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

AIDI 1002: Machine Learning Programming — Assignment - 2

Due Date : November 23, 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_2_firstname_lastname.pdf/.ipynb*)

1. Consider this dataset from kaggle. (Download the dataset from following link : `https://www.kaggle.com/shrutimechlearn/step-by-step-kmeans-explained-in-detail/data`) and answer the following questions :

   1.1 Perform k-means clustering over this dataset using Manhattan distance as the distance-measure. (10 Points)

   1.2 After performing k-means clustering, extract the groups or clusters and add a separate column in your dataset as 'Labels' and fill it with cluster number assigned by k-means algorithm. (5 Points)

   1.3 Now, you should be ready with your labeled dataset. Perform standard classification task using logistic regression, decision trees, random forest, and Naive Bayes algorithm. (25 Points)

   1.4 Compare the performance of these various supervised learning algorithm and comment on the homogeneity of clusters, like is the clusters or groups are making sense or not ? (10 Points)

2. Consider the breast_cancer dataset given in the sklearn library and answer the following questions.

   2.1 Import the breast_cancer dataset from sklearn.datasets library. (5 Points)

   2.2 Perform PCA (2 components) and LDA (1 components) over the dataset. (20 Points)

   2.3 Visualise the components and see if its able to segregate the class label in breast_cancer dataset. (10 Points)

   2.4 What is the maximum variance explained by both the components in PCA and LDA. (10 Points)

   2.5 Comment on the working of PCA and LDA and which one is better for breast_cancer dataset. (5 Points)