University of Cape Town

Course Code

STA5077Z

Cluster analysis and association rule mining

Author:
Lungile P. Nkuna

Student Number:
NKNPRA001

September 15, 2024

Plagiarism Declaration I, Prayer L. Nkuna, hereby declare that the work on which this documentary is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university. I authorize the University to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

Signature: P.L.Nkuna.

Date: 29 August 2024

Contents

1 Overview

2 Cluster Analysis

2.1 Introduction

2.2 Data Wrangling

In the data wrangling phase, the histogram tendency variable was converted into a factor in order to obtain the proportion of the number of foetal health cases in each class (Appendix A, Figure).This was done to enhance the categorisation of the histogram tendency variable. for instance, the values -1. 0 and 1 represent three classes namely normal, suspect and pathological not respectively. Additionally, missing values were checked using the is.na command and the data had zero missing values. following the missing values, the duplicates were checked in this data and it was found that there are 28 observation with exact duplicate values. However since this is a medical data obtained through examinations and some of these values like accelerations, fetal movement, uterine contractions, light deccelerations, severe decceleratins and prolonged deccelerrations being obtained in second intervals along with other values being obtained within a short period of time, these duplicates were not removed. following these steps, box plots along with a table of summary statistics were obtained to variations in the raw data and histograms were also generated to understand how the data was distributed.furthermore, a correlation matrix was also generated to understand the relationship between numeric variables in the data. Observing how the raw data had variabilities and unscaled, the data was standardized to ensure that the results obtained are not obscured and misleading. following the standarsations of the data, PCA was done to further transform the data by reducing the variables and creating three new representative variables that explains most of the variability in the original dataset. Through these sequential steps, a comprehensive understanding and transformation of the datset was achieved laying the groundwork for subsequent analysing and modelling.

2.3 Exploratory Data Analysis

2.3.1 Description of data

small paragraph on the data

Variable name	**Description of varibale**	**Variable type**
baseline value	Baseline FHR (beats per minute)	Metric
accelerations	Number of accelerations per second	Metric
fetal movement	Number of fetal movements per second	Metric
uterine contractions	Number of uterine contractions per second	Metric
light decelerations	Number of light decelerations per second	Meric
severe decelerations	Number of severe decelerations per second	Metric
prolonged decelerations	Number of prolonged decelerations per second	Metric
abnormal short-term variability	Percentage of time with abnormal short-term variability	Metric
mean value of short-term variability	Average value of short-term variability	Metric
percentage of time with abnormal long-term variability	Time long-term variability outside normal range	Metric
mean value of long-term variability	Average value of long-term variability	Metric
histogram width	Width/range of FHR histogram (generated from the exam)	Metric
histogram min	Minimum of FHR histogram(from the exam)	Metric
histogram max	Minimum of FHR histogram(from the exam)	Metric
histogram number of peaks	Number of FHR histogram peaks (from the exam)	Metric
histogram number of zeroes	Number of FHR with zeroes (from the exam)	Metric
histogram mode	Most FHR values (generated from the exam)	Metric
histogram mean	Average FHR values (generated from the exam)	Metric
histogram median	FHR Values at center of sorted values	Metric
histogram variance	Variability in FHR values	Metric
histogram tendency	Direction of FHR changes	Nominal

Table 1: Description and type of data variables used, where FHR stands for foetal heart rate.

2.3.2 Unscaled Data

Variable	Min	Median	Mean	Max	SD
Baseline value	106.0	133.000	133.30	160.000	9.8408
Accelerations	0.0	0.002	0.00	0.019	0.0039
Fetal movement	0.0	0.000	0.01	0.481	0.0467
Uterine contractions	0.0	0.004	0.00	0.015	0.0029
Light decelerations	0.0	0.000	0.00	0.015	0.0030
Severe decelerations	0.0	0.000	0.00	0.001	0.0001
Prolonged decelerations	0.0	0.000	0.00	0.005	0.0006
Abnormal short-term variability	12.0	49.000	46.99	87.000	17.1928
Mean value of short-term variability	0.2	1.200	1.33	7.000	0.8832
% time abnormal long-term variability	0.0	0.000	9.85	91.000	18.3969
Mean value of long-term variability	0.0	7.400	8.19	50.700	5.6282
Histogram width	3.0	67.500	70.45	180.000	38.9557
Histogram min	50.0	93.000	93.58	159.000	29.5602
Histogram max	122.0	162.000	164.03	238.000	17.9442
Histogram number of peaks	0.0	3.000	4.07	18.000	2.9494
Histogram number of zeroes	0.0	0.000	0.32	10.000	0.7061
Histogram mode	60.0	139.000	137.45	187.000	16.3813
Histogram mean	73.0	136.000	134.61	182.000	15.5936
Histogram median	77.0	139.000	138.09	186.000	14.4666
Histogram variance	0.0	7.000	18.81	269.000	28.9776
Histogram tendency	-1.0	0.000	0.32	1.000	0.6108

Table 2: Summary statistics for unstandardised data

The baseline value shows a near-normal distribution, with a mean (133.30), median (133.00), and low standard deviation (9.84), indicating stable readings (Table 2). This is reinforced by the histogram, which peaks around the same range (Appendix A, Figure). In contrast, variables such as accelerations, fetal movement, and uterine contractions are right-skewed, with most values near zero and rare high outliers (Table 2). For example, fetal movement has a maximum of 0.481 but a very low mean (0.01) and standard deviation (0.0467), indicating infrequent but significant occurrences (Table 2). Greater variability is seen in features like abnormal short-term variability (SD = 17.19) and percentage of time with abnormal long-term variability (SD = 18.39), which suggests wider fluctuations in these metrics (Table 2). The histograms for these variables reflect this broader spread (Appendix A, Figure). Lastly, Table 2 also shows that histogram-derived statistics show a nearly normal distribution for central tendency measures, such as mean (134.61), median (136), and standard deviation (15.59), indicating consistent central values in the data's histogram representation (Appendix A, Figure). However, variables such as histogram variance show significant right skewness and greater dispersion, with a mean of 18.81 and a large standard deviation of 28.97, implying considerable variability across the dataset (Table 2). This mix of skewed and normal distributions highlights the contrast between stable baseline measures and occasional extreme events in the dataset.

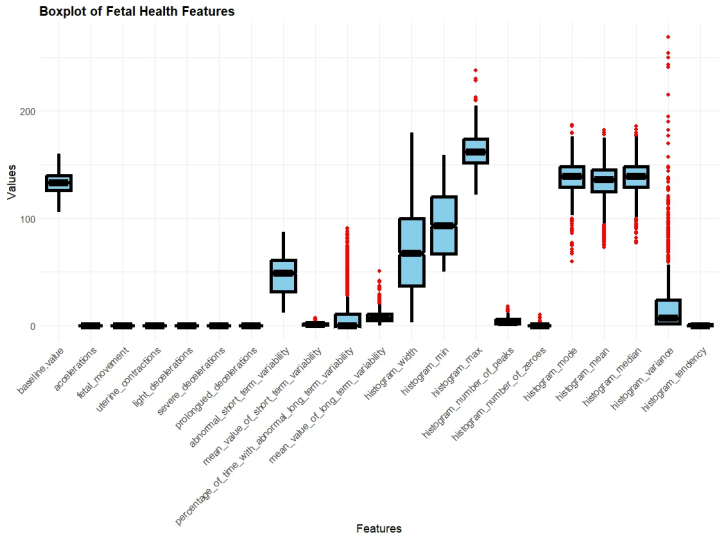


Figure 1: Pre standardised Data

In the boxplot of unstandardised foetal health features, a wide range of values across different features are evident, with some features having vastly different scales. For example, the baseline value and histogram-related features like histogram mode, mean, and median have much higher spread (values ranging between 110 and 150) compared to other features like accelerations, foetal movement, and uterine contractions, which have a little variation and small ranges with values close to zero. There are several features with a significant number of outliers, such as abnormal short-term variability and prolonged decelerations, which could represent critical clinical events or anomalies in the data.

Overall, this variability across different features suggests that scaling (e.g., standardisation or normalisation) is necessary to bring all variables onto a comparable scale before applying clustering/machine learning algorithms.

2.3.3 Scaled Data

The boxplot of standardized fetal health features in Figure 3 offers insight into the distribution of different variables after scaling. Standardization has transformed the features to have a mean of 0 and a standard deviation of 1, helping eliminate the impact of differing scales across variables. important clinical insights or potential noise. Overall, the boxplot demonstrates the effects of standardization and highlights variability and outliers within specific features, suggesting possible areas for further analysis or transformation.

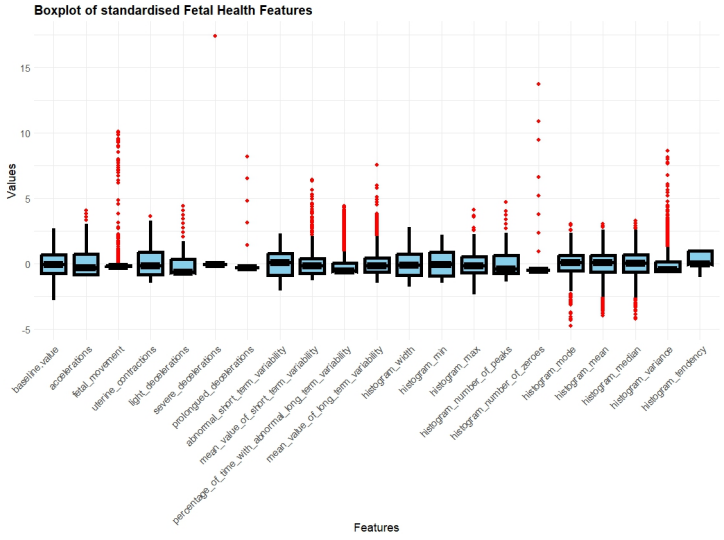


Figure 2: unscaled boxplots

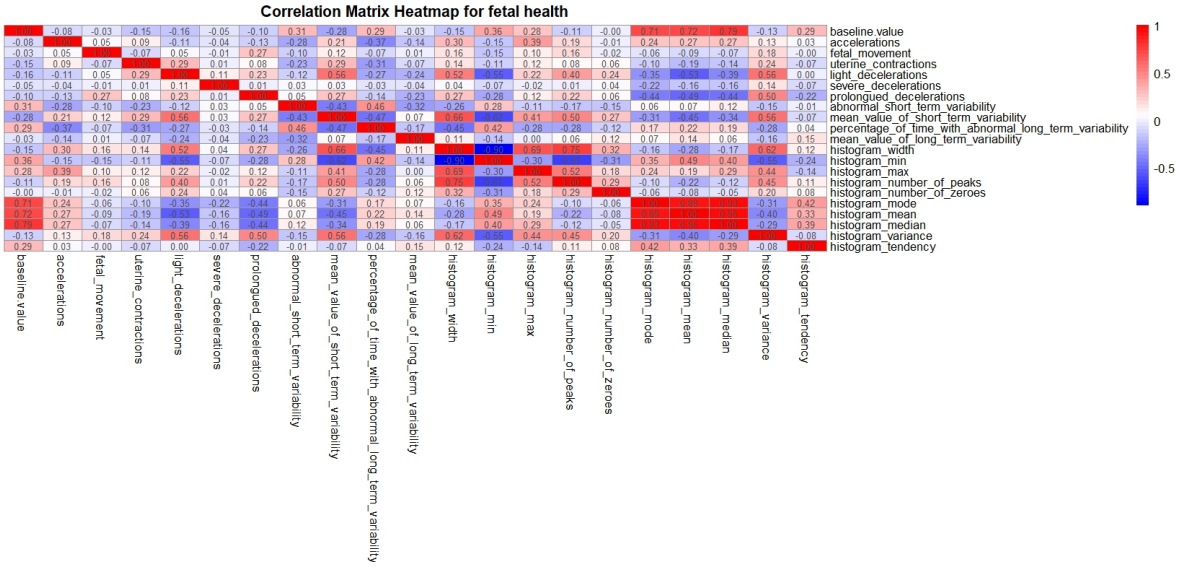


Figure 3: correlation

The correlation matrix heatmap (Figure 4) provides a detailed view of the relationships between foetal health features. The colour scale ranges from dark blue (strong negative correlations) to bright red (strong positive correlations), with white representing near-zero correlations. Several important relationships stand out. For instance, baseline value is strongly positively correlated with histogram mean (0.77), histogram mode (0.74), and histogram median (0.71), suggesting multicollinearity among these features.

