CS 4038D Data Mining - Assignment (20 marks) – Winter 2020-2021
Due by 24/04/2021 at 10PM

Instructions:

1. You must create a single PDF file containing screenshots along with necessary descriptions of each step executed in completing the assignment.
2. You must also submit all the source code files you have created.
3. All the above mentioned files must be put into a folder, zip the folder and upload the zipped file. Zipped and PDF files must be named in the format <<Firstname_Rollno>>
4. You MUST submit your original work only. If plagiarism is found, then all the submissions involved (irrespective of any claim that *it is my original work* etc) will be getting **zero** marks.

Download a dataset (that require data cleaning) of your interest from any of the following websites for classification studies.

1. https://www.kaggle.com/datasets

2. https://archive.ics.uci.edu/ml/index.php

A. Give a brief description about your dataset like – about the problem, how may attributes, how may samples, how many classes, distribution of the classes.

B. Analyse your data and apply data cleaning techniques wherever required and transform the data if required. You must use **OpenRefine** (https://openrefine.org/) or any other similar tools for this purpose.

Divide the cleaned dataset (D) into training and test sets according to some standard techniques. Give the details of the method you have followed and the reason also.

Now execute the following questions and create a detailed result analysis for each. You may use Python/R for implementations.

Q1. Run Decision Tree, Naïve Bayes, KNN and ANN algorithms and compare their performances using confusion matrix and other metrics like precision, recall and F-score. Report the training and test errors also. For each classifier, mention the values of various parameters used, wherever applicable.

Q2. For ANN plot the loss function values against epochs. Compare the performance of ANN with any three different activation functions. What are your inferences?

Q3. In case of ANN, plot the test set error for different numbers of hidden nodes, like 1, 2, 3, square root of number of features, and half of the number of features. Compare the results and what inference you can make from this.

Q4. Start with your original cleaned dataset. Now create three datasets, D1, D2 and D3 containing 1/4, 2/4, and 3/4 portions of samples from the original dataset. Now create your training and test datasets and repeat **Q1**. Compare the results obtained for D1, D2, D3 and D. Give your inferences with justifications.