**MODEL SUMMARY**

**BACKGROUND TEAM**

* Competition Name: Azubi Africa Hackathon
* Team Name: Cluster Champs
* Team Led: Abanise Orojo
* Programmer:
* Programmer:

**METHODOLOGY**

The study data was obtained from the Source Company (Company name withheld for confidentiality purposes). The analysis was conducted on 891 samples. The geochemical data for The analysis was carried out following processes and conventional approach adopted and recognised.

The samples were processed in the Organic Laboratory of the Source Company, where they were oven-dried at 110°C overnight and pulverized using pestle and mortar. With the use of total organic carbon contents and Rock-Eval pyrolysis methods, crushed rock samples were geochemically evaluated.

According to McCarthy et al. (2011), a direct heating method can be used to obtain the total organic carbon values that only needs 1 g of rock samples. The samples are crushed and processed to extract any carbon discovered in samples of carbonate, and/or contaminants.

The samples are heated up with a high-frequency induction furnace at 1,200˚C. Carbon in kerogen is transformed into CO and CO2. The developed carbon fractions in an infrared cell are assessed, transformed to TOC and recorded as a percentage of rock mass weight

Nuñez- Betelu and Baceta (1994) indicated that the Rock-Eval tool’s working parameters are similar to those described in Clementz et al. (1979). Samples of ground whole rock measuring up to 100 mg are pyrolyzed at 300 ° C for 3-4 min., accompanied by calibrated pyrolysis at 25 ° C / min to 550 ° C, both in a helium atmosphere. It takes about 20 minutes for each assessment to cool the oven.

If there is no oxygen, heating of organic matter produces organic compounds. In the first stage of pyrolysis when the sample is kept at 300˚C, the free organic compounds (bitumen) present in the rock are refined. The insoluble organic matter (kerogen) is split into pyrolytic products in the second stage of increasing heating to 600˚C. During the two phases, flame ionization and thermal conductivity sensors senses any organic compounds and CO2 produced.

The helium gas flow sweeps the volatile products from the oven to a splitter in the first stage. The first half of the divided effluent is sent to the water scrubber and a CO2 trap. The second half is directed into a hydrogen flame ionisation detector where hydrocarbons volatile at 300˚C are identified and analyzed quantitatively. The oven raises its temperature at 25˚C / minute to 600˚ C after 2 minutes. After reaching this temperature, the CO2 trap is dumped onto a thermal conductivity detector and the amount of thermally evolved organic CO2 is measured.

According to Peters (1986), there are several measurements in the Rock-Eval II method (Espitalie et al, 1977). A flame ionization detector (FID) senses all organic compounds generated all through pyrolysis.

The first peak (S1) represents milligrams of hydrocarbons which can be distilled thermally from one gram of rock. The second peak (S2) reflects milligrams of hydrocarbons generated by kerogen pyrolytic disintegration. (Although the literature represents S1 and S2 in milligrams of "hydrocarbons" per gram of rock, the FID also detects non-hydrocarbons when carbon atoms are present.)The third peak (S3) represents milligrams of carbon dioxide produced during temperature programming up to 390˚ C from a gram of rock and is evaluated by thermal conductivity detection (TCD).

A thermocouple monitors the temperature during pyrolysis. The temperature that generates the highest quantity of S2 hydrocarbons is called Tmax (Tissot and Welte, 1984). Vitrinite reflectance and temperature are included in the calibration information for the thermal models. Vitrinite reflectance values ranged from 0.27% to 0.65%.

The parameters obtained from the pyrolysis were used to calculate the following:

* Oxygen Index
* Hydrogen Index: The HI vs. OI plot helps deduce the type of kerogens in the rock source (Peters and Moldowan, 1993)
* Production Index
* Hydrogen richness in the kerogen
* Generation potential of the source rock

**DATASET:**

The dataset used is a sample of wells from around the world. The dataset that was used for this project is a subset of a much larger datasets as prescribed by . The number of samples were restrict to the samples provided by has the SOURCE.

* S1 - This represents the number of free hydrocarbons that can be easily flushed out of the rock during the early part of pyrolysis. It is an important measurement for the detection of hydrocarbons.
* S2 - The “hydrocarbon yield from kerogen cracking”. This is a direct measurement of the rock's potential to generate hydrocarbons.
* S3
* TOC - Total Organic Carbon, this indicates the richness of the organic matter in the rock. It includes both the insoluble organic matter (kerogen) and the soluble organic
* PP
* Tmax - This is the pyrolysis temperature of the S2 peak. It is a useful back-up to vitrinite reflectance, particularly in the late immature to strongly mature stage.
* Classes

In the dataset, there are 890 example vectors. Expert Systems have been used in the field

of geoscience to assist the specialist in classifying each well, and this can help save money.

