University of Cape Town

Course Code

STA 5077Z

Cluster Analysis for Fetal Health Classification and Association Rule Mining for Coronary Artery Disease Diagnosis

Author:
Khuliso Mmbi

Student Number:
MMBKHU001

September 16, 2024

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: **STA5077Z**

COURSE NAME: Unsupervised Learning

STUDENT NAME: Khuliso Mmbi

STUDENT NUMBER: MMBKHU001

TUTORS NAME: _____ TUT. 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/policies/plagiarism_students.pdf).

Signature: khuliso mmbi Date: 16/09/2024

Contents

1 Introduction

The aim of this report is to provide an analysis of two heath care datasets, focusing on fetal health monitoring and the Coronary Artery Disease (CAD) diagnosis by using two Unsupervised learning techniques namely, the Clustering analysis methods and the Association rule mining methods .

This will be explored using two section in which section 1 will explore the fetal health monitoring using cluster analysis and later on on section 2, CAD will be analysed using Association rule mining techniques. The findings of these analyses are expected to give valuable information regarding the public health initiatives, particularly in refining fetal health classifications and enhancing CAD diagnosis.

2 Fetal Health Analysis

Fetal health status is one of the important components of prenatal care. This is assessed using the cardiotocogram (CTG) exams, which monitor the fetal heart rate and uterine contractions to ensure the well-being of the fetal throughout the pregnancy. Ensuring accurate assessment of the status could help in identifying complications early and coming up with ways to improve the health of the fetal.

To effectively analyse and interpret the fetal health data, clustering methods will be explored. Clustering helps in grouping similar observations together, and thereby identifying patterns and anomalies in fetal health metrics. By segmenting the data into meaningful clusters, healthcare professionals can gain insights into the various health states of the fetal and make informed decisions based on the data.

The aim of this section is to explore whether the fetal health data set can be categorized into three classes: Normal, Suspect and Pathological as done by the obstetricians. This will be explored by first touching on the explanatory data analysis, followed by clustering analysis to group the fetal health data into distinct categories based on the observed metrics. Lastly, all the results will be summarised and the effectiveness of each clustering method will be evaluated..

2.1 Explanatory data analysis

2.1.1 Data description

Table 1 below shows the description of all the variables that are in the fetal health data set. There are 21 variables and 2126 observations, all of which are numeric variables except for the histogram tendency which is a factor. The dataset also had 15 duplicates which were removed. All the variables will be referred by their abbreviations through out this report.

Variable	**Abbreviation**	**Description**	**Variable Type**
baseline value	LB	Baseline heart rate by SisPorto	Numerical
accelerations	AC	Number of accelerations detected by SisPorto	Numerical
fetal movement	FM	Number of fetal movements detected by SisPorto	Numerical
uterine contractions	UC	Number of uterine contractions detected by SisPorto	Numerical
light decelerations	DL	Brief drops in fetal heart rate	Numerical
severe decelerations	DS	Significant drops in fetal heart rate	Numerical
prolongued decelerations	DP	Extended period of decelerations	Numerical
abnormal short-term variability	ASTV	Time short-term variability outside normal range	Numerical
mean value of short-term variability	MSTV	Average short-term variability	Numerical
percentage of time with abnormal long-term variability	ALTV	Time long-term variability outside normal range	Numerical
mean value of long-term variability	MLTV	Average long-term variability	Numerical
histogram width	Width	Range of heart rate values	Numerical
histogram min	Min	Lowest frequency in histogram	Numerical
histogram max	Max	Highest frequency in histogram	Numerical
histogram number of peaks	nMax	Number of peaks in histogram	Numerical
histogram number of zeroes	nZeroes	Times frequency is zero	Numerical
histogram mode	Mode	Most frequent heart rate value	Numerical
histogram mean	Mean	Average heart rate value	Numerical
histogram median	Median	Middle value of heart rate data	Numerical
histogram variance	Variance	Variability in heart rate values	Numerical
histogram tendency	Tendency	Tendency: -1=left; 0=symmetric; 1=right	Categorical

Source: `https://archive.ics.uci.edu/dataset/193/cardiotocography`

Table 1: Variables descriptions

2.1.2 Summary statistics

Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
LB	106.000	126.000	133.000	133.300	140.000	160.000
AC	0.000	0.000	0.002	0.003	0.006	0.019
FM	0.000	0.000	0.000	0.009	0.003	0.481
UC	0.000	0.002	0.004	0.004	0.007	0.015
DL	0.000	0.000	0.000	0.002	0.003	0.015
DS	0.000	0.000	0.000	0.000	0.000	0.001
DP	0.000	0.000	0.000	0.000	0.000	0.005
ASTV	12.000	32.000	49.000	47.000	61.000	87.000
MSTV	0.200	0.700	1.200	1.333	1.700	7.000
ALTV	0.000	0.000	0.000	9.847	11.000	91.000
MLTV	0.000	4.600	7.400	8.188	10.800	50.700
Width	3.000	37.000	67.500	70.450	100.000	180.000
Min	50.000	67.000	93.000	93.580	120.000	159.000
Max	122.000	152.000	162.000	164.000	174.000	238.000
nMax	0.000	2.000	3.000	4.068	6.000	18.000
nZeroes	0.000	0.000	0.000	0.324	0.000	10.000
Mode	60.000	129.000	139.000	137.500	148.000	187.000
Mean	73.000	125.000	136.000	134.600	145.000	182.000
Median	77.000	129.000	139.000	138.100	148.000	186.000
Variance	0.000	2.000	7.000	18.810	24.000	269.000

Table 2: Descriptive Statistics of Fetal Health Data

Table 2 above summaries the fetal dataset. There are 21 variables which are used for Cardiotocogram exams and has been summarised below:

Fetal Heart metrics

- The foetal heart rate has been summarised by using the Baseline Value, Accelerations Foetal movements and Decelerations. The baseline heart rate has a mean and median of 133, with the minimum heart rate being 106 and maximum heart rate being 160. Accelerations, contractions and decelerations (light, severe and prolonged decelerations) all have low median values. This suggest that our data is left skewed and possibly that the heart rate is more stable with low occurrences of accelerations or deceleration.

Variability metrics

- For variability, the abnormal short-term variability is intermediate, with a mean of 46.99%, and ranges from a minimum of 12% to a maximum of 87%. In contrast, the abnormal long-term variability has a mean of 9.85%. This suggests that, within this dataset, the fetus is more likely to experience faster oscillations in heart rate rather than slower ones.

Histogram metrics

- The differences in the mean, median, mode, maximum and minimum for Histogram is not noticeable and this can indicate that there is consistent heart rate and minimal skewness. Similarly, the central tendency metrics (mean, mode and median) show a stable heart rate distribution. There is a tendency towards the higher heart rate with a tendency of 165 for left symmetry and 846 for right symmetry.

2.1.3 Correlation matrix

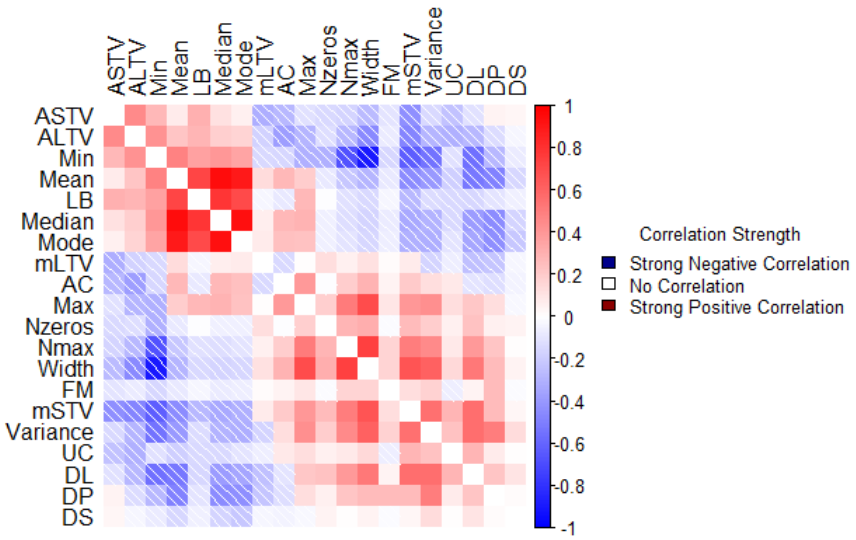


Figure 1: Correlation matrix

The correlation matrix has been summarise in Figure 1 above, highlighting key relationships within the dataset.

Strong positive Correlation (Dark red)

- There is a strong positive correlation between the Baseline values and the central tendency measures (mean, mode and median and min) suggesting that as the heart rate increases, these measures also increase
- Severe Decelerations and prolonged decelerations are also strongly correlated indicating that the significant drop in heart rate is often associated with prolonged decelerations.

Positive correlation (Red)

- The correlation between the variance and mean short term variability is positive but nit strong which suggest that as the heart rate increases the overall variance also increases which could potentially lead to more fluctuations in the heart rate.

Weak negative correlation (Light blue or Red)

- A weak negative correlation exists between abnormal long-term variability and accelerations which suggest that the changes in the heart beat does not really affect heart rate.

Negative correlation (Blue)

- There is no significant correlation between the Number of zero frequencies and the number of peaks implying that the heart rate is more uniform regardless of the gaps in heart rate.

2.1.4 Data visualisation

Figure 2 below summarises all the numeric variables in boxplots. The plots for LB, AC, FM, UC, DL, DS and DP have values which are closer to 0 and do not show any outliers. This could indicate that the is a stable baseline heart rate and values are close to the mean. The small values close to 0 are an indication that most of the observations have little to no occurrences of accelerations, foetal movements, contractions or decelerations. The rest of the variables show outliers which means for example, for the short-term variability and long-term variability, there are times where the heart rate may deviate from the usual heart rate. Another potential reason for the outliers could be errors in measurements.

