Assignment 1

Name: Snehal Sudhir Bhole **Student ID: 300318105**

In assignment 1, I worked on a supervised classification problem by analyzing users' drug consumption patterns to classify them as Drug users and non-user. Given Dataset consists of 12 input and 18 output variables which are central nervous system psychoactive drugs with 1885 samples. Out of these 18 target variables, I am considering the following drugs: Heroin, Alcohol, Meth, LSD, Cocaine, Cannabis, and Ecstasy.

Pre-processing data for binary classification (i.e. User, Non-User Category) makes the dataset relatively more balanced than the original dataset. Since the dataset considered is categorical data, the Chi-Square test is performed on a dataset for feature selection after splitting data. The Chi-square test finds a correlation between each non-negative feature and class, Label Encoder is used to convert negative values to positive. The function returns the f-score and p-value. We consider the p-value score to determine which features to select. The lesser the p-value score more important the feature is. For every classification model, the attributes are selected based on their p-value score.

I have used Decision Tree, Random Forest, K-nearest Neighbour, and Support Vector Machine classification models. Models are trained against train data and evaluated against test data with the help of accuracy, precision and recall data. I analyzed the performance of each model for each drug and analyse the performance of models using the ROC curve and confusion matrix. For Alcohol, all the models tend to behave in a similar manner giving 0 non-users which shows data for the non-user is not sufficient or skewed.

When compared to the paper given[1], writers have used all the classification models except SVM in their analysis. The accuracy of models developed varies with the final result by 15-20 %. The models trained by the authors perform well as they have considered various measures like Euclidean Distance, Adaptive distance, 1-30 nearest neighbours, information gain, Gini gain, DKM gain, Fisher’s discriminant, etc.

The main key learning from this analysis is, accuracy is not sufficient to measure the model performance. Precision, specificity and Recall/sensitivity measures can also be considered to study the model behaviour. The balance between specificity and sensitivity is important. When given data samples are skewed (eg. in Alcohol), it is difficult to determine which model works the best. In that scenario, the class which is of more interest to the business problem helps to determine which model to choose from as the best solution.