Data Mining Project

By:

Deena Fathi Mohamed 20201447217

Amgad Mohamed Ahmed 20201446888

Ahmed Mohamed Hassan 20201321583

The required imported libraries:

- import numpy as np -> used for working with arrays.
- import pandas as pd -> used for importing multiple file formats and converting an entire data table into a NumPy matrix array
- import seaborn as sns -> used for making statistical graphics
- import matplotlib.pyplot as plt -> used for data visualization and graphical plotting
- from sklearn.metrics import accuracy_score -> for calculating accuracy
- from sklearn.naive_bayes import GaussianNB -> for calculating Naïve Bayes
- from sklearn_extra.cluster import KMedoids -> for calculating K-Medoids
- from sklearn.cluster import AgglomerativeClustering -> for calculating Agglomerative Clustering
- from sklearn.model_selection import train_test_split -> for splitting the data into test and train

Dataset used:

Rice types

About the data set:

There is two types of rice and many attributes differentiating between them, the types are 0s and 1s since there is only two types.

Data:

We first check on the data and clean it from any nulls (missing values) after that we split the data into x and y where x the attributes about the types of the data and y is the types of the rice.

Code:

We boxplot the data before and after cleaning it, then splitting the data into train and test so that we send most of the data to it to train the model then the rest of the data to test the data.

We then calculate the naïve bayes, k-medoids and agglomerative clustering after calculating each, we calculate the accuracy score for each model to see which one did best. Lastly, we plot the y-pred and y-test to se how the model did.

Conclusion:

We found out that naïve bayes algorithm is the best one with 98.5 accuracy then agglomerative clustering with 95.3 accuracy then k-medoids with 93.5 accuracy.