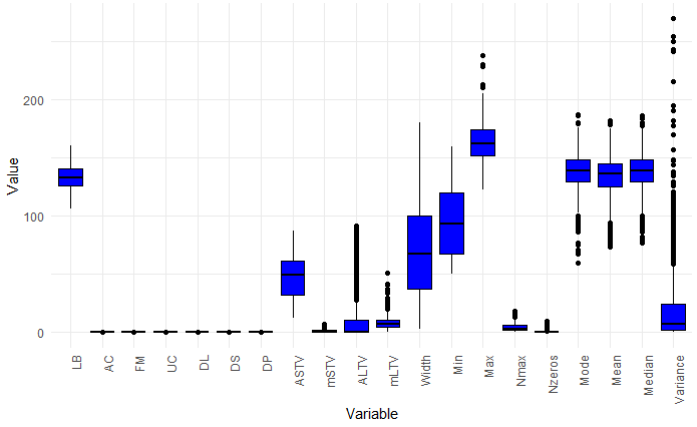


Figure 2: Box Plots for numeric variables

Figure 3 below summarises the Histogram tendency. From the graph, there is a tendency towards the higher heart rate with a tendency of 165 for left symmetry and 846 for right symmetry.

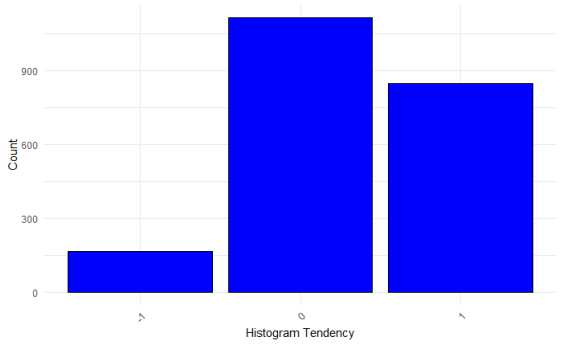


Figure 3: Histogram tendency

Figure 4 below shows how the fetal health data is structured on a lower dimensional space using Principal Component Analysis (PCA). Two principal components is used to capture the variance in the data. From the scatter plot it is evident that the points are spread out in certain patterns which suggests the potential natural grouping of the data. There are also a lot of outliers, but as we are working with a health dataset, these outliers will not be removed as they might represent some important information about the fetal dataset.

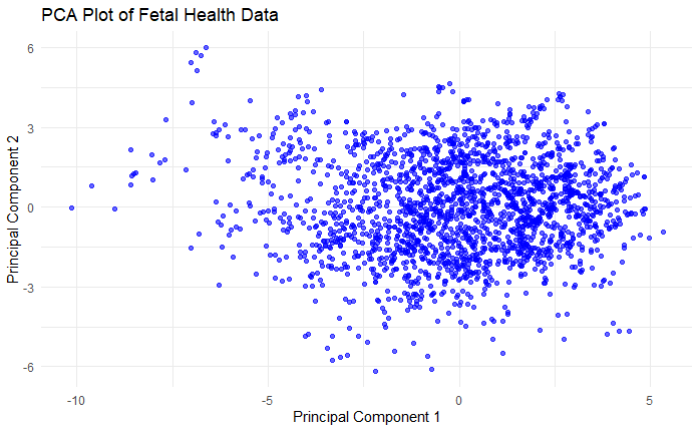


Figure 4: Fetal health dataset structure

2.1.5 Data standardisation

Based on the information provided above, it is clear that the variables in the dataset are on different scales. To mitigate potential errors and ensure more accurate results, the data has been standardized. This will also ensure that the distance measures are accurately calculated.

Figure 5 below summarises the box plots for all the numeric variables after scaling. In contrast to Figure 2 above, the box plots are now more in a uniform scale and are easier to compare. For instance, variables such as LB and Variance are now compressed in the same range. Outliers are also now more visible in certain variables such as LB, AC, FM, UC, DL, DS and DP which previously had no visible outliers.

All the variables in the data set now have the same scale and are comparable. The next section will now cover the different clustering analysis that will be used to explore the fetal health dataset.

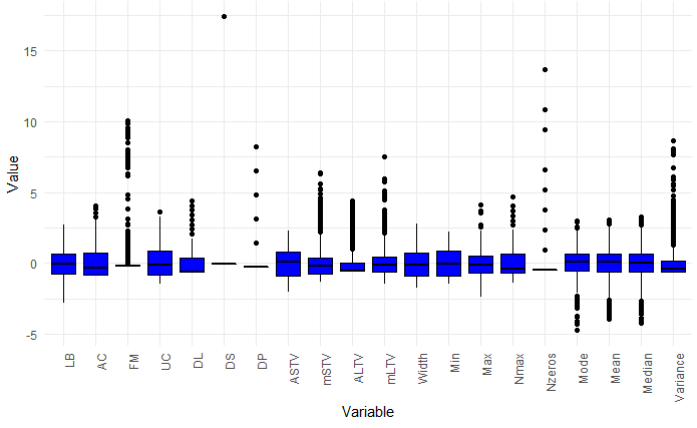


Figure 5: Box Plots for numeric variables after standardisation

2.2 Cluster Analysis

Clustering methods involves grouping observations in a dataset into groups which are similar to each other (James et al., 2013, p. 385). This similarity is determined by measures such as distance metrics or similarity functions, depending on the specific clustering algorithm used. This analysis will be explored using the fetal health data set and will focus on 4 clustering methods, namely the partitioning methods, Hierarchical methods, density based methods, and lastly the probabilistic methods. The choice of these methods is based on their individual abilities to handle different data structures effectively. Additionally, Principal components analysis (PCA) will only be used to visualise the results of these methods by providing a low-dimensional representation that captures the maximum amount of variation.

2.2.1 Partitioning methods

Partitioning methods involves grouping the data points into a certain number of groups(K) where each observation falls within one group. The aim of this method is to divide the data set into different groups or K partitions.

The fetal health data set contains outliers and the best partitioning methods chosen were the K Medoids and K means. The K-Means categorises the data points by minimising the sum distances between the data points and their assigned clusters. This proved to be essential for the data set as it aims to minimize the variance within-cluster. The K Medoids on the other hand, is efficient when working with data that contains outliers and also requires less steps to converge.

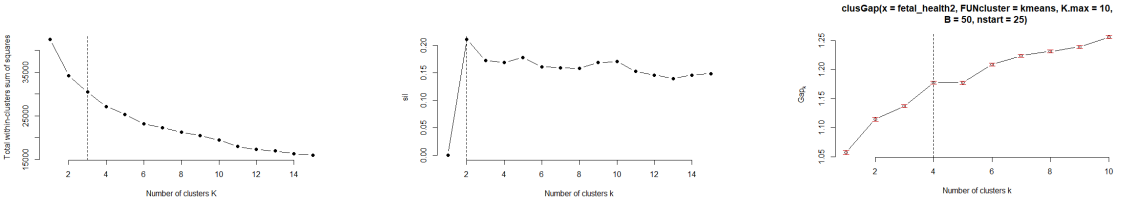
K Means

Determining Optimal number of clusters

The suggested K as per the obstetricians is 3, corresponding to Normal, Suspect, and Pathological categories.To ensure accuracy, other methods of determining K were also explored. These methods included:

- **Silhouette Method**: Evaluates the degree of separation between clusters.
- **Elbow Method**: Assesses the Sum of Squares between data points and their centroids.
- **Gap Statistic**: Compares the spread between clusters to what we would expect if no real clusters existed.

Figure 6 illustrates below the recommendations from each method: the Elbow Method recommended K=2, the Silhouette Method offered K=3, and the Gap Statistic proposed K=4.



Elbow method Silhouette Method Gap statistic

Figure 6: Methods for determining K

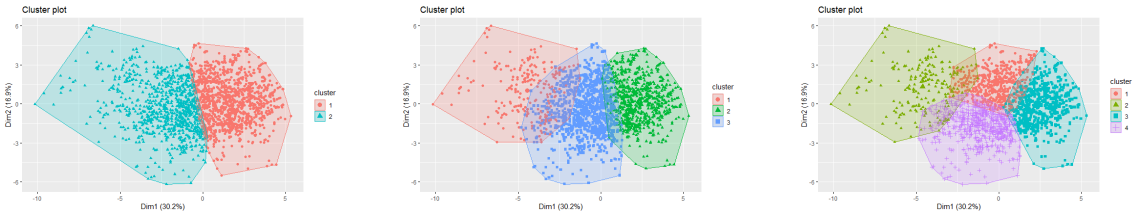
The K means model was explored with the different level of k (2,3,4) with maximum iterations of 100. Table 3 below summarises the performance of each model. The Silhoette method, had the

highest silhoette score of 0.211 which suggest that when k is 2, the model is able to achieve well-defined clusters, however it's WCSS of 34201.2, which is the highest amongst the three methods, could suggest that the clusters are not well separated or compact. The model with K = 3, had lower values for both the silhoette score (0.172) and the WCSS (30438.15) as compared to all the other models, which indicating a good balance between quality and separation. For K =4, the WCSS decreased further to 27122.22 reflecting how the model is improving at fitting the data as the number of clusters increase., however, the decrease in Silhouette score to 0.168, may indicate that the clusters are more compact and may not be well separated.

Method	**Number of K**	**Silhouette score**	**WCSS**
Silhouette	2	0.211	34201.2
Elbow	3	0.172	30438.15
Gap Statistic	4	0.168	27122.32

Table 3: Evaluation of Partitioning Methods

Figure 7 below further shows the cluster plots when using the different levels of K. With K = 2, the data is well separated with a little overlap, which could suggest that this is the natural grouping or the model is over-fitting. And with a K = 3, the preferred clusters, there is little overlapping between the 3 groups and for K = 4, all the groups are overlapping and not well separated. Overall, the optimal number of clusters when using the K means model proved to be 3 as it provides a good balance between quality and separation.



