IMPLEMENTATION OF RANDOM FOREST FOR LITHOFACIES CLASSIFICATION

VU NGUYEN

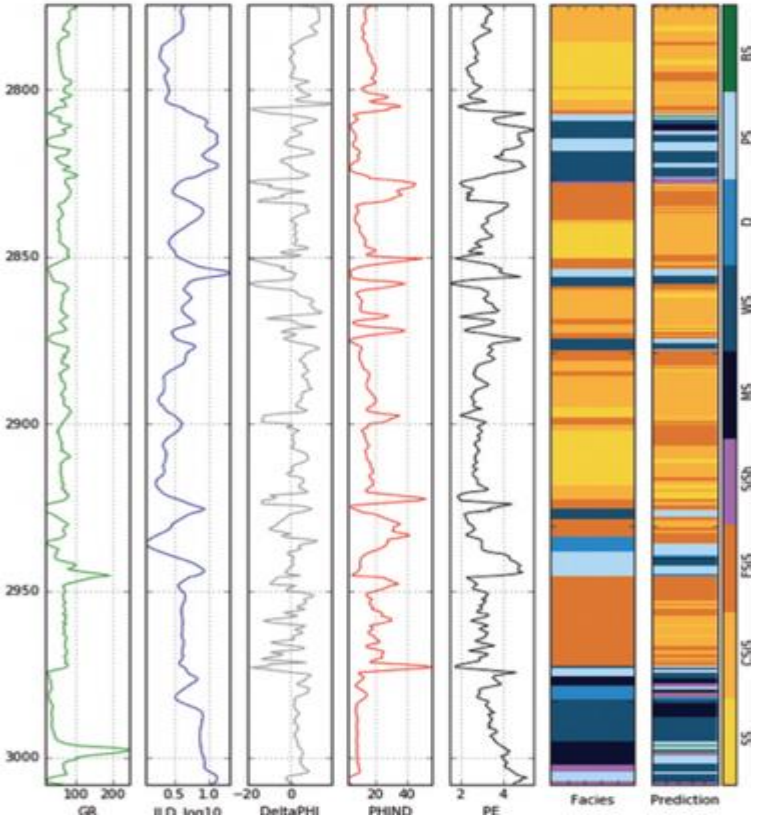
OBJECTIVE: TO DEVELOP A MACHINE LEARNING MODEL TO CLASSIFY LITHOFACIES

Outlines:

- Introduction
- Data Acquisition/Data Wrangling
- Data Visualization
- Random Forest
- Conclusions

INTRODUCTION

- Lithofacies identification is a process that allows the determination of the hydrocarbon bearing zone.
- The ideal sources for lithofacies classification are core samples of rock extracted from wells within the field. However, core samples are not always available due to the associated cost
- Indirect measurement such as well logging emerges as an alternative method to classify facies
- The application of machine learning is a promising method facilitate the lithofacies classification process.



DATA ACQUISITION/DATA WRANGLING

- The data obtained from Alberta Geological Survey. it is a data collection of 2193 wells to map the McMurray Formation and the overlying Wabiskaw Member of the Clearwater Formation in the Athabasca Oil Sand Area
- Two csv files “Picks” and “Intellog” were loaded into Pandas Dataframes. The columns of “Picks” dataframe were renamed to “SitID”, “HorID”, “Depth” and “Quality”.
- The “Depth” column was converted from object to numeric data type
- The “RW” column of “Intellog” dataframe was converted to numeric data type
- Merged “Intellog” and “Picks” dataframe together to become “Main_File” dataframe by inner joint method by the common columns “SitID” and “Depth”
- Dropped any N/A values in “Main_File”
- Created features matrix by dropping columns “SitID”, “HorID”, “Depth” and “LithID” from “Main_File” dataframe
- Finally, created target column by choosing column “LitID” only from “Main_File” Dataframe

DATA ACQUISITION/DATA WRANGLING (CONT.)

6 Features were used to characterize different classes of lithology:

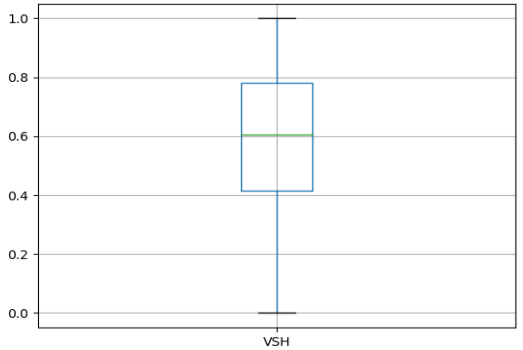
- PHI: The porosity of the rock.
- Quality: The degree of confidence placed on the correlation of a pick.
- RW: Water resistivity at water temperature
- SW: Water saturation
- VSH: Volume of Shale
- W_tar: Mass percent of bitumen

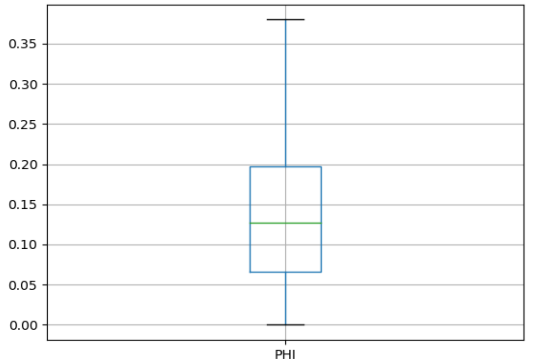
2 engineer features were computed:

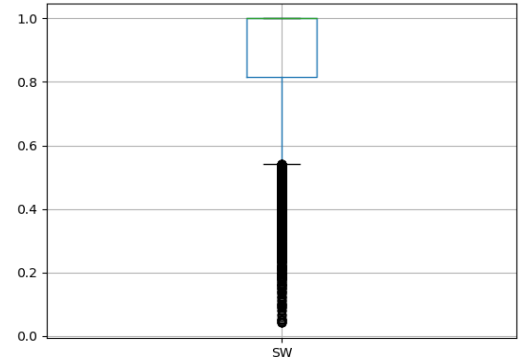
- VS: Volume of sand
- Shc: Hydrocarbon saturation

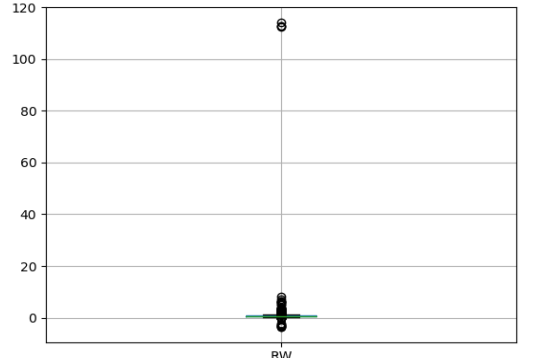
DATA VISUALIZATION

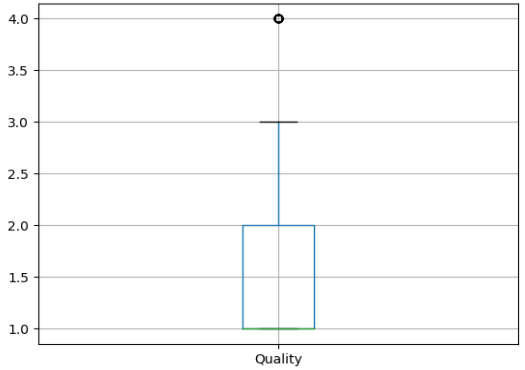
- All the features, except for VSH, have outliers.
- Leave the outliers alone

