UNIVERSITY OF CAPE TOWN

DEPARTMENT OF STATISTICAL SCIENCES

STA5077Z - Unsupervised Learning Assignment 1

MSHTSA009

September 16, 2024

Contents

1 Cluster Analysis

1.1 Introduction

CChild mortality continues to be a critical global health issue, and efforts to reduce it are encapsulated within the United Nations' Sustainable Development Goals (SDGs). One of the key strategies in reducing child mortality is improving neonatal care, which includes monitoring and assessing foetal health. Cardiotocograms (CTGs) provide a non-invasive, cost-effective method for evaluating foetal well-being during pregnancy. CTGs measure vital indicators such as the foetal heart rate and uterine contractions, which offer real-time insights into potential health complications.

The dataset used in this analysis, titled "fetal_health.csv", contains 2,126 records extracted from CTG examinations. Obstetricians typically classify foetal health into three categories: Normal, Suspect, and Pathological. These categories assist in identifying foetuses at risk and enable timely interventions. Accurate classification of these health statuses is essential for directing medical resources and minimizing preventable complications in newborns.

This report aims to perform a comprehensive cluster analysis to investigate whether the three predefined foetal health categories—Normal, Suspect, and Pathological—are appropriate, based on the features provided in the dataset. The clustering methods will include K-Means, K-Medoids, DBSCAN, and Hierarchical Clustering. To further facilitate the analysis, Principal Component Analysis (PCA) will be applied for dimensionality reduction. The results of each method will be evaluated using silhouette analysis and other clustering validation metrics to determine the quality of the clustering.

1.2 Data Exploration and Preprocessing

Before applying clustering methods, it is essential to thoroughly inspect and preprocess the data to ensure accuracy and reliability in the analysis. The dataset used in this analysis contains 2,126 records, derived from cardiotocogram (CTG) exams, which monitor various fetal health indicators. These include the baseline fetal heart rate, accelerations, uterine contractions, decelerations, and several histogram-based attributes. Below is a detailed description of the data exploration and preprocessing steps undertaken.

1.2.1 Data Inspection

The dataset was initially checked for its structure, dimensions, and the presence of any missing values. It contains 21 columns, each representing a unique feature related to fetal health. A key part of data cleaning is ensuring there are no missing values or duplicated entries. Upon inspection, the dataset revealed the following:

- **No missing values** were present in any of the columns, which eliminates the need for imputation or handling of NaNs.
- **Duplicates:** Upon inspection, a few duplicate entries were identified and subsequently removed. Retaining duplicates can lead to biased clustering, as it may overweight certain patterns that are not

representative of the broader dataset. Therefore, duplicates were excluded to enhance the validity of the clustering results.

- Data Types: All columns were appropriately typed, with numerical features represented as continuous variables, making them suitable for clustering algorithms. No additional type conversion was required.

1.2.2 Descriptive Statistics

To understand the underlying characteristics of the data, descriptive statistics were calculated for each feature, including minimum, median, mean, standard deviation, and maximum values. This step provides insight into the central tendencies and variability within the dataset, highlighting which features may have more influence in clustering. For example:

- The baseline fetal heart rate ranged from **106** to **160** beats per minute (bpm), with a median of **133 bpm**.
- The accelerations showed a mean value of **0.003** with a standard deviation of **0.004**, indicating very low variation in fetal movement accelerations within the dataset.

Feature	**Min**	**Median**	**Mean**	**Std Dev**	**Max**
baseline.value	106	133.000	133.304	9.841	160
accelerations	0	0.002	0.003	0.004	0.019
fetal_movement	0	0.000	0.009	0.047	0.481
uterine_contractions	0	0.004	0.004	0.003	0.015
light_decelerations	0	0.000	0.002	0.003	0.015
severe_decelerations	0	0.000	0.000	0.00006	0.001

Table 1: Descriptive statistics for the fetal health dataset

These statistics help identify which features may be more discriminative during the clustering process. For instance, features with high variability, such as fetal movement, could play a more significant role in distinguishing clusters.

1.2.3 Data Standardization

Since clustering algorithms are sensitive to differences in scale, particularly distance-based methods like K-Means, it is crucial to standardize the data. Features with larger numerical ranges can dominate the clustering process, leading to biased results. To prevent this, all features were standardized using z-score normalization, ensuring each feature contributes equally to the analysis.

- The `scale()` function was applied to normalize each feature by centering the data (mean = 0) and scaling to unit variance (standard deviation = 1).
- The scaled dataset was converted back into a DataFrame for ease of manipulation and visualization in subsequent steps.

By preparing the dataset through cleaning, descriptive analysis, and standardization, we ensure that the features are appropriately scaled for accurate clustering results.

1.3 Exploratory Data Analysis (EDA)

A thorough Exploratory Data Analysis (EDA) was conducted to uncover the distributions of features, relationships between variables, and potential anomalies that could impact clustering performance.

1.3.1 Distribution of Features

To gain insights into the spread and variability of each feature, histograms were plotted for all 21 variables. This helped to detect any skewness or outliers and offered an understanding of the data distributions.

Figure 1: Distribution of Features

Key findings from the distribution analysis:

- **Baseline Value:** Displayed a relatively normal distribution centered around 130-140 bpm, indicating a balanced range of values.
- **Accelerations and Fetal Movement:** Both showed highly skewed distributions, with most values clustered near zero. This suggests that these features have low variation, which could affect their usefulness in clustering.
- **Histogram-based Features:** Features such as histogram mode and variance exhibited broader distributions, indicating more variation and potentially higher discriminative power in cluster formation.

These findings further support the need for standardization and will inform the choice of clustering methods.

1.3.2 Correlation Analysis

To examine the relationships between features, a correlation matrix was calculated and visualized through a heatmap. This matrix helps identify any multicollinearity (high correlation between features) that could influence clustering results.

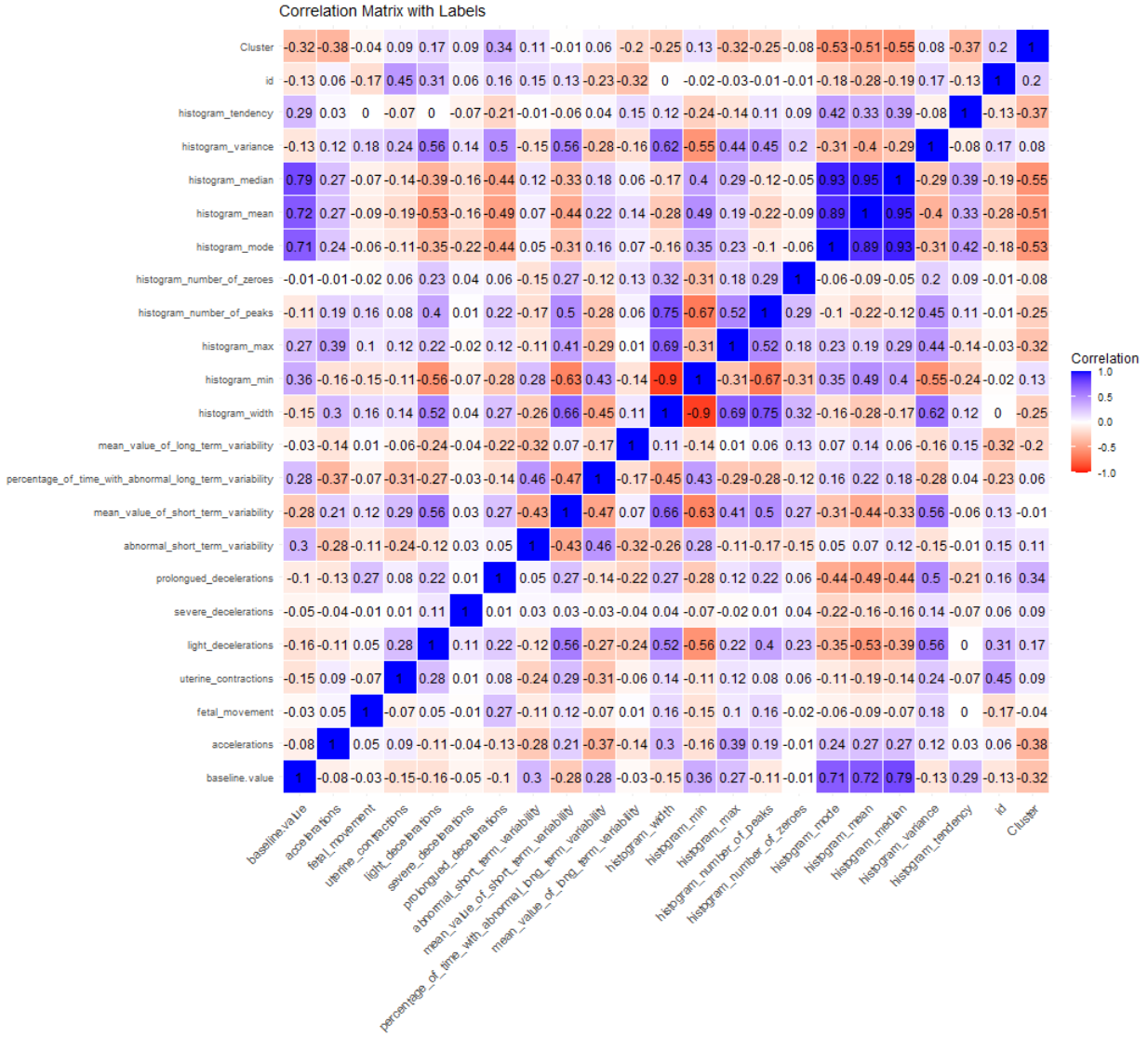


Figure 2: Correlation Heatmap

Key observations:

- Strong correlations were found between **histogram mean**, **histogram median**, and **histogram mode**, indicating these variables capture similar information about the data.
- Weak correlations were seen between features like **Accelerations** and **Uterine Contractions** with other variables, suggesting these may have minimal influence on the clustering process.
- The **percentage of time with abnormal long-term variability** showed very low correlations

with most other variables, suggesting it captures a unique aspect of fetal health that may contribute distinctively to clustering.

This correlation matrix provides a comprehensive view of the interdependencies between variables, helping to anticipate which features might drive the formation of clusters.

1.3.3 Outlier Detection

Box plots were generated to identify potential outliers across all features. Outliers can heavily influence clustering results and may need to be addressed depending on their nature and frequency.

