Lithofacies classification is a process that allows geologists to identify the hydrocarbon-bearing zone. So, the potential resource can be developed and produced. Core samples of rocks extracted from the wells provide reliable results for lithofacies classification. However, these core samples are not always available due to the associated cost. Indirect measurement such as well logging is another alternative method to classify facies. Nevertheless, the interpretation of these well logs is a mammoth task which highly needs experience expert to interpret the measurement and convert it to the meaningful data. Machine learning is a well-suited method to address this challenge. A well-trained machine learning model can predict the facies correctly from the well logs measurements; therefore, it facilitates the process. In this project, a variety of machine learning models with different resampling techniques will be evaluated to determine the best model.

The data used in this project is obtained from the 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. There are two CSV files that contain all the information. They will be loaded into dataframe to reprocess the data. Exploratory data analysis (EDA) will be performed to visualize the data. Plotting box plot for each feature to understand the data distribution. A heat map will be constructed to examine the correlation between each feature. After EDA completed, the data will be standardized to get ready for training the models.

The data will be divided into training and test sets. A variety of supervised machine learning models with default model structures will be used to train and fit the data. K-fold cross-validation will be implemented to evaluate model evaluation. This will establish the baseline performance of the models. Because the dataset exhibits the unequal distribution of classes; hence, it is necessary to investigate whether or not the resampling technique will improve the performance of the model. Several resampling techniques such as “Random Over Sampler”, “SMOTE”, “ADASYN”, “SMOTEENN”,” SMOTETOMEK” and “Random Under Sampler” are combined with the best three baseline models to fit the data. Comparing the model performance before and after resampling technique will help to identify the best model for this problem. Finally, the project report and slide deck will be prepared to show my work along with the codes.