Optimizing Blood Donation Campaigns through Data Exploration of the Blood-Transfusion Dataset

Data Mining

International Masters of Advanced Methods in Particle Physics

Names: **Edgar Eduardo MATA MENDOZA**
Syeda Eman Zahra

Date: 2023.12.22

UNIVERSITÉ Clermont Auvergne

Analysis of the Blood-Transfusion Dataset

Edgar Eduardo MATA MENDOZA, Syeda Eman Zahra

December 22, 2023

Contents

1 Chapter 1: Data Set Characteristics and Exploration **4**
- 1.1 Dataset Overview . . . 4
  - 1.1.1 Attributes Information . . . 4
- 1.2 Data Preprocessing . . . 5
  - 1.2.1 Data Normalization/Standardization . . . 5
- 1.3 Data Visualization . . . 5
  - 1.3.1 Summary Statistics . . . 6

2 Chapter 2: Attribute Selection/Extraction **7**
- 2.1 Normalization . . . 7
- 2.2 PCA Implementation . . . 7
  - 2.2.1 PCA Results . . . 7

3 Chapter 3: Clustering Results and Analysis **9**
- 3.1 Clustering Algorithms . . . 9
- 3.2 Elbow Method . . . 9
- 3.3 K-Means Implementation with 2 Clusters . . . 10
  - 3.3.1 Implementation Details . . . 10
  - 3.3.2 Visualization of Clustering Results . . . 10
  - 3.3.3 Internal Validation Metrics . . . 11
- 3.4 K-Means Implementation Using Functions . . . 11
  - 3.4.1 Procedural Approach . . . 11
  - 3.4.2 Results of Functions-Based K-Means . . . 12
  - 3.4.3 Discussion on Clustering Outcomes . . . 12
- 3.5 Gaussian Mixture Models (GMM) Implementation . . . 13
  - 3.5.1 Implementation Details . . . 13
  - 3.5.2 Visualization of GMM Clustering . . . 13
  - 3.5.3 Internal Validation Metrics for GMM . . . 14
  - 3.5.4 Discussion on GMM Clustering Outcomes . . . 14
- 3.6 Hierarchical Clustering Algorithm (HCA) Implementation . . . 14
  - 3.6.1 Implementation Details . . . 14
  - 3.6.2 Visualization of HCA Clustering . . . 14
  - 3.6.3 Internal Validation Metrics for HCA . . . 15

3.6.4 Discussion on HCA Clustering Outcomes 15
3.6.5 Dendrogram Analysis of HCA 16
3.6.6 Analysis of Individual Dendrograms 16
3.6.7 Internal Validation Metrics for HCA Dendrograms 17

4 Chapter 4: External Index Validation 18
4.1 Dataset Labels as External Index 18
4.2 Visualization via Scatterplots 18
4.3 Pairplot Analysis 19
4.4 Accuracy of Clustering Methods 20
4.5 Pairplot for K-Means (OOP) Labels 20

5 Chapter 5: Performance Evaluation Using Confusion Matrix 21
5.1 Confusion Matrix Analysis 21
5.2 Interpretation of Results 21
5.3 Model Accuracy 22
5.4 Critical Evaluation 22

6 Chapter 6: Conclusion and Discussion 22

A Appendices 24
A.1 Python Code 24
A.2 References 24

Introduction

Setting the Scene for Data-Driven Donor Segmentation
The lifeline of healthcare systems globally, blood donation, requires a steady and reliable supply to meet the urgent and varied needs of patients. Ensuring this supply involves not just encouraging generous individuals to donate but also understanding the patterns and preferences of these donors. This is where the power of data mining and machine learning comes to the forefront, offering robust tools to dissect complex datasets and glean actionable insights.

The Imperative of Efficient Donor Segmentation

Blood donation campaigns have long grappled with the challenge of effectively targeting potential donors. The traditional one-size-fits-all approach has given way to the need for a more nuanced strategy—donor segmentation. By classifying donors into distinct groups based on their donation behaviors, campaigns can tailor their outreach, making it more personal, persuasive, and ultimately, successful.

Objective of the Study

This study aims to leverage unsupervised learning algorithms to analyze the Blood-Transfusion Service Center Dataset from Hsin-Chu City in Taiwan. The objective is twofold: to understand the underlying donation patterns and to identify characteristics of high-value donors who are crucial to the sustainability of blood donation systems.

Methodological Overview

Employing a methodical approach, the project will:

- Preprocess the data to ensure it is primed for analysis, involving steps such as cleaning, normalization, and dimensionality reduction.
- Utilize prominent unsupervised learning algorithms such as K-Means, Gaussian Mixture Models, and Hierarchical Clustering to segment the donor data.
- Compare these methods using internal validation metrics like the Silhouette Index and external indices provided within the dataset to validate the models.
- Visualize the results using a range of techniques, from scatter plots to heatmaps, to offer a clear view of the clustering outcomes.

Relevance to Blood Donation Campaigns

The relevance of this analysis cannot be overstated. As blood donation organizations strive for efficiency, understanding donor segments allows for more focused and cost-effective campaigns. The insights derived from this study have the potential to revolutionize how donation drives are conceptualized and executed.

Structure of the Report

The report unfolds over several chapters, each addressing critical aspects of the analysis:

- Following this introduction, Chapter 1 focuses on analyzing the characteristics of the dataset.
- Chapter 2 delves into data collection, preprocessing, and transformation.
- Chapter 3 presents the data analysis, visualization, and the application of clustering algorithms.
- Chapter 4 tackles the external validation of the clustering results and their comparative analysis.

- Chapter 5 evaluates the performance of the clustering algorithms using a confusion matrix.
- The report culminates in a comprehensive conclusion and discussion in Chapter 6, synthesizing the findings and their implications for blood donation campaigns.

Anticipated Contributions

The anticipated contributions of this report are substantial. By presenting a detailed comparison of clustering algorithms and their implications on donor segmentation, the report aims to serve as a guide for data-driven decision-making in the domain of healthcare and public welfare.

This introduction sets the stage for the exploration and analysis that will follow, detailing the objectives, methodology, and anticipated contributions of the study. The structured approach outlined here prepares the reader for an understanding of the potential of data mining in enhancing blood donation campaigns.

1 Chapter 1: Data Set Characteristics and Exploration

1.1 Dataset Overview

Introduction to the Blood-Transfusion dataset.

Data Set Characteristics: Multivariate
Number of Instances: 748
Area: Business
Attribute Characteristics: Real
Number of Attributes: 5
Date Donated: 2008-10-03
Associated Tasks: Classification
Missing Values: None

1.1.1 Attributes Information

- R (Recency): Months since last donation.
- F (Frequency): Total number of donations.
- M (Monetary): Total blood donated in c.c. (cubic centimeters)
- T (Time): Months since first donation.
- Binary Variable: Representing whether the donor donated blood in March 2007 (1 = yes, 0 = no).

1.2 Data Preprocessing

- Data reading using pandas.
- Initial dataset exploration using df.head().
- Checking for Missing Values: There are no missing values in the dataset, which is consistent with the dataset description.

1.2.1 Data Normalization/Standardization

- The data (excluding the binary variable indicating whether the donor donated blood in March 2007) has been standardized. Standardization is important for PCA and clustering algorithms because these methods are sensitive to the scale of the data.
- The resulting standardized data is stored in a new DataFrame.

1.3 Data Visualization

Using matplotlib.pyplot and seaborn for visual exploration.

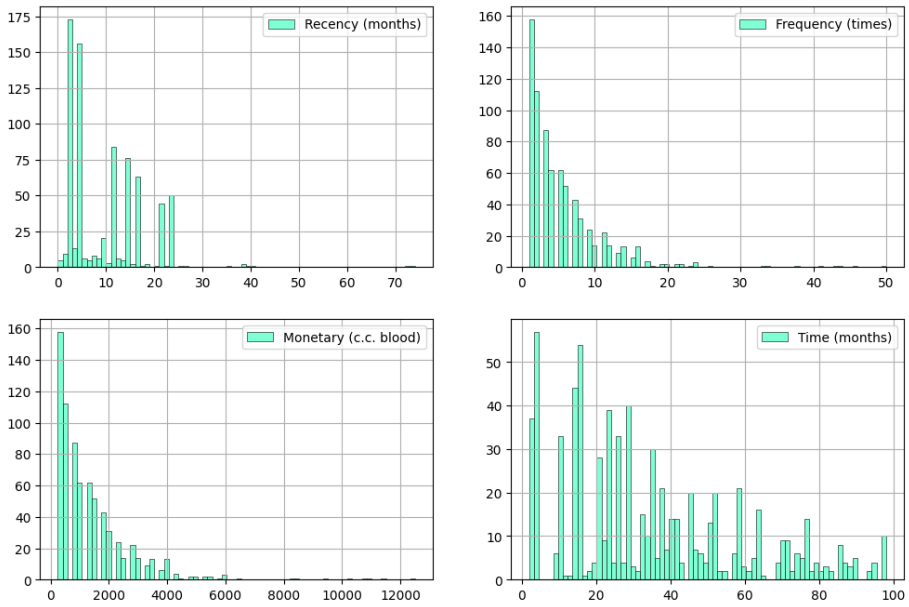

Figure 1: Distribution of Attributes

1.3.1 Summary Statistics

The histograms for each attribute of the dataset have been plotted. These histograms provide a visual representation of the distribution of each attribute:

1. Recency (months)
2. Frequency (times)
3. Monetary (c.c. blood)
4. Time (months)

2 Chapter 2: Attribute Selection/Extraction

2.1 Normalization

Employing StandardScaler for data normalization.

2.2 PCA Implementation

Utilizing PCA for dimensionality reduction.

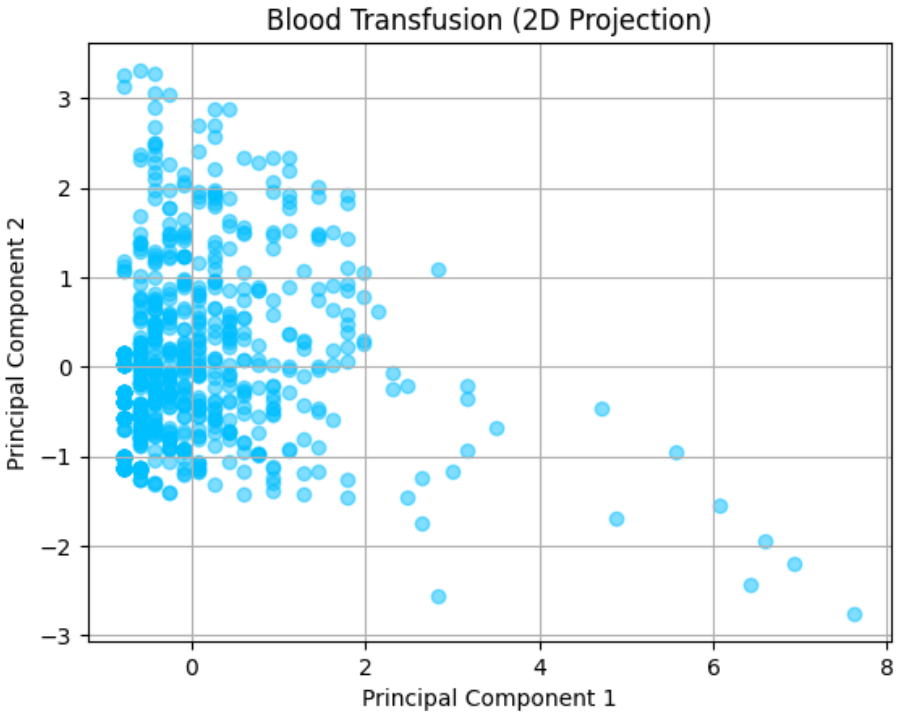

Figure 2: Visualization of principal components analysis of Blood Transfusion Dataset.