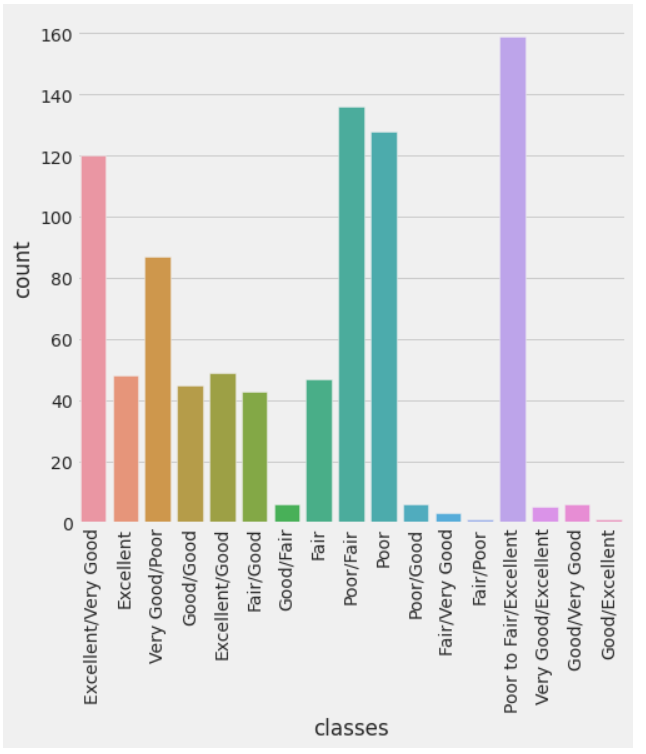
* How it helped save money
* Importance of model

**SUMMARIZE DATA**

1. **DESCRIPTIVE STATISTICS**



1. **DATA VISUALISATION**



**Counting Plot of Classes Column**

**PREPARE DATA**

1. **DATA CLEANING**

No major data cleaning needed to be done to the dataset. 3 missing rows were dropped from the data set which accounted for less the 1% of the entire data set

1. **FEATURE ADDITION/SELECTION**

Majority feature addition was done in excel, creating the following new columns:

* S1 + S2 - is a measure of the Genetic potential of the source rock
* S2/S3 - This ratio is an indicator of hydrogen richness in the kerogen.
* matter (bitumen).
* Hydrogen index (HI) =S2/TOC \*100
* Oxygen Index (OI) = S3/TOC \*100
* Production Index (PI) =S1/(S1+S2)
* R0 = (0.018 \* Tmax) - 7.16

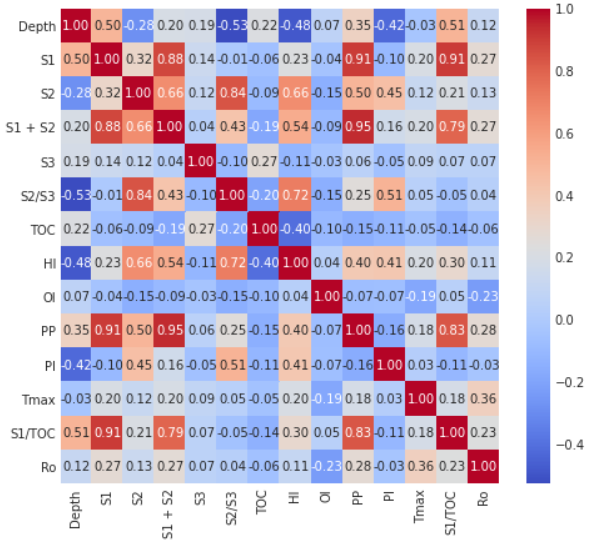
**FEATURE ENGINEERING**

All numeric columns were normalized with a MaxAbsScaler which scaled each feature by its maximum absolute value. This estimator scaled and translated each feature individually such that the maximal absolute value of each feature in the training set will be 1.0. It does not shift/center the data, and thus does not destroy any sparsity.

