CS4092D Machine Learning Lab
Module4(PCA & LogR): Exercise
PCA and Logistic Regression
S4 MCA: Winter 2022-23

Date: 03-Feb-2023

Q.

a. Implement the classical dimensionality reduction technique called **Principle Component Analysis (PCA) from scratch**. Compare your implementation of PCA with the PCA() function available in the Python library module ‘sklearn.decomposition’. You may compare both by plotting the percentage loss in data as you keep deleting the principal components.

b. On the reduced dataset using PCA (you may reduce the dimensions till you have at most *20%* loss in the data variance), perform a binary-classification task on the shared dataset using the **Logistic Regression** technique. For the same, **implement the Logistic Regression model from scratch**. Compare your classification model’s performance (use Precision, Recall & F1-score) with respect to the LogisticRegression() function available in the Python library module ‘sklearn.linear_model’. You must also compare the performance of running the LogisticRegression() function defined in Python on your implementation of PCA *Vs.* Python’s PCA() function.

Dataset: We have shared the ‘adult.csv’ file which is to be used for implementing this problem. The dependent variable in the dataset is the feature ‘income’ with binary values. **NB:** There are missing values in many of the features, and you may impute them using the *mode* values of the respective features. Also note that, some features may not carry any meaning/weightage in predicting the dependent variable, and such features you may remove from the dataset. As always, the train-test split remains the same: 70%-30%.