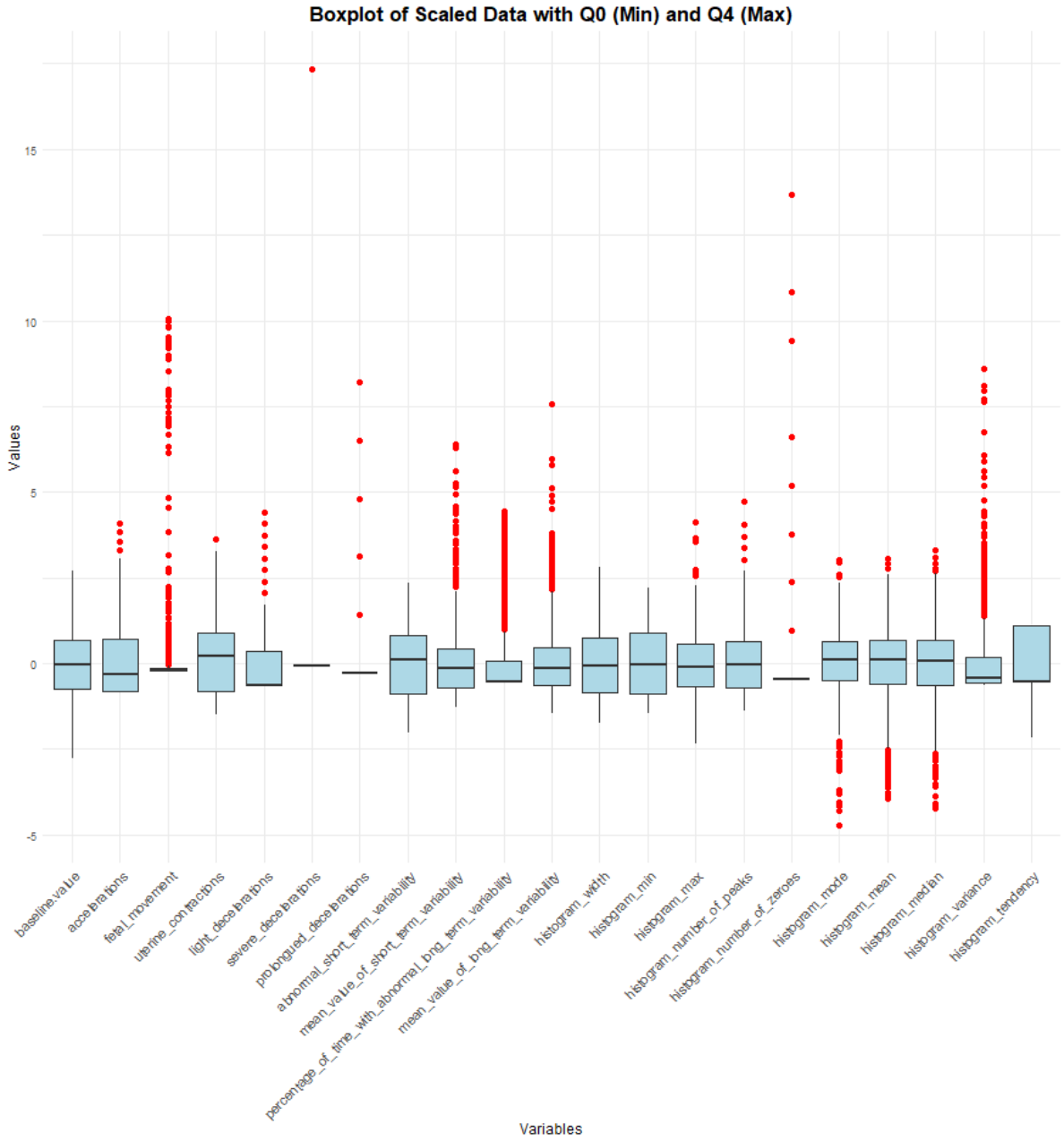


Figure 3: Boxplot for Outliers

Key findings:

- **Significant Outliers:** Features like Severe Decelerations and Percentage of Time with Abnormal Long-Term Variability exhibited considerable outliers. Given the clinical nature of the data, these outliers likely represent important health patterns rather than errors. Therefore, these outliers were retained for clustering, as they may reveal critical distinctions in fetal health.
- **Impact on Clustering:** Retaining outliers ensures that the analysis reflects real-world variations in fetal health, which is particularly relevant in detecting pathological conditions. However, their presence may influence the performance of certain clustering algorithms like K-Means, which are sensitive to such deviations.

1.4 Dimensionality Reduction with Principal Component Analysis (PCA)

To address potential issues of multicollinearity and high dimensionality, Principal Component Analysis (PCA) was applied. The goal was to reduce the dataset's dimensionality while retaining as much variance as possible for subsequent clustering.

1.4.1 Results of PCA

The PCA results indicate that the first few principal components (PCs) account for a significant proportion of the variance in the data. The output shows that the first principal component (PC1) explains **27.77%** of the variance, while the second principal component (PC2) explains **16.12%**. Together, these two components account for **43.90%** of the total variance, making them suitable for clustering analysis. The following table summarizes the results for the top PCs:

Component	Standard Deviation	Proportion of Variance	Cumulative Proportion
PC1	2.4719	0.2777	0.2777
PC2	1.8832	0.1612	0.4390
PC3	1.3731	0.0525	0.5246

Table 2: Summary of Principal Components

Given that nearly 44% of the variance is explained by the first two principal components, these were selected for further analysis using clustering methods such as K-Means, K-Medoids, DBSCAN, and Hierarchical Clustering.

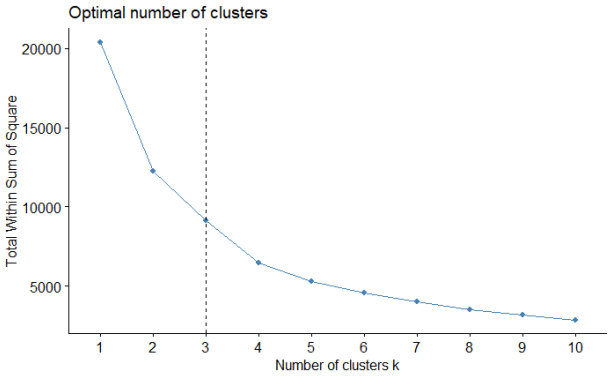
1.5 K-Means Clustering Analysis

In this section, we explore the application of K-Means clustering to the fetal health dataset, with the goal of determining if the data naturally forms distinct groupings that may correspond to known fetal health categories. To ensure robust results, both the *Elbow Method* and *Silhouette Analysis* were used to select the optimal number of clusters. The clustering performance is then evaluated and visualized to validate the results.

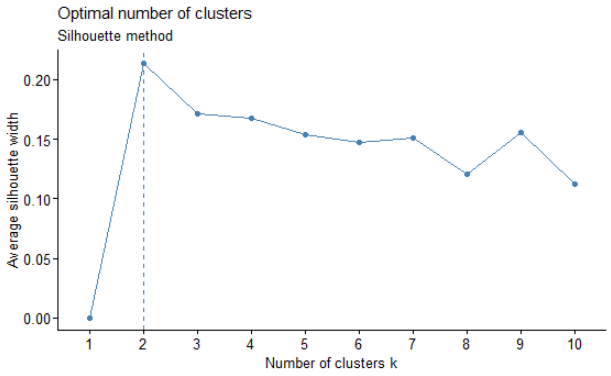
1.5.1 Determining the Optimal Number of Clusters

Selecting the correct number of clusters (k) is crucial for meaningful results. We employed two widely recognized methods to assist in this decision-making process: the *Elbow Method* and *Silhouette Analysis*.

- Elbow Method
  - The Elbow Method evaluates the *Total Within-Cluster Sum of Squares (WSS)* for different values of k. The plot of WSS against the number of clusters typically shows a clear "elbow" point, after which the marginal gain in reducing WSS diminishes. This point signifies the ideal number of clusters, as adding more clusters beyond this point yields diminishing returns in model performance.
- Silhouette Analysis
  - Silhouette Analysis provides an additional validation of the optimal number of clusters by measuring how well each data point fits within its assigned cluster. The *average silhouette width* is computed, with values closer to 1 indicating better-defined clusters.



(a) Elbow Method showing optimal number of clusters at k = 3. Suggesting that three clusters provide the best balance between simplicity and accuracy.



(b) Silhouette Analysis showing optimal number of clusters at k = 2, ollowed by a slightly lower value for k = 3. Despite this, k = 3 was selected to align with the findings from the Elbow Method

Figure 4: Combined figures showing the Elbow Method and Silhouette Analysis

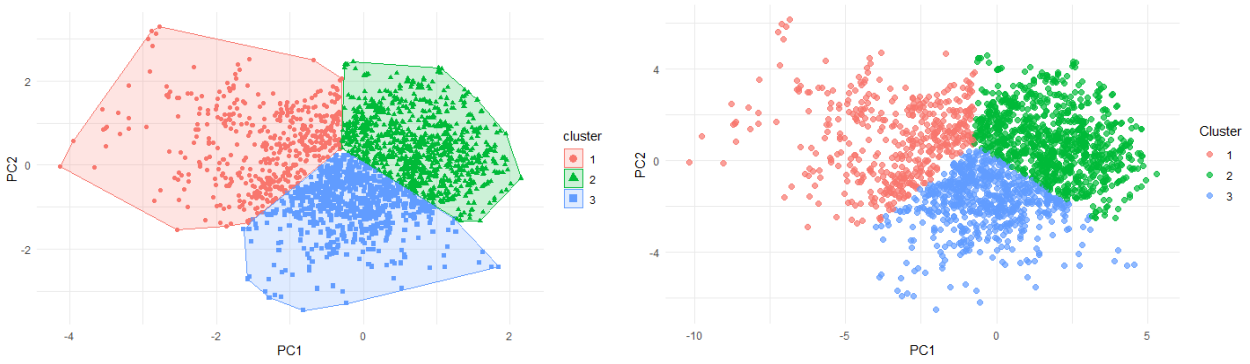
1.5.2 Results

The final clustering was performed using $k = 3$, and the clusters were visualized based on the first two principal components (PC1 and PC2). The clustering results reveal three distinct groups in the data, corresponding to the clusters identified by K-Means. These clusters are well-separated in the two-dimensional PCA space, with some overlap between cluster 3 and the other two clusters, as indicated by the silhouette analysis.

- **Summary of Cluster Sizes**

Cluster	Size
1	698
2	1029
3	384

Table 3: Cluster Sizes



(a) (b)

Figure 5: Comparison of K-Means clustering results (k = 3) using the first two principal components, illustrating cluster separation and overlap. (a) K-Means clustering results (k = 3) based on the first two principal components. (b)Clusters formed by K-Means (k = 3) using the first two principal components.

1.5.3 Silhouette Analysis

The silhouette plot offers a graphical representation of how well-separated the clusters are. The silhouette width, which ranges from -1 to 1, measures how similar an observation is to its own cluster (cohesion) compared to other clusters (separation). A silhouette width closer to 1 indicates well-clustered points, while values near 0 imply points are on the boundary between clusters. Negative values suggest that points are likely misclassified.

- Cluster 1: The silhouette width for cluster 1 is relatively low at 0.18, indicating that the points within this cluster are not well separated from other clusters. This suggests some degree of overlap with the other clusters, likely with cluster 3, as observed in the PCA scatter plot. The narrow width reflects a mix of well-clustered points and others that may be on the border between clusters.

