Using Machine Learning to Predict Lithostratigraphic Facies

Advikar Ananthkumar

Acton Boxborough High School

ABSTRACT

This paper aims to use rapidly growing machine learning applications in geology to predict vertical layers in rock based on properties. These layers in rock with similar chemical and physical properties are referred to as facies. Understanding the underlying strata and various facies informs geologists about the nature of a particular area. The order and nature of the layers in the ground can represent both how the location formed, as well as its evolution over time. This paper takes commonly analyzed wells from a block in the Dutch sector of the North Sea and shows methodology in selected particular models and parameters for prediction. Visual representation of the parameters allows for influence on the facies to be determined. My approach filters through extraneous properties and applies a Butterworth low-pass filter. Because depth is a continuous data parameter that cannot be pieced apart for training data, splitting the training data was an obstacle. However, this problem was circumvented by using a stratified k-fold split. Six different models of supervised learning were directly compared both visually and analytically. Results from these comparisons from the F02-1 well indicate that a K-Nearest-Neighbors model is most accurate and should be used by lithostratigraphic drillers. Results on the test data yielded a prediction accuracy of 99%, but prediction accuracy is yet to be extensively applied to other wells. Finally, a visual reconstruction of the facies of a nearby F02-3 well presents the results of the application and reveals the geographic history of the North Sea.

INTRODUCTION

Although the conjunction of machine learning and geology is relatively bare, geologic mapping is one of the few explored areas of application. The specified facies of given rocks is key in physical and historic characteristics. The knowledge of these facies provides information on the composition, environment, and geologic record of the underlying strata. The relationship between a facies and its composition and environment can be seen through both sedimentary and metamorphic facies. This paper aims to predict the facies of an area given well data.

Sedimentary facies form under certain conditions of sedimentation, reflecting a particular process or environment. Different adjacent facies represent distinct depositional environments and compositions. To classify distinct facies groups, unique characteristics must be seen in the description of composition, texture, sedimentary structures, bedding geometry, nature of bedding contact, fossil content, and color. Metamorphic facies are the set of mineral assemblages in metamorphic rocks formed under similar pressures and temperatures. Similar to sedimentary facies, metamorphic facies provide stratigraphic information. The different metamorphic facies are defined by the mineral composition of a rock. Both sedimentary and metamorphic facies provide invaluable information to geomappers, drillers, etc. If facies were able to be predicted by chemical and physical properties, the aforementioned information on the environment and earth could be learned.

A visual mapping of the vertical layers of rock can help with interpretations of the geologic evolution of an area. Firstly, the law of superposition provides the framework for which layers in the ground are evaluated with. In its plainest form, it states that in undeformed stratigraphic sequences, the oldest strata will lie at the bottom of the sequence, while newer material stacks upon the surface to form new deposits over time. This allows us to derive age from depth since deposition flows in a stack-like manner. Walther's law of facies, or simply Walther's law, named after Johannes Walther, states that the vertical succession of facies reflects lateral changes in the environment. This reflects the vertical stratigraphic succession that typifies marine transgressions and regressions. These changes occur when for example, the water level shifts, laterally shifting erosion and deposition on the bank. These laws allow us to map together unconformities in the terrain, which hint at geological evolutionary events that changed the local geography over time. An unconformity is a buried erosional or non-depositional surface separating two rock masses or strata of different ages, indicating that sediment deposition was not continuous. Within the general label of unconformity, geologists can dissect more information. A disconformity is revealed when there is an unconformity between parallel layers of sedimentary facies. A nonconformity exists between sedimentary and metamorphic or igneous rocks when the sedimentary rock lies above and was deposited on the pre-existing and eroded metamorphic or igneous rock. An angular unconformity is where horizontal parallel strata of sedimentary rock are deposited on tilted and eroded later.

