Mustafa Yaşar
21702808
GE461 - Introduction to Data Science
Project 3 Report

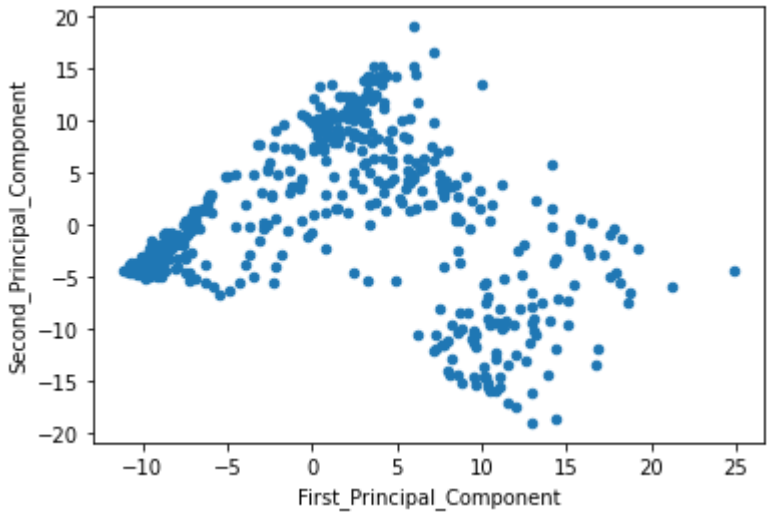
(To run the code, the csv dataset file and the uploaded ipynb must be in the same directory)

Part A

In order to be able to visualize the dataset, PCA with two principal components has been used. According to the analysis, the best two principal components explain 40.726% variance of the features dataset.

Note that to use PCA, the dataset must be standardized so that the mean of the features becomes 0 and the standard deviation becomes 1.

After projecting the dataset onto those two principal components, the following plot has been obtained.