- Cluster 2: This cluster has the highest silhouette width at 0.37, indicating a good separation from the other clusters. The points in this cluster are more compact, meaning they are more homogenous and distinct from the other clusters. This suggests that cluster 2 is the most well-formed and clearly defined cluster in the analysis.

- Cluster 3: The silhouette width for cluster 3 is 0.09, which is quite low. This indicates significant overlap with the other clusters, particularly cluster 1. The low silhouette score suggests that many points in cluster 3 may be close to or potentially misclassified into other clusters. This can be further investigated by analyzing the cluster centroids and the nature of the points within this cluster.

- The average silhouette width across all three clusters is 0.25, which indicates that the clustering struc-

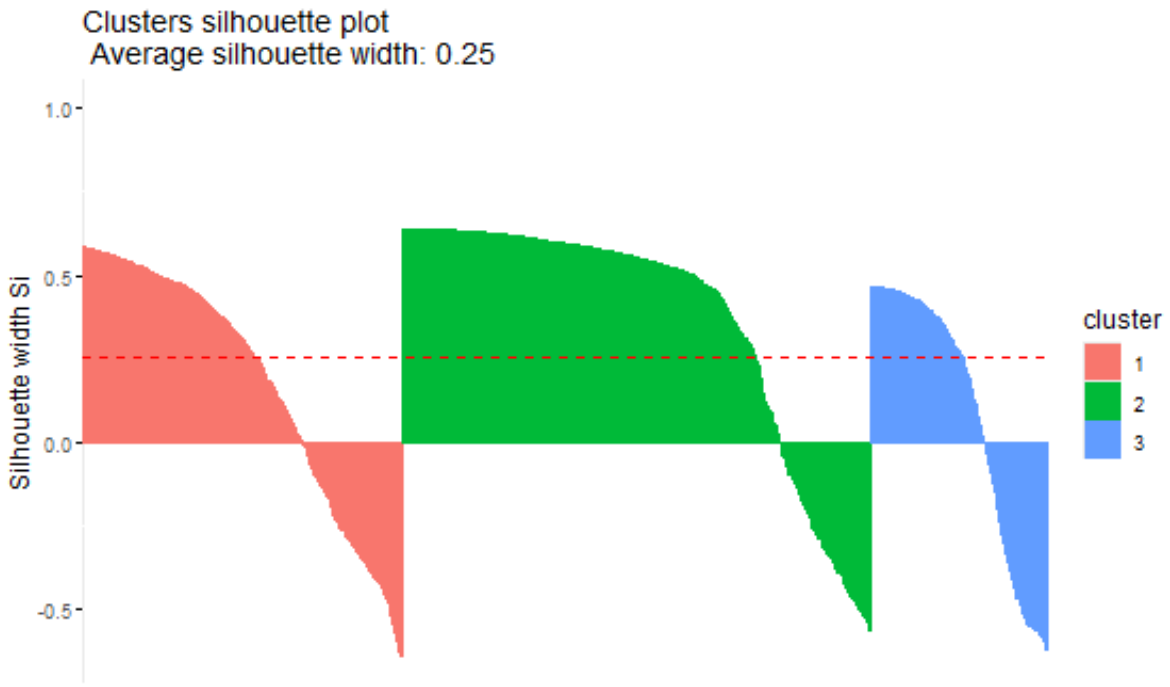


Figure 6: Silhouette plot for K-Means clustering with k=3, showing the silhouette widths for each cluster.

ture is somewhat weak. While cluster 2 demonstrates good cohesion and separation from the others, clusters 1 and 3 show significant overlap. This is consistent with the results from the PCA visualization, where clusters 1 and 3 appear closer to one another.

The table below presents a summary of the K-Means clustering results, showing the size of each cluster and the corresponding average silhouette width. This offers insights into how well the points within each cluster are grouped and how distinct each cluster is from others.

Cluster	**Size**	**Average Silhouette Width**
1	698	0.18
2	1029	0.37
3	384	0.09

Table 4: Cluster Sizes and Average Silhouette Widths for K-Means.

The K-Means clustering with k=3 gave us useful insights into the structure of the fetal health data. Cluster 2 was well-defined, with a silhouette width of 0.37, meaning the algorithm successfully identified a distinct group. However, clusters 1 and 3 were less clear, with lower silhouette widths of 0.18 and 0.09, showing some overlap between them.

While K-Means did a good job capturing the general structure, the separation between the clusters, particularly between clusters 1 and 3, could be improved. To get better results, further steps like feature engineering or trying different clustering methods might help. Overall, K-Means provided a solid foundation for analyzing

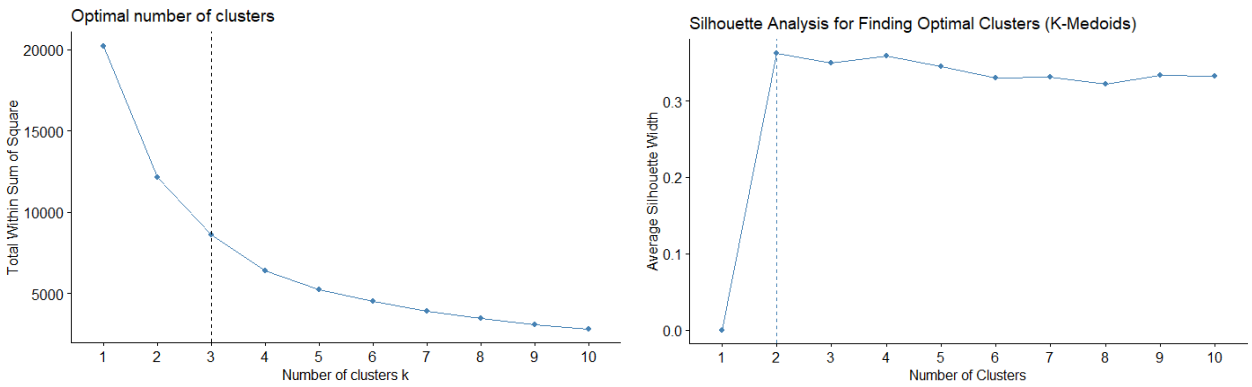
fetal health categories,

1.6 K-Medoids Clustering Analysis

After applying K-Means, K-Medoids clustering was performed to assess if a more robust clustering method could provide better-separated and more distinct clusters. K-Medoids, being less sensitive to outliers compared to K-Means, is often preferred for handling datasets with noise or anomalies.

1.6.1 Determining the Optimal Number of Clusters

- Elbow Method
  - The elbow method was again used to determine the optimal number of clusters for K-Medoids. The plot of the total within-cluster sum of squares (WSS) against the number of clusters shows an "elbow" at $k = 3$, indicating that three clusters are the most appropriate for this dataset, consistent with the K-Means findings.
- Silhouette Method
  - Silhouette analysis was also used to assess the quality of the clustering and to further validate the choice of k=3.

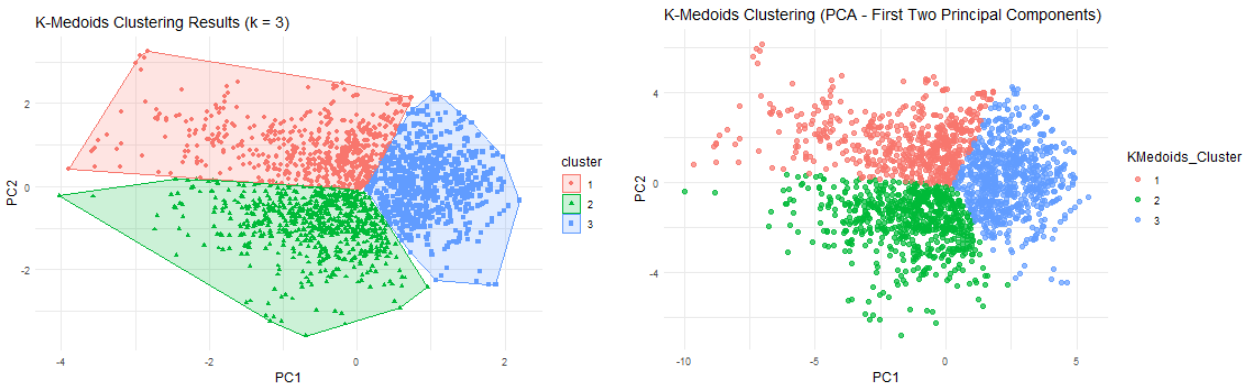


(a) (b)

Figure 7: (a) Elbow Method showing optimal number of clusters at k = 3. (b) Silhouette Method showing optimal number of clusters at k = 3.

1.6.2 Results

The results we get from K-Medoids clustering with k=3 offer insights into how the dataset is grouped based on the first two principal components. As shown in Figure 8, Cluster 2 demonstrates clear separation and cohesion, while Clusters 1 and 3 exhibit overlap and wider dispersion. These visualizations help illustrate the performance of K-Medoids in identifying distinct patterns within the data, though some ambiguity remains between clusters.



(a) (b)

Figure 8: (a)K-Medoids clustering results (k = 3) based on the first two principal components. (b)Clusters formed by K-Medoids (k = 3) using the first two principal components.

1.6.3 Silhouette Analysis

The average silhouette width for the K-Medoids clustering was **0.24**, slightly lower than K-Means (**0.25**). This suggests that K-Medoids clustering, while somewhat effective, did not significantly outperform K-Means in terms of overall cluster separation. See in Figure 9 and Table 5.

