Daqian Dang
Proposal
DATS 6103: Introduction to Data Mining
Dr. Amir Jafari
June 22, 2022

Classification problems I selected are supervised learning problems, which the training dataset consists of data related to independent variables and target class. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to help a bank judging credit card applicants as low, medium, or high risk on the approval of issuing credit card.

The Kaggle website I chosen is because there are lots of public classification datasets, and some of the datasets came from the real world. The classification datasets from Kaggle website are the great and useful dataset for data analysis beginner to practice machine learning and improve analysis skill. Python will be primarily utilized for analyzing the large dataset. Python is an interpreted, general purpose, and high-level language with an objected oriented approach and this is a useful tool for analyzing big data.

Since the dataset I selected is a classification problem, the Python libraries could bring me several important analysis tools, such like data mining, data preprocessing, and data modeling along with visualization, which support me to clean, transform, and visualize the data as well as build the model for the data. The resources from Python libraries are Pandas, Numpy, Seaborn, Matplotlib, and Sklearn. In addition, five main model algorithms, logistic regression, naïve bayes, decision tree, random forest, k-nearest neighbors, will be applied to the classification dataset.

To validate and judge the performance of the five models, I will use accuracy score, F1-score, AUC/ROC, and confusion matrix.