1   Instructions

1. You can use any programming language of your choice with the suitable input format for each of the question. Each question should have a readme file and the detailed report.

2. A readme file should precisely tell how to compile and run your program. Give the exact commands with respect to the datasets provided.

3. The marks will be given on the basis of quality of code, use of innovative data structures, scalability, correctness, and completeness of the report. Please submit your code along with the readme file and the report.

4. You are supposed to submit the assignment on google classroom no later than **9th November 2020**. This is a strict deadline and any assignment submitted later will not be consider for evaluation unless you take prior permission (at least 4 days before the submission deadline).

   Datasets to be used:

1. Adult Dataset: https://www.kaggle.com/qizarafzaal/adult-dataset

2. Credit card fraud dataset: https://www.kaggle.com/mlg-ulb/creditcardfraud

Problem 1:

Credit card fraud dataset is a highly imbalance dataset meaning the number of examples having class 1 (fraud) is much less as compared to the number of examples having class 0 (not fraud). The goal of this problem is to understand given this database which of the following three probabilistic algorithms work better. Implement the three probabilistic classifiers and choose the appropriate metric to determine the efficacy of your classifiers. Clearly explain the results obtain and give the insights.   [25 Marks]

1. Bayes classifer

2. Naive Bayes Classifier

3. K-NN classifier with appropriate value of K

2   Problem 2:

Implement decision tree algorithm to perform classification on Adult dataset to classify whether the income of a particular user exceeds $50K per year or not. Divide your data into two sets: Training data (80%) and Test data (20%). Plot the training error and test error against the number of nodes present in the decision tree. Present the optimal decision tree in your final report.   [20 Marks]