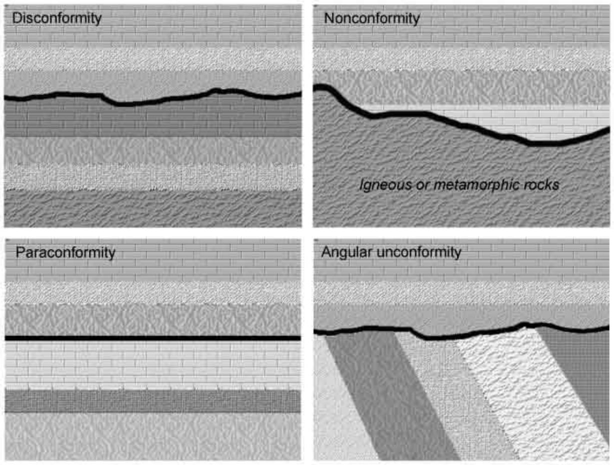


Figure 1: Unconformity types (Disconformity, Nonconformity, Paraconformity, Angular unconformity) visually shown (Source: Department of Earth, Ocean and Atmospheric Sciences, University of British Columbia)

This challenge of predicting facies given well data begs examination of the type of data and crisis. The well data which will be outlined further was strictly numerical and focused on objective measurable properties of the well extractions. Because the particular area of the F-01 was already studied, the facies are known. This makes this problem best solved by supervised learning, in which the data set is already tagged with correct values.

BACKGROUND

A good amount of literature has been published on lithology and facies prediction using machine learning. Thomas Martin, Ross Mayer, and Zane Jobe tackled the same problem in their paper in a METHODS article published in Frontiers in Earth Science. The approach taken by Martin et al followed a similar supervised learning model in which all the correct facies were manually labeled. The data used by Martin et al investigated Paleocene deep-marine strata within Quadrants 204 and 205 of the Faroe-Shetland Basin, West of Shetland, United Kingdom. A potential issue with this data set is that the geologic rifting of the area during the Devonian and Mesozoic periods applies complications in the labeling of the facies. Rifting can be associated with contact metamorphism along the sheer lines, chemically altering the facies horizontally. This would be an issue when labeling, as homogenous horizontal strata would be ideal. Furthermore, labeling was determined by core-color images. Also, color images can provide insight into the data, and small changes in lighting or interpretation of color can offset possible conclusions. For example, the interpretation of a well core could be either laminated sandstone or interbedded sandstone and mudstone depending on the lighting. The article was able to achieve maximum accuracy of 35.7% (nine training wells) and minimum accuracy of 18.2%.

A LinkedIn article written by Yohanes Nuwara outlined his full process from data cleanup to results. This article heavily influenced this paper and was generally followed for almost all steps. Nuwara uses data from the Netherlands F3 Block, which I also chose to use over the Faroe-Shetland and Kansas well data logs. A key difference between Nuwara's article and Martin et al. is the use of a low-pass Butterworth filter. Nuwara explains the benefits and implementation, which I used exactly. For the model, Nuwara uses KNN (K-Nearest-Neighbors) but never discusses why a KNN was specifically chosen. For this reason, I decided to examine multiple models with the same data used by Nuwara to determine if there was possibly a better model. The data, which will be examined in the next section, was available thanks to Nuwara's article. In the article, Nuwara can achieve an accuracy of 98%, which leaves very little room for improvement. However, my goal was to take Nuwara's methods and further improve them.

DATASET

The data used for this project comes from the Netherlands F3 block. This dataset is a seismic survey of approximately 384 km2 in the Dutch offshore portion of the Central Graben basin, The formation of The Central Graben Basin follows the opening of the North Atlantic and posterior division of Pangea. This area contains ten main lithostratigraphic groups: The Carboniferous, Lower-Upper Rotliegend, Zechstein, Germanic Trias, Altena, Schieland, Scruff, Niedersachsen, Rijnland, and Chalk Group. This data is publicly available and tends to be labeled as seismic data.