The PCA (Principal Component Analysis) has been successfully applied to the dataset, reducing it to two dimensions. This is useful for visualization purposes and can also help in understanding the data's structure.

2.2.1 PCA Results

- Visualization of the dataset: The scatter plot above shows the dataset projected onto the first two principal components. This visualization can help in identifying patterns or clusters in the data.

- Explained Variance:
  - The first principal component explains approximately 63.53% of the variance.
  - The second principal component accounts for about 27.53% of the variance.
  - Together, they explain about 91.06% of the total variance in the dataset.

3 Chapter 3: Clustering Results and Analysis

3.1 Clustering Algorithms

The following clustering algorithms will be used:

- K-Means and Gaussian Mixture Models.
- Implementation of Hierarchical Clustering Algorithm.

For the K-Means, 2 different implementations from scratch will be developed. One using Object-Oriented Programming, and the other one defining functions on Python.

3.2 Elbow Method

The "Elbow Method" is an algorithm that allows us to define the optimal number of clusters to be used for a specific dataset.

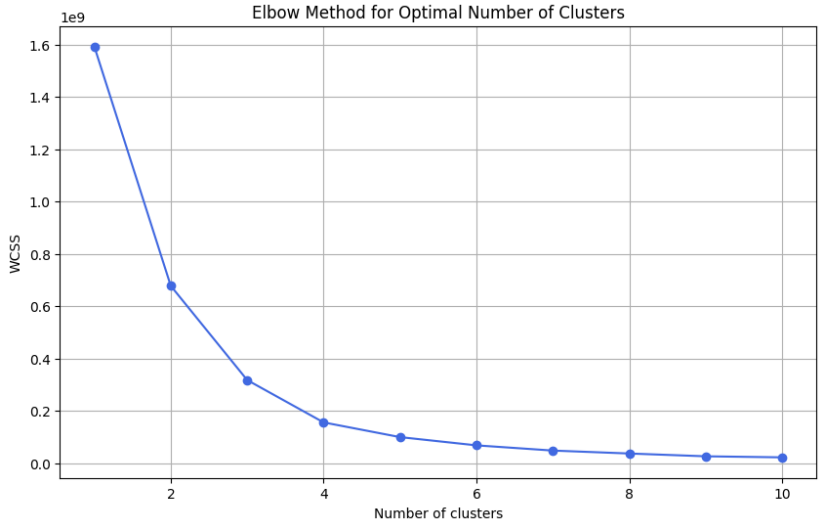

Figure 3: The Elbow Method

The plot from implementing "The elbow method" on our data set indicates the within-cluster sum of squares (WCSS) for different numbers of clusters. The "elbow" point in the plot is where the WCSS begins to decrease at a slower rate, suggesting that adding more clusters does not significantly improve the fit of the model. This point represents the most appropriate balance between the number of clusters and the variance explained.

From the plot, it looks like the elbow point could be at around 2-4 clusters, as the line starts to flatten after this point. We choose 2 clusters because we want a finer clustering.

3.3 K-Means Implementation with 2 Clusters

3.3.1 Implementation Details

- A K-Means clustering algorithm was implemented from scratch using object-oriented programming principles.
- The class `KMeansCustom` encapsulates all the functionalities required for clustering, including initialization of centroids, computation of distances, assignment of clusters, and updating centroids after each iteration.
- This custom implementation allows for a deeper understanding of the inner workings of the algorithm and provides flexibility in the clustering process.

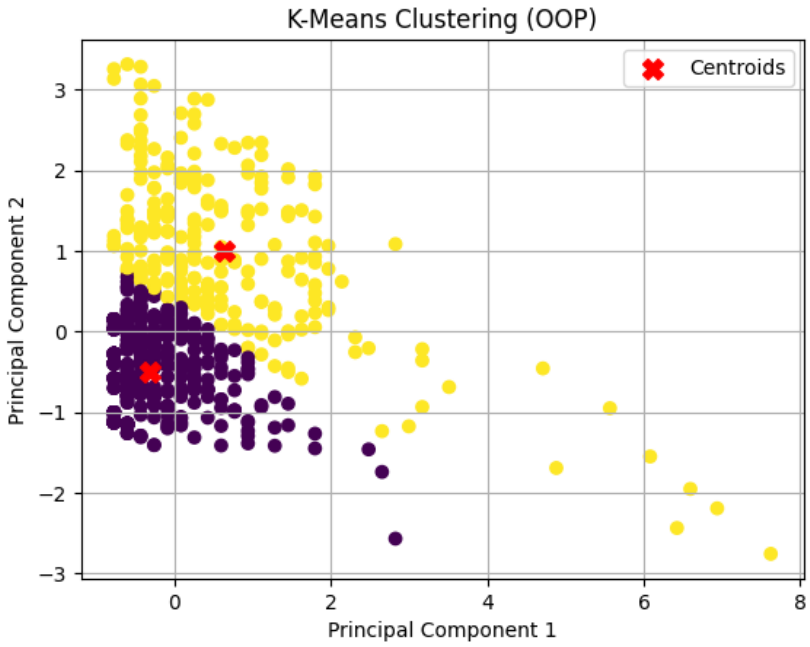

Figure 4: K-Means Implementation with two clusters.