Similarly, histogram variance is highly correlated with these features, indicating redundancy in the information they provide. Additionally, the mean value of short-term variability and the percentage of time with abnormal long-term variability has a strong positive correlation (0.75), indicating that both capture similar aspects of variability in foetal heart rates. On the other hand, features like uterine contractions, foetal movement, and accelerations exhibit low or near-zero correlations with most other variables, which suggests they may provide independent and unique information, making them valuable for predictive modelling.

The strong correlations between several histogram metrics and baseline values highlight the need for dimensionality reduction or feature selection techniques such as PCA (principal component analysis) to avoid multicollinearity in predictive modelling. This heatmap is a crucial tool for identifying relationships that could affect model performance and guiding further steps in feature engineering.

2.4 Data Transformation

2.4.1 PCA

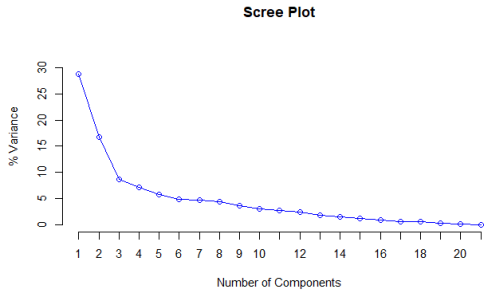


Figure 4: The scree plot used for PCAs.

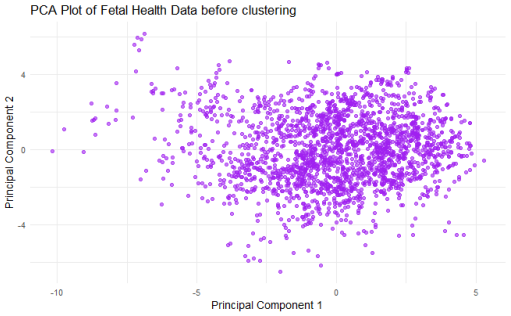


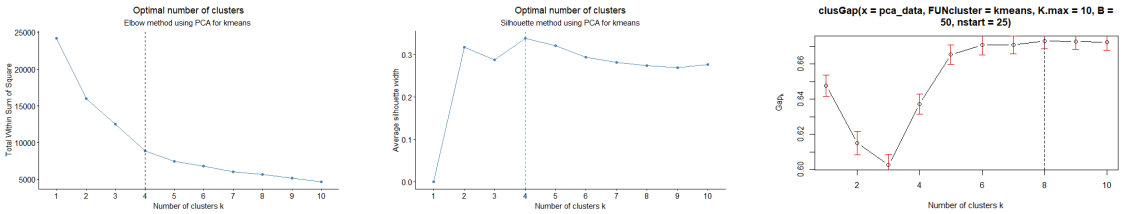
Figure 5: Structure of the data

Based on the elbow criterion, the scree plot (Figure 5) shows that the first three components might be optimal since they retain approximately 60 % of the total variance (35%, 15%, and 10% respectively). Beyond this point, additional components contribute significantly less variance, indicating that they add little meaningful information. Therefore, using the first three components would capture most of the data's structure (Figure 6), while including more components may be unnecessary for analysis. As a result, the first 3 components are chosen and they will be used throughout the analysis of this project.

2.5 Analysis

2.5.1 Distance based methods: Partitioning algorithms and Hierarchical algorithms

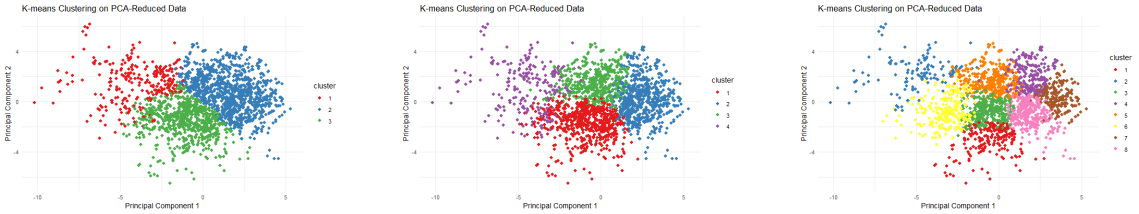
Kmeans



K =3 K =4 K = 8

Figure 6: Determining optimal number of clusters (K) for Kmeans

According to the Elbow Method, the best optimal K is 4 with a WSS (Within-Cluster Sum of Squares) of approximately 25000, which means adding more clusters beyond K=4 does not significantly reduce the WSS which must be kept as small as possible since it measures compactness, suggesting diminishing returns on further partitioning. Using the Silhouette Method, the optimal K is also 4, with a silhouette width of the range 0.34-0.35, indicating that clusters are not dense and well-separated. Lastly, the Gap Statistic suggests an optimal K of 8, with a value of 0.675, which indicates that 8 clusters are significantly better than a random uniform distribution.

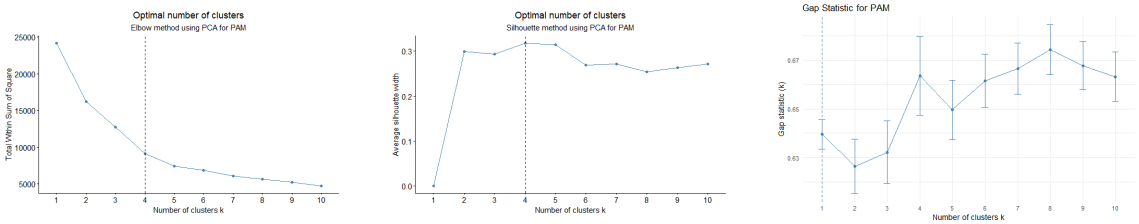


K =3 K =4 K = 8

Figure 7: kmeans clustering plots showing the distribution of clusters in PCA 2-dimensional space

The distribution analysis of KMeans clusters in fetal health data shows that K = 4 is the most compact, indicating homogeneous and well-defined clusters. This suggests that fetal health conditions are effectively grouped into four categories with minimal overlap. K = 8 captures more detail but results in less compact clusters with increased overlap, making it harder to differentiate between conditions and potentially leading to confusion. K = 3 provides well-separated clusters but lacks compactness, potentially combining diverse health conditions within the same cluster. Overall, K = 4 strikes the best balance between compactness and separation, making it the most reliable for identifying distinct fetal health outcomes.

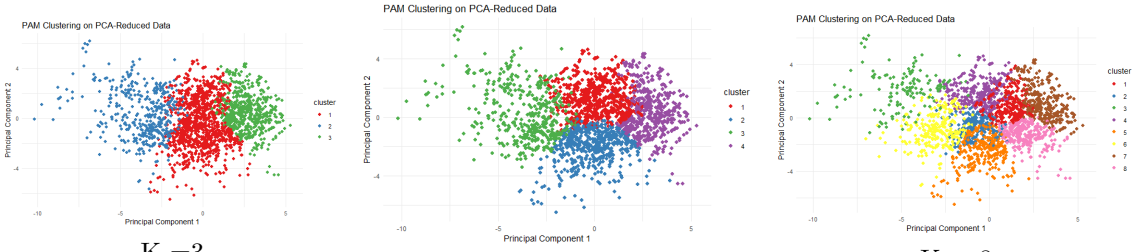
PAM



K =4 K =4 K = 8

Figure 8: Determining the optimal value of K using PAM

The optimal number of K in PAM (Figure 9) is the same as K-means, with K = 4 identified by the Elbow Method (WSS 25000), K = 8 using the Gap Statistic (0.68), and K = 4 using the Silhouette Method. However, PAM's silhouette method shows a silhouette width of range 0.31-0.33, which is slightly lower compared to K-means, suggesting that PAM's clusters are defined by actual data points (medoids), and they do not capture central tendency as effectively as K-means, which uses the mean of the points.

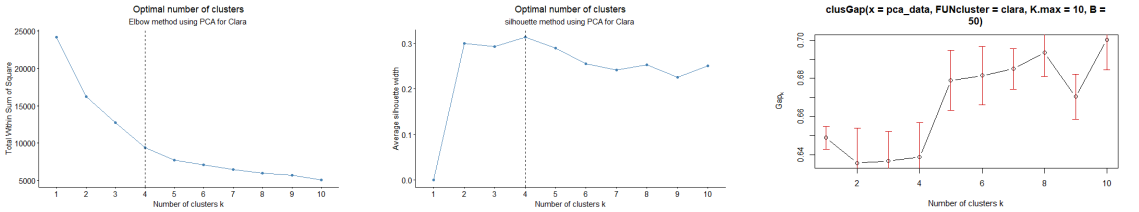


K =3 K =4 K = 8

Figure 9: PAM clustering plots showing the distribution of clusters in PCA 2-dimensional space

The distribution using PAM (Figure 10) reveals that When k is 4, the clusters are well-defined and capture meaningful distinctions. At K = 3, the clusters are more distinct but lack finer details, potentially missing subtle variations in health classes. At K = 8, the clusters become overly scattered and less compact, suggesting over-segmentation. Overall, the K = 4 provides the most effective clustering for clear differentiation of fetal health classes' and it shows interpretability.