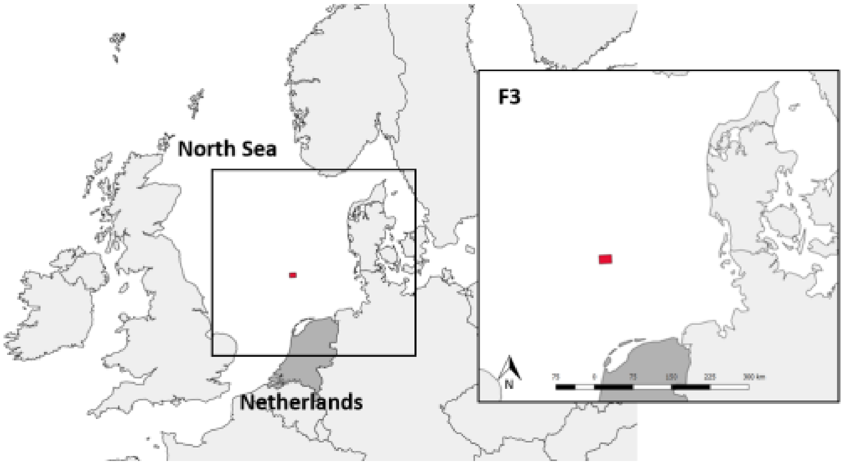


Figure 2: Location of the F3 3D survey in the North Sea, Netherlands offshore. (Source: Emilio Vital Brazil)

The main components of the data set are "Depth, Density, Sonic, Gamma-ray, P-impedance, P-impedance_rel, and Porosity." Depth and density are fairly straightforward, but the other parameters may require some definitions. Sonic refers to the P-wave travel time versus depth and is recorded as microseconds per foot. P waves, Primary waves, or compressional waves are the first seismic waves to arrive at a seismograph. Gamma Ray logs are used to measure the radioactivity of rocks and are scaled in American Petroleum Institute (API) units. The gamma-ray API unit is defined as 1/200 of the difference between the count rate recorded by a logging tool in the middle of the radioactive bed and that recorded in the middle of the nonradioactive bed. The P-Impedance is defined as density * P-wave velocity. Lastly, porosity is defined as the fraction of void space over the total space.

The data used for this project was taken directly from Nuwara's GitHub so it is unclear how much preprocessing went into it. Nevertheless, some data cleaning was needed at the beginning. Together these properties collectively make up the seismic well data. In the training data, there is a total of 4096 samples, some of which may be incomplete. To fix this, all data marked as "-999" was switched to NaN, or "not a number". The assigned facies were separated from the main dataset and only had marker lengths. To assign each sample its facies, a dictionary was created with each sample value having a facies key.

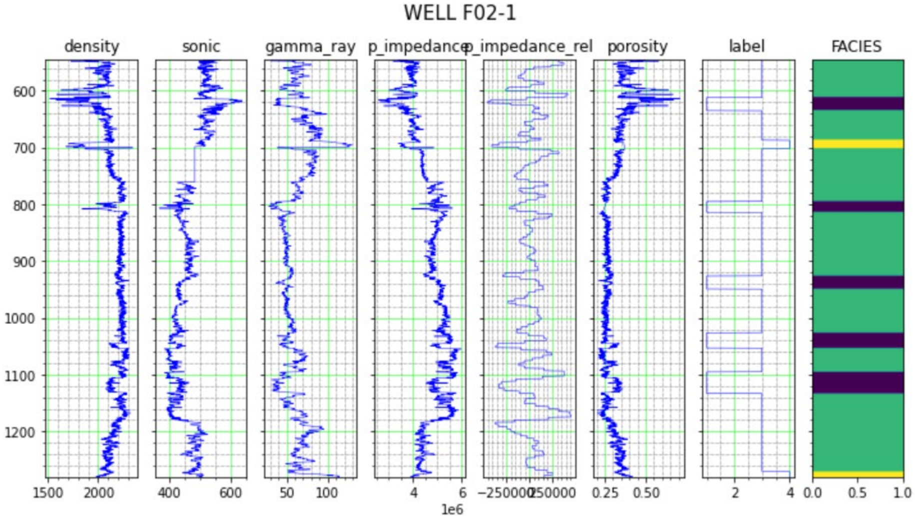


Figure 3: Graphs of all the parameters against depth alongside a labeled vertical interpretation of facies for well F02-1

