Assignment 2

Submitted by,
Swati Verma

Steps performed during Stage Classification of Liver Cancer

1. Training and testing dataset are read.
2. For normalizing the data in both training and testing dataset, formula used is logBase2(1+ value of each cell)
3. Min max scaling is also performed so that the data can range between 0 and 1.
4. After that variance of each column have been calculated and all those columns where variance was lower than 0.025 are dropped from both the training and testing dataset.
5. For feature selection, Select k Best function from sklearn library is used and features with top 100 scores are selected.
6. For dimensionality reduction Principle Component Analysis i.e. PCA has also been applied.
7. At the end several classification models like SVM, Random Forest, Multilayer Perceptron Model and Decision tree classifiers are applied.
8. I am getting the best score using Decision Tree Classifier.
9. After applying classifiers predicted values are stored in the output file and is uploaded into Kaggle submissions.

To run the python notebook file:

1. Install python version greater than 3.7.
2. Install jupyter notebook and other modules like panda and numpy.
3. Open jupyter notebook and the run all the cells one by one