K = 2 K =3 K = 4

Figure 7: K Means plots using different K

The Welch sample T-test was also performed to assess the statistical significance between clusters for the different levels of k. Table 4 summarises the results. This table provides the t-values, degrees of freedom, p-values, and 95% confidence intervals for each pair of clusters across different k values.

From the table, for k =2, the difference between the clusters was not statistically significant (p=0.6552). For K = 3, significant differences were between clusters 1 and 2 (p = 0.0045) and between clusters 1 and 3 (p = 0.04121). Whilst for clusters 2 and 3, they were insignificant with p = 0.8342. K = 4 on the other hand, most clusters were statistically significant, except for clusters 2 and 4 which did not show any statistically significant difference.

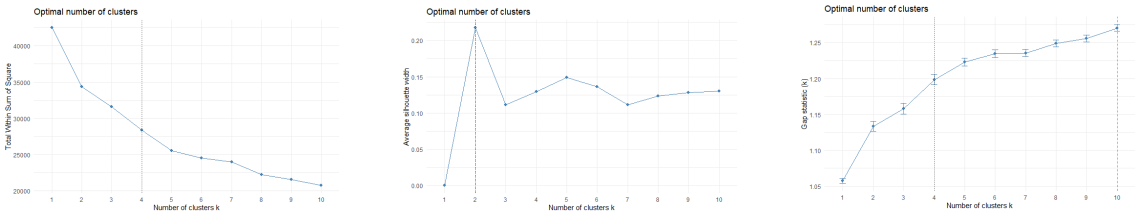
Test Number	Clusters Compared	t-Value	p-Value
k = 2			
1	1 vs 2	0.446	0.655
k = 3			
1	1 vs 2	2.845	0.004
2	1 vs 3	2.042	0.041
3	2 vs 3	0.209	0.834
k = 4			
1	1 vs 2	-8.480	¡ 2.2e-16
2	1 vs 3	-3.017	0.003
3	1 vs 4	-8.593	¡ 2.2e-16
4	2 vs 3	3.065	0.002
5	2 vs 4	-0.144	0.885
6	3 vs 4	-3.164	0.002

Table 4: Welch Two Sample t-Test Results for Different Numbers of Clusters

Overall, for the K means, K =3 proved to be the best choice as it offers a good balance between cluster and quality and has a relative high silhouette score and lower WCSS as compared to k =2. Additionally, the visualisations showed well defined clusters with minimal overlap which aligns with the silhouette and t-test results in table 4. This further support the obstetricians preference of categorizing fetal health into three (3) classes: Normal, Suspect and Pathological

K Medoids

The K Medoids method was evaluated using different levels of k and assessed through the WSS method, the Silhouette method, and the Gap statistic method. Figure 8 displays the results, where the WCSS suggested k = 4, while both the Silhouette and Gap statistic methods favored k = 2 and k = 4. This is consistent with the results obtained from K-means clustering above.



WCSS Silhouette Method Gap Statistic

Figure 8: Methods for determining k

Table 5 summarizes the performance of K Medoids for k = 2 and k = 4 , based on the PAM algorithm used for this dataset. For k = 2 , the silhouette score was the highest at 0.217, and the objective function value was 7.591. In contrast, the silhouette score for k = 4 was notably lower at 0.129, suggesting that k = 4 is less effective in terms of cluster cohesion and separation. These findings imply that clustering the data into three clusters is not suitable.

The results indicate that while the silhouette score for k = 2 was lower than other methods, it still offers a good balance between cluster separation and cohesiveness, as evidenced by its silhouette score and objective function value.

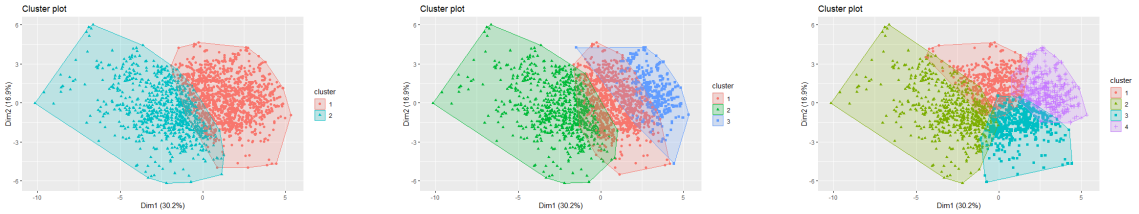
Method	**Number of k**	**Silhouette**	**Objective Function**
Silhouette	2	0.217	7.591
Obstetrician Preference	3	0.111	7.179
Gap Statistic / WSS	4	0.129	6.740

Table 5: Evaluation of clustering methods based on different metrics

The medoids identified for different values of k were as follows:

- Medoids for $k = 2$: Data points with IDs 876 and 1328
- Medoids for $k = 3$: Data points with IDs 850, 1328, and 546
- Medoids for $k = 4$: Data points with IDs 850, 1328, 546, and 580

Figure 9 illustrates the separation of data using various k values identified above. Similar to K-means clustering, k = 2 provides near-perfect separation, which diminishes as k increases. The decreasing objective function values with increasing k suggest that clustering quality improves, but this improvement is balanced against the simplicity of fewer clusters.



K = 2 K = 3 K = 4

Figure 9: K Medoids using different k values

The Welch Two Sample t-Test results in Table 6 below show that for k = 2, there is a strong separation between clusters 1 and 2 (p-value 1.692×10^{-11}), indicating statistically significant clustering. For k = 3 , all cluster pairs were statistically significant, which suggests well-defined clusters. However, for k = 4, the difference between clusters 2 and 3 was not statistically significant, while other pairs were significant.

The optimal number of K proved to be 2 when using the Medoids model. This is because it gave a clear and interpretable structure with strong separation clusters, a good silhouette score and a good balance in the objective function.

Summary

For the partitioning methods, the K means showed the optimal number of clusters for the fetal health data set is 3 suggesting that the data could be well clustered into three (3) classes: Normal, Suspect and Pathological. However, for the K Medoids, the optimal number of clusters is 2 which could suggest that the natural group of the data could be just either Normal or not Normal.

Test Number	Clusters Compared	t-Value	p-Value
k = 2			
1	1 vs 2	-22.544	1.692×10^{-111}
k = 3			
1	1 vs 2	-20.615	1.311×10^{-93}
2	1 vs 3	-4.261	2.057×10^{-5}
3	2 vs 3	11.647	3.107×10^{-31}
k = 4			
1	1 vs 2	-36.448	8.264×10^{-282}
2	1 vs 3	-44.258	0.000
3	1 vs 4	-19.705	1.638×10^{-85}
4	2 vs 3	1.880	0.060
5	2 vs 4	17.657	2.791×10^{-69}
6	3 vs 4	19.363	1.068×10^{-82}

Table 6: Welch Two Sample t-Test Results for Different Numbers of Clusters

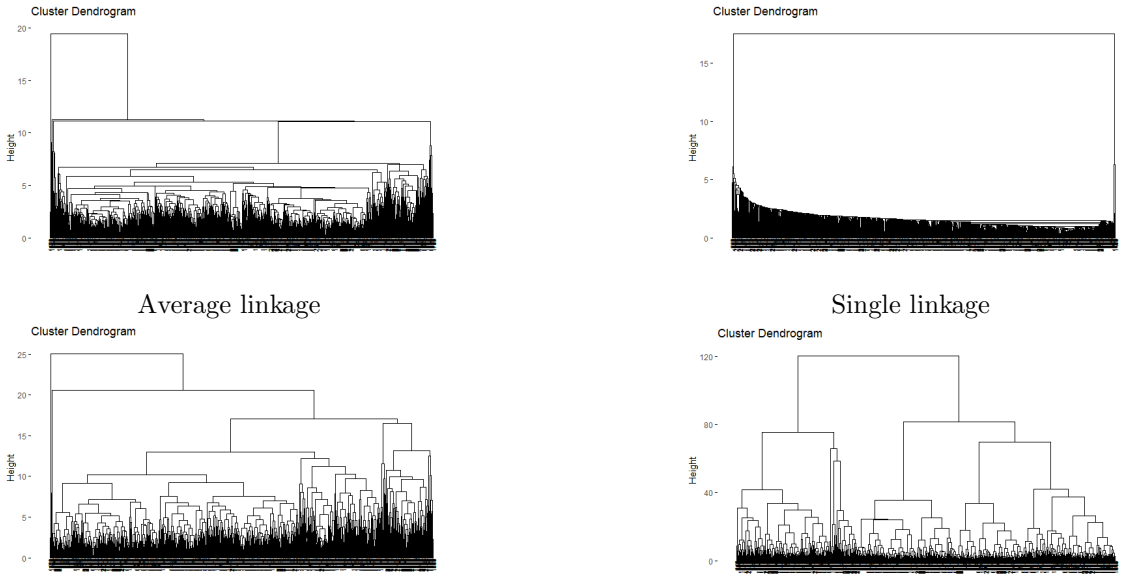
2.2.2 Hierarchical methods

The hierarchical method is another type of the distance-based methods that was used for the fetal health dataset. It aims to create a hierarchical representation of clusters in a dataset by initially treating each data point as it's own separate cluster. Then, using the selected distance metric, the algorithm keeps combining the nearest clusters until a predefined stopping threshold is satisfied.