A visual representation of all the data parameters can be seen in figure 2. Some clear correlations can be spotted through the examination of the graphs. Facies 1 is associated with relatively higher porosity and P-wave impedance with its surroundings. From these relative visual graphs, we can guess that facies 1 (blue) is most likely a fractured limestone formation of sorts, facies 3 (Green) may be sandstone, and facies 4 (yellow) could be a shale formation. The test data is unlabeled, so we won't know how accurate the model truly is, as the result cannot be evaluated.

METHODS

After manipulating the data to contain no problematic values and linking the facies to the samples, the training features were split from the target. Parameters of density, porosity, P-wave impedance, relative P-wave impedance, and gamma-ray were chosen. These parameters were chosen just by playing around and experimenting with combinations until one with the highest accuracy was found. It is important to note that this combination of parameters differs from Nuwara's selection. To avoid overfitting on the F-01 well, cross validator was used in the form of a stratified k-fold from SKLearn. This cross-validation object is a variation of KFold that returns stratified folds. The folds are made by preserving the percentage of samples for each class. After playing around with a couple more parameters, a value of 5 folds was used for the cross-validation, with the shuffle set to true. When selecting a model to use for prediction, there is a multitude of options. Because I'm not experienced with the various models, I decided to try a few and pick the one with the highest accuracy. The selected prediction models for this trial were Ridge Classifier, Support Vector Machine, Gaussian Naive Bayes, Classification Decision Tree, Random Forest Classifier, and K-Nearest-Neighbors.

Ridge classification differs from other classification models by adding a penalty term to the cost function that discourages complexity. This penalty term is usually the sum of the squared coefficients of the features in the model. By using a penalty term, the coefficients are forced to remain low, which prevents overfitting. When fit and tested on Well F02-1, the Ridge Classifier had an accuracy of 86.1%. Support Vector Machines work by taking data and mapping it to a high-dimensional feature space so that data points can be categorized, even when the data are not otherwise linearly separable. Typically, in a 2D space, a decision line could be used to separate data points. However, usually, a line becomes difficult when the data does not allow for a linear split. SVMs use kernel functions, in which dot products are used to find separators in higher dimensions. Following this, characteristics of new data can be used to predict the facies to which a new sample should belong. When fit and tested, the SVM had an accuracy of 89.8%. Gaussian Naive Bayes is a type of Naive Bayes classification. Naive Bayes is a probabilistic algorithm that is based on the Bayes theorem. The name "naive" is used because the algorithm uses features in its model that are independent of each other. Naive Bayes is a probabilistic machine learning algorithm that can be used in several classification tasks. Typical applications of Naive Bayes are the classification of documents, filtering spam, prediction, and so on. This algorithm is based on the discoveries of Thomas Bayes and hence its name. Gaussian Naive Bayes assumes that each class follows a Gaussian distribution. For Well F02-1, Gaussian NB had an accuracy of 84.2%. A Classification Decision Tree is a non-parametric algorithm with a hierarchical tree structure. This structure consists of a root node, branches, internal nodes, and leaf nodes. The algorithm works by starting at the root node and making decisions that drive classification down the tree. Decision tree learning employs a divide-and-conquer strategy by conducting a greedy search to identify the optimal split points within a tree. This process of splitting is then repeated in a top-down, recursive manner until all, or the majority of records have been classified under specific class labels. Whether or not all data points are classified as homogeneous sets is largely dependent on the complexity of the decision tree. The decision tree had an accuracy of 96.0%. Random Forest classifiers consist of numerous individual decision trees which work together. Each tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction. These individual trees are not correlated to each other, which results in the trees protecting each other from their errors. This classifier ultimately led to an accuracy of 96.7%. Lastly, the K-Nearest-Neighbors classifier uses each sample's neighbors in a 2D space to draw boundaries to classify. Distances between the points are used to determine proximity. KNN had an accuracy of 96.2%. The predicted facies of each algorithm next to the actual facies can be seen in figure 4.

