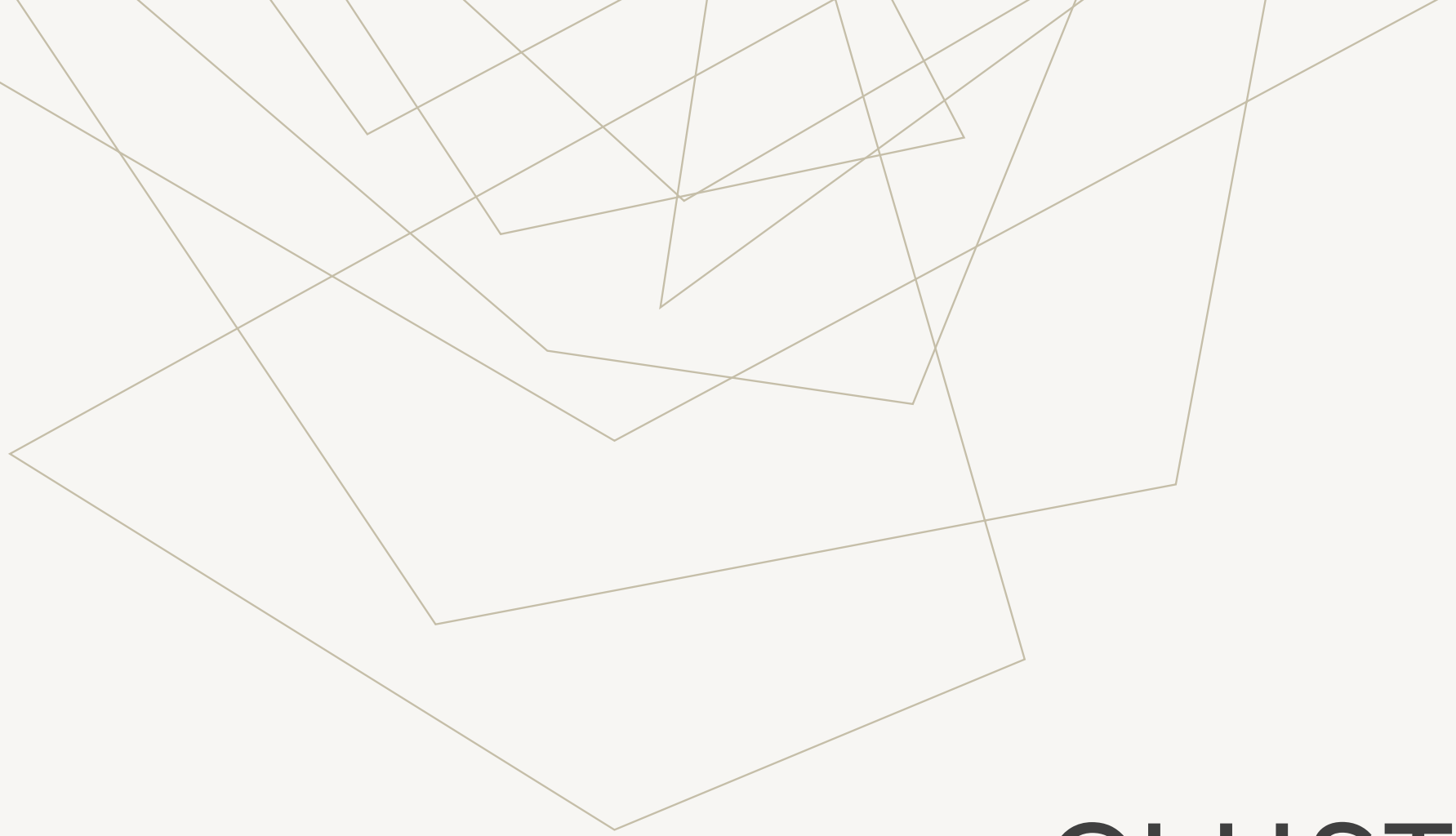


CLUSTERING

Introduction to Data Analysis

Anjali Savlani

100844738

AGENDA

- Introduction
- Analysis Description
- Identification and Justification of Number of Clusters
- Insights from the clusters created

INTRODUCTION

This presentation focuses on K Means clustering of the chosen dataset along with two clustering methods to determine optimum number of clusters in which the date should be divided to make it insightful.

