Rocks Lithofacies Classification Using Machine Learning

1. Background and dataset

1.1 Background

In oil & gas exploration one of the crucial risk factors is reservoir quality. High quality reservoirs are the ones with a large amount of pores and good permeability (permeability measures how easy oil or gas flows out of the pores). Formations in different depositional environments (facies) have different levels of porosity and permeability. For example, sandstones are better reservoirs than siltstones because they have more pores and higher permeability. Therefore obtaining lithofacies information is essential in reservoir evaluation.

Accurate lithology data can be obtained from core sampling during drilling. However, this procedure requires a special tool and is extremely expensive. A common approach used by most oil companies is to have an experienced geologist interprete lithofacies using wireline log measurements, for example Gammaray log (GR), Density logs, Neutron logs, Sonic logs etc.. However, the drawback of this approach is obvious - it's time-consuming and the interpreter needs to be highly skilled. Machine learning, to be more specific, supervised machine learning naturally becomes a great tool here as it can automate lithofacies classification of the new wells by learning from the interpretations manually done by geologists for the existing wells.

In this project we will investigate how different machine learning techniques work in lithofacies classification. We will use one well as the "test dataset", and try to build a model based on the logs from all the other wells ("training dataset"), and use the derived model to predict lithofacies for the test well.

1.2 Dataset

This dataset comes from 10 wells in the Hugoton and Panoma Fields in North America. Each well contains 8 logs as listed below. Logs 1-5 are direct measurements or derived logs, 6 and 7 are indicators derived from geological information. The last one **"**Facies**"** is the lithofacies interpreted by geoscientists and the target feature in this study.

1) GR: measure natural radioactivity.
2) ILD_log10: measures rock resistivity or conductivity; displays in a 10-based logarithm scale.

3) PE: photoelectric effect log. A supplementary measurement from density tool. Less sensitive to porosity than density log, but more sensitive to minerals.
4) DeltaPHI: Phi is a porosity index in petrophysics. Density-neutron porosity difference
5) PNHIND: Average of neutron and density log.
6) NM_M:nonmarine-marine indicator
7) RELPOS: relative position
8) Facies: interpreted litho facies based on data from 1-7.

For the Facies column, there are 9 lithofacies values presented as integers in the dataset. The mapping is as following:

1 - SS: Nonmarine sandstone
2 - CSiS: Nonmarine coarse siltstone
3 - FSiS: Nonmarine fine siltstone
4 - SiSH: Marine siltstone and shale
5 - MS: Mudstone (limestone)
6 - WS: Wackestone (limestone)
7 - D: Dolomite
8 - PS: Packstone-grainstone (limestone)
9 - BS: Phylloid-algal bafflestone (limestone)

One note to point out is that change of the depositional environment is often transitional. For example, we may see sediments gradually change from coarse to finer grains without a sharp change. This is seen in this dataset. For example, facies 2 (Coarse Siltstone) usually occurs next to facies 1 (Sandstone, which is coarser than siltstone) or facies 3 (Fine Siltstone). The gradual and subtle changes may cause ambiguity in facies classification. This will be considered in the evaluation of model performance later in the document.

Key data source: Data source is from Kaggle:
https://www.kaggle.com/datasets/imeintanis/well-log-facies-dataset

2. Data visualization and wrangling

The table below provides a snapshot of the data. Each column is a log curve, or a feature in the language of data science. Each row gives the features values at the corresponding depth for a particular well. The dataset is organized by “Well Name”, and ordered by “Depth” from shallow to deep within each well.

	Facies	Formation	Well Name	Depth	GR	ILD_log10	DeltaPHI	PHIND	PE	NM_M	RELPOS
0	3	A1 SH	SHRIMPLIN	2793.0	77.45	0.664	9.9	11.915	4.6	1	1.000
1	3	A1 SH	SHRIMPLIN	2793.5	78.26	0.661	14.2	12.565	4.1	1	0.979
2	3	A1 SH	SHRIMPLIN	2794.0	79.05	0.658	14.8	13.050	3.6	1	0.957
3	3	A1 SH	SHRIMPLIN	2794.5	86.10	0.655	13.9	13.115	3.5	1	0.936
4	3	A1 SH	SHRIMPLIN	2795.0	74.58	0.647	13.5	13.300	3.4	1	0.915

2.1 Visualization

As log data are recorded by depth for each well, it's a common practice to plot each log curve in depth. Figure 1 displays the data for "SHANKLE" well. Shown from left to right are GR, ILD_log10, DeltaPHI, PHIND, PE, NM_M, RELPOS, and color-coded target feature Facies. NM_M has two values only: 1 for non-marine and 2 for marine. RELPOS is the relative position in a linear fashion from one facies to the next. Data from SHANKLE looks normal by visualizing it.

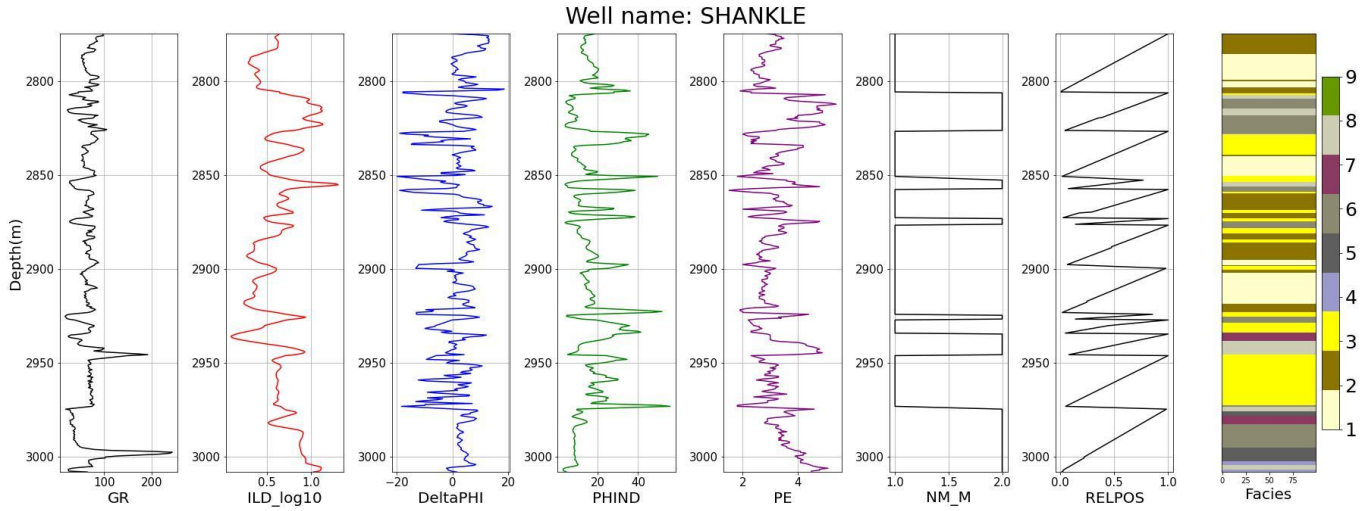

Figure 1: Log data plotted in depth for well "SHANKLE"

From a quick visualization of all 10 wells we noticed that well "RECRUITE_F9' has suspicious data in all features (Figure 2) for reasons we are not aware of. This well should be removed from the dataset.

