Cairo University
Faculty of Computers and Artificial Intelligent

pattern project

Heart Attack & Prediction

Name	ID	section
Basel youssef Abd-ElAziz Hamed	20200111	S1,2
Aya Mohamed Ehab Abbas	20201042	S1,2
Eman Mohamed Sobhy Abdulghany	20201039	S1,2
Mahmoud Ahmed Mahmoud El-komy	20200491	S1,2
Ahmed Essam Ahmed Moustafa	20200028	S1,2

Under supervision

Dr. Mona Soliman & Dr. Moustafa Hosny

Dept: Information Technology “IT”

Analysis of Heart Disease Dataset using K-means Clustering and PCA

The aim of this report is to describe the methodology used for analyzing a heart disease dataset using K-means clustering and PCA (Principal Component Analysis). The dataset contains various features related to heart health, and our goal is to identify patterns and clusters within the data. The implemented code provides a step-by-step approach to perform K-means clustering and visualize the results using PCA.

So let`s go through the explanation of the methodology used to solve this problem:

1. **Data Preprocessing:**
- The initial step involves loading the dataset and performing necessary preprocessing steps.
- One-hot encoding is applied to categorical variables in the dataset using the pandas get_dummies function.
- The target variable is separated from the features.

2. **K-means Clustering:**
- The preprocessed data is split into training and testing sets using the train_test_split function from scikit-learn.
- Standardization is performed on the features using the StandardScaler from scikit-learn.
- K-means clustering is applied using the KMean_Fitter function to find clusters within the data.
- The optimal number of clusters (K) is provided as an input parameter.

3. **PCA Visualization:**
- Principal Component Analysis (PCA) is applied to reduce the dimensionality of the data.
- The PCA function from scikit-learn is used to extract the principal components.
- The reduced data is plotted using scatter plots to visualize the clusters obtained from K-means clustering.
- The centroids of the clusters are also plotted in the same scatter plots.

4. **Model Evaluation:**
- Accuracy scores are calculated for the K-means clustering model using the accuracy_score function.
- The accuracy is computed separately for the training and testing datasets.
- Additionally, the accuracy of other models, such as Random Forest and Decision Tree, is evaluated as a point of reference benchmark.

5. **Comparison:**

- Random Forest and Decision Tree classifiers are also used.
- After testing the data and applying four different classifiers, we compared them with each other to see which classifier is better for solving the problem.
- Then we plot the four classifier accuracies` to see which method will give us the highest accuracy and is better for this problem.

In this part we will show the results of our implementation:

a) K-means Clustering:

- The K-means algorithm is applied to the heart disease dataset, resulting in the identification of clusters.
- The centroids and labels of the clusters are stored in KMean_Centroids and KMean_Labels, respectively.

b) PCA Visualization:

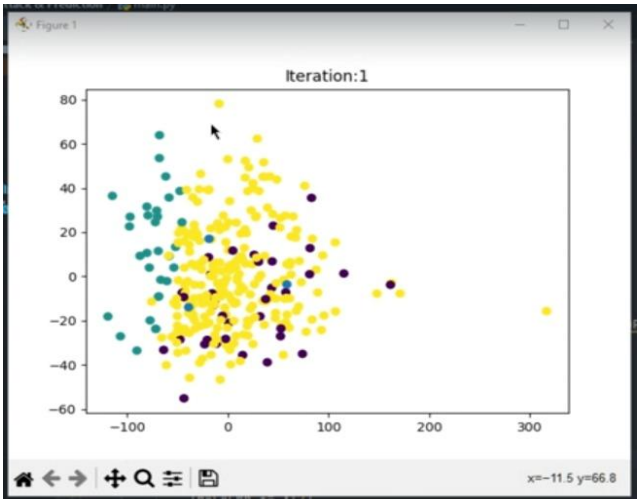
- Principal Component Analysis (PCA) is applied to the preprocessed data for dimensionality reduction.
- The reduced data is plotted using scatter plots, with each data point colored according to its assigned cluster.
- The centroids of the clusters are also plotted in the same scatter plots, providing visual insights into the clustering results.

c) Model Evaluation:

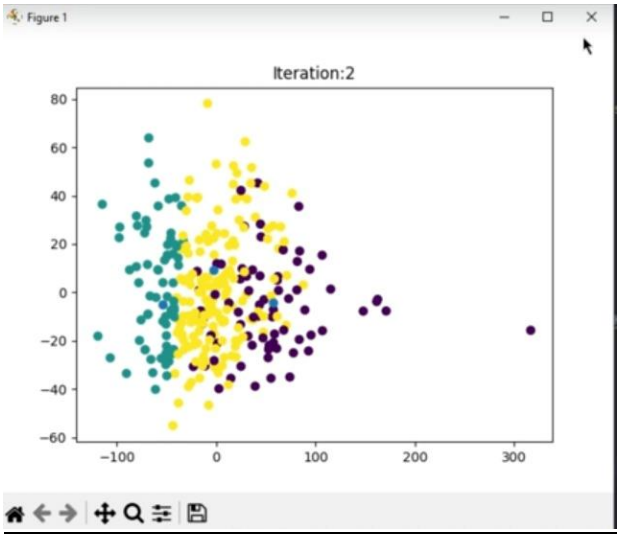
- The accuracy scores of the K-means clustering model are calculated for both the training and testing datasets.
- The accuracy scores of benchmark models, including Random Forest and Decision Tree, are also computed.
- The accuracies of these models are compared to test assess their performance on the heart disease dataset.

This is an example of the output, each plot represents each iteration where the cluster data compared to a new centroid every time and it will stop iterating when it reaches the optimal centroid (old clusters = current clusters).

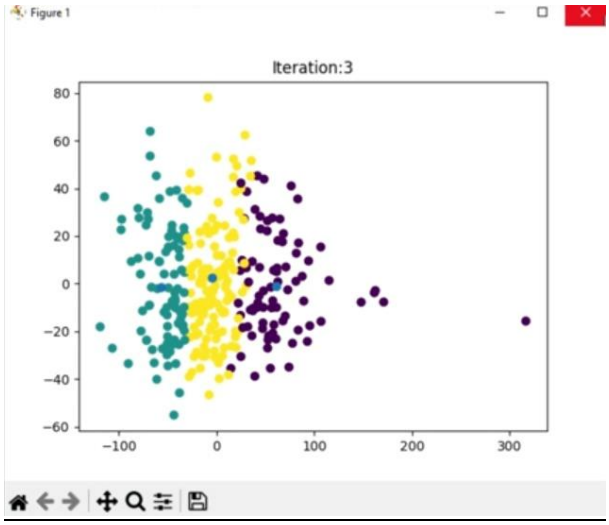
The first iteration:



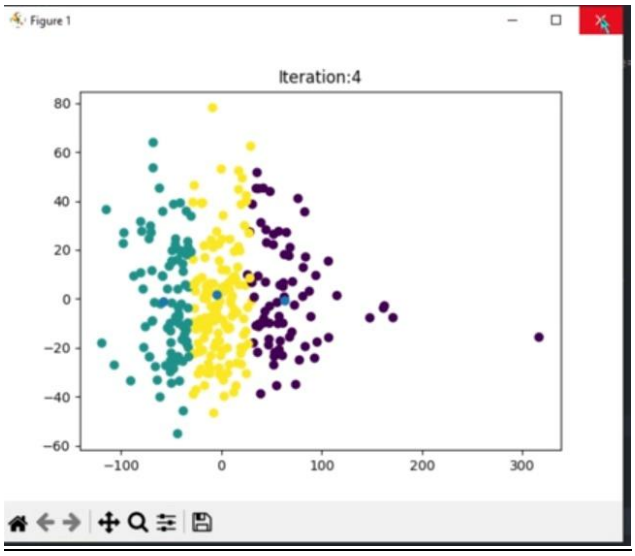
The second iteration:



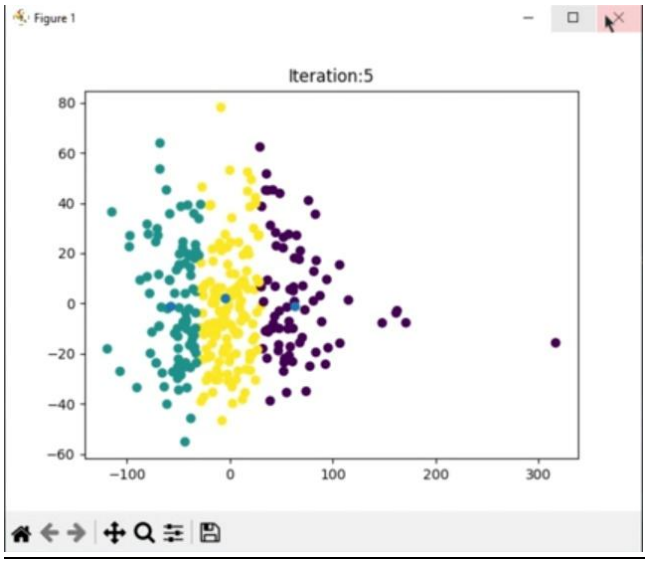
The third iteration:



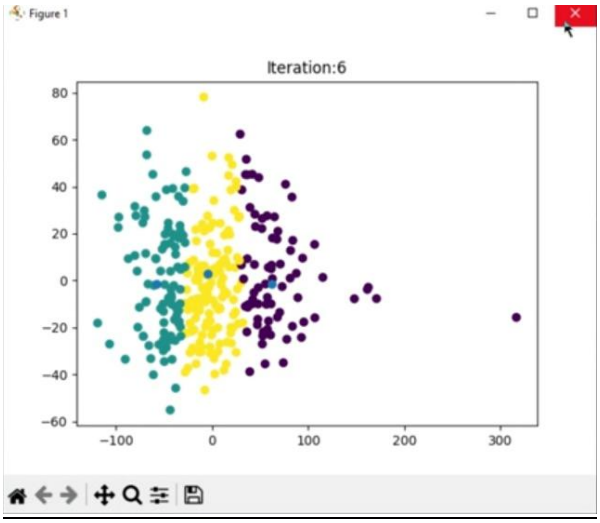
The fourth iteration:



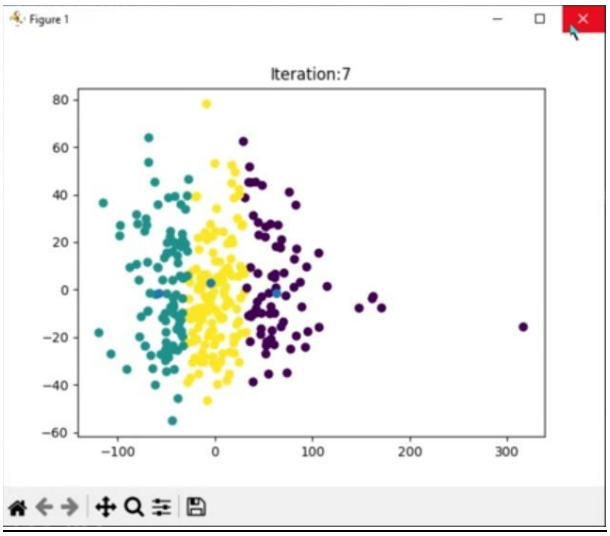
The fifth iteration:



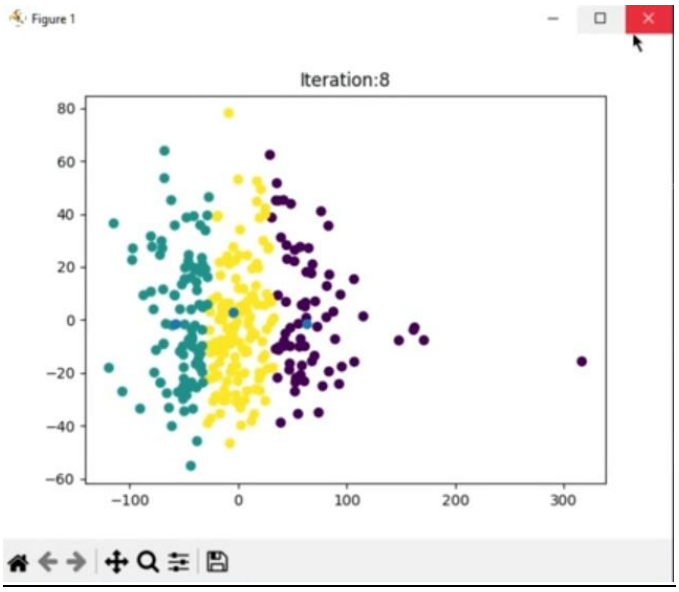
The sixth iteration:



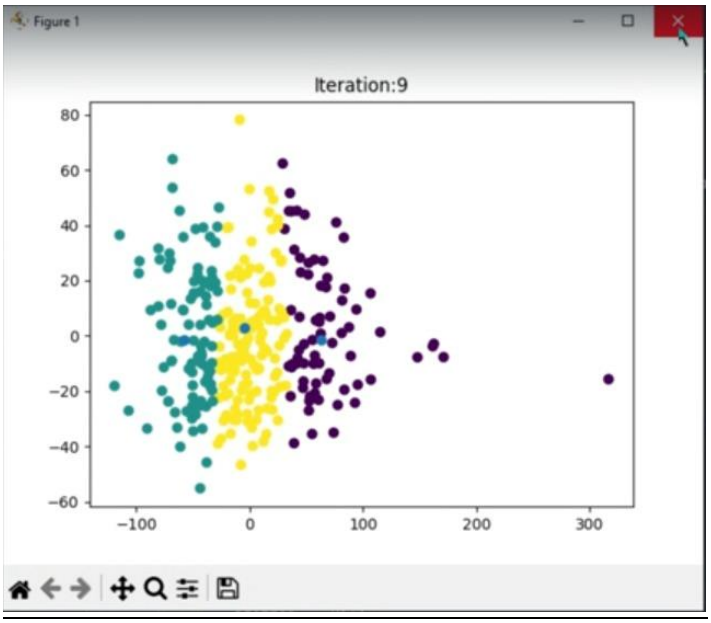
The seventh iteration:



The eighth iteration:



The ninth iteration:



After reaching the optimal centroid, the accuracy plot will be shown.

The accuracy of the K-means clustering model ,Principal component analysis(PCA) ,Random Forest and Decision Tree is evaluated for predicting heart disease using the heart disease dataset. Here's a summary of what happens in terms of accuracy:

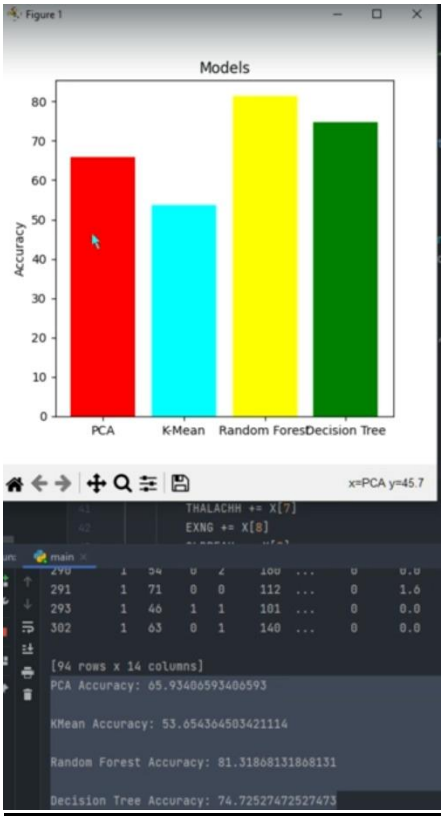
PCA Accuracy:

- PCA is applied to the feature matrix of the heart disease dataset using the PCA class from scikit-learn.
- The dimensionality of the dataset is reduced by transforming the features into a lower-dimensional space.
- The transformed features are used as input to a K-nearest neighbors (KNN) classifier.
- The KNN model is trained on the training dataset and used to predict the target variable for the testing dataset.
- The accuracy of the KNN model is calculated by comparing the predicted labels with the true labels of the testing dataset.
- The accuracy score is the proportion of correct predictions out of the total number of predictions.

K-means Accuracy:

- K-means clustering is applied to the preprocessed heart disease dataset.
- Random initial cluster centroids are selected.
- The algorithm iteratively assigns data points to the nearest centroid and updates the centroids until convergence.
- The predicted cluster labels for both the training and testing datasets are obtained using the K-means model.
- The accuracy of the K-means model is calculated separately for the training and testing datasets using the accuracy_score function.
- The accuracy scores for the training and testing datasets are then averaged to obtain the overall K-means accuracy.

The accuracy plot:



The accuracy results of these models provide an indication of their performance in predicting heart disease. Higher accuracy values suggest better predictive capability. By comparing the accuracy scores of the K-means clustering model, Principal component analysis(PCA) ,Random Forest, and Decision Tree, we can see that the random forest method gives the highest accuracy .

The detailed accuracy results for each model are available in the code implementation, giving a comprehensive evaluation of their performance.