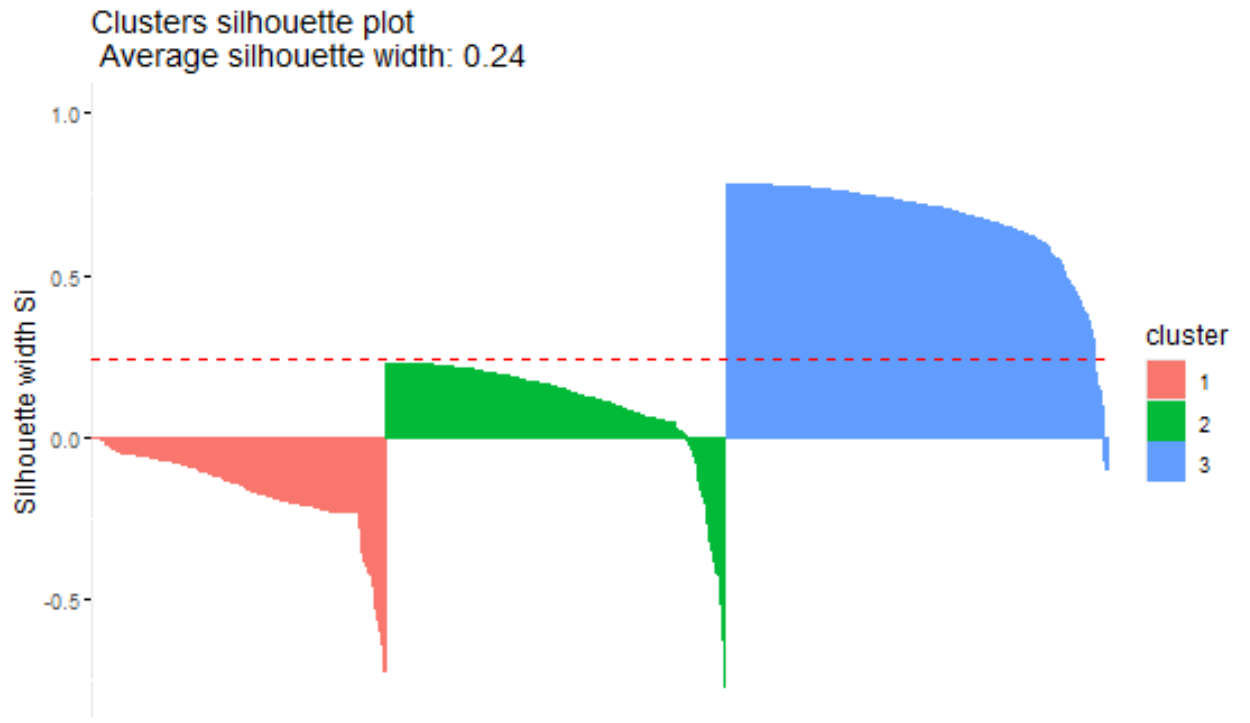


Figure 9: Silhouette plot for K-Medoids clustering with k=3, showing the silhouette widths for each cluster.

Cluster	Size	Average Silhouette Width
1	610	-0.16
2	707	0.11
3	794	0.67

Table 5: Cluster sizes and silhouette widths for K-Medoids clustering.

- **Cluster 1:**
  - **Average Silhouette Width**: **-0.16**, which indicates poor cohesion and separation. Many points within this cluster are likely misclassified, with significant overlap with other clusters. A negative silhouette score suggests that these points may actually belong in other clusters.
- **Cluster 2:**
  - **Average Silhouette Width**: **0.11**, which is low, showing weak separation from other clusters. Much like in K-Means, this cluster struggles with clear boundaries and contains points near the edges of other clusters.
- **Cluster 3:**
  - **Average Silhouette Width**: **0.67**, which indicates strong cohesion and good separation from the other clusters. This cluster is the most well-defined and distinct, similar to the K-Means results where one cluster exhibited much better performance than the others.

In contrast to K-Means, K-Medoids is generally more robust to outliers, so it was expected to yield better clustering results. However, in this case, the performance was not significantly better overall, and the negative silhouette width in cluster 1 suggests that this algorithm struggled to separate some of the more complex patterns in the dataset.

1.7 Hierarchical Clustering Analysis

Building on the earlier analysis using K-Means and K-Medoids, hierarchical clustering was applied to further evaluate the clustering structure of the fetal health dataset. Unlike K-Means and K-Medoids, hierarchical clustering does not require a pre-specified number of clusters and offers a visual representation of how the data points are grouped. The dendrograms generated using three different linkage methods — complete, average, and single linkage — provide additional insights into the dataset's structure and cluster formation.

1.7.1 Complete Linkage Dendrogram

Complete linkage focuses on minimizing the maximum distance between points in different clusters, which tends to create more compact and well-separated clusters. The dendrogram generated through this method reveals three distinct clusters, with notable separation between them, particularly at the higher levels.
In comparison to K-Means and K-Medoids, complete linkage also achieves strong separation, similar to what we observed with the clear boundary of Cluster 2 in both K-Means and K-Medoids. However, the overlap between Clusters 1 and 3 that we saw in the previous analyses is still present, although reduced.

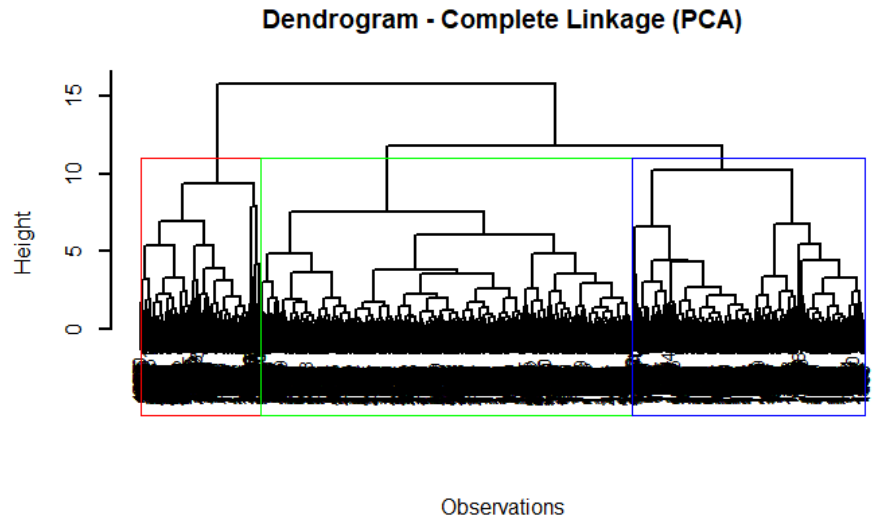


Figure 10: Dendrogram using complete linkage based on PCA, showing distinct clusters with reduced overlap, particularly between Clusters 1 and 3.

1.7.2 Average Linkage Dendrogram

Average linkage takes a more balanced approach by minimizing the average distance between all pairs of points in different clusters. The resulting dendrogram shows moderate separation of clusters, but not as compact as the complete linkage method. There's some overlap, especially between Clusters 1 and 3, mirroring the challenge seen in both K-Means and K-Medoids where these two clusters had difficulty maintaining separation. Interestingly, while Cluster 2 was clearly separated in K-Means and K-Medoids, the average linkage method reveals a bit more spread in Cluster 2, indicating that it may not be as tightly grouped as initially suggested by the centroid-based methods.

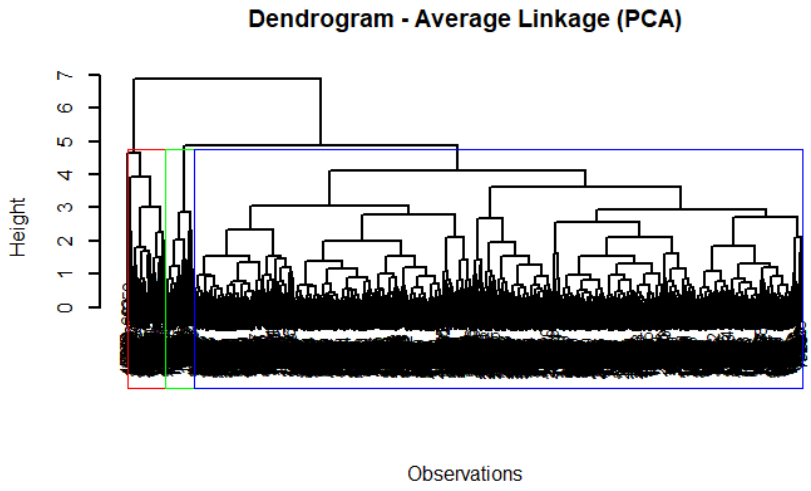


Figure 11: Dendrogram using average linkage based on PCA, illustrating moderate cluster separation but with some overlap, particularly between Clusters 1 and 3.

1.7.3 Single Linkage Dendrogram

Single linkage minimizes the distance between the closest points in different clusters, often resulting in elongated, chain-like clusters. The single linkage dendrogram confirms this, as the clusters appear much more elongated and less distinct. This is most evident with Cluster 1, which dominates the lower section of the dendrogram. This result mirrors the poorer performance seen in Cluster 1 during K-Medoids clustering, where the silhouette score was negative, indicating misclassification. Single linkage exacerbates this issue, with Cluster 1 extending broadly and creating a less meaningful separation.

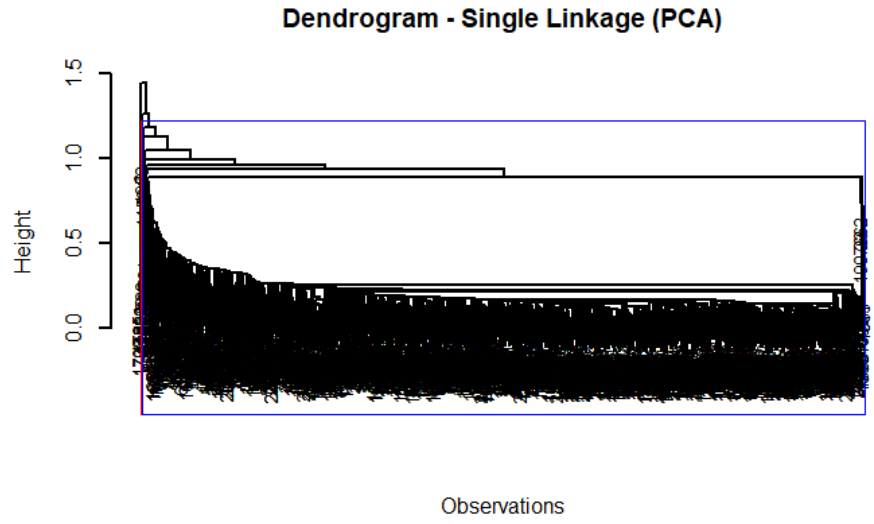


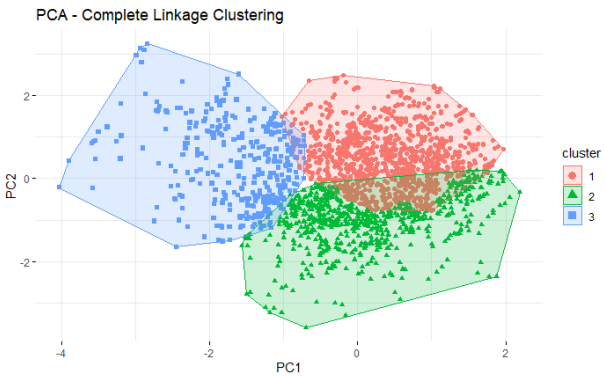
Figure 12: Dendrogram using single linkage based on PCA, showing elongated clusters with significant overlap and chaining, particularly within Cluster 1.

1.7.4 Cluster Visualization Comparison

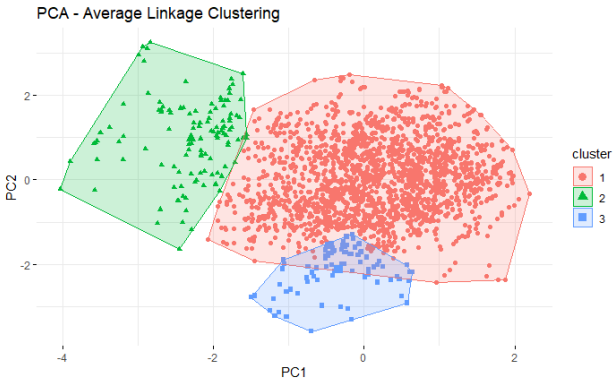
To further understand the hierarchical clustering results, we visualized the clusters based on the PCA components for each linkage method.