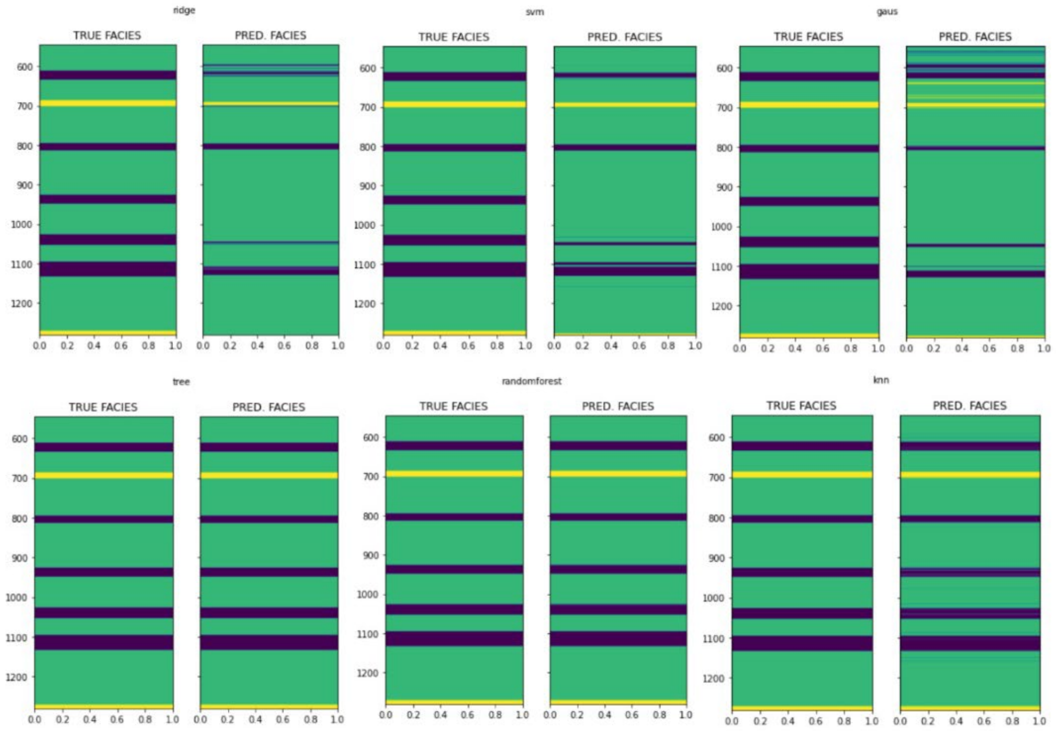
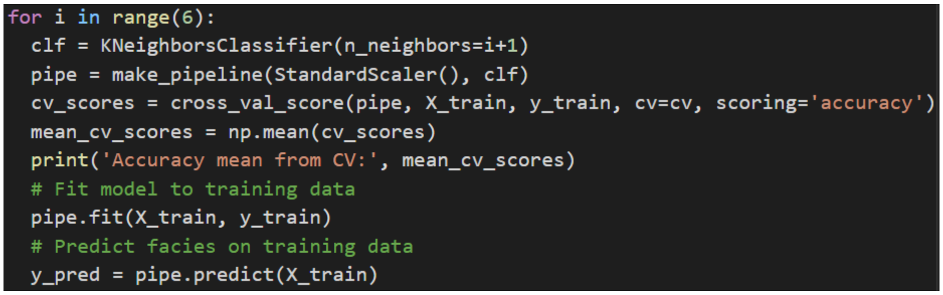


Figure 4: Various models predicted facies vs. true facies (Ridge, SVM, Gaussian, Tree, Random Forest, KNN)