3.3.2 Visualization of Clustering Results

- The clusters resulting from the custom K-Means algorithm were visualized on the PCA-reduced data.
- The scatter plot distinguishes two clusters with the centroids marked, showcasing the grouping determined by the algorithm.

3.3.3 Internal Validation Metrics

Silhouette Index Analysis

- The Silhouette Index was calculated for the K-Means clustering results to measure the quality of the clustering.
- A Silhouette Index of 0.4525 suggests a moderate separation between clusters, indicating that while there is some overlap, the clusters are relatively distinct.
- The value implies that the clustering has been reasonably effective, with most points being closer to their own cluster centroid than to others.

3.4 K-Means Implementation Using Functions

3.4.1 Procedural Approach

- A functions-based implementation of K-Means clustering was carried out as an alternative to the OOP approach.
- This procedural method involved creating functions for calculating distance matrices, determining the membership of each point (i.e., the cluster they belong to), and updating cluster centers.
- Two random points from the PCA-reduced data were chosen as initial centroids, and the algorithm was iterated for 20 iterations to refine the cluster centers.

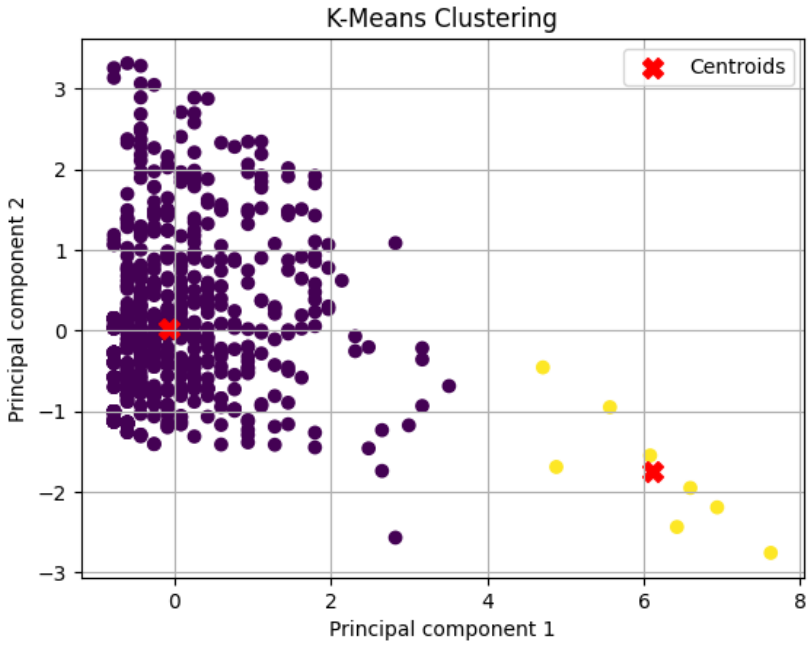

Figure 5: K-Means Implementation based on functions.

3.4.2 Results of Functions-Based K-Means

- The clusters resulting from this functions-based implementation were visualized, demonstrating the algorithm's ability to group data points effectively.
- The new cluster centers were identified and plotted, corresponding to the most representative points of the segmented donor behaviors.
- The labels generated by this implementation allowed for the visualization of data points and their respective clusters, providing an alternative perspective to the OOP-based K-Means results.

3.4.3 Discussion on Clustering Outcomes

Analysis of Clusters

- The PCA-reduced data led to two distinct clusters, suggesting a separation in the dataset that could correlate with different donor behaviors.
- The centroids indicate the average location of each cluster in the reduced feature space, which can be interpreted as the central profile of each donor segment.

- Observations from the plot may suggest that one cluster represents more frequent donors (lower recency and higher frequency), while the other may consist of less frequent donors.
- These findings are critical for identifying potential high-value donors and understanding the overall distribution of donation patterns.

3.5 Gaussian Mixture Models (GMM) Implementation

3.5.1 Implementation Details

- The GMM algorithm was applied to the PCA-reduced data using the `GaussianMixture` class from the `sklearn.mixture` library.
- The model was created specifying two components, reflecting the assumption of two clusters, and was fitted to the data to predict the labels.

3.5.2 Visualization of GMM Clustering

Cluster Visualization

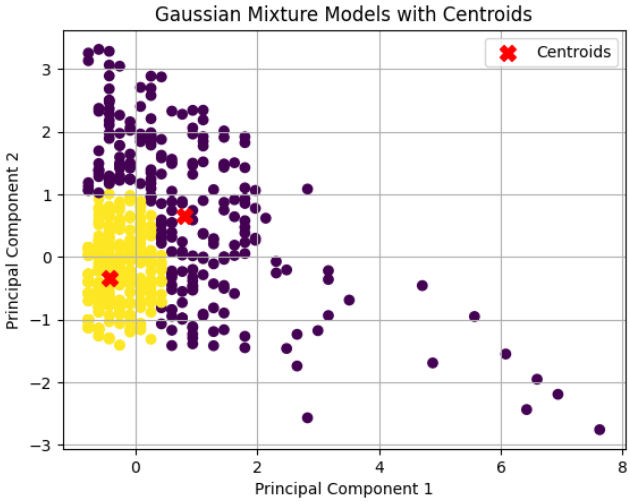

Figure 6: Gaussian Mixture Models (GMM) Implementation with two clusters

- The clusters from the GMM were visualized, with each point colored according to the cluster it was most likely to belong to.
- The centroids of each Gaussian component, representing the mean of the clusters, were plotted to visualize the centers of the data distributions.