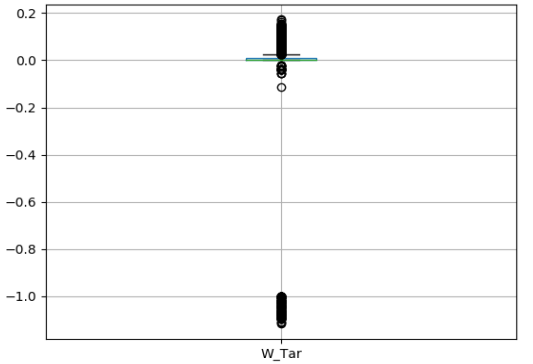






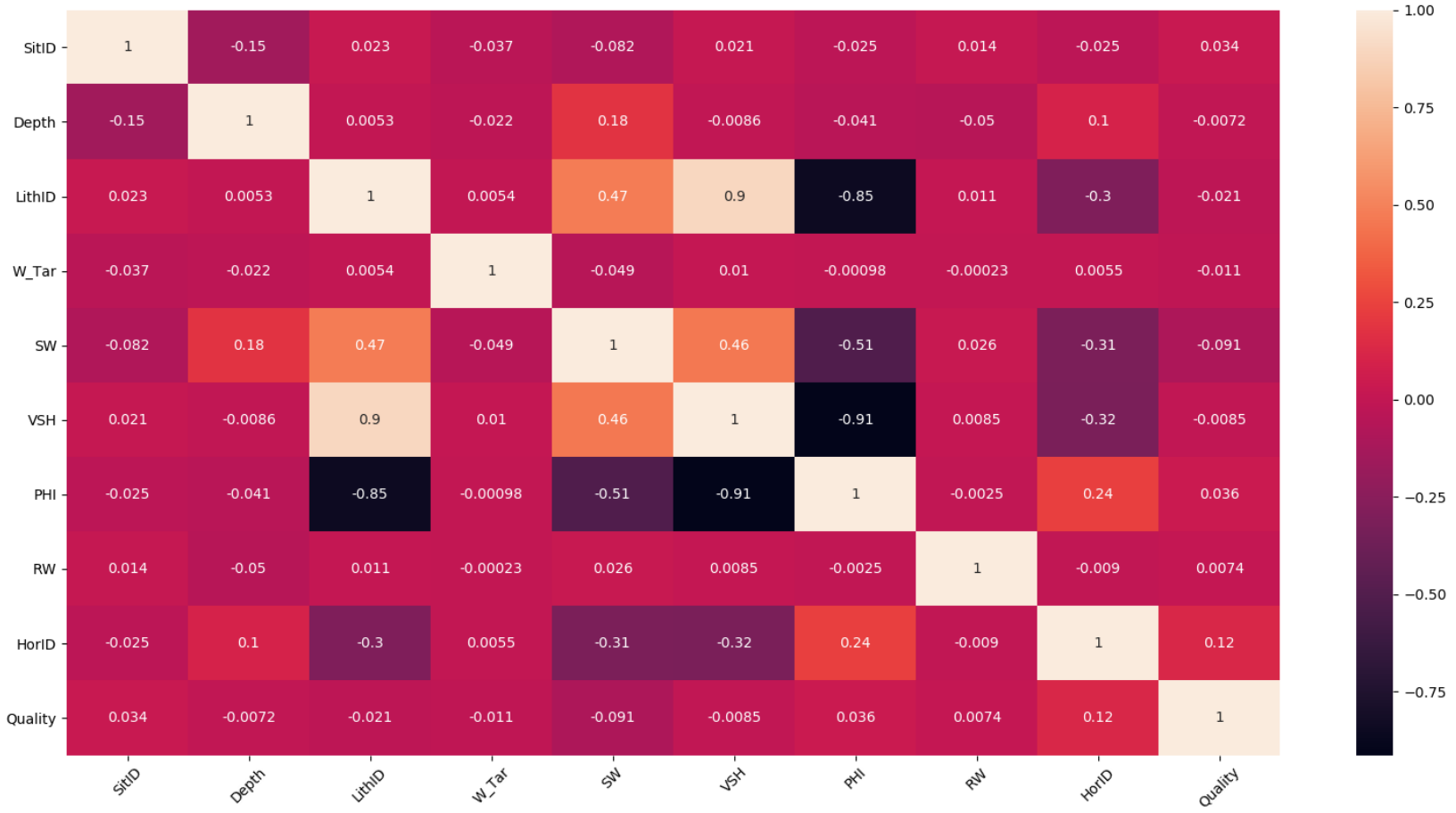






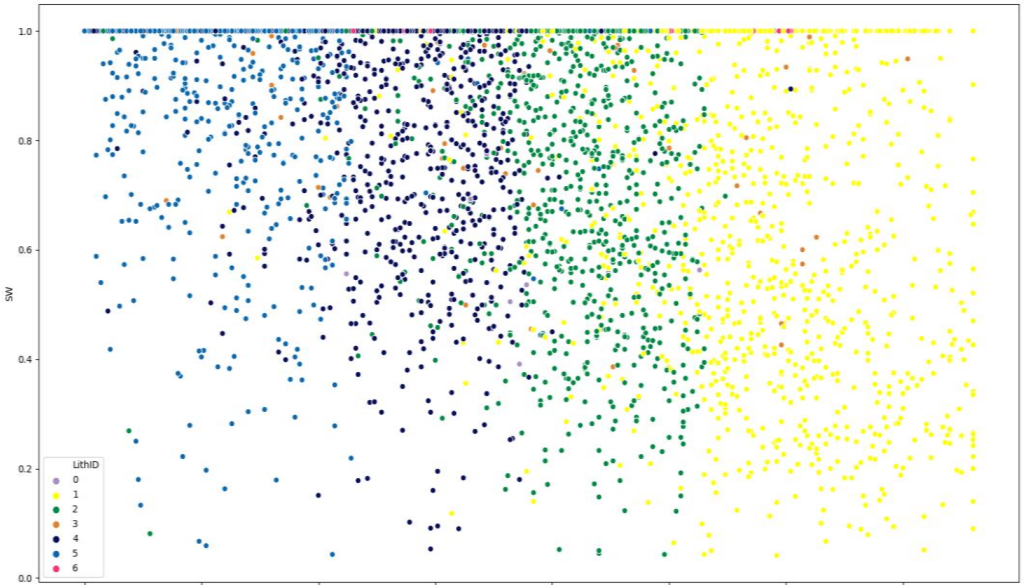
DATA VISUALIZATION (CONT.)

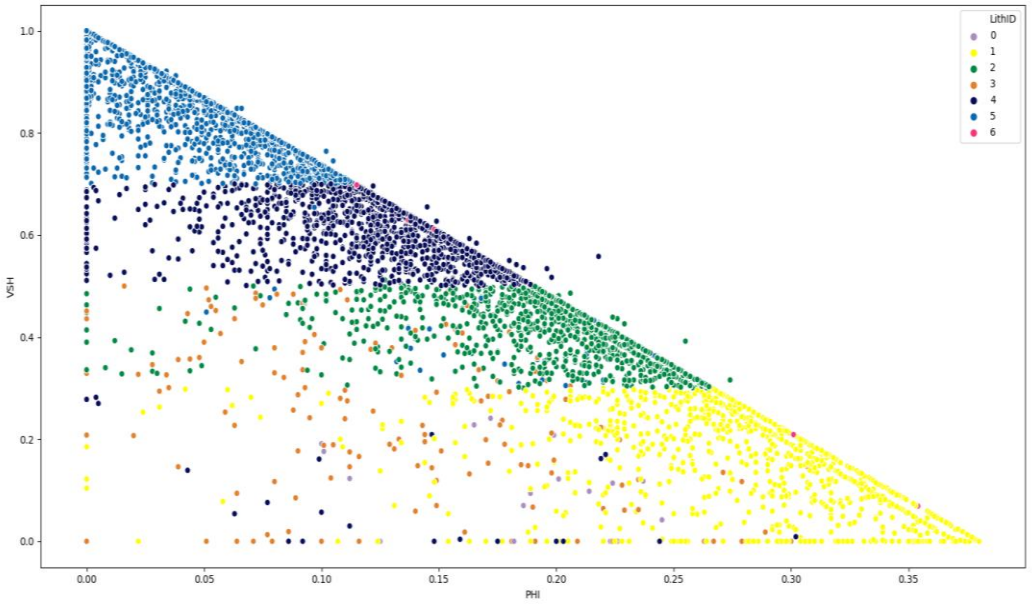
Strong correlation between lithofacies classes vs. VSH, PHI and SW

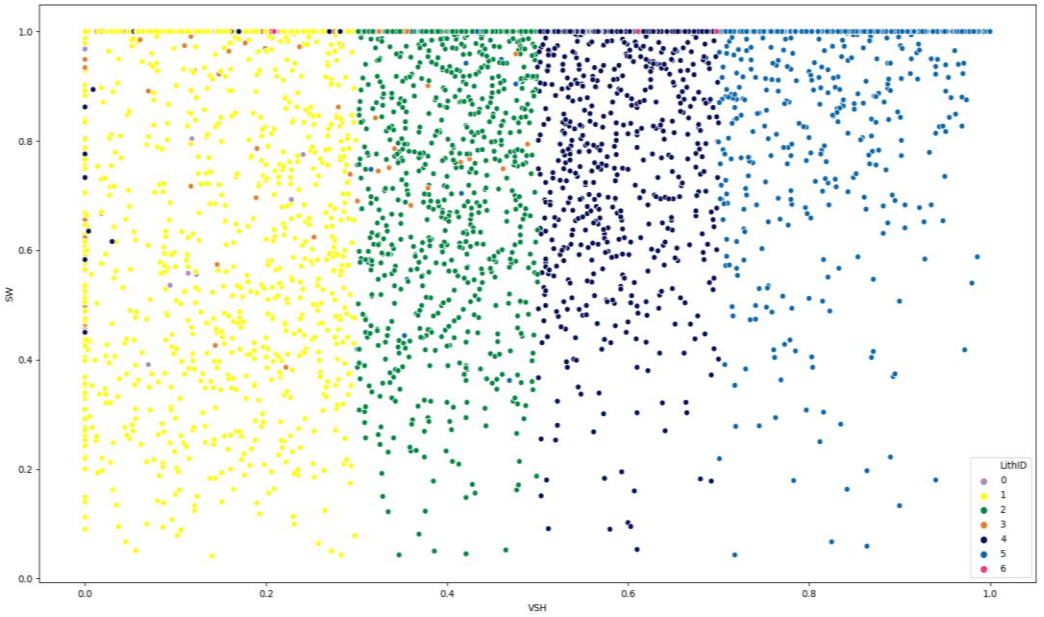


DATA VISUALIZATION (CONT.)