CLARA

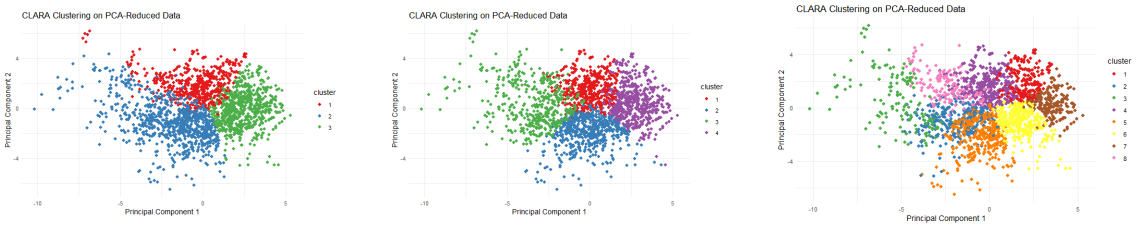


K =4 K =4 K = 8

Figure 10: Determining the optimal value of K using CLARA

The optimal number of K determined by the Elbow Method, Silhouette Method, and Gap Statistic

is consistently K = 4, k = 4, and k = 8, respectively, across K-Means, PAM, and CLARA, with WSS and Gap Statistic values of approximately 2500 and approximately 0.68 (respectively) aligning closely among these methods. However, while K-means and PAM yield similar silhouette widths, CLARA shows a lower silhouette width ranging from approximately 0.30–0.31. This lower value for CLARA is due to its use of sampling, which can lead to less distinct cluster separation compared to the full-data approaches of K-Means and PAM.



K =3 K =4 K = 8

Figure 11: Distribution and size of clusters in Clara clustering algorithm

The CLARA method, like PAM, uses medoids for clustering but incorporates data sampling, which can introduce variations in cluster definitions. For fetal health data, this affects the compactness and separation of clusters (as shown in Figure 13) compared to PAM's full-data clusters (Figure 10). The K = 4 scatter plot provides the most effective clustering, balancing compactness and separation to differentiate between fetal health states (normal, suspect, and pathological). In contrast, K = 3 clusters are distinct but less detailed, while K = 8 clusters are scattered and less compact. Therefore, K = 4 is the optimal choice for CLARA, effectively representing the data.

For the fetal health dataset, K-Means is the most effective clustering method, with an optimal number of 4 clusters and the highest silhouette width of 0.34. It provides better-defined clusters than PAM and CLARA, which have slightly lower silhouette scores due to their methods and sampling variations. K-Means is the best choice for distinguishing between fetal health states for this data.

Hierarchical Clustering algorithms

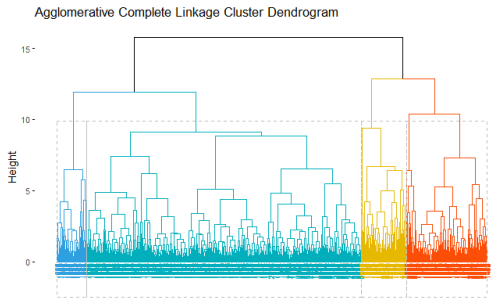


Figure 12: Complete linkage dendogram

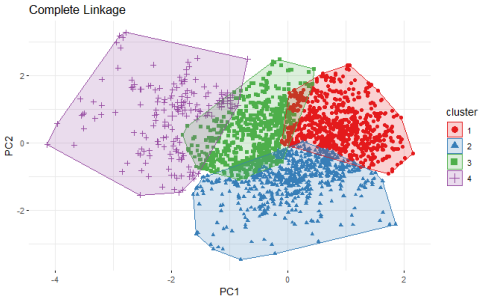


Figure 13: Distribution of data using complete linkage

The average linkage dendrogram (Appendix A, Figure) reveals clusters formed by averaging distances between points in different clusters, creating moderately dense groupings. This clustering method allows some overlap, which could be suitable for foetal health classification, especially when there is a gradual transition between categories. However, in the scatter plot (Appendix A, Figure..), clusters exhibit some overlap, reducing clarity in the separation between categories, which could make it harder to distinguish between different foetal health outcomes effectively.

The centroid linkage dendrogram (Appendix A, Figure...) forms clusters by calculating the distance between each cluster's centroids. While this method can identify broader groupings in fetal health data, the clusters are looser compared to complete linkage. The scatter plot (Appendix A, Figure.) reflects some overlap between clusters, which can blur the lines between health categories. However, centroid linkage may still work well when foetal health data has more variability or a general trend rather than strict classification.

The single linkage dendrogram (Appendix A, Figure.) shows elongated clusters due to the chaining effect, where clusters are formed based on the nearest neighbouring points. For foetal health data, this method is not ideal as it artificially connects distinct health conditions, potentially confusing healthy, at-risk, and pathological categories. The single linkage scatter plot (Appendix A, Figure.) confirms this, displaying significant overlap between clusters, which undermines its effectiveness in clearly differentiating between foetal health statuses.

The median linkage dendrogram (Appendix A, Figure.) offers a middle ground between average and centroid linkage methods, resulting in more compact clusters without the chaining effect seen in a single linkage. For foetal health data, this method provides moderately clear distinctions between categories. The scatter plot (Appendix A, Figure.) shows tighter clusters compared to the average and single linkage, with less overlap. While it's an improvement, the separation is still not as distinct as complete linkage, making it a less ideal choice for applications requiring precision in health classification.

The complete linkage dendrogram (Figure 15) produces the most tightly packed clusters, ensuring minimal variance within each cluster. In the context of foetal health, this method ensures distinct separations between health conditions, such as normal, at-risk, and pathological categories. The scatter plot (Figure 16) reflects this, with well-defined, non-overlapping clusters. This makes it highly suitable for datasets where precise classification of health outcomes is crucial, allowing for better differentiation between health conditions.

Among the linkage methods analysed, complete linkage stands out as the most effective for foetal health data. It produces well-defined, non-overlapping clusters both in the dendrogram and scatter plot (Figures 15 and 16), which is essential for clearly separating different health outcomes.

2.5.2 Density-based methods : DBSCAN

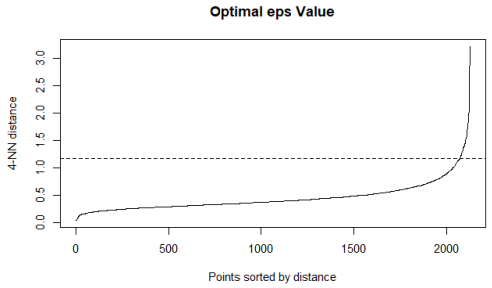


Figure 14: K-distance plot: eps= 1.2

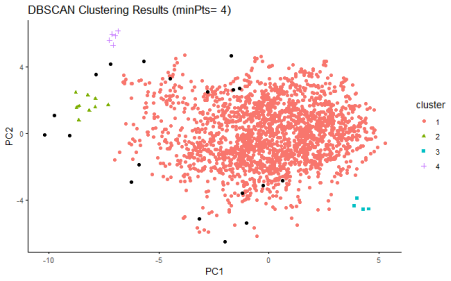


Figure 15: DBSCAN plot

The eps value is crucial for DBSCAN clustering in fetal health data analysis. The optimal eps value is around eps = 1.2 (figure 17), with MinPts of 4. This MinPts was chosen because the data has three components, and the rule of thumb states we should determine k by adding 1 to the number of dimensions. The eps at 1.2 means the algorithm will separate clusters of fetal health conditions and identify outliers, which may indicate rare or anomalous fetal monitoring cases..

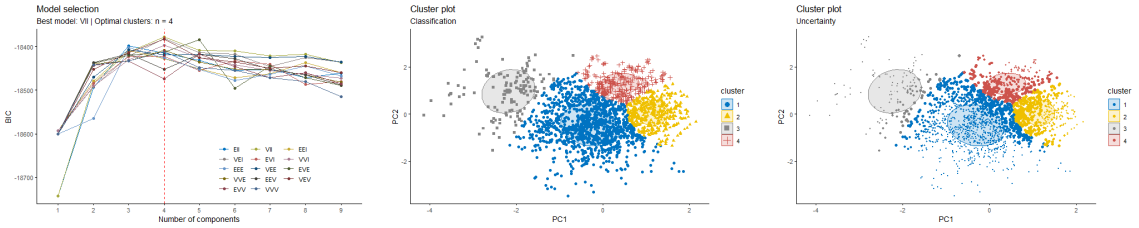
When using minPts = 4 (Figure 16) with eps = 1.2, the DBSCAN algorithm identifies four clusters. The majority of points fall into one large cluster. However, the smaller two clusters may correspond to suspected or pathological fetal health cases with the last small cluster potentially indicating rare conditions that could have been misclassified by the doctors. The black points, representing noise, highlight data points that do not fit into any cluster and could be rare or abnormal fetal monitoring cases that do not follow typical patterns or measurement errors.

Using MinPts = 5 (Appendix A,Figure 28), three distinct clusters emerge, aligning well to determine if fetal health data can be naturally grouped into three categories: normal, suspected, and pathological. The presence of noise is higher compared to when Minpts is 4. However, this still highlights some highly anomalous cases, suggesting that while this configuration is effective, it may not capture all the complexity of fetal health conditions compared to MinPts 4.

Using MinPts = 6 (Appendix A, Figure 28), only two clusters are identified, suggesting that some of the subtler distinctions between suspected and pathological cases are lost. This also shows that as MinPts becomes more stringent, the algorithm prioritises clustering points with stronger similarities, but this may also classify some borderline cases as noise. The black points in this case might represent outliers that could be overlooked in important cases.

The findings with eps = 1.2, using MinPts = 4 seem to strike the best balance. It identifies three clusters, which align with the standard fetal health classification classified by doctors, and an extra fourth cluster that may be a rare case. Additionally, it highlights a manageable number of noise points (outliers) compared to when MinPts is 5 and 6 (Appendix A, Figure 28) which proves how it is the best. This configuration suggests that four fetal health classes are possible.

2.5.3 Probabilistic methods : Mixture of Gaussians



Model selection plot Classification plot Uncertainty plot

Figure 16: Mixture of Gaussians plots for the the optimal model

The Bayesian Information Criterion (BIC) plot indicates that the model with 4 clusters is optimal, minimizing the BIC score and balancing model complexity with data fit. Lower BIC values imply better model performance, meaning the 4-cluster model adequately represents the fetal health data without overfitting or underfitting.

The classification plot for 3 clusters separates the data into distinct groups based on principal components, representing different fetal health states. These clusters are relatively well-defined, but the limited number of clusters may oversimplify the data, potentially combining fetal health states that should be distinct, reducing granularity in understanding health risks. The uncertainty plot for the 3 clusters shows large gray circles around the overlapping areas, indicating high uncertainty in these regions. This suggests that the model struggles to confidently classify fetal health data points near cluster boundaries, reflecting a lack of clarity in differentiating between certain fetal health states, especially where the conditions may transition or overlap.

With 4 clusters, the classification is more refined, better capturing distinct fetal health states. The separation between clusters is clearer, which may allow for more specific diagnosis of different fetal conditions. The additional cluster introduces more detail, dividing groups that may have previously been lumped together in the 3-cluster model. The uncertainty plot for 4 clusters shows smaller and more confined areas of high uncertainty (gray circles), indicating improved confidence in the classification. The model still shows some ambiguity near cluster borders, but the 4-cluster solution provides a clearer distinction between states compared to the 3-cluster model, especially in regions that were previously uncertain.