The agglomerative approach was used to select the linkage method. The aggloromtive method is best when working with a large data set as it provides a good visualisation of how the data is clustered which could help in understanding how the fetal health conditions are grouped based on CTG features. Five methods explored are:

- Complete: calculates the maximum distance between clusters
- Average linkage: Combines clusters based on the distance between points in different clusters
- Single linkage: Utilizes the minimum distance between points to merge clusters
- Median linkage: Provides a balance between the single and complete linkage by considering the median between clusters
- Ward's method: best for producing well separated clusters by minimising the total within-cluster variance.

The number of clusters found using the different methods was 3 for most of the linkage methods as can be seen by figure 10 below. The methods were further evaluated using the average silhouette which measures how well the clusters are separated,the Inertia (WCSS) which measures the compactness of the clusters and the Cophenic correlation which evaluates how the dendogram fits the distance data.



Average linkage Single linkage

Complete linkage Ward's method

Figure 10: Number of clusters K using different linkage methods

Additionally, the results from Table 7, the single linkage method had the highest silhoette score of 0.6912, which could suggest that it was best at separating the clusters. This was followed by the complete linkage and average linkage method with scores of 0.5237 and 0.4910 respectively. On the Cophenic correlation, the Average linkage has the highest score of 0.827 which means that it is able to keep the original distances between data points when clustering. Similarly with the WCSS, all the models have the same values except for the Ward's method, this suuggest that it might not be suitable for this dataset. Overall, the average linkage proved to be the best as it has the highest Cophenic correlation and the silhoette score of 0.4910. Given the nature of the dataset, the average linking dataset will also be able to effectively separate the clusters without being influenced by the outliers.

Method	**Average Silhouette**	**Cophenetic Correlation**	**WCSS**
Average Linkage	0.4910	0.8269	45638.28
Single Linkage	0.6912	0.7432	44577.7
Complete Linkage	0.5237	0.6543	45585.64
Ward's Method	0.1009	0.4269	50617.59

Table 7: Comparison of linkage Methods

Figure 11 below shows the structure of the dendogram when using the average linkage and the tree being cut into 3. There are three visible clusters, with cluster 1 (blue) and cluster 2 (yellow) interacting at lower heights. This is an indication of how the two clusters are similar to each other. Cluster 3 (black) on the other hand, has a higher height suggesting it's disssimilar to the other two clusters. This height difference could be an indication that the data can be naturally grouped into three clusters as suggested by the obstetricians.

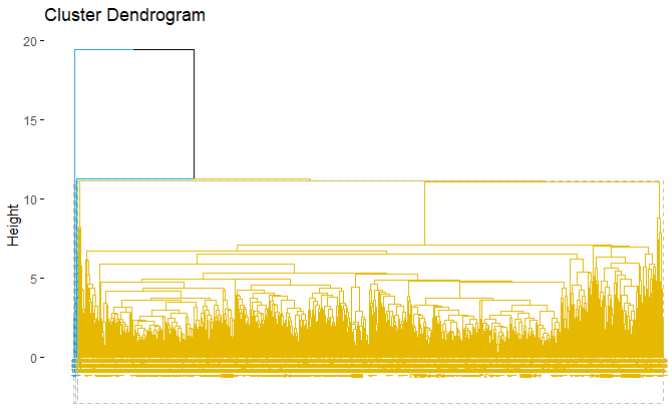


Figure 11: Cluster dendongram

Table 8, shows the results from the Welch table which show that clusters 1 and 3, as well as clusters 2 and 3, reveal statistically significant differences, with very low p-values (0.000 and 0.001, respectively). However, the difference between clusters 1 and 2 is not statistically significant, with a p-value of 0.085, which suggests that there might not be a distinct separation between these clusters. This lack of statistical significance between clusters 1 and 2 could also imply that a clustering configuration with K=3 may provide a more meaningful and refined structure compared to K=2. Therefore, the optimal number of clusters when using the Hierarchical method is three and the fetal health data can be divided into 3 classes as currently being done by the obstetricians.

Clusters Compared	T-Value	P-Value
1 vs 2	2.056	0.085
1 vs 3	-9.887	0.000
2 vs 3	-5.613	0.001

Table 8: Welch Two Sample t-Test Results for Different Numbers of Clusters

2.2.3 Density Based Spatial Clustering of Applications with Noise

Th Density Based Spatial Clustering of Applications with Noise (DBSCAN), which is part of the density based methods of clustering was the forth clustering technique explored. DBSCAN works by grouping the points in the data set that are close together based on the distance density between other points. The K-Nearest Neighbour is one approach that can be used to determine the distances between the points.

The parameters used for the fetal health dataset is the epsilon(eps) which determines what the maximum distance between points should be for them to be considered as neighbours and the minimum points(minPts) which specifies the minimum number of points that is needed to form a cluster. The kNNdistplot was used to determine the ideal eps parameter with a minPts of 4 determined by the rule of thumb of adding 1 to the number of dimensions which yielded an eps value of 3 (figure 12). The DBSCAN was then tested using these parameters and they did not produce any significant clusters as can be seen by figure 13. The model gave 8 clusters, with cluster 1 having more than half of the total data set. The average silhoette score for this model was 0.1380.

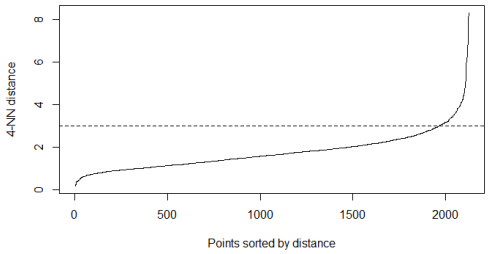


Figure 12: KNN Plot

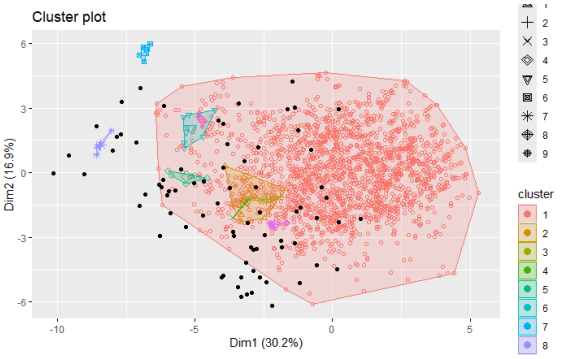


Figure 13: DBSCAN Plot 1

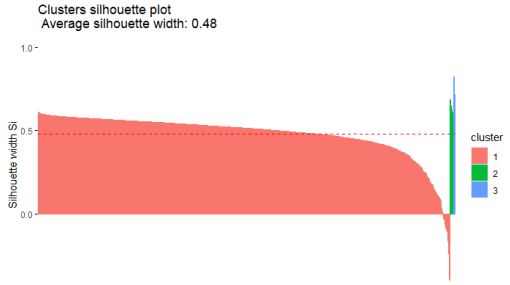


Figure 14: Silhoette plot

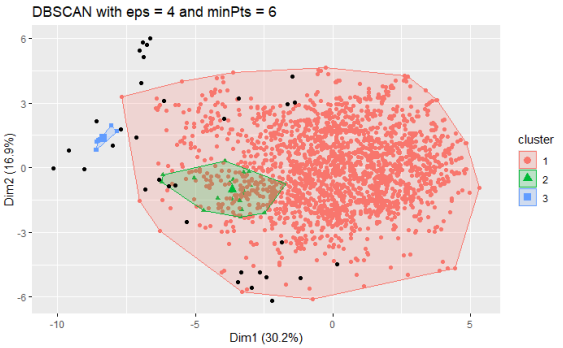


Figure 15: DBSCAN Plot 2

Both eps and minPts were re-evaluated within the following range of values to improve the clustering: eps values of c(2, 2.5, 3, 3.5, 4, 5) and minPts values of c(4, 5, 6). Effective clustering was attained with this parameter adjustment at an eps value of 4 and minPts of 6. This can be seen by figure 15, where the black dots represents the noise, which are the points which could not be captured by the model. The average silhouette improved to 0.48(figure 14). This indicates that the model was able to cluster more effectively when using the eps value of 4 and minPts of 6. The results further confirms that the fetal health dataset can be categorised into 3 classes as suggested by the obstetricians

2.2.4 Gaussian mixed methods

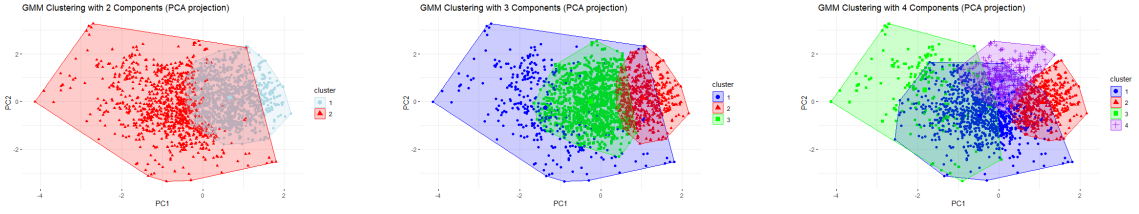
The Gaussian mixture model is used for determining the probability that the data points in the data set belong to a cluster. This is achieved by applying a soft assignment of data points to clusters based on their likelihood. This is different from the k-means explored above where a more hard assignment is used.

For this method, the number of clusters (k of 2,3,4) used were the same as determined in section 2.2.1 above for the partitioning methods. Table 6 shows the results of the Gaussian finite mixture model fitted by EM algorithm. The log-likelihood for a k of 2 is -54873.23, k of 3 is -47789.36 and fir a k of 4 is -44674.22, this suggests that the model is improving with the number of clusters. The Bayesian Information Criterion (BIC) aims to capture the model complexity, and therefore lower values are preferred. From the table, the BIC started to decrease with an increase in K which shows that at k of 4, the mode is more effective at finding the balance between fit and complexity. Similarly for the ICL, lower values are preferred, where the model with k of 4 performed better, however this could also indicate that the model is over-fitting due to the increase in clusters. And lastly, the silhouette score, is highest for a k of 2 which shows that as more and more clusters are added, there is less distinct separation.