With the results from this experiment in mind, I decided to move forward with KNN. Although the KNN model had lower accuracy than the Random Forest, the KNN model can be optimized by changing the value of k neighbors. Furthermore, results with random forest-modeled lithostratigraphy that was not likely. The value of k determines how many neighbors are taken into account when learning, so optimizing this value can have profound effects on the results. Similar to the small simulation done to choose the model, I simulated the prediction of k values 1 through 6. Ultimately, I chose a value of 4.



```
for i in range(6):
  clf = KNeighborsClassifier(n_neighbors=i+1)
  pipe = make_pipeline(StandardScaler(), clf)
  cv_scores = cross_val_score(pipe, X_train, y_train, cv=cv, scoring='accuracy')
  mean_cv_scores = np.mean(cv_scores)
  print('Accuracy mean from CV:', mean_cv_scores)
  # Fit model to training data
  pipe.fit(X_train, y_train)
  # Predict facies on training data
  y_pred = pipe.predict(X_train)
```

Figure 5: Testing ascending K nearest neighbors values 1-6

Lastly, the data from the wells contain a lot of noise which potentially introduces problems in the predictions. Well-log data has a high resolution because of its highest frequency in the range of 20 to 40 kHz. This usually is good, as this frequency can capture small contacts between two different lithofacies as accurately as 10-20 centimeters. But this accuracy level is not needed for our data. Which contains facies thickness that ranges from 20-100 meters. Filtering out the high frequency will help accuracy levels. Filtering the high frequency and retaining the low frequency can be done by implementing a Butterworth filter. Implementation for the Butterworth filter was done by following Nuwara's article tutorial.

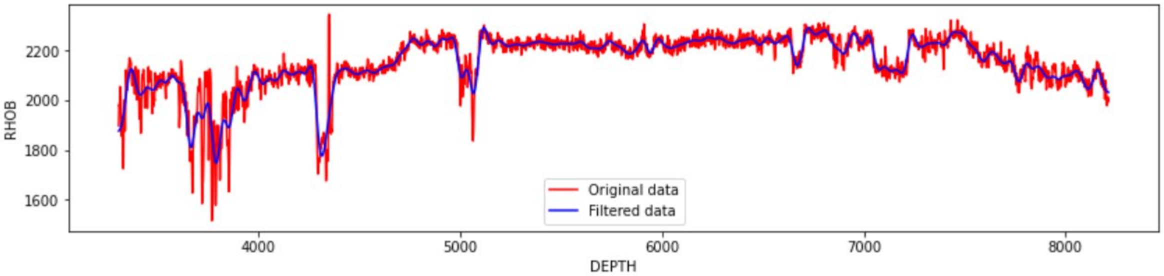


Figure 6: Line graph showing original density data (Red) and filtered data (Blue) through the Butterworth filter

RESULTS

The results consist of the preliminary results on the stratified k-fold as well as the prediction of the neighboring F02-3 well. The result of the testing on stratified k-fold Well F02-1 was an accuracy of 99%. This is a 1% increase over the results in Nuwara's article. The predicted facies of Well F02-3 can be seen in figure 8. Figure 7 provides a confusion matrix that deconstructs the few incorrect predictions.

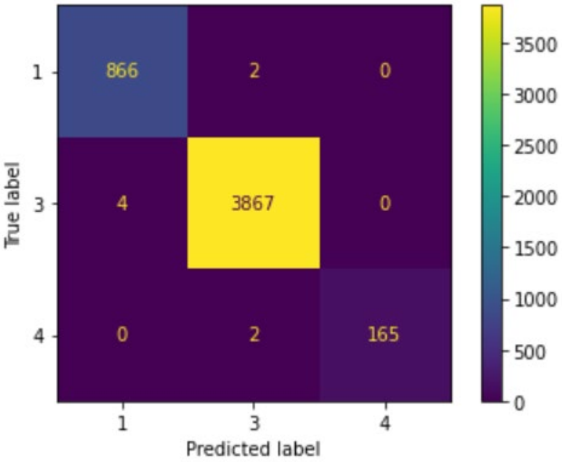


Figure 7: Confusion Matrix that visualizes the number of correct samples as well as the number of incorrect samples

This prediction is different in comparison to the one in Nuwara's article but is more likely closer to the actual lithostratigraphy of the well. Our prediction is more consistent with the thickness of facies in the Central Graben Basin. Our results were different from Nuwara due to two key changes: The parameters of the model, and the Butterworth low pass filter settings. Instead of the parameters of density, sonic, gamma ray, and porosity, we selected density, P-wave impedance, gamma ray, relative P-wave impedance, and porosity. This selection of parameters was done based on visual analysis, research, and trials. The low pass filter settings were selected using trial and error. The initial data value of 5 was too high, so I brought it down to 4.