3.5.3 Internal Validation Metrics for GMM

Silhouette Index Analysis

- The Silhouette Index for the GMM clustering was computed, resulting in a score of 0.4307.
- This score is slightly lower than that of the K-Means clustering, indicating that the clusters are not as distinctly separated.
- The Silhouette Index suggests that while there is some definition between clusters, there may be some overlap or the clusters are not as cohesive as could be desired.

3.5.4 Discussion on GMM Clustering Outcomes

Analysis of Clusters

- GMM provides a probabilistic clustering approach, which might be more appropriate for datasets with overlapping clusters or non-spherical distributions.
- The centroids in the GMM plot help to interpret the Gaussian components and the distribution of the data.
- The results suggest that there are two main distributions of donor behavior within the dataset, which could be explored further to understand the underlying characteristics of each group.

3.6 Hierarchical Clustering Algorithm (HCA) Implementation

3.6.1 Implementation Details

- The Hierarchical Clustering Algorithm was implemented using different linkage criteria: single, complete, average, and ward.
- The linkage function from the `scipy.cluster.hierarchy` module was used to create linkage matrices, which describe the hierarchical clustering process.

3.6.2 Visualization of HCA Clustering

Cluster Visualization

- Clusters resulting from HCA using the four different linkage criteria were visualized to assess their structures.
- Each set of clusters was plotted with centroids to observe the effects of different linkage criteria on the clustering outcome.

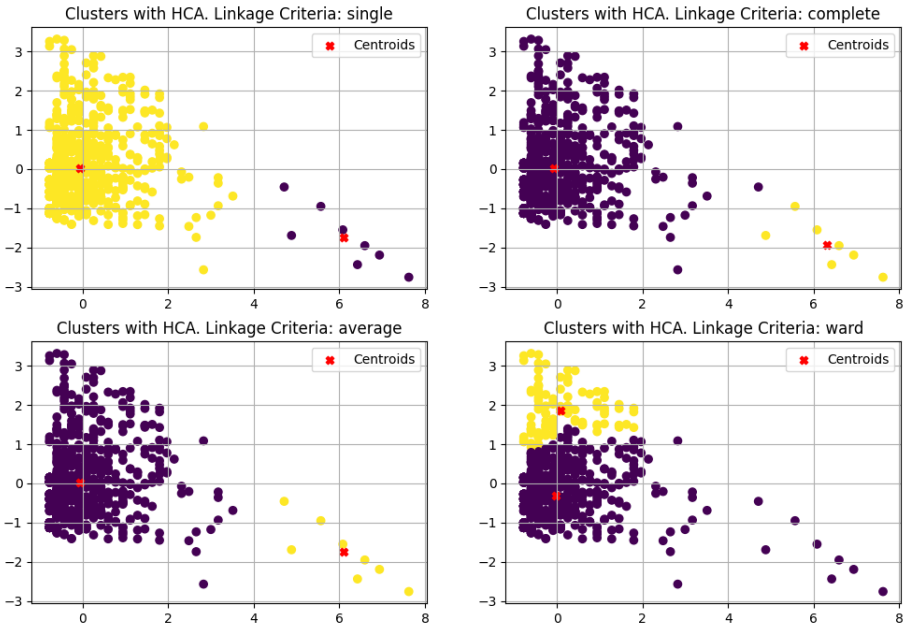

Figure 7: Hierarchal Clustering Algorithm (HCA) Implementation with two clusters

3.6.3 Internal Validation Metrics for HCA

Silhouette Index Analysis

- The Silhouette Index was calculated for each set of clusters formed by the different linkage criteria.
- The scores for single, complete, and average criteria were notably high, suggesting well-separated clusters. In contrast, the ward criterion had a lower score, indicating less distinction between clusters.
- The high Silhouette Index for single and complete linkage criteria indicates that clusters are more clearly delineated, with single linkage forming elongated clusters and complete linkage resulting in more compact clusters.

3.6.4 Discussion on HCA Clustering Outcomes

Analysis of Clusters

- Different linkage criteria led to varied clustering structures, which can significantly affect the interpretation of donor behaviors.

- The clusters obtained with single and complete linkage show a high degree of separation, which may be beneficial for identifying distinct donor groups.
- The clusters formed by the ward criterion are more cohesive, which could be advantageous for identifying homogeneous donor segments.

