DSCI-633 Final Individual Report, gk5951

The main part that I have played in this Acute Ischemic Stroke Prediction project is mainly to research the machine learning models that are best suitable for this project. As part of my research, I came across Machine Models like GaussiannNb, BernoulliNB, and Extreme Gradient decent Boosting.

GaussianNB is Supervised Machine Learning and can be imported from the sci-kit learn library from python. To use GaussiannNb, The features that we are using should be Continuous and it is also assumed that the features should be following gaussian distribution. Gaussian Navies Bayes need small training data to estimate the parameters needed for classification. GaussianNB Classifiers have simple design and implementation and they can be applied to many real-life situations. This model can be adjusted by simply finding the mean and standard deviation of the points in each label, which is all that is needed to determine such a distribution. At each data point, the z-score distance between that point and each class mean is calculated, i.e. the distance from the class mean divided by the standard deviation of that class. By using the GaussianNB we got an accuracy score of 65 percent and a K-Fold Validation Mean Accuracy of 71.25 percent and has a Standard Deviation of 8.93 percent, ROC AUC Score of 6.77, Precision 0f 0.39, Recall is 1, and has F1 of 0.56.

BernoulliNB is also a supervised Machine Learning Model and can be imported from the sci-kit Learn library from python. Bernoulli Naive Bayes is used for Boolean and Binary features. This is used for discrete data and it works on Bernoulli distribution. This main reason for using Naive Bayes Algorithms is because it works with small data as we only have 200 data points for predicting the severity. The Accuracy score for BernoulliNB is 75 percent and a K-Fold Validation Mean Accuracy of 82.50 percent and has a Standard Deviation of 11.11 percent, ROC AUC Score of 0.68, Precision 0f 0.45, Recall is 0.56, and has F1 of 0.50. BernoulliNB Seems to be working better than the GaussianNB because the accuracy of the BernoulliNB is Higher than the accuracy score of GaussianNB. But he recalls of the first model has better recall than the second one.

I have also done my part helping my teammates to use the XGBoost Machine Learning model. DMatrix is an internal data structure that is used by XGBoost, which is optimized for both memory efficiency and training speed.XGBoost is an advanced implementation of a gradient boosting algorithm with a tree model as the base model. Boosting algorithms iteratively learn weak classifiers and then add them to a final strong classifier. XGBoost Tree is very flexible and provides many parameters. XGBoost model that we used for this project gave An accuracy of 95 percent and a K-Fold Validation Mean Accuracy of 97.50 percent and has a Standard Deviation of 3.06 percent, ROC AUC Score of 0.89, Precision 0f 1, Recall is 0.78, and has F1 of 0.88.

Apart from building the Machine Models, I have helped my teammates with One Hot encoding and Data preprocessing.