ANALYSIS DESCRIPTION

K means Clustering:

K Means Clustering is one of the most basic and often used unsupervised machine learning algorithms. It divides the dataset into meaningful classes called ‘clusters’ to further aid the statistical studies and machine learning.

In terminology: k is number of clusters
n is the observations in dataset

In this assignment we have taken the illnessstudy dataset which has 30 independent variables (numerical) and one dependent named diagnosis. The dependent variable ‘diagnosis’ is categorical variable with two categories.

In our case, we have n = 569 observations and we shall divide this into optimal k number of clusters, where, we will utilize two methods to determine this optimal number of clusters.

Methods to be used are; i) Elbow Method
ii) Silhouette Method

IDENTIFICATION AND JUSTIFICATION OF NUMBER OF CLUSTERS

In the Illnessstudy dataset, we are performing clustering with the help of two methods; Elbow Method and Silhouette method.

Elbow Method

In Elbow method, we first calculate the Inertia,

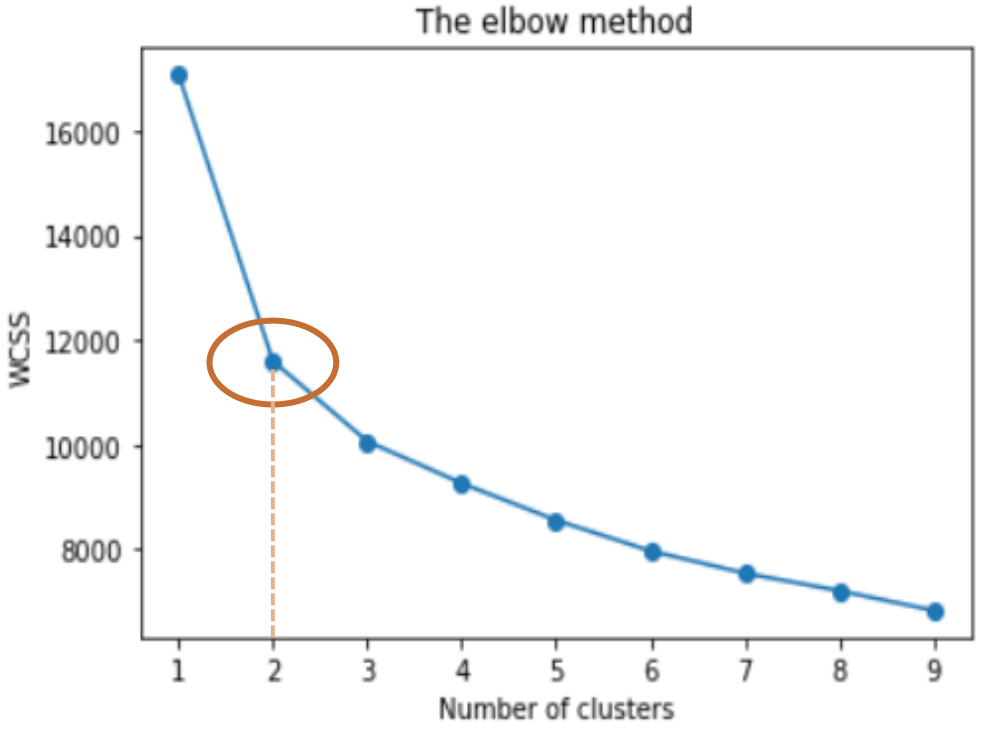
Which is the sum of squared distances of the samples to their closest cluster center. Secondly, it is plotted against the number of clusters.

Now, as per the name ‘Elbow’, points on this visualization form an elbow. According to definition of this method, at the cluster point where the elbow forms is considered to be optimal value for the number of clusters.