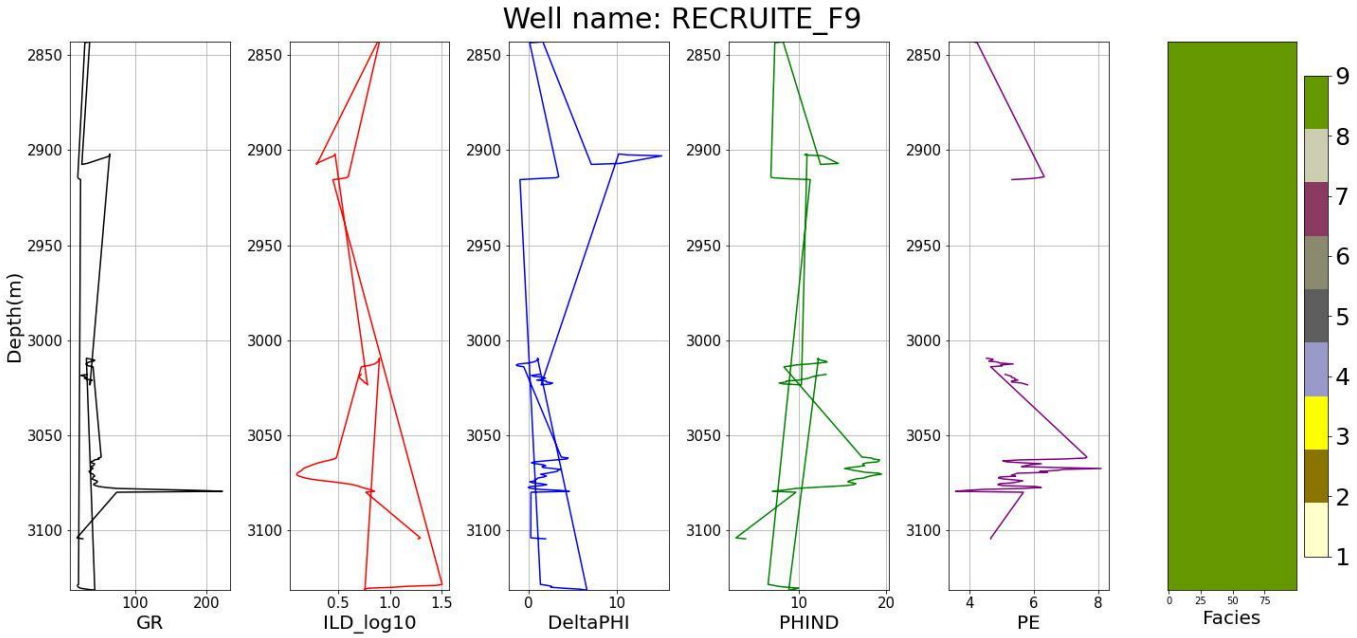

Figure 2: Data from well "RECRUITE_F9" looks suspicious.

What we also discovered from visualization was that the "PE" feature is missing in two wells "ALEXANDER D" and "KIMZEY A", as seen in Figure 3 for "ALEXANDER D'. How we should handle this missing information will be discussed in the data pre-processing step.

Conclusions from visualization: other than the missing PE in two wells and poor data quality in one well, the rest of the data looks good.

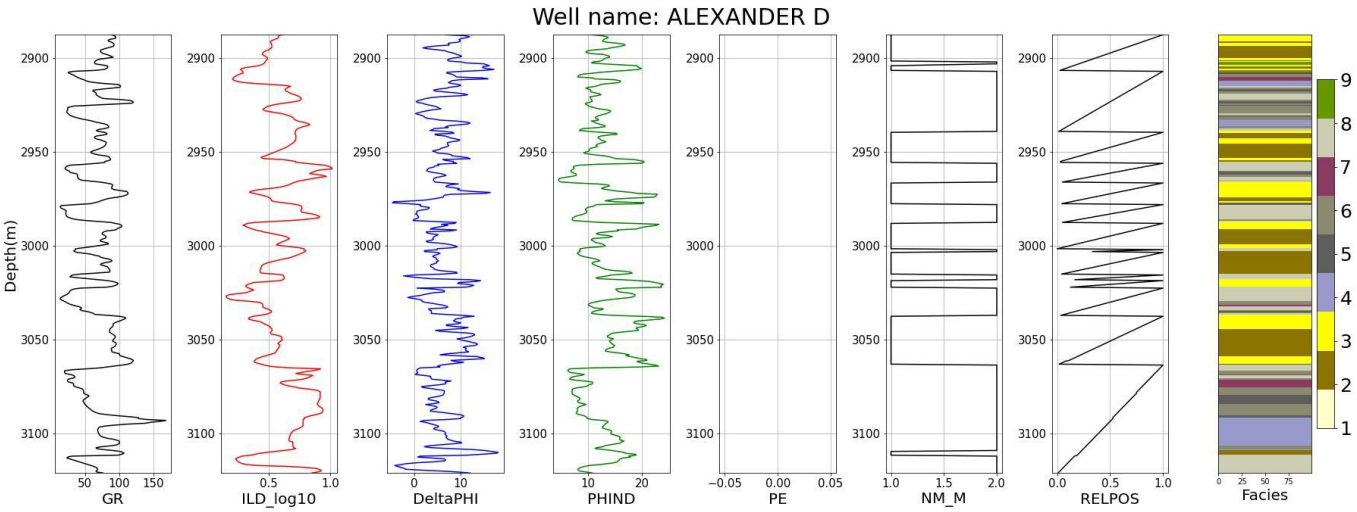

Figure 3: "PE" feature is missing from well "ALEXANDER D"

2.2 Data wrangling

In this step, we first removed the data from well ‘RECRUITE_F9”. Next we used the LabelEncoder function in sklearn to convert “Formation” names from strings to numeric values, and created a new numeric column for feature “Well Name”. This was done because some procedures in EDA and modeling require all features to be in numeric format.

With the remaining 9 wells, the distribution of the Facies values is shown in Figure 4. Facies 2, 3 , 6 and 8 have the most presence. Facies 2 and 3 are both siltstone while 2 is coarser than 3. Facies 6 and 8 are both limestone and the difference between them is how much mud or grains in the rocks. They are considered adjacent facies with small differences in their characteris.

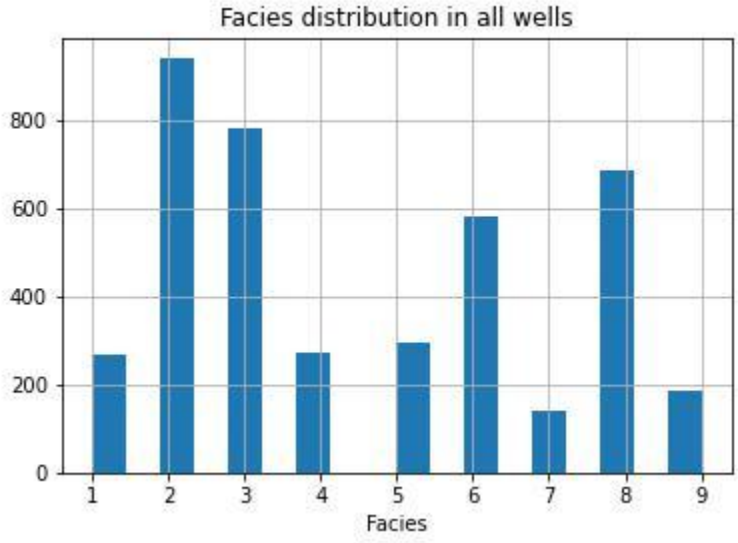

Figure 4: Distribution of the values of target feature “Facies” in the remaining 9 wells.

3. Exploratory data analysis and pre-processing

3.1 EDA

In this step, our main task was to handle the missing PE feature in “ALEXANDER D” and “KIMZEY A” wells. We had three options: drop these two wells entirely, drop PE data from all wells, or impute PE for these two wells. In order to choose the best approach it was essential to first investigate the importance of PE in Facies prediction. A quick assessment can be done using the heatmap to plot the correlation between features. From Figure 5 we can see that Facies and PE are highly correlated, implying that we should try to impute PE.

One way of imputing PE is to use other features that have strong correlation with PE. The heatmap shows that among the 4 continuous logs GR, ILD_log10, PHIND and DeltaPHI, PHIND has the highest correlation with PE. So one way of estimating PE is to model it using PHIND data. This will be discussed in the data pre-processing section, along with two other approaches.

