UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD

Statistical Sciences

Natalie Bianca Alexander

The University of Cape Town

The Department of Statistical Sciences

STA5077Z Unsupervised Learning

Assignment 1: Question 2

Due Date: 6 October 2023

Table of Contents

Question 2

CAD

Abstract

The aim of this project is to identify the important features associated with coronary artery disease (CAD), which may assist in the timely diagnosis of CAD. The objectives of this project are to: (1) use association rule mining methods to determine the features that are mostly associated with CAD and (2) to find any other interesting association rules.

Introduction

Many people suffer from cardiovascular diseases, which has a detrimental effect on the global mortality rate. For this reason, accurate and timely diagnosis of coronary artery disease (CAD) is essential. Angiography was proposed as a method for CAD diagnoses. Angiography produces the most accurate results among other CAD diagnostic tools. Researchers are, therefore, seeking novel modalities for CAD diagnosis. This project assesses the CAD dataset, which contains angiography results and other measurements obtained from 303 patients. The data set also includes information on whether a patient had CAD or not.

The data presented in this project has been previously analysed by R. Alizadehsani et al (2013) [1]. This project uses the methods outlined by R. Alizadehsani et al (2013) to determine the features that are mostly associated with CAD and identify some interesting association rules.

Data

The CAD data set was accessed in XLSX format from the VULA site. The data set may also be downloaded on the Mendeley Data website [2].

Exploratory Data Analysis

A. Checking the data dimensions

The dimensions of the data was assessed using the dim() function, which suggested that this is a 303 row by 56 column dataset, where 55 columns represent the independent variables and 1 column represents the dependent variable. I then check for missing data and found that there were no rows with missing data.

Table 1 in **Appendix B** shows the column names or "independent variables" which were identified using the colnames() function. **Table 1** in **Appendix B** also provides the description of each independent variable. It is important to note that the data provided on VULA does not include the "Rhythm" independent variable used by R. Alizadehsani et al (2013) [1]. This data set has a variable called "BBB" in its place. In addition, the data set used in this project has an additional independent variable called "Length" which is not included in the data set used by R. Alizadehsani et al (2013).

B. Head and tail of the data

Table 2 to **Table 7** in **Addendum B** shows the head of the CAD data set**,** while **Table 8** to **Table 13** in **Addendum B** shows the tail of the CAD data set.

C. Data types

Next, I assessed the data types of each column using the str() function. The data types of each variable can be seen in **Table 1** in **Appendix B**. In all cases, where the categorical variable has "yes" and "no" classes, I coded the class to 1 and 0, respectively. Thereafter, all categorical and binary variables

shown in **Table 1** of **Appendix B** were converted to factors. Thereafter, the summary statistics of each independent variable was computed.

D. Summary statistics

Table 14 below shows the summary statistics for each numeric variable, where summary statistics such as the minimum value, first quartile, median, mean, third quartile and maximum value are provided. In **Table 14** we see that the numeric variables differ in terms of their range, where certain independent variables such as "ESR" have a wider range (min of "ESR" = 1 and max of "ESR" 90) relative to other independent variables such as "Length" with a narrower range (min of "Length" = 140 and max of "Length" = 188). In addition, certain independent variables have discrete values such as "Age", while other independent variables have continuous values such as "CR". The mean values in **Table 14** also differ significantly among these independent variables, with mean values ranging from 1.056 for "CR" to 7562.046 for "WBC". The median values in **Table 14** also differ significantly among these independent variables, with median values ranging from 1 for "CR" to 7100 for "WBC". The difference in the mean and median values suggest differences in the distributions of the observations across the numeric independent variables. The large differences in quartile 1 and quartile 3 (in **Table 14)** for each of the numeric independent variables indicate differences in the interquartile range (IQR = Q3 – Q1) which further suggests differences in the spread of the data across these variables.

For ease of formatting, **Table 15** shows the variance and standard deviation for each variable. Dark green cells suggest that the numeric independent variable has relatively large variance and standard deviation values. Lighter green cells indicate numeric independent variables with moderately large variance and standard deviation values. Dark orange cells indicate numeric independent variables with relatively small variance and standard deviation values. Lighter orange cells indicate numeric independent variables with moderately small variance and standard deviation values. Large variance values suggest that these variables have observations which are, on average, far from the mean of the variable. On the other hand, small variance values indicate variables where the observations are on average, close to the mean of the variable. We look at the standard deviation for a more representative and comparable value which explains the average distance of the observations to the mean. The standard deviation has a unit of measurement which is the same as the unit of measurement for the variable in question. Numeric independent variables such as "FBS", "TG", "LDL", "WBC", and "PLT" have large variance and standard deviation values, while "CR", "HB" and "K" have smaller variance and standard deviation values.

Table 14: Summary statistics of the numeric independent variables of the CAD data set. This table includes the minimum value, first quartile, median, mean, third quartile and maximum value.

Variables	Min	Q1	Median	Mean	Q3	Max
Age	30	51	58	58.898	66	86
Weight	48	65	74	73.832	81	120
Length	140	158	165	164.716	171	188
BMI	18.115	24.514	26.776	27.248	29.412	40.901
BP	90	120	130	129.554	140	190
PR	50	70	70	75.142	80	110
FBS	62	88.5	98	119.185	130	400
CR	0.5	0.9	1	1.056	1.2	2.2
TG	37	90	122	150.343	177	1050
LDL	18	80	100	104.644	122	232
HDL	15.9	33.5	39	40.234	45.5	111
BUN	6	13	16	17.502	20	52
ESR	1	9	15	19.462	26	90
HB	8.9	12.2	13.2	13.153	14.2	17.6
K	3	3.9	4.2	4.231	4.5	6.6
Na	128	139	141	140.997	143	156
WBC	3700	5800	7100	7562.046	8800	18000
Lymph	7	26	32	32.399	39	60
Neut	32	52.5	60	60.149	67	89
PLT	25	183.5	210	221.488	250	742
EF-TTE	15	45	50	47.231	55	60

Table 15: Variance and standard deviation of the CAD dataset. Dark green cells suggest that the independent variable has a relatively large variance and standard deviation values. Lighter green cells indicate independent variables with moderately large variance and standard deviation values. Dark orange cells indicate independent variables with relatively small variance and standard deviation values. Lighter orange cells indicate independent variables with moderately small variance and standard deviation values.

Variables	Variance	Std
Age	107.999	10.392
Weight	143.697	11.987
Length	87.005	9.328
BMI	16.801	4.099
BP	358.652	18.938
PR	79.42	8.912
FBS	2712.29	52.08
CR	0.07	0.264
TG	9596.054	97.959
LDL	1252.926	35.397
HDL	111.494	10.559
BUN	48.397	6.957
ESR	253.971	15.936
HB	2.594	1.61
K	0.21	0.458
Na	14.5	3.808
WBC	5826137.521	2413.739
Lymph	99.453	9.973
Neut	103.683	10.182
PLT	3696.178	60.796
EF-TTE	79.695	8.927

E. Visualizing the distribution of the numeric independent variables by means of box plots

Figure 1 below shows the box plots for each numeric independent variable 1 to 9, while **Figure 2** shows the box plots of each numeric independent variable 10 to 18 and **Figure 3** shows the box plots of the numeric independent variables 19 to 21.

We see that most numeric independent variables have almost symmetric distribution, with some variables having a slight positive skew in their distribution. There are few cases where a slight negative skew exists, e.g., for the "EF-TTE" variable. It is important to note that certain independent variables also have outliers, which can be seen in the box plots in **Figures 1-3.** Below I discuss the distribution of the numeric independent variables in terms of their box plots and their mean and median values found in **Table 14**.

In the box plot in **Figure 1**, the "Age" independent variable has a symmetric shape in accordance with **Table 14**, where the mean of 58.898 is approximately equal to the median of 58. In the box plot in **Figure 1**, the "Weight" independent variable has an almost symmetric shape in accordance with **Table 14**, where the mean of 73.832 is approximately equal to the median of 74. In the box plot in **Figure 1**, the "Length" independent variable has a symmetric shape in accordance with **Table 14**, where the mean of 164.716 is approximately equal to the median of 165. In the box plot in **Figure 1**, the "BMI" independent variable has an almost symmetric shape in accordance with **Table 14**, where the mean of 27.248 is approximately equal to the median of 26.776. In the box plot in **Figure 1**, the "BP" independent variable has an almost symmetric shape in accordance with **Table 14**, where the mean of 129.554 is approximately equal to the median of 130. In the box plot in **Figure 1**, the "PR" independent variable has a positive skew, in accordance with **Table 14**, where the mean of 75.142 > median of 70. In the box plot in **Figure 1**, the "FBS" independent variable has a positive skew, in accordance with **Table 14**, where the mean of 119.185 > median of 98. In the box plot in **Figure 1**, the "CR" independent variable has a positive skew, in accordance with **Table 14**, where the mean of 1.056 > median of 1. In the box plot in **Figure 1**, the "TG" independent variable has a positive skew, in accordance with **Table 14**, where the mean of 150.343 > median of 122.

In the box plot in **Figure 2**, the "LDL" independent variable has a positive skew, in accordance with **Table 14**, where the mean of 104.644 > median of 100. In the box plot in **Figure 2**, the "HDL" independent variable has a positive skew, in accordance with **Table 14**, where the mean of 40.234 > median of 39. In the box plot in **Figure 2**, the "BUN" independent variable has a positive skew, in accordance with **Table 14**, where the mean of 17.502 > median of 16. In the box plot in **Figure 2**, the "ESR" independent variable has a positive skew, in accordance with **Table 14**, where the mean of 19.462 > median of 15. In the box plot in **Figure 2**, the "HB" independent variable has a symmetric shape, in accordance with **Table 14**, where the mean of 13.153 is approximately equal to the median of 13.2. In the box plot in **Figure 2**, the "K" independent variable has a symmetric shape, in accordance with **Table 14**, where the mean of 4.231 is approximately equal to the median of 4.2. In the box plot in **Figure 2**, the "Na" independent variable has a symmetric shape, in accordance with **Table 14**, where the mean of 140.997 is approximately equal to the median of 141. In the box plot in **Figure 2**, the "WBC" independent variable has a positive skew, in accordance with **Table 14**, where the mean of 7562.046 is greater than the median of 7100. In the box plot in **Figure 2**, the "Lymph" independent variable has a symmetric shape, in accordance with **Table 14**, where the mean of 32.399 is approximately equal to a median of 32.

In the box plot in **Figure 3**, the "Neut" independent variable has a symmetric shape, in accordance with **Table 14**, where the mean of 60.149 is approximately equal to the median of 60. In the box plot in **Figure 3**, the "PLT" independent variable has a right skew, in accordance with **Table 14**, where the mean of 221.488 is greater than the median of 210. **Figure 3**, the "EF-TTE" independent variable has a negative skew, in accordance with **Table 14**, where the mean of 47.231 is less than the median of 50.

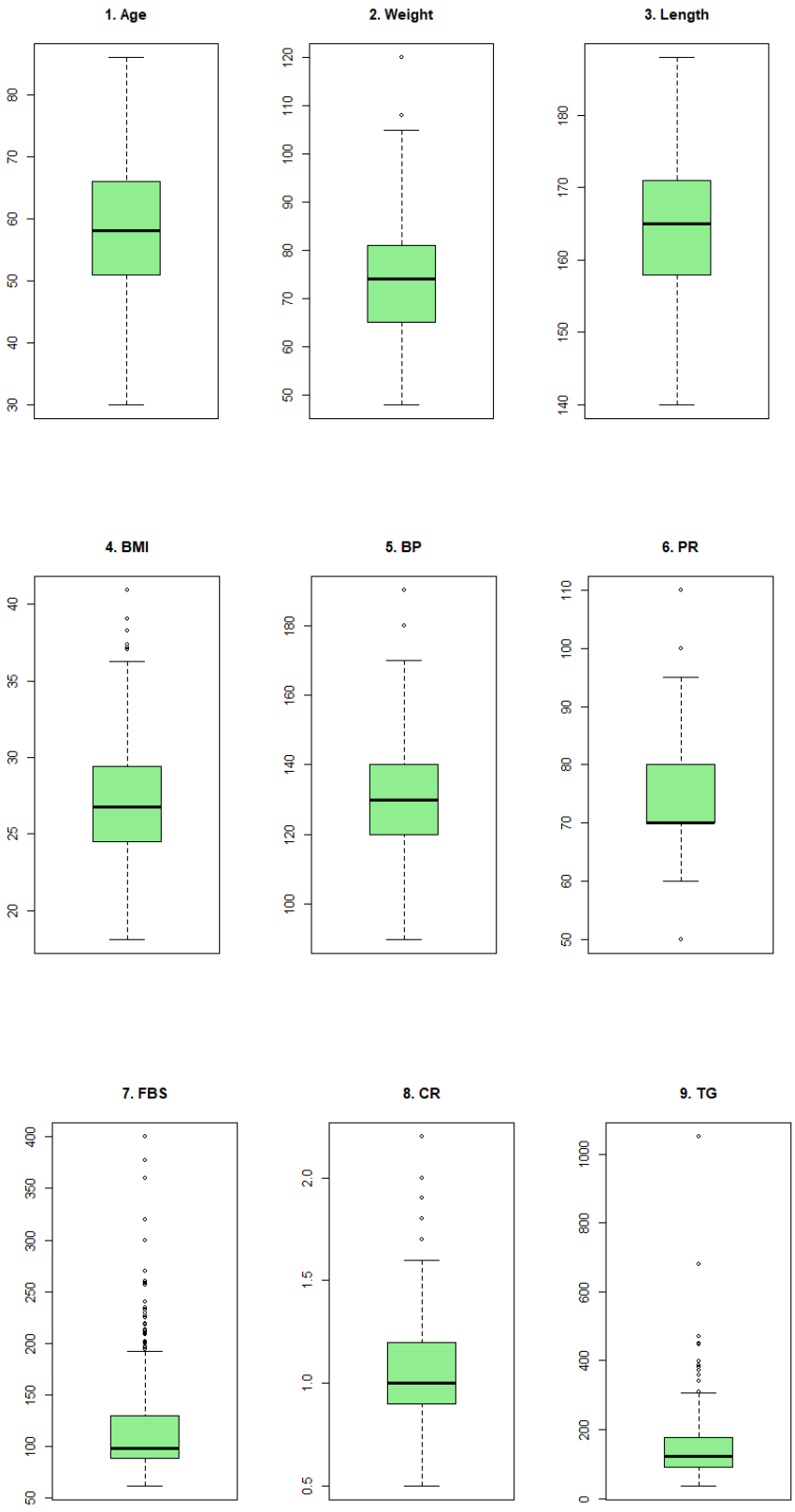

Figure 1: Box plots showing the distribution of observations for numeric independent variables 1 to 9.

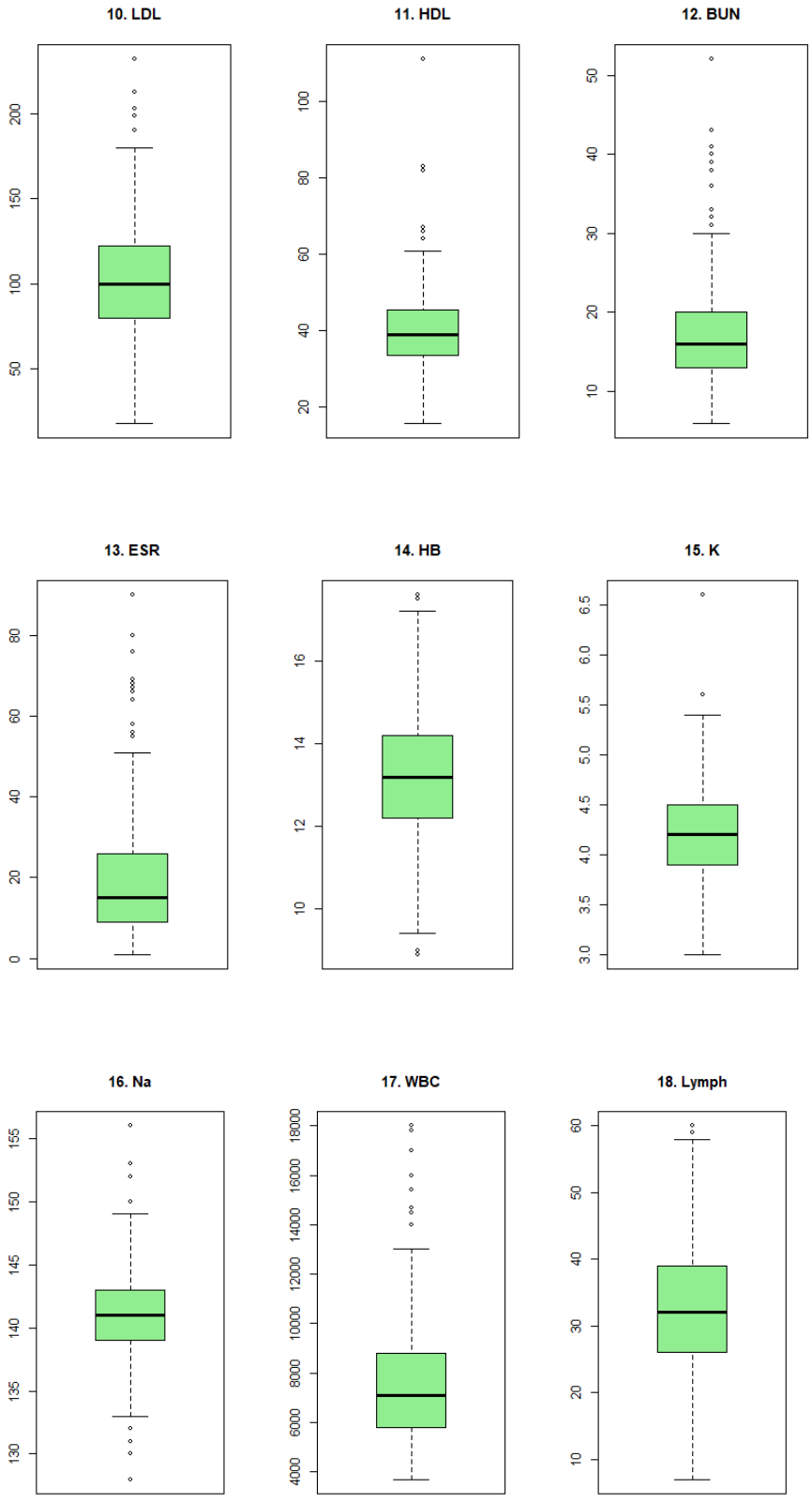

Figure 2: Box plots showing the distribution of observations for numeric independent variables 10 to 18.

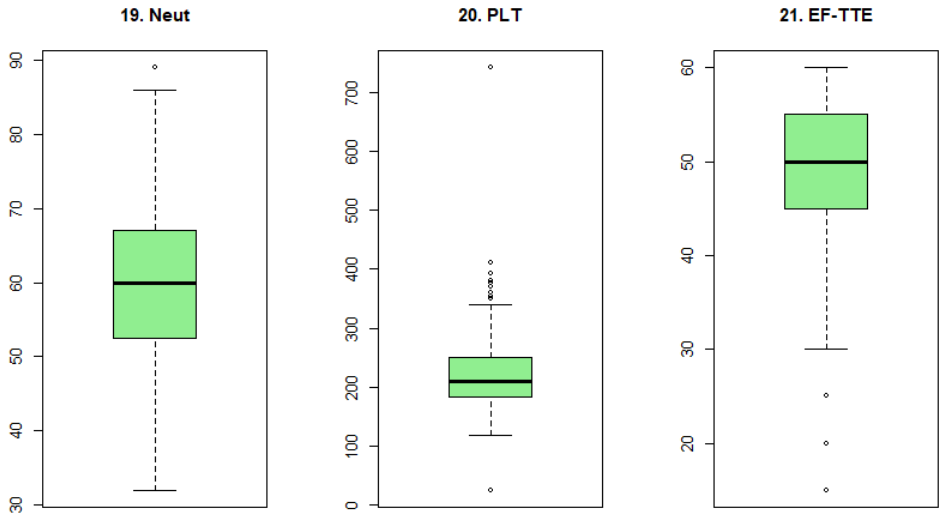

Figure 3: Box plots showing the distribution of observations for numeric independent variables 19 to 21.

F. Visualizing the distribution of the numeric independent variables by means of histograms.

The histograms in **Figures 4-6** below show the distribution of observations for each numeric independent variable. The histograms displayed are in agreement with the box plots above in **Figures 1-3**. We see strong, positive skewed distributions for "PR", "FBS", "CR", "TG", "LDL", "HDL", "BUN", "ESR", "WBC" and "PLT". In addition, we see strong, negative skewed distributions for "EF-TTE". All other histograms show approximate symmetric or "bell" shapes for the remaining independent variables.

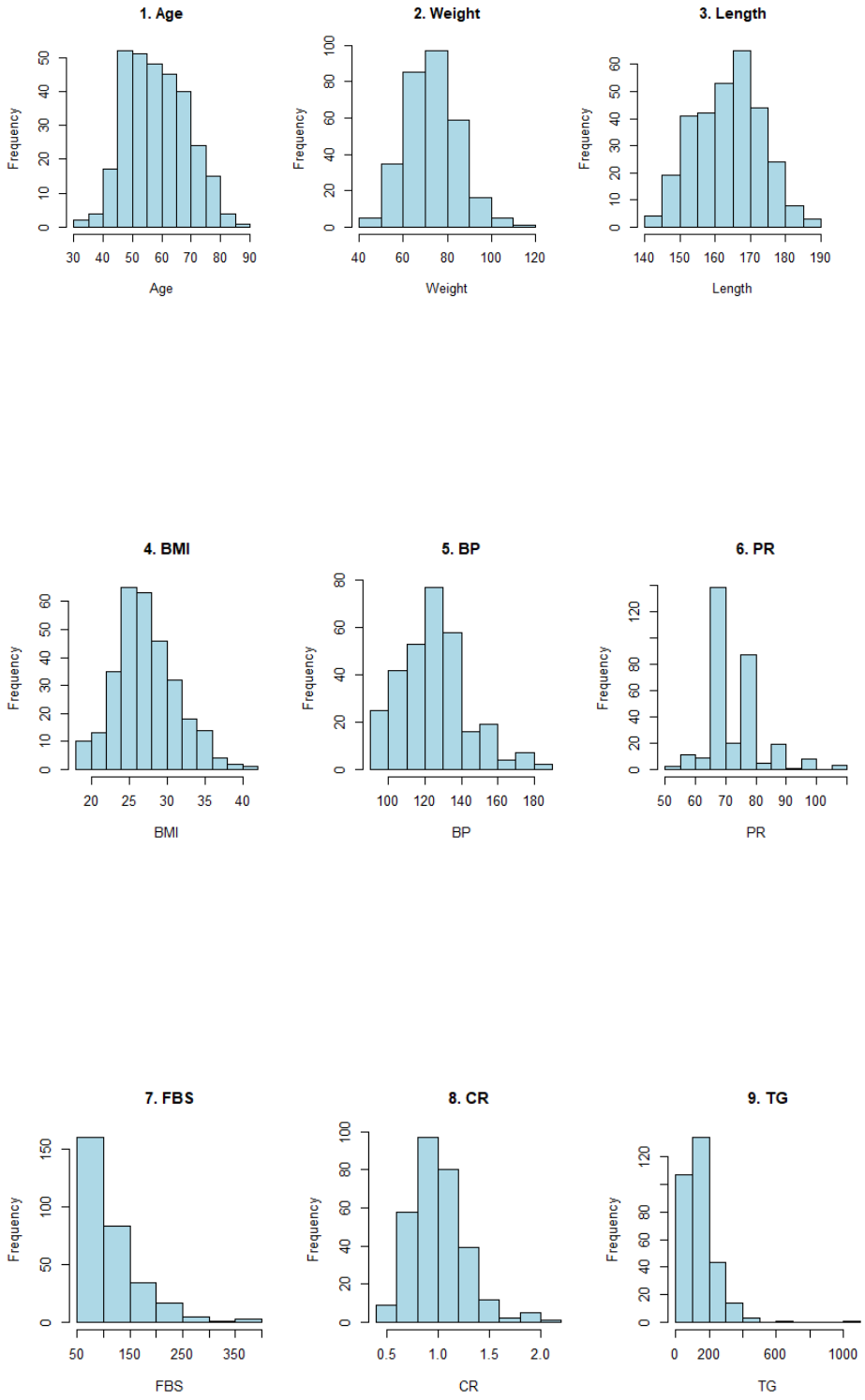

Figure 4: Histograms showing distribution of observations for each numeric independent variable 1 to 9.

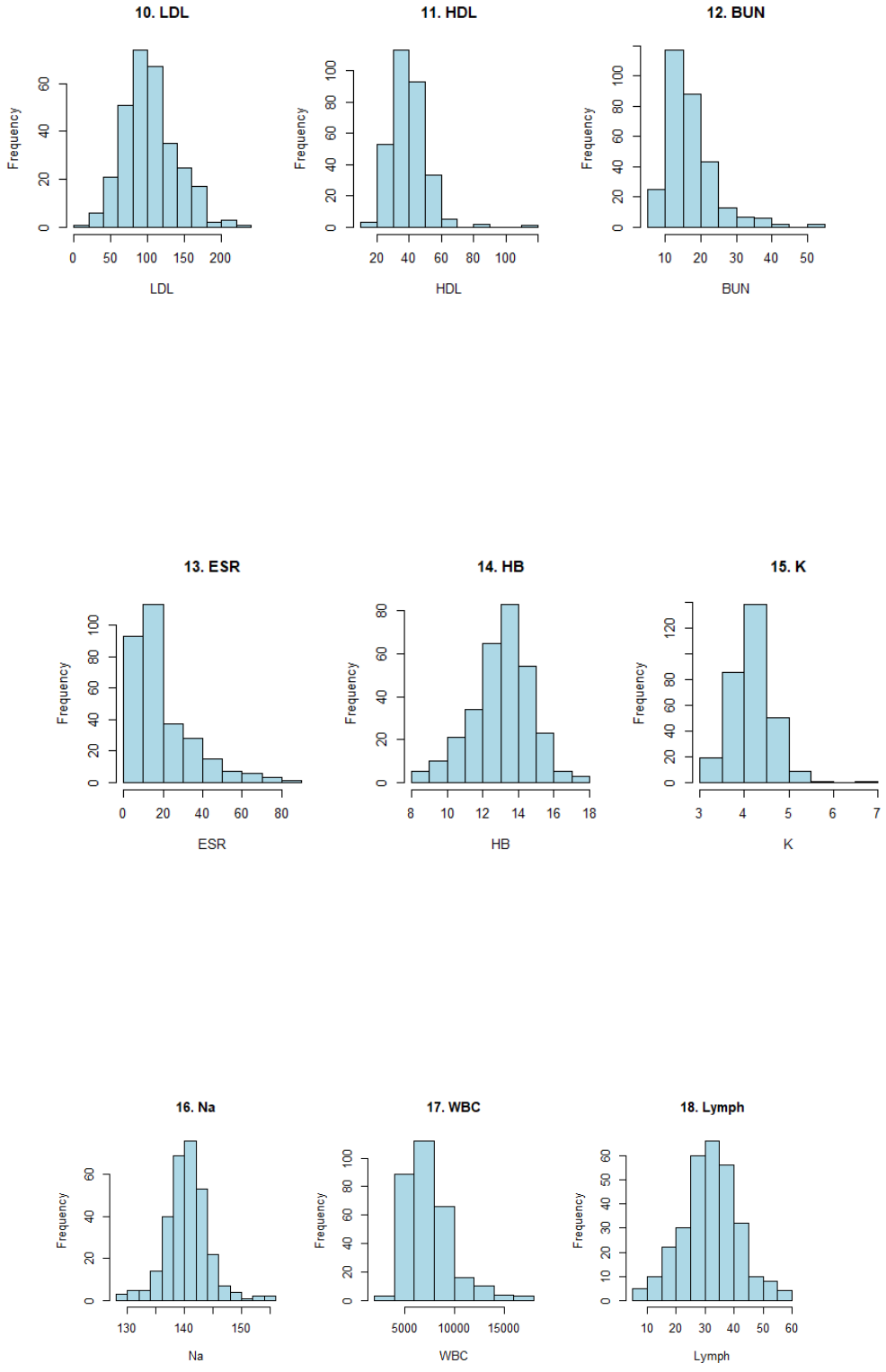

Figure 5: Histograms showing the distribution of observations for each numeric independent variable 10 to 18.

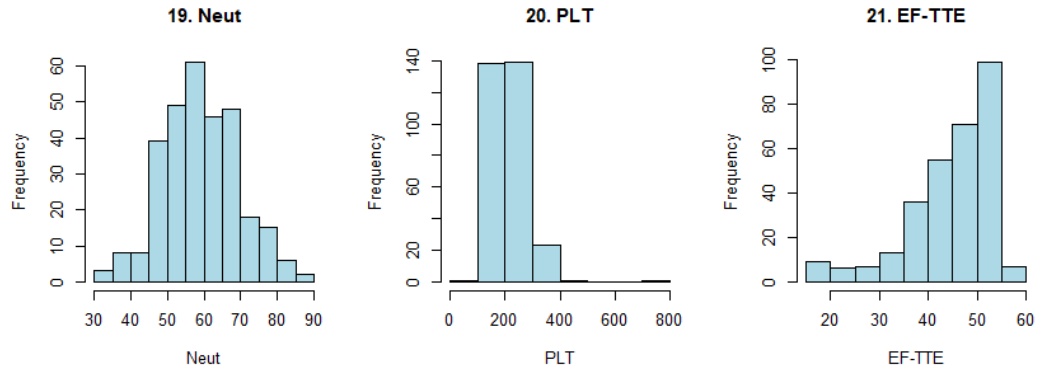

Figure 6: Histograms showing the distribution of observations for each numeric independent variable 19 to 21.

G. *Visualizing the distribution of the categorical and binary independent variables by means of bar plots.*

The bar plots in **Figure 7** to **Figure 10** show the frequency of each class for the binary and categorical independent variables. "Yes" and "no" classes are coded as 1 and 0, respectively.

The bar plots show a trend, where in most cases, the independent variables have a greater number of persons in the "no" or 0 class compared to the "yes" or 1 class. In addition, I noticed that the "Exertional CP" independent variable (shown in the bar plot in **Figure 9 - Frame 24**) has all its observations grouped into the "no" class (in other words, the 0 coded class).

Below I discuss the bar plots in detail.

In the bar plot in **Figure 7,** we see that the number of males (176) exceeded the number of females (127). In the bar plot in **Figure 7,** we also see that the number of non- diabetic individuals (DM = 0 with a frequency of 213) surpassed the number of diabetic persons (DM = 1 with a frequency of 90). In the bar plot in **Figure 7,** we also see that the number of persons with hypertension (HTN = 1 with a frequency of 179) is greater than the number of persons without hypertension (HTN = 0 with a frequency of 124). In the bar plot in **Figure 7,** we also see that the number of non-smokers (current smoker = 0 with a frequency of 240) is greater than the number of current smokers (current smoker = 1 with a frequency of 63). In the bar plot in **Figure 7,** we also see that the number of people who are not ex-smokers (Ex-smoker = 0 with a frequency of 293) are greater than the frequency of people who are ex-smokers (Ex-smoker = 1 with a frequency of 10). In the bar plot in **Figure 7,** we also see that the number of people without a family history of CAD (FH = 0 with a frequency of 255) is greater than the number of people with a family history of CAD (FH = 1 with a frequency of 48). In the bar plot in **Figure 7,** we also see that the number of people with obesity (Obesity = 1 with a frequency of 211) is greater than the number of people without obesity (Obesity = 0 with a frequency of 92). In the bar plot in **Figure 7,** we also see that the number of people without chronic renal failure (CRF = 0 with a frequency of 297) is greater than the number of people with chronic renal failure (CRF = 1 with a frequency of 6). In the bar plot in **Figure 7,** we also see that the number of people without a cerebrovascular accident (CVA = 0 with a frequency of 298) is greater than the number of people with a cerebrovascular accident (CVA = 1 with a frequency of 5).

In the bar plot in **Figure 8,** we see that the number of people without an airway disease (coded 0 with a frequency of 292) is greater than the number of people with an airway disease (coded 1 with a frequency of 11). In the bar plot in **Figure 8,** we also see that the number of people without a thyroid

disease (coded 0 with a frequency of 296) is greater than the number of people with a thyroid disease (coded 1 with a frequency of 7). In the bar plot in **Figure 8,** we also see that the number of people without a congestive heart failure (CHF, coded 0 with a frequency of 302) is greater than the number of people with a congestive heart failure (CHF, coded 1 with a frequency of 1). In the bar plot in **Figure 8,** we also see that the number of people without Dyslipidaemia (DLP = 0 with a frequency of 191) is greater than the number of people with Dyslipidaemia (DLP = 1 with a frequency of 112). In the bar plot in **Figure 8,** we also see that the number of people without Edema (coded 0 with a frequency of 291) is greater than the number of people with Edema (coded 1 with a frequency of 12). In the bar plot in **Figure 8,** we also see that the number of people without a weak peripheral pulse (coded 0 with a frequency of 298) is greater than the number of people with a weak peripheral pulse (coded 1 with a frequency of 5). In the bar plot in **Figure 8,** we also see that the number of people without lung rales (coded 0 with a frequency of 292) is greater than the number of people with lung rales (coded 1 with a frequency of 11). In the bar plot in **Figure 8,** we also see that the number of people without systolic murmur (coded 0 with a frequency of 262) is greater than the number of people with systolic murmur (coded 1 with a frequency of 41). In the bar plot in **Figure 8,** we also see that the number of people without diastolic murmur (coded 0 with a frequency of 294) is greater than the number of people with diastolic murmur (coded 1 with a frequency of 9).

In the bar plot in **Figure 9,** we see that the number of people with typical chest pain (coded 1) with a frequency of 164 is greater than the number of people without typical chest pain (coded 0) with a frequency of 139. In the bar plot in **Figure 9,** we also see that the number of people without dyspnea (coded 0) with a frequency of 169 is greater than the number of people with dyspnea (coded 1) with a frequency of 134. In the bar plot in **Figure 9,** we see that the number of people with function class of 0 (freq = 211) > number of people with a function class of 2 (freq = 73) > number of people with a function class of 3 (freq = 18) > number of people with a function class of 1 (freq = 1). In the bar plot in **Figure 9,** we see that the number of people who are not atypical (coded 0) with a frequency of 210 is greater than the number of people who are atypical (coded 1) with a frequency of 93. In the bar plot in **Figure 9,** we see that the number of people who are anginal (coded 0) with a frequency of 287 is greater than the number of people who are nonanginal (coded 1) with a frequency of 16. In the bar plot in **Figure 9,** we see that all 303 people do not have Exertional Chest Pain i.e., Exertional CP coded as 0. In the bar plot in **Figure 9,** we see that the number of people without a low threshold angina (coded as 0) having a frequency of 301 is greater than the number of people with a low threshold angina (coded as 1) having a frequency of 2. In the bar plot in **Figure 9,** we see that the number of people who are Q Wave negative (coded as 0) with a frequency of 287 is greater than the number of people who are Q Wave positive (coded as 1) with a frequency of 16. In the bar plot in **Figure 9,** we see that the number of people who are St elevation negative (coded as 0) with a frequency of 289 is greater than the number of people who are St Elevation positive (coded as 1) with a frequency of 14.

In the bar plot in **Figure 10,** we see that the number of people without St Depression (coded 0) having a frequency of 232 is greater than the number of people with St Depression (coded 0c) having a frequency of 71. In the bar plot in **Figure 10,** we see that the number of people without Tinversion (coded 0) having a frequency of 213 is greater than the number of people with Tinversion (coded 1) having a frequency of 90. In the bar plot in **Figure 10,** we see that the number of people without left ventricular hypertrophy (LVH = 0) having a frequency of 283 is greater than the number of people with left ventricular hypertrophy (LVH = 1) having a frequency of 20. In the bar plot in **Figure 10,** we see that the number of people without Poor R Progression (coded 0) having a frequency of 294 is greater than the number of people with Poor R Progression (coded 1) having a frequency of 9. In the bar plot in **Figure 10,** for the variable BBB, we see that the number of people in the "N' class (282) > the number of people in the LBBB class (13) > the number of people in the RBBB class (8). In the bar plot in **Figure 10,** for the variable "Region RWMA", we see that the number of people in class 0 (n = 217) > the number of people in the class 2 (n = 32) > the number of people in class 1 (n =

26) > the number of people in the class 3 and class 4 (n = 14 for both classes 3 and 4). In the bar plot in **Figure 10,** for the variable "VHD", we see that the number of people in the "mild' class (n = 149) > the number of people in the "N" class (n = 116) > the number of people in the "Moderate" class (n = 27) > the number of people in the "Severe" class (n = 11).

For the dependent variable "Cath" in **Figure 10**, we see that the number of people with CAD is 216 which is greater than the number of "normal" people without CAD (n = 87). This suggests the presence of class imbalance.

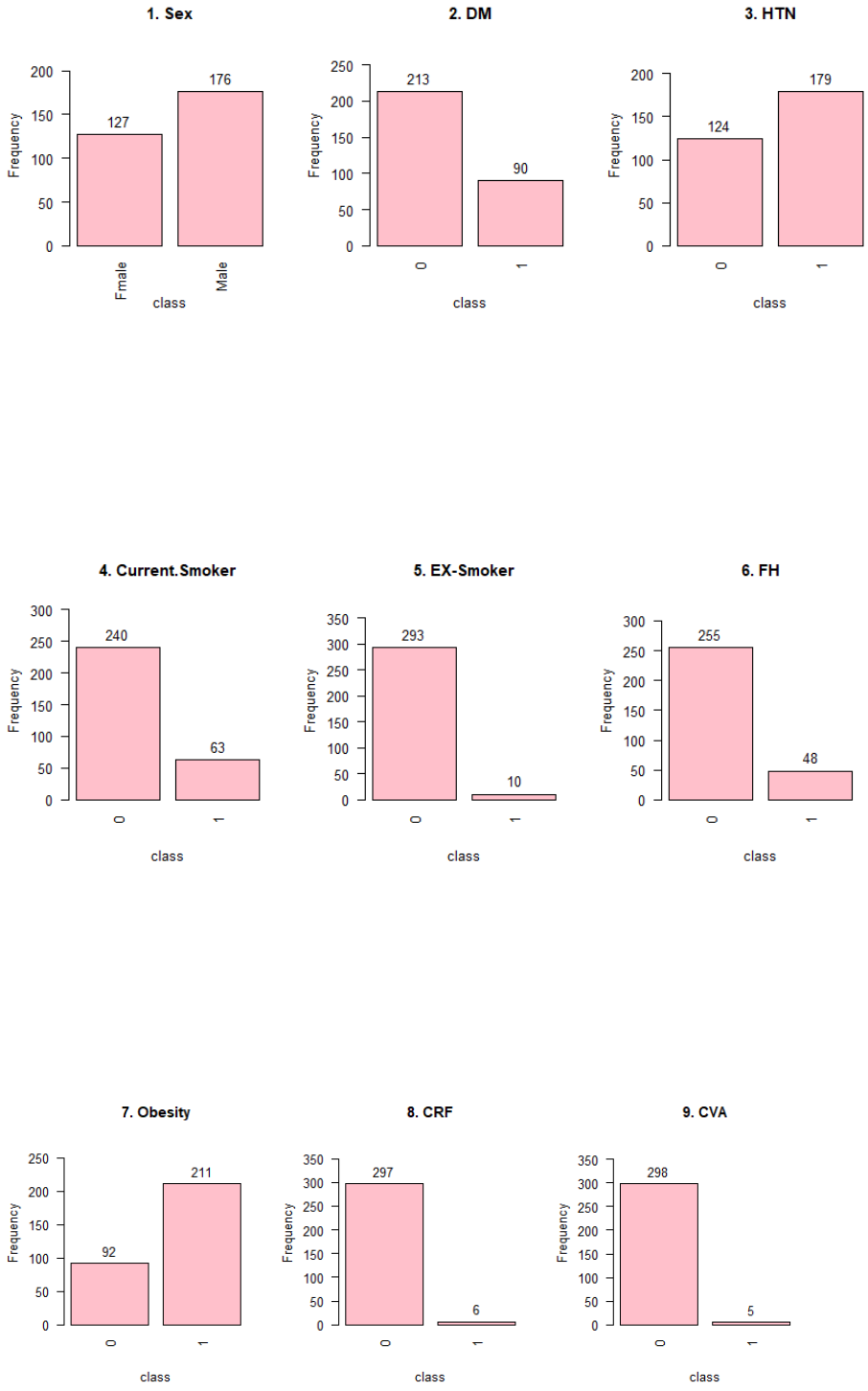

Figure 7: Bar plots showing the per class frequency for the categorical and binary independent variables 1 to 9. "Yes" and "no" classes are coded as 1 and 0, respectively.

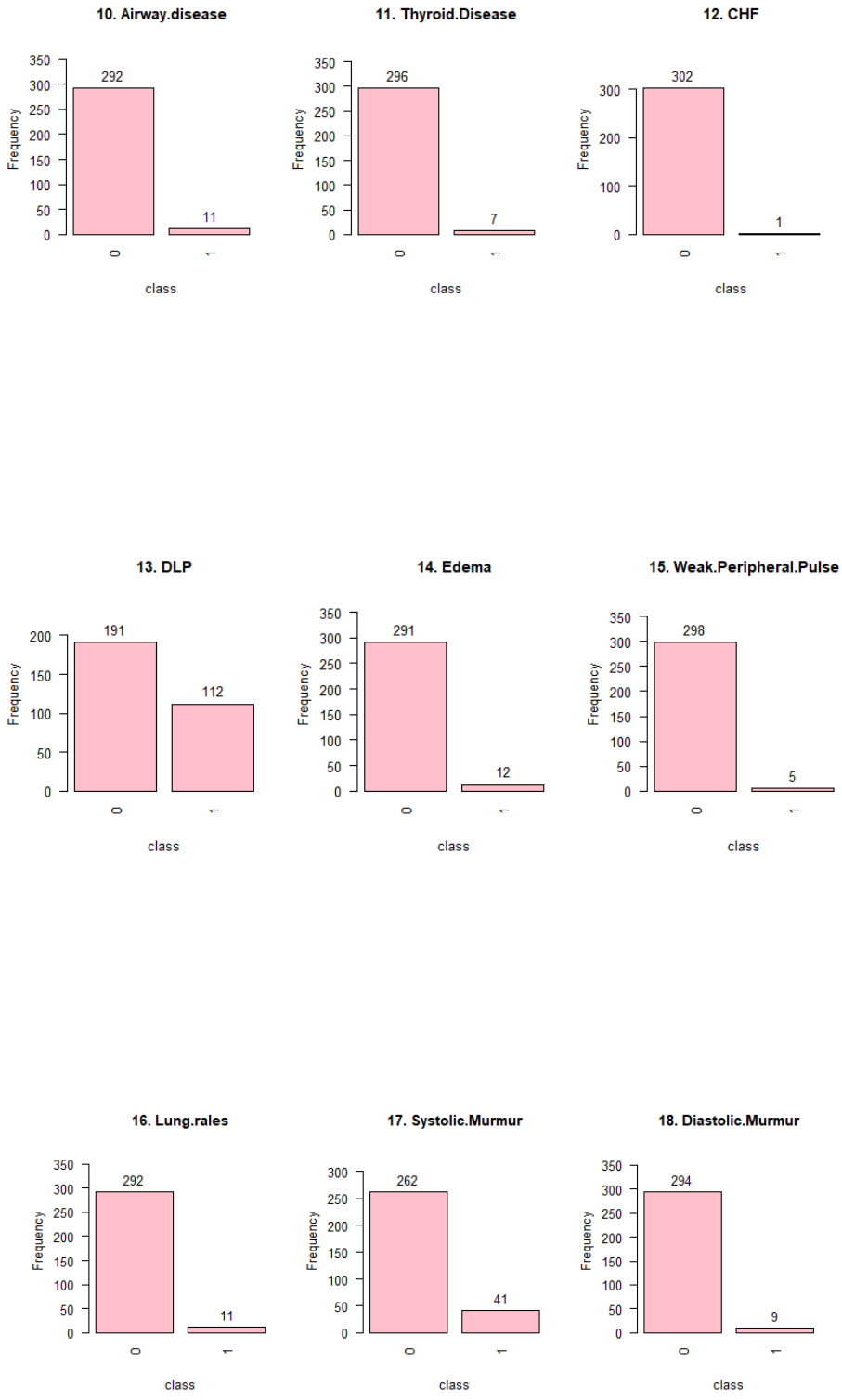

Figure 8: Bar plots showing the per class frequency for the categorical and binary independent variables 10 to 18. “Yes” and “no” classes are coded as 1 and 0, respectively.