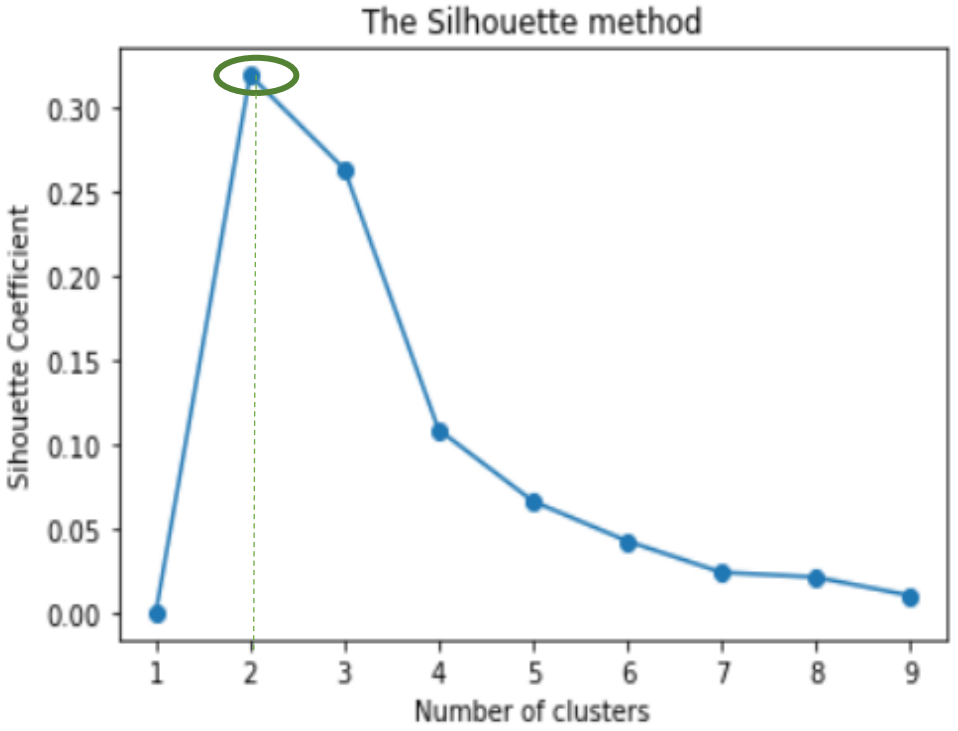
Silhouette Method

The term "silhouette" refers to a way of interpreting and validating consistency inside data clusters.

The method generates a simple graphical depiction of how effectively each object was classified. In Silhouette method, silhouette coefficient is calculated using the Euclidean method and plotted against the number of clusters, the cluster point at which the coefficient is maximum is considered to be the optimal value for the number of clusters.



Insight: In the visualization of ‘within clusters sum of squared distances (WCSS) and number of clusters, at cluster value ‘2’, sum of squared distances from the respective centroid starts to decrease forming an elbow.
We can infer from this method that 2 segments of the illnessstudy dataset is the optimal numbers of clusters.

