapter 2's workflow, not all of these 336 features were useful to the supervised algorithms for classification. These features were reduced as in chapter 2's data-driven workflow. The reduction methods are similar to what was used in chapter 2, but the F-value is not utilized and a new parameter called mutual information (MI) was used in its place. The MI approach can be applied to both discrete and continuous variables, in the case of this described feature-space there were discrete and continuous features (Estévez et al., 2009). For a set of two continuous variables MI was calculated by examining the joint probability density function *p(x, y)*, and marginal probability density functions *p(x)* and *p(y)*. MI between *X* and *Y* is defined by *equation 9:*

$$I(X;Y) = \int \int p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)}\right) dxdy \qquad \text{(eq. 9)}$$

While MI between two discrete variables was calculated by a joint probability mass function *p(x, y)* and marginal probabilities *p(x)* and *p(y)* defined as (*eq. 10*) (Estévez et al., 2009):

$$I(X;Y) = \sum_{x \epsilon X} \sum_{y \epsilon Y} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)}\right) \qquad \text{(eq. 10)}$$

In summary, the MI algorithm calculated the dependency shared between two variables. A higher value indicates high dependency, while a low value indicates a more independent relationship between the features. For the purpose of feature selection, it was preferable to have features with low MI scores. Thus, to reduce the number of features being used for training a threshold is set for p-values from an ANOVA F-test and a threshold for MI score was also used. The reduction process reduced the feature set down to 5% of the data set.

3.5.2 Supervised Algorithms Utilized

The same three successful supervised models from chapter 2 were also used to differentiate HWPs and LWPs with the new features from unsupervised clustering. Interestingly, the models seemed to show differing preferences for the number of features used for training. For example, the KNN algorithm seemed to favor a lower feature count with more restrictive p-value thresholds than logistic regression or SVC.

3.6 Results and Interpretations

The results of supervised classification models, from features extracted from the clustered well logs, are shown for all three algorithms in *figures 13 – 15*. The first important note about the three distributions of MCC scores is that they all generate a median score of 0.90. This is an improvement for the KNN algorithm's performance in *Chapter 2*, which was originally scoring 0.84 median MCC. The second key observation from *figures 12 – 14* is that none of the three algorithms produce an MCC score less than 0.70, over 100 iterations. This means that even the worst possible training data from the cross-validation spits performs relatively well for all three algorithms. Increasing the scope of the data set, in terms of relative depth to the KOP, allows the user to gain a more comprehensive understanding of what is happening petrophysically within the rock formations.

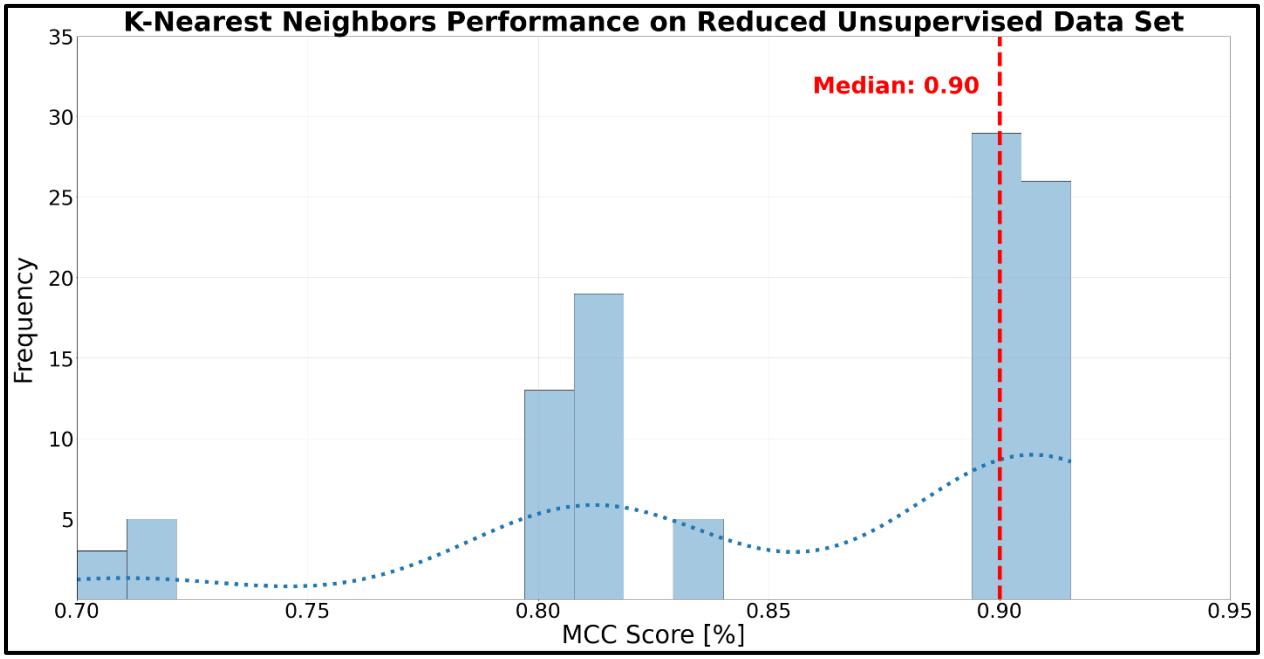

Fig. 13: Histogram of the performance of KNN algorithm on the reduced feature set from the unsupervised workflow for 100 cross-validation iterations. The median MCC score is represented by the red dashed line at 0.90.

