Project 2: K-NN, PCA, and Clustering

Group Members : Sruti Munukutla and Mahvash Maghrabi

Shortlisted Datasets -
https://www.kaggle.com/datasets/catherinerasgaitis/mxmh-survey-results (Finalized dataset)

https://www.kaggle.com/datasets/fedesoriano/company-bankruptcy-prediction

https://www.kaggle.com/datasets/pritsheta/diabetes-dataset

https://www.kaggle.com/datasets/iamhungundji/covid19-symptoms-checker

Github - https://github.com/mahvashmaghrabi/CS6140_Project2

Working Pattern -

Task 1, 4 - Sruti Munukutla
Task 2, 3 - Mahvash Maghrabi

About the Dataset

From the datasets shortlisted above, we decided to work on the dataset '**Music & Mental Health Survey Results'.** The owner of the dataset conducted a survey via Google Form to assess the Mental health of people who listen to various genres of music. We found this survey quite interesting and since most of the data apart from Age, Hours per day (listening to music) is categorical, we felt this would be perfect for Clustering and Nearest Neighbor analysis. The data was cleaned and the string values of "ALWAYS", "SOMETIMES", "NEVER" etc reflecting the frequency of listening to a genre of music was changed to an integer ranging from 0-5. This is for ease of calculation of distance metric. The dependent variable is Anxiety.

Nearest Neighbor analysis:

Nearest Neighbor algorithm is used to classify a datapoint based on how its neighbors are classified. The distance metric is used to assess the similarity with the neighboring data class types. For this Project, Euclidean distance has been implemented which is an L2 Norm.

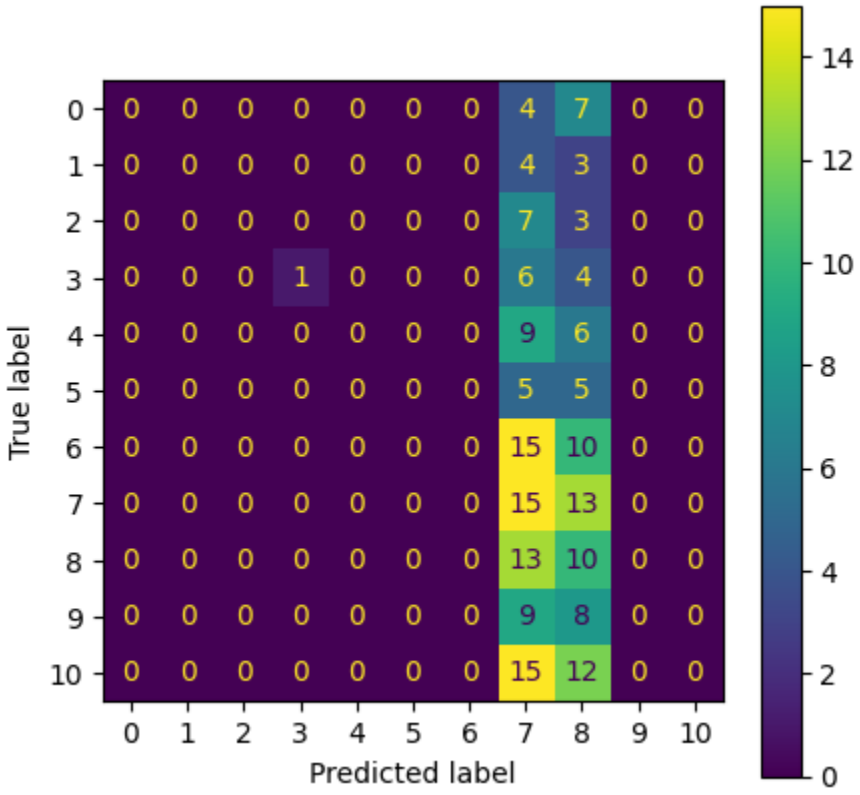
Other widely used distance metrics include:

- Hamming Distance
- Manhattan Distance
- Minkowski Distance

The get_euclidean_distance assumes that X_test and X_train will be a set of data points. For a single test data point, distance to each training data point is calculated. From the list of distance per test data point, the shortest distance is recorded. The respective y value of the shortest distance is taken to be the predicted dependent value (In this case, level of Anxiety experienced).