Number of clusters	log-likelihood	df	BIC	ICL	Silhoette score
2	-54873.23	43	-110075.9	-110235.7	0.2474
3	-47789.36	84	-96222.34	-96303.13	0.1611
4	-44674.52	106	-90161.2	-90258.14	0.1329

Table 9: Cluster Evaluation Metrics

Figure 16 below further shows how the data points are separated when using the different levels of k. For all the plots, there seems to be an overlap between the clusters. With a k of 2, there separation between the clusters is clear but the 2 clusters are overlapping. Similarly, with a k of 3 and a k of 4, the clusters are still overlapping, and this could be a result of the soft assignment when using the gmm method. However, it is also noticeable that although there is a lot of overlapping, the separation becomes more and more complex with the increase in k. Given the evaluation metrics in table 6 above, it is evident that the optimal number of clusters when using the GMM model is 4 as it provides a balance between the model performance and the interpretability.



K =2 K = 3 K = 4

Figure 16: Gaussian mixed models

2.3 Comparative Analysis

Several techniques for evaluating the clustering methods were applied. The evaluation aimed to identify which clustering approach best captured the underlying patterns in the dataset and provided the most insightful results for further analysis.

Internal Methods

The internal methods uses the information that is in the clustering process to evaluate the effectiveness of the clustering method without using any external information. For the fetal dataset, this was done using the Silhouette score and the Dunn Index. Table 10 below summarises the results.

Method	Optimal K	Silhouette	Dunn Index
K Means	3	0.2135	0.0152
K Medoids	2	0.2177	0.0161
Hierarchical	3	0.4910	0.2272
DBSCAN	3	0.4812	0.1320
Gaussian Mixed	3	0.1485	0.0138

Table 10: Clustering Methods Evaluation

Hierarchical Clustering

- This model was the top performing, with the highest Silhoette score of 0.4910 and a Dunn Index of 0.2272. This suggests that the use of the average linkage method was effective in creating well separated and compact clusters. The high performance could have also been a result of using distance in forming clusters.

DBSCAN

- As a reults of it's ability to detect different shapes and handle noise effectively, DBSCAN was also able to perform well with a Silhoette score of 0.4812, almost the same with Hierarchical and a Dunn Index of 0.1320.

K means and K medoids

- The partition methods showed an average performance with a with similar Silhoette scores of 0.2177 and 0.2135 respectively. Although the two methids successfully managed to partition the data sets into different clusters, the were not able to clearly separate the clusters as effective as the Hierarchical model.

GMM

- The GMM was the least performing model with silhoette scores of 0.1485 and a Dunn Index of 0.013. This could have been affected by the soft clustering aprroach of the model

Optimal number of clusters

- The majority of the models indicated 3 clusters to be the optimal number of clusters for the fetal health dataset. This could suggest that the dataset could be grouped into three (3) classes: Normal, Suspect and Pathological as currently being done by the obstetricians. However, the highest silhoette score being less than 0.5 shows an area of improvement for the models.

2.4 Summary

In conclusion, the different clustering techniques conducted in this paper confirmed that the fetal health dataset can be grouped into three classes. The hierarchical clustering method proved to be the best performing by offering well-separated clusters.

3 Coronary Artery Disease

Coronary artery disease (CAD) is one of the leading causes of mortality in the world. It mostly develops when cholesterol accumulates in the artery wall resulting in plaques (Medical News Today, 2023). This build up then restricts the blood flow, and causes potential heart attacks and other complications. Early detection of the symptoms is crucial for early diagnosis and management of CAD.

The aim of this section is to use the Association rule mining to determine the features or symptoms that are mostly associated with CAD. By understanding these symptoms, researchers will be able to use more targeted screening protocols and be able to detect early symptoms of CAD in patients.

3.1 Explanatory data Analysis

3.1.1 Data description

The CAD dataset has 56 features, of which 21 are numerical and 35 are categorical and 303 observations. Appendix A, Table 20 contains all the data descriptions.

3.1.2 Correlation matrix

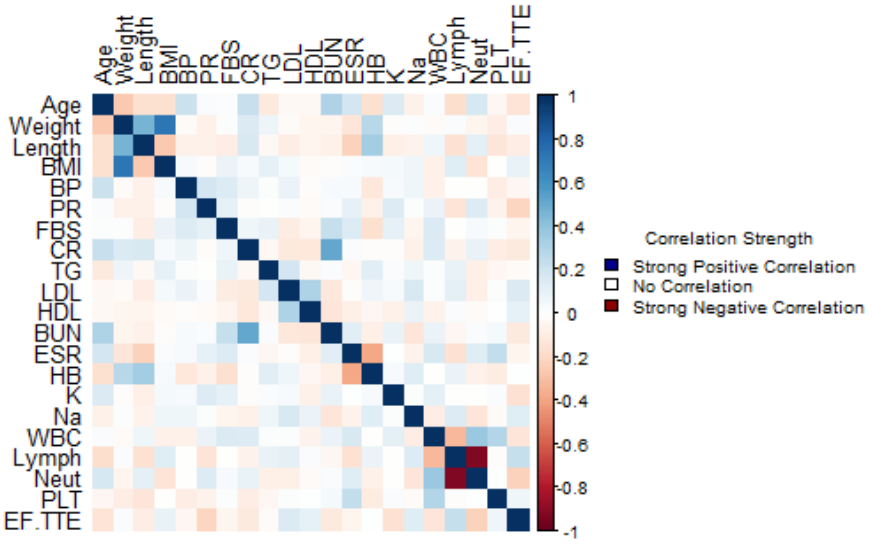


Figure 17: Correlation matrix

Figure 17 above summarises the correlation between the numeric variables. There is a strong positive correlations (dark blue) between Age, Weight, and Length. Some variables, like WBC, Lymph, and Neut, form a cluster of moderate to strong correlations. Other variables such as EF and TTE show a strong negative correlation (dark red). Overall, most variables exhibit weak correlations with each other, as indicated by the pale colours.

3.1.3 Summary statistics

Statistic	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Age	30.0	51.0	58.0	58.9	66.0	86.0
Weight	48.00	65.00	74.00	73.83	81.00	120.00
Length	140.0	158.0	165.0	164.7	171.0	188.0
BMI	18.12	24.51	26.78	27.25	29.41	40.90
BP	90.0	120.0	130.0	129.6	140.0	190.0
PR	50.00	70.00	70.00	75.14	80.00	110.00
FBS	62.0	88.5	98.0	119.2	130.0	400.0
CR	0.500	0.900	1.000	1.056	1.200	2.200
TG	37.0	90.0	122.0	150.3	177.0	1050.0
LDL	18.0	80.0	100.0	104.6	122.0	232.0
HDL	15.90	33.50	39.00	40.23	45.50	111.00
BUN	6.0	13.0	16.0	17.5	20.0	52.0
ESR	1.00	9.00	15.00	19.46	26.00	90.00
HB	8.90	12.20	13.20	13.15	14.20	17.60
K	3.000	3.900	4.200	4.231	4.500	6.600
Na	128	139	141	141	143	156
WBC	3700	5800	7100	7562	8800	18000
Lymph	7.0	26.0	32.0	32.4	39.0	60.0
Neut	32.00	52.50	60.00	60.15	67.00	89.00
PLT	25.0	183.5	210.0	221.5	250.0	742.0
EF.TTE	15.00	45.00	50.00	47.23	55.00	60.00

Table 11: Descriptive Statistics of CAD Data

Table 11 above is the sumamry statistics for the CAD data set.

Demographics metrics

- Most of the demographic metrics have a symmetric distribution. For instance, age, patients ate middle aged to elderly with a median of 58 and a weight range of 48kg to 120kg and a height if 140cm to 188cm. The median of 26.78 for the BMI also shows that most of the patients are overweight.

Cardiovascular metrics

- The BP has a mean of 139.6 and a median of 130 while the PR is mostly normal with a median of 70. EF has a median of 50%, with some patients showing a reduced heart function.

Blood Profile and Chemistry

- There are some elevated values for LDL (18 to 232, median: 100), HDL (15.9 to 111, median: 39), and FBS (62 to 400, median: 98) which could indicate cardiovascular risk factors. Additionally, triglycerides range from 37 to 1050 (median: 122) and creatinine ranges from 0.5 to 2.2 (median: 1.0), with some patients experiencing higher values that may suggest metabolic or kidney issues.

Other metrics

- The ESR also shows a broad range (1 - 90), with a median of 15, indicating a wide variety of inflammation level amongst patients. RWMA vary from 0 to 4, with a median of 0, indicating that most patients do not exhibit substantial heart motion difficulties, whilst a minority do. PLT range from 25 to 742 and a median of 210, with some abnormalities.

3.1.4 Data visualisation

The histograms below in Figure 18 depicts the distribution of the categorical variables. Most of the variables such as Obesity and Thyroid disease have 2 categories. There is also a visible imbalance in the distributions with one category being more high than others and for some variables like Sex, there is a relative balance between the categories.

