Indian Institute of Technology Delhi
Department of Mathematics
II Semester 2021-2022
Assignment
Weightage 30% Due Date 28th Match 2022

Each question carries 15 marks.

You need to form a group of 3 members and jointly complete the assignment using Python Jupyter Notebook.

Q1. Choose a data set from UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets.php?format=&task=cla&att=&area=&numAtt=greater100&numIns=&type=&sort=nameUp&view=list) (or any other Source) for Multi class classification problems.

(i) Your first task is characterize the data set. Answer the following questions about the data: [3]

1) What the data is about.
2) What type of benefit you might hope to get from data mining.
3) Discuss data quality issues: For each attribute,
   a) Are there problems with the data?
   b) What might be an appropriate response to the quality issues.

(ii) Implement (1) Decision Tree, (2) Random Forest, (3) Naïve Bayes Classifier and (4) KNN classifier and compare the performances using k-fold cross validation and other tuning techniques. [12]

Q2. (a) Implement Apriori and FP-growth algorithm. Cite any sources helpful to you for implementing the algorithms. [12]

(b) Modify the algorithms to achieve the same task (preferably with some improvement) . Clearly mention the difference in the modified algorithm. [3]

Note: You may use the datasets at http://fimi.ua.ac.be/data/retail.dat to evaluate your implementations

---