Last Name: ____________________ First Name: ____________________ Student ID: __________________

AIDI 1002: Machine Learning Programming — Assignment - 3

Due Date : December 10, 2022, 11:59 PM

Note : Submit two files in the submission folder. First is your colab notebook including your code and outputs and second is the pdf of colab notebook with the following naming convention for both the files.
(File name : *Assignment_3_firstname_lastname.pdf/.ipynb*)

1. Design a deep learning experiment for a multi class classification dataset `https://www.kaggle.com/datasets/abisheksudarshan/customer-segmentation`. It is a multi class classification task where "var_1" is a class label column having 3 categories as Cat_6: 65%, Cat_4: 13%, Other: 22%. There is a slight imblance in class distribution. This link contains two files 'train.csv' and 'test.csv'. You need to divide the 'train.csv' in appropriate percentage to get the validation set. Your experiment should involve following step in appropriate order.

   1.1 Shuffling of the data before training (2 points)

   1.2 Design and train a neural network model (e.g. you can use DNN network or if you want to use any other models it is also acceptable) (10 points)

   1.3 Use validation data for model tuning and monitor the f1-score while applying the early stopping logic from keras library (10 points)

   1.4 Use test data to calculate the appropriate classification metrics. (5 points)

   1.5 Explain the significance of each metrics. e.g what recall denotes in terms of multi class classification. (3 points)

   1.6 Generate the loss and f1-score curve for training and validation set. (10 points)

   1.7 Generate a ROC-AUC curve and comment on your model accuracy and find the optimal threshold from the curve. (10 points)

   1.8 Repeat the steps from 1.1 to 1.7 with sampling in training set. (you can do over sampling to increase the instances of majority class in training set) Compare and comment on the results you get from sampled data and original data distribution. (50 points) (You are expected to do some research on how to apply sampling over a dataset and which libraries usually employed to do so.)