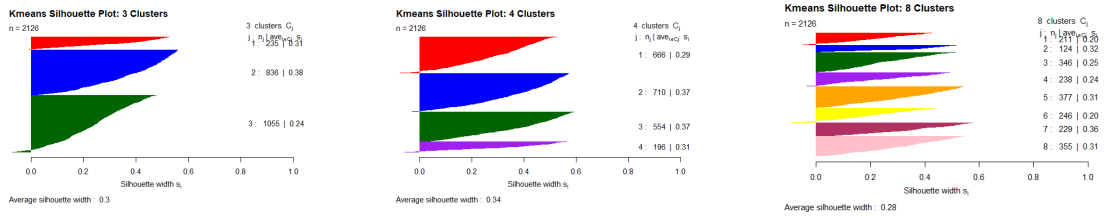
Increasing the number of clusters to 5 introduces even more granularity, potentially capturing subtler variations in fetal health data. However, this may lead to overfitting, as the clusters start to overlap visually, and some groups may represent noise or minor variations rather than distinct health conditions, complicating the interpretation of fetal health. The uncertainty plot for 5 clusters highlights larger and more dispersed uncertainty regions (large gray circles), indicating increased ambiguity. This suggests that the model is struggling to confidently assign data points to a specific cluster, likely due to overfitting, where the addition of more clusters leads to more overlapping regions rather than clearer distinctions.

Based on model selection and uncertainty analysis, the 4-cluster model emerges as the best solution for the fetal health data. It strikes a balance between classification clarity and manageable uncertainty, without overfitting. The smaller uncertainty regions and optimal BIC value indicate that this model provides the most reliable and interpretable classification of fetal health conditions.

2.5.4 Neural net-based methods: Self Organising Maps

2.6 Evaluating clusters

KMEANS



K =3 K =4 K = 8

Figure 17: Silhouette for Kmeans

K	Compactness	Separation	Connectivity	Dunn Index	Silhouette
3	12212.128	0.4954649	193.2349	0.0078854	0.3019457
4	8874.655	0.6333501	280.8532	0.0099785	0.3380750
8	5500.038	0.7727699	450.6107	0.0106590	0.2758730

Table 3: Clustering Metrics for Different K's in K-Means

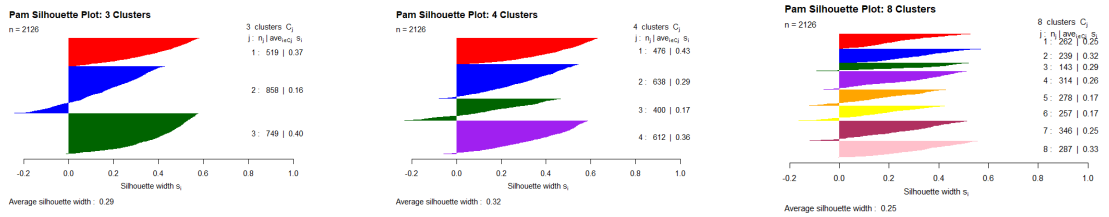
K	Cluster	Compactness	Separation	Silhouette	Size
3	1	12343.504	3.677	0.312	235
3	2	7771.221	2.745	0.376	836
3	3	6508.912	3.170	0.241	1055
4	1	4729.754	2.742	0.288	666
4	2	7266.387	2.474	0.366	710
4	3	6020.599	2.302	0.371	554
4	4	11825.105	3.587	0.315	196
8	1	3185.810	2.326	0.205	211
8	2	9784.892	3.281	0.323	124
8	3	700.424	1.865	0.254	346
8	4	1310.332	1.924	0.239	238
8	5	4853.386	1.956	0.313	377
8	6	4152.175	2.570	0.200	246
8	7	2630.118	1.661	0.360	229
8	8	3777.391	1.745	0.310	355

Table 4: Detailed Clustering Metrics for Each Cluster for kmeans

For the fetal health dataset, K-Means clustering with 4 clusters (K = 4) is the most effective, achieving a good balance between compactness (8,874.655) and separation (0.6334) and the highest overall silhouette score (0.3381), indicating well-defined and cohesive clusters. The cluster sizes range from 196 to 710, reflecting meaningful distinctions among the three fetal health states (normal, suspect, and pathological). In contrast, K = 3 clusters show lower compactness (12,212.128) and a lower silhouette score (0.3019), with uneven cluster sizes (235 to 1,055), suggesting less detailed separation and potential oversimplification. With K = 8, while the clusters capture finer details, they display the lowest silhouette score (0.2759) and the most scattered distribution, with several small clusters (124 and 211), indicating over-segmentation and reduced interpretability.

Therefore, K = 4 provides the most robust and interpretable clustering for analysing fetal health data.

PAM



K = 3 K =4 K = 8

Figure 18: Silhouette for PAM

K	Connectivity	Dunn Index	Silhouette width
3	248.3869	0.0068204	0.2932475
4	266.7917	0.0053117	0.3182592
8	519.7488	0.0076082	0.2536294

Table 5: Clustering Metrics for Different K Values in PAM

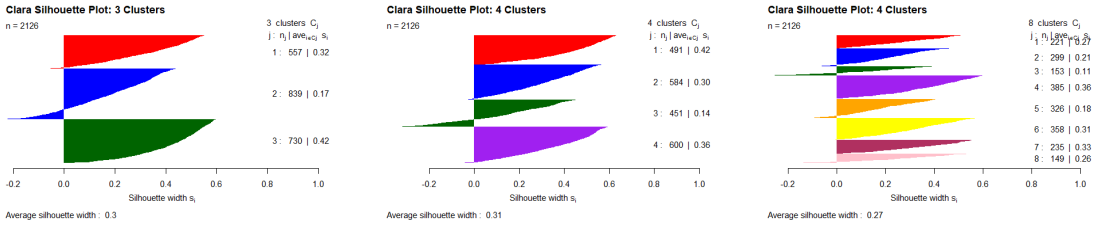
K	Cluster	Compactness	Silhouette	Connectivity	Size
3	1	7071.930	0.371	248.387	519
3	2	10842.266	0.156	248.387	858
3	3	7787.868	0.397	248.387	749
4	1	5516.556	0.430	266.792	476
4	2	5013.563	0.291	266.792	638
4	3	12156.238	0.166	266.792	400
4	4	5884.984	0.359	266.792	612
8	1	1115.679	0.249	519.749	262
8	2	433.667	0.318	519.749	239
8	3	9954.274	0.287	519.749	143
8	4	4710.994	0.264	519.749	314
8	5	2992.828	0.175	519.749	278
8	6	3618.882	0.165	519.749	257
8	7	2608.627	0.251	519.749	346
8	8	3812.409	0.334	519.749	287

Table 6: Detailed Clustering Metrics for Each Cluster in PAM

PAM clustering with 4 clusters (K = 4) proves to be the most effective. It achieves a well-balanced silhouette width of 0.3183 and a moderate Dunn index of 0.0053117, indicating well-defined and cohesive clusters with suitable separation. The clusters are reasonably compact, with sizes ranging from 400 to 638, reflecting meaningful distinctions among the fetal health states (normal, suspect, and pathological). In contrast, K = 3 clusters have a lower silhouette width (0.2932) and higher connectivity (248.387), indicating less distinct clustering and more overlap between clusters. The sizes of these clusters range from 519 to 858, which may result in less detailed separation. With K = 8, while capturing finer details, the silhouette width drops to 0.2536, and the connectivity increases to 519.749, suggesting that the clusters are less cohesive and more scattered, with

sizes varying widely and some clusters being very small (e.g., 28). Thus, K = 4 provides the best balance of compactness, separation, and interpretability for clustering fetal health data using PAM.

CLARA



K =3 K =4 K = 8

Figure 19: Silhouette plots using Clara algorithm

K	Connectivity	Dunn Index	Silhouette
3	238.4183	0.0072597	0.2985415
4	288.7476	0.0056655	0.3109323
8	502.9845	0.0085591	0.2655459

Table 7: Clustering Metrics for Different K Values in CLARA

K	Cluster	Compactness	Silhouette	Connectivity	Size
3	1	8026.472	0.321	238.418	557
3	2	10165.345	0.174	238.418	839
3	3	8001.066	0.424	238.418	730
4	1	5041.978	0.419	288.748	491
4	2	4849.665	0.305	288.748	584
4	3	11066.974	0.137	288.748	451
4	4	5305.409	0.359	288.748	600
8	1	1091.398	0.272	502.985	221
8	2	1059.370	0.209	502.985	299
8	3	9314.658	0.111	502.985	153
8	4	4621.485	0.359	502.985	385
8	5	2881.749	0.180	502.985	326
8	6	3719.809	0.312	502.985	358
8	7	2579.056	0.333	502.985	235
8	8	2588.915	0.257	502.985	149

Table 8: Detailed Clustering Metrics for Each Cluster in CLARA

Using CLARA, the optimal number of clusters is 4. This configuration provides a good balance with a silhouette width of 0.3109 and a Dunn index of 0.0056655, indicating well-separated and cohesive clusters. The clusters have sizes ranging from 491 to 600, reflecting meaningful differentiation among fetal health states. In comparison, K = 3 clusters have a slightly lower silhouette width (0.2985) and a higher connectivity score (238.4183), suggesting less distinct clustering and potential overlap between clusters. The cluster sizes vary from 557 to 839, which might result in less precise separation. With K = 8, although the clusters capture more detail, the silhouette width drops to 0.2655 and connectivity rises to 502.9845, indicating less cohesion and more scattered clusters,

with sizes ranging widely and some being very small (e.g., 149). Therefore, K = 4 offers the most balanced clustering solution for fetal health data using CLARA, providing effective separation and interpretability and it suggests the possibility of four clusters.

Linkage Methods

Method	Silhouette Width	Dunn Index	Compactness	Cophenetic Correlation
Single	0.3877333	0.09606733	25392949	0.6077778
Complete	0.1924180	0.01988139	28807215	0.6011884
Average	0.2520026	0.04173202	24537766	0.6820154
Centroid	0.2091903	0.04348808	24746396	0.6786111
Median	0.1961027	0.01448791	25891114	0.5495343

Table 9: Clustering Metrics for Different Linkage Methods

Method	Cluster	Cluster Size	Compactness	Avg Distance to Center	Silhouette Width
Single	1	2121	23785.9962	3.034511	0.3875772
Single	2	3	525.7573	13.103695	0.7566026
Single	3	1	0.0000	0.000000	0.0000000
Single	4	1	0.0000	0.000000	0.0000000
Complete	1	1356	11707.2546	2.736313	0.1726953
Complete	2	148	1248.2885	2.564360	0.3600217
Complete	3	226	6429.0708	4.761766	0.2181218
Complete	4	396	9534.8181	4.634689	0.1826442
Average	1	1967	18120.8773	2.829025	0.2460675
Average	2	18	3001.5060	12.731299	0.4244431
Average	3	137	8159.6826	6.962695	0.2961292
Average	4	4	233.1321	5.652074	0.8832526
Centroid	1	2004	18907.4681	2.856147	0.1987364
Centroid	2	120	10046.5278	8.877005	0.3872557
Centroid	3	1	0.0000	0.000000	0.0000000
Centroid	4	1	0.0000	0.000000	0.0000000
Median	1	1823	16370.6130	2.750011	0.1363114
Median	2	52	6331.9702	9.610724	0.4045343
Median	3	250	1647.1125	2.358404	0.5895322
Median	4	1	0.0000	0.000000	0.0000000

Table 10: Linkage Methods Metrics for Individual Clusters Across All Linkage Methods

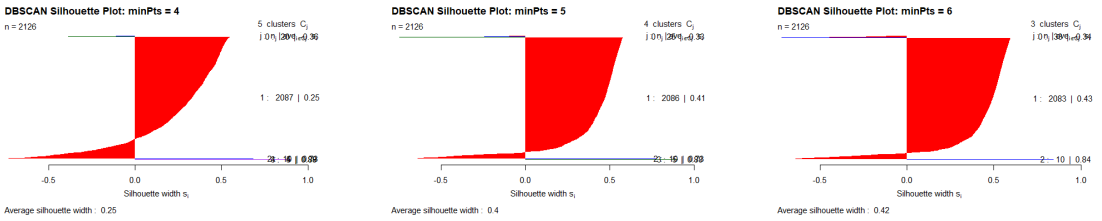
Single linkage shows strong cluster separation, with a high silhouette width (0.3877) and Dunn index (0.096), but it tends to over-separate the data, forming very small clusters like Cluster 2 (3 points, silhouette width 0.7566) and Clusters 3 and 4 (1 point each, silhouette width 0.0000). This over-separation can result in clusters that are too granular to capture broader trends in the data.

