Work Set 3 Report

Wojciech Lotko
Department of Physics and Astronomy and Computer Science,
Jagiellonian University, Cracow, Poland

(Dated: April 25, 2025)

Abstract

Below report describes analysis of patients fingerprints and patient zero characteristic. Having clustered the data could help fit patient zero characteristics to broader population. This may increase the chance of finding more individuals with the same resistance to new SARS-CoV-2 variant.

I. INTRODUCTION

This report details the development of an unsupervised machine learning model used for clustering data points into distinct groups using *KMeans* algorithm. Project aimed to identify possible groups of patients based on provided fingerprints.

II. DATA EXPLORATION

Data was hard to explore. Fingerprints were provided as numpy arrays. Each fingerprint was represented by 512 values and whole array contained 16929 samples.

III. DATA PREPROCESSING

In this project, I implemented two preprocessing techniques to prepare the genetic fingerprint data for clustering analysis:

- **Feature Scaling**
  I applied StandardScaler[1] to normalize all 512 features in the genetic fingerprint's dataset. It transformed features to have zero mean and unit variance, ensured all genetic markers contributed equally to the distance calculations and improved the stability and performance of the KMeans[2] algorithm, which is sensitive to feature scaling
- **Dimensionality Reduction**
  To address the high dimensionality of data, I used Principal Component Analysis (PCA)[3]. It reduced dimensions from 512 to 50 principal components and preserved approximately 74.44% of the original variance. Mitigated the curse of dimensionality[4] for more effective clustering. Finally reduced dimensionality to 2 which enabled better visualization and interpretation of the results.

IV. MODEL SELECTION AND FEATURE IMPORTANCE

In this case we are dealing with unsupervised machine learning problem. Best approach for solving finding groups of people sharing similar characteristics is using clustering techniques. KMeans algorithm is one of the best due to its interpretability. Our responsibility is to identify the optimal number of clusters for given problem. This was solved using elbow technique based on graph representing inertia values for given k.

[1] Standard Scaler [https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html]

[2] KMeans [https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html]

[3] PCA [https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html]

[4] The Curse of dimensionality in Machine Learning [https://www.datacamp.com/blog/curse-of-dimensionality-machine-learning]

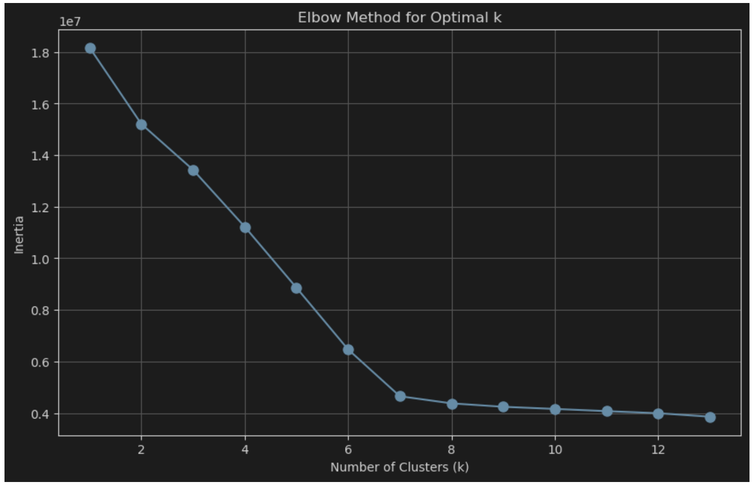

Figure 1 Intertia value based on number of clusters k

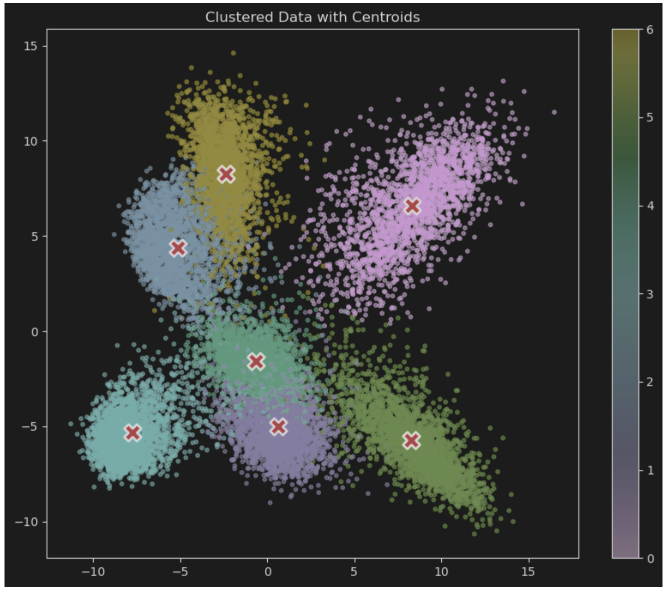

Figure 2 Clusters with centroids

As we can see k=7 is the optimal number. Further increasing is pointless (cluster coherency is slowing down). Lower k would lead to incorrectly assigned patients.

V. MODEL VISUALISATION

Below you can find graph how different clusters are separated and where our patient zero fits. He is placed in cluster 4 according to our model.

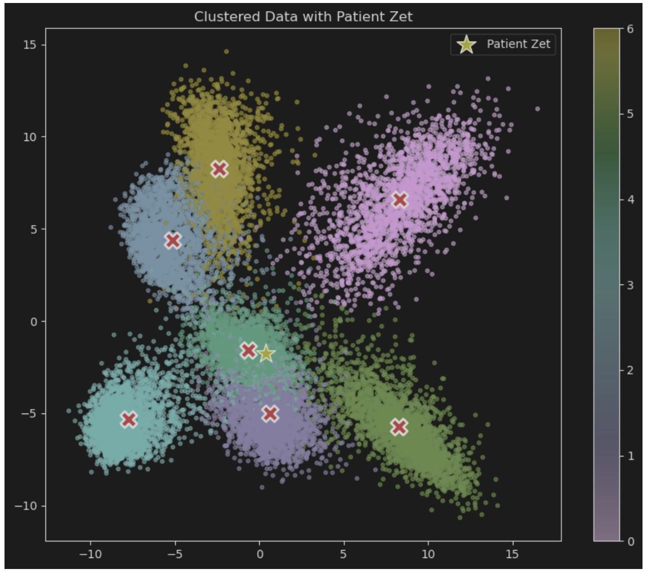

Figure 3 Patient zero assigned to cluster 4

Further model evaluation was impossible. Provided sample was not containing labels (that's one of points why we used unsupervised learning).

VI. CONCLUSIONS

If mentioned prediction is used to determine what people should be examined to create new vaccine, then it would be best to look for people assigned to cluster 4.

They share the most similarities with patient zero which gives us most chances to identify more people already immune to disease. Above analysis identified 2427 people in cluster 4.