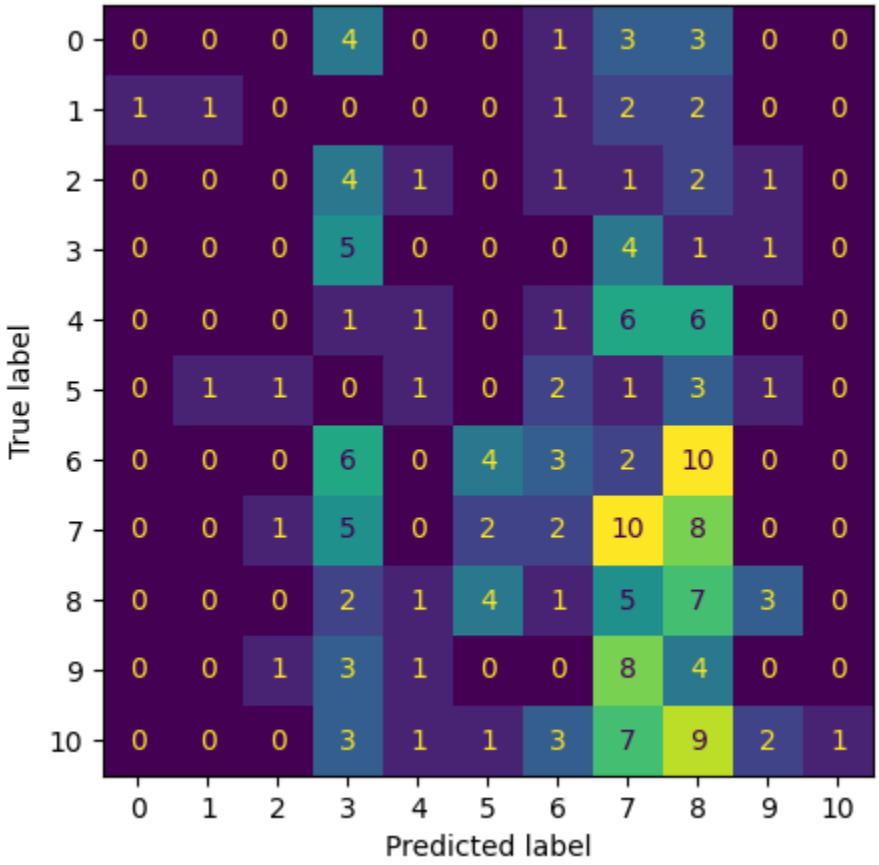
Since our data was not whitened/scaled/weighted, we did not account for it in the calculations.

The Accuracy of 1NN model was: 14.67%
The confusion Matrix is as shown below:



It can be seen that most of the predicted y value - Anxiety level is 7, this could be because the dataset is skewed - more number of users have recorded to have anxiety levels around 7. Hence, when more number of neighboring classes are not included, the results are skewed too.

After implementing knn, the highest accuracy was achieved for k = 15, which is equal to 15.76%. The confusion matrix reflects better results (less sparse) where the true value and predicted values for each category of anxiety experienced are equal.



Clustering - It is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. Clustering is a collection of objects on the basis of similarity and dissimilarity between them.
Different types of clustering methods :

1. Partitioning Clustering
2. Density-Based Clustering
3. Distribution Model-Based Clustering
4. Hierarchical Clustering
5. Fuzzy Clustering

Task 2 : Experiment with clustering and PCA on two structured data sets

A. Plotting and Visualizing the Clusters

The given datasets Set A and Set B were downloaded and plotted on a scatter plot. The datasets had three columns X1, X2 and ID. Out of which only the X1 and X2 columns were used for plotting and understanding the clusters of the dataset. There are two scatter plots one for set A and one for set B

Set A
The set A scatter plot has six clusters and the shape of the clusters are elliptical.
Set B
The set B scatter plot has six clusters two whereas the clusters are in linear shape.