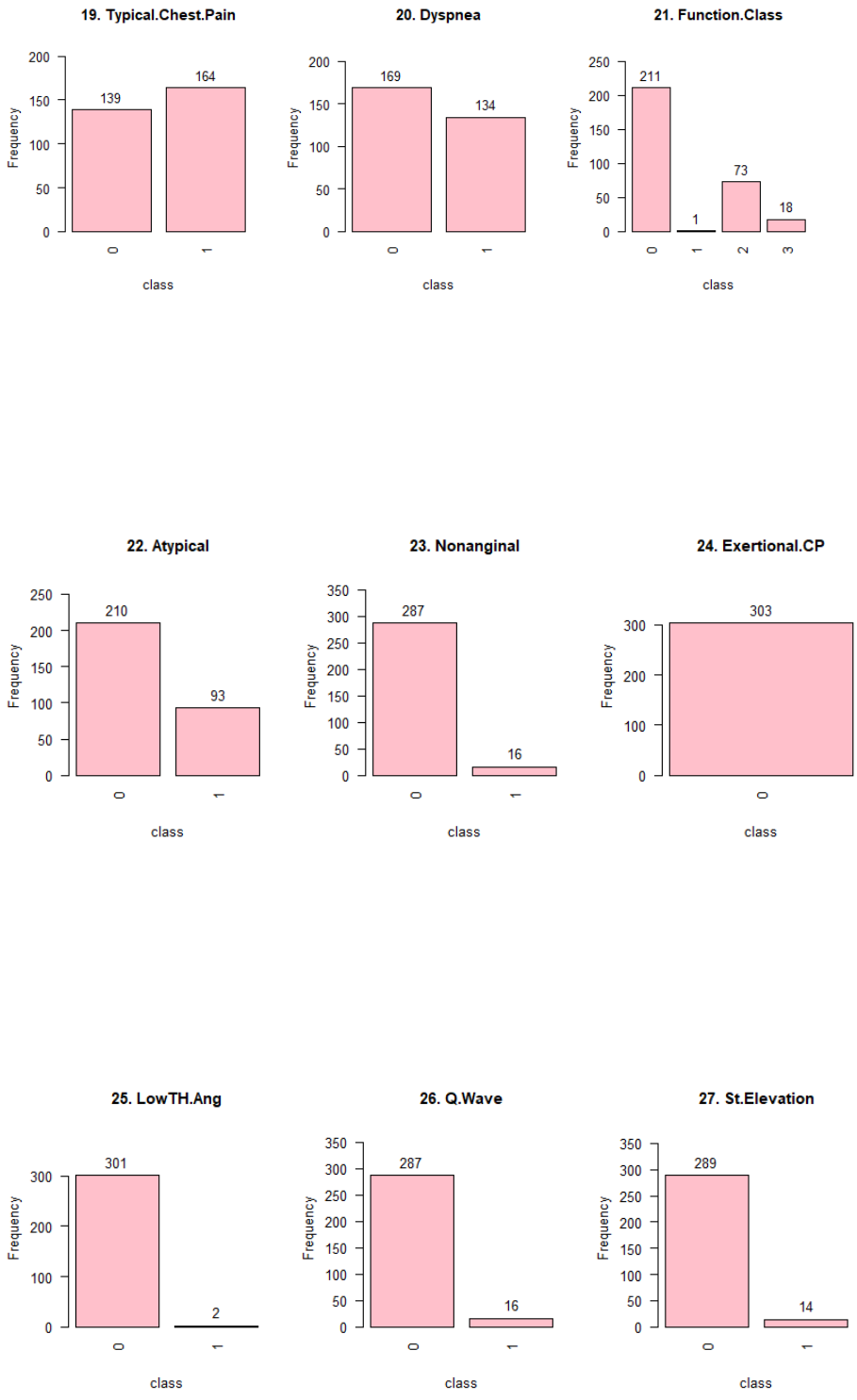

Figure 9: Bar plots showing the per class frequency for the categorical and binary independent variables 19 to 27. “Yes” and “no” classes are coded as 1 and 0, respectively.

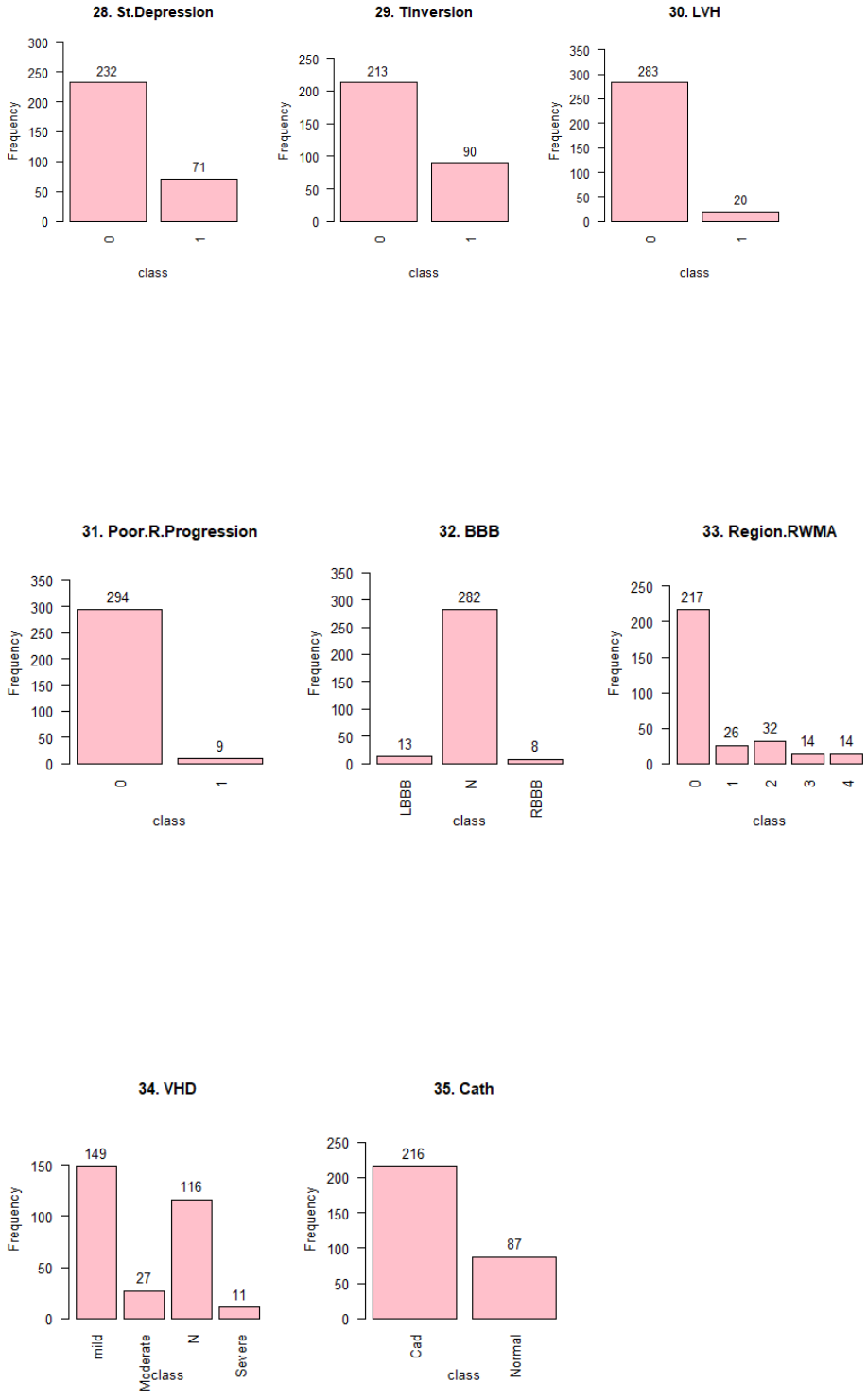

Figure 10: Bar plots showing the per class frequency for the categorical and binary independent variables 28 to 34 and the dependent variable "Cath" in frame 35. "Yes" and "no" classes are coded as 1 and 0, respectively.

H. Pearson correlation between numeric independent variables

In the Pearson's correlation plot in **Figure 11** below, we can see the pairwise correlation between the numeric independent variables. The numeric independent variables numbered 1 to 21 are: "Age", "Weight", "Length", "BMI", "BP", "PR", "FBS", "CR", "TG", "LDL", "HDL", "BUN", "ESR", "HB", "K", "Na", "WBC", "Lymph", "Neut", "PLT", and "EF-TTE". The Pearson's correlation metric is colour coded, where perfect, negative correlation (cor = -1) is shown in dark red and perfect, positive correlation (cor = +1) is shown in dark blue.

We can see that strong, positive correlation (shown in dark blue) exists between the numeric independent variables 2 and 4, namely Weight and BMI, respectively (which makes intuitive sense). We can also see that strong, negative correlation (shown in dark red) exists between the numeric independent variables 18 and 19, namely Lymph and Neut, respectively (which makes intuitive sense again).

Moderate, positive correlation (shown in a lighter shade of blue) exists between numeric independent variables 2 and 3, namely Weight and Length, respectively (which makes intuitive sense). Moderate positive correlation (shown in a lighter shade of blue) also exists between numeric independent variables 8 and 12, namely CR and BUN, respectively.

Moderate negative correlation (shown in a lighter shade of red) exists between numeric independent variables 13 and 14, namely ESR and HB, respectively. Moderate negative correlation (shown in a lighter shade of red) also exists between numeric independent variables 17 and 18, namely WBC and Lymph, respectively.

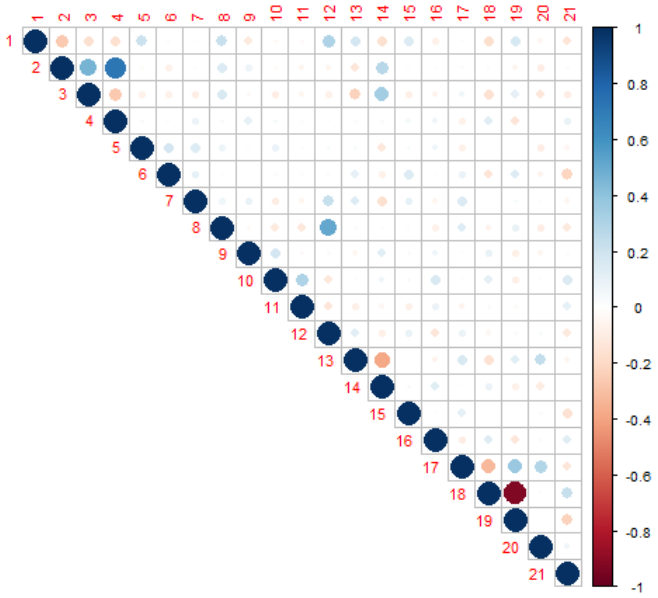

Figure 11: Correlation plot showing the pairwise correlation between the numeric independent variables. The numeric independent variables are numbered 1 to 21, namely: "Age", "Weight", "Length", "BMI", "BP", "PR", "FBS", "CR", "TG", "LDL", "HDL", "BUN", "ESR", "HB", "K", "Na", "WBC", "Lymph", "Neut", "PLT", and "EF-TTE". The Pearson's correlation metric is colour coded, where perfect, negative correlation (-1) is shown in dark red and perfect, positive correlation (+1) is shown in dark blue.

Methods

Data processing

A. *Removal of correlated variables*

Based on the correlation plot in **Figure 11** (above in the ***Exploratory Data Analysis section*** under ***H. Pearson correlation between numeric independent variables)*** and by using the findCorrelation package in r, I removed one numeric independent variable within a pair of correlated variables. For a pairwise correlation to be highly correlated, the Pearson correlation coefficient must be greater than 0.7. This was necessary since highly correlated variables introduce redundancy into the data set [3]. "Weight" was removed because it is highly correlated with "BMI" (cor = +0.725). In addition, "Lymph" was removed because it is highly correlated with "Neut" (cor = -0.923). As a result, I am left with 53 independent variables at this point, not 55 as in the original dataset.

B. *Removal of constant variables.*

In addition, the "Exertional.CP" independent variable was removed because all observations were grouped into 1 class, which can be seen in the histogram (above in **Figure 9: Frame 24** in the ***Exploratory Data Analysis*** section under ***G. Visualizing the distribution of the categorical and binary independent variables by means of bar plots)***. A variable where all its observations are grouped into one class does not have much variability in the dataset and is referred to a as a "constant" variable. This constant variable may not provide useful information to the analysis. As a result, I am left with 52 independent variables at this point, not 55 as in the original dataset.

C. *Number of features after data pre-processing*

I am left with 52 independent variables after removing highly correlated and constant variables.

D. *Discretization*

The numeric independent variables were divided into discrete classes. Thereafter, these independent variables were converted to factors. Certain categorical independent variables were also divided into discrete classes. **Table 16** below shows which independent variables were discretized. **Table 16** also shows the classes into which these independent variables were partitioned.