- Complete Linkage Clustering
  - The complete linkage clustering, projected onto the first two principal components, **Figure 13 (a)** shows well-separated clusters with Cluster 3 (blue) remaining distinct, and Cluster 2 (green) retaining much of its compactness. However, Cluster 1 (red) continues to overlap with the other clusters, though the overlap is less pronounced than in the K-Medoids or single linkage methods.
- Average Linkage Clustering
  - The average linkage clustering is more balanced but still shows noticeable overlap between Cluster 1 (red) and Cluster 3 (blue) on **Figure 13 (b)**. Cluster 2 (green), while still separate, is more spread out compared to the results from K-Means and K-Medoids. This suggests that average linkage captures broader patterns within the dataset but may lack the tighter grouping seen with centroid-based methods.

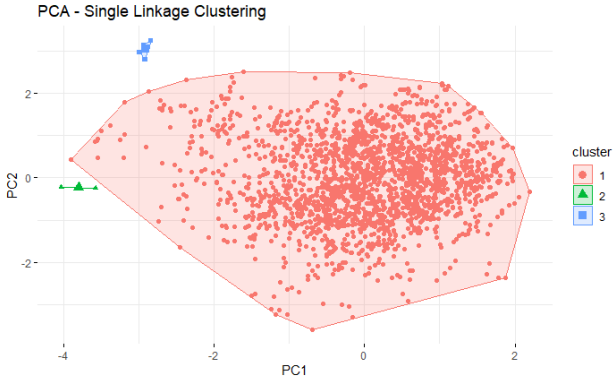
- Single Linkage Clustering
  - The single linkage clustering projection highlights the chaining effect observed earlier. Cluster 1 (red) dominates, spreading widely, while Cluster 2 (green) and Cluster 3 (blue) appear as smaller, more isolated groups. This visualization clearly shows the weakness of single linkage in this context, where clusters are not well-defined, and separation is poor, shown in **Figure 13 (c)**



(a) PCA-based clustering using complete linkage, showing compact and distinct clusters with some overlap, particularly between Clusters 1 and 3.



(b) PCA-based clustering using average linkage, displaying balanced clusters but with visible overlap between Clusters 1 and 3, and a more spread-out Cluster 2.



(c) PCA-based clustering using single linkage, where Cluster 1 dominates and overlaps significantly with the other clusters, showing weak separation and highlighting the chaining effect of single linkage.

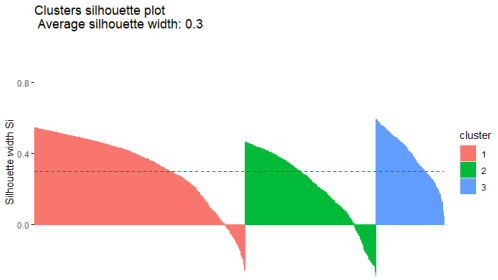
Figure 13: Comparison of clustering results using different linkage methods.

1.7.5 Silhouette Analysis

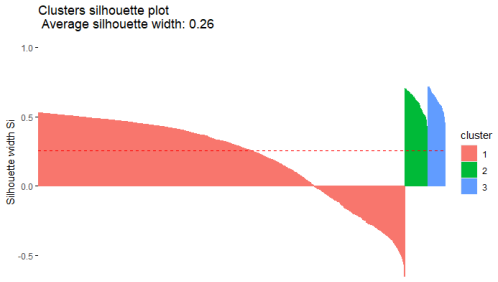
We now evaluate the clustering performance using silhouette scores for each linkage method. This provides insight into the cohesion and separation of clusters, offering a comparison to the previous K-Means and K-Medoids results.The table below and Figure 20 present the clustering results for complete, average, and single linkage methods, along with their respective silhouette scores.

Clustering Method	Cluster	Cluster Size	Silhouette Width
Complete Linkage	Cluster 1	1084	0.32
	Cluster 2	677	0.21
	Cluster 3	350	0.39
	Average		**0.30**
Average Linkage	Cluster 1	1901	0.22
	Cluster 2	121	0.59
	Cluster 3	89	0.62
	Average		**0.26**
Single Linkage	Cluster 1	2104	0.53
	Cluster 2	2	0.83
	Cluster 3	5	0.93
	Average		**0.53**

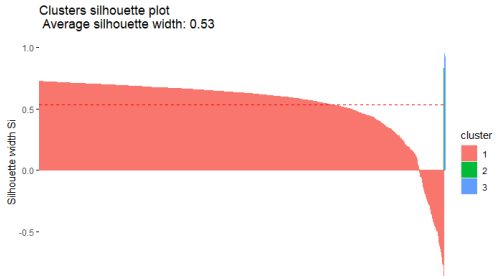
Table 6: Clustering results for different linkage methods.



(a) The silhouette plot for complete linkage.



(b) The silhouette plot for average linkage.



(c) The silhouette plot for single linkage.

Figure 14: Silhouette plots for different linkage methods.

- The silhouette plot for complete linkage shows an average silhouette width of 0.30, which indicates moderate clustering quality. Complete linkage shows reasonable clustering performance, with Cluster 3 being the most distinct and Cluster 2 having the weakest separation.
- The silhouette plot for average linkage shows an average silhouette width of 0.26, slightly lower than that of complete linkage, indicating a bit more overlap among clusters.Average linkage produced good clustering results for Clusters 2 and 3, but Cluster 1, which is the largest, suffers from poor separation and significant overlap.
- The silhouette plot for single linkage shows an average silhouette width of 0.53, which is higher than

both complete and average linkage.The high silhouette scores for Clusters 2 and 3 are misleading because of their very small sizes. This clustering result suggests that single linkage struggled to find meaningful groups, leading to a large dominant cluster and a few very small, isolated clusters.

1.8 Conclusion

In this assignment, we applied three clustering methods — **K-Means**, **K-Medoids**, and **Hierarchical Clustering** — to analyze fetal health data and evaluate the appropriateness of the three foetal health classes: *Normal*, *Suspect*, and *Pathological*.

1.8.1 Clustering Performance

K-Means performed moderately well, particularly for **Cluster 2**, but showed overlap between **Clusters 1 and 3**, which affected the overall performance. **K-Medoids** showed better handling of outliers and noise, offering slightly improved results, but still experienced overlap between **Clusters 1 and 3**. **Hierarchical Clustering**, using **complete linkage**, provided the most balanced results, with **Cluster 3** being well-separated and **Cluster 1** showing moderate separation. **Single linkage** produced poor results due to the chaining effect, and **average linkage** struggled with separating larger clusters.

1.8.2 Best Fit for the Assignment

K-Medoids emerges as the most suitable method, striking a good balance between handling outliers and producing cohesive clusters, despite some overlap between **Clusters 1 and 3**. It performed more consistently across clusters compared to hierarchical clustering, making it the best fit for identifying distinct groups in the fetal health dataset.

1.8.3 Appropriateness of Three Foetal Health Classes

The clustering results consistently revealed the presence of three distinct clusters, but there was noticeable overlap between **Clusters 1 and 3**, particularly in **K-Means** and **K-Medoids**. **Cluster 2** (representing the *Suspect* class) appeared well-separated and easily identifiable. However, **Cluster 1** (likely *Normal*) and **Cluster 3** (likely *Pathological*) showed some overlap, suggesting that the separation between normal and pathological cases is not entirely clear-cut.

While the three foetal health classes are broadly appropriate, the overlap between *Normal* and *Pathological* suggests that additional features or diagnostic criteria may be needed to improve separation. The classes are generally suitable but could benefit from refinement to better distinguish between these overlapping groups.

2 Association Rule Mining for Coronary Artery Disease (CAD)

2.1 Introduction

Coronary artery disease (CAD) is a major health concern worldwide, responsible for numerous deaths due to its progressive nature, causing narrowing of the coronary arteries. Timely and accurate diagnosis is critical in reducing its impact. Data mining techniques, especially association rule mining, have emerged as a promising approach to identifying critical patterns and features associated with CAD. The dataset used for this analysis consists of angiographic measurements of 303 patients, including various factors such as age, blood pressure (BP), and body mass index (BMI), with a column indicating whether the patient has CAD (Cath). The aim of this report is to determine the features that are most associated with CAD using association rule mining, specifically the Apriori and FP-Growth algorithms.

The primary objective of this analysis is to use **association rule mining methods** to determine the features that are mostly associated with coronary artery disease (CAD), utilizing the dataset 'CAD dataset.xls'. This dataset contains various medical and demographic information obtained via angiography from 303 patients. The goal is to uncover associations between these features and the presence of CAD, as indicated by the target column "Cath."

2.2 Data Exploration and Preprocessing

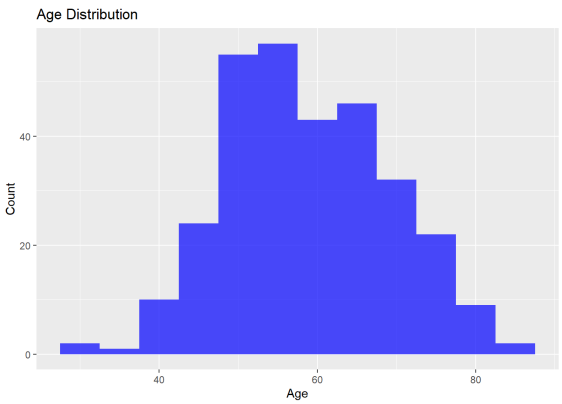
- **Data Overview**: The dataset was loaded, and summary statistics were calculated for each variable. Key variables include:
  - **Age**: ranging from 30 to 86 years.
  - **BMI**: ranging from 18.12 to 40.90, with a mean of 27.25.
  - **Blood Pressure (BP)**: ranging from 90 to 190, with a mean of 129.6.
  - **Other medical conditions** such as hypertension, smoking status, and diabetes mellitus were also summarized.
- **Target Variable (Cath)**: The target column, **Cath**, was summarized to identify the distribution of CAD cases. The two categories include:
  - **Cad (216 cases)**: indicating the presence of CAD.
  - **Normal (87 cases)**: indicating no CAD.
- **Data Preprocessing**: Several continuous variables such as **Age**, **BMI**, and **Blood Pressure** were binned into categorical variables to facilitate the mining of association rules. For example:
  - **Age** was categorized as *Young*, *Middle-aged*, *Senior*, and *Elderly*.
  - **BMI** was categorized as *Underweight*, *Normal*, *Overweight*, and *Obese*.
  - **Blood Pressure (BP)** was categorized as *Normal*, *Pre-hypertension*, *Stage 1 Hypertension*, and *Stage 2 Hypertension*.