Strong correlation between lithofacies classes vs. VSH, PHI and SW

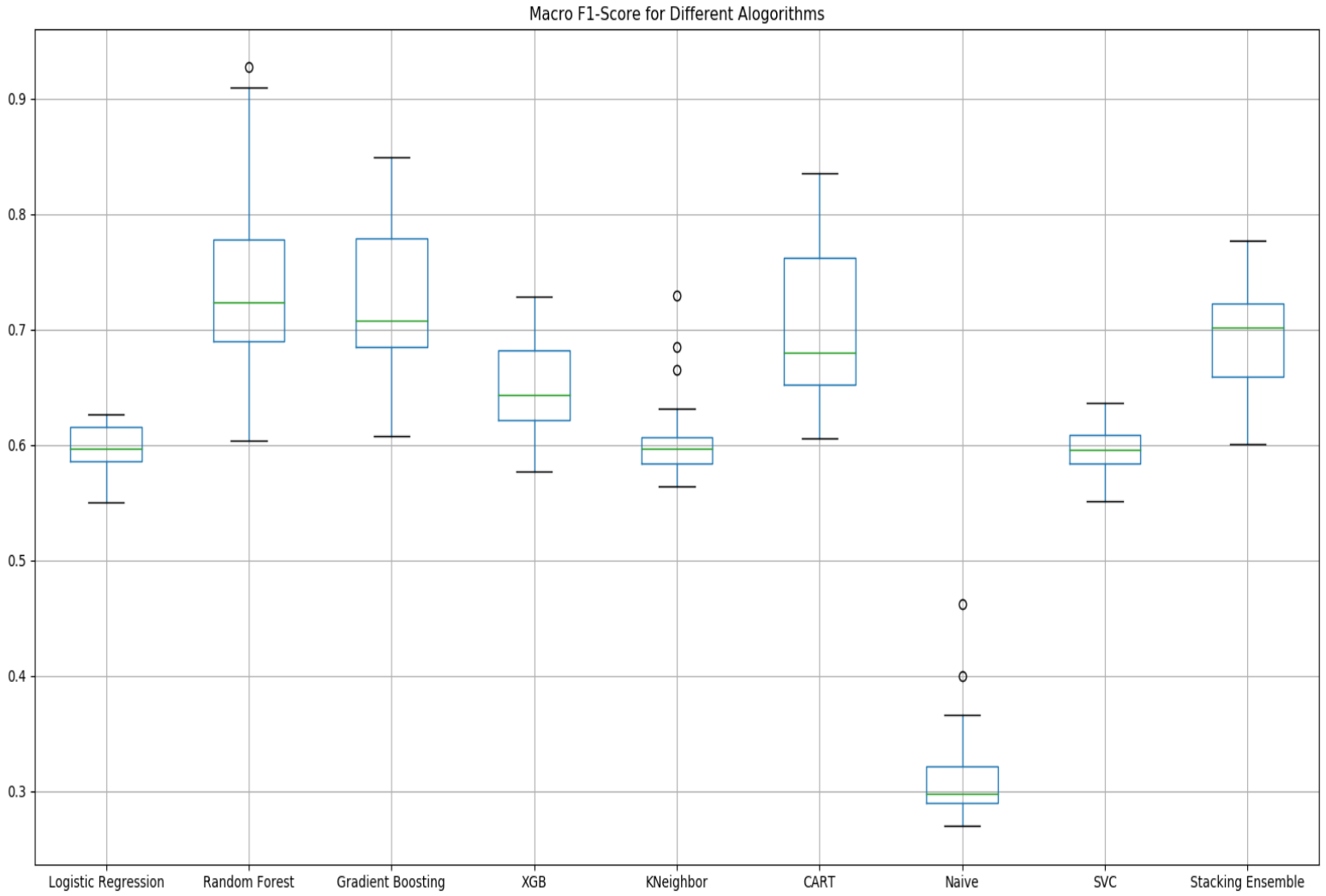




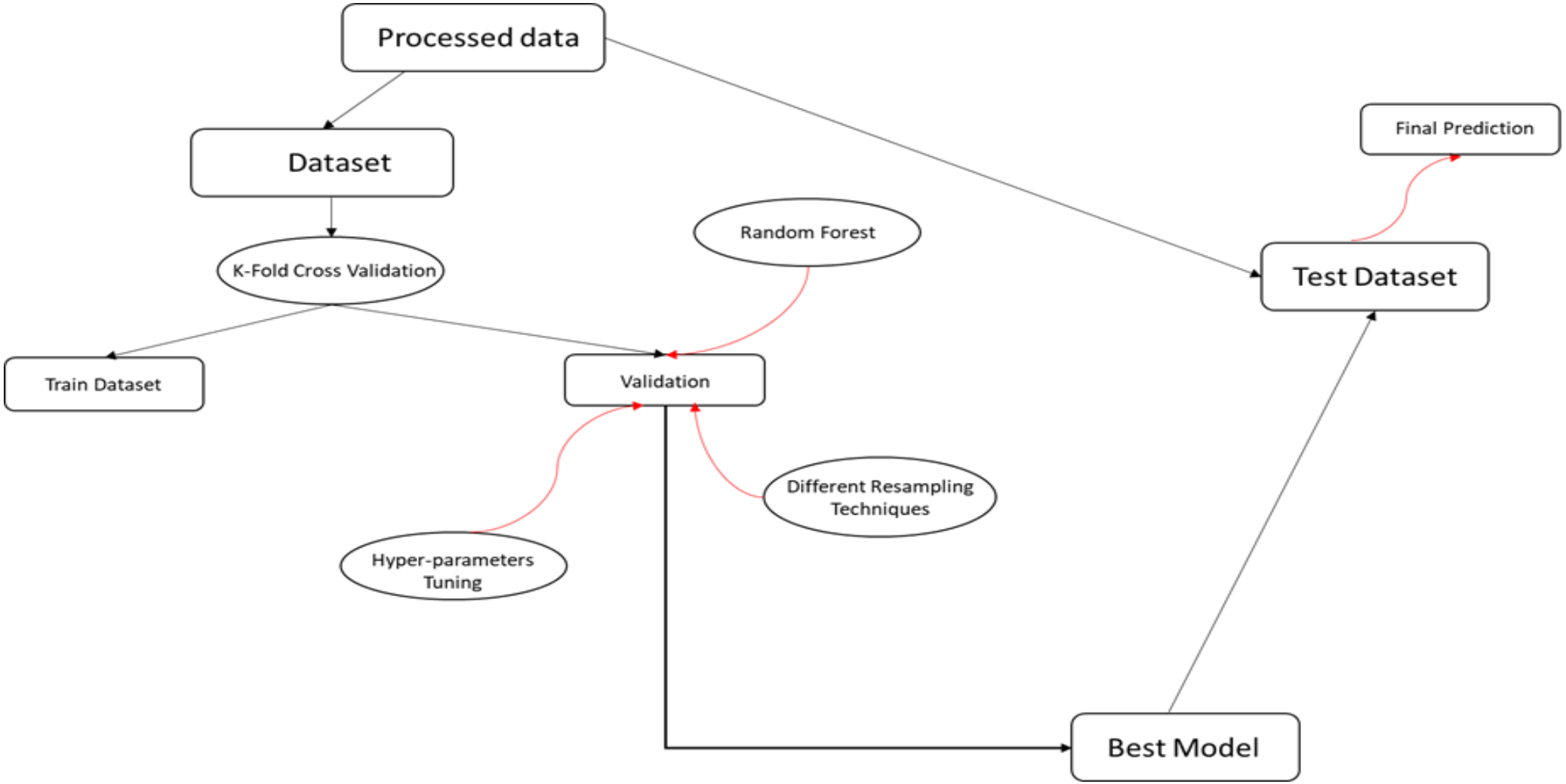


MACHINE LEARNING ALGORITHMS SCREENING

- Various algorithms were screened to identify the potential methods.
- Ensemble learning, including Random Forest, Gradient Boosting and Stacking, outperformed other methods
- Random Forest was chosen for further investigation