Table 16: Discretised features and their range of values[a]

Feature	Low	Normal	High
Cr	Cr < 0.7	0.7 ≤ Cr ≤ 1.5	Cr > 1.5
FBS	FBS < 70	70 ≤ FBS ≤ 105	FBS > 105
LDL		LDL ≤ 130	LDL > 130
HDL	HDL < 35	HDL ≥ 35	–
BUN	BUN < 7	7 ≤ BUN ≤ 20	BUN > 20
ESR		If male and ESR ≤ age/2 or if female and ESR ≤ age/2 + 5	If male and ESR > age/2 or if female and ESR > age/2 + 5
HB	If male and HB < 14 Or If female and HB < 12.5	If male and 14 ≤ HB ≤ 17 or if female and 12.5 ≤ HB ≤ 15	If male and HB > 17 or if female and HB > 15
K	K < 3.8	3.8 ≤ K ≤ 5.6	K > 5.6
Na	Na < 136	136 ≤ Na ≤ 146	Na > 146
WBC	WBC < 4000	4000 ≤ WBC ≤ 11,000	WBC > 11,000
PLT	PLT < 150	150 ≤ PLT ≤ 450	PLT > 450
EF	EF ≤ 50	EF > 50	
Region with RWMA2	–	Region with RWMA = 0	Region with RWMA ≠ 0
Age[b]		If male and age ≤ 45 or if female and age ≤ 55	If male and age > 45 or if female and age > 55
BP	BP < 90	90 ≤ BP ≤ 140	BP > 140
PulseRate	PulseRate < 60	60 ≤ PulseRate ≤ 100	PulseRate > 100
TG		TG ≤ 200	TG > 200
Function Class		1	2, 3, 4
Neut[c]	Neut < 40	40 ≤ Neut ≤ 60	Neut > 60
Underweight	**Healthy**	**Overweight**	**Obesity**
BMI[d] < 18.5	18.5 ≤ BMI < 25	25 ≤ BMI < 30	BMI ≥ 30
Length	Length < 160	160 ≤ Length < 180	Length ≥ 180

a. Classes were determined by the Braunwald heart book [4], unless otherwise specified.
b. Given that women under 55 years and men under 45 years are less affected by CAD, the range of age is partitioned at these values.
c. Classes for "Neut" were determined by UCSF Health [5].
d. Classes for "BMI" were determined by the Centres for Disease Prevention and Control [6].

E. Stratified sampling

The dataset is imbalanced in terms of the dependent variable, "Cath". The bar plot above in **Figure 10 (frame 35)** shows that the "Cad" class has 216 records, whereas the "normal" class only has 87 records. For this reason, the dataset was down sampled so that each class "Cad" and "normal" have 87 samples each. As a result, there are 174 observations which remained after down sampling.

F. Feature selection and variable importance

Various feature selection techniques were used. Agreement among all methods, suggests that the variable is not important and should be removed. However, if the variable is found to be unimportant but has class imbalance, and theoretically the variable is important – then I do not discard the variable. In all cases a score < 0.6 suggests that the variable is "unimportant" for that criterion.

i. Support Vector Machines (SVMs)

As suggested by R. Alizadehsani et al. (2013) [1], a linear-kernel SVM model was built using the processed CAD data. As stated by R. Alizadehsani et al. (2013) the SVM model

takes in numeric x-data, and so factors were converted back to numeric data types. Default parameters were used in the svm() function. The coefficients for the variables were then extracted and variables with a coefficient < 0.6 were assigned as the "unimportant" variables for criterion 1.

ii. Random Forests

A random forest was built on the processed CAD data, using the randomForest() function with default parameters. The varImp() function was then used to extract the variable importance scores and variables with scores < 0.6 were assigned as the "unimportant" variables for criterion 2.

iii. Area under Receiver-operating curve (ROC)

The maximum area under the curve (ROC AUC) was computed. The AUC for variables with a value < 0.6 were assigned as the "unimportant" variables for criterion 3.

iv. Classification tree model

A tree model was built using the rpart() function on the processed CAD data, The varImp() function was then used to determine the important variables. Variables with a score < 0.6 were assigned as the "unimportant" variables for criterion 4.

Finally, an "unimportant" variables list was created with the following variables identified by all of the variable selection criteria (above in points **i to iv**):

- "Ex.Smoker" *(removed from dataset)*
- "CRF" *(removed from dataset)*
- "CVA" *(removed from dataset)*
- "Airway.disease" *(removed from dataset)*
- "Thyroid.Disease" *(removed from dataset)*
- "CHF" *(Incorrectly identified as unimportant due to class imbalance)*
- "PR" *(Incorrectly identified as unimportant due to data sparsity)*
- "Weak.Peripheral.Pulse" *(Incorrectly identified as unimportant due to class imbalance)*
- "Lung.rales" *(Incorrectly identified as unimportant due to class imbalance)*
- "LVH" *(removed from dataset)*
- "Poor.R.Progression" *(removed from dataset)*
- "BBB" *(removed from dataset)*
- "Na" *(removed from dataset)*
- "PLT" *(removed from dataset)*

It is important to note the "CHF", "PR", "Weak.Peripheral.Pulse", and "Lung.rales" identified as unimportant here, are in fact important because they belong to the "Symptom and examination" group of independent variables. These variables are identified as unimportant because of class imbalance. "CHF" only has 1 positive case, "Weak.Peripheral.Pulse" has only 5 positive cases, and "Lung.rales" only has 11 positive cases (seen in the bar plot in **Figure 8**), while all other remaining cases out of the 303 cases are in the negative class. In addition, "PR" has a sparse distribution with 2 peaks (seen in the histogram in **Figure 4**). As a result, I did not remove "CHF", "PR", "Weak.Peripheral.Pulse", and "Lung.rales" due to their theoretical significance to CAD. I did however remove all other "unimportant" variables.

Additionally, I also removed the variables identified as unimportant by R. Alizadehsani et al. (2013) [1], namely:

- "Atypical" *(removed from dataset)*
- "Nonanginal" *(removed from dataset)*

- "FBS" *(removed from dataset)*
- "Diastolic Murmur" *(removed from dataset)*
- "Current Smoker" *(removed from dataset)*

It is important to note that these "unimportant" variables were identified by individual variable selection methods mentioned above in **i to iv.**

G. *Variables for further analysis*

As a result of data processing and feature selection, I am left with 37 independent variables in total, namely:

- "Age"
- "Length"
- "Sex"
- "BMI"
- "DM"
- "HTN"
- "FH"
- "Obesity"
- "CHF"
- "DLP"
- "BP"
- "PR"
- "Edema"
- "Weak Peripheral Pulse"
- "Lung.rales"
- "Systolic.Murmur"
- "Typical.Chest.Pain"
- "Dyspnea"
- "Function.Class"
- "LowTH.Ang"
- "Q.Wave"
- "St.Elevation"
- "St.Depression"
- "Tinversion"
- "CR"
- "TG"
- "LDL"
- "HDL"
- "BUN"
- "ESR"
- "HB"
- "K"
- "WBC"
- "Neut"
- "EF"
- "Region.RWMA"
- "VHD"

H. *Conversion of polynomial variables to binomial variables*

The Apriori association mining algorithm assumes a binary transaction database [7]. Consequently, it was necessary to convert polynomial variables to binomial variables.

As suggested by R. Alizadehsani et al. (2013) [1], all independent variables with greater than 2 classes were converted to variables with 2 classes, i.e. binary conversion. A rule was followed whereby, extreme classes such as "Low" and "High" were coded as 1. On the other hand, classes identified as "Normal", or which fall at the median or mean point are coded as 0.

In addition, all other independent variables that are already in binary form are coded as 0 and 1, where necessary. Finally, the "Sex" variable was coded so that Female = 1 and Male = 0.

Rule:
A "1" indicates an extreme class (either "low" or "high") which may have an effect on the dependent variable, whether the individual has CAD or not.
In contrast, a "0" indicates a healthy, normal or steady state that would not necessarily impact the dependent variable, whether an individual has CAD or not.

Apriori Association Rule Mining

A. *Transaction data*

After data processing, 174 records, 37 independent variables and 1 dependent variable remained, where all independent variables are binary coded. As stated before, extreme values which are within the "high" or "low" class are coded as 1 since these classes may have an effect on CAD. On the other hand, "normal", mean or median classes are coded as 0 since they will not have an effect on CAD.

The processed data set was then converted to a transaction data set.

B. *Exploratory data analysis on the transaction data*

i. Inspecting the data

I inspected the first 5 item sets in ascending order of their transaction ID. This was necessary to assess the binary nature and length of transactions.

ii. Support and Frequency

I then checked the support values of the first 5 single items. Thereafter, I assessed the frequency plots to determine the distribution of items according to their frequency and support values.

C. *Objective 1: Use association rule mining to determine the features that are mostly associated with CAD.*

i. Generating rules

In order to generate rules, I applied the Apriori algorithm on the binary transaction data, where minimum support = 0.01 and confidence = 0.8 in accordance with the order of magnitude specified by R. Alizadehsani et al. (2013) [1]. I then ensured that the RHS of all transactions is always "Cath = Cad", in order to determine features that are associated with CAD. I also ensured that rules have a maximum length of 4 to ensure that the maximum length of items on the LHS of the rule is always 3. This provides us with information on features that coexist and are associated with CAD, in addition to reducing computational intensity. Redundant rules were then removed so that only non-redundant rules are assessed. The rules were then sorted and assessed based on their support value, confidence, and lift values.

The resultant rules can be seen in the ***Results*** section below under ***B. Objective 1: Use association rule mining to determine the features that are mostly associated with CAD.***

D. *Objective 2: Find any other interesting association rules.*

i. Generating rules

In order to generate rules, I applied the Apriori algorithm on the binary transaction data, where minimum support was decreased from 0.01 (in the previous objective) to 0.001 in the current objective. This was necessary, in order to find rare association rules, which is commonly sort after in the medical field. In addition, the minimum confidence level was increased from 0.8 in the previous objective to 0.9 in this objective to produce high confidence rules. I then ensured that the RHS of all transactions is always "DM = 1", in order to find rules associated with diabetes mellitus. This is especially interesting because DM is highly associated with CAD [8]. In this implementation of the Apriori algorithm, I increased the minimum length of the rule until 0 rules were obtained. This indicated that the minimum, rule length should be 6, which is in fact also the maximum rule length. This is necessary because I am not only interested in the features that are associated with DM, but also the largest combination of features that interact and are cumulatively associated with DM.

The resultant rules can be seen in the ***Results*** section below under ***C. Objective 2: Find any other interesting association rules.***

Results

A. Exploratory data analysis on transaction data

i. Inspecting the data

Inspection of the first 5 itemsets suggest that transactions are long and contain many variables which are coded as 0. Below is an example of an itemset belonging to transaction 1:

{Age=1, Length=0, Sex=0, BMI=1, DM=1, HTN=1, FH=0, Obesity=1, CHF=0, DLP=1, BP=0, PR=0, Edema=0, Weak.Peripheral.Pulse=0, Lung.rales=0, Systolic.Murmur=0, Typical.Chest.Pain=0, Dyspnea=1, Function.Class=1, LowTH.Ang=0, Q.Wave=0, St.Elevation=0, St.Depression=0, Tinversion=0, CR=0, TG=1, LDL=0, HDL=1, BUN=0, ESR=0, HB=0, K=0, WBC=0, Neut=0, EF=1, Region.RWMA=1, VHD=1, Cath=Cad}

ii. Support and Frequency

Inspection of the frequency values for the first 5 single items, suggest that there is a large variance in the frequencies of single items throughout the database. The following frequencies were obtained from the first 5 items: (1) Age = 0 : 0.276, (2) Age = 1 : 0.724, (3) Length = 0 : 0.684, (4) Length = 1 : 0.316 and (5) Sex = 0 : 0.546 , where Sex = 0 indicates male and Sex = 1 indicates female.

The frequency plot in **Figure 12** below shows items with a support value greater than 0.7, where minimum support = 0.7 for ease of plotting. Features such as “Age”, and “Function Class” are coded as 1. I then checked the relative frequency plot of the first 25 items in the plot in **Figure 13** below. It is evident that items coded as 0 have the highest frequencies.

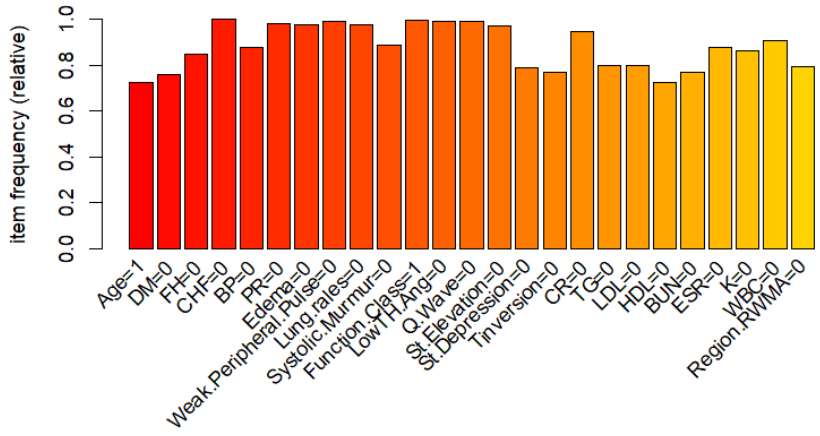

Figure 12: Frequency plot of items with a support value greater than 0.7. Minimum support = 0.7 for ease of plotting.

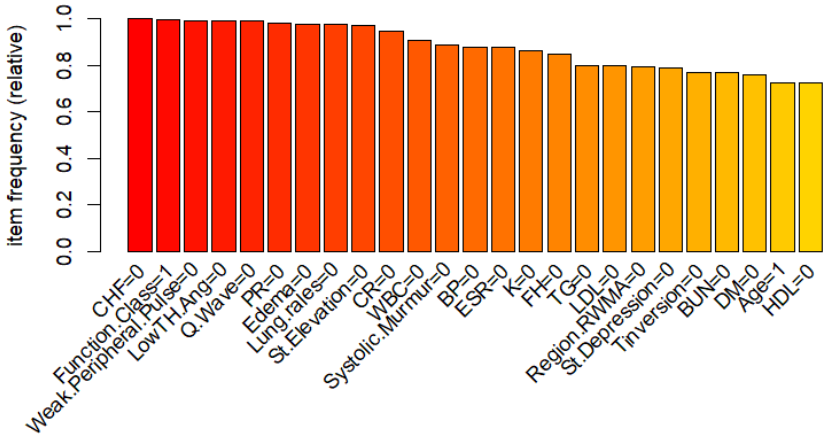

Figure 13: Relative frequency plot of first 25 items

B. Objective 1: Use association rule mining to determine the features that are mostly associated with CAD.

i. Resultant rules of association mining

I obtained a total of 8239 rules, however this was reduced to 516 rules after removing redundant rules.

By assessing the summary statistics in **Table 17** below, we can see that the minimum support is 0.011, while the maximum support is 0.339. The median support value of 0.029 is less than the mean support value of 0.038. In addition, the Q3 value of the support (0.046) is greater than the Q1 value of the support (0.023).

Table 17 also shows the summary statistics for the confidence values. The minimum confidence is 0.8, while the maximum confidence is 1, suggesting that high confidence rules were generated. The median confidence value of 0.857 is less than the mean confidence value of 0.886. In addition, the Q3 value of the confidence (1) is greater than the Q1 value of the confidence (0.8).