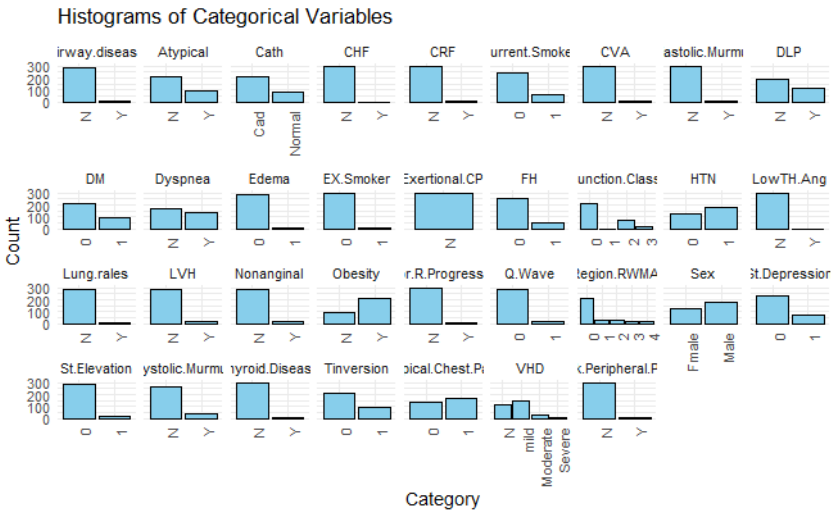


Figure 18: Histograms for categorical vaiables

Figure 19 below visualises the unscaled numeric variables giving insights into their distributions. Most of the variables have values close to 0 and only one variable WBC, has values ranging from 4000 to 16000. Many variables also have outliers have outliers.

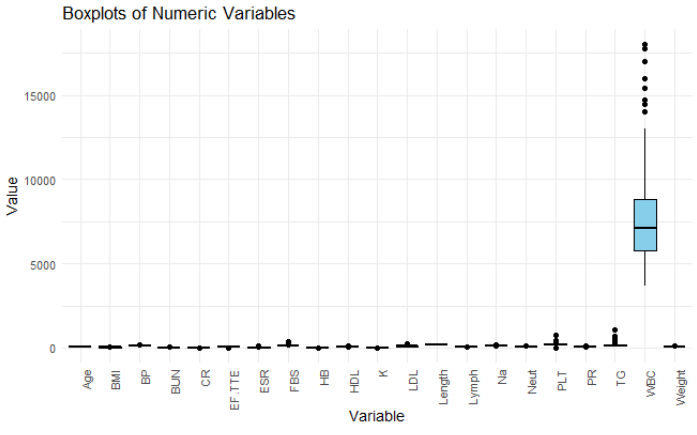


Figure 19: Box plots for numeric variables before scaling

3.1.5 Data standardisation

Figure 20 below shows the box plot for all the numeric variables after standardisation. Most variables are now centred around zero, which shows they are now standardised. There is also a relatively symmetric distribution amongst the majority of the variables like Age and BMI, whilst variable like ESR and ETTE are more skewed. Most of the variables have outliers and they will not be removed as they might provide some valuable insight about CAD diagnosis.

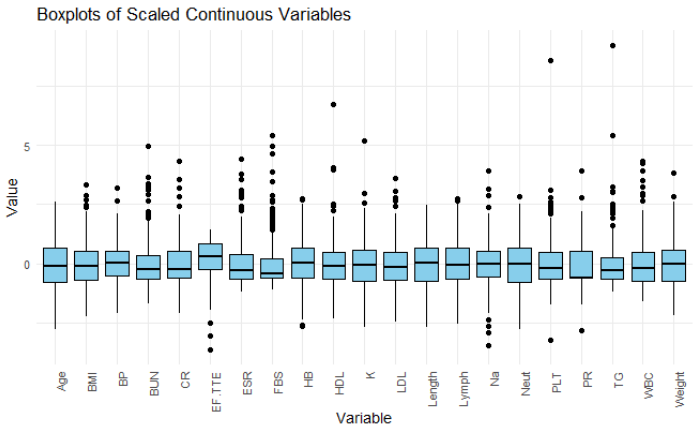


Figure 20: Box plots for numeric variables after scaling

3.2 Feature selection

The CAD data set has 56 features and to redundancy, Random Forest model was utilized to determine feature importance. This was done by fitting the model in the CAD dataset with the Cath Variable as the response variable. Figure 21 below highlights the top 23 features based on their mean decrease in Gini index, with the top 3 features being Typical chest pains, Age and Atypical. The mean decrease in Gini index is best for this analysis as it highlights how each feature is able to contribute to the model's ability to differentiate between classes. The choice to use only the top 23 features (half of the features) was to ensure a good balance between having enough features to capture the important parts of the cad dataset and to avoid having a complex model.

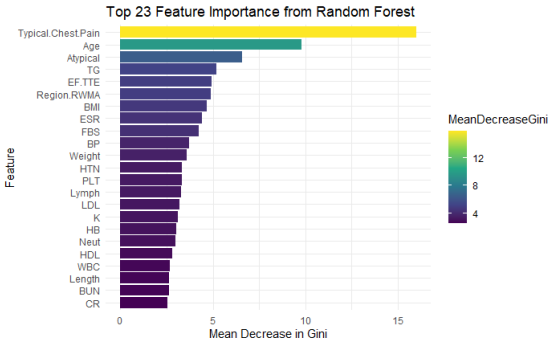


Figure 21: Feature importance using Random Forest

3.3 Association rule mining

3.3.1 Apriori Algorithm

Data preparation

To prepare the data for the Apriori algorithm, all the numeric variables were discretized into categorical bins to ensure the data is suitable for the mining process. This involved dividing the numeric columns into 5 discrete intervals and converting them to categorical factors. Table 12 below summarizes the most frequent items identified in these transactions with TG and PLT having 265 and 263 respectively, occurrences indicating a strong presence across transactions.

Item	**Occurrences**
TG=1	265
PLT=2	263
FBS=1	224
Region.RWMA=1	217
(Other)	6,303

Table 12: Most Frequent Items

Modelling and Parameter tuning

The default Apriori model was run on this data with 303 transactions and generated 23890 association rules. To reduce redundancy and focus on the most significant rules, parameters were refined to simplify the analysis.

Table 13 shows the parameters used. The support of 0.2 was used to ensure that only items that appear in at least 20% of the transactions are considered. As a result of the importance of this data set in determining Cardiovascular diseases, a higher confidence of 0.8 was used to ensure all the significant rules are supported by the data. The number of rules after parameter tuning reduced to 197 rules. To further refine the rules, redundant rules were further identified based on their confidence values and a total of 140 rules were found to be redundant. After removing the 140 redundant rules, a total of 57 rules were remaining and were sorted based on their lift values.

Parameter	**Value**
Minimum Length (minlen)	2
Maximum Length (maxlen)	10
Support	0.2
Confidence	0.8

Table 13: Parameters Used for Apriori model

Table 14 below shows the results of the top 6 rules from the Apriori algorithm, summarising how the certain symptoms (lhs) lead to the diagnosis of CAD (rhs). For example, the first rules shows that in 20.1% of the transactions where a patient shows signs of typical chest pains and over the age of 3 , CAD is predicted at a high confidence of 98.4%. The high confidence indicates how reliable the symptoms are at predicting CAD. Furthermore, the lift of 1.380 shows that there is a 38.1% of CAD occurring when a patient shows these symptoms than at random.

Similarly, the second rules demonstrates that when a patient has typical chest pains(1), PLT(2), and HTN(1), CAD is predicted at 97.8% and with a confidence of 29.4% of the transactions. Furthermore, the lift value of 1.372 shows that CAD is 37.2% more likely to happen when a patient has these symptoms.

From this summary, it is evident that some symptoms(features) are very influential at predicting CAD. The occurance of Typical chest pain (1) in all of the top 6 rules shows how important it is at diagnosing CAD. The other significant symptoms that patients of healthcare professionals should be aware of included PLT, EF.TTE and TG accompanied by typical chest pains.

lhs	rhs	support	conf	coverage	lift	count
{Typical.Chest.Pain=1, Age=3}	{Cath=Cad}	0.201	0.984	0.205	1.380	61
{Typical.Chest.Pain=1, PLT=2, HTN=1}	{Cath=Cad}	0.294	0.978	0.300	1.372	89
{Typical.Chest.Pain=1, Neut=3}	{Cath=Cad}	0.234	0.973	0.241	1.364	71
{Typical.Chest.Pain=1, EF.TTE=4}	{Cath=Cad}	0.218	0.971	0.224	1.362	66
{Typical.Chest.Pain=1, TG=1, BMI=2}	{Cath=Cad}	0.211	0.970	0.218	1.360	64
{Typical.Chest.Pain=1, HDL=2, HTN=1}	{Cath=Cad}	0.208	0.969	0.215	1.360	63

Table 14: Top 6 rules using Apriori

Figure 22 and Figure 23 below illustrate the network graph and the parallel coordinates plot for the 57 rules generated from the dataset.

From the Network graph(figure 22), the nodes represent the symptoms, with their size indicating how frequent they are in the rules. The dark red colour indicates a strong association between features, which is measured by the lift. For instance, the association between Typical Chest Pain = 1 and Cath shows a strong lift (big red node), suggesting that patients with typical chest pains are more likely to have CAD. On the other hand, there is a weak association between EF.TTE = (45,55) and CAD, as indicated by the low lift and a light red node.

Similarly in Parallel coordinates plot, the relationship between the features is represented by the lines which corresponds to the association rule. The stronger the support or frequency of the rule in the dataset, the darker the line. A thick, dark red line, indicating a frequent and strong link, indicates that variables such as TG and HDL have significant relationships across different positions. This suggests these factors play crucial roles in the rules generated from the dataset. There are also some variable such as HTN and K that changes positions alot, this could indicate that these symptoms had a big influence when the rules were generated.