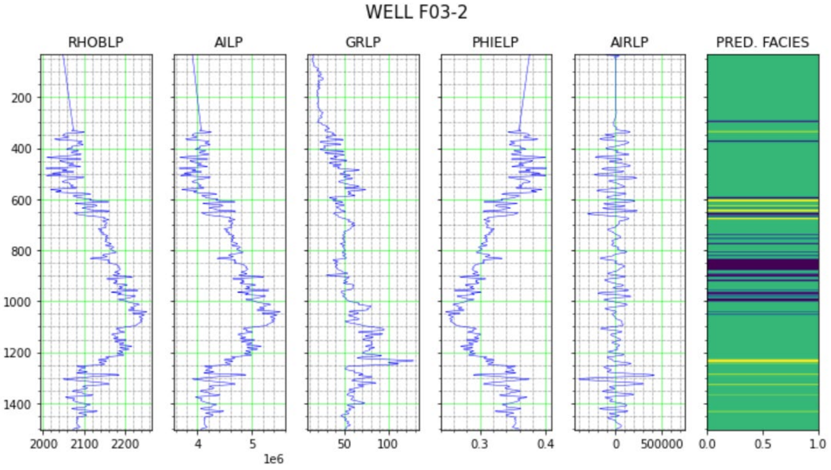


Figure 8: Final predicted facies for Well F02-3 along with the parameters graphed against depth

To check the validity of the KNN model, I also ran the same parameters on the Random Forest model. Recall that initially, the Random Forest model had outperformed the KNN in stratified testing. The prediction results from the Random Forest are visible in figure 8. The issue with this prediction is the likelihood of the well lithostratigraphy. We had previously identified in the data section that facies 1 (blue) was likely similar to limestone, facies 3 (green) was likely similar to sandstone, and facies 4 (yellow) was similar to shale. In the Random Forest model, facies 1 forms a thick layer. Given the thickness and formation environment of the limestone beds in well F02-1, the RF model is inconsistent. The depositional environment for well F02-1 was mainly sandstone. The interpretation of the data is also more consistent with the North Sea formation aforementioned in the dataset section. Nuwawa's final prediction is ultimately less likely than the prediction I present because of the nature of the prediction. The number of facies changes is unrealistic. Considering Walther's law of Facies, for Nuwara's prediction to be true the terrain would've moved laterally back and forth repeatedly in an extremely short time. Although our prediction contains similar thin facies, it is less dramatic than Nuwara.

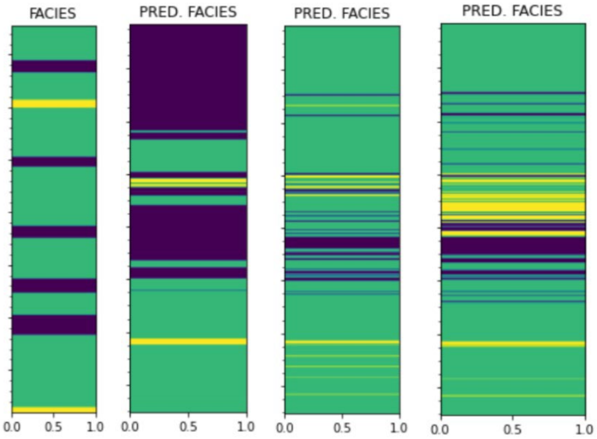


Figure 9: Left to right: Well F02-1 true facies, Random Forest prediction for well F02-3, KNN prediction for well F02-3, Nuwara's KNN prediction for well F02-3

