CSE382 Data Mining and Business Intelligence

Project Proposal

Submitted to:

Dr. Nourhan Zayed

Eng. Mahmoud Soheil

Submitted by:

Youssef George Fouad	19P9824
Kerollos Wageeh Youssef	19P3468
Jannah Ayman Amir	19P1728
CESS	Senior-1

1. GOAL

The goal of this project is to compare the accuracy of several classification models that classifies the chosen dataset "**Fruit 360**". Throughout this project, we are going to visit different descriptive and predictive data mining tasks starting from the Data exploration and preprocessing till visualizing the results.

2. THE DATASET

The dataset is a high-quality, dataset of images containing 118 types of fruits and vegetables, with a total of over 60K, 100x100 pixels pictures. It would be divided into test, train, and validation sets throughout the followed project approach.

3. APPROACH

As mentioned above, several data mining tasks will take place throughout the process as follows:

- **Data Exploration:** exploring our dataset by visualizing it using different basic graphical statistical description of the data.
- **Data Preprocessing:** where data needs to be processed on several steps.
    - *Data cleaning*: to remove noise, outliers found during exploration.
    - *Data reduction*: including dimensionality reduction, where we will use principal component analysis PCA, numerosity reduction, and data compression to reduce processing time.
    - *Data transformation*: to transform each training sample image to linear features vector.
- **Training Classifiers:** train models using different classification techniques SVM, Decision trees, and KNN, while tuning their hyperparameters to reach the best results.
- **Testing Classifiers:** test the trained models using different testing approaches including a simple 75%-25% approach and a 5-Fold cross validation approach, to determine the different models' accuracy.
- **Results Visualization:** to compare between the obtained accuracy results in order to deduce the best possible approach for our classification problem on the given dataset.