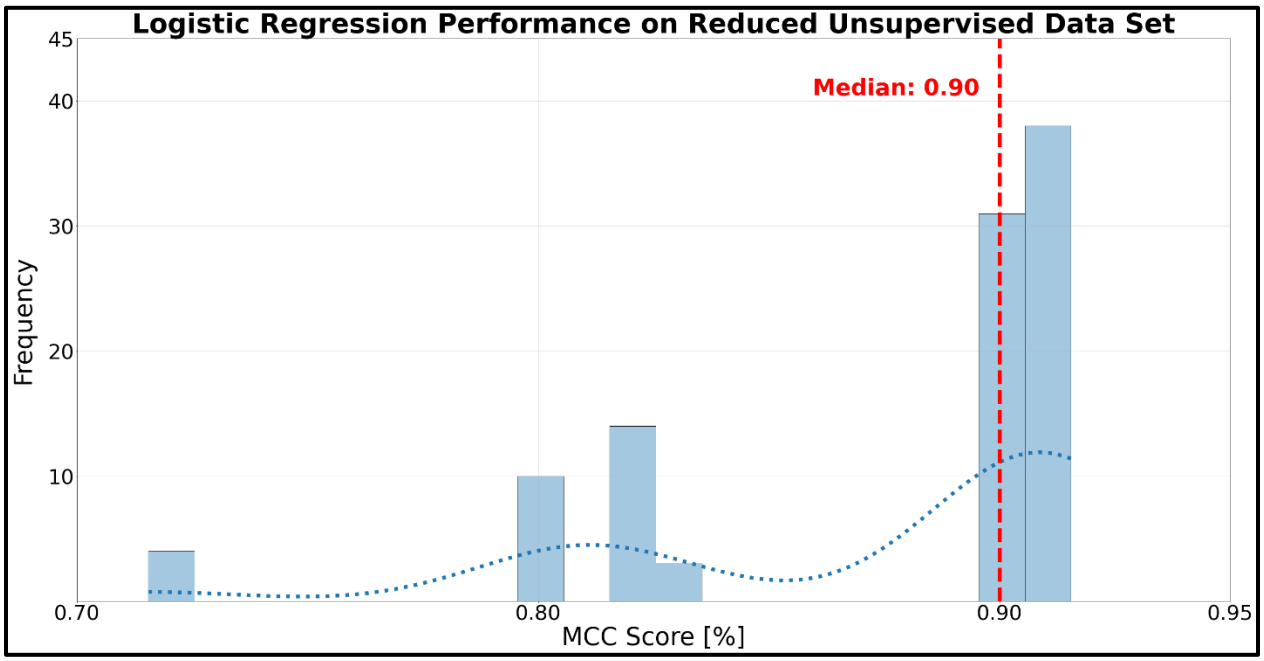

Fig. 14: Histogram of the performance of logistic regression algorithm on the reduced feature set from the unsupervised workflow for 100 cross-validation iterations. The median MCC score is represented by the red dashed line at 0.90.

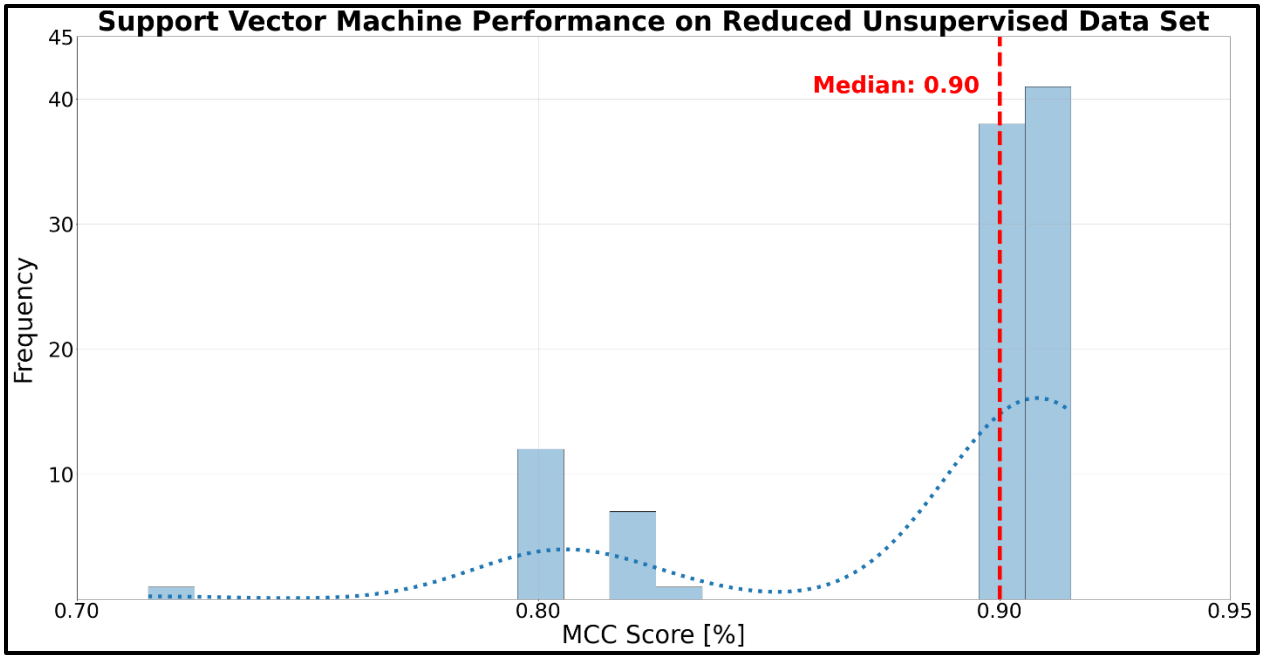

Fig. 15: Histogram of the performance of support vector machine classifier algorithm on the reduced feature set from the unsupervised workflow for 100 cross-validation iterations. The median MCC score is represented by the red dashed line at 0.90.

3.6.1 Ranked Features from Unsupervised Workflow

Due to the fact that no supervised algorithm was relevantly superior to the other two algorithms, it was appropriate to use models from all three supervised algorithms for feature ranking. The top 10 features for each algorithm were determined using a modified sampling version to the methods described in section *2.8*. In this unsupervised clustering workflow, 20 models generated from each of the 3 algorithms (K-Nearest neighbors, support vector machine, and logistic regression) were used for permutation testing and feature ranking. Upon calculation of the top 10 features for each algorithm, it was apparent that the overlap of significant features was very large between the three supervised algorithms. It was then possible to take these 20 models from all 3 supervised methods and

rank each algorithm's significant features individually first. Following the individual ranked significant feature generation for all three algorithms, the ranked features were then brought together for a cumulative ranking. There were 11 significantly informative features in total, across all methods, which are useful for differentiating HWPs and LWPs (*Table 2*).

Table 2: Cumulative ranked features for KNN, SVM, and logistic regression methods resulting from the prediction performance on Delaware basin reduced feature set(s). Features are ranked from most important (1) to least important (10), but still much more important than features not included in the list. The features are illustrated with the following structure: “Cluster alias” + “feature or well log + statistical parameter.” The “present” feature is a binary feature which checks if the cluster is present in a given well. “gr30” checks if there are at least 30 samples of the said cluster in a given well. The "icount#” feature counts the number of samples which belong to a cluster in a given depth interval with respect to the KOP (interval #).

Delaware Basin Cumulative Ranked Features for All 3 Supervised Methods	
Feature	**Rank**
A0_NPHI_Kurtosis	1
A1Present	2
B0_DPHI_Kurtosis	3
B1_GR_rms	4
B1icount3	5
B1_DPHI_rms	6
B1_GR_Variance & B1_ILD_mean	7
B1gr30	8
B1_ILD_median	9
B1_ILD_rms	10

3.6.2 Discussion of Ranked Features from Unsupervised Workflow

At a glance of the ranked features across all supervised methods, there are a few characteristics which are quite apparent. The first of which is that out of the 10 features found to be important across all models, 7 of them are related to the ‘B1’ cluster. The ‘B1’ lithology occupies the most rankings of the top 10 most informative features, but this cluster is still trumped statistically by features taken from A0, A1, and B0.

The kurtosis of the neutron porosity within lithology A0 is by far the most influential feature on all three algorithms. Prior to culmination of features into the cumulative ranked feature set, this feature was easily rank 1 for all three supervised methods. Why this cluster's NPHI curve's kurtosis is so significant needs to be further investigated. However, the fact that NPHI appears significant is not surprising. This well log is very sensitive to hydrogen atoms, which belong to water and hydrocarbon molecules in the case of the subsurface. As this well log device fires neutrons into the formation, the number of neutrons which return back to the device is monitored and logged. The number of neutrons which are captured in the formation is influenced greatly by the fluid which is saturating the pore space of the formation. Thus, the kurtosis of the distribution of NPHI within a cluster provides the insight that the way this well log reading, within the A0 lithology, is varying between HWPs and LWPs is so significantly different that the supervised ML algorithms have found it the most important factor in differentiating the well classes.

The second most important feature, which is again ranked as second most important for all models, is 'A1Present.' This is a binary-type feature which check whether or not the cluster named 'A1' is present in the well. Due to the unsupervised methods, the physicality of the lithology is somewhat diminished, from an interpretive sense, but for whatever reason the simple presence of one sample of this cluster in a well is quite statistically significant in differentiating HWPs and LWPs.

The third highest ranked feature is 'B0_DPHI_Kurtosis.' This feature only appears as significant in the SVM and logistic regression models, but since it is consistently ranked as 3rd for both models it is still considered quite significant across all three supervised methods. Again, this is another porosity log which is considered to be quite statistically

significant. The kurtosis component of both ‘A0_NPHI’ and ‘B0_DPHI’ proving to be quite statistically significant for differentiating well class is peculiar. This statistical parameter is related to the curvature of the distribution for the porosity well logs involved, which can be summarized as involving the number of samples which seem to stack up at or around one value.

The next 8 features are all related to the cluster ‘B1.’ All of these features are more subjective in terms of relative importance to each supervised methods, with some which do not appear significant at all for a given algorithm. However, all of them are considering some factor of the ‘B1’ cluster. It could be stated that due to the fact that the ‘B1’ cluster appears so frequently across all models at different levels of importance, its presence is perhaps is more important than the top 3 features. This of course is unlikely, due to the wide statistical significance of at least the top 2 features across all algorithms. Another observation one might make is that the ‘A2’ cluster does not appear to have any statistical significance whatsoever to the supervised models, as it does not appear once in the top ranked features.

3.7 Chapter III Conclusions

In chapter 3, a second data-driven workflow is proposed for separating HWPs from LWPs in a given set of wells. This workflow utilizes unsupervised clustering methods to assign well log samples a predicted lithology, based on similar characteristics between well log samples. 5 lithologies were determined using multi-layer unsupervised clustering for the well log data taken from the Delaware basin. From these 5 unique lithologies, features were extracted based on well log data for each lithology. These features were then used to train supervised ML algorithms to differentiate high water producing wells and low water

producing wells. Three algorithms display comparable success in differentiating the two well classes in this proposed workflow: K-nearest neighbors, support vector machine, and logistic regression. Each of these algorithms produced 100 models, using cross-validation methods, which result in histograms of MCC scores with median values of 0.90 for all three algorithms.

As performed in chapter 2, the most informative features in training models generated in chapter 3's workflow were ascertained using permutation testing methods. The top 3 most informative features to models generated across all 3 supervised algorithms are: the kurtosis of the neutron porosity (NPHI) of lithology A0, the presence of lithology A1, and the kurtosis of the density porosity (DPHI) of lithology B0. Out of the 5 lithologies generated by unsupervised ML clustering, four of them indicate significance when predicting whether or not a well will be a HWP or an LWP. The lower 7 of the top 10 most informative features all belong to lithology B1, which provides good indication that this lithology can be used as a key component for differentiating the two well classes.

CHAPTER IV

REVIEW OF WATER SATURATION EMPIRICAL ESTIMATION METHODS

4.1 Archie's Equation

One of the primary rock features associated with water production is water saturation, S_w. Over approximately 80 years of experimentation and development has gone into improving the accuracy of empirical estimations of S_w. In 1942, the work Archie performed in determining water saturation was published (Archie, 1942). In this work, Archie presented multiple relationships between resistivity, porosity, and water saturation which he was able to identify experimentally using a combination of well log data and core data from fully brine-saturated sandstone core samples. One key relation Archie identified was the relationship between the formation or rock resistivity and the resistivity of the brine saturating the rock (*eq. 11*):

$$R_o = FR_w \qquad \text{(eq. 11)}$$

Where R_o is the resistivity of the sand when all of its pores are filled with brine, R_w is the resistivity of the brine and F is a "formation resistivity factor (Archie, 1942)." The formation resistivity factor, F, was what Archie described as a function of the type and character of the formation being investigated, which varies with porosity and permeability of the reservoir rock. Upon further investigation comparing values of F to permeability and porosity values of sandstones, Archie determined that porosity, θ, had a more direct relationship to F related by the following *eq. 12*:

$$F = \frac{1}{\theta^m} \qquad \text{(eq. 12)}$$

Where Archie defined the value of *m* as the slope of the line found on the porosity vs. formation resistivity factor plots, such as *fig. 16.*

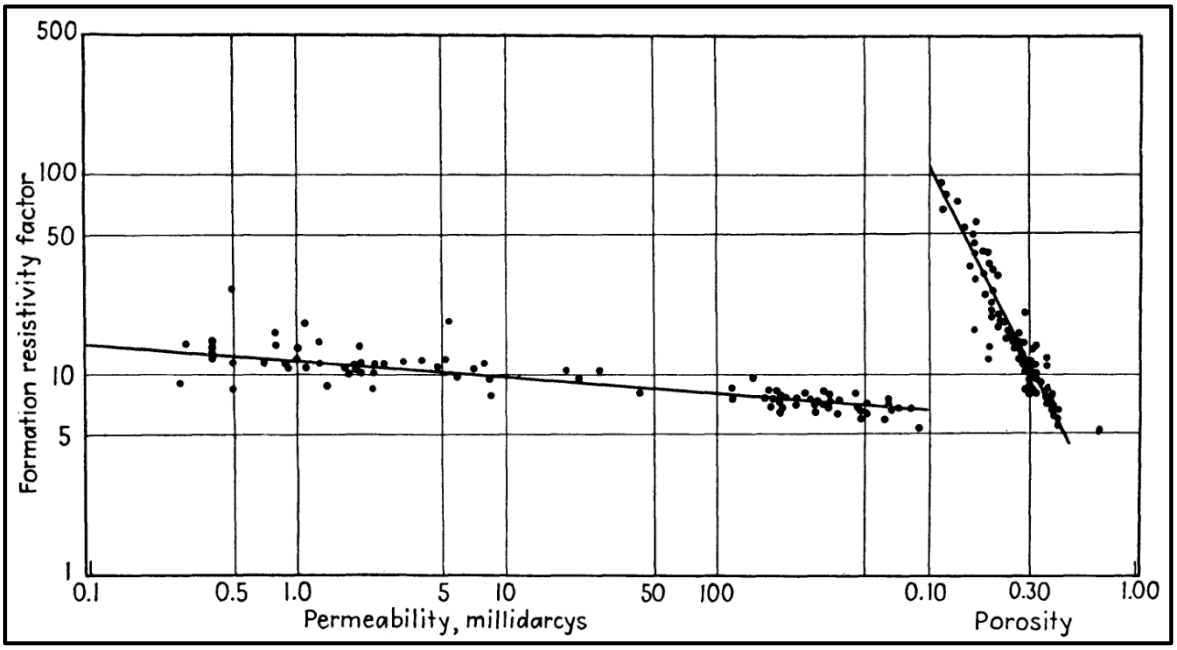

Fig. 16: Graphical representation of petrophysical data published in Archie's original 1942 paper. Samples taken from consolidated sandstone cores from the Gulf Coast (Archie, 1942).

Multiple investigators performed experiments studying the variation of resistivity of sands with varying water saturation percentages (Archie, 1942). These experiments were done by displacing the conductive water from the sands with non-conductive fluid. The resulting relationship they identified became the precursor to what is now known as Archie's equation, which even to this day is the most influential empirical relationship for the estimation of water saturation in a rock body (*eq. 13*).

$$S_w^{-n} = \frac{R_o}{R} \quad \text{or} \quad S_w^{-n} = \frac{FR_w}{R} \qquad \text{(eq. 13)}$$

Where *R* represents the apparent resistivity of the formation and *n* is referred to as the "saturation exponent" (Doveton, 2001).

4.2 Evolution of Archie's Equation

Later investigators replaced the numerator in *eq. 12* with a parameter named '*a'* (Winsauer et al., 1952). This parameter is polarizing among users of *eq. 14* as to what exactly '*a'* represents physically, if it represents any physical characteristic at all (Doveton, 2001). Many who argue that *a* does have a physical meaning, refer to it as the "tortuosity index." This factor *a* has transformed *eq. 13* into what is now most commonly referred to as Archie's Equation (*eq. 14*):

$$S_w = \sqrt[n]{\frac{a*R_w}{R*\theta^m}} \qquad \text{(eq. 14)}$$

This finalized version of Archie's equation boasts very accurate results for both resistivity and water saturation, if the rock being analyzed is a clean sandstone. In modern times and especially in onshore US, the amount of production stemming from clean sandstones has been dropping greatly. This means that Archie's equation cannot be applied to unconventional targets of the current day (Doveton, 2001).

Expanding upon this, many investigators have gone on to construct multiple models with varying assumptions. In the most general form, the majority of transformations of Archie's equation can be written as *eq. 15*:

$$\frac{1}{R_t} = \frac{S_w^2}{F*R_w} + X \qquad \text{(eq. 15)}$$

The *X* variable in *eq. 15* represents the influence which clay rock has on the system. Many models of this general scheme have been generated with varying results and assumptions

(Doveton, 2001; Bardon and Pied, 1969; Simandoux, 1963). Typically, these modified versions of Archie's equation require the calculation of properties such as the resistivity of shale or the volume of shale in the target interval.

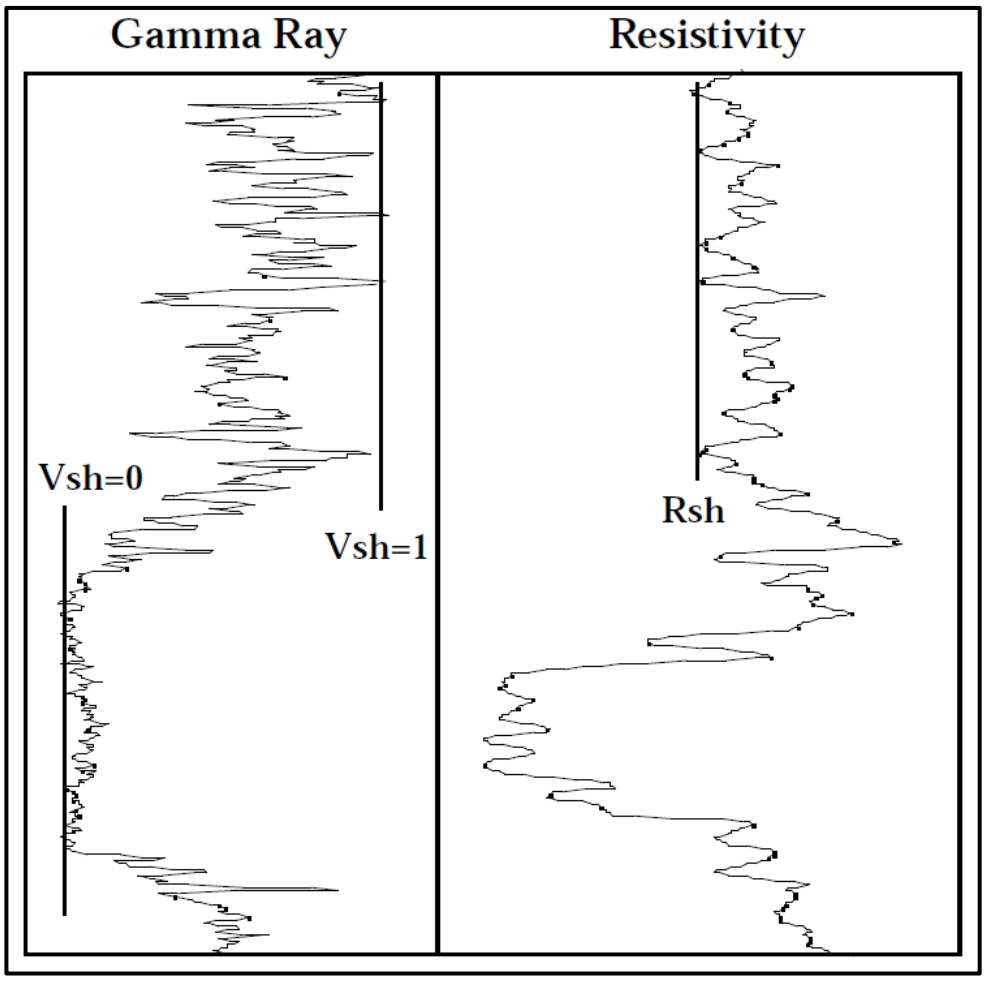

Fig. 17: Example of workflow used in (Doveton, 2001) to determine resistivity of shale, R_{sh} and the volume of shale, V_{sh} to be used for water saturation calculations in shaly sands.

As illustrated above, this method involves applying a cut-off line for 100% shale rocks in the gamma ray curve and then translating this interpretation over to the resistivity log for an approximation of the resistivity of the interpreted shales. This approach is intended for

shaley sandstones, not for pure shale target reservoirs water saturation approximations. Typically, the values for R_{sh} are taken from a shale interval separate from the target reservoir interval. This means that that not only are these shales likely different geologically from the target reservoir interval, but the values of R_{sh} are not going to be entirely transferable to the target interval and thus will result in slight differences in S_w than reality (Doveton, 2001).

4.3 Water Saturation in Shale Reservoirs

In 1972, another model branching off of the Archie's equation emerged from Schlumberger specifically to describe water saturation in shale formations (Zhang and Xu, 2016). This equation, the total shale model, adds the effect of the non-shale rock volume to the resistivity of water in the denominator of Archie's equation (*eq. 16*):

$$\frac{1}{R_t} = \frac{\theta^m S_w^n}{a * R_w * (1 - V_{sh})} + \frac{V_{sh} S_w}{R_{sh}} \qquad \text{(eq. 16)}$$

Where V_{sh} and R_{sh} are the volume of shale and resistivity of shale, respectively. In more modern times, petrophysicists have learned the influence that TOC (total organic carbon) content is having on the determination of water saturation percentages in shales (Zhang and Xu, 2016). An increase in TOC content has been shown to result in an increase in resistivity for shale rocks, in both core and conventional water saturation equations. Due to this relation, corrections must be applied to accurately estimate water saturation in shales. One method designed for gas producing shales separates the rock body into two separate categories: inorganic and organic. As a result of many studies illustrating positive correlation of gas content with TOC, it can be inferred that TOC has a negative correlation with water saturation (Zhang and Xu, 2016). Ultimately, this results in what Zhang et. al

has referred to as the "TOC correction method." This method uses graphical relationships of TOC and water saturations from core samples and conventional water saturation equations (*fig. 17*), which the authors consider any equations ranging from Archie's to Simandoux and beyond. This method is based on the difference between water saturation calculated from conventional equations, S_{w_con} and the observed difference in water saturation within organic matter of the shales, S_{w_d} (Zhang and Xu, 2016). The authors relate S_{w_d} directly to TOC content in the shale and estimate water saturation for a shale formation using *eq. 17*:

$$S_w = S_{w\_con} * (1 - \frac{TOC}{TOC_x}) \quad \text{(eq. 17)}$$

Where TOC_x was defined as a constant value of TOC (%) determined graphically by *Fig. 18.* The investigators stressed that this method may only be used when a linear relationship is shown between TOC and $(1-\frac{S_{w\_core}}{S_{w\_con}})$, where S_{w_core} is the water saturation determined from core samples.

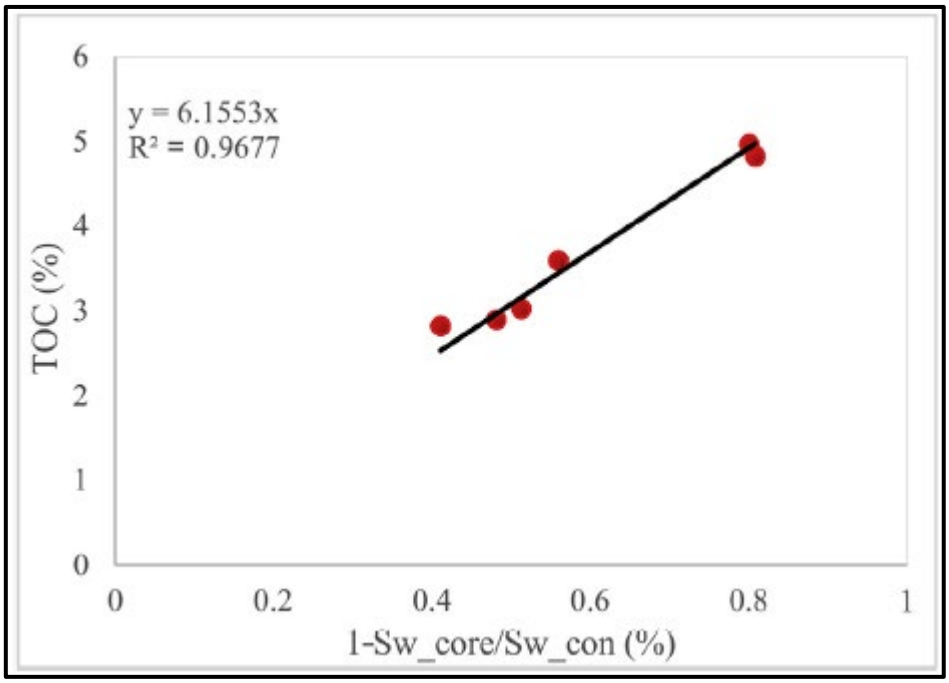

Fig. 18: Graphical representation of the relationship between TOC (%) and $S_{w\ core}/S_{w\ con}$ (Zhang and Xu, 2016).

4.4 Chapter IV Conclusions

Archie's equation was established through empirical experiments and published in 1942. This equation still functions to this day as the basis for the majority of water saturation calculations using well logs. The original Archie's equation assumed the rock being analyzed is a clean sandstone. As target reservoirs became more shaley over time, the equations used to calculate water saturation from well logs had to evolve over time. In more modern models, an estimation of the volume % of shale in a target reservoir and the resistivity of shale needed to be calculated for accurate water saturation estimations. Expanding on this idea, the different properties within a shale separating organic matrix and inorganic matrix have become increasingly important when calculating water

saturation in shale reservoirs. Thus, the influence of total organic content (TOC) must be taken into account when estimating water saturation.

CHAPTER V

CONCLUSIONS

The final products of this thesis project are the two proposed workflows described in chapters 2 & 3 of this manuscript. The intended goal of these workflows is to provide operators who are concerned with high water-cut wells with a quick and accurate determination of whether or not a well will produce a high amount of water relative to the target produced hydrocarbon. The workflows proposed use data which is already a common part of drilling and logging processes. This allows operators to incorporate both of these workflows seamlessly into their already existing standard production processes.

