Assignment 2

Name: Snehal Sudhir Bhole Student ID: 300318105

In assignment 2, I have worked on various datasets, performed K-Fold validation on those datasets and using Friedman's test and Nemenyi post hoc test, critical differences in performance is determined. First dataset used was drug-consumption dataset. For drug-consumption dataset, I chose the target output column "Cannabis". Feature section was done on the basis of chi-square test. Two of the datasets were derived from over sampling and under sampling the drug-consumption dataset. Other 2 datasets were Labour-Negotiation dataset and heart disease dataset.

Labour data and heart disease data was cleaned and pre-processed. For Labour data, all the columns with missing value percentages greater than 20 are dropped. Missing values for other columns are replaced with mode of the value. Later, chi-sqare test is performed for feature selection on the input data. Similarly, heart Disease data was processed and scaled. Every dataset is evaluated against following 6 algorithms: Random Forest, Decision Tree, Support Vector Machine, K-nearest Neighbour, Multi-layer perceptron and Gradient boost algorithm.

The dataset DB1 was derived by oversampling the Dataset D using SMOTE while the Dataset DB2 was generated by under sampling the dataset D using Edited Nearest Neighbours approach. The K-value for experiment is 10. The data is split in 10 partitions and model is trained and evaluated on each partition (9:1). Finally, the mean accuracy for each algorithm is measured and their corresponding graphs are analysed. I noticed that, performing under or over sampling on the skewed dataset improved the model accuracy.

For the K value 6, and N value 5 Friedman's statistic value generated is 12.133 and the chi-squred statistic for alpha 0.05 is 10.49, thus there was significant difference observed between the algorithms. Hence, Nemenyi Test is performed to determine which algorithms performed similarly and did not show significant difference in their accuracies. Critical difference after Nemenyi test is 3.371. From the final graph, it can be seen that KNN does not lie in the critical difference range and rest 5 algorithms demonstrate similar performance. Also, if we leave Random Forest, as per the graph, the rest 5 algorithms have similar performance.