Homework 3 Data Mining

Davide Fortunato 1936575

December 2024

1 Introduction

This analysis aims to evaluate the impact of feature engineering and preprocessing on clustering performance using a specific algorithm. I have chosen KMeans++ which, briefly, partitions the data into k clusters by iteratively minimizing the within-cluster sum of squared distances (WCSS) and, to enhance cluster initialization and convergence, the k-means++ method was used, ensuring a better spread of initial centroids and improving the overall cluster quality. To evaluate clustering performance, I preferred the silhouette score over the elbow method because the first provides a straightforward numeric measure of cluster quality, making it faster and more convenient for comparing different cases. In contrast, the elbow method requires a more subjective interpretation of a plot to determine the "elbow point" and there were cases in which I had difficulties to understand which was the elbow point. In addition, I visualized clusters employing Principal Component Analysis after having centered and normalized data.

2 Feature engeneering

I experimented with several feature engineering techniques, but the most effective ones in the final solution were one-hot encoding and scaling. One-hot encoding was used to transform categorical features, such as ocean_proximity, into numerical representations suitable for clustering. Then, since the dataset contained variables measured on different scales, I applied normalization, testing both StandardScaler and MinMaxScaler and the latter yielded better results. Other methods, such as generating polynomial features, adding ratios for correlated variables (seen by plotting the correlation matrix), and using PCA for dimensionality reduction beyond visualization, were also tested but did not lead to improved clustering performance.

3 Results and Observations

Cluster Analysis without Preprocessing:

- Optimal k = 2, based on silhouette scores.
- Silhouette score: 0.6288
- Running time: 0.0356 seconds

With Preprocessing:

- Optimal k = 4, based on silhouette scores.
- Silhouette score: 0.6370
- Running time: 0.0214 seconds

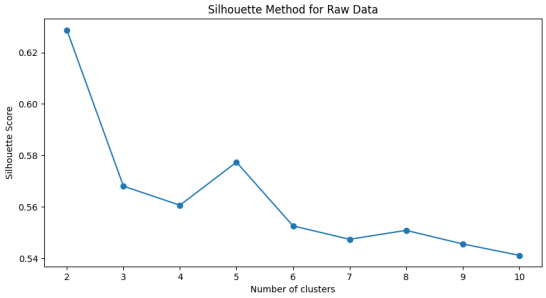


Figure 1: Silhouette method for raw data

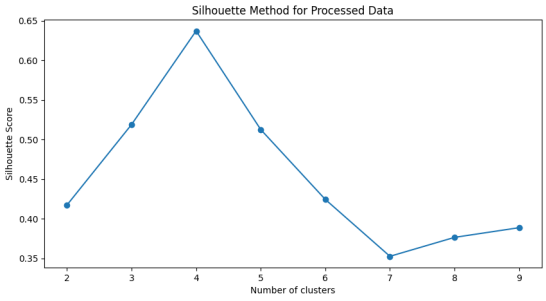
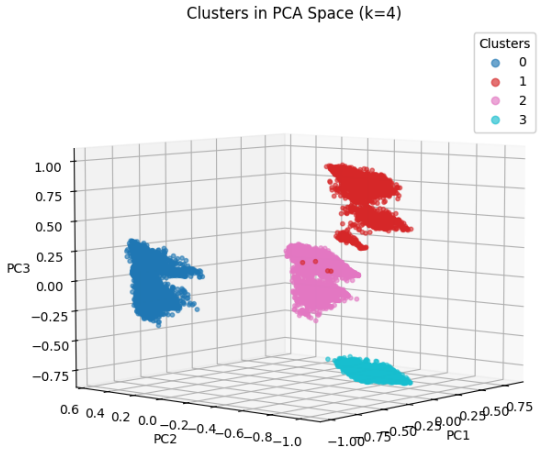
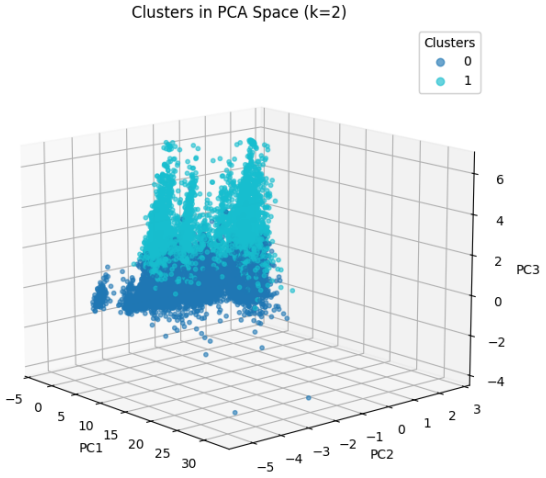


Figure 2: Silhouette method for processed data

We can observe that, although the silhouette score and running time showed only a slight improvement, the cluster plots highlight how in the processed data, the clusters are more compact and well-separated, indicating a clearer delineation between groups while, in the raw data, clusters appear diffuse and overlapping, with less distinct boundaries.



(a) Clusters on processed data



(b) Clusters on raw data

Figure 3: Clusters visualizations