The first data-driven workflow described utilizes supervised machine learning algorithms, trained on well log data taken from 20 horizontal wells from the Delaware basin. The well log data was taken from 300 feet below the kick-off point and divided into six 50-ft bands. From these six bands 7 statistical parameters were extracted from all five well logs used in the data set, which resulted in 210 features in total per well. These features were reduced by 90% using an ANOVA F-test with relation to well class (high water producer and low water producer) and reduction based on collinearity amongst the features. The reduced feature set was then used to train a set of supervised machine learning algorithms, of which 3 proved the best at differentiating well class: K-Nearest neighbors, support vector machine, and logistic regression. Of these three best algorithms, the logistic regression performed the best based on the classifier evaluation metrics F1 score and Matthew's Correlation Coefficient (MCC). The logistic regression produced a score distribution over 100 cross-validation training iterations with a median F1 score and MCC score of 0.96 and 0.92, respectively. This workflow was also evaluated on two separate proof-of-concept data sets, one from the Fort Worth basin and

one from the Gulf Coast region. The median F1 and MCC scores over 100 cross-validation training iterations proved to be satisfactory for both outside data sets.

The second data-driven workflow discussed in chapter 3 of this work incorporates unsupervised methods to extract features, then train supervised models to differentiate well class. To accomplish this, 17 horizontal wells from the Delaware basin which have sufficient well log data are used. This workflow incorporates 500 total feet of well log data surrounding the kick-off point of the wellbore, 200 feet from above the kick-off point and 300 feet from below the kick-off point. This well log data is scaled, transformed, and purged of outlier data before being subject to the multilayer clustering methods. The data is then reduced by permutation testing, using silhouette score and agreement between K-Means and spectral clustering as a guide. Using all well log data in one data set, the unsupervised methods showed that the data was best split into two major clusters: cluster A and cluster B. Cluster A and B are the product of layer one of the multilayer clustering process used. Due to large number of samples in both clusters remaining, it was deemed suitable to further divide these two major clusters into sub-clusters. Using similar evaluation metrics as in layer one, cluster A was divided into 3 sub-clusters, while cluster B was divided into 2 sub-clusters. These 5 unique clusters generated represent different lithology types, purely generated by clustering algorithms based on similar characteristics between samples. From these lithologies generated, 336 features were extracted for each well based on lithology presence, statistical parameters of well logs specific to a cluster, frequency of cluster presence in a well, and cluster features based on distance from the kick-off point of each well. Similar to methods described Chapter 2, supervised methods were trained on a reduced set of these features to differentiate well classes. Three supervised algorithms produced very similar results. The

three best algorithms were again: K-nearest neighbors, support vector machine, and logistic regression. All three of these algorithms produce MCC score distributions with median values of 0.90.

In the fourth chapter of this thesis, literature regarding empirical water saturation estimation methods is reviewed. A brief history of the evolution of Archie's equation (Archie, 1942), which to this day is one of the most influential empirical relationships between resistivity of a rock formation and water saturation percentage. This simple relationship between petrophysical properties of clean sandstones and water saturation has slowly been expanded to more shaley rocks. Relationships such as the Simandoux equation and total shale model were developed from Archie's equation to account for the influence which clay content has on conductivity in shaley sandstone reservoir intervals. In recent studies the influence of TOC in shale reservoirs has become a key factor in determining water saturation percentages. The necessity to calculate TOC from shale targets has made core data from shales even more important than previous years to properly assess a shale's average TOC content.

CHAPTER VI

FUTURE WORK

The work done in this thesis can be expanded upon by examining well data from outside of Texas. As mentioned previously in the thesis, it is not required of oil and gas operators to report barrels of water produced in Texas. This fact leads to estimations of water production by Enverus using available data, such as well tests. Although these estimations methods have proven to be accurate, they are still approximations compared to thorough reporting of water production. There is an abundance of well log data available from the Delaware basin taken from New Mexico based wells, horizontal and vertical. New Mexico's governing body requires operators to report water produced from hydrocarbon wells, which presents potential to expand the workflows used in this thesis to data which is likely more accurate. The methods could be even extended to change the approach from a classification problem to a regression problem to predict water cut percentages from hydrocarbon wells.

The evolution of Archie's equation has also made it apparent how crucial to the workflows presented the expansion of data may be. As this thesis is focused on the production from unconventional reservoirs, it is most appropriate to have some inclusion of TOC data in the workflows to predict water production which has shown to have a strong correlation to water saturation in shalier reservoirs. If core data from shalier or shale reservoirs can be incorporated into the workflows discussed in chapters 2 and 3, there could be great improvement in the prediction of relative water production for unconventional wells in the data set.

REFERENCES

Archie, G.E., (1942). The electrical resistivity log as an aid in determining some reservoir characteristics. *Transactions of the AIME*, *146*(01), pp.54-62.

Bardon, C. and Pied, B., 1969, May. Formation Water Saturation in Shaly Sands, Paper Z. In *10th Annual Logging Symposium Transactions: Society of Professional Well Log Analysts* (pp. 21-19).

Bergstra, J. and Bengio, Y., 2012. Random search for hyper-parameter optimization. *Journal of machine learning research*, *13*(2).

Chicco, D. and Jurman, G., 2020. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics*, *21*(1), pp.1-13.

Cortes, C. and Vapnik, V., 1995. Support-vector networks. *Machine learning*, *20*(3), pp.273-297.

Doveton, J.H., 2001. All Models Are Wrong, but Some Models Are Useful:" Solving" the Simandoux Equation. In *J. of the International Association for Mathematical Geology Conference, Cancun, Mexico*.

Duman*, R., 2019, October. Permian Produced Water: Impact of Rising Handling Costs and Larger Water Cuts on Wolfcamp Growth. In *Unconventional Resources Technology Conference, Denver, Colorado, 22-24 July 2019* (pp. 4453-4460). Unconventional Resources Technology Conference (URTeC); Society of Exploration Geophysicists.

