Heart Disease Prediction Based on Clinical Variables

Grace Brownlie, Bridget Fijma, Aidan Ang, Hannah Collier, Thomas Rowley

20th October 2023

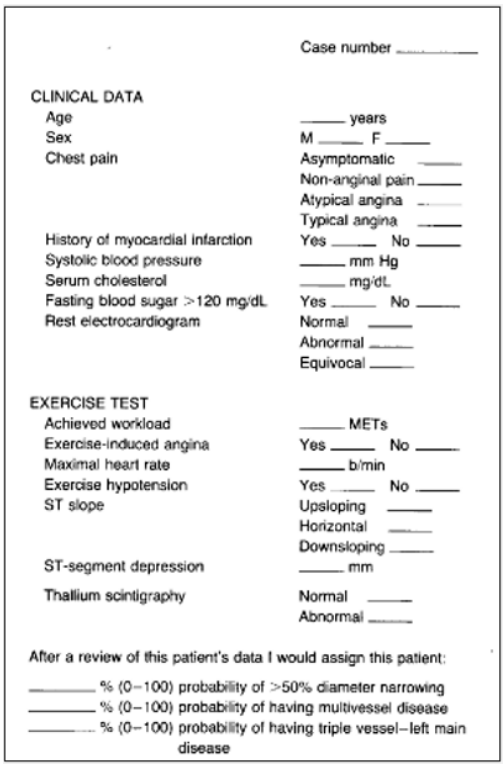
Contents

Introduction

Heart disease is the leading cause of death in New Zealand accounting for almost 1 in 3 fatalities. A New Zealander dies from heart disease every 90 minutes with many of these deaths being preventable [1]. Understanding the clinical variables that make people vulnerable to heart disease is crucial to implementing effective prevention strategies and informing policy decisions.

The data we used for this report contains observations of 270 patients who have been referred for coronary angiography at Cleveland Clinic in Cleveland, Ohio. Given that they have been referred for further testing we know that each of these patients have been determined to have some possible indications of heart disease. Patients were diagnosed with heart disease if they had more than 50% diameter narrowing of a coronary artery [2].

The article "Clinical Assessment of the Probability of Coronary Artery Disease: Judgemental Bias from Personal Knowledge" [3] assessed the bias of physicians when basing probability of heart disease on personal knowledge. The study consisted of 510 patients in Long Beach California referred for the first time for coronary angiography. Presented with case summaries of these patients in Figure 1 physicians were asked to estimate the probability of that patient having heart disease. The variables used in this study are almost the same as those in the Cleveland dataset with three additional variables; "History of myocardial infarction", "Achieved workload" and, "Exercise hypotension". The results showed that the physicians consistently overestimated the probability of coronary artery disease. This indicates that relying solely on the personal knowledge of physicians may not be adequate for prediction of heart disease.



Case number _____

CLINICAL DATA

Age	_____ years
Sex	M _____ F _____
Chest pain	Asymptomatic _____
	Non-anginal pain _____
	Atypical angina _____
	Typical angina _____
History of myocardial infarction	Yes _____ No _____
Systolic blood pressure	_____ mm Hg
Serum cholesterol	_____ mg/dL
Fasting blood sugar >120 mg/dL	Yes _____ No _____
Rest electrocardiogram	Normal _____
	Abnormal _____
	Equivocal _____

EXERCISE TEST

Achieved workload	_____ METs
Exercise-induced angina	Yes _____ No _____
Maximal heart rate	_____ b/min
Exercise hypotension	Yes _____ No _____
ST slope	Upsloping _____
	Horizontal _____
	Downsloping _____
ST-segment depression	_____ mm
Thallium scintigraphy	Normal _____
	Abnormal _____

After a review of this patient's data I would assign this patient:

_____ % (0–100) probability of >50% diameter narrowing
_____ % (0–100) probability of having multivessel disease
_____ % (0–100) probability of having triple vessel–left main disease

Figure 1: Example of Patient Summary shown to physicians in Long Beach study

This leads to our research questions:

1. Using empirical techniques, can we accurately predict whether heart disease is present based on clinical variables?

2. How does sex and heart disease presence interact? Are there significant differences in predicting the presence of heart disease in females compared to males?

In order to answer these questions we first investigated the data using Exploratory Data Analysis. We then used Linear Discriminant Analysis and Bayesian classification using Gaussian models to evaluate the accuracy of our model in predicting the presence of Heart Disease. We then repeated this now using interaction between sex and heart disease to examine how this changes the accuracy of the model and whether the sensitivity and specificity of the model differs between classes.

Methodology

The Dataset

Our dataset "Heart_Disease_Prediction" was created by Andras Janosi, William Steinbrunn, Matthias Pfisterer and Robert Detrano and was donated to the University of California Irvine data repository [4]. It was made publicly available by Robert Hoyt MD and can be found at link. The original database contained 76 attributes but the published data uses a subset of just 14 of these attributes.

Variables

The variables in our dataset can be categorised into Clinical Data and Stress Test Data. Clinical Data are those which could be observed or measured at a local GP clinic. Stress Test Data are observations taken to measure the ability of the heart to pump blood under increased stress conditions, usually this is during exercise using a treadmill or stationary bicycle.

Clinical Data

- Age (*numeric*): The age of the patient
- Sex (*categorical*): Sex of the patient (0 = Female, 1 = Male)
- Chest pain type (*categorical*): Type of chest pain experienced by the patient (1 = Typical angina, 2 = Atypical angina, 3 = Non-anginal pain, 4 = Asymptomatic)
- BP (*numeric*): The blood pressure level of the patient in mmHg
- Cholesterol (*numeric*): The cholesterol level of the patient in mg/dl
- