Assignment 2 - PySpark Data Preprosessing and Clustering

Submission Deadline: ***20th April 2022***

Submit the code and a Report in the form of a word document on google classroom.

Question 1: Explore and preprocess the "Leaves" data given along with this assignment

Load the given dataset into ***Spark Data-Frames*** and perform the following preprocessing on it.

a) Handle missing values
b) Identify if an attribute has outliers or noise
c) Apply measures of the central tendency and dispersion to **analyze numeric attributes**. That is, compute the mean, median, mode, range, variance, correlation for the attribute. Don't just give values explain analyze them.
d) Would you apply preprocessing techniques like discretization or normalization on any attribute? Explain your answer. If yes, then apply the technique and share the results.

Question 2: CLUSTER THE DATA using the PySpark built-in K-means and bisecting K-means clustering algorithm (this is provided in the Spark ML Library).

a) Cluster the given dataSet using atmost **3-4** attributes to avoid curse of dimensionality. You can select the attribute based on the preprocessing.
b) Cluster the dataset for different value of K

Run your algorithm for various values of K and different values of convergence, show the results in your report.

Use measures such as SSE(the sum of square error), silhouette co-efficient, and NMI (normalized mutual Index) to analyze the clustering results.

See Scikit for more info on the above measures
https://scikit-learn.org/stable/modules/generated/sklearn.metrics.normalized_mutual_info_score.html