3.6.5 Dendrogram Analysis of HCA

Creation of Dendrograms

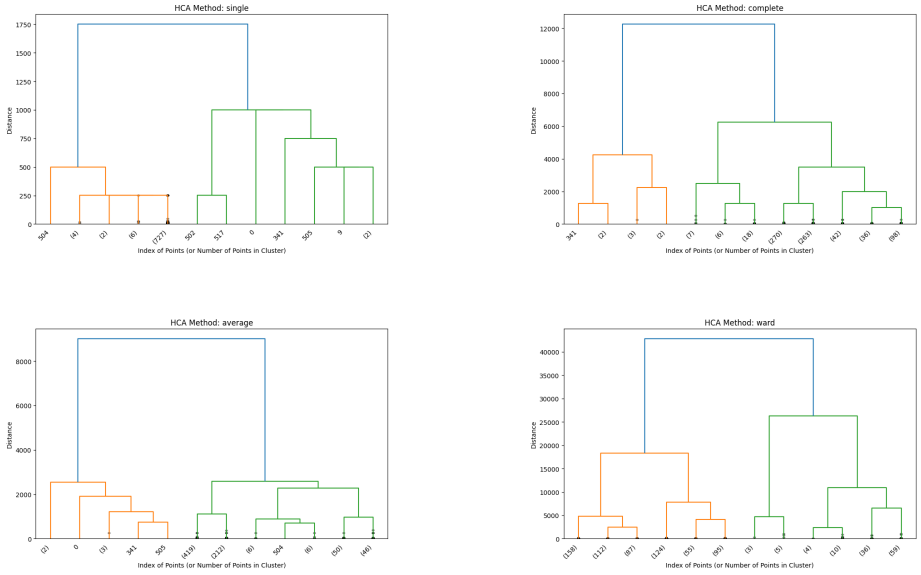

Figure 8: Hierarchal Clustering Algorithm (HCA) Dendograms

- Dendrograms for each linkage criterion (single, complete, average, and ward) were created to visualize the hierarchical clustering process.
- Each dendrogram illustrates the agglomerative process, starting with individual points and merging them into clusters based on the distance or similarity measures.

3.6.6 Analysis of Individual Dendrograms

Single Linkage

- Exhibits a tendency to create long chains that can lead to less balanced cluster sizes.

- The dendrogram indicates that some clusters are merged at a very low distance, suggesting high similarity between points within the same cluster.

Complete Linkage

- Tends to create more uniform clusters, merging clusters based on the maximum distance between points in different clusters.
- This linkage criterion resulted in relatively balanced clusters, as seen in the dendrogram where the height at which clusters merge is more uniform.

Average Linkage

- Averages the distance between all pairs of points in any two clusters before merging, leading to a balance between the single and complete linkage criteria.
- The dendrogram shows that clusters are merged at intermediate distances, reflecting a compromise between single and complete linkage properties.

Ward Linkage

- Minimizes the total within-cluster variance, tending to create more compact clusters.
- The dendrogram for the ward linkage shows larger distances at which clusters merge, indicating a preference for groupings that minimize within-cluster variance.

3.6.7 Internal Validation Metrics for HCA Dendrograms

Silhouette Index Analysis

- The Silhouette Index scores for each linkage criterion were calculated to validate the quality of the clusters formed.
- The scores for single and complete linkage were notably high, suggesting distinct and well-separated clusters, while the ward criterion had a lower score, indicating less separation.
- The dendrograms complement these scores by visually representing the compactness and separation of clusters, with varying heights indicating how clusters are merged at different levels of similarity.

4 Chapter 4: External Index Validation

4.1 Dataset Labels as External Index

- The original dataset labels indicate whether an individual donated blood in March 2007. These labels serve as an external index for validating the clustering results.
- The comparison of clustering labels with the original labels allows for an assessment of how well the clustering algorithms have performed in replicating the natural grouping within the dataset.

Comparative Visualization of Clustering Results

4.2 Visualization via Scatterplots

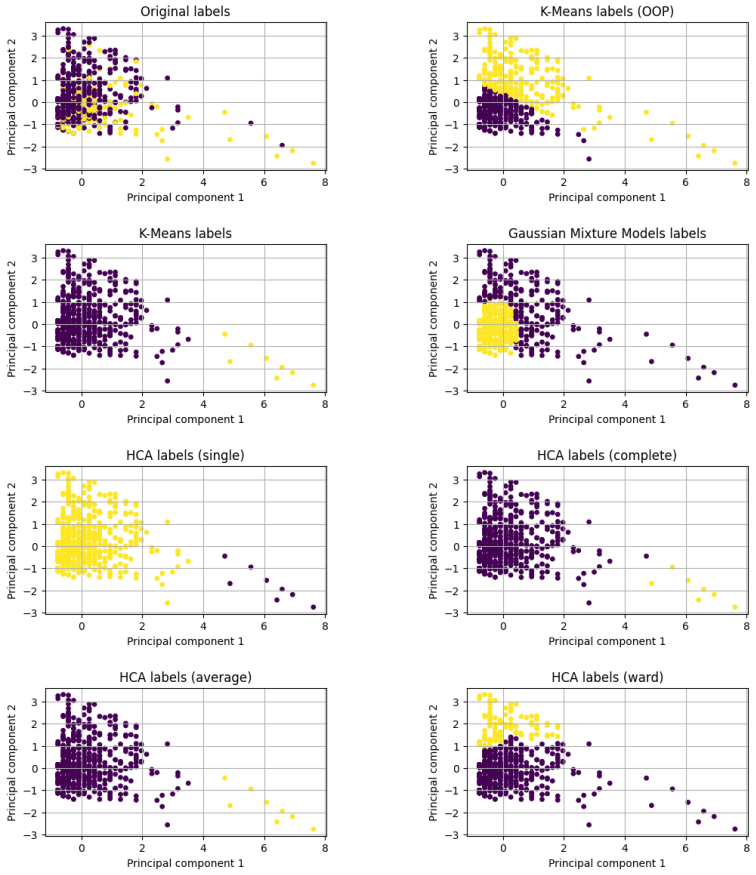

Figure 9: Pairplot Analysis

- A series of scatterplots were created to visualize the PCA-reduced data points, with each plot showing the clusters formed by a different algorithm against the original labels.
- The scatterplots demonstrate how closely each clustering method aligns with the original dataset's inherent groupings.

Analysis of Pairplot Visualizations

On Python, the Seaborn Library was used to get an understanding of the correlation and behavior of the features involved in the dataset.