RANDOM FOREST - METHODOLOGY

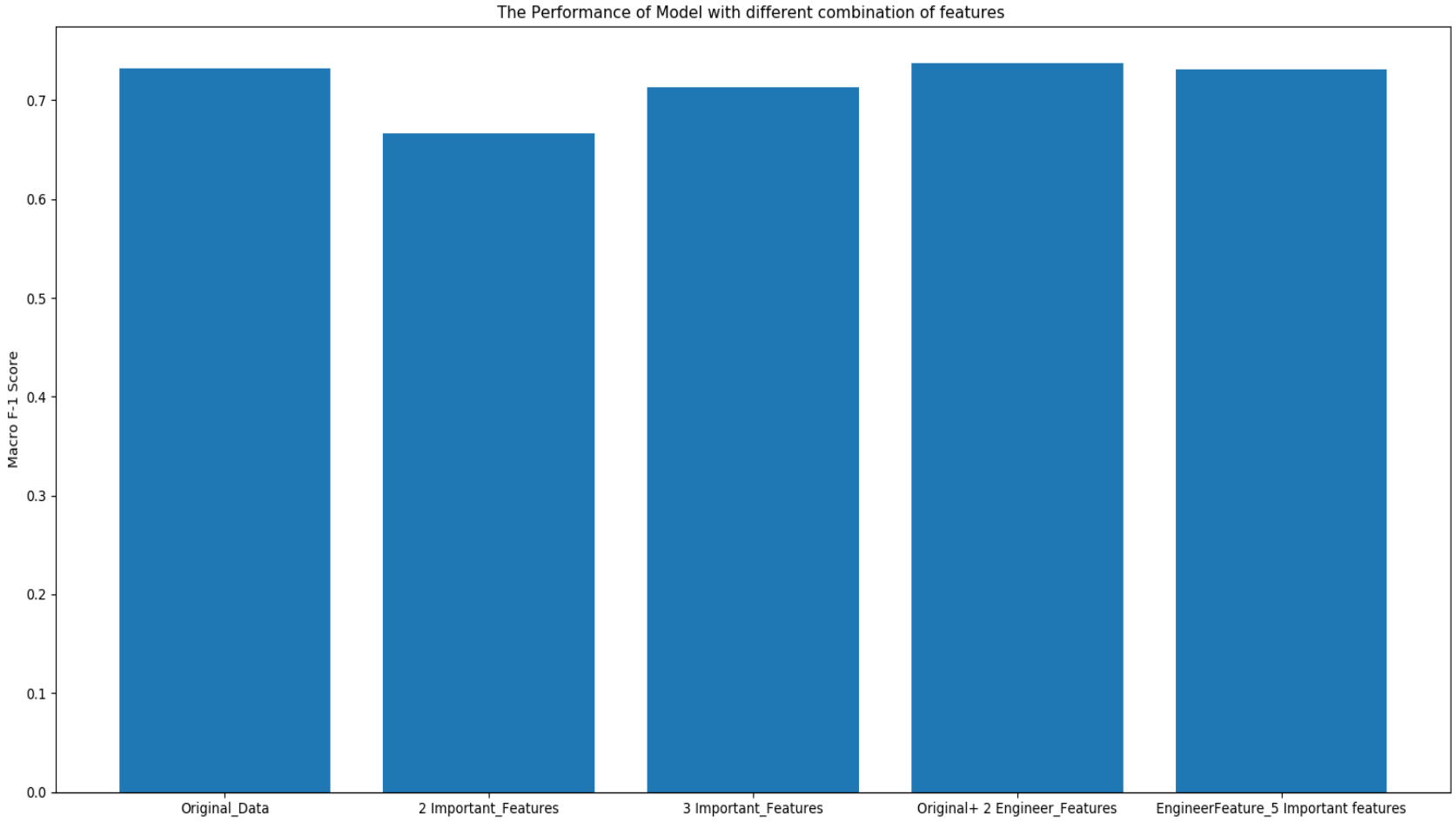


RANDOM FOREST – THE PERFORMANCE OF MODEL WITH DIFFERENT FEATURES

A baseline Random Forest were trained with the original dataset but with different chosen features:

- Top two most important features
- Top Three most important features
- With Engineer Features