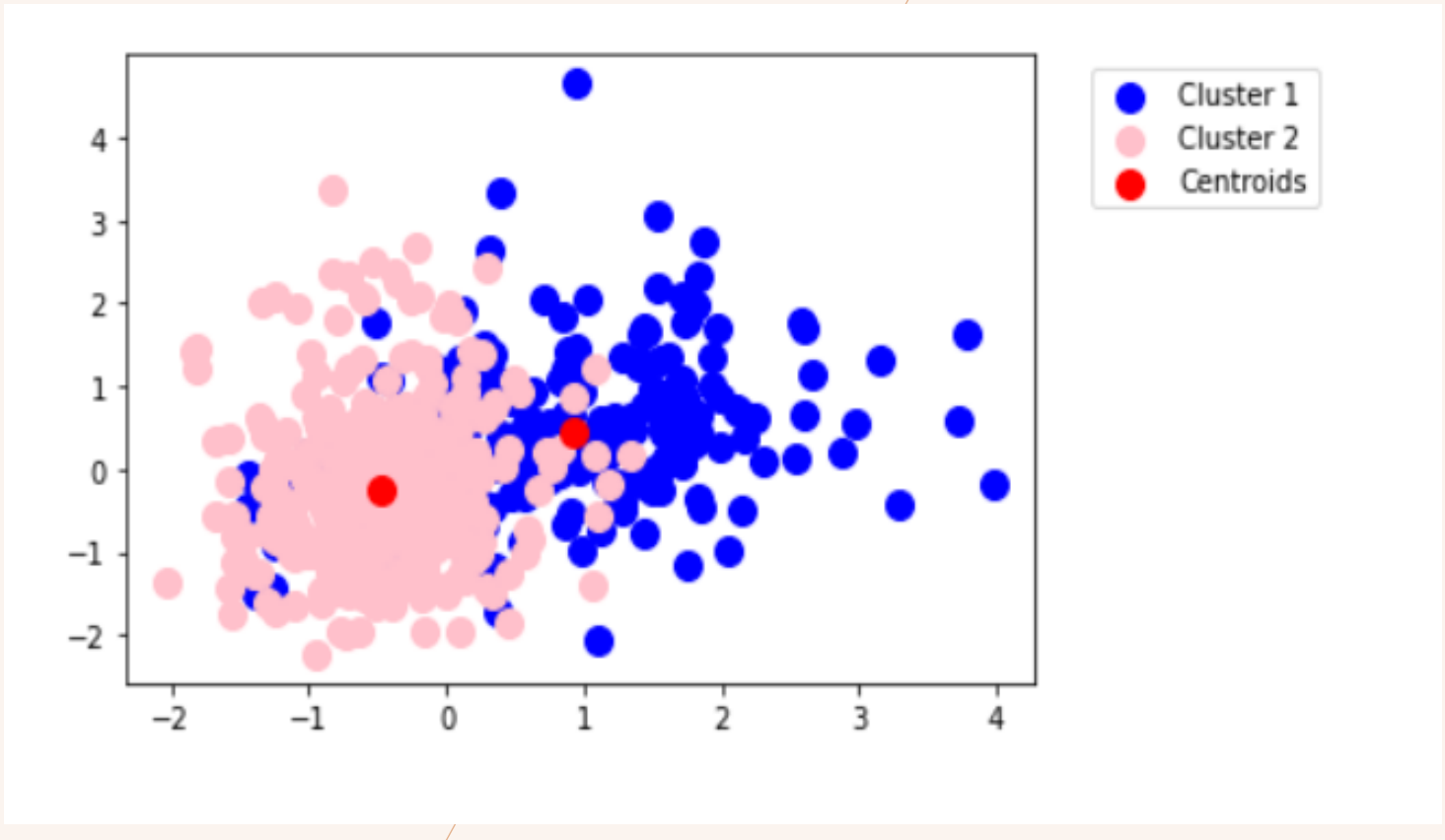


Insight: From the visualization of Silhouette coefficient (or Silhouette score) and the number of cluster, at cluster value ‘2’, Silhouette coefficient is maximum.
It can be deduced that 2 clusters of the illnessstudy dataset is the optimal number of clusters the dataset.