Finally, **Table 17** also shows the summary statistics for the lift values. The minimum lift is 1.6, while the maximum lift is 2, suggesting that all rules have L > 1 and so X occurs together with Y more often than expected. The median lift value of 1.714 is less than the mean lift value of 1.773. In addition, the Q3 value of the lift value (2) is greater than the Q1 value of the lift value (1.6).

Table 17: Summary statistics of rules generated for objective 1.

	Support	**Confidence**	**Lift**
Minimum	0.011	0.800	1.600
Q1	0.023	0.800	1.600
Median	0.029	0.857	1.714
Mean	0.038	0.886	1.773
Q3	0.046	1.00	2.00
Maximum	0.339	1.00	2.00

ii. Features mostly associated with CAD

In **Table 18** below, we see the most commonly associated rules with CAD when sorting rules by decreasing value of support. In **Figure 14** below we can see the graph of the first 5 rules mostly

associated with CAD. It is clear that the rules with large lift values (dark orange) have smaller support values (small circle sizes) that point towards the centre, i.e., pointing to CAD.

Based on **Table 18**, the most commonly associated features with CAD are:

- *Typical chest pain = True*
- *Age = True.*
  In other words, classified as high-risk based on discretization, where the individual is male with an age > 45 or if the individual is female with an age > 55.
- *Hypertension ("HTN") = True*
- *EF (ejection fraction) (%) = True.*
  In other words, EF ≤ 50% (classified as "low" based on discretization).
- *Region with RWMA (regional wall motion abnormality) = True.*
  In other words, the individual should fall in class 1, 2, 3 or 4 (classified as "High" based on discretization).
- *Diabetes ("DM") = True.*
- *Tinversion = True*
- *Haemoglobin ("HB" measured in g/dl)* = True.
  In other words, "high" or "low" based on discretization, where "high" indicates males with HB > 17 g/dl or females with HB > 15 g/dl and "low" indicates males with HB < 14 g/dl or females with HB < 12.5 g/dl.
- *Valvular heart disease ("VHD") = True.*
  In other words, the individual has VHD, with a status of mild, moderate, or severe.
- *DLP (*Dyslipidemia*) = True*
- *Triglycerides ("TG" measured in mg/dl) = True.*
  In other words, TG > 200 mg/dl based on discretization.
- *Sex = Male*
- *St Depression = True*
- *Neutrophil) ("Neut"measured in %) = True.*
  In other words, Neut > 60% ("High") or Neut < 40% ("Low") based on discretization.

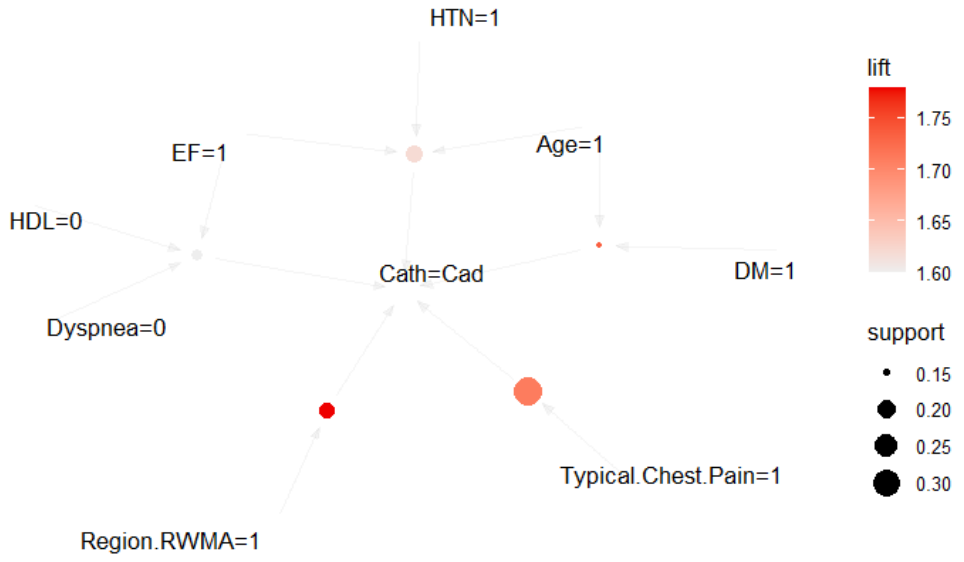

Figure 14: Graph showing relationship between first 5 rules sorted by descending order of support value

In **Table 18** below, we see that rules with high support values (and high counts) generally have lower confidence values (close to the minimum confidence level of 0.8). We still see that all lift values > 1, suggesting that X occurs together with Y more often than expected. In addition, we see that the coverage for these rules are relatively small, suggesting that the antecedent (LHS of the rule) do not occur frequently in the transaction database, i.e., P(LHS of rule) is small.

Table 18: Twenty-five most commonly associated rules with CAD, sorted by decreasing order of support value. Here minimum support = 0.01 and minimum confidence = 0.8.

Rules	Support	Confidence	Coverage	Lift	Count
{Typical.Chest.Pain=1} => {Cath=Cad}	0,339	0,855	0,397	1,710	59
{Age=1,HTN=1,EF=1} => {Cath=Cad}	0,195	0,810	0,241	1,619	34
{Region.RWMA=1} => {Cath=Cad}	0,184	0,889	0,207	1,778	32
{Dyspnea=0,HDL=0,EF=1} => {Cath=Cad}	0,161	0,800	0,201	1,600	28
{Age=1,DM=1} => {Cath=Cad}	0,149	0,867	0,172	1,733	26
{DM=1,HTN=1} => {Cath=Cad}	0,138	0,800	0,172	1,600	24
{Length=0,DM=1} => {Cath=Cad}	0,138	0,800	0,172	1,600	24
{Tinversion=1,EF=1} => {Cath=Cad}	0,132	0,821	0,161	1,643	23
{DM=1,HB=1} => {Cath=Cad}	0,126	0,846	0,149	1,692	22
{Length=0,Tinversion=1} => {Cath=Cad}	0,121	0,808	0,149	1,615	21
{DM=1,EF=1} => {Cath=Cad}	0,121	0,840	0,144	1,680	21
{PR=0,EF=1,VHD=0} => {Cath=Cad}	0,115	0,800	0,144	1,600	20
{HTN=1,HB=1,EF=1} => {Cath=Cad}	0,115	0,833	0,138	1,667	20
{DM=1,VHD=1} => {Cath=Cad}	0,109	0,826	0,132	1,652	19
{HTN=1,DLP=1,HB=1} => {Cath=Cad}	0,109	0,905	0,121	1,810	19
{HTN=1,DLP=1,Dyspnea=0} => {Cath=Cad}	0,109	0,864	0,126	1,727	19
{HTN=1,TG=1} => {Cath=Cad}	0,103	0,947	0,109	1,895	18
{Sex=0,DM=1} => {Cath=Cad}	0,103	0,857	0,121	1,714	18
{HTN=1,DLP=1,EF=1} => {Cath=Cad}	0,103	0,857	0,121	1,714	18
{HTN=1,St.Depression=1} => {Cath=Cad}	0,098	0,895	0,109	1,789	17
{St.Depression=1,EF=1} => {Cath=Cad}	0,098	0,810	0,121	1,619	17
{Tinversion=1,Neut=1} => {Cath=Cad}	0,098	0,944	0,103	1,889	17
{Sex=0,Tinversion=1} => {Cath=Cad}	0,098	0,810	0,121	1,619	17
{DM=1,Neut=1} => {Cath=Cad}	0,098	0,944	0,103	1,889	17
{Age=1,St.Depression=1,HDL=0} => {Cath=Cad}	0,098	0,810	0,121	1,619	17

iii. Rules with high confidence, high lift and small support values

Interesting rules are found at the high confidence, low support boundary, which is also where the lift values are high. This makes intuitive sense since biomedical research looks for associations that are rare but may be the causal associations.

In **Figure 15 (a)** below, we see that interesting rules have a high lift value (shown by darker orange colours) and are found where confidence is high, and support is low. As a result, the graph of association rules in **Figure 15 (b)** shows the first 5 rules sorted by descending order of confidence, where we can see that all rules point to CAD as expected and all rules have a lift value of 2 (shown by a dark shade of orange) and confidence = 1 shown in **Figure 15 (a)**. The rule with the largest support (largest circle size) is:

1. {DM=1, St.Depression=1}→ {Cath=Cad}, followed by
2. {DM=1, TG=1}→ {Cath=Cad},
3. {BP=1, HB=1, K=0}→ [Cath=Cad},
4. {FH=1, Tinversion=1}→ {Cath=Cad} and
5. {DM=1, Tinversion=1} → {Cath=Cad}

In **Figure 15 (c),** we can see the grouped plot of rules, where support is shown by the size of the circle and lift is shown by the intensity of the colour. We can see a trend where many rules with large support (large circle sizes) and high lift values (dark orange colour) contain Edema = 1, PR = 1, HTN= 1, TG= 1, Neut=1, DM=1 and HB=1 or Tinversion = 1 on the LHS of the rule.

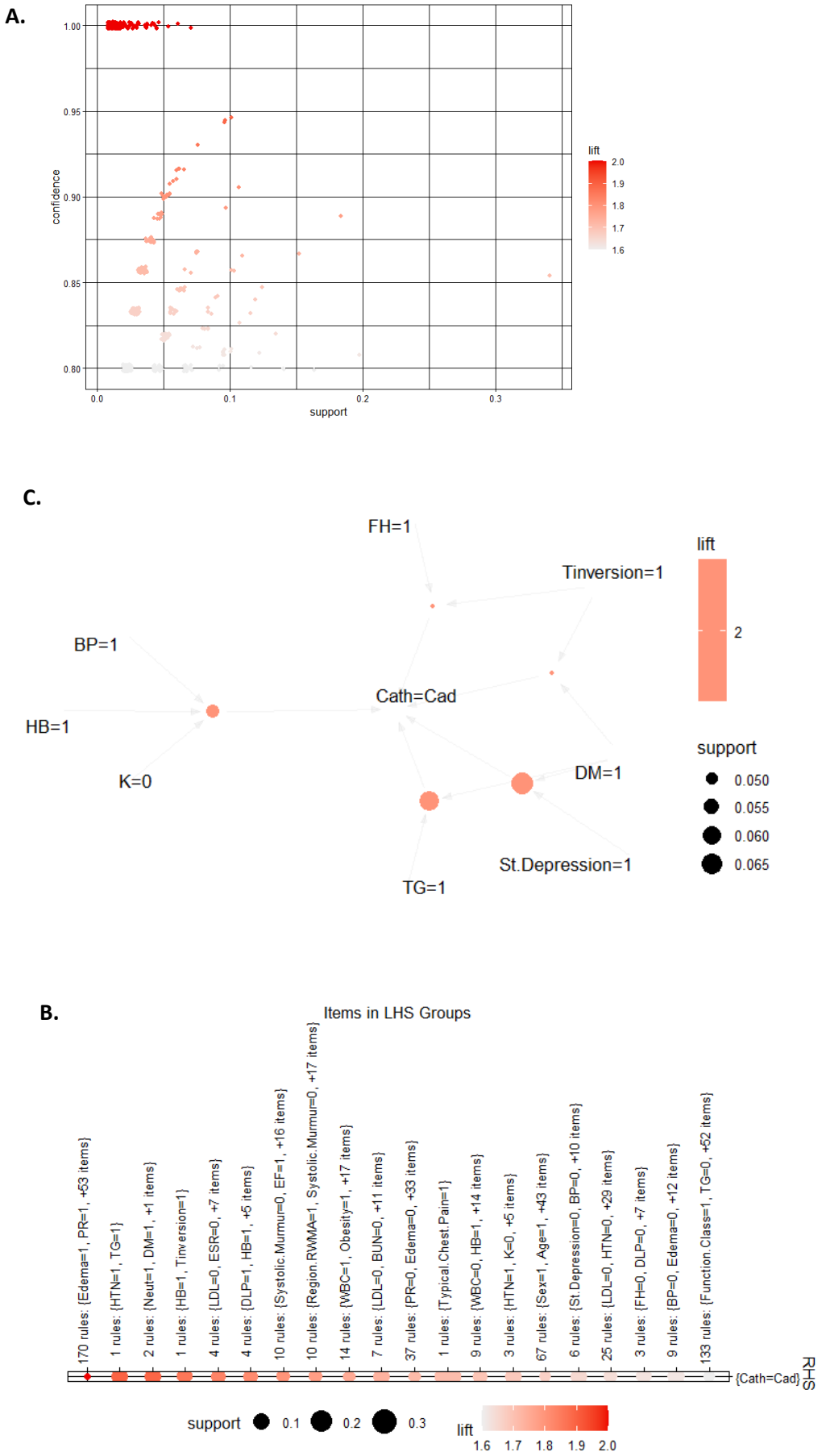

Figure 15: Finding interesting rules associated with CAD. **(a)** Support vs confidence plot showing lift values by colour intensity. **(b)** Graph showing relationship between first 5 rules sorted by descending order of confidence **(c)** Grouped plot showing rules, where support is shown by the size of the circle and lift is shown by the intensity of the colour.

iv. Interesting rules

Table 19 below shows interesting rules associated with CAD, sorted by decreasing order of lift values. We see that the rules sorted by decreasing order of confidence or lift values produce the same rules. We also see that high confidence/high lift rules have a small support value and low count. In addition, we see that the coverage for these rules are relatively small, suggesting that the antecedent (LHS of the rule) do not occur frequently in the transaction database, i.e., P(LHS of rule) is small. Worth noting is that all lift values = 2 and so L > 1, suggesting that X occurs together with Y more often than expected. Finally, we can see that all rules have confidence = 1 suggesting that these are high confidence rules.

