Breast Cancer Dataset Analysis Report

1. Introduction

K clustering is Unsupervised Machine Learning is the process of teaching a computer to use unlabelled, unclassified data and enabling the algorithm to operate on that data without supervision. The goal of clustering is to divide the population or set of data points into a number of groups so that the data points within each group are more comparable to one another and different from the data points within the other groups. It is essentially a grouping of things based on how similar and different they are to one another. In the context of the dataset I chose , which includes information about breast cancer dataset aimed to helping diagnose patients with breast cancer given certain features. In this project the K-means clustering is achieved using the scikit-learn library in Python. The goal was is to primarily investigate if inherent patterns or clusters exist within the data or not.

2. Data Exploration

To prepare my data set I extracted the relevant columns from the dataset.
I started by checking the dataset for any missing or erroneous data. Ensure that there are no missing values in the columns I plan to use for clustering. If I found any missing data,I either remove those rows or impute missing values using methods like mean, median, or interpolation, depending on dataset and the extent of missing data. In my dataset, I had no missing values. The dataset was loaded and examined, containing 569 samples with 30 features each. The analysis concentrated on the underlying data patterns, side lining the actual diagnosis labels for the time being.

3. Data Processing

The features were standardized using StandardScaler. Standardization is essential for PCA, ensuring that all features contribute equally to the analysis by removing the influence of their magnitudes.

4. Principal Component Analysis (PCA)

PCA was performed to reduce the dimensionality of the dataset. The analysis revealed that 90% of the dataset's variance could be retained by considering the first 7 principal components out of the 30 available features. This reduction allowed for a more manageable and interpretable dataset.

5. K-means Clustering

K-means clustering was employed on the reduced dataset. The elbow method was utilized to determine the optimal number of clusters, indicating that 2 clusters were suitable for the data. Visualization of the clusters showed a clear separation, indicating inherent patterns within the data. Perform kmeans clustering on the projected data or in PCA space with 4 components. To determine optimal k value we do a parameter sweep over values in the range [1,11], the elbow kink occurs at 4. So we use 4 as number of clusters in the dataset. The features that accurately capture these clusters are the Annual Income and the Spending Score of a customer. The clusters can be described as :he clusters can be described as - 0 - Medium to High spending score (typically above 50), aged 40 and below with annual income below 60k - 1 - High spending score (typically above 70) , aged between 30 and 40 and annual income above 80k - 2 - Spending score below 45 and annual income above 60k - 3 - Spending score below 60 and annual income below 60k

6. Conclusion and Implications

The analysis successfully identified 2 distinct clusters within the breast cancer dataset. While the analysis was performed without considering the labels, the identified clusters might correspond to different subtypes of breast cancer or distinct patient profiles. Further research and domain expertise are required to interpret these clusters effectively. However, the analysis showcases the potential of unsupervised learning techniques in exploring complex datasets.

7. Limitations

While does reduce dimensionality we lose interpret-ability i.e I discarded a component but do not necessarily know which feature it corresponds to in original feature space.
The analysis focused on exploring inherent patterns without considering the diagnosis labels. While this approach provides insights into the data's structure, interpreting the clusters in the context of breast cancer subtypes or patient profiles requires domain-specific knowledge. Future work could involve integrating the diagnosis labels to validate the identified clusters and explore their clinical relevance. Additionally, further research could include other unsupervised learning algorithms and techniques for a comprehensive understanding of the dataset.