INSIGHTS FROM NEWLY CREATED CLUSTERS

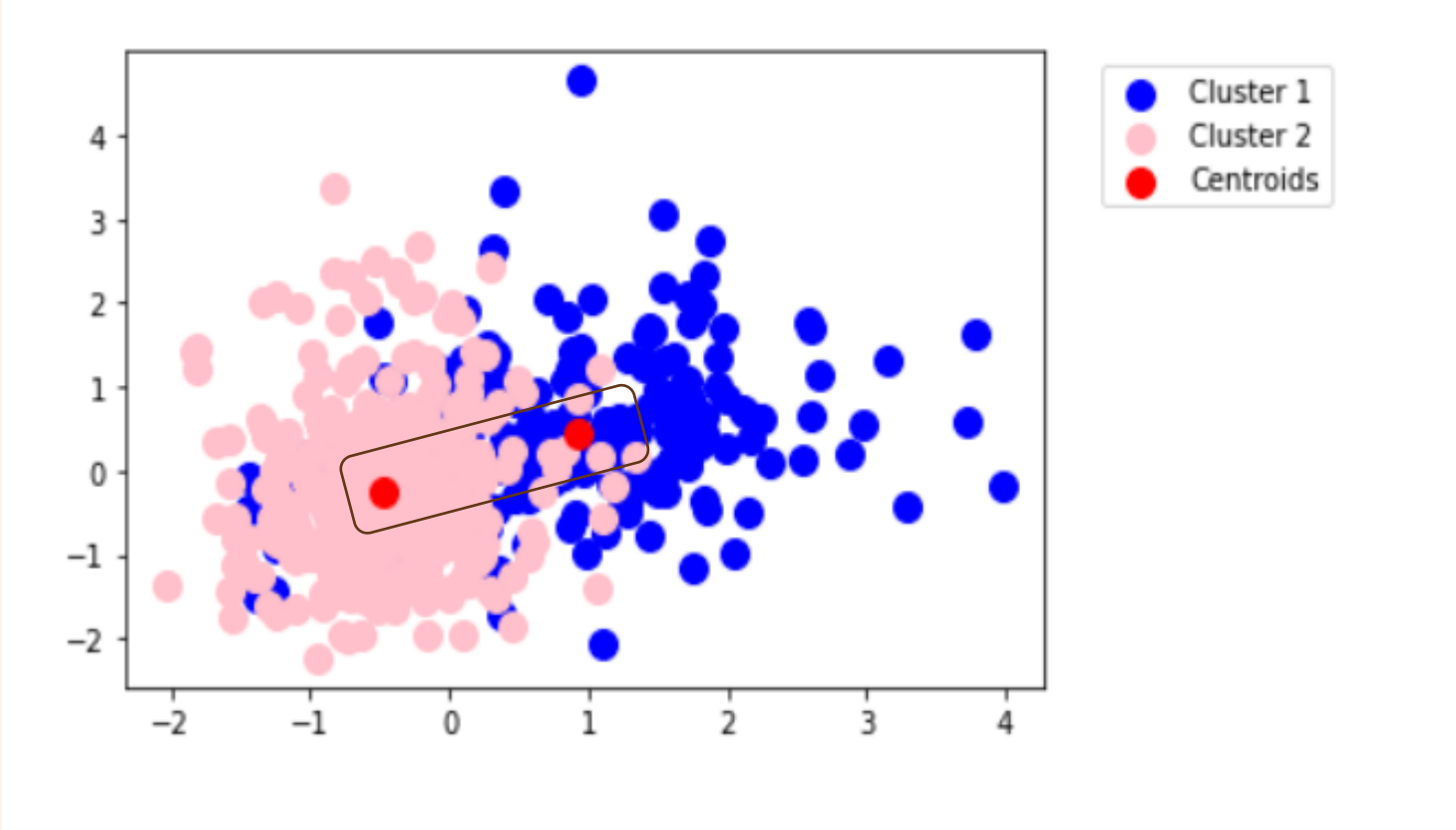
In this visualization we have plotted 2 clusters by showing them in two different clusters with their centroids coloured as red as follows;

1. In these newly formed clusters, we can see that cluster consistency is not well-defined and they are not finely apart from each other.
    - Some of the points in the cluster to the right (Blue) overlap the points in the cluster to the left (Pink).
    - This implies that the variable is not good enough to be taken as the base for clustering the data points. In our case, categorical variable diagnosis is not insightful to be taken for reference to cluster the illnessstudy dataset.