4.3 Pairplot Analysis

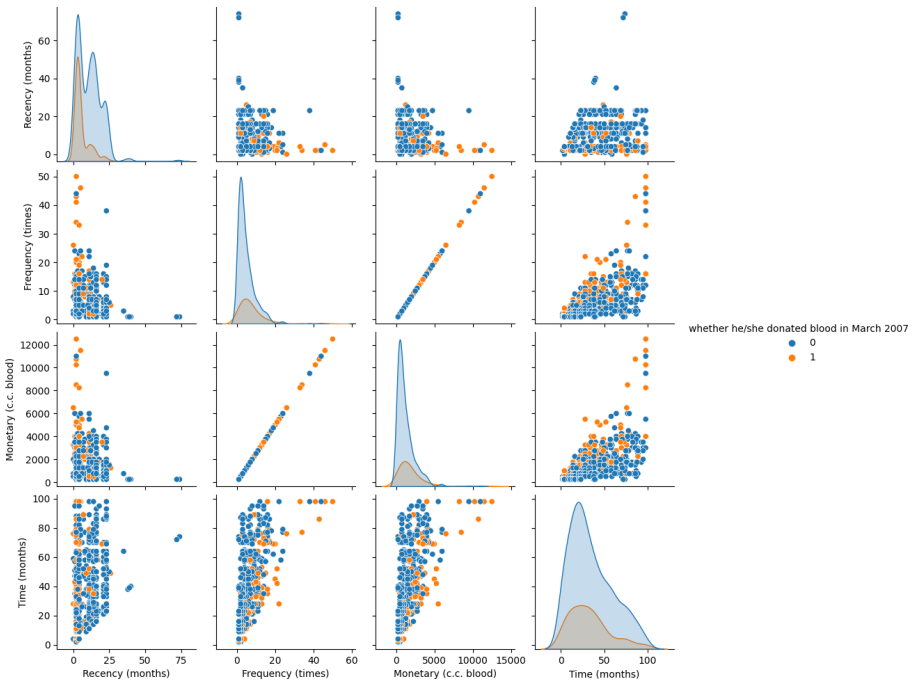

Figure 10: Pairplot Analysis

- The seaborn pairplot visualizations provide a matrix of all possible pairwise plots between variables, incorporating the hue based on the clustering labels.
- These visualizations offer insights into the relationships between features and how they correlate with the clustering results.

Evaluation of Clustering Labels

4.4 Accuracy of Clustering Methods

- The K-Means (OOP) and HCA with ward criterion labels closely match the original labels, suggesting a higher level of accuracy in capturing the natural groupings within the data.
- The comparison reveals that the K-Means (OOP) method, in particular, aligns well with the original labels, indicating its effectiveness in segmenting the dataset.

Insights from Seaborn Pairplot

4.5 Pairplot for K-Means (OOP) Labels

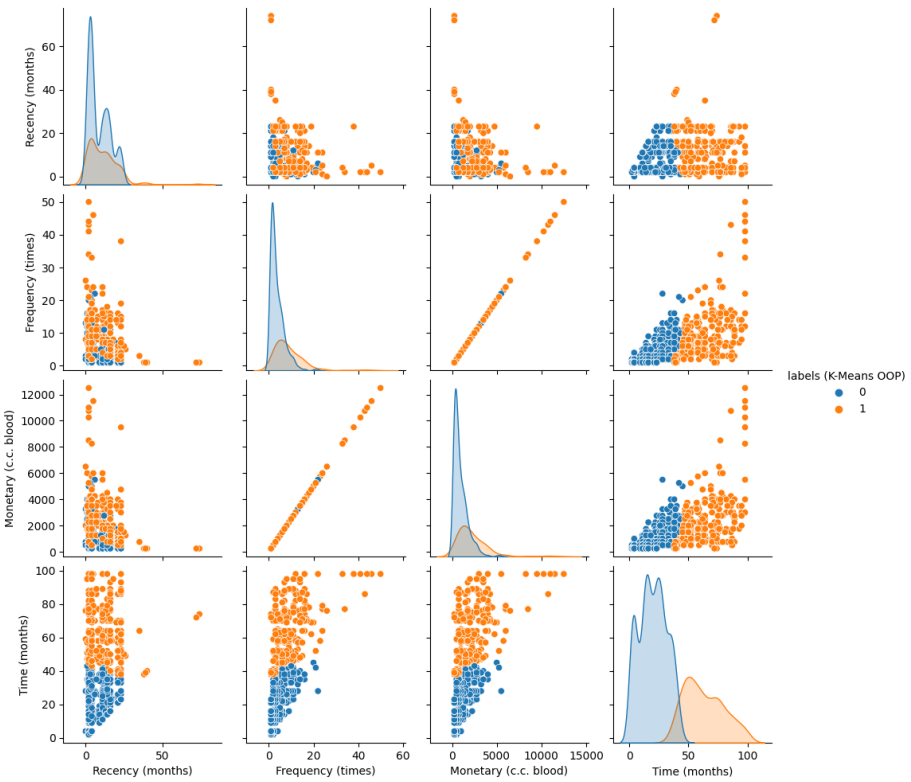

Figure 11: Pairplot for K-Means (OOP) Labels

- A seaborn pairplot was also created for the K-Means (OOP) labels, allowing for a direct comparison with the original labels in terms of distribution and separation across different features.
- The pairplot reveals that while some clusters are distinctly separated, others may overlap, providing a view of the dataset's structure.

5 Chapter 5: Performance Evaluation Using Confusion Matrix

5.1 Confusion Matrix Analysis

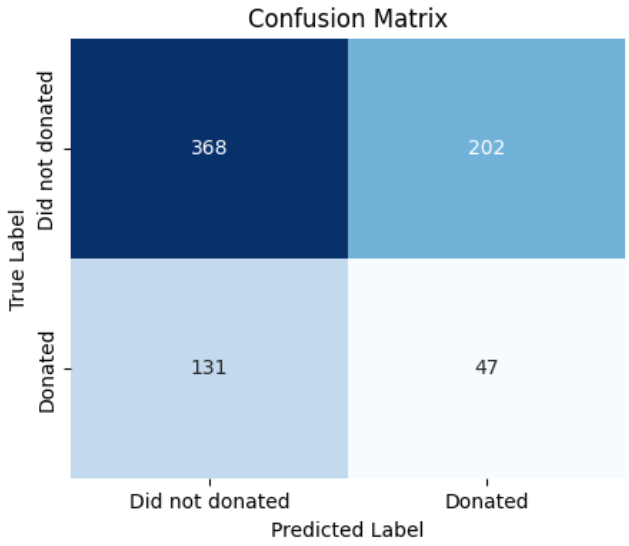

Figure 12: Confusion Matrix

- A confusion matrix was constructed to evaluate the performance of the K-Means (OOP) clustering model.
- The matrix provides a tabular representation of the actual versus predicted labels, offering insights into the true positives, false positives, true negatives, and false negatives.

5.2 Interpretation of Results

- The confusion matrix shows that the model correctly identified 368 instances where individuals did not donate and 47 instances where they did, as per the original dataset labels.

- However, there are a significant number of false positives (202 instances) and a smaller number of false negatives (131 instances), indicating some level of misclassification by the model.

5.3 Model Accuracy

- The accuracy of the model, calculated as the sum of true positives and true negatives over the total number of instances, was approximately 55.48
- This level of accuracy suggests that while the model has a moderate ability to distinguish between donors and non-donors, there is room for improvement.

5.4 Critical Evaluation

- The false negatives and false positives highlight potential areas where the model's clustering mechanism does not align perfectly with the actual donation behavior.
- The dataset's suitability for clustering tasks may be questioned, as the inherent noise and overlap in the feature space could lead to challenges in accurately segmenting donors based on the available attributes.

6 Chapter 6: Conclusion and Discussion

The study conducted a comprehensive analysis of a blood transfusion dataset using various unsupervised learning algorithms to uncover patterns and segments within donor behaviors. The primary clustering algorithms evaluated were K-Means, Gaussian Mixture Models (GMM), and Hierarchical Clustering Analysis (HCA), each with different linkage criteria.

Comparison of Clustering Algorithms

- The K-Means algorithm, particularly the Object-Oriented Programming (OOP) implementation, demonstrated moderate effectiveness with an accuracy of approximately 55.48%. It provided distinct clusters that corresponded reasonably well with the actual labels, suggesting its potential utility in segmenting donors.
- The GMM offered a probabilistic approach to clustering, revealing overlapping clusters that could correspond to ambiguous cases in donor behavior. However, the Silhouette Index for GMM was slightly lower than K-Means, indicating less distinct separation between clusters.
- HCA with various linkage criteria presented different clustering structures, with the 'ward' linkage producing more cohesive clusters. Notably, the 'complete' and 'single' linkage methods resulted in high Silhouette Index

scores, suggesting well-separated clusters, though these may not perfectly align with the actual donation patterns as indicated by the original labels.

Relative Performance of Clustering Algorithms

- K-Means (OOP) and HCA with the 'ward' criterion had a better alignment with the original labels, indicating their suitability for datasets with distinct groupings or where the goal is to identify clear-cut segments.
- The high Silhouette Index scores for 'complete' and 'single' linkage in HCA suggest that these methods are capable of detecting well-separated clusters, but may not align with actual donor behaviors as effectively as K-Means or 'ward' linkage.
- The confusion matrix for K-Means highlighted the presence of false positives and false negatives, revealing limitations in the algorithm's ability to accurately classify all individuals.

Significance for Blood Donation Campaign Strategies

- The analysis is significant for blood donation campaigns as it provides a foundation for developing targeted strategies based on identified donor segments.
- Algorithms that offer clear segmentation, like K-Means, can help campaigns to tailor their messaging and outreach efforts towards specific groups, potentially increasing donor turnout and retention.
- Understanding the characteristics of different donor segments allows for personalized engagement, which is crucial in encouraging repeat donations and optimizing resource allocation.
- The probabilistic nature of GMM could be leveraged to address uncertain or borderline cases, offering nuanced insights into donor behavior that may be masked by more rigid clustering approaches.

Recommendations

- Future strategies for blood donation campaigns should consider not only the clustering results but also the nature of the dataset and the specific campaign goals.
- It is recommended to further explore clustering algorithms that can handle the complexities and nuances of donor behavior, possibly integrating machine learning techniques that can learn from donor engagement over time.

- Additional features, such as demographic data or historical donation frequencies, could be incorporated to enhance the clustering models and provide more granular insights into donor profiles.

In conclusion, the application of unsupervised learning algorithms to the blood transfusion dataset has revealed valuable patterns and segments within donor behaviors, although, there is still room for improvement on identifying properly the high value donors. The relative performance of these algorithms provides actionable insights for blood donation campaigns, informing strategies that could lead to more effective donor segmentation and engagement. The study underscores the potential of data mining in transforming operational approaches within the domain of healthcare and public service initiatives.

A Appendices

A.1 Python Code

The version of Python in which the Jupyter Notebook code was made is 'Python 3.9.13'.

A.2 References

The idea to use the function *pairplot* from the Seaborn library was obtained from the following link: `https://inria.github.io/scikit-learn-mooc/python_scripts/datasets_blood_transfusion.html`, which also manipulates the Blood Transfusion Dataset.