Estévez, P.A., Tesmer, M., Perez, C.A. and Zurada, J.M., 2009. Normalized mutual information feature selection. *IEEE Transactions on neural networks*, *20*(2), pp.189-201.

Guevara, J., Kormaksson, M., Zadrozny, B., Lu, L., Tolle, J., Croft, T., Wu, M., Limbeck, J. and Hohl, D., 2017. A data-driven workflow for predicting horizontal well production using vertical well logs. *arXiv preprint arXiv:1705.06556*.

Guyon, I. and Elisseeff, A., 2003. An introduction to variable and feature selection. *Journal of machine learning research*, *3*(Mar), pp.1157-1182.

Hajizadeh, Y., 2019. Machine learning in oil and gas; a SWOT analysis approach. *Journal of Petroleum Science and Engineering*, *176*, pp.661-663.

Hosmer Jr, D.W., Lemeshow, S. and Sturdivant, R.X., 2013. *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Jain, A., Nandakumar, K. and Ross, A., 2005. Score normalization in multimodal biometric systems. *Pattern recognition*, *38*(12), pp.2270-2285.

Khan, N.A., Engle, M., Dungan, B., Holguin, F.O., Xu, P. and Carroll, K.C., 2016. Volatile-organic molecular characterization of shale-oil produced water from the Permian Basin. *Chemosphere*, *148*, pp.126-136.

Liu, F.T., Ting, K.M. and Zhou, Z.H., 2008, December. Isolation forest. In *2008 eighth ieee international conference on data mining* (pp. 413-422). IEEE.

MacQueen, J., 1967, June. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, No. 14, pp. 281-297).

Male, F., 2019. Using a segregated flow model to forecast production of oil, gas, and water in shale oil plays. *Journal of Petroleum Science and Engineering*, *180*, pp.48-61.

Miah, M.I., Zendehboudi, S. and Ahmed, S., 2020. Log data-driven model and feature ranking for water saturation prediction using machine learning approach. *Journal of Petroleum Science and Engineering*, *194*, p.107291.

Mohamed, A., Hamdi, M.S. and Tahar, S., 2015, August. A machine learning approach for big data in oil and gas pipelines. In *2015 3rd International Conference on Future Internet of Things and Cloud* (pp. 585-590). IEEE.

Ng, A., Jordan, M. and Weiss, Y., 2001. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, *14*, pp.849-856.

Osogba, O., Misra, S. and Xu, C., 2020. Machine learning workflow to predict multi-target subsurface signals for the exploration of hydrocarbon and water. *Fuel*, *278*, p.118357.

Pesarin, F. and Salmaso, L., 2010. The permutation testing approach: a review. *Statistica*, *70*(4), pp.481-509.

Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, pp.53-65.

Scanlon, B.R., Reedy, R.C., Male, F. and Walsh, M., 2017. Water issues related to transitioning from conventional to unconventional oil production in the Permian Basin. *Environmental science & technology*, *51*(18), pp.10903-10912.

Simandoux, P., (1963). Dielectric measurements in porous media and application to shaly formation. *Rev. L'Institut Français du P'etrole 18*, 193–215.

Stone, M., 1974. Cross-validation and multinomial prediction. *Biometrika*, *61*(3), pp.509-515.

Visa, S., Ramsay, B., Ralescu, A.L. and Van Der Knaap, E., 2011. Confusion Matrix-based Feature Selection. *MAICS*, *710*, pp.120-127.

Wagstaff, K., Cardie, C., Rogers, S. and Schroedl, S., 2001, June. Constrained k-means clustering with background knowledge. In *Icml* (Vol. 1, pp. 577-584).

Winsauer, W.O., Shearin, H.M., Masson, P.H. and Williams, M., 1952. Resistivity of brine-saturated sands in relation to pore geometry. *AAPG bulletin*, *36*(2), pp.253-277.

Wu, Y., Misra, S., Sondergeld, C., Curtis, M. and Jernigen, J., 2019. Machine learning for locating organic matter and pores in scanning electron microscopy images of organic-rich shales. *Fuel*, *253*, pp.662-676.

Yeo, I.K. and Johnson, R.A., 2000. A new family of power transformations to improve normality or symmetry. *Biometrika*, *87*(4), pp.954-959.

Zhang, B. and Xu, J., 2016. Methods for the evaluation of water saturation considering TOC in shale reservoirs. *Journal of Natural Gas Science and Engineering*, *36*, pp.800-810.

APPENDIX

A DATA COLLECTION AND PREPARATION

All data used in this project was taken from Enverus, a database for oil and gas data. The number of wells, or sample count, and types of wells included in the data sets utilized has varied greatly over several iterations. The set which was represented in this thesis is the final version, which has had the best results for prediction of relative water production. Primarily through a series of trial and error, it had become apparent that the more homogeneous the samples in each data set are, the better the machine learning (ML) algorithms were able to differentiate high water producers (HWPs) and low water producers (LWPs). Geologic heterogeneity was found to be significant by attempting to use various mixtures of wells with varying production types, with the final data set utilizing only oil-producing wells in the Delaware basin. The varying production types of these horizontal wells can be interpreted as some relative change in geology which made petrophysical signatures vary slightly, which may contribute to less accurate predictions from ML algorithms. Another feature of the data set that made a major difference is the presence of the kick-off point (KOP) within the analyzed logged interval for each well. In this analysis, we define the KOP as the point of the wellbore which it begins to transition from vertical into a horizontal wellbore. Through this analysis the region surrounding the KOP has demonstrated statistical significance in differentiating the two well classes.