Complete linkage and median linkage perform worse overall, with lower silhouette widths (0.1924 and 0.1961) and poor cohesion, as seen in Complete Linkage's Cluster 1 (1,356 points, silhouette width 0.1727) and Cluster 3 (226 points, silhouette width 0.2181). Centroid linkage similarly suffers from over-separation, creating very small clusters like Clusters 3 and 4 (1 point each, silhouette width 0.0000), which limits its ability to represent the data effectively.

In contrast, average linkage strikes a better balance, offering moderate silhouette width (0.2520) and lower compactness (24,537,766), which results in more cohesive and meaningful clusters. For in-

stance, Cluster 1 (1,967 points, silhouette width 0.2461) shows good balance, while smaller clusters like Cluster 4 (4 points, silhouette width 0.8833) demonstrate very tight clustering. Additionally, the average linkage has the highest cophenetic correlation (0.6820), preserving the original data structure, and making it the best overall method for clustering fetal health data due to its effective balance of meaningful separation, compact clusters, and well-preserved data structure.

DBSCAN



minPts= 4 minPts= 5 minPts= 6

Figure 20: Silhouette plots for DBSCAN with eps=1.8

minPts	Silhouette Width	Compactness	Dunn Index	Outliers
4	0.2513947	5108.260	0.04400580	20
5	0.3993695	6357.786	0.04006125	25
6	0.4215839	8391.366	0.04046015	33

Table 11: Clustering Metrics for DBSCAN Using Different Values of minPts

minPts	Cluster	Number of Points	Silhouette	Compactness
4	1	2087	0.2522223	21720.7849
4	2	10	0.7262165	1555.0001
4	3	4	0.8878981	233.1321
4	4	5	0.8836991	901.8781
5	1	2086	0.4054173	21683.1043
5	2	10	0.7262165	1555.0001
5	3	5	0.8836991	901.8781
6	1	2083	0.4316607	21595.79
6	2	10	0.8431337	1555.00

Table 12: Clustering Metrics for Different minPts Values in DBSCAN

For different values of minPts, the silhouette width, which measures how similar an object is to its cluster compared to other clusters, increases from 0.251 to 0.422. This suggests that with higher minPts, the clusters become more distinct and well separated, indicating better clustering quality. Compactness, which measures the total within-cluster variance, decreases as minPts increases, reflecting that clusters are becoming more tightly packed.

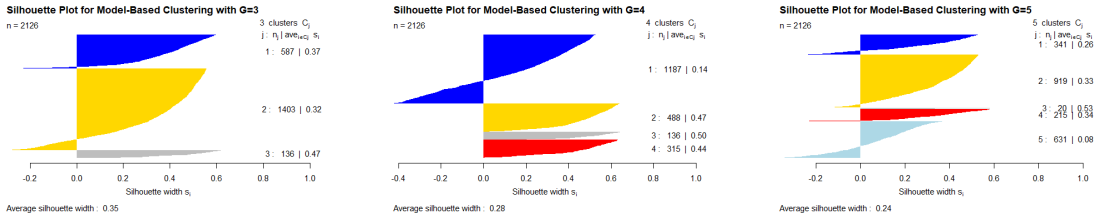
For minPts = 4, the clusters show considerable variation in silhouette widths, with some clusters like Cluster 2 having a high silhouette width of 0.726, suggesting that this cluster is well-separated from others. However, there is a significant number of outliers (20) across all clusters, which might indicate that some data points are not well represented in any cluster.

When minPts is increased to 5 or 6, silhouette widths improve further, indicating better-defined clusters. For example, the silhouette width for minPts = 6 is 0.431, and the compactness values

are also quite low, indicating well-formed clusters. However, the number of outliers increases with minPts = 6, which suggests that some data points are still not fitting well into the clusters.

Overall, DBSCAN appears to perform better with higher values of minPts in terms of cluster separation and compactness, making the clustering more meaningful for fetal health data. The increase in silhouette width and decrease in compactness with higher minPts indicate improved clustering quality. However, the increase in outliers with higher minPts suggests a trade-off between cluster quality and the number of data points not fitting neatly into clusters.

Gaussian Mixture Models



(a). G= 3 (a). G= 4 (c). G=5

Figure 21: Silhouette plots for Gaussian Mixture Models

G	Silhouette Score	Log-Likelihood	BIC	ICL
3	0.3455568	-9165.010	-18398.98	-19386.89
4	0.2850725	-9130.978	-18384.55	-19623.17
5	0.2387927	-9116.258	-18378.09	-19489.51

Table 13: Clustering Metrics for Different Number of Clusters in Gaussian mixture models

Clusters (G)	Cluster	Compactness	Silhouette Width	Size
3	1	1.799684	0.3714895	587
	2	2.173521	0.3223865	1403
	3	7.350542	0.4726554	136
4	1	2.024995	0.1410299	1187
	2	2.907821	0.4674880	488
	3	7.301661	0.5011419	136
	4	1.764763	0.4441779	315
5	1	2.031718	0.25827242	341
	2	2.561495	0.32604863	919
	3	10.911381	0.53105689	20
	4	4.284562	0.33566862	215
	5	1.923682	0.07797826	631

Table 14: Cluster Metrics for the individual cluster in Gaussian mixture models (G3, G4, G5)

The 4-cluster model generally provides the best balance between cluster quality and fit. This model achieves high silhouette scores, with Cluster 2 and Cluster 3 having widths of 0.4675 and 0.5011, respectively, indicating excellent separation and cohesion. The 3-cluster model shows a maximum silhouette width of 0.4727, suggesting good separation but not as strong as the 4-cluster model. Additionally, the 4-cluster model has a more favourable Log-Likelihood (-9130.978) and BIC (-18384.55) compared to the 3-cluster model's -9165.010 and -18398.98, respectively, indicating a

better balance between fit and complexity. The ICL for the 4-cluster model (-19623.17) also shows superior separation compared to the 3-cluster model's -19386.89. The 5-cluster model, while having the best log-likelihood (-9116.258) and BIC (-18378.09), includes very small clusters, such as Cluster 3 with only 20 data points. This raises concerns about stability and interpretability, as small clusters can be less reliable. Although it offers detailed clustering, the potential instability could impact practical use in foetal health scenarios.

Overall, the 4-cluster model is the most balanced and practical choice for foetal health data, providing clear, meaningful, and stable clustering. It delivers well-defined clusters with good separation and fit. The 5-cluster model has a good fit but is less stable due to small clusters, while the 3-cluster model is simpler, and does not match the detailed performance of the 4-cluster model.

Self Isolating Maps

From the evaluating and analysis on this data, the best algorithm for classifying foetal health classes is K-means clustering with K= 4 clusters. This is based on reasons that it has a good balance in compactness, separation and overall silhouette score (Figure 8: k =4). Moreover, the Silhouette plot also shows well separated clusters with minimal misclassification(Figure 8: k =4).

2.7 Cluster Profiling: K-Means with K= 4

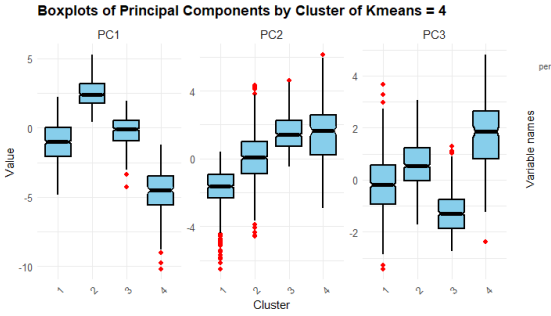


Figure 22: Histograms of each cluster in k-meams= 4

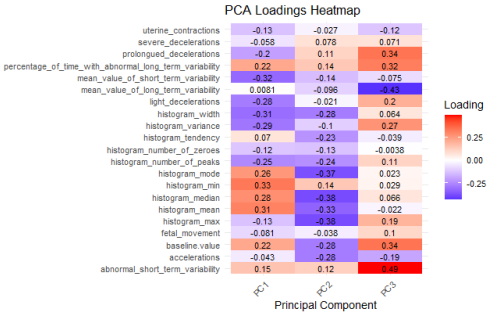


Figure 23: Heatmap showing contribution of variable in PCAs

The foetal health data can be categorised into four distinct groups based on PCA and K-means clustering. Cluster 1 represents a group of foetuses at risk due to consistently negative values across multiple principal components. The negative mean in PC1 (-1.12) and PC2 (-1.79), as well as a moderate negative mean in PC3 (-0.18), suggest underlying health concerns. This cluster may represent foetuses with multiple risk factors or complications, warranting closer monitoring and possible intervention.

Cluster 2 captures a healthier foetal profile, showing consistently positive or near-zero values, with a strong positive mean in PC1 (2.53) and a slight positive mean in PC3 (0.62), while PC2 remains close to zero (0.10). The relatively narrow spread in PC1 (SD: 0.94) and PC3 (SD: 0.87) suggests a concentrated distribution of values potentially reflecting a group of generally stable and normal health parameters. This group may represent normal or low-risk foetuses with fewer complications, indicating a baseline group against which other clusters can be compared.

Cluster 3 appears to represent a more mixed group, possibly indicating foetuses with intermediate or borderline health conditions. This cluster has a negative mean in PC1 (-0.24) and PC3 (-1.22)

but a positive mean in PC2 (1.51), suggesting variability across different health indicators. This group may require moderate observation, as it could represent cases with some risk factors that are not immediately critical but could evolve.

Cluster 4 represents a diverse group, capturing a wide range of health states from normal to moderately abnormal. This cluster reflects significant variation in means in PC1 (-4.71), PC3 (1.82), and PC2 (1.44) across different health dimensions, suggesting the need for individualised assessment, as foetuses in this cluster could require different levels of monitoring and care.

