Machine Learning Project 1

A) Investigate Clustering (70 points)

Use

- Kmeans,
- DBSCAN,
- BisectingKmeans(`https://scikit-learn.org/stable/modules/generated/sklearn.cluster.BisectingKMeans.html`)
- HDBSCAN (`https://scikit-learn.org/stable/modules/generated/sklearn.cluster.HDBSCAN.html`)
- OPTICS (`https://scikit-learn.org/stable/modules/generated/sklearn.cluster.OPTICS.html`) to cluster the Iris data, cybersecurity malware data, and breast_wisc_data, data. Then
  - compare their Silhouette Coefficients, V-measure
  - adjusted random index (ARI),
  - Normalized Mutual Information (NMI)
  - and adjusted mutual info_score (AMI).
- What kind of conclusion you can get?
- You are required to use at least 4 normalization methods to preprocess data

B) Yellow brick (50 points)

- Yellowbrick is powerful visualization package for ML. Develop at least 20 pages slides to demonstrate how to use different visualization and data analysis methods in yellow brick to
  - visualize iris, cybersecurity malware data, HFT APPL data, breast_wisc_data, and a dataset you pick.
  - Rank and visualize the feature importance of the datasets

What should you turn in?

- 1. A folder that contains
  - A ppt to show details of your analytics (at LEAST 40 pages)
  - your data
  - source files
  - corresponding related output.
  - A link to your presentation video
- 2. Send the zipped file (.zip instead of ,rar)