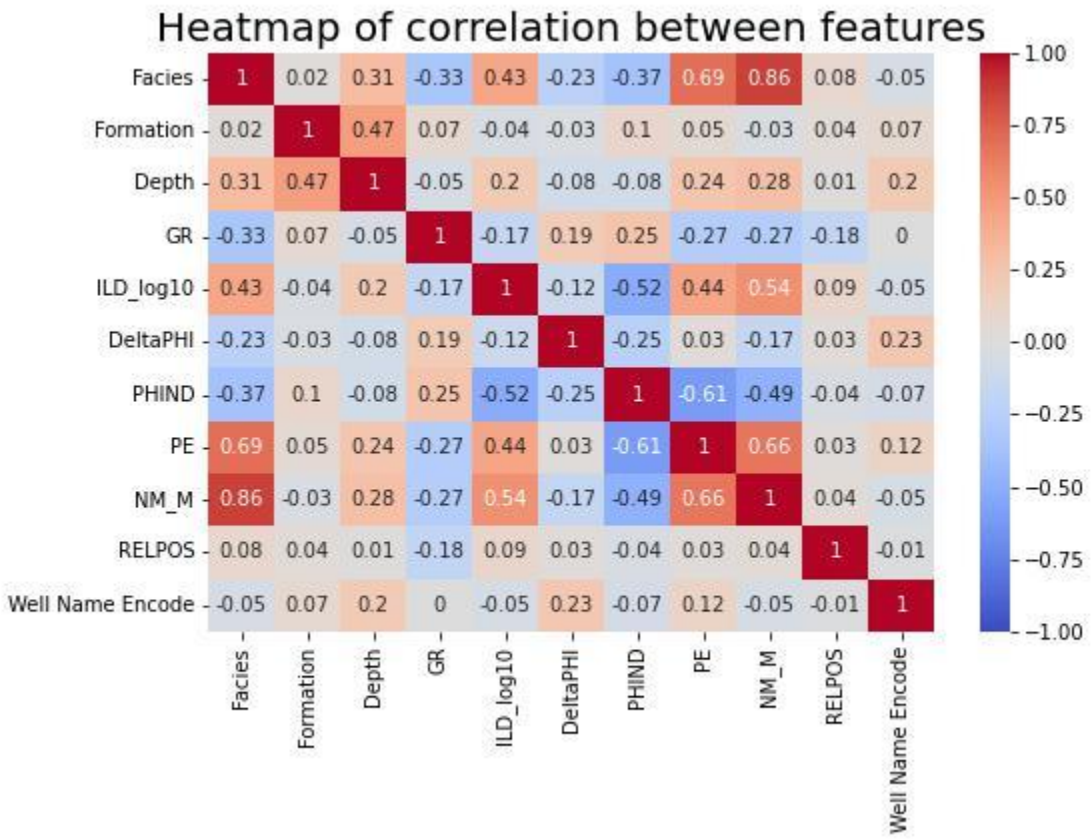

Figure 5: Heatmap showing correlation between various features.

3.2 Data pre-processing

In this step we will investigate the best way of estimating PE for “ALEXANDER D” and “KIMZEY A”. We will compare three methods:

1) polynomial regression using PHIND
2) multilayer perceptron regressor using all features
3) KNN in “fancyimpute” package

Among the 9 wells, 7 have PE data. We split the data from these 7 wells into a training set and a test set. The same split was used in all three models. Modeling results were evaluated using the RMSE and R2 scores. Figure 6 shows the scatter plots of the true vs. predicted PE values for the test set, and the scores of each model. Note that the scatter plots are for the test set and the scores are for the training set (test set gives similar scores). Among the three models, KNN from the “fancyimpute” package gave the lowest error and best match on the scatter plot, and was chosen as the final method for PE estimation. Figure 7 plots the data from “ALEXANDER D” well after imputation.

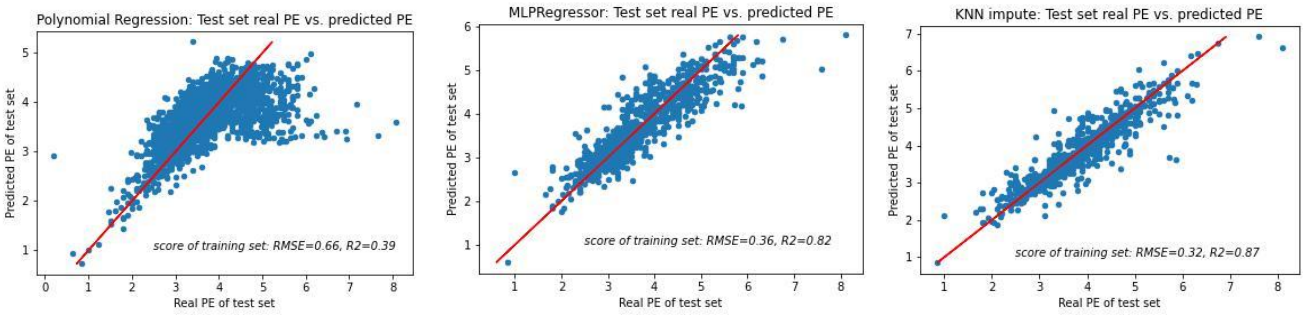

Figure 6: Comparison of three PE imputation methods (polynomial regression, MLPR, KNN). Scatter plots are real vs. predicted PE data for the test dataset randomly chosen from the 7 wells.