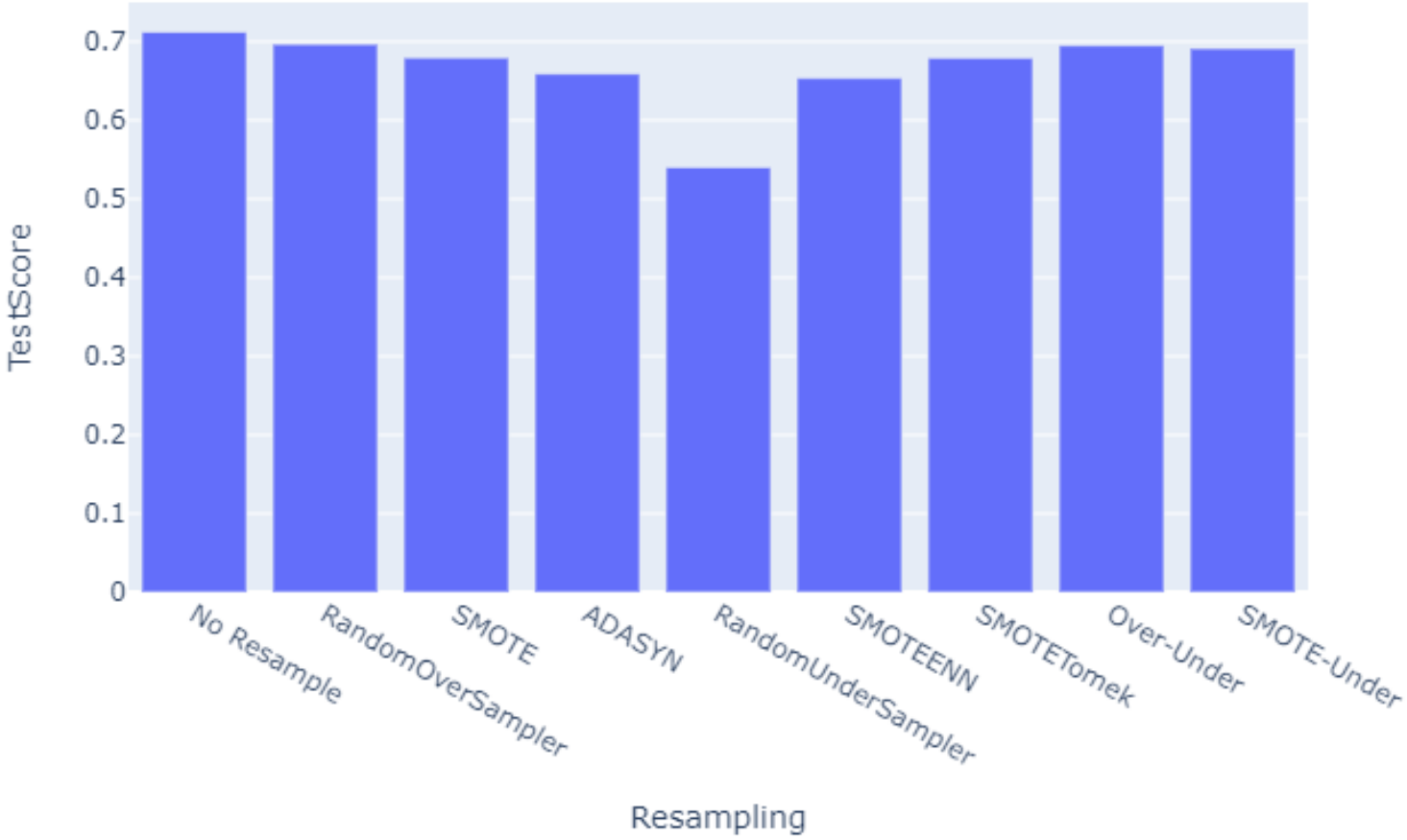
A baseline Random Forest performs the best with original data



RANDOM FOREST – DIFFERENT RESAMPLING TECHNIQUES

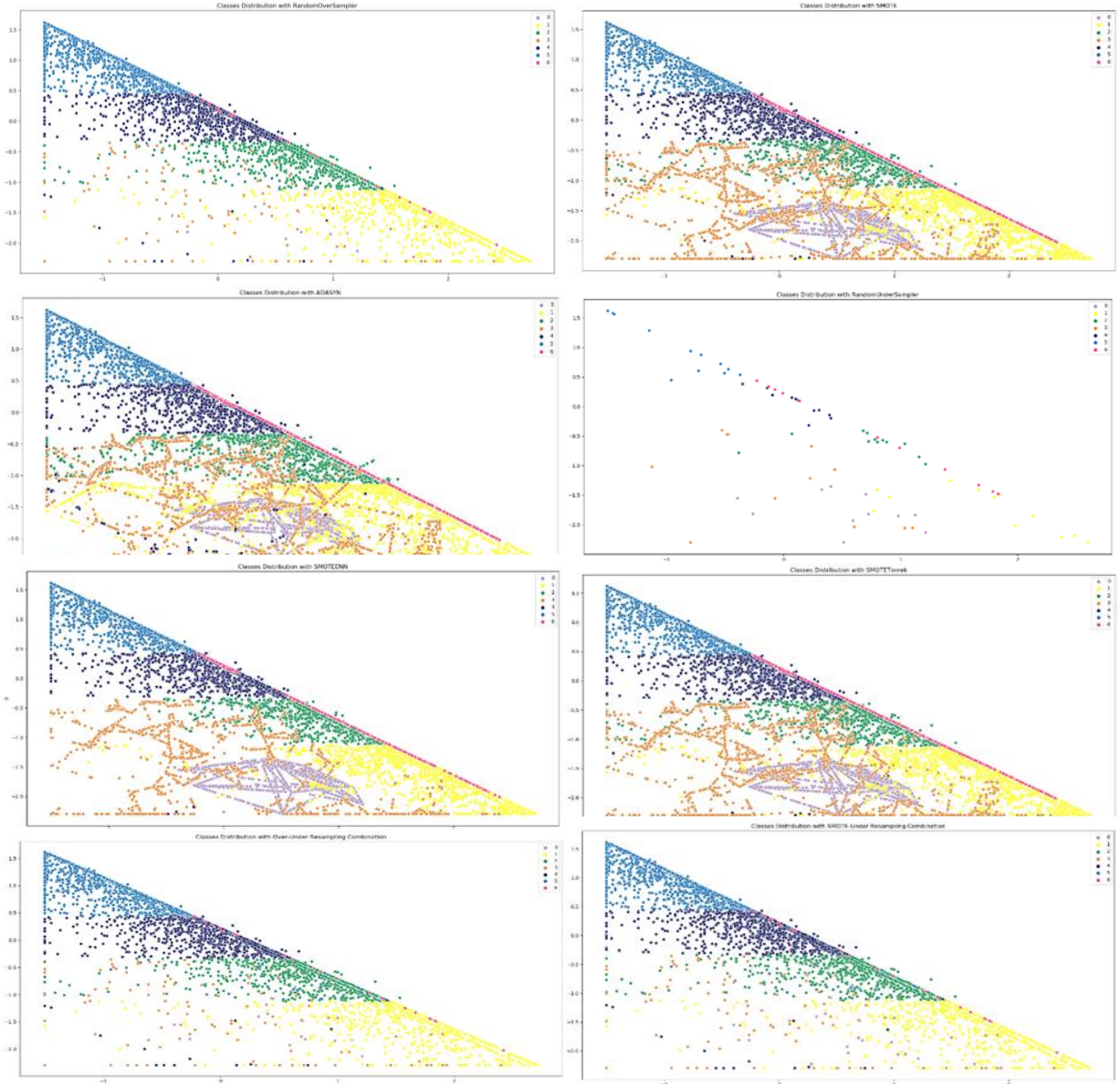
Different resampling technique were applied to the original dataset to see if it will improve the baseline Random Forest

