ProgDS-2025 Exercise – 8:

Concepts covered:

1. Data imbalance
2. Correlation coefficients and correlation heat-map
3. Variance Inflation Factor (VIF)
4. Principal Component Analysis
5. Visualization using PCA
6. t-SNE

Part-1 : Data Imbalance

1. Review the Notebook **Handling_Data_Imbalance.ipynb**
2. In a document explain the code in each cell, and interpret the output of each cell
3. Experiment by making your modifications to the code

Part-2 : Correlation, VIF and Dimensionality Reduction using PCA

You are given the following two data sets:

1. **patient-data.csv**
2. **curse-of-dimensionality.xlsx**

(Note: the 'curse of dimensionality' data set is an xlsx file with multiple tabs, each representing a situation. Understand the situation prior to performing your analysis)

Carry out the following steps on both these data sets:

1. Decrease the number of observations considered for model building and observe the impact on model metrics like R2, MAE, precision, recall, accuracy, F1-score, etc. (whichever metrics are applicable to each of the data sets)
2. Perform pair-wise correlation analysis on all the independent features, create the correlation heatmap, and state your observations
3. Perform VIF analysis on all the independent features and state your observations.
4. Based on progressive VIF analysis, drop the features one at a time, keep track of the its impact on the model metrics (by plotting the metrics' trends). State your final conclusions.
5. Perform PCA on the independent variables and create a bar chart of the contributions of the PCs to the overall variance in the data set.
6. Create ML models, regression or classification, as the case may be, using appropriate subset(s) of PCs. Compare the metrics generated by PC based models (eg. ML models with 3 PCs v/s ML models with 5 PCs, etc.). Also compare these metrics with the metrics of ML models created using the original features. What can you conclude?

Part – 3 : Visualization (2D) Using PCs and t-SNE

Use the data set **mnist_test_nolabels.csv** in the following tasks:

1. Carry out Principal Component Analysis on the data set and create a bar-plot of variances explained by the PCs. Also create a second bar-plot, this time showing the *cumulative variance explained curve* in addition to the bars.
2. Visualize the data set by creating a scatter plot: PC1 v/s PC1
3. Pass the data set through t-SNE to create two t-SNE components. Create a scatter plot of these two components to visualize the data set.
4. Sate your observations and learnings.

oooOOOooo

This file is meant for personal use by aashishktrivedi@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.