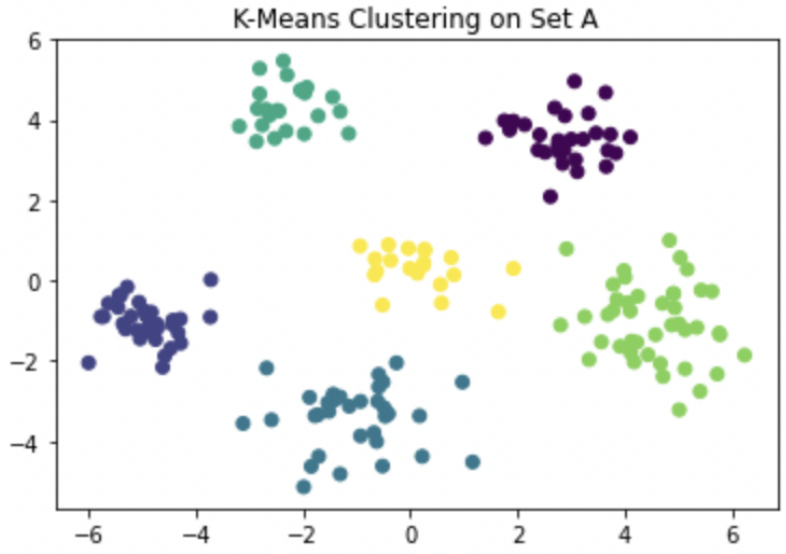
B. Clustering both the datasets using K-means Algorithm.

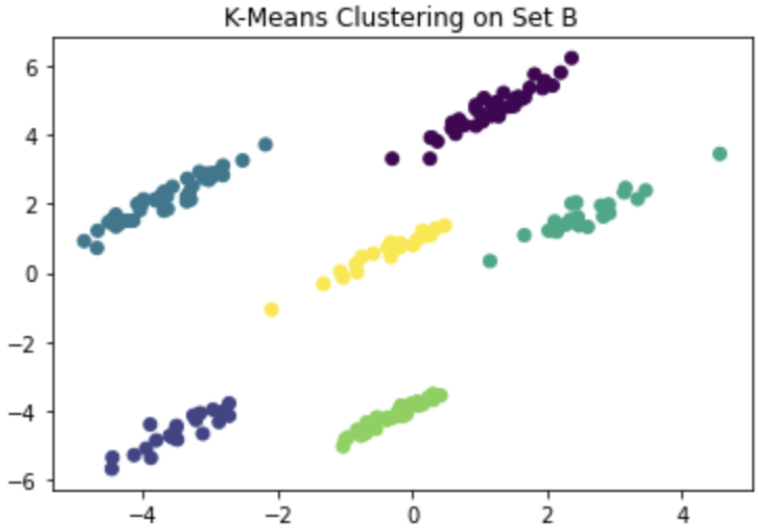
K-means algorithm

To process the learning data, the K-means algorithm in data mining starts with a first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative (repetitive) calculations to optimize the positions of the centroids It halts creating and optimizing clusters when these conditions are met :

1) The centroids have stabilized — there is no change in their values because the clustering has been successful.
2) The defined number of iterations has been achieved

The k means algorithms is applied on both the datasets keeping k = 6 which indicates the dataset should be divided into six clusters. According to the visual observation there are six elliptical clusters each of different colour for set A nad six linear shaped clusters each of different colour for set B. The k-means algorithm groups similar data points together and separate dissimilar data points. The second clustering method that we used was DBSCAN.





DBSCAN

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is an algorithm for clustering a dataset into groups too but it is based on density. It considers the number of data points within a certain distance called epsilon of a given data point rather than using the number of clusters like the ‘k’ in k-means algorithm. The algorithm identifies core points which have a large number of nearby data points, and also border points which have a few nearby data points as well as noise points which do not have any nearby data points.