These patterns suggest that the foetal health data can be effectively categorised into four distinct groups, each with unique characteristics across the principal components. furthermore, the percentage of time with abnormal long-term variability and severe decelerations are major drivers of variance in PC1, while baseline value and histogram mode significantly impact PC2. Accelerations and abnormal short variability are key contributors to explain the variability captured by PC3.

Cluster	**Description**	**PC1**	**PC2**	**PC3**	**Potential Health Status**	**Cases**
Cluster 1	High-risk group with consistently negative values.	Mean: -1.12	Mean: -1.79	Mean: -0.18	Potentially at risk; requires monitoring.	666
Cluster 2	Low-risk group with positive/near-zero values.	Mean: 2.53	Mean: 0.10	Mean: 0.62	Likely normal or low risk.	710
Cluster 3	Intermediate group with mixed values.	Mean: -0.24	Mean: 1.51	Mean: -1.22	Borderline health conditions; some risk factors.	554
Cluster 4	Diverse group with a wide range of values.	Mean: -4.71	Mean: 1.44	Mean: 1.82	Varied health states; requires individualized assessment.	196

Table 15: Cluster Descriptions and Potential Health Status

2.8 Inference

	Df	**Pillai**	**Approx F**	**num Df**	**den Df**	**Pr(>F)**
cluster	3	1.7245	956.31	9	6366	< 2.2e-16 ***
Residuals	2122					

Table 16: Multivariate Analysis of Variance (MANOVA) Results. Significance code: 0 '***'.

The MANOVA analysis reveals a statistically significant effect of the cluster factor on the multivariate response variables, with a Pillai's Trace value of 1.7245 and an approximate F-statistic of 956.31 ($p < 2.2e\text{-}16$). This extremely low p-value indicates strong evidence against the null hypothesis, suggesting that the cluster factor significantly affects the multivariate response variables. Consequently, the differences between clusters are highly significant in terms of the overall variance of the dependent variables.

2.9 Conclusion

3 Association Rule Mining

3.1 Introduction

3.2 Data Wrangling

3.3 Exploratory data analysis

3.3.1 Description of data: Used Features

Feature	Description	Type
Typical Chest Pain	1 = Yes typical chest pain, 0 = No typical chest pain	Factor
Age	Patient's age into risk levels: "Low-Risk Age", "Moderate Risk Age", "High-Risk Age"	Factor
Atypical	N = No atypical chest pain, Y = Atypical chest pain present	Factor
TG	Triglyceride levels in the blood, classified as Normal or High	Factor
EF.TTE	Ejection Fraction from TTE categorised as Low or Normal	Factor
BMI	Body Mass Index categorised based on weight classification: "Underweight", "Normal", "Overweight"	Factor
FBS	Fasting Blood Sugar levels categorised as Low, Normal, or High	Factor
ESR	Erythrocyte Sedimentation Rate categorized into levels of inflammation: "$Very_Low$", "Low", "$Normal$", "$Elevated$", "$High$"	Factor
BP	Blood pressure categorised as Low, Normal, or High	Factor
Weight	Patient's weight classes: "Underweight", "Light", "Normal", "Heavy", "Very Heavy", "Obese", "Severely Obese"	Factor
Region.RWMA	Regional Wall Motion Abnormality categorised as Normal or Abnormal	Factor
LDL (Low Density Lipoprotein)	LDL levels categorized as Normal or High	Factor
HTN (Hypertension)	1 = Hypertension present, 0 = No hypertension	Factor
K (Potassium)	Potassium levels categorized as Low, Normal, or High	Factor
HB (Hemoglobin)	Hemoglobin levels categorized as Low, Normal, or High	Factor
PLT (Platelets)	Platelet count categorized as Low, Normal, or High	Factor
Lymph	Lymphocyte percentage in blood categorised as Low, Normal, or High	Factor
HDL (High-Density Lipoprotein)	HDL levels categorized as Low or Normal	Factor
Length	Patient height classes: "Very Short", "Short", "Average", "Tall", "Very Tall", "Extremely Tall"	Factor
Neut (Neutrophils)	Neutrophil percentage in blood categorised as Low, Normal, or High	Factor
Na (Sodium)	Sodium levels categorized as Low, Normal, or High	Factor
BUN (Blood Urea Nitrogen)	BUN levels categorized as Low, Normal, or High	Factor
CR (Creatinine)	Creatinine levels categorized as Low, Normal, or High	Factor
WBC (White Blood Cells)	White blood cell count are categorised as Low, Normal, or High	Factor
Cath	Target variable indicating whether CAD is present (Cad) or not (Normal)	Factor

Table 17: Description and type of data variables used for Association mining rules after descritizining.

3.3.2 Unstandardized data: Summary statistics

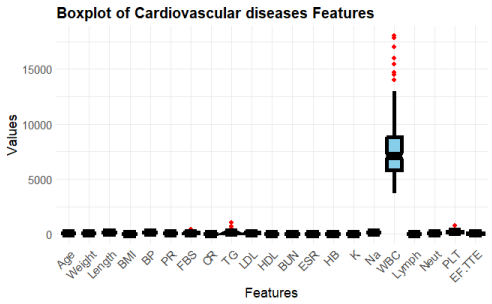


Figure 24: Boxblots displaying the distribution of unstandardised CAD data.

The dataset reflects a population with varying distribution patterns across the variables. Age follows a normal distribution with a mean of 58.90 years, centered around the middle-aged to elderly population, and some outliers in the older range (SD = 10.39). Weight and BMI, which represent a key risk factor for cardiovascular diseases, have a right-skewed distribution, where the majority of patients cluster around higher weights and BMI values, indicating overweight or obesity (see boxplot). Blood pressure (BP) also exhibits a right-skewed distribution, with most patients in the pre-hypertensive range around 129.55 mmHg, but a few outliers pushing the maximum to 190 mmHg (see table).

Variable	Min	Median	Mean	Max	SD
Age	30.00	58.00	58.90	86.00	10.39
Weight	48.00	74.00	73.83	120.00	11.99
Length	140.00	165.00	164.72	188.00	9.33
BMI	18.12	26.78	27.25	40.90	4.10
BP	90.00	130.00	129.55	190.00	18.94
PR	50.00	70.00	75.14	110.00	8.91
FBS	62.00	98.00	119.18	400.00	52.08
CR	0.50	1.00	1.06	2.20	0.26
TG	37.00	122.00	150.34	1050.00	97.96
LDL	18.00	100.00	104.64	232.00	35.40
HDL	15.90	39.00	40.23	111.00	10.56
BUN	6.00	16.00	17.50	52.00	6.96
ESR	1.00	15.00	19.46	90.00	15.94
HB	8.90	13.20	13.15	17.60	1.61
K	3.00	4.20	4.23	6.60	0.46
Na	128.00	141.00	141.00	156.00	3.81
WBC	3700.00	7100.00	7562.05	18000.00	2413.74
Lymph	7.00	32.00	32.40	60.00	9.97
Neut	32.00	60.00	60.15	89.00	10.18
PLT	25.00	210.00	221.49	742.00	60.80
EF.TTE	15.00	50.00	47.23	60.00	8.93

Table 18: Summary statistics for selected variables

In terms of lipid profiles, triglycerides (TG) and low-density lipoprotein (LDL) are also right-skewed, with the majority of patients showing moderately elevated levels, but some extreme cases pushing the TG levels up to 1050 and LDL up to 232 (see boxplot). White blood cell count (WBC)

shows a bimodal distribution, with two distinct groups: one in the normal range and another with elevated values indicating possible inflammation or infection (see boxplot). Overall, the data reveals a combination of normal and right-skewed distributions, particularly in key cardiovascular risk factors like BP and lipid levels, highlighting elevated health risks in the patient population.

3.3.3 Standardized data and correlation

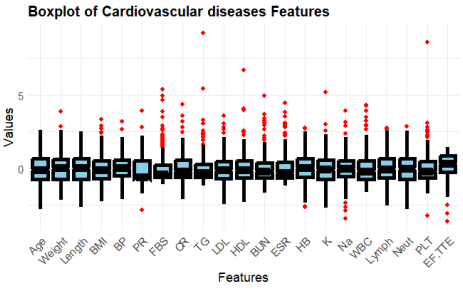


Figure 25: Boxblots displaying the distribution of standardised CAD data.

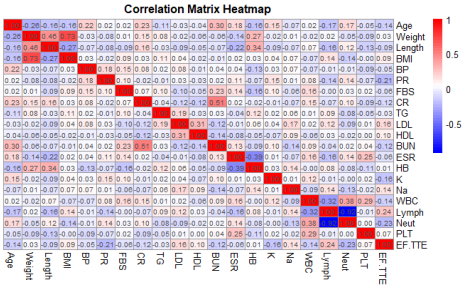


Figure 26: correlation heatmap of CAD data.

The boxplot of standardized cardiovascular disease (CAD) features (Figure 25) illustrates that most variables are centred around their means, indicating that the data have been standardized with a mean of zero. Despite standardization, several variables show a notable number of outliers, such as weight, BMI, LDL, and triglycerides (TG), which maintain a right-skewed distribution. These outliers suggest extreme values that are still significantly higher than the majority of the population even after adjusting for scale.

The correlation matrix heatmap (Figure 26) highlights key relationships between cardiovascular disease (CAD) features. Strong positive correlations are observed between BMI and weight, as expected, since BMI is derived from weight and height. Similarly, LDL cholesterol (LDL) and total cholesterol (TC) are highly correlated, reflecting LDL's significant contribution to overall cholesterol levels. Age shows a moderate positive correlation with systolic blood pressure (BP), indicating that BP tends to rise with age. Triglycerides (TG) also correlate moderately with total cholesterol, while HDL cholesterol shows some correlation with other lipid variables like LDL, highlighting the interrelatedness of lipid profiles.

On the other hand, variables such as white blood cell count (WBC) display weak or near-zero correlations with lipid profiles, suggesting little linear relationship between inflammation markers and cholesterol levels. This suggests white blood cell cont and lipid levels are influenced by different factors.

3.3.4 Feature selection: Random Forest

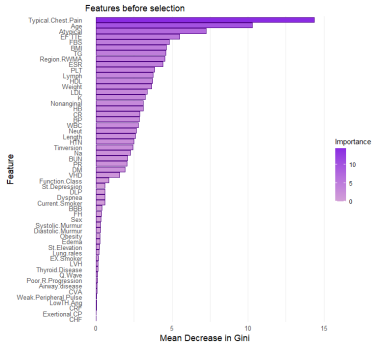


Figure 27: Importance of all Features in CAD dataset

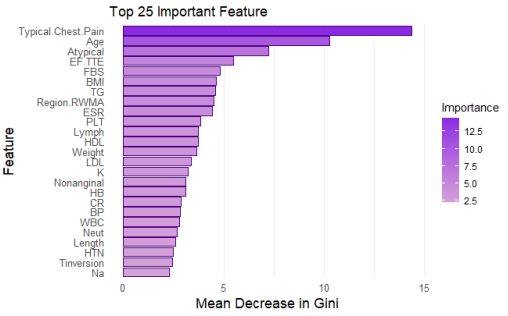


Figure 28: Features selected

Random Forest was used in this dataset to assess the importance of various features in predicting Coronary Artery Disease (CAD). Figure 3 helps identify which features can be discarded due to their lesser importance, guiding the creation of a more efficient model. It also suggests that focusing on key features may improve predictive accuracy. After pinpointing the most important features, the dataset was refined to include the top 25 features, as illustrated in Figure 4. This graph showcases these top 25 features, ranked by their "Mean Decrease in Gini" values. It is significant as it demonstrates the benefits of narrowing down to the most influential features. By concentrating on predictors, the graph shows how reducing the number of features enhances the model's interpretability and performance, making it more robust and effective.

