

Data Mining CSE 385

**Final Project**

Submitted To: **Engineer Ahmed Hesham**

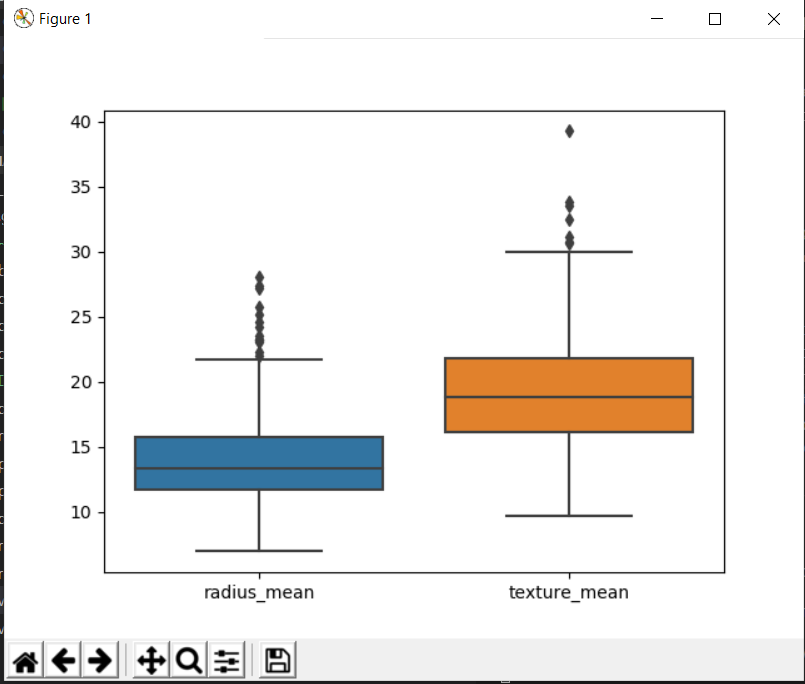