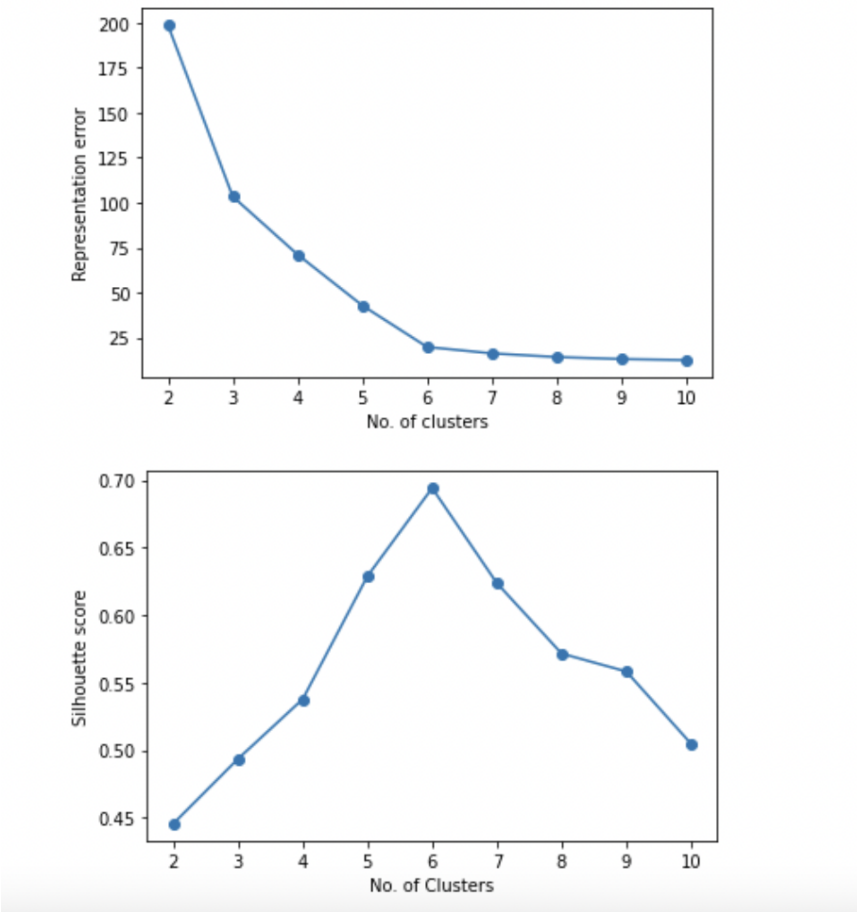
After observing the plots received on plotting both the cluster algorithms, it shows that the two cluster results are entirely different. The difference makes sense as k-means algorithms divide the dataset in the specified number od=f clusters whereas the DBSCAN groups the data points into clusters based on their density relative to other data points. It makes no assumptions about the shape or size of the clusters.

C. Comparison using different numbers of clusters

For this task only the setA dataset was used. K-means clustering was applied on the dataset and the values of k ranged k=2 to k=10. The data was standardized first. Standardizing the data is transforming the variables so that they have a mean of 0 and a standard deviation of 1. This ensures that all variables are on the same scale and this also improves the algorithm performance. A for loop was used to apply the k-means algorithm for values of k from 2 to 10.

Implemented Krzanowski and Lai for cluster quality metric . The KL index is a cluster quality metric that compares the within cluster sum of squares (WSS) for values of K. The index is calculated as the difference between the WSS for k-1 clusters and the WSS for k clusters, divided by the WSS for k-1 clusters. The value of K that maximizes the KL index is considered to be the optimal number of clusters for the data. The optimal value obtained for the given dataset set A was 9 clusters.

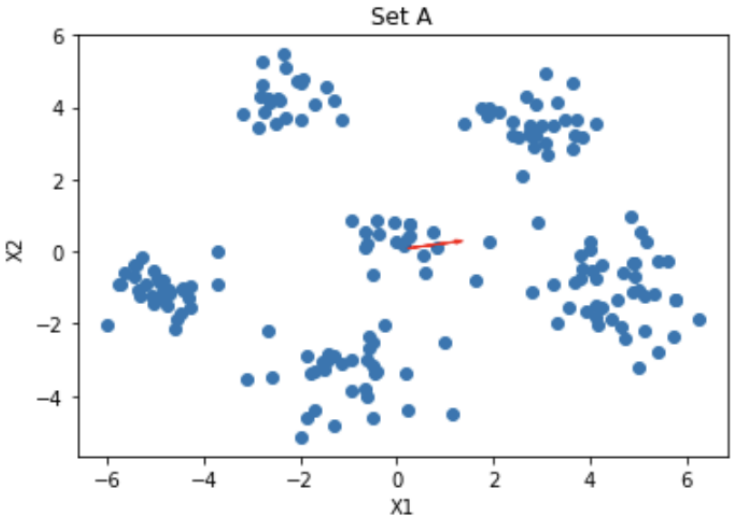
Silhouette Score cluster quality metric was also used to evaluate the quality or performance of a clustering algorithm. It quantifies the similarity of data points within each cluster as well as the dissimilarity between different clusters.It ranges from -1 to 1 with values closer to 1 indicating a better clustering solution. The value obtained for the dataset was 0.70 and the corresponding clusters for that value was 6 clusters.