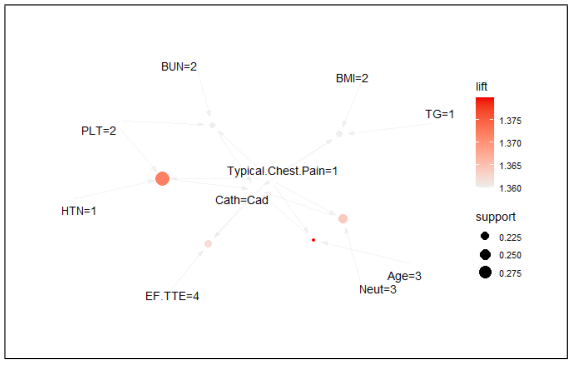


Figure 22: Network graph

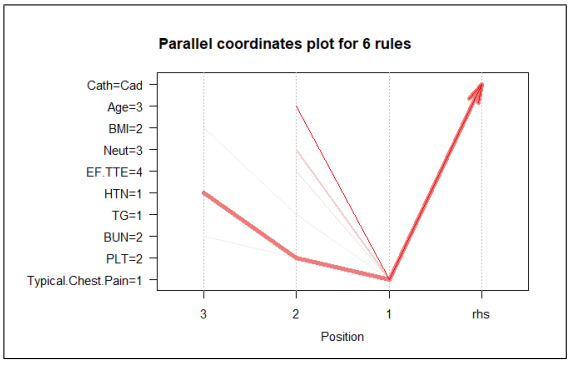


Figure 23: Parallel cooordinates plot

Summary

Overall, the Apriori algorithm was able to identify the key symptoms that are associated with CAD by providing invaluable insights and uncovering relationships.

3.3.2 Frequent Pattern Growth Algorithm

The second association rule mining algorithm explored was the Frequent Pattern (FP) Growth Algorithm. The FP growth algorithm is efficient in discovering frequent items sets and it differs from Apriori as it uses a compact tree structure to reduce time and costs associated with generating candidate items (Bala et al., 2016, p. 282).

Data Preparation

The FP grow model requires the same data preparation as the one used for the Apriori algorithm and therefore, the same results are the same as in table 9 above.

Modelling and Parameter Tuning

The FP-Growth algorithm was run using the rCBA package. The table below highlights the parameters used for the model. Similar to the Apriori algorithm, the support was set to 0.2 and the confidence was set to 0.8. This configuration was chosen to capture only the rules that appear in at least 30% of the transactions and have an 80% confidence level, meaning that only those rules with a high likelihood of occurrence were included. A total of 33 rules were generated of which 21 were redundant and only leaving us with 12 rules.

Parameter	**Value**
Support	0.2
Confidence	0.8

Table 15: Parameters Used for FP growth model

Table 16 below highlights the top 6 rules generated from the FP growth model. Similar to the Apriori above, one of the symptoms in the first rule is Typical chest pains but accompanied by HTN. This indicates that in 33.7% of the transactions were a patient displays these symptoms, CAD is predicted with a confidence of 96.2%. The lift of 1.350 suggests that there is a 35% chance that a patient will have CAD when these symptoms are present. Some of the symptoms that are influential to the predicting of CAD in patients are HDL, PLT and HTN.

All the top 6 rules had high confidence ranging form 88.1% to 96.2%. This suggests that there is a strong possibility that a patient will be diagnosed with CAD when these symptoms or a combination of these symptoms are present. Additionally, the high lift values also reinforces the significance of these symptoms. Overall, the FP model is effectively capturing strong patterns in the data.

lhs	**rhs**	**support**	**confidence**	**lift**
{Typical.Chest.Pain=1, HTN=1}	{Cath=Cad}	0.337	0.962	1.350
{PLT=2, Typical.Chest.Pain=1}	{Cath=Cad}	0.436	0.950	1.332
{Typical.Chest.Pain=1, HDL=2}	{Cath=Cad}	0.320	0.942	1.321
{Typical.Chest.Pain=1}	{Cath=Cad}	0.508	0.939	1.317
{PLT=2, Atypical=N, HTN=1}	{Cath=Cad}	0.340	0.896	1.256
{Atypical=N, HTN=1}	{Cath=Cad}	0.389	0.881	1.235

Table 16: Top 6 Rules using FP Growth

Figures 24 and 25 below shows the network graph and parallel coordinates plot generated for the 12 rules. From the network graph, there is a strong association between CAD and HTN(1), Typical Chest pains and PLT(2), this is indicated by the darker and larger nodes, highlighting higher

support and lift values. Typical chest pains and HTN are some of the frequently seen symptoms in patients as highlighted by the high lift values.

Similarly in the Parallel coordinates plot, when a patient has symptoms such as HTN = 2, and Typical Chest Pains =1, there is a high probability that they will have CAD. Both the Network graph and the Parallel coordinates plot have HTN and Typical Chest pains as the strongest predictors of CAD.

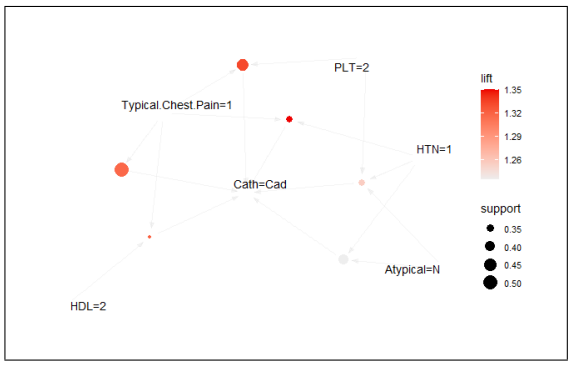


Figure 24: Network graph

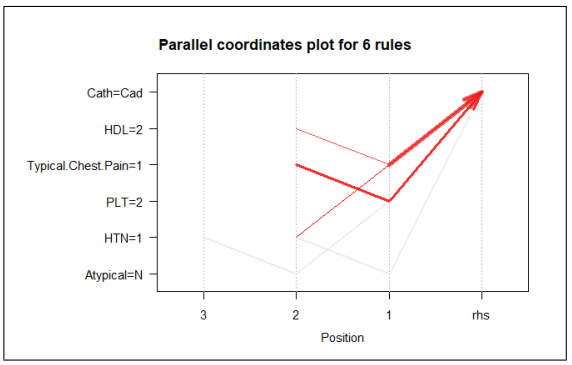


Figure 25: Parallel cooordinates plot

3.3.3 Equivalence Class Clustering and bottom-up Lattice Traversal Algorithm (ECLAT)

The final association rule mining technique explored in this paper was the Eclat algorithm. Unlike the Apriori algorithm, the Eclat uses the vertical data format to find frequent items

Modelling and Parameter Tuning

In order to identify frequent item-sets that are related to CAD, the Eclat algorithm was employed with a support of 0.2 and a minimum length of 2, as reflected in Table 17 below. The support of 0.2 was chosen to ensure the model is able to identify frequent items combinations that are highly associated with CAD without being overwhelmed by irrelevant patterns.

Parameter	Value
Support	0.2
Minlen	2

Table 17: ECLAT Parameters

The ECLAT algorithm initially generated a total of 1279 item-sets, which were narrowed down to 363 by focusing item-sets that included CAD-related features. Table 18 below presents the top 6 item-sets based on the support. From the table, it is evident that item-sets such as TG and PLT have a strong association with CAD, as can be seen by their support values of 0.531. This suggests that these are some of the combination of symptoms that are frequently observed in CAD analysis.

Some symptoms such as Typical chest pains and Atypical appear in most of the item-sets, this indicate the relevance of these symptoms in CAD. Monitoring of these symptoms in patience could provide good insights in early detection and diagnosis of CAD.

To further understand the relationships between symptoms and CAD, we visualized the ECLAT results using a network graph (Figure 26) and a parallel coordinates plot (Figure 27). From

Itemset	Support	Count
{TG=1, PLT=2, Cath=Cad}	0.531	161
{Typical.Chest.Pain=1, Atypical=N, Cath=Cad}	0.508	154
{Atypical=N, PLT=2, Cath=Cad}	0.498	151
{Atypical=N, TG=1, Cath=Cad}	0.495	150
{Typical.Chest.Pain=1, Atypical=N, TG=1, Cath=Cad}	0.439	133
{Typical.Chest.Pain=1, TG=1, Cath=Cad}	0.439	133

Table 18: Top 6 Itemsets associated with CAD

the network graph, certain itemsets, such as TG, PLT, and atypical chest pain, appear more frequently, suggesting these symptoms are often present in patients with CAD. The strength of these associations is indicated by the size and positioning of the nodes within the graph.

The parallel coordinates plot emphasizes these relationships, providing a clearer comparison across the itemsets. Features like typical chest pain and atypical chest pain are shown to occur frequently, reinforcing their significance in the context of CAD. These visualizations collectively highlight the importance of monitoring these symptoms as they frequently co-occur with CAD, potentially aiding in early detection and diagnosis.

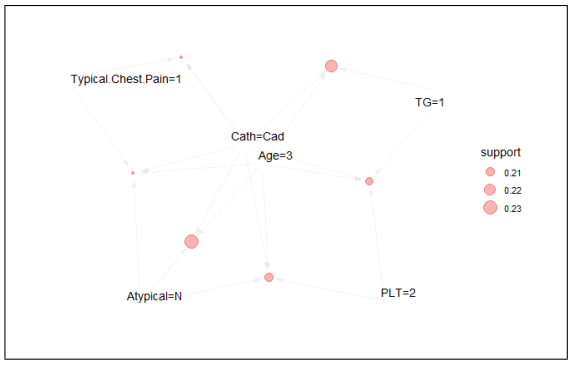


Figure 26: Network graph

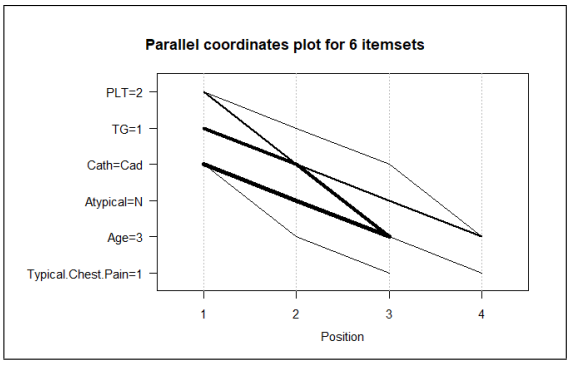


Figure 27: Parallel cooordinates plot

Overall, the visualisations in Figure 25 and 26 reinforces the initial findings in table 15. Typical chest pain still remains as the most frequently appearing item in the CAD-related item sets followed by HTN. PLT and TG also appear as important sympotms as they have significant nodes in the network graph and the parallel coordinates plot.