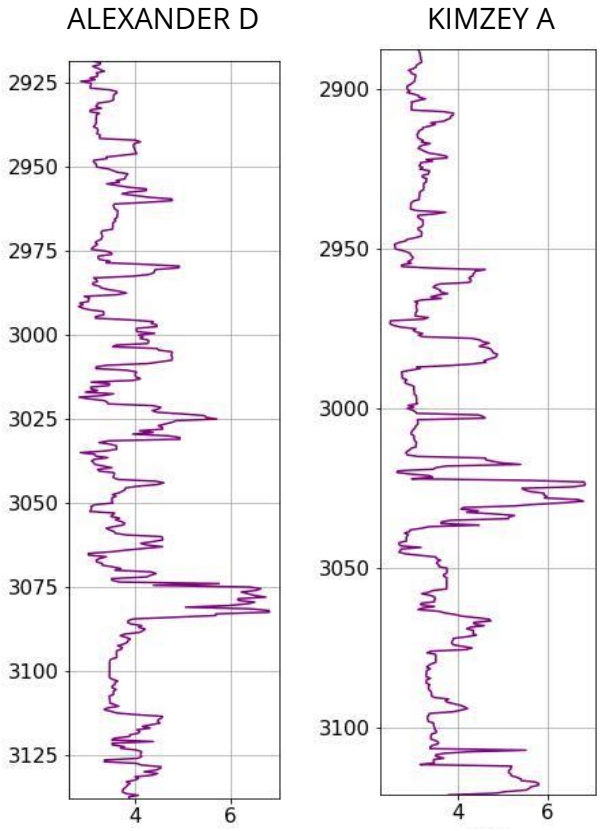

Figure 7: Imputed PE for "ALEXANDER D" and "KIMZEY A"

3.3 Features selection

The dataset was ready, and the next step was to figure out what features to be included in the modeling. From the heatmap in Figure 5 we can see the target feature "Facies" has little correlation with "Formation", "Well Name" and "RELPOS", therefore these features were dropped from the input data. Data from 9 wells and 7 features was used in the subsequent modeling process.

4. Modeling

In this stage we tested three different classifiers, evaluated their performance and chose the best model on the test well.

4.1 Modeling performance

Three classifiers were tested and compared: Random Forest, KNN, and SVM. Cross-validation and hyperparameter grid search were applied to achieve the optimal parameters and reliable performance evaluation for each model. The table below compares the modeling output.

Accuracy	Random Forest	KNN	SVM
Cross-validation accuracy	0.53	0.45	0.5

Out of the three classifiers Random Forest has the best performance and was chosen as the final model for this particular dataset. Its average accuracy under cross-validation is 0.53. We will further discuss model accuracy later in this chapter.

4.2 Training and test dataset split

In this study we wanted to leave 1 well as the test well and use the other 8 wells for model training. As not all facies are present in every well, we should try to avoid the situation where the test well has a facies value that has no or little presence in the other 8 wells. After examining Facies distribution in each well, “SHANKLE” was chosen as the test well.

4.2 Test well output

As the final step, the Random Forest model was constructed using the best parameters given from the hyperparameters grid search, and applied to the test well “SHANKLE”. Figure 8 shows the comparison of real and predicted Facies for this well. The output accuracy is 0.55, which is consistent with the cross-validation accuracy of the Random Forest model.