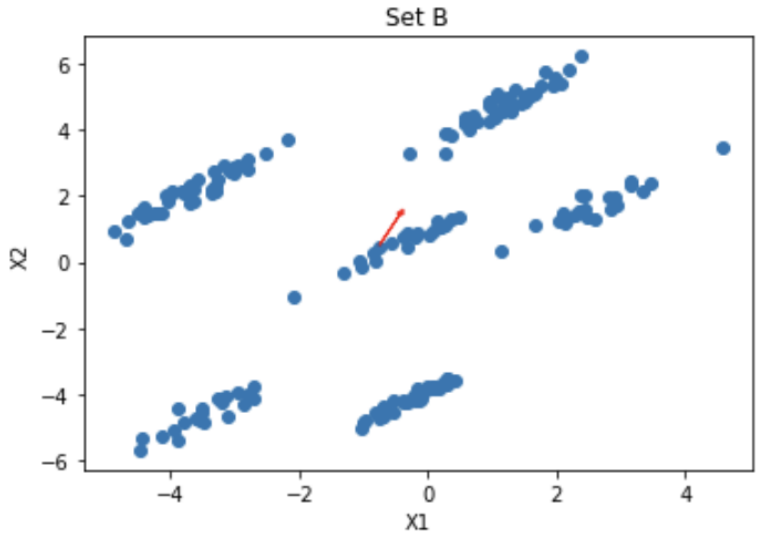
D. Applying Principal Component Analysis to both the datasets

Principal Component Analysis It is used to convert a large set of variables into a smaller set of principal components that capture as much of the variation in the original data as possible. The process of PCA involves finding the eigenvectors of the covariance matrix of the original data and they are used to transform the data into a new coordinate system with the eigenvectors as the new axes. Whitening refers to a technique that is used to decorrelate the data by transforming it whereas non-whitened data is the data that has not been transformed to decorrelate the features. The Eigenvalues for non-whitened data are larger than the eigenvalues of whitened data. This is because eigenvalues are a measure of the amount of variation. In a whitened dataset, the eigenvalues will be equal to the variances of the individual features, since the data has been decorrelated transformed to have zero mean and unit variance it has a lower value. The eigenvalues of non-whitened data is larger because it is not transformed to decorrelate the features and the value is a combination of both variance and covariance of the features.

The PCA function code from project 1 was used to implement this on the dataset A and B. The direction of the first eigenvector is shown as an arrow on the plot. The direction makes sense as the eigenvectors points to the direction of the maximum variance in the data. It represents the most important patterns in the data and can help determine the relationships between the variables in the data. The direction in the below plot, points in the direction of the line that best captures the maximum variance in the data. A high variance in the data indicates that the data is spread out over a large range of values, while a low variance indicates that the data is tightly clustered around the mean value. When the clusters in the plot are compact and have a small spread then the variance of the data is low. If the clusters are spread out over a large area means that the variance of the data is high.
Here in Set A the eigenvector points to the cluster that has maximum spread indicating high variance.