  The data was then converted into a transaction format required for association rule mining, with 303 transactions and 58 items.

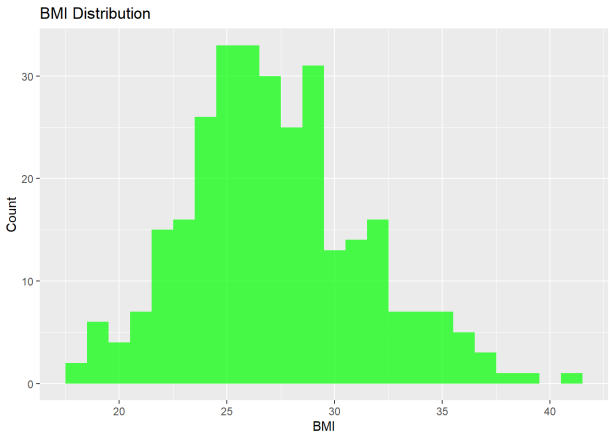
2.3 Exploratory Data Analysis (EDA)

Before conducting association rule mining, an exploratory data analysis was performed to understand the distribution of the variables and the relationships between them.

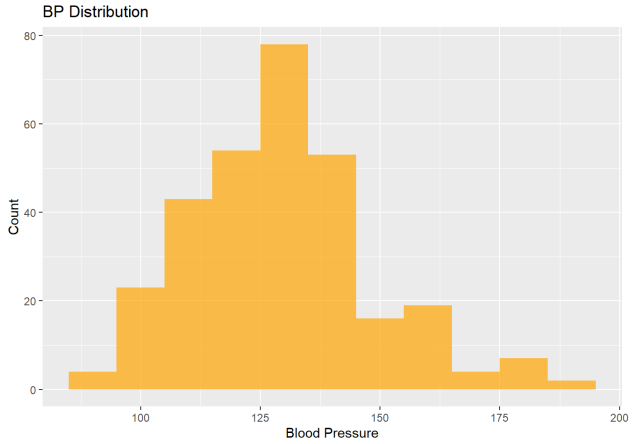
2.3.1 Distribution of Variables



(a) Age Distribution. Most patients fall within the *50-70 years* range, with a noticeable skew towards older age groups.A histogram revealed that younger individuals were underrepresented in this sample.



(b) BMI Distribution. The majority of patients have a *BMI* between *25 and 30*, indicating that many patients are either *overweight* or *obese*.



(c) BP Distribution. Most patients have *normal to slightly elevated blood pressure* (120-150 mmHg)

Figure 15: Distribution plots for Age, BMI, and BP.

2.3.2 Correlation Analysis

A correlation heatmap was generated to explore relationships between continuous variables. Key findings include:

- A strong correlation between *Age* and *Blood Pressure (BP)*.
- Moderate correlations between *BMI* and various health conditions like hypertension.

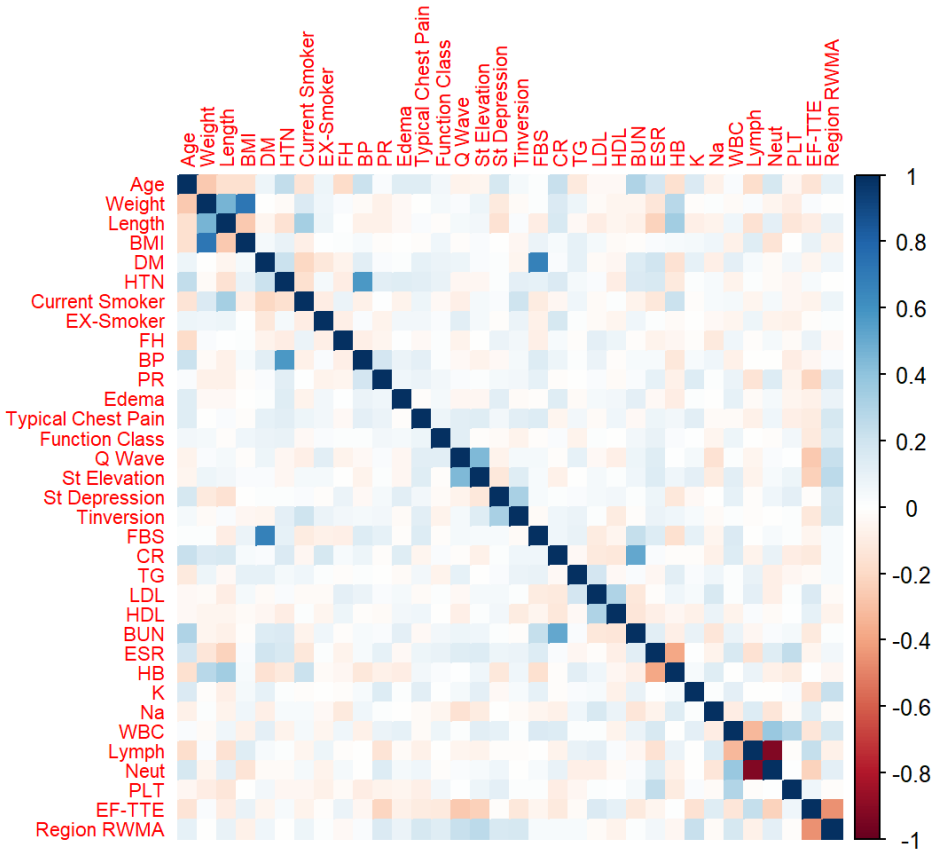
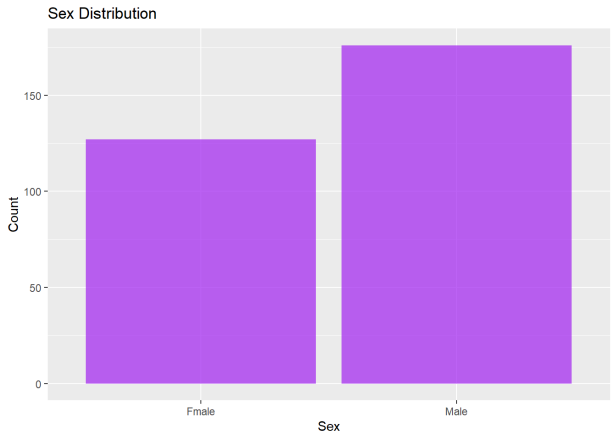
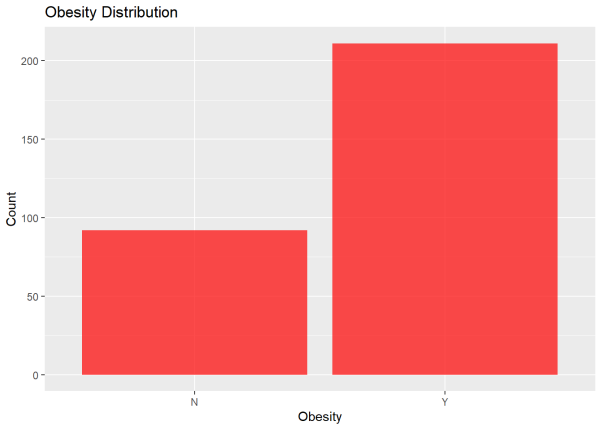


Figure 16: Correlation Heatmap

2.3.3 Categorical Variables



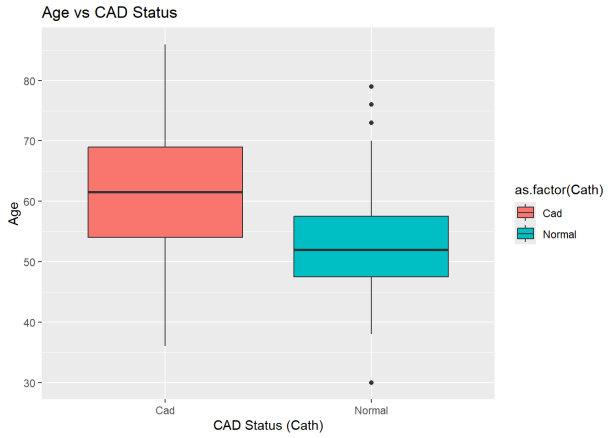
(a) Sex Distribution. The dataset is skewed towards males, who formed the majority of the patient group.



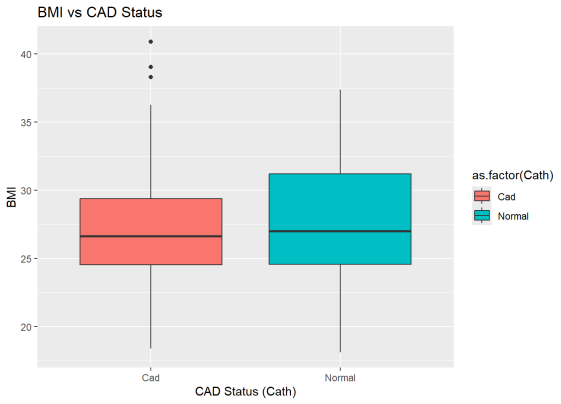
(b) Obesity. A significant number of patients are categorized as obese, which is an important risk factor for CAD.

Figure 17: Demographic and health-related distributions.

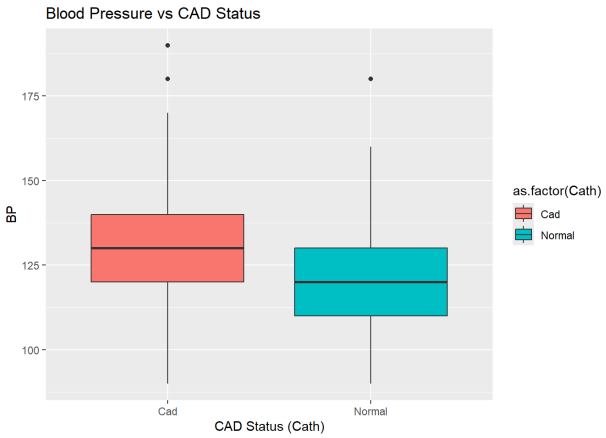
2.3.4 Boxplots for Continuous Variables by CAD Status



(a) Age: The dataset is skewed towards males, who formed the majority of the patient group.



(b) BMI. A significant number of patients are categorized as obese, which is an important risk factor for CAD.



(c) Blood Pressure. A significant number of patients are categorized as obese, which is an important risk factor for CAD.

Figure 18: Boxplots for Continuous Variables by CAD Status

2.4 Association Rule Mining Using Apriori Algorithm

The Apriori algorithm is a popular data mining technique used to discover interesting patterns or associations between items in a dataset. In this case, it was applied to the CAD dataset to identify associations between patient attributes (such as age, sex, blood pressure, and BMI) and the presence or absence of coronary artery disease (CAD). The goal was to find the most frequent and significant patterns that help in predicting CAD diagnosis based on the relationships between features.