INSIGHTS FROM NEWLY CREATED CLUSTERS

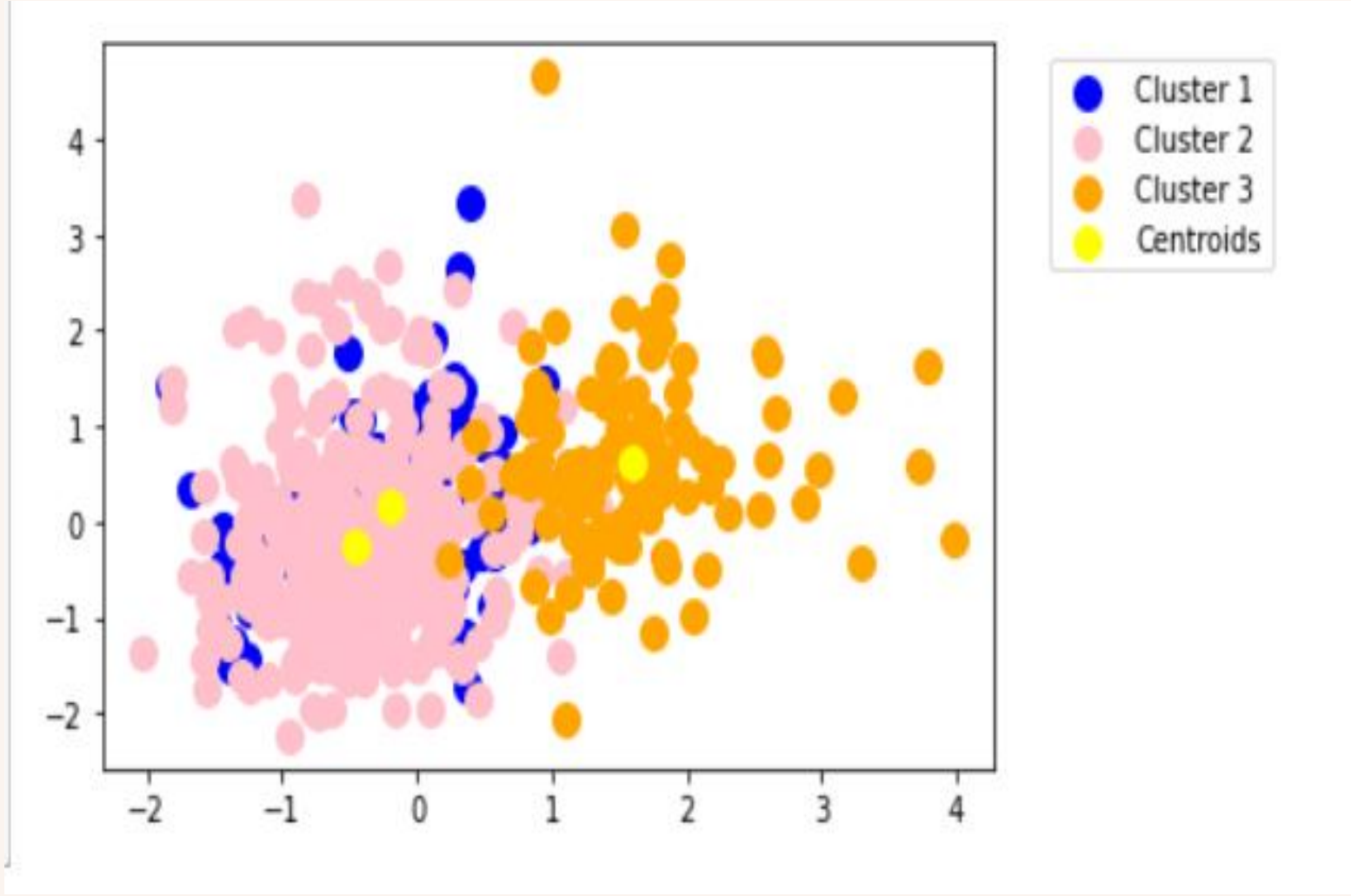
2. We can deduce from the cluster visualization that, centroids of the two clusters are finely apart. Moreoever, since the silhouette score measures the seperability between the clusters be taking the mean difference within the clusters, it implies that there is sufficient consistency withtin the clusters.



FURTHER JUSTIFICATION

Inference: If we increase the number of clusters and plot them, we can see that the clusters become more chaotic and the points increasingly overlap onto each other, indicating the non-optimized number of clusters.

Secondly, we can also see that the centroids(yellow dots) of cluster number 1 and 2 are not sufficiently apart from each other. This is an additional evidence to confirm that more than 2 clusters are not the optimal solution for the cluster analysis.



SUMMARY

From the Clustering Analysis, we created 2 sufficiently consistent clusters of the illnessstudy dataset which took into account the dependent variable named ‘diagnosis’, which is a categorical variable.

THANK YOU