Random Forest with not-resampling data still performed the best.



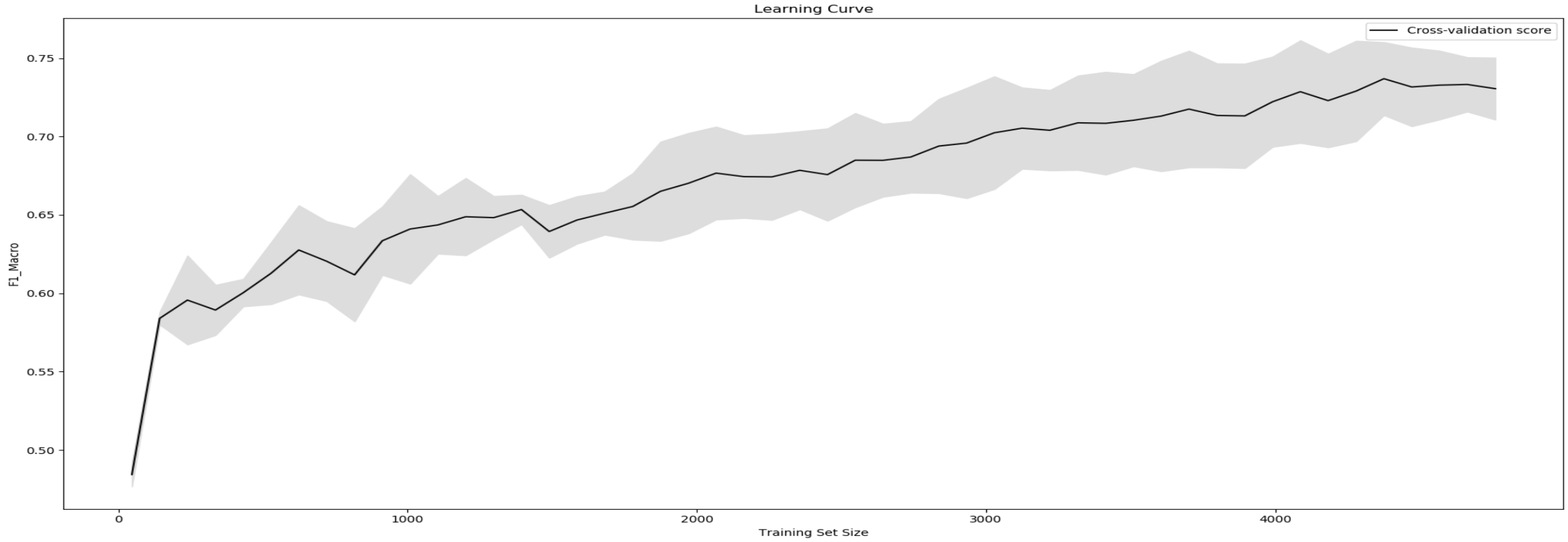
RANDOM FOREST – DIFFERENT RESAMPLING TECHNIQUES (CONT.)

- Minority classes seem to be blended between the groups of majority classes as their physical properties are mixed in between majority classes
- Therefore, resampling technique would not work effectively



RANDOM FOREST – HYPER-PARAMETERS TUNING

Random Forest	N_Estimators	Max_Depth	Max_Feature	Min_Sample _Leaf	Min_Sample _Split	Bootstrap	Macro F1-score
	557	50	2	1	3	False	0.74



CONCLUSION

- Various machine learning algorithms were evaluated to determine the potential algorithm for this dataset.
- Ensemble learning, which includes Random Forest, Gradient Boosting and, Stacking outperformed other algorithms.
- Random Forest were trained with the same dataset but different combination of important features and engineer features. However, Random Forest with the original dataset still performs the best.
- Different resampling techniques were applied to the original dataset. However, they were ineffective because the physical properties of minority classes were very similar to the physical properties of majority classes.
- Random Forest was further fined tune to obtain a macro F1-score of 0.74 was obtained with the optimized model.

RECOMMENDATION

- More data should be acquired, especially the data for the minority classes, if it is important to predict the minority classes.

- Because the physical properties of minority classes are blended between the majority classes, discuss with experienced geologist to see if more features can be used to describe the minority classes better. Therefore, the model can significantly improve.