2.4.1 Aprori Algorithm Implementation

- Support Threshold: 0.03
  - Chosen to capture itemsets that appear in at least 3% of the transactions. Balances between

identifying frequent patterns and avoiding overly rare associations, ensuring that important yet infrequent medical associations are not missed in the relatively small dataset (303 records).

- Confidence Threshold: 0.8
  - A rule must be correct at least 80% of the time to be considered. Ensures reliability in rule prediction while allowing for flexibility to discover meaningful associations that are not overly strict.
- Rule Length: Minlen = 2, Maxlen = 10
  - Minimum of 2 items per rule and a maximum of 10 items. Captures multi-item interactions (e.g., multiple risk factors), which are common in medical conditions like CAD, while preventing overly complex rules that may be hard to interpret.

Key Findings: After mining the dataset for rules, redundant rules were removed, and the top 10 rules were extracted based on *lift*. The top rules highlight the significant associations between specific features and the absence or presence of CAD (Cath). The following table summarizes the top 10 association rules generated using the Apriori algorithm, sorted by *lift* (a measure of the strength of association compared to random chance).

LHS (Rule Antecedent)	**RHS (Rule Consequent)**	**Support**	**Confidence**	**Coverage**	**Lift**	**Count**
{Age=40-60, Sex=Female, DLP=Y, Atypical=Y}	{Cath=Normal}	0.0330	1.00	0.0330	3.48	10
{Age=40-60, Sex=Female, BP=Normal, Atypical=Y}	{Cath=Normal}	0.0429	1.00	0.0429	3.48	13
{Sex=Female, BP=Normal, Dyspnea=N, Atypical=Y}	{Cath=Normal}	0.0363	1.00	0.0363	3.48	11
{Sex=Female, DLP=Y, Dyspnea=N, Atypical=Y, LVH=N}	{Cath=Normal}	0.0363	1.00	0.0363	3.48	11
{Sex=Female, BP=Normal, Atypical=Y, Poor R Progression=N, VHD=N}	{Cath=Normal}	0.0363	1.00	0.0363	3.48	11
{Age=40-60, BMI=Overweight, Lung rales=N, Atypical=Y, VHD=mild}	{Cath=Normal}	0.0330	1.00	0.0330	3.48	10
{Age=40-60, DLP=N, Lung rales=N, Dyspnea=Y, Atypical=Y, LowTH Ang=N}	{Cath=Normal}	0.0363	1.00	0.0363	3.48	11
{Age=40-60, Obesity=Y, Systolic Murmur=N, Dyspnea=Y, Atypical=Y, LowTH Ang=N}	{Cath=Normal}	0.0363	1.00	0.0363	3.48	11
{Age=40-60, Obesity=Y, Lung rales=N, Dyspnea=Y, Atypical=Y, LowTH Ang=N}	{Cath=Normal}	0.0429	1.00	0.0429	3.48	13
{Sex=Female, BP=Normal, Atypical=Y, Poor R Progression=N}	{Cath=Normal}	0.0561	0.94	0.0594	3.29	17

Table 7: Top 10 Association Rules with Support, Confidence, Coverage, Lift, and Count

2.4.2 Discussion of the Top 10 Rules

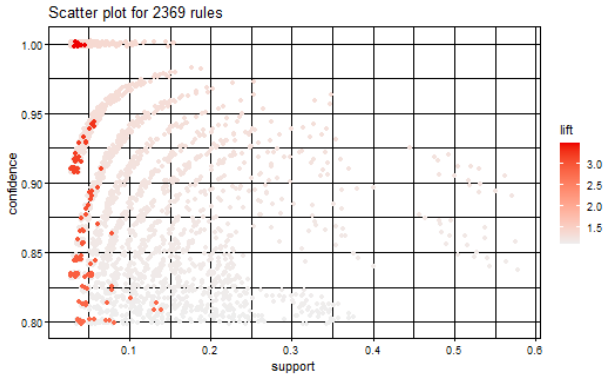
The top 10 rules generated by the Apriori algorithm were identified based on their **support**, **confidence**, and **lift**. These rules effectively meet the objectives of the assignment by identifying key patterns that are

strongly associated with the absence or presence of coronary artery disease (CAD).

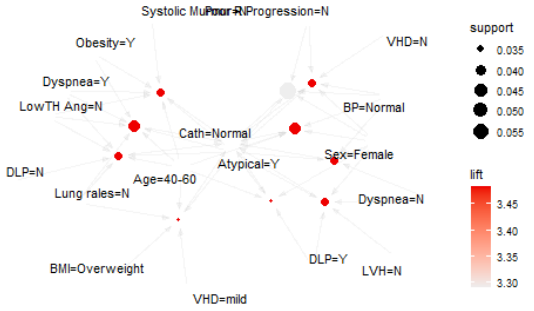
- **High Confidence**
  - All of the top rules have a **confidence of 1.0** (100%), except for two rules with a confidence of 0.94 (94%). This means that in the majority of cases, these combinations of attributes reliably predict whether a patient has or does not have CAD (specifically, *"Cath = Normal"*).
- **Key Predictive Factors**
  - **Age 40-60**: The age range of 40-60 appears frequently, indicating that this is a critical age group for CAD diagnosis. It shows up in 6 out of the top 10 rules, highlighting that middle-aged individuals have important attributes that determine their CAD risk.
  - **Sex = Female**: Female patients are associated with a higher likelihood of being CAD-negative, particularly when combined with normal blood pressure and atypical chest pain.
  - **BP = Normal**: Patients with normal blood pressure are consistently classified as CAD-negative, suggesting that normal BP is a protective factor against CAD.
  - **Atypical Chest Pain**: The presence of atypical chest pain (as opposed to typical) frequently appears in rules predicting CAD-negative outcomes, indicating that not all chest pain is equally indicative of CAD.
- **Lift Values**
  - All rules have **lift values above 3.0**, with the highest being 3.48. This demonstrates that the patterns found significantly increase the likelihood of the consequent (i.e., CAD-negative) compared to random chance. High lift values suggest these rules are meaningful in identifying important associations in the data.
- **Obesity and CAD**
  - While obesity (Obesity = Y) and overweight BMI (BMI = Overweight) show up in the top rules, they are still associated with CAD-negative outcomes in these specific cases, especially when combined with normal BP or mild ventricular hypertrophy (VHD).

2.4.3 Visualization of Apriori Rules

The graphs generated from the Apriori algorithm provide additional insight into the associations found and help to visually interpret the top rules. Below is a concise discussion of the graphs used to represent the association rules.

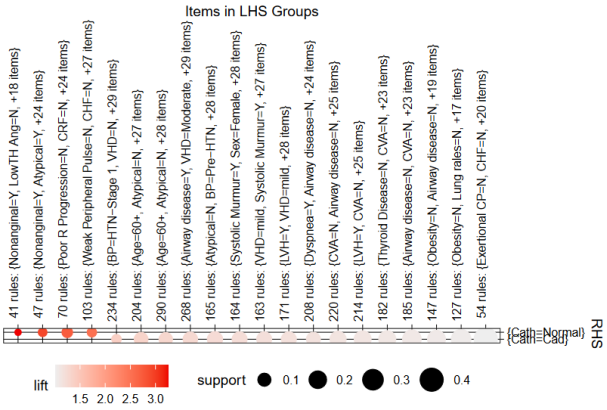


(a) Scatterplot of Rules by Support, Confidence, and Lift.The scatterplot visualizes the relationship between the support, confidence, and lift of all the rules generated by the Apriori algorithm. Each point in the plot represents a rule, and the size of the points is proportional to the lift.

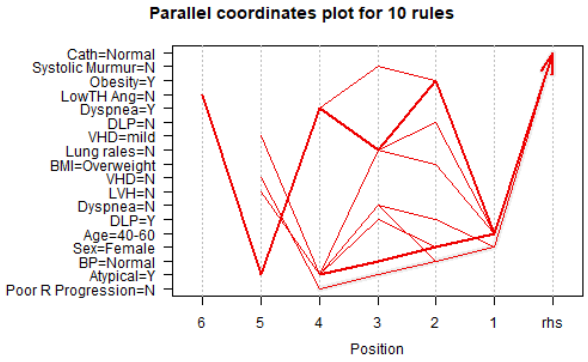


(b) Graph Visualization of Top 10 Rules. This graph visualizes the top 10 rules, showing how the left-hand side (antecedents) and right-hand side (consequent) are connected. Each node represents an item (e.g., "Age=40-60," "Cath=Normal"), and edges represent the association rules.

Figure 19: Scatterplot and Graph Visualization of Association Rules.



(a) Grouped Matrix Plot. This plot reveals that the top rules are highly dependent on a few key variables, such as age, sex, and BP, which consistently predict CAD outcomes.



(b) Parallel Coordinates Plot. The plot highlights which variables consistently appear in the rules and their relationship to the consequent (e.g., "Cath=Normal").

Figure 20: Comparison of Grouped Matrix and Parallel Coordinates Plots.

Conclusion

The generated rules align well with the objectives of the assignment, as they provide clear associations between patient characteristics and CAD diagnosis. The high confidence and lift values confirm the strength of these patterns, making them reliable indicators of CAD status in this dataset. Each rule provides actionable insights into which patient attributes are most predictive of a CAD-negative diagnosis, fulfilling the assignment's goal of identifying key features associated with CAD.

2.5 Association Rule Mining Using FP-Grwoth

The FP-Growth algorithm is an alternative to the Apriori algorithm for finding frequent itemsets in transactional data. It is especially useful when dealing with large datasets because of its efficiency. Like Apriori, FP-Growth was used to identify the relationships between features in the *CAD dataset* and the diagnosis of coronary artery disease (CAD). The analysis aimed to generate frequent patterns and rules, with a focus on those most associated with a positive or negative CAD diagnosis.

2.5.1 FP-Growth Algorithm Implementation

The FP-Growth algorithm was applied with the following parameters:

- **Minimum Support**: Set to 0.03, meaning that each rule must appear in at least 3% of transactions (approximately 9 transactions).
- **Minimum Confidence**: Set to 0.8, meaning the rule must be correct at least 80% of the time.
- **Consequent**: The right-hand side (RHS) of the rules was fixed to predict CAD status, either "Cath=Normal" or "Cath=CAD."

The algorithm generated 13,723 rules, significantly more than Apriori. This is a hallmark of FP-Growth's efficiency, as it processes larger datasets more quickly and produces a higher number of rules. To focus on the most relevant rules, redundant rules were removed, leaving the top rules based on lift.

2.5.2 Discussion of the Top 10 Rules

Similar to Apriori, the FP-Growth algorithm's rules were sorted by lift, and the top 10 rules were inspected in Table 8.