CONCLUSION

We were able to successfully deconstruct machine learning-based lithostratigraphy, and understand the decisions made in model selection and implementation. This was done through visual data analysis, trial and error, and geographic research. Visual analysis of data is extremely important, and choosing the right parameters for input can seriously change the output. Perhaps the most crucial takeaway is the questionable nature of the result. Because the F02-3 well is not labeled, it is unclear how well our model truly performed. This problem is overshadowed by the larger, pressing issue in the earth science machine learning domain: The lack of data. In a search for applicable datasets, there were very few labeled and previously used wells. My options were limited to 3-4 popular wells that had already been extensively studied. The well data is by no means in a shortage either; There are a large number of wells that could be used to further enhance and fine-tune existing models. During this project, I reached out to a couple of drilling companies to no avail. Although the final predictions must be taken with a grain of salt, the process demonstrated in selecting a KNN is still valid. The success of the KNN may be because of the nature of the problem. The chemical and physical properties of one sample are heavily influenced by the adjacent neighboring samples above and below.

References

Alaudah, Y., Michalowicz, P., Alfarraj, M., & AlRegib, G. (2019, April 10). *A machine learning benchmark for facies classification*. arXiv.org. https://arxiv.org/abs/1901.07659

Bestmann, I. (2020, August 24). *Facies classification using unsupervised machine learning in geoscience*. Medium. https://towardsdatascience.com/facies-classification-using-unsupervised-machine-learning-in-geoscience-8b33f882a4bf

Bisla, D., Wang, J., & Choromanska, A. (2022, February 4). *Low-pass filtering SGD for recovering flat optima in the Deep Learning Optimization Landscape*. arXiv.org. https://arxiv.org/abs/2201.08025

Chen, J. (2018, August 29). *Application of machine learning in rock facies classification with physics-motivated feature augmentation*. Papers With Code. https://paperswithcode.com/paper/application-of-machine-learning-in-rock

Dwihusna, N. (1970, January 1). *Seismic and well log based machine learning facies classification in the PANOMA-hugoton field, Kansas and Raudhatain Field, North Kuwait*. The Mines Repository. https://repository.mines.edu/handle/11124/174200

Iykekings. (2019, March 13). *Facies classification with machine learning*. Kaggle. https://www.kaggle.com/code/iykekings/facies-classification-with-machine-learning

Kaur, H., Pham, N., Fomel, S., Geng, Z., Decker, L., Gremillion, B., Jervis, M., Abma, R., & Gao, S. (2023, February 1). *A deep learning framework for seismic facies classification*. Interpretation. https://pubs.geoscienceworld.org/interpretation/article/11/1/T107/619761/A-deep-learning-framework-for-seismic-facies

Lee, A.-S., Enters, D., Huang, J.-J. S., Liou, S. Y. H., & Zolitschka, B. (2022, November 26). *An automatic sediment-facies classification approach using machine learning and feature engineering*. Nature News. https://www.nature.com/articles/s43247-022-00631-2

Martin, T., Meyer, R., & Jobe, Z. (2021, June 4). *Centimeter-scale lithology and facies prediction in cored wells using machine learning*. Frontiers. https://www.frontiersin.org/articles/10.3389/feart.2021.659611/full

Melo, A., & Li, Y. (2021, December). *Geology differentiation by applying unsupervised machine learning to multiple independent geophysical inversions*. Academic.oup.com. https://academic.oup.com/gji/article/227/3/2058/6346571

Nuwara, Y. (n.d.). *PDA series #2 facies classification from well logs*. LinkedIn. https://www.linkedin.com/pulse/pda-series-2-facies-classification-from-well-logs-yohanes-nuwara

Palkovic, M. (2021, September 7). *Exploring use cases of machine learning in the Geosciences*. Medium. https://towardsdatascience.com/exploring-use-cases-of-machine-learning-in-the-geosciences-b72ea7aafe2