Subject: 21DS602

Lab Session: 12

Lab Session Date: 04/01/2023

Notes:

1. Please read the assignment notes carefully and comply to the guidelines provided.
2. Report should be submitted to TurnItIn. Once done, please submit your assignments in Teams.

Main Section (Mandatory):

Please use the data associated with your own project.

Refer:

- **https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html**
- **https://scikit-learn.org/stable/modules/generated/sklearn.manifold.MDS.html**

Main Section (Mandatory):

A1. Perform correlation analysis between the features for your project data. Plot the correlation matrix as a heatmap plot. Study the heatmap plot to understand the correlation existing between the features.

A2. Employ PCA for feature decorrelation. Find the count of attributes in PCA transformed space which cover 100% variance.

A3. Employ these features to perform your classification or regression work. Check the scores of classification / regressions and compare them to the same obtained from the raw data (original vector space).

A4. Find the count of features covering 95% of the variance. Take these features and repeat A3.

A5. Perform correlation between each individual feature with the output variable. Study the correlation and understand the importance of each feature for output prediction.

A6. Perform PCA on the feature set after Z-score normalization of the features. Repeat exercises A1 to A5 on z-score normalized features. Note your observations.

Optional Section

O1. Study MDS and its effect on observation vectors. Study the impact it creates for attaining your goal (classification or regression). Compare its performance for goal attainment against other transformation (scaling, feature selection, feature reduction with PCA etc.) techniques employed so far.

Report Assignment:

1. Describe a scenario where PCA would not be able to help reduce the dimensionality of the available training data. [1]
2. Compare and summarize the effects of various data transformation techniques on your project. Which is the most suitable technique for your project data. Justify your answer. [3]