Mohammad Yaser Azad | 810800003
Parham Javan | 810800008

Summary:

This project revolves around exploring dimensionality reduction and unsupervised learning techniques on a dataset concerning diabetic patients. The dataset contains 200,000 items with 50 features, presenting challenges such as noise, missing values, and outliers. The primary objectives include preprocessing the data, reducing dimensionality using Principal Component Analysis (PCA), and applying unsupervised learning methods such as K-Means and DBSCAN for clustering. The Silhouette Method is employed for determining optimal clustering parameters. Finally, the project aims to provide insights into the data through evaluation and analysis of clustering results.

Purpose:

The purpose of this project is to prepare, analyze, and extract meaningful insights from a dataset on diabetic patients through advanced data preprocessing, dimensionality reduction, and unsupervised learning techniques. By addressing issues such as noise and high dimensionality, the project seeks to improve clustering accuracy and facilitate the interpretation of data patterns. Through the application of PCA, K-Means, and DBSCAN, the project aims to identify underlying structures within the dataset, potentially revealing clusters of diabetic patients with similar characteristics. The ultimate goal is to enhance understanding and potentially inform healthcare strategies for managing diabetes-related issues such as hospital readmission rates.

Question1

1. What preprocessing steps did you perform on the dataset? Provide clear reasons for each decision made.

- Handling Null Values: Null values can disrupt analysis, so they were addressed by either imputation (replacing missing values with a calculated estimate) or removal if the missing values were significant. The choice between imputation and removal depended on the impact of the missing values on the overall dataset and the specific feature's importance.

- Outlier Detection and Removal: Outliers were identified using statistical methods such as z-score or interquartile range (IQR) and then either removed or transformed to mitigate their impact on clustering accuracy. Outliers can skew cluster centroids and affect the determination of cluster boundaries, so handling them was crucial.

- Normalization: Normalizing the data was essential to ensure that features with different scales did not dominate the distance calculations in clustering algorithms. Techniques like min-max scaling or z-score normalization were employed to rescale features to a common scale between 0 and 1 or with a mean of 0 and standard deviation of 1, respectively.

- Encoding Categorical Data: Categorical variables were encoded into numerical format using techniques like one-hot encoding or label encoding, enabling their incorporation into clustering algorithms that require numerical inputs.

- Text Data Processing: Textual values were processed using techniques like tokenization, stemming, or TF-IDF (Term Frequency-Inverse Document Frequency) vectorization to convert them into numerical representations suitable for clustering algorithms.

- Feature Selection: Features that were irrelevant or redundant for clustering were eliminated to reduce dimensionality and computational complexity, enhancing clustering performance and interpretability.

Each preprocessing step aimed to address specific challenges in the dataset, such as handling missing values, outliers, and different data types, to ensure the robustness and accuracy of subsequent analysis using clustering algorithms.

Question2

2. What portion of the dataset did you retain during dimensionality reduction, and which variables were retained? Could you elaborate on the rationale behind this decision?