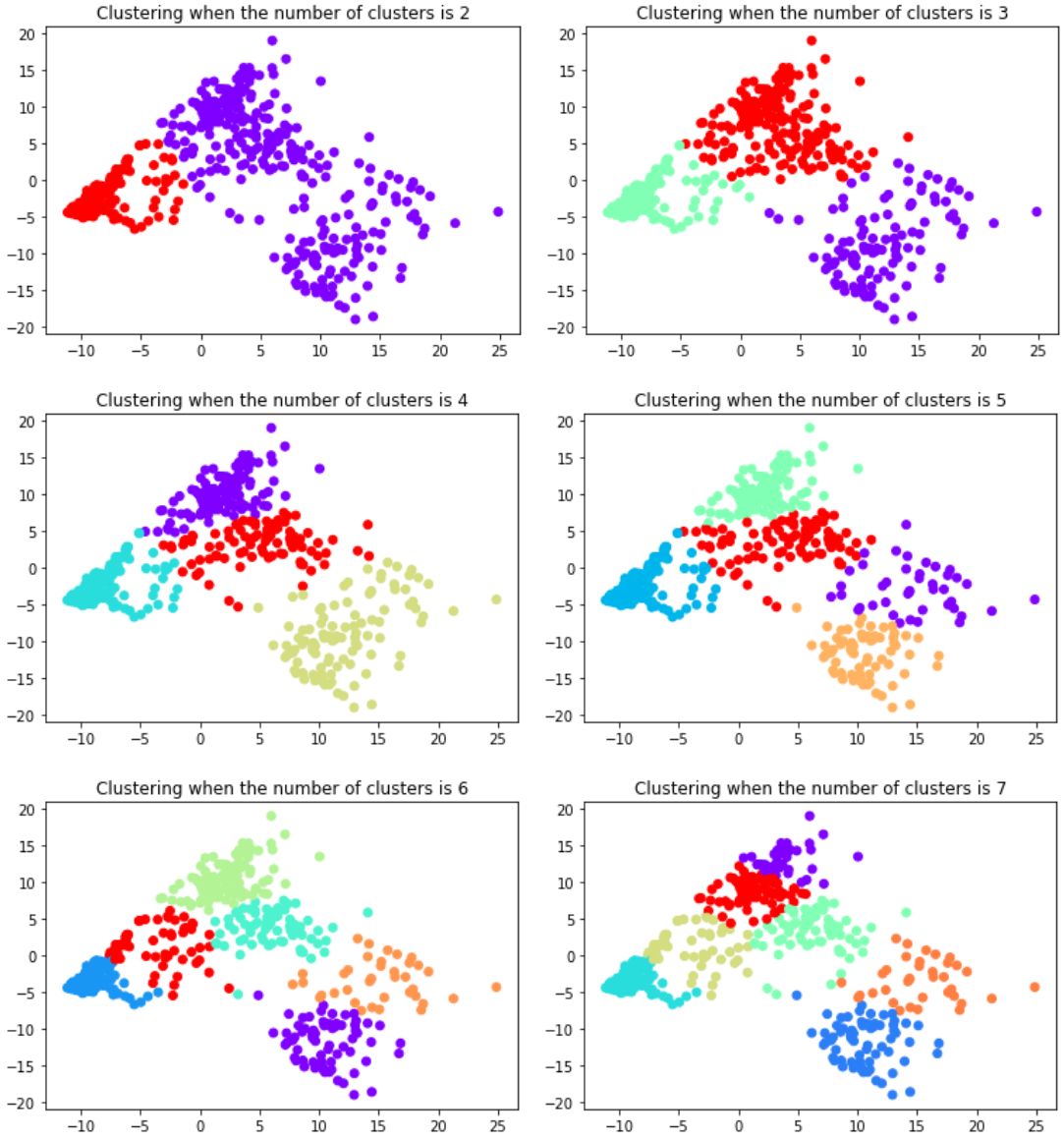


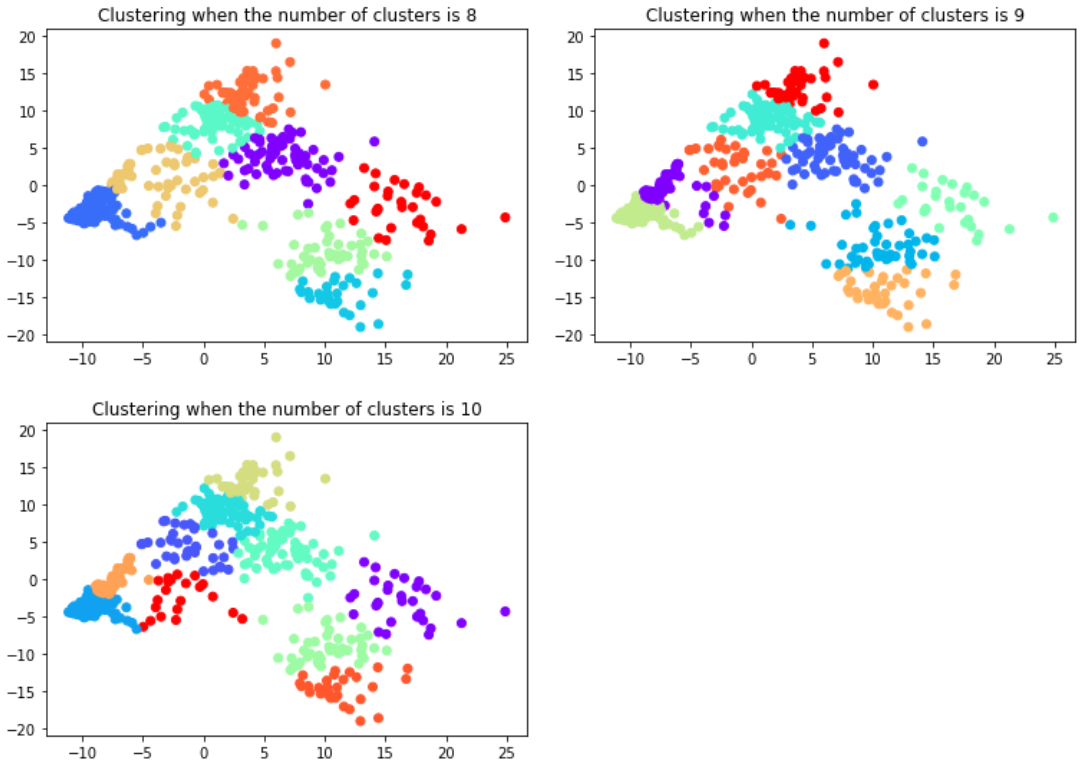
After obtaining this projection, the K-Means algorithm has been run with a different number of clusters (N = 2, 3, 4, …, 10).

Since our response is a binary response, i.e., fall or not fall, it's most logical to use the K-Means with 2 clusters.

The degree of percentage consistency between the cluster memberships and original labels when there are two clusters is 81.27%. According to this accuracy, it's concluded that the K-Means performed well and can be used to detect falls.

Note: In some runs, because of the randomness, 0s and 1s are flipped when the number of classes is equal to 2 in the K-Means algorithm, therefore, the accuracy is calculated as 100 - 81.27 = 18.73%.





Part B

In this section, SVM and MLP classifiers have been trained and tested. In order to train efficient classifiers, hyper-parameters should be estimated by using the validation dataset. Therefore, the dataset has been splitted into 70% training, 15% validation, and 15% testing sets.

Hyper-parameters to be estimated for the SVM classifier are regularization parameter (C), gamma (kernel coefficient for rbf, poly, and sigmoid), and the kernel type to be used in the algorithm.

To calculate the best hyper-parameters, GridSearchCV of sklearn has been used and the following hyper-parameters have been obtained with the validation set.

```
Best hyper-parameters for the SVM classifier
{'C': 10, 'gamma': 0.001, 'kernel': 'rbf'}
```

After training the classifier with the training set and testing it with the testing set, 100% accuracy has been obtained.

As for the MLP classifier, the hyper-parameters to be estimated are activation function for the hidden layer, hidden layer sizes, maximum number of iterations, and the solver for weight optimization.

After using GridSearchCV again, the following hyper-parameters have been obtained.

```
Best hyper-parameters that are determined by using validation set
{'activation': 'identity', 'hidden_layer_sizes': (1,), 'max_iter': 500, 'solver': 'lbfgs'}
```

When those hyper-parameters were estimated, the MLP classifier was trained on the training set and tested on the testing set. After testing, 98.82% accuracy has been obtained.

To conclude, both two models have achieved perfect accuracy, therefore, they can be used to detect falls.

Note that the determining the best hyper-parameters for MLP classifier takes 1-2 minutes depending on the system. Secondly, the dataset has been randomized before splitting into training, validation, and testing sets, therefore, best hyper-parameters may change in different runs.