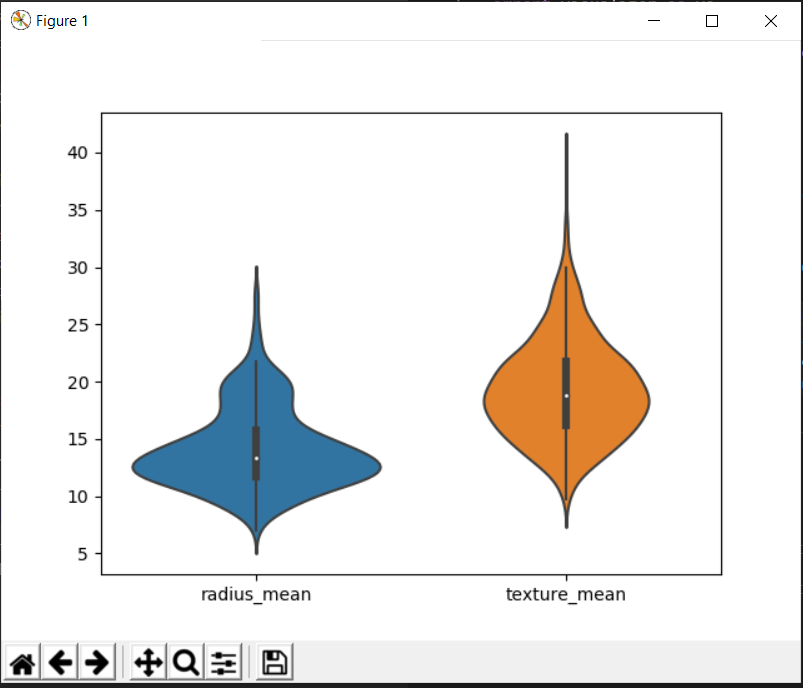
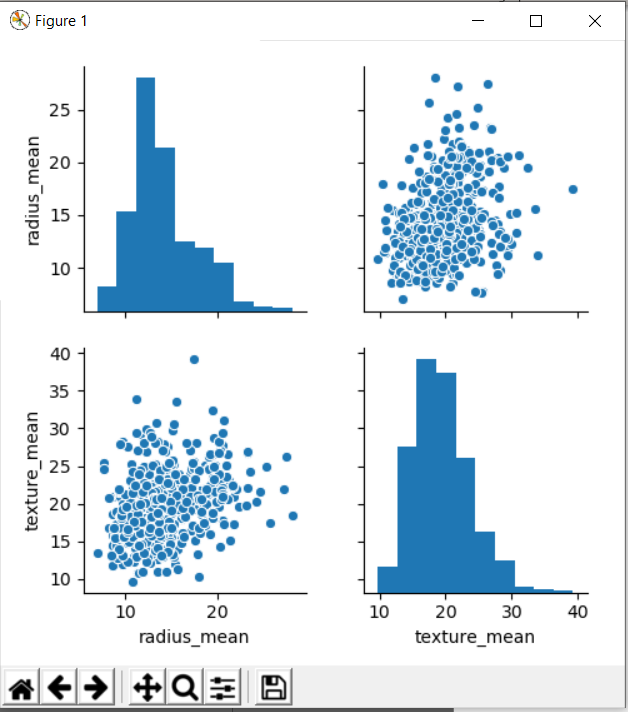
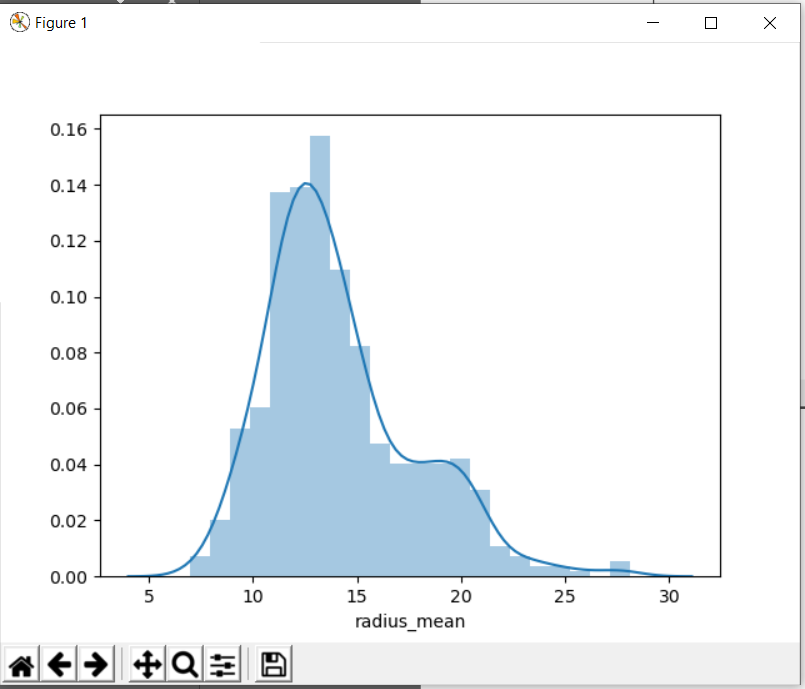
Submitted By:

**Mohamed Sameh 16p3061**

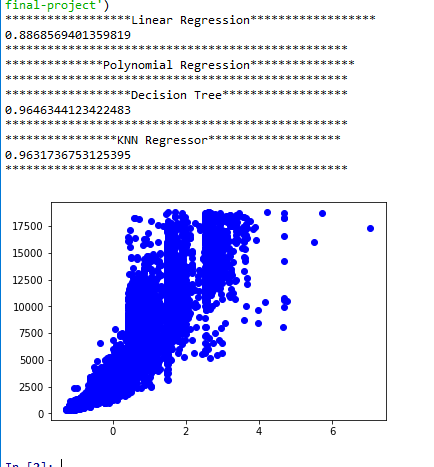
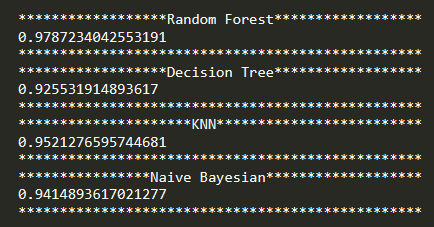
**Nour E-din Osama 16p6043**

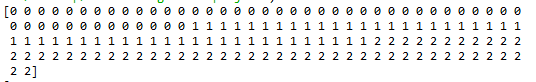
# Bonus Features

We have implemented the following bonus features to our project

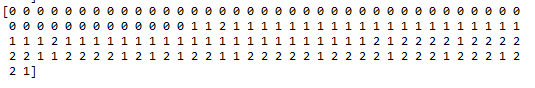
1. **Visualization:** (box plot, violin plot, pair plot, histogram)  
     
     
     
     
     
     
   
2. **Preprocessing:** We have added two extra options to **handling missing values**, insert mean and insert mode.
3. **Regressor:** We have implemented an extra type called Decision Tree Regressor
4. **Classifier:** We have implemented the **Random Forest** classifier, which is the same as decision trees except it uses multiple decision trees as to avoid the overfitting problem of a normal decision tree.

# Results

1. **Regressors:** for the regressors we used the Diamonds dataset from Kaggle and got the following results.  
   
2. **Classifiers:** for the classifiers we used the Wisconsin Breast Cancer dataset from Kaggle **after** removing the irrelevant columns like the ID and got the following results.  
   
3. **Clustering:** for the clustering we used the IRIS dataset after removing the class label and got the following results.

**Real Clustering after encoding**  


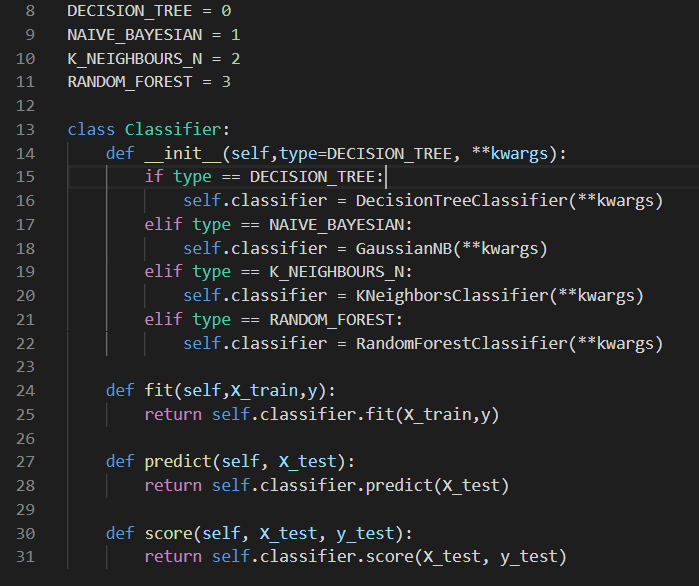
**Output Clustering**



# Implementation

## Classification

The classifier was built as a wrapper around the SKLEARN library implementation for various classifiers which are: Decision tree, Naïve Bayesian, K Nearest Neighbors, and Random Forest. The classifier class passed a dictionary of arguments which have to be submitted by the user according to the selected classifier type.



## Clustering

//write clustering here

## Regression

//write clustering here

## Visualization

The visualizer class used predefined graphs from SEABORN and MATPLOTLIB that provides a reliable API for data exploration and extracting important features and discarding noise.

We have used 4 different plots:

1. Pair plot: allows to explore relations between different attributed in a dataset.
2. Box plot: represents the interquartile range for all variables to be able to determine if outliers exist and gain important information about the distribution of the data
3. Violin plot: shows the interquartile ranges for each variable while providing some representation for the attribute distribution at different values.
4. Histogram: shows a distribution of each attribute on it’s own for deeper analysis of each attribute.