LHS (Rule Antecedent)	RHS (Rule Consequent)	Support	Confidence	Coverage	Lift	Count
{Age=40-60, Sex=Female, Atypical=Y, BP=Normal}	{Cath=Normal}	0.0429	1.00	0.0429	3.48	13
{Age=40-60, DLP=Y, Sex=Female, Atypical=Y}	{Cath=Normal}	0.0330	1.00	0.0330	3.48	10
{Dyspnea=N, Sex=Female, Atypical=Y, BP=Normal}	{Cath=Normal}	0.0363	1.00	0.0363	3.48	11
{Poor R Progression=N, Sex=Female, Atypical=Y, BP=Normal}	{Cath=Normal}	0.0561	0.94	0.0594	3.29	17
{Obesity=Y, Age=40-60, DLP=Y, Atypical=Y}	{Cath=Normal}	0.0462	0.93	0.0462	3.25	14
{BMI=Overweight, Age=40-60, DLP=Y, Atypical=Y}	{Cath=Normal}	0.0396	0.92	0.0396	3.21	12
{Nonanginal=Y, DLP=N}	{Cath=Normal}	0.0363	0.92	0.0363	3.19	11
{Nonanginal=Y, DLP=N, Atypical=N}	{Cath=Normal}	0.0363	0.92	0.0363	3.19	11
{Nonanginal=Y, DLP=N, Atypical=N, Airway disease=N}	{Cath=Normal}	0.0363	0.92	0.0363	3.19	11
{Exertional CP=N, Nonanginal=Y, DLP=N}	{Cath=Normal}	0.0363	0.92	0.0363	3.19	11

Table 8: Top 10 FP-Growth Association Rules with Support, Confidence, Coverage, Lift, and Count

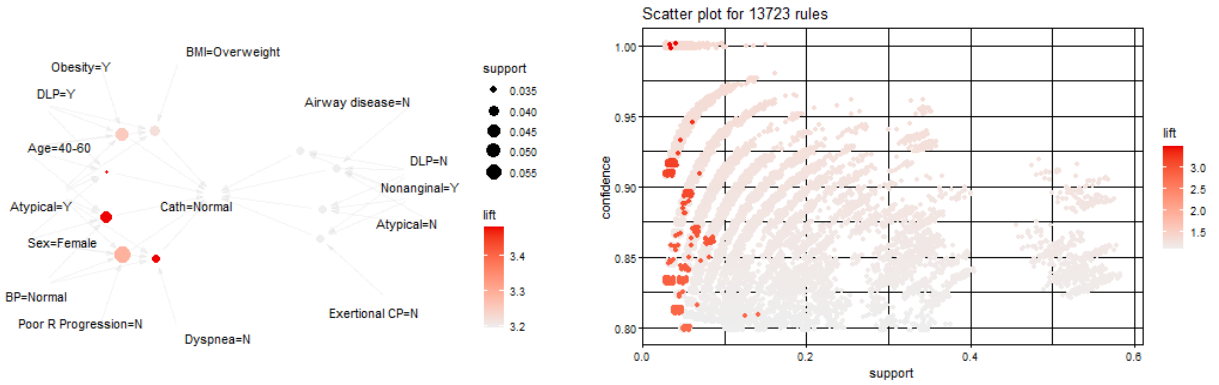
The top rules reflect similar patterns seen with Apriori, particularly the importance of attributes like age, sex, blood pressure, and atypical chest pain in predicting a CAD-negative diagnosis. The lift values above 3.0 indicate that these are strong associations, and the confidence values of 1.0 show the reliability of these rules in predicting the absence of CAD.

2.5.3 Key Insights from FP-Growth Rules

- **Age Group 40-60**: Similar to Apriori, the 40-60 age group frequently appears in the top rules, suggesting that this is a key age range for CAD risk analysis.
- **Sex=Female**: Female patients with specific combinations of factors (e.g., normal BP, atypical chest pain) are consistently predicted to be CAD-negative, reaffirming that gender plays a crucial role in CAD diagnosis in this dataset.
- **Blood Pressure**: Patients with normal blood pressure are more likely to be CAD-negative, and this factor appears consistently in the top rules.
- **Atypical Chest Pain**: Atypical chest pain, rather than typical chest pain, is often linked to a CAD-negative diagnosis, indicating that not all chest pain is equally indicative of heart disease.

2.5.4 Visualizations of FP-Growth Rules

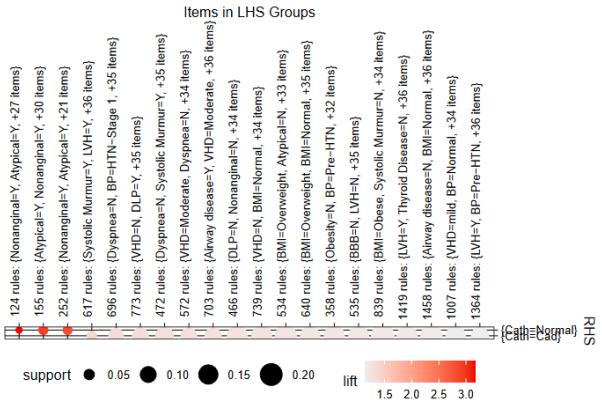
The visualizations generated by FP-Growth are similar to those from the Apriori analysis, offering additional insight into the relationships between features.



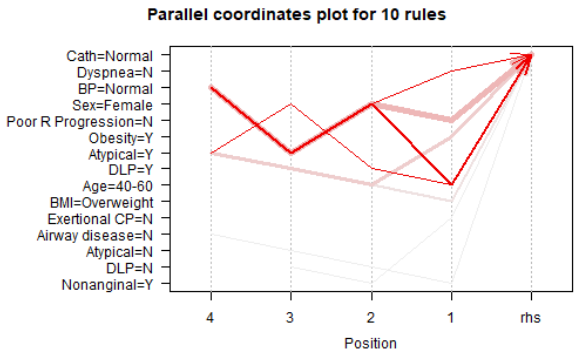
(a) Graph Visualization of the Top 10 Rules. This graph shows the relationships between the antecedents (LHS) and the consequent (RHS) for the top 10 rules, with nodes representing items and edges representing rules.

(b) The scatterplot illustrates the relationship between support, confidence, and lift for rules generated by FP-Growth. Rules with high confidence and lift are positioned in the upper-right corner, indicating their strong reliability.

Figure 21: Scatterplot and Graph Visualization of Association Rules.



(a) The grouped matrix plot highlights clusters of similar rules, such as "Age=40-60" and "BP=Normal," making it easier to identify common patterns and providing insight into which variables consistently co-occur in the rules, emphasizing their importance in predicting CAD status.



(b) The Parallel Coordinates Plot for FP-Growth highlights how key features like "Age=40-60," "Sex=Female," and "BP=Normal" consistently combine to predict Normal CAD status, visualizing the relationships across multiple dimensions.

Figure 22: Grouped Matrix Plot and Parallel Coordinates Plot for FP-Growth rules.

2.6 Pruning of Apriori and FP-Growth

Apriori Algorithm Pruning:

In the Apriori algorithm, rule pruning significantly improved the quality and relevance of the rules. Initially, 3,257 rules were generated, many of which were redundant or insignificant. By removing redundant rules and focusing on those with higher lift and confidence, the final set of rules was refined, leading to improved classification accuracy from 84.16% to 97.03%. This demonstrates that pruning effectively enhanced the algorithm's precision by retaining only the most meaningful associations.

FP-Growth Algorithm Pruning:

FP-Growth generated 13,723 rules, many of which were redundant. After pruning, focusing on support, confidence, and lift, accuracy improved from 81.52% to 93.4%. Though faster and more scalable, FP-Growth's final accuracy was slightly lower than Apriori's, suggesting that Apriori delivers more refined results after pruning.

Summary of Pruning Outcomes:

Algorithm	Initial Rules Generated	Accuracy Before Pruning (%)	Accuracy After Pruning (%)
Apriori	3,257	84.16	97.03
FP-Growth	13,723	81.52	93.40

Table 9: Comparison of Pruning for Apriori and FP-Growth

The table above summarizes the results of pruning for both algorithms. While both algorithms showed improvements after pruning, Apriori achieved higher accuracy post-pruning compared to FP-Growth, though FP-Growth generated rules more efficiently.

2.7 Conclusion

Both Apriori and FP-Growth algorithms effectively identified key associations between patient attributes and coronary artery disease (CAD) status. As seen in Table 7 and Table 8, the Apriori algorithm provided highly accurate predictions, particularly after pruning, with rules that had strong lift values and high confidence. The scatterplots and graph visualizations in Figure 19a and Figure 19b illustrate the strength of these associations, especially for key features like age, sex, and blood pressure. Despite being slower, Apriori produced more precise rules, making it better suited when accuracy is a priority.

In contrast, the FP-Growth algorithm, outlined in Table 8, demonstrated greater efficiency, quickly generating a larger set of rules. While its precision was slightly lower, as shown in Figure 21a, the consistency of its results is confirmed by the parallel coordinates plots in Figure 22b. FP-Growth is ideal for larger datasets due to its speed, but Apriori's accuracy makes it preferable when the goal is high precision in CAD diagnosis. Together, the two algorithms provide complementary insights into feature importance for CAD prediction.

References

1. R. Alizadehsani et al. (2013), "A data mining approach for diagnosis of coronary artery disease," Computer Methods and Programs in Biomedicine, vol.111, no.1, pp.52-61.

2. Kumbhare, T.A. and Chobe, S.V., 2014. An overview of association rule mining algorithms. International Journal of Computer Science and Information Technologies, 5(1), pp.927-930.

3. Nagaraj, S. and Mohanraj, E., 2020. A novel fuzzy association rule for efficient data mining of ubiquitous real-time data. Journal of Ambient Intelligence and Humanized Computing, 11(11), pp.4753-4763.

A Appendix

List of Tables

List of Figures

Department of Statistical Sciences Plagiarism Declaration Form

A copy of this form, completed and signed, to be attached to all coursework submissions to the Statistical Sciences Department. Submissions without this form will not be marked.

COURSE CODE: STA5076Z
COURSE NAME: Supervised Learning
STUDENT NAME: Tsapang Masheego
STUDENT NUMBER: MSHTSA009
TUTOR'S NAME:
TUTOR GROUP #:

PLAGIARISM DECLARATION:

1. I know that plagiarism is wrong. Plagiarism is to use another's work and pretend that it is one's own.
2. I have used a generally accepted citation and referencing style. Each contribution to, and quotation in, this tutorial/report/project from the work(s) of other people has been attributed, and has been cited and referenced.
3. This tutorial/report/project is my own work.
4. I have not allowed, and will not allow, anyone to copy my work with the intention of passing it off as his or her own work.
5. I acknowledge that copying someone else's assignment or essay, or part of it, is wrong, and declare that this is my own work.

Note that agreement to this statement does not exonerate you from the University's plagiarism rules (http://www.uct.ac.za/uct

Signature: T. Mashego **Date:** 16/09/24