Table 19: Twenty-five most interesting rules found associated with CAD, sorted by decreasing order of lift values*, where minimum support = 0.01 and minimum confidence = 0.8

Rules	Support	Confidence	Coverage	Lift	Count
{DM=1,St.Depression=1} => {Cath=Cad}	0,069	1,000	0,069	2	12
{DM=1,TG=1} => {Cath=Cad}	0,063	1,000	0,063	2	11
{BP=1,HB=1,K=0} => {Cath=Cad}	0,052	1,000	0,052	2	9
{FH=1,Tinversion=1} => {Cath=Cad}	0,046	1,000	0,046	2	8
{DM=1,Tinversion=1} => {Cath=Cad}	0,046	1,000	0,046	2	8
{TG=1,Neut=1,EF=1} => {Cath=Cad}	0,046	1,000	0,046	2	8
{Tinversion=1,WBC=1} => {Cath=Cad}	0,040	1,000	0,040	2	7
{DLP=1,WBC=1} => {Cath=Cad}	0,040	1,000	0,040	2	7
{DM=1,K=1} => {Cath=Cad}	0,040	1,000	0,040	2	7
{Age=1,FH=1,EF=1} => {Cath=Cad}	0,040	1,000	0,040	2	7
{Age=1,FH=1,VHD=1} => {Cath=Cad}	0,040	1,000	0,040	2	7
{Length=0,St.Depression=1,Neut=0} => {Cath=Cad}	0,040	1,000	0,040	2	7
{BP=1,TG=1} => {Cath=Cad}	0,034	1,000	0,034	2	6
{Tinversion=1,ESR=1} => {Cath=Cad}	0,034	1,000	0,034	2	6
{Length=0,Sex=1,LDL=1} => {Cath=Cad}	0,034	1,000	0,034	2	6
{Age=1,TG=1,BUN=1} => {Cath=Cad}	0,034	1,000	0,034	2	6
{Sex=0,St.Depression=1,Neut=0} => {Cath=Cad}	0,034	1,000	0,034	2	6
{St.Elevation=1} => {Cath=Cad}	0,029	1,000	0,029	2	5
{CR=1,Neut=1} => {Cath=Cad}	0,029	1,000	0,029	2	5
{CR=1,K=0} => {Cath=Cad}	0,029	1,000	0,029	2	5
{HB=0,WBC=1} => {Cath=Cad}	0,029	1,000	0,029	2	5
{Length=0,ESR=1,EF=1} => {Cath=Cad}	0,029	1,000	0,029	2	5
{Length=0,ESR=1,VHD=1} => {Cath=Cad}	0,029	1,000	0,029	2	5
{Age=1,Length=0,ESR=1} => {Cath=Cad}	0,029	1,000	0,029	2	5
{Length=0,FH=1,LDL=1} => {Cath=Cad}	0,029	1,000	0,029	2	5

***Sorting by decreasing order of lift values produces the same rules as sorting by decreasing order of confidence.**

C. *Objective 2: Find any other interesting association rules.*

i. Resultant rules

A total of 554258 rules were mined in total. **Table 20** below provides the summary statistics for mined rules.

By assessing the summary statistics in **Table 20** below, we can see that the minimum support is 0.006, while the maximum support is 0.069. The median support value of 0.006 is less than the mean support value of 0.007. In addition, the Q3 value of the support (0.006) is equal to the Q1 value of the support (0.006).

Table 20 also shows the summary statistics for the confidence values. The minimum confidence is 0.9, while the maximum confidence is 1, suggesting that high confidence rules were generated. The median confidence = mean confidence = 1. In addition, the Q3 value = Q1 value = 1.

Finally, **Table 20** also shows the summary statistics for the lift values. The minimum lift is 3.729, while the maximum lift is 4.143, suggesting that all rules have L >>> 1 and so X occurs together with Y more often than expected. The median lift value of 4.143 is equal to the mean lift value of 4.143. In addition, the Q3 of the lift value = Q1 of the lift value = 4.143.

Table 20: Summary statistics of rules generated for objective 2.

	Support	**Confidence**	**Lift**
Minimum	0.006	0.900	3.729
Q1	0.006	1.00	4.143
Median	0.006	1.00	4.143
Mean	0.007	1.00	4.143
Q3	0.006	1.00	4.143
Maximum	0.069	1.00	4.143

ii. Rules with high confidence, high lift and small support values

In **Figure 16 (a)** below we see the scatter plot showing the support values along the x-axis vs the confidence along the y-axis. The lift values are colour coded based on the colour bar on the right. It is evident that rules with large lift values (L > 1), also have large confidence values at confidence= 1, but these rules have small support values.

The graph in **Figure 16 (b)** below shows the relationship between the first 5 rules sorted in descending order of confidence. All five rules have a lift value of L = 4.143. A lift value greater than 1 (L > 1) implies that X occurs together with Y more often than expected. All five rules also have confidence = 1, suggesting that these rules are of high confidence. In other words, the rule containing X and Y is likely correct. All five rules also have support = 0.006, which is relatively small suggesting that the fraction of transactions which contain X and Y is quite rare.

The rules in **Figure 16(b**) can be assessed by the first 5 rows of **Table 21** below.

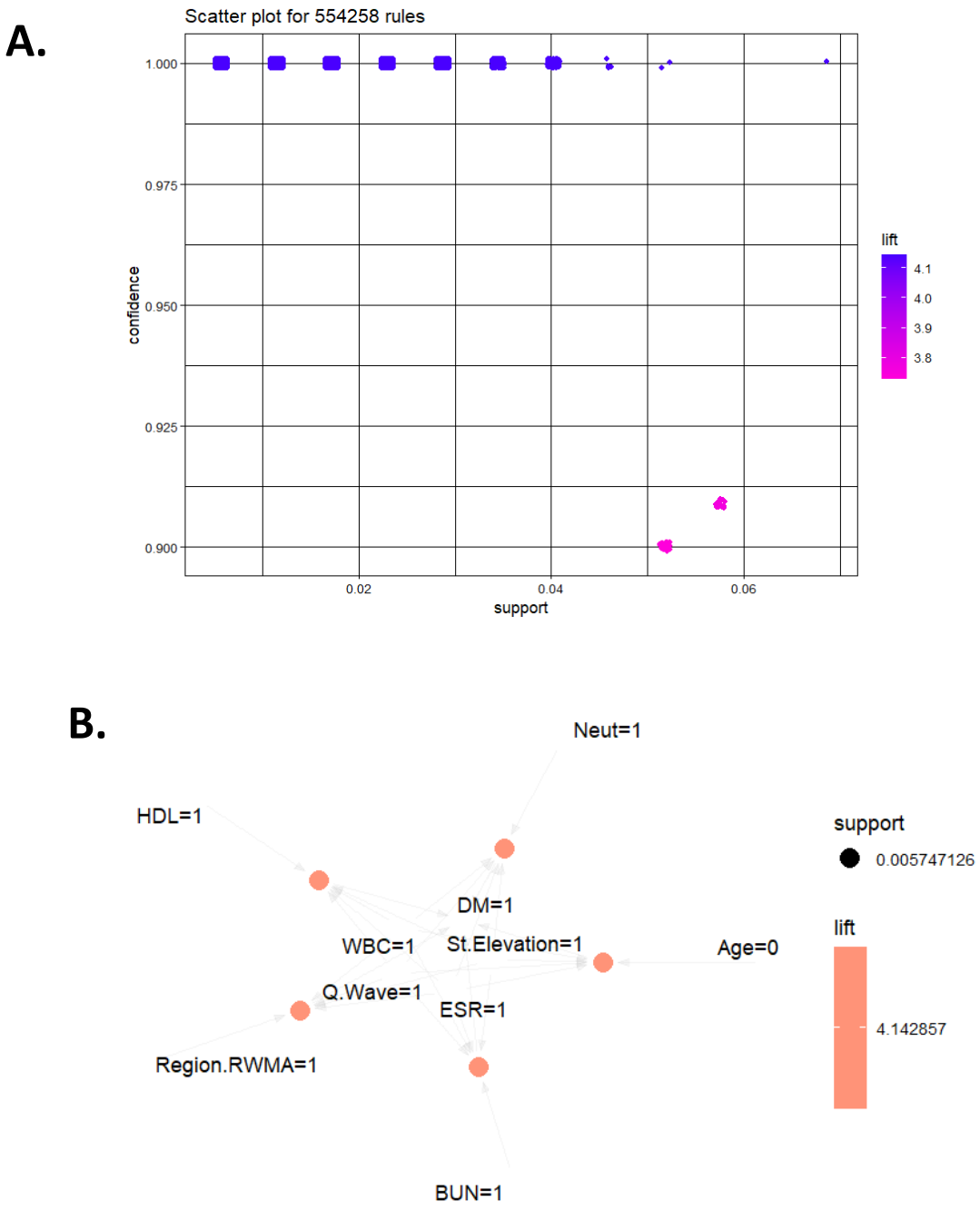

Figure 16: Finding interesting rules associated with DM, where DM is associated with CAD. **(a)** Support vs confidence plot showing lift values by colour intensity **(b)** Graph showing relationship between first 5 rules sorted by descending order of confidence

iii. **Interesting rules**

Table 21 below shows interesting rules, sorted by descending order of lift. We see that all 25 rules have a confidence value = 1.00, suggesting that high confidence rules were generated. In other words, the large confidence values suggest that rules containing X and Y are likely correct. In addition, all 25 rules have a small support value of 0.006, suggesting that there is a small fraction of transactions that contain both X and Y. These rules are not as common. All 25 rules also have a large lift value of 4.143, where L > 1, suggesting that X occurs together with Y more often than expected and X and Y are highly correlated. All 25 rules also have a coverage = 0.006 which is relatively low, suggesting that the LHS of the rule occurs quite rarely in the database, i.e., LHS has low support.

Although my rules are sorted in descending order of lift value, it is evident that the same rules are presented when the rules are sorted by decreasing order of confidence. We also see a trend where these high confidence rules have high lift values but small support values, and low counts.

Table 21: Twenty-five most interesting association rules, sorted by descending order of lift values*, where minimum support = 0.001 and minimum confidence = 0.9.

LHS	RHS	Support	Confidence	Coverage	Lift	Count
{Q.Wave=1, St.Elevation=1, ESR=1, WBC=1, Region.RWMA=1}	{DM=1}	0,006	1,000	0,006	4,143	1
{Q.Wave=1, St.Elevation=1, BUN=1, ESR=1, WBC=1}	{DM=1}	0,006	1,000	0,006	4,143	1
{Age=0, Q.Wave=1, St.Elevation=1, ESR=1, WBC=1}	{DM=1}	0,006	1,000	0,006	4,143	1
{Q.Wave=1, St.Elevation=1, HDL=1, ESR=1, WBC=1}	{DM=1}	0,006	1,000	0,006	4,143	1
{Q.Wave=1, St.Elevation=1, ESR=1, WBC=1, Neut=1}	{DM=1}	0,006	1,000	0,006	4,143	1
{Sex=1, Q.Wave=1, St.Elevation=1, ESR=1, WBC=1}	{DM=1}	0,006	1,000	0,006	4,143	1
{Dyspnea=1, Q.Wave=1, St.Elevation=1, ESR=1, WBC=1}	{DM=1}	0,006	1,000	0,006	4,143	1
{Q.Wave=1, St.Elevation=1, ESR=1, HB=1, WBC=1}	{DM=1}	0,006	1,000	0,006	4,143	1
{Q.Wave=1, St.Elevation=1, ESR=1, WBC=1, Cath=Cad}	{DM=1}	0,006	1,000	0,006	4,143	1
{HTN=1, Q.Wave=1, St.Elevation=1, ESR=1, WBC=1}	{DM=1}	0,006	1,000	0,006	4,143	1
{Q.Wave=1, St.Elevation=1, ESR=1, WBC=1, EF=1}	{DM=1}	0,006	1,000	0,006	4,143	1
{DLP=0, Q,Wave=1, St.Elevation=1, ESR=1, WBC=1}	{DM=1}	0,006	1,000	0,006	4,143	1
{Q.Wave=1, St.Elevation=1, ESR=1, WBC=1, VHD=1}	{DM=1}	0,006	1,000	0,006	4,143	1
{Typical,Chest,Pain=0, Q.Wave=1, St.Elevation=1, ESR=1, WBC=1}	{DM=1}	0,006	1,000	0,006	4,143	1
{Length=0, Q.Wave=1, St.Elevation=1, ESR=1, WBC=1}	{DM=1}	0,006	1,000	0,006	4,143	1
{BMI=1, Q.Wave=1, St.Elevation=1, ESR=1, WBC=1}	{DM=1}	0,006	1,000	0,006	4,143	1
{Obesity=1, Q.Wave=1, St.Elevation=1, ESR=1, WBC=1}	{DM=1}	0,006	1,000	0,006	4,143	1
{Q.Wave=1, St.Elevation=1, Tinversion=0, ESR=1, WBC=1}	{DM=1}	0,006	1,000	0,006	4,143	1
{Q.Wave=1, St.Elevation=1, St,Depression=0, ESR=1, WBC=1}	{DM=1}	0,006	1,000	0,006	4,143	1
{Q.Wave=1, St.Elevation=1, LDL=0, ESR=1, WBC=1}	{DM=1}	0,006	1,000	0,006	4,143	1
{Q.Wave=1, St.Elevation=1, TG=0, ESR=1, WBC=1}	{DM=1}	0,006	1,000	0,006	4,143	1
{FH=0, Q.Wave=1, St.Elevation=1, ESR=1, WBC=1}	{DM=1}	0,006	1,000	0,006	4,143	1
{Q.Wave=1, St.Elevation=1, ESR=1, K=0, WBC=1}	{DM=1}	0,006	1,000	0,006	4,143	1
{BP=0, Q.Wave=1, St.Elevation=1, ESR=1, WBC=1}	{DM=1}	0,006	1,000	0,006	4,143	1
{Systolic,Murmur=0, Q.Wave=1, St.Elevation=1, ESR=1, WBC=1}	{DM=1}	0,006	1,000	0,006	4,143	1

*The rules presented are the same when sorting by decreasing order of confidence

Discussion

The Apriori algorithm requires a binary transaction database as input. For this reason, it was necessary to discretize numeric variables into discrete classes, convert polynomial classes to binomial classes, and convert classes to binary variables. For example, “high” and “low” classes which may influence CAD were coded as 1, while “normal” or average classes were coded as 0 since they will not influence CAD. In addition, “yes” was coded as 1 and “no” was coded as 0. For the sex variable, “Female” was coded as 1 and “Male” was coded as 0.

It is important to note that association does not infer causality when assessing the association rules.

In objective 1, I found that 14 features among the first 25 rules sorted by decreasing order of support, were mostly associated with CAD. Although these rules have large support values (i.e., a large fraction of rules containing both X and Y), these rules had relatively small confidence values at the boundary of the minimum support value (minsup = 0.8). However, the confidence is still greater than 0.5 suggesting that the rules containing X and Y are likely correct. Interestingly, these rules have lift values greater than 1, which suggests that X and Y are highly correlated, and X occurs together with Y more often than expected.

Among the 14 features with a large support value and identified as the features most commonly associated with CAD: having typical chest pain, senior age, and having hypertension were the top three features associated with CAD, among other features. These features along with the other features are in fact theoretically associated with CAD.

I then investigated the most interesting rules associated with CAD and found that the following rule was the most interesting: {DM=1,St.Depression=1} → {Cath=Cad} with a support of 0.069, confidence = 1, and lift = 2. Diabetes mellitus (DM) is a well-known risk factor of CAD, where patients with DM have a greater likelihood of developing CAD. In addition, ST segment depression on an ECG may indicate myocardial ischemia, which is the reduced blood flow to the heart muscle (a common symptom of CAD).