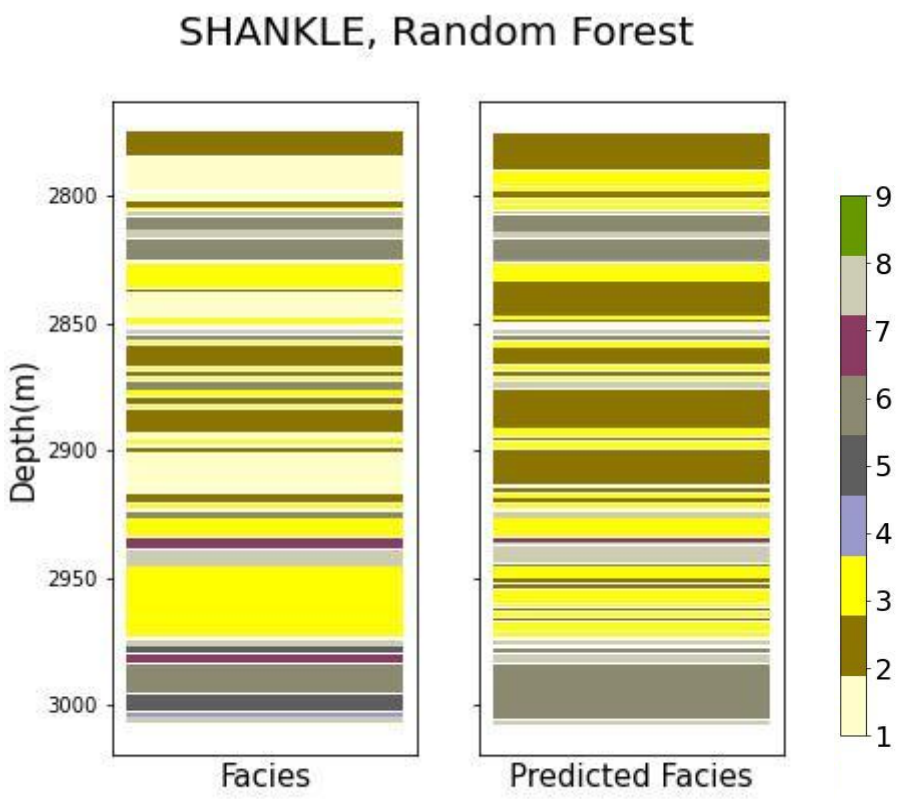

Figure 8: True Facies and predicted Facies for test well "SHANKLE"

The model accuracy is low compared with many machine learning applications we've seen in other fields. To understand the result better, we took a further look at how the mis-classifications were distributed. Figure 9 displays the confusion matrix of the modeling output with the test well. It shows for each facies how many data points are classified correctly or misclassified. It was noticed that even though the classification accuracy is only 0.55, the majority of the errors were made in mislabeling with the neighboring facies. Take the example of true Facies value 3. Among all data points, 82 were predicted correctly, 35 were mislabeled as Facies 2.

As we mentioned in previous sessions, a transition zone is commonly seen between two adjacent facies, and there is often no clear separation. In addition, both Facies 2 and 3 are siltstones which are fine grains sediments. The difference between the sub-classification of coarse and fine siltstone is not significant. So it is understandable that machine learning algorithms were not able to accurately separate these two features.

If we could use some tolerance in mislabeling with the neighboring facies here and re-calculate the prediction accuracy, it increases from 0.55 to 0.91, which implies that machine learning is able to separate facies with large difference.

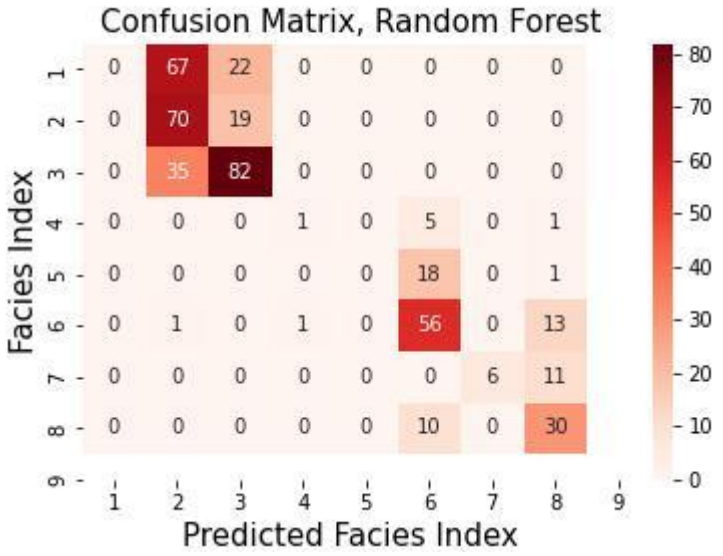

Figure 9: Confusion matrix for test well "SHANKLE" modeling output

5. Summary and future work

Facies classification from the gamma ray, sonic and density logs using machine learning techniques were examined with the dataset in the Hugoton and Panoma Fields in Kansas. The classification accuracy is in the range of 50's%. The errors in the results lie mainly in the mis-labeling with the neighboring facies, which is not entirely surprising because the sediments litho-facies often gradually transition from one facies to another, and sometimes the adjacent facies do not have significant difference. Although machine learning is not able to accurately separate adjacent rock facies, it classifies the non-adjacent facies quite well, and can be used as a good start point for the geoscientists in their interpretation workflow.

In this project three models were tested. We are confident that there are models that would perform better with this dataset, and people are encouraged to explore more and make machine learning a more useful tool in facies classification.