3.4 Association Rule Mining Algorithms

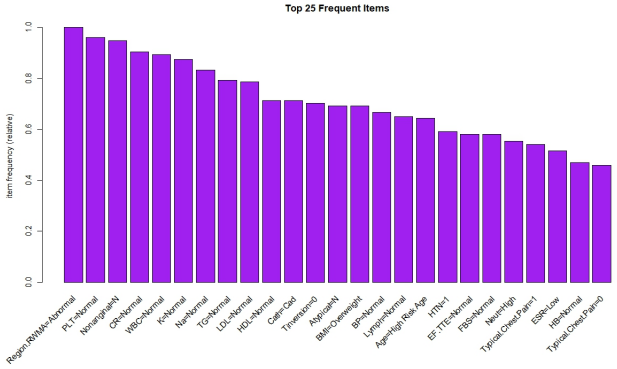
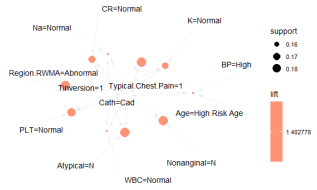


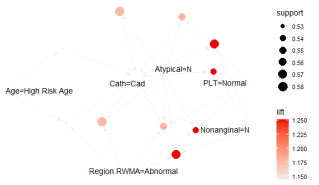
Figure 29: Item Frequency graph.

Figure 29 displays frequently occurring items or feature combinations discovered during rule mining. It is important because it reveals key patterns that can be used for further rule-based predictions or decision-making processes. For instance, the frequent co-occurrence of symptoms like chest pain types and the RWMA region highlights significant associations within the dataset. These patterns can inform rule mining approaches such as association rule learning, aiding in the development of more precise and actionable insights.

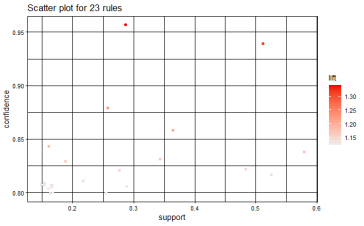
3.4.1 Apriori



(a). Lift graph for top 10 rules indicating how likely conditions appear together.



(b). Support graph for top 10 rules indicating how frequently a rule appears.



(c). Rules

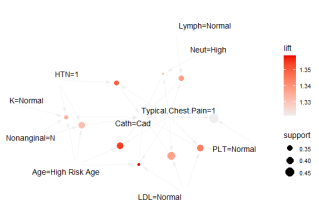
Figure 30: Apriori graphs

The lift graph (Fig a) shows Cath=CAD in the central position in a lighter colour (lift around 1.40) meaning that the presence of CAD is moderately more likely when the associated symptoms. the typical chest is closely associated with the Cath=CAD which means that patients with typical chest pain have a significantly increased likelihood of being diagnosed with CAD. The lift value of 1.30-1.40 means that chest pain makes a patient 30–40% more likely to have CAD than without it. Region.RWMA=Abnormal has a lift value of 1.40 and it means that Patients with abnormal regional wall motion on imaging have a 40% increased likelihood of having CAD, showing that structural heart abnormalities are strong indicators of disease. Lastly, BP=High and Age=High Risk show a moderate lift value, suggesting that being in a high-risk age group or having high blood pressure increases the chance of CAD by around 20–30%.

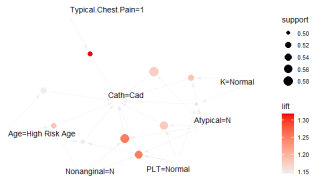
The support graph (Figure b), Cath=CAD: node has a moderate support value around 0.55 indicating that 55% of the patients have CAD, either as a condition or diagnosis. PLT=Normal, Nonanginal=N, and Atypical=N nodes show support values around 0.58, meaning that 58% of the patients have normal platelet counts and do not present atypical or non-anginal chest pains. This suggests that these are common characteristics, but they are not strong indicators of CAD. The typical chest pain node with 0.18 and abnormal wall motion of 0.17 does not have a high frequency. However, it indicates that 18% and 17% of the patients with these symptoms have a strong chance of CAD when it occurs.

A few rules appear in the top-right corner of Figure C, showing high confidence (0.95) and moderate support (0.5). This means that these rules are both frequent and reliable for predicting CAD. For example, a rule with Typical Chest Pain and Cath=CAD likely has high confidence, meaning that when chest pain is present, CAD is present 95% of the time, making chest pain a reliable predictor. Many rules cluster around a confidence of 0.85 and support values between 0.2 and 0.4. These rules have moderate predictive value and could include associations like BP=High or Age=High Risk with CAD, indicating that while these are common factors, they do not guarantee the presence of CAD like chest pain or abnormal RWMA.

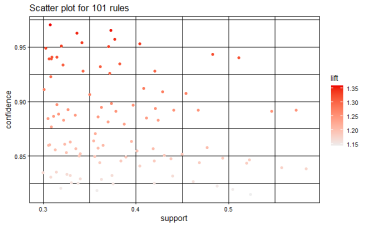
From the Apriori graphs, we can see that chest pain and RWMA abnormalities are strong diagnostic indicators of CAD. Additionally, high BP and high-risk age are moderate predictors but are still important in the overall risk assessment.



(a). Lift graph for top 10 rules



(b).Support graph for top 10 rules



(c). Rules

Figure 31: FP-growth graphs

3.4.2 FP-growth

In figure 1, The typical Chest Pain = 1: node has a higher lift value of 1.35. This suggests that when typical Chest Pain is present, the likelihood of CAD increases by 35% making it strongly associated with CAD. The Nonanginal=N, Atypical=N, PLT = Normal nodes are associated with CAD but with lower lift values (1.20–1.25). This means that not having non-anginal or atypical symptoms and having a normal platelet count increases the likelihood of CAD by 20–25%. Hypertension and high neutrophil (HTN=1 (Hypertension) and Neut = High) counts are moderately associated with CAD, with lift values around 1.25–1.30. These conditions increase the likelihood of CAD by 25–30%, meaning they are important predictors but not as strong as chest pain. the K=Normal and LDL=Normal nodes have lower lift values of range 1.15–1.20, indicating a weaker association with CAD.

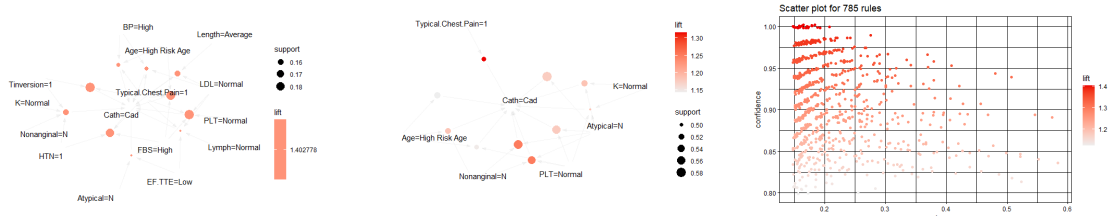
The support graph (Figure B), shows Cath=CAD has moderate support (0.52) meaning that 52% of the patients in the dataset are diagnosed with CAD. The Typical Chest Pain = 1 node has a support value around 0.40, meaning 40% of the patients exhibit chest pain. While not the most frequent condition, it is a strong predictor when present. Nonanginal=N and Age=High Risk nodes have moderate support (0.50–0.55), suggesting that half of the patients in the dataset do not have non-anginal chest pain and are in a high-risk age (above 55) category.

The rules Scatter Plot (Figure C) shows high confidence values (0.95) and moderate support values (0.4–0.5). These rules are both frequent and reliable. For instance, rules combining Typical Chest Pain and Cath=CAD likely fall in this range, with 95% confidence that when chest pain is present, CAD is present. Moderate Support and Confidence rules cluster around 0.85–0.90 confidence with support values of 0.3–0.4. This means these rules are frequent and moderately reliable in predicting CAD. For example, rules involving Hypertension or High Neutrophils with CAD may fall into this range. Low confidence rules Rules with lower confidence (around 0.80) likely involve conditions with weaker associations with CAD, such as LDL=Normal or K=Normal, where the prediction of CAD is less certain.

From FP-Growth, typical chest pain, Hypertension, and abnormal laboratory values (e.g., neutrophil count) are strong predictors of CAD, whereas normal potassium or LDL levels are less predictive.

3.4.3 Eclut

Figure A shows that Cath=CAD is the central node, and has a lift value indicating that the likelihood of CAD increases by 30–40% when associated with many other variables. Typical Chest Pain = 1 node has a lift value of around 1.40 suggesting that when chest pain is present, the



(a). Lift graph for top 10 rules

(b).Support graph for top 10 rules

(c). Rules

Figure 32: Eclut graphs

likelihood of CAD is 40% higher than what would be expected by chance. The Age = High-Risk Age node has a lift of 1.40, meaning that patients in the high-risk age group are 40% more likely to be diagnosed with CAD than random chance. Nonanginal Chest Pain = N (lift =1.25) suggests that patients with nonanginal chest pain are also more likely to have CAD. Those with LDL = Normal (1.25) and K = normal (1.12) have the likelihood of CAD by 15% and 12% respectively, showing a weaker positive association.

Figure B nonanginal=N and atypical = N nodes show a higher support of 0.57-0.58, highlighting how commonly patients with above 55 chest pains pain are diagnosed with CAD. BP = High, node have the highest support, around 0.56–0.58, meaning they appear in about 56–58% of the patients. These are common patterns, but not necessarily the most predictive on their own. Nonanginal = N, Atypical = N, and Age = High-Risk nodes also have high support (around 0.54–0.55), suggesting these clinical features are quite common. Typical Chest Pain = 1 has a support value of around 0.40, meaning 40% of the patients exhibit typical chest pain.

Figure C shows rules with high confidence (above 0.95) and moderate support (around 0.1–0.2) at the top of the scatter plot. These rules are both frequent and highly reliable, most likely involving variables like Typical Chest Pain = 1 and Age = High-Risk Age where the confidence suggests that when typical chest pain is present, CAD is diagnosed in more than 95% of cases. Moderate Support, Moderate Confidence rules fall between 0.85 and 0.95 confidence, with support between 0.30 and 0.40. These represent less common combinations of clinical features that still provide reasonably strong predictive power for CAD. Low Confidence rules have lower confidence (around 0.80–0.85) and lower support (around 0.3), meaning these rules involve variables that appear less frequently in the dataset and have a weaker predictive association with CAD. These could include variables like Abnormal RWMA or PLT, which, while predictive, are not as universally reliable as chest pain or age.