In the top 25 rules ordered in descending order of lift value, I found that DM occurred quite frequently in the top few rules. This encouraged me to investigate DM in objective 2.

In objective 2, I decreased the minimum support to 0.001 and increased the minimum confidence to 0.9 to find rare associations which are most likely correct. Among the 25 interesting rules found and ordered by descending order of lift, the most interesting rule was:

{Q.Wave=1, St.Elevation=1, ESR=1, WBC=1, Region.RWMA=1} → {DM=1}

Q waves are abnormal patterns observed on an electrocardiogram (ECG) and can indicate previous myocardial infarction (heart attack), which is a risk factor of CAD. Individuals with diabetes have higher prevalence of Q waves, relative to healthier groups [9]. St segment changes on an ECG may indicate myocardial ischemia, which is the reduced blood flow to the heart muscle and is a common symptom of CAD. Diabetic individuals present St Elevation myocardial infarction [10]. ESR (erythrocyte sedimentation rate) is a blood test which measures body inflammation, where inflammation plays a primary role in developing atherosclerosis, which is an underlying cause of many CAD cases. Diabetic individuals have elevated ESR levels in the absence of overt infection [11]. WBC (white blood cell) count is another blood test which may indicate inflammation in the body and may indicate the progression of atherosclerosis which is an underlying cause of CAD. WBC count is associated with macro- and microangiopathic complication in individuals with type 2 diabetes [12]. Region RWMA (Regional wall motion abnormality) refers to the abnormal movement of the heart muscle in response to contraction which is also associated with CAD. Region RWMA was also found to be associated with individuals with diabetes mellitus [13]

Conclusion

In this project, I identified 14 features most commonly associated with CAD at a minimum support of 0.01 and minimum confidence of 0.8. I also found some interesting association rules by decreasing the minimum support from 0.01 to 0.001 and increasing the minimum confidence from 0.8 to 0.9, and by looking for features associated with DM, since DM is highly associated with CAD.

References

1. R. Alizadehsani et al. (2013), "A data mining approach for diagnosis of coronary artery disease," Computer Methods and Programs in Biomedicine, vol.111, no.1, pp.52-61.
2. Roohallah Alizadehsani (2017). Z-Alizadeh Sani. Mendeley Data, [online]
3. Andrews.edu. (2023). Correlation Coefficients. [online] Available at: https://www.andrews.edu. [Accessed 7 Oct. 2023].
4. Google Books. (2013). Braunwald's Heart Disease E-Book. [online] Available at: https://books.google.co.za. [Accessed 7 Oct. 2023].
5. ucsfhealth.org. (2019). Blood Differential. [online] Available at: https://www.ucsfhealth.org/medical-tests [Accessed 7 Oct. 2023].
6. CDC (2022). Defining Adult Overweight & Obesity . [online] Centers for Disease Control and Prevention. Available at: https://www.cdc.gov [Accessed 7 Oct. 2023].
7. Kononenko, I. (2007). Symbolic Learning. Elsevier eBooks, [online] pp.227–258.
8. Matias Mäenpää, Kujala, I., E Harjulahti, Iida Stenström, Wail Nammas, Juhani Knuuti, Antti Saraste and Teemu Maaniitty (2023). The impact of diabetes on the relationship of coronary artery disease and outcome: a study using multimodality imaging. Cardiovascular Diabetology, [online] 22(1).
9. Li, Y., Dawood, F.Z., Chen, H., Jain, A., Walsh, J., Alonso, Á., Lloyd-Jones, D.M. and Soliman, E.Z. (2013). Minor Isolated Q Waves and Cardiovascular Events in the MESA Study. The American Journal of Medicine, [online] 126(5), pp.450.e9–450.e16.
10. Megaly, M., Schmidt, C., Dworak, M., Garberich, R., Stanberry, L., Sharkey, S., Brilakis, E.S., Aguirre, F.V., Pacheco, R., Tannenbaum, M., Coulson, T., Smith, T.D., Henry, T.D. and García, S. (2022). Diabetic Patients Who Present With ST-Elevation Myocardial Infarction. Cardiovascular Revascularization Medicine, [online] 38, pp.89–93.
11. Elias AN;Domurat E (2021). Erythrocyte sedimentation rate in diabetic patients: relationship to glycosylated hemoglobin and serum proteins. Journal of medicine, [online] 20(3-4). Available at: https://pubmed.ncbi.nlm.nih.gov. [Accessed 8 Oct. 2023].
12. Veronelli, A., Laneri, M., Ranieri, R., Koprivec, D., Vardaro, D., Paganelli, M., Franco Folli and Pontiroli, A.E. (2004). White Blood Cells in Obesity and Diabetes. Diabetes Care, [online] 27(10), pp.2501–2502.
13. Lee, S.-H. and Park, J. (2023). The Role of Echocardiography in Evaluating Cardiovascular Diseases in Patients with Diabetes Mellitus. Diabetes & Metabolism Journal, [online] 47(4), pp.470–483.

Appendix B

Table 1: Table showing the variable names and its description for the CAD dataset.

Independent variables			
	Variable name	**Variable range**	**Variable type**
Demographic	Age	30–86	Discrete numeric
	Weight	48–120	Continuous numeric
	Length	140-188	Continuous numeric
	Sex	Male, female	Binary variable
	BMI (body mass index Kg/m^2)	18.12–40.90	Continuous numeric
	DM (Diabetes Mellitus)	Yes, no	Binary variable
	HTN (hyper tension)	Yes, no	Binary variable
	Current smoker	Yes, no	Binary variable
	Ex-Smoker	Yes, no	Binary variable
	FH (family history)	Yes, no	Binary variable
	Obesity	Yes if MBI > 25, no otherwise	Binary variable
	CRF (chronic renal failure)	Yes, no	Binary variable
	CVA (*Cerebrovascular Accident*)	Yes, no	Binary variable
	Airway disease	Yes, no	Binary variable
	Thyroid Disease	Yes, no	Binary variable
	CHF (congestive heart failure)	Yes, no	Binary variable
	DLP (*Dyslipidemia*)	Yes, no	Binary variable
Symptom and examination	BP (blood pressure: mmHg)	90–190	Continuous numeric
	PR (pulse rate) (ppm)	50–110	Continuous numeric
	Edema	Yes, no	Binary variable
	Weak peripheral pulse	Yes, no	Binary variable
	Lung rales	Yes, no	Binary variable
	Systolic murmur	Yes, no	Binary variable
	Diastolic murmur	Yes, no	Binary variable
	Typical Chest Pain	Yes, no	Binary variable
	Dyspnea	Yes, no	Binary variable
	Function class	1, 2, 3, 4	Categorical variables with 4 classes
	Atypical	Yes, no	Binary variable
	Nonanginal CP	Yes, no	Binary variable
	Exertional CP (Exertional Chest Pain)	Yes, no	Binary variable
	Low Th Ang (low Threshold angina)	Yes, no	Binary variable
ECG	BBB	LBBB, N or RBBB	Categorical variables with 3 classes
	Q Wave	Yes, no	Binary variable
	ST Elevation	Yes, no	Binary variable
	ST Depression	Yes, no	Binary variable
	T inversion	Yes, no	Binary variable
	LVH (left ventricular hypertrophy)	Yes, no	Binary variable

	Poor R progression (poor R wave progression)	Yes, no	Binary variable
Laboratory and echo	FBS (fasting blood sugar) (mg/dl)	62–400	Continuous numeric
	Cr (creatine) (mg/dl)	0.5–2.2	Continuous numeric
	TG (triglyceride) (mg/dl)	37–1050	Continuous numeric
	LDL (low density lipoprotein) (mg/dl)	18–232	Continuous numeric
	HDL (high density lipoprotein) (mg/dl)	15–111	Continuous numeric
	BUN (blood urea nitrogen) (mg/dl)	6–52	Continuous numeric
	ESR (erythrocyte sedimentation rate) (mm/h)	1–90	Continuous numeric
	HB (hemoglobin) (g/dl)	8.9–17.6	Continuous numeric
	K (potassium) (mEq/lit)	3.0–6.6	Continuous numeric
	Na (sodium) (mEq/lit)	128–156	Continuous numeric
	WBC (white blood cell) (cells/ml)	3700–18,000	Continuous numeric
	Lymph (Lymphocyte) (%)	7–60	Continuous numeric
	Neut (neutrophil) (%)	32–89	Continuous numeric
	PLT (platelet) (1000/ml)	25–742	Continuous numeric
	EF (ejection fraction) (%)	15–60	Continuous numeric
	Region with RWMA (regional wall motion abnormality)	0, 1, 2, 3, 4	Categorical variables with 5 classes
	VHD (valvular heart disease)	Normal, mild, moderate, severe	Categorical variables with 4 classes
Dependent variable			
Cath	Did the patient have CAD or not?	Cad, or Normal	Binary variable

Table 2: Head of CAD data set, showing columns 1-9.

Age	Weight	Length	Sex	BMI	DM	HTN	Current Smoker	EX Smoker
53	90	175	Male	29.38776	0	1	1	0
67	70	157	Fmale	28.39872	0	1	0	0
54	54	164	Male	20.07733	0	0	1	0
66	67	158	Fmale	26.83865	0	1	0	0
50	87	153	Fmale	37.16519	0	1	0	0
50	75	175	Male	24.48980	0	0	1	0

Table 3: Head of CAD data set, showing columns 10-18.

FH	Obesity	CRF	CVA	Airway disease	Thyroid Disease	CHF	DLP	BP
0	Y	N	N	N	N	N	Y	110
0	Y	N	N	N	N	N	N	140
0	N	N	N	N	N	N	N	100
0	Y	N	N	N	N	N	N	100
0	Y	N	N	N	N	N	N	110
0	N	N	N	N	N	N	N	118

Table 4: Head of CAD data set, showing columns 19-26.

PR	Edema	Weak Peripheral Pulse	Lung rales	Systolic Murmur	Diastolic Murmur	Typical Chest Pain	Dyspnea
80	0	N	N	N	N	0	N
80	1	N	N	N	N	1	N
100	0	N	N	N	N	1	N
80	0	N	N	N	Y	0	Y
80	0	N	N	Y	N	0	Y
70	0	N	N	N	N	1	N

Table 5: Head of CAD data set, showing columns 27-34.

Function Class	Atypical	Nonanginal	Exertional CP	LowTH Ang	Q Wave	St Elevation	St Depression
0	N	N	N	N	0	0	1
0	N	N	N	N	0	0	1
0	N	N	N	N	0	0	0
3	N	Y	N	N	0	0	1
2	N	N	N	N	0	0	0
3	N	N	N	N	0	0	0

Table 6: Head of foetal CAD data set, showing columns 35-44.

Poor R Progression	LVH	Tinversion	BBB	FBS	CR	TG	LDL	HDL	BUN
N	N	1	N	90	0.7	250	155	30	8
N	N	1	N	80	1.0	309	121	36	30
N	N	0	N	85	1.0	103	70	45	17
N	N	0	N	78	1.2	63	55	27	30
N	N	0	N	104	1.0	170	110	50	16
N	N	0	N	86	1.0	139	119	34	13

Table 7: Head of CAD data set, showing columns 45-56 where column 56 is the dependent variable.

K	HB	ESR	Na	WBC	Lymph	Neut	PLT	EF TTE	Region RWMA	VHD	Cath
4.7	15.6	7	141	5,700	39	52	261	50	0	N	Cad
4.7	13.9	26	156	7,700	38	55	165	40	4	N	Cad
4.7	13.5	10	139	7,400	38	60	230	40	2	mild	Cad
4.4	12.1	76	142	13,000	18	72	742	55	0	Severe	Normal
4.0	13.2	27	140	9,200	55	39	274	50	0	Severe	Normal
4.2	15.6	18	141	7,300	26	66	194	50	0	N	Cad

Table 8: Tail of the CAD data set, showing columns 1-9.

Age	Weight	Length	Sex	BMI	DM	HTN	Current Smoker	EX Smoker
30	100	172	Male	33.80206	0	0	1	0
58	84	168	Male	29.76190	0	0	0	0
55	64	152	Fmale	27.70083	0	0	0	0
48	77	160	Fmale	30.07812	0	1	0	0
57	90	159	Fmale	35.59986	1	0	0	0
56	85	170	Fmale	29.41176	0	1	1	0

Table 9: Tail of the CAD data set, showing columns 10-19.

FH	Obesity	CRF	CVA	Airway disease	Thyroid Disease	CHF	DLP	BP	PR
1	Y	N	N	N	N	N	N	110	60
0	Y	N	N	N	N	N	N	100	76
0	Y	N	N	N	N	N	N	100	60
1	Y	N	N	N	N	N	N	130	70
0	Y	N	N	N	N	N	N	100	60
0	Y	N	N	N	N	N	N	120	80

Table 10: Tail of the CAD data set, showing columns 20-26

Edema	Weak Peripheral Pulse	Lung rales	Systolic Murmur	Diastolic Murmur	Typical Chest Pain	Dyspnea
0	N	N	N	N	0	N
0	N	N	N	N	1	N
0	N	N	Y	N	0	Y
0	N	N	N	N	0	N
0	N	N	N	N	0	Y
0	N	N	N	N	1	N

Table 11: Tail of the CAD data set, showing columns 27-34.

Function Class	Atypical	Nonanginal	Exertional CP	LowTH Ang	Q Wave	St Elevation	St Depression
0	Y	N	N	N	0	0	0
0	N	N	N	N	0	0	0
0	Y	N	N	N	0	0	0
0	N	Y	N	N	0	0	0
0	Y	N	N	N	0	0	0
0	N	N	N	N	0	0	0

Table 12: Tail of the CAD data set, showing columns 35-44

Tinversion	LVH	Poor.R.Progression	BBB	FBS	CR	TG	LDL	HDL	BUN
0	N	N	N	83	1.0	205	97	53	20
0	N	N	N	92	1.0	112	115	44	13
0	N	N	LBBB	86	0.9	111	40	23	23
0	N	N	RBBB	83	1.0	93	112	42	13
0	N	N	N	96	1.0	116	130	49	14
1	N	N	N	78	0.7	139	124	34	16

Table 13: Tail of the CAD data set, showing columns 45-56, where column 56 represents the dependent variable.

ESR	HB	K	Na	WBC	Lymph	Neut	PLT	EF TTE	Region. RWMA	VHD	Cath
16	13.1	4.0	143	9,100	39	60	294	55	1	N	Normal
13	12.3	4.8	146	8,500	34	58	251	45	0	N	Cad
3	12.4	4.0	139	11,400	16	80	377	40	0	mild	Normal
20	12.8	4.0	140	9,000	35	55	279	55	0	N	Normal
31	10.1	3.8	141	3,800	48	40	208	55	0	N	Normal
13	14.7	4.4	147	6,000	32	55	302	55	0	N	Cad