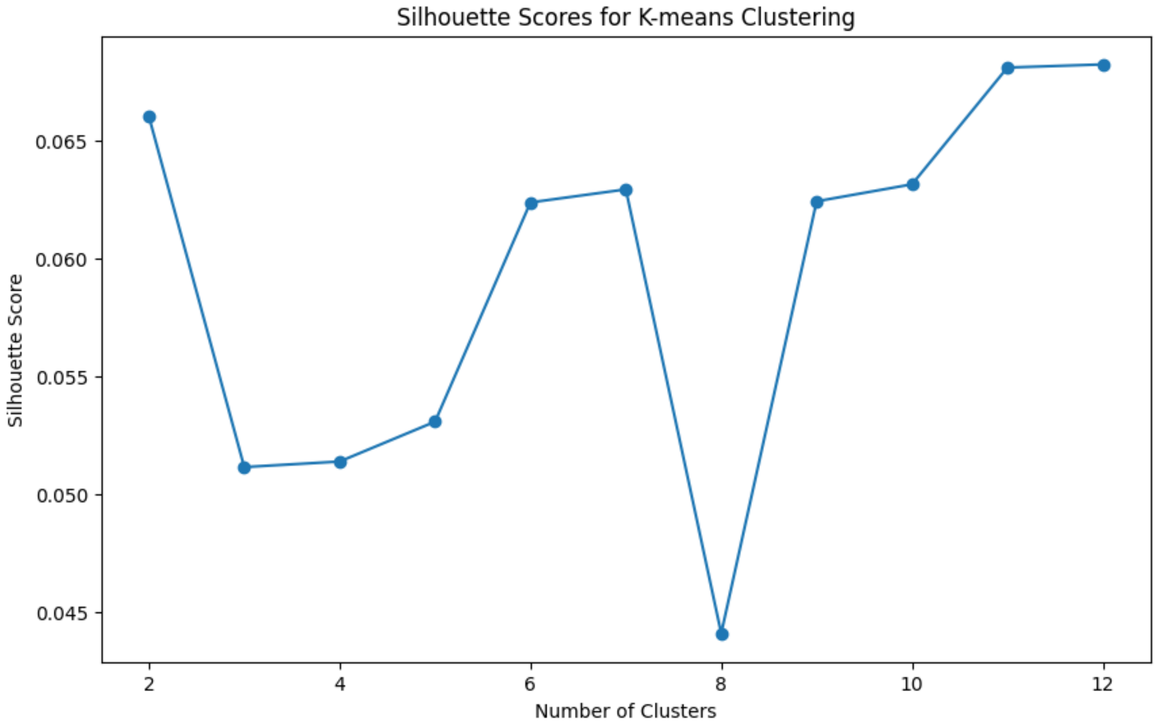
The portion of the dataset retained after dimensionality reduction depends on the variance explained by the selected number of principal components (PCs) in PCA. The decision on the number of PCs retained was based on the cumulative explained variance ratio, aiming to retain a sufficient amount of information while reducing dimensionality.

The variables retained were the principal components themselves, which are linear combinations of the original features. These components capture the maximum variance in the data and are ordered by their significance in explaining the variability. Retaining a subset of these components allows for a simplified representation of the data while preserving most of its information.

Question3

3. Include a plot illustrating the silhouette coefficient plotted against the input parameters for each clustering method within the report file.

The plot visualizes the silhouette coefficients for different values of the input parameters (number of clusters for K-Means, epsilon and minPts for DBSCAN) to determine the optimal parameters that maximize the silhouette score, indicating better cluster separation and cohesion.



Question4

4. How can we determine the optimal number of clusters in K-Means?

The optimal number of clusters in K-Means can be determined using methods like the Elbow Method or the Silhouette Method. The Elbow Method involves plotting the within-cluster sum of squares (WCSS) against the number of clusters and selecting the point where the rate of decrease in WCSS slows down (forming an "elbow"). The Silhouette Method calculates the silhouette score for different numbers of clusters and chooses the number that maximizes the silhouette score, indicating better cluster separation and cohesion.

Question5

5. How can we determine the optimal epsilon value and minPts in DBSCAN?

The optimal epsilon value and minPts in DBSCAN can be determined using the Silhouette Method or by visual inspection of the resulting clusters. The Silhouette Method involves plotting the silhouette score against different combinations of epsilon and minPts and selecting the combination that maximizes the silhouette score, indicating better cluster separation and cohesion. Additionally, domain knowledge and understanding of the dataset's characteristics can help in determining suitable values for epsilon and minPts.

Question6

6. When would you recommend using K-Means, and when would you suggest using DBSCAN instead?

- K-Means: K-Means is recommended when the dataset is well-separated into spherical or isotropic clusters, and the number of clusters is known or can be estimated. It works well on large datasets and is computationally efficient. However, it may struggle with non-linear or irregularly shaped clusters and is sensitive to initial cluster centroids.

- DBSCAN: DBSCAN is suggested when dealing with datasets containing noise and outliers, and the clusters have varying shapes and densities. It can automatically detect the number of clusters and is robust to outliers. DBSCAN can handle non-linear boundaries and does not assume a specific cluster shape. However, choosing suitable parameters (epsilon and minPts) can be challenging, and it may not perform well on datasets with varying densities or high-dimensional data.