**MODEL**

Baseline Classifiers:

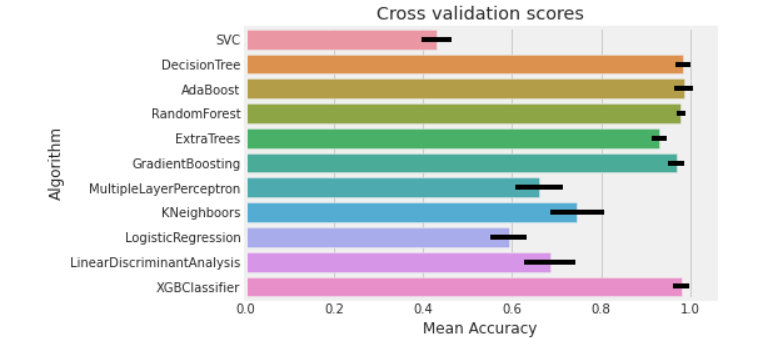
We first performed a feature correlation to check how independent each feature was from each other:



This shows that most features are independent of each other expect for features like S1 and S1 + S2, where a feature was reused to calculate a new feature.

As the baseline classifiers, we chose 11 classifiers, we chose SVC, Decision Tree, AdaBoost, Random Forest, Extra Trees, Gradient Boosting, Multiple Layer Perceptron, K Nearest Neighbors, Logistic Regression, Linear Discriminant Analysis, XG Boost because they are relatively easy to compute, and the features in the given dataset are all aspects independent of each other it would be easier to classify each case. Each algorithm was scored based on accuracy and was plotted in a bar chart.

A stratified K Fold shuffled the data then splits the data into 10 parts, using each part as a test set.

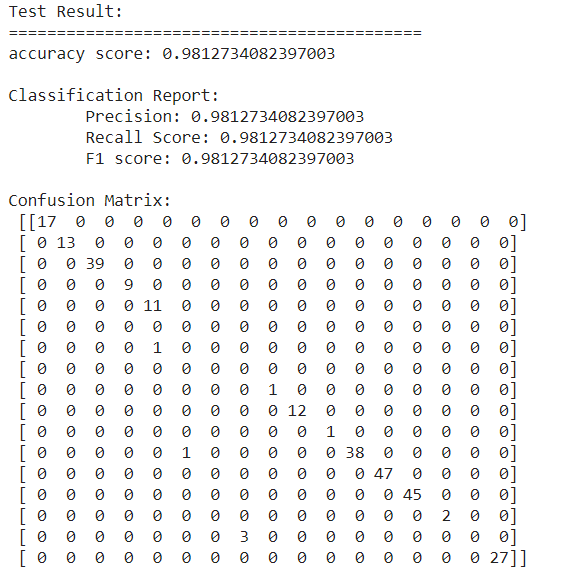


As we can see that Decision Tree, Ada Boost and XG Boost were able to best classify the data with over 95% in both cases.

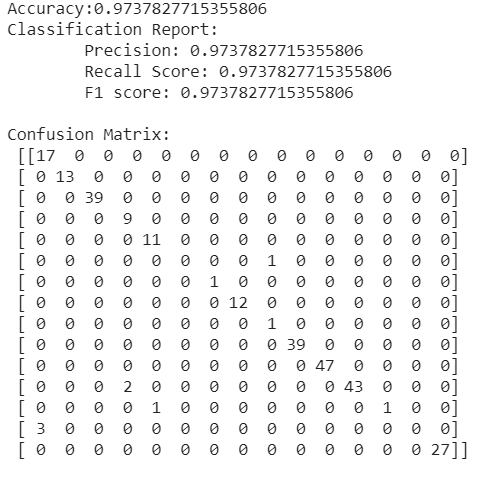
**HYPERPARAMETER TUNING:**

A further deep dive was done into the top 3 models to see which would produce the best model. The gird search method was used to find the best parameters for each model and the result of each are shown below:

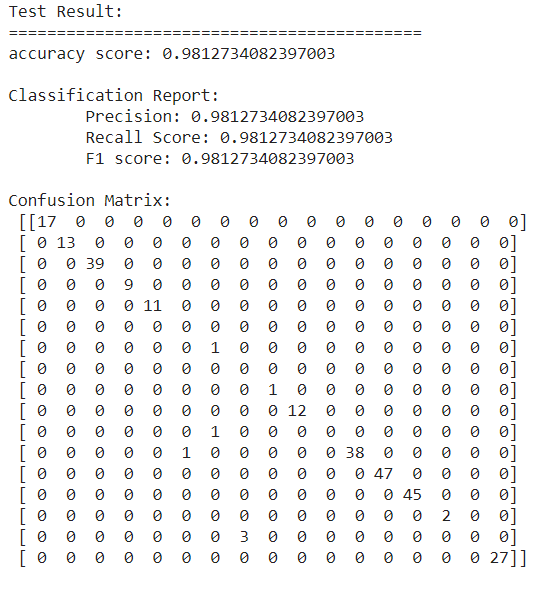
**AdaBoost**



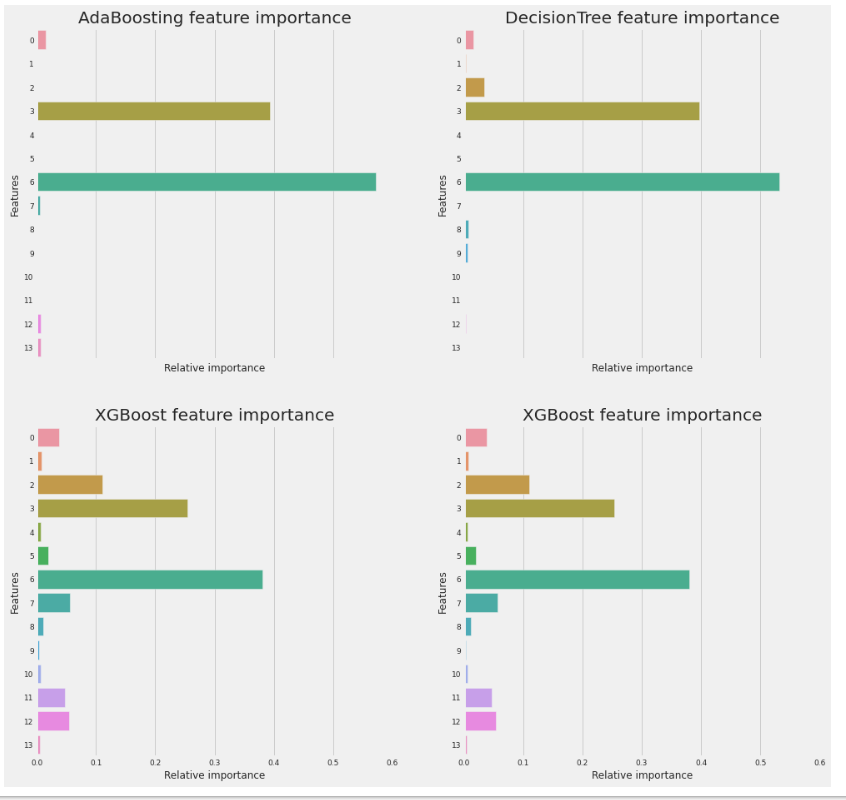
**XGBoost**



**DECISION TREE**



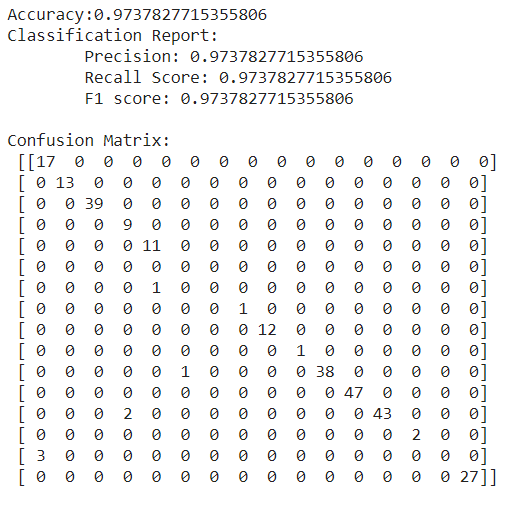
As can be seen AdaBoost performed the best out of all the models used. We plotted a feature importance graph to identify the most important feature(s) in the AdaBoost classifier:

****

TOC is the most important feature for the in splitting the data

**ENSEMBLES**

The 3 classifiers give more or less the same prediction but there are some differences. These differences between the 3 classifier predictions are sufficient to consider an ensembling vote. We choose a voting classifier to combine the predictions from the 3 classifiers. We preferred to pass the argument "soft" to the voting parameter to take into account the probability of each vote.



**CONCLUSION**

To conclude the best classifier was the AdaBoost as it performed better than the ensemble vote.