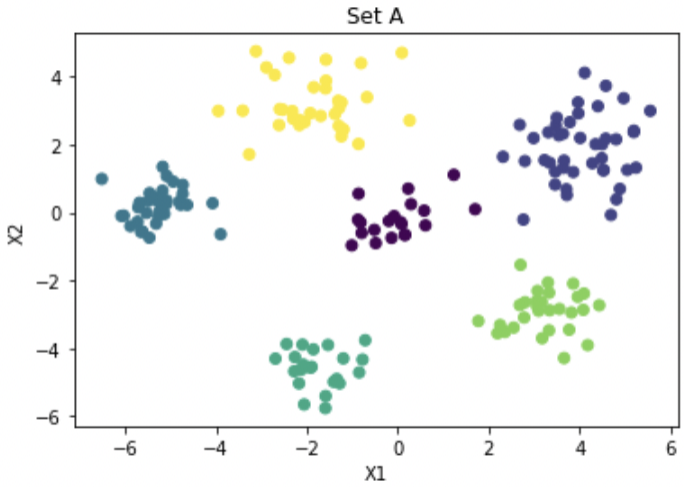


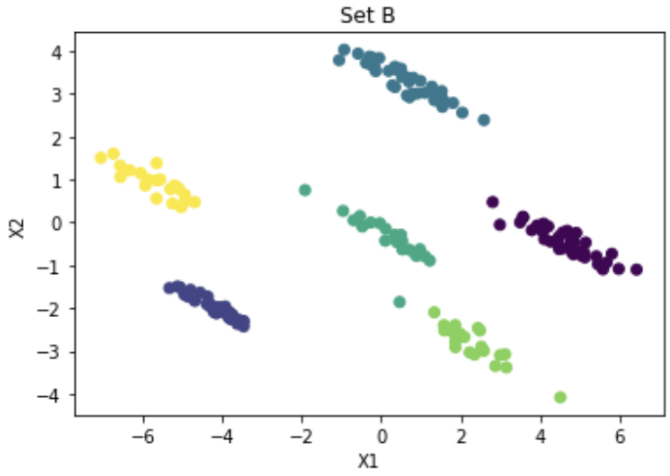
Whereas, in Set B the eigenvector points to the cluster that is compact and tightly clustered indicating low variance.



E. Reclustering using the projected data

This step required clustering on the projected data that was obtained by performing PCA on the datasets. The reclustering was done using k-means algorithm and DBSCAN. The results obtained by using clustering on the projected data were different from the results obtained by using the original data for both the algorithms. By using the projected data the clustering process is improved as the dimensionality of the data is reduced which also helps to overcome the curse of dimensionality. The curse of dimensionality is faced when the number of dimensions in the data increases, the number of data points required to describe the data also increases exponentially. Weighing the eigenvectors differently also affects the clustering process. As we assign different weights to the eigenvectors, we are in control of the importance of each eigenvector. For instance, we can assign higher weights to the eigenvectors corresponding to the largest eigenvalues as that captures the most important patterns and structures in the data, and lower weights to the eigenvectors corresponding to the lower eigenvalues, which will capture less important information.





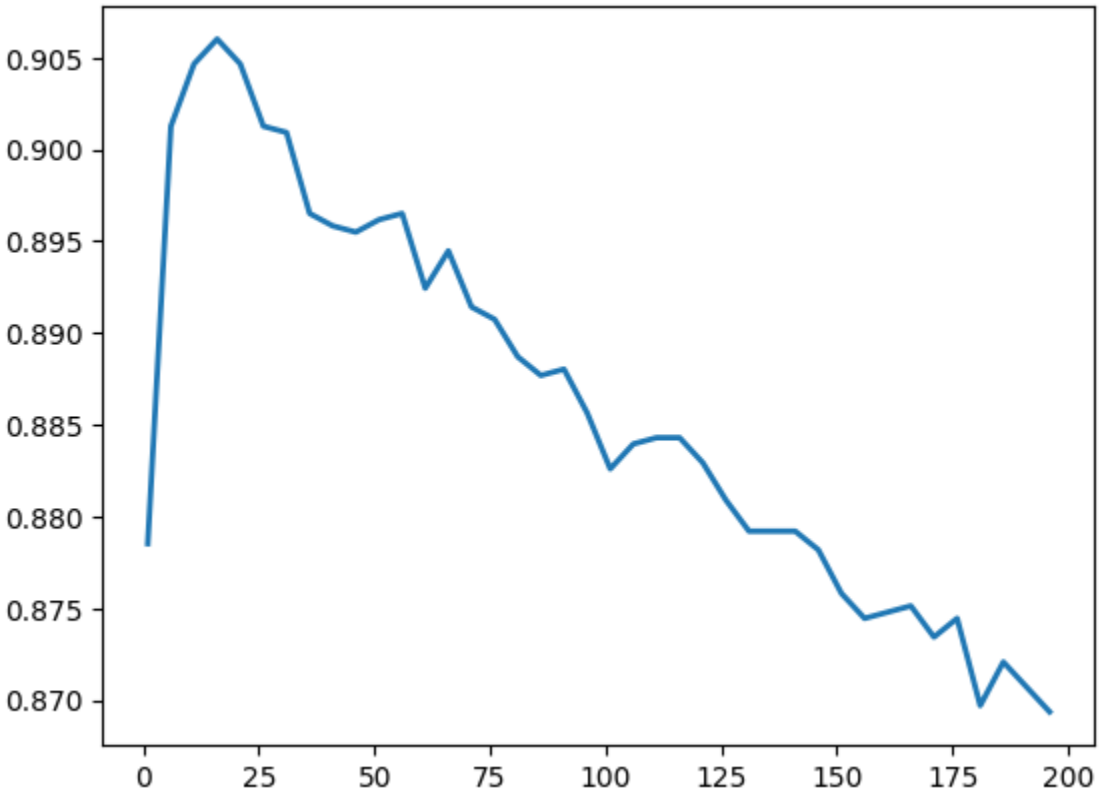
Task 3 : Apply K-Means Clustering to your data set