From eclat, typical chest pain, age and hypertension are highly associated with CAD.

3.5 Evaluation

Algorithm	Support	Confidence
Apriori	0.15	0.8
FPGrowth	0.3	0.8
Eclut	0.15	0.8

Table 19: Parameters used during different association rules algorithms.

Algorithm	Initial rules	Redundant rules	Used rules
Apriori	51082	51059	23
Fpgrowth	935	834	101
Eclut	51082	NA	51082

Table 20: Redundancy table comparing the number of initial, redundant, and cleaned rules for different association rules algorithms.

LHS (Antecedent)	RHS (Consequent)	Support	Confidence	Lift
{Atypical=N, Region.RWMA = Abnormal}	{Cath=Cad}	0.5809	0.8381	1.1757
{Atypical=N}	{Cath = Cad}	0.5809	0.8381	1.1757
{Atypical=N, Region.RWMA = Abnormal, Nonanginal=N}	{Cath = Cad}	0.5710	0.8918	1.2509
{Atypical=N, Nonanginal =N}	{Cath = Cad}	0.5710	0.8918	1.2509
{Atypical=N, Region. RWMA = Abnormal, PLT=Normal}	{Cath = Cad}	0.5545	0.8400	1.1783

Table 21: Association rule results for Apriori.

LHS (Antecedent)	RHS (Consequent)	Support	Confidence	Lift
{Typical.Chest.Pain = 1, LDL = Normal, Age=High Risk Age}	{Cath=Cad}	0.3069	0.9688	1.3589
{Typical.Chest.Pain=1, Age = High Risk Age}	{Cath=Cad}	0.3729	0.9658	1.3548
{Typical.Chest.Pain=1, HTN =1}	{Cath =Cad}	0.3366	0.9623	1.3498
{PLT=Normal, Typical.Chest.Pain =1, LDL=Normal}	{Cath = Cad}	0.3795	0.9583	1.3443
{Typical.Chest.Pain =1, Lymph = Normal}	{Cath=Cad}	0.3399	0.9537	1.3378

Table 22: Association rule results for FPGROWTH.

LHS (Antecedent)	RHS (Consequent)	Support	Confidence	Lift
{Atypicalv = N, Region.RWMAv = Abnormal}	{Cath = Cad}	0.5809	0.8381	1.1757
{Atypical= N}	{Cathv= Cad}	0.5809	0.8381	1.1757
{Atypical= N, Region.RWMA = Abnormal, Nonanginal =N}	{Cath = Cad}	0.5710	0.8918	1.2509
{Atypical=N, Nonanginal =N}	{Cath=Cad}	0.5710	0.8918	1.2509
{Atypical = N, Region. RWMA = Abnormal, PLT =Normal}	{Cath = Cad}	0.5545	0.8400	1.1783

Table 23: Association rule results showing antecedent (LHS), consequent (RHS), support, confidence, and lift values eclut.

3.6 Best Partitioning rule

3.7 Conclusion

4 References

5 Appendix A

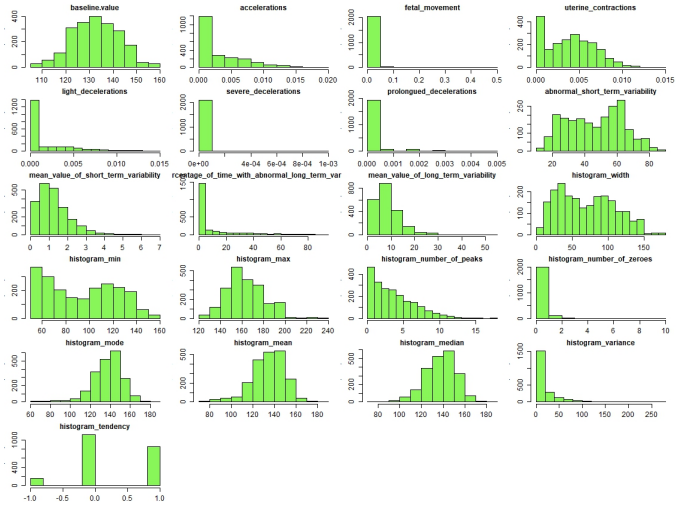


Figure 33: unscaled boxplots

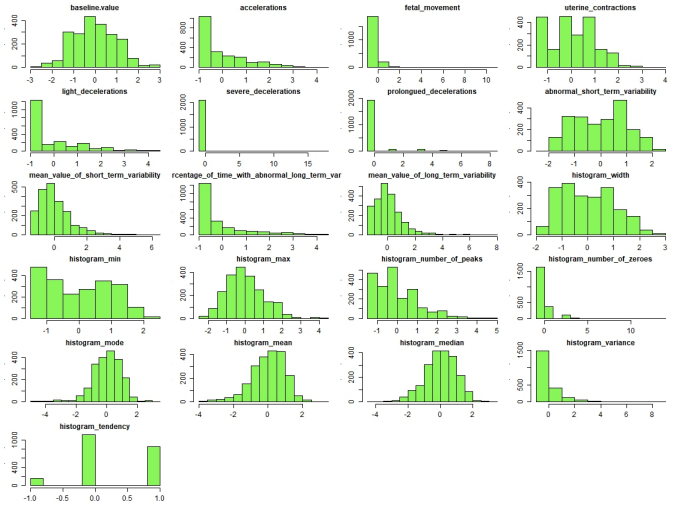
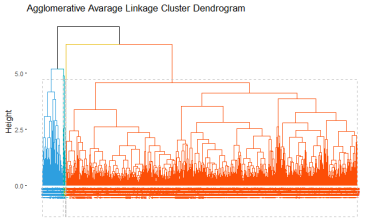
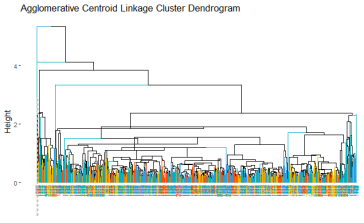


Figure 34: unscaled boxplots

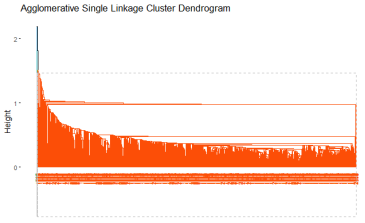
Hierarchical



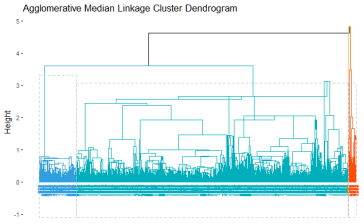
Avarage linkage



Centroid linkage

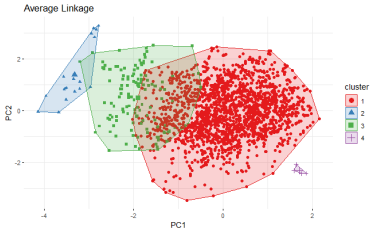


Single linkage

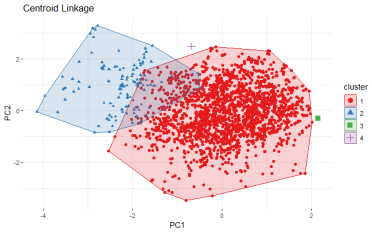


Median linkage

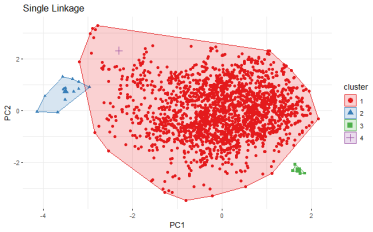
Figure 35: Dendogram plots using different linkage methods



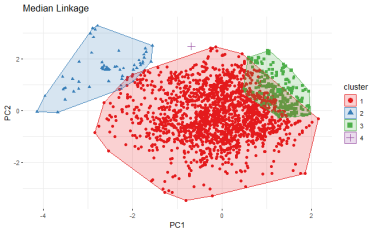
Avarage linkage distribution



Centroid linkage distribution



Single linkage distribution



Median linkage distribution

Figure 36: Number of clusters K using different linkage methods

DBSCAN

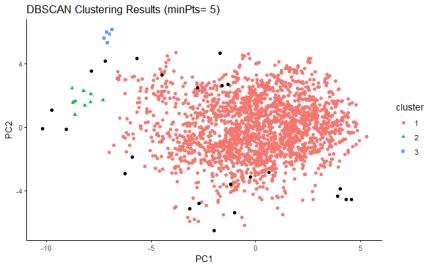


Figure 37: DBSCAN plot: eps= 1.2

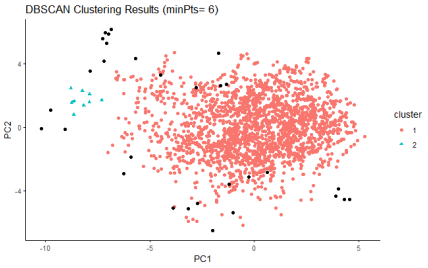
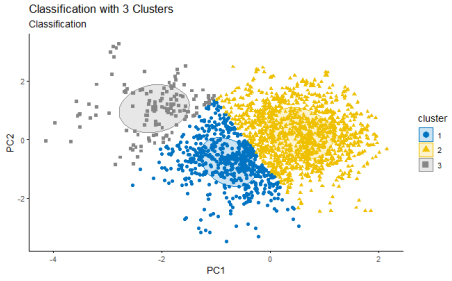
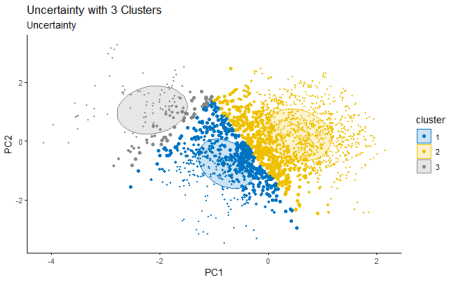


Figure 38: DBSCAN plot:eps = 1.2

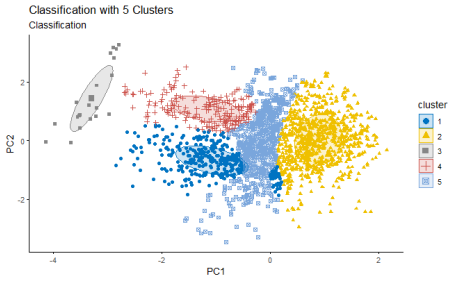
Gaussian



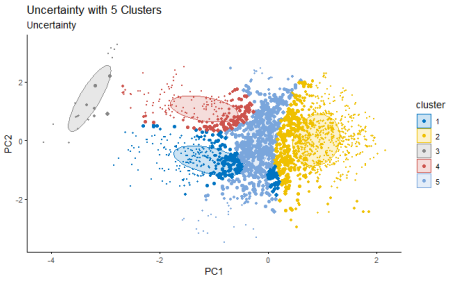
(a)Plot for Classification with 3 clusters



(b) Uncertainty for 3 clusters



(c) Classification for 5 clusters



(d) Uncertainty for 5 clusters

Figure 39: Classification and Uncertainty plots for Gaussian mixture models