CMP 466 – Machine Learning and Data Mining- Spring 2021

Team Project- Assignment 2

Please submit a report including the following by the deadline:

1- [10 points] Read your data into your program. Save it in 2 parallel matrices, one for features (X), and one for labels (y).

2- [20 points] Perform random splits on the dataset you have to divide it into training set (80% of the data) and testing set (20% of the data). Repeat the random splitting of the data over and over (at least 10 times) to make sure the results you get are not based on a single split. Report the results for each split (point 4 below)

3- [20 points] Perform k-fold cross validation on the data with k= 5 and/or 10. Report the results for each fold (point 4 below).

4- [40 points] Apply decision tree (DT) classifier on your dataset and report the training and testing accuracy using the models built as described in points 2 and 3 above. Choose the best tree size by tweaking the DT hyper parameters. Make sure your DT is not overfitted or underfitted. Produce a plot showing the DT complexity (no of nodes) vs. the training and testing accuracy. Comment on the best number of nodes for your tree.

5- [10 points] Plot your best fitted tree. Summarize and discuss your results.

Submission: Please submit your assignment to iLearn by the deadline. Late submissions will be penalized according to the syllabus late policy. No submissions will be accepted after four days of the deadline.