3.4 Model Evaluation

Algorithm	N. Rules	Support	Confidence	Lift
Apriori	91	0.268	0.901	1.263
FP-Growth	18	0.420	0.883	1.239
ECLAT	709	0.260	NA	NA

Table 19: Comparison of rule algorithms

Table 19 above summarises he metrics of the 3 association rule mining algorithms explored in the previous sections.

Number of rules

- Eclat produced the highest number of frequent item-sets, 709. This could mean that it has a wide coverage in the dataset. FP growth had 18 rules, which could mean that it was more restrictive and Apriori on the other had had 91 rules, and as compared to the FP growth there is a possibility that it captured effective patterns and could also mean it might have captured a lot of noise as well.

Support

- Apriori and Eclat had values close to each other, 26.8% and 26.0%, which means about 27% of the dataset contains patterns that were generated by these algorithms. FP growth on the other hand had support of 42% which was the highest amongst the three.

Confidence and Lift

- Apriori had confidence of 90.1% and a lift value pf 1.263. The high confidence values suggest that it was able to generate strong rules and that 90% of the symptoms on the LHS resulted in the occurrence of CAD. The high lift also suggest that there is a strong relationship between the symptoms(LHS) generated and the occurrence of CAD.
- FP growth had a high confidence of 88.3% which is slightly lower than Apriori and a Lift value of 1.239 which also suggest the high likelihood of the symptoms on the LHS being good predictors of CAD than if it was at random.

Summary

The Apriori provided a good balance between the number of rules, confidence and lift values. This makes it the best algorithm for generating reliable rules and not really being influenced by noise. For Eclat, the high number of rules and support value which is almost the same as for the Apriori algorithm shows that it was capturing a good coverage of the item-sets but as it does not use confidence and lift, it is difficult to compare the quality of the item sets against the other algorithms. And lastly, the low values for rules for FP growth, accompanied by high confidence a lift values could be a good indicator of the quality of the rules that were generated.

3.5 Comparison to previous work

This section will compare the highlights from R. Alizadehsani et al. (2013), "A Data Mining Approach for Diagnosis of CAD" with the findings from my models. The goal of Alizadehsani et al. was to develop a model for diagnosing CAD using data mining techniques, while the objective of this paper was to apply association rule mining.

Key Insights

Dataset, Model and Feature selection

- The dataset used for this paper was the same in Z. Alizadeh's study. Due to the data imbalance, Alizadeh applied different SMOTE algorithms to balance the Cath category.
- Random forest was used in this paper to select the top 23 features for association rule mining, whilst Alizadeh used LightGBM algorithm and the Logistic Regression model (The best amongst all that were evaluated).

Evaluation metrics

- Alizadehsani et al. used metrics such as accuracy, F1 score, and AUC. In contrast, I used confidence, lift, and support for evaluation.

Findings (Symptoms)

- Alizadehsani 's paper frequently mentioned typical chest pain as one of the symptoms for CAD, which aligns with my findings. In my analysis, typical chest pain appears in most rules generated by the Apriori and FP-Growth algorithms, as well as in many item sets identified by the Eclat model.

3.6 Summary

To conclude this section, it is evident that the association rule mining algorithms (Apriori, FP growth and Eclat) were successful in identifying the key features/symptoms that are mostly associated with CAD. From the results, the most common symptoms proved to be Typical chest pains, HPT and PLT which were consistent in all the three models.

Overall, the use of feature selection (Random forest) and the association rule mining techniques provided invaluable insight into the CAD dataset by uncovering patterns and relationships between the features. This information will play a pivotal role in assisting researchers make early diagnosis and treatment for CAD.

4 Conclusion

In this report, two datasets were successfully analyzed using two unsupervised learning techniques to investigate the monitoring of fetal health and the diagnosis of CAD. For the fetal health, the cluster analysis confirmed that the dataset can be grouped into three distinct classes: Normal, Suspect, and Pathological. And in the case for coronary artery disease, association rule mining revealed that typical chest pains, hypertension (HPT), and platelet count (PLT) were the most influential features associated with CAD.

Overall, the results of these analyses provided a good insight that can be used in improving the health care sector by accurately monitoring the health status of the fetal and detecting early symptoms of coronary artery disease.

5 References

1. Alizadehsani, R., et al., 2013. 'A data mining approach for diagnosis of coronary artery disease', *Computer Methods and Programs in Biomedicine*, 111(1), pp. 52-61.
2. Bala, A., Shuaibu, M.Z., Lawal, Z.K. and Zakari, R.Y., 2016. 'Performance Analysis of Apriori and FP-Growth Algorithms (Association Rule Mining)', *International Journal of Computer Technology Applications*, 12(3), pp. 270-290.
3. James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013. *An introduction to statistical learning: with applications in R*. Springer Texts in Statistics.
4. Medical News Today, 2023. Coronary artery disease (CAD): Symptoms, causes, and treatment. Retrieved from `https://www.medicalnewstoday.com/articles/184130` (Accessed: 13 September 2024).
5. Zhao, X., Yang, J., Zhang, X. and Liu, Q., 2021. 'A novel clustering method based on fuzzy C-means and genetic algorithm', *Soft Computing*, 25(9), pp. 1-15. Available at: `https://link.springer.com/article/10.1007/s10726-021-09758-7` (Accessed: 28 August 2024).

6 Appendix A

Table 20: Description of Variables in CAD Dataset

Variable Name	**Description**	**Variable Type**
Age	Age of the patient in years	Numeric
Weight	Weight of the patient in kg	Numeric
Length	Height of the patient in cm	Numeric
Sex	Gender of the patient	Categorical (Character)
BMI	Body Mass Index of the patient	Numeric
DM	Diabetes Mellitus (0: No, 1: Yes)	Numeric
HTN	Hypertension (0: No, 1: Yes)	Numeric
Current Smoker	Current smoking status (0: No, 1: Yes)	Numeric
EX-Smoker	Ex-smoker status (0: No, 1: Yes)	Numeric
FH	Family history of heart disease (0: No, 1: Yes)	Numeric
Obesity	Obesity status (Y: Yes, N: No)	Categorical (Character)
CRF	Chronic renal failure (Y: Yes, N: No)	Categorical (Character)
CVA	Cerebrovascular accident (Y: Yes, N: No)	Categorical (Character)
Airway disease	Airway disease (Y: Yes, N: No)	Categorical (Character)
Thyroid Disease	Thyroid disease (Y: Yes, N: No)	Categorical (Character)
CHF	Congestive Heart Failure (Y: Yes, N: No)	Categorical (Character)
DLP	Dyslipidemia (Y: Yes, N: No)	Categorical (Character)
BP	Blood pressure (mmHg)	Numeric
PR	Pulse rate (bpm)	Numeric
Edema	Edema presence (0: No, 1: Yes)	Numeric
Weak Peripheral Pulse	Weak peripheral pulse (Y: Yes, N: No)	Categorical (Character)
Lung rales	Lung rales (Y: Yes, N: No)	Categorical (Character)
Systolic Murmur	Systolic murmur (Y: Yes, N: No)	Categorical (Character)
Diastolic Murmur	Diastolic murmur (Y: Yes, N: No)	Categorical (Character)
Typical Chest Pain	Typical chest pain (0: No, 1: Yes)	Numeric
Dyspnea	Dyspnea (shortness of breath) (Y: Yes, N: No)	Categorical (Character)
Function Class	Functional class (0-4) based on symptom severity	Numeric
Atypical	Atypical chest pain (Y: Yes, N: No)	Categorical (Character)
Nonanginal	Nonanginal chest pain (Y: Yes, N: No)	Categorical (Character)
Exertional CP	Exertional chest pain (Y: Yes, N: No)	Categorical (Character)
LowTH Ang	Low-threshold angina (Y: Yes, N: No)	Categorical (Character)

Variable Name	**Description**	**Variable Type**
Q Wave	Presence of Q wave (0: No, 1: Yes)	Numeric
St Elevation	ST segment elevation (0: No, 1: Yes)	Numeric
St Depression	ST segment depression (0: No, 1: Yes)	Numeric
Tinversion	T wave inversion (0: No, 1: Yes)	Numeric
LVH	Left ventricular hypertrophy (Y: Yes, N: No)	Categorical (Character)
Poor R Progression	Poor R wave progression (Y: Yes, N: No)	Categorical (Character)
BBB	Bundle branch block (Y: Yes, N: No)	Categorical (Character)
FBS	Fasting blood sugar (mg/dL)	Numeric
CR	Creatinine (mg/dL)	Numeric
TG	Triglycerides (mg/dL)	Numeric
LDL	Low-density lipoprotein cholesterol (mg/dL)	Numeric
HDL	High-density lipoprotein cholesterol (mg/dL)	Numeric
ESR	Erythrocyte sedimentation rate (mm/hr)	Numeric
HB	Hemoglobin (g/dL)	Numeric
K	Potassium (mEq/L)	Numeric
Na	Sodium (mEq/L)	Numeric
Lymph Platelet count (per µL)	Lymphocyte percentage (Neut Numeric	Neutrophil percentage (PLT
EF-TTE	Ejection fraction (percent) from TTE	Numeric
Region RWMA	Regional wall motion abnormalities (0-4)	Numeric
VHD	Valvular heart disease (N: No, mild, Severe)	Categorical (Character)
Cath	Angiographic results (CAD: CAD, Normal)	Categorical (Character)