The dataset used for the project depicts the music and mental health relationship. The dataset has 20 features out of which the first feature index is dropped prior to performing k-means clustering. The dataset is standardized to ensure that all variables are on the same scale and to improve the algorithm performance. The algorithm was performed keeping value of k=4. As the dataset contains 19 features it is difficult to plot it in a two-dimensional array. So PCA is used as it identifies the underlying structure in the data and reduces the number of features to a smaller number of components that capture most of the information in the original data. The results are then plotted on the graph and it is noticed that it has natural clusters. The cluster quality metric is performed for k value in the range 2 to 11 and it depicts that for value of k =2 it has the highest Silhouette score making k=2 an ideal value to perform k-means clustering.

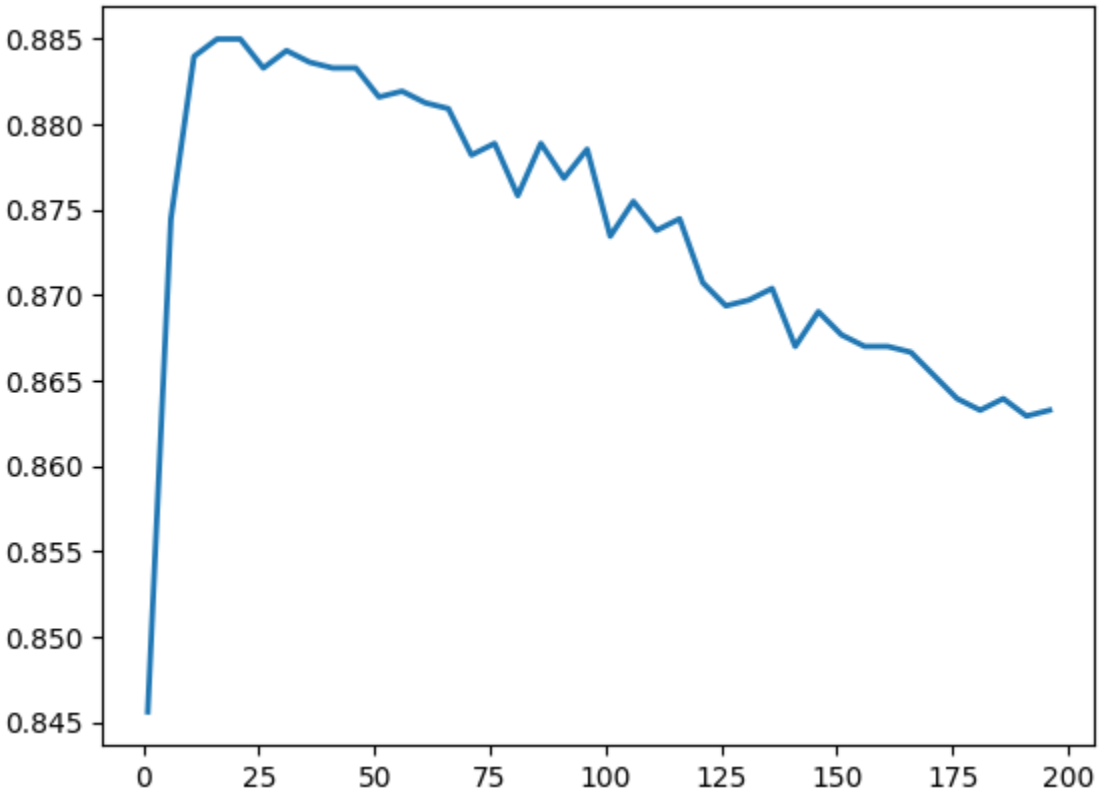
Task 4 : Use K-Nearest Neighbor and PCA to classify activity from phone measurements

The dataset provided for this part records walking activity from the phones of 30 users. The data was recorded when the users were either WALKING, WALKING_UPSTAIRS, WALKING_DOWNSTAIRS, SITTING, STANDING, LAYING. So, each data point belongs to either of the categories. Using KNN, the model can be trained on the training data X_train, Y_train and predict the category of the user for test data X_test.

Accuracy of the model was calculated for various values of n_neighbours in the range 1 - 200. It has been observed that the accuracy peaks around between 15-25 neighbors after which it starts dropping with increasing nearest neighbor. This is shown in the plot below:

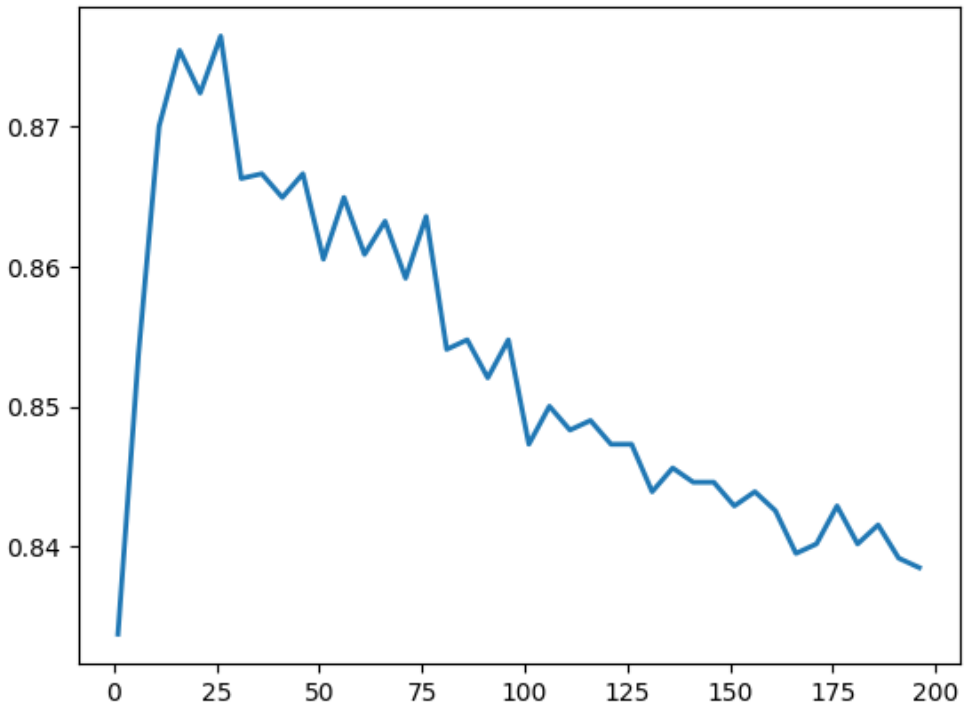


Next, Principal Component Analysis (PCA) was performed on the data to filter out the crucial dimensions and analyze if this improves the model's accuracy. PCA was performed on training data and eigenvectors, eigenvalues and standard deviation and mean was used to project the test data into the same eigenspace. Also, the data was not whitened for PCA. PCA returned the crucial dimension count as 34 for 90% variance of the data and the results from KNN on reduced dimensions for test is as follows:



As it can be seen, the peak accuracy is lesser for the data after PCA for non-whitened data and occurs around the same range.

Performing PCA after whitening data yielded better accuracy similar to KNN without PCA with higher levels of accuracy for n_neighbors data between 15-25.



EXTENSIONS

1. **Additional Clustering Methods on dataset A and B : AgglomerativeClustering**

   AgglomerativeClustering is a hierarchical clustering algorithm it builds a hierarchy of clusters by merging smaller clusters into larger ones till the defined number of clusters is achieved. It follows a bottom-up approach where individual data points are treated as single clusters, and then merged together to form bigger clusters. The merging is based on a similarity metric between the data points such as the Euclidean distance. There is a difference between the output of clusters between k-means and agglomerative clustering as both follow different approaches.

2. **Additional Cluster Quality Metric : Davies-Bouldin Index**

   The Davies-Bouldin Index measures the similarity between each cluster and its most similar cluster. It basically evaluates how well separated the clusters are from each other. It is opposite to the Silhouette score as the DBI ranges between 0 and infinity with a lower value indicating better clustering results and the graph obtained by performing is DBI looks like an upside down silhouette score graph.

3. **Confusion Matrix**

Confusion matrix for Part 1 (mentioned as extension) is plotted to observe the model's performance.

REFERENCES

1. https://www.kaggle.com/code/josluizfjunior/eda-music-and-mental-health-relationship
2. https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1
3. https://en.wikipedia.org/wiki/Cluster_analysis
4. https://www.javatpoint.com/clustering-in-machine-learning
5. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html
6. https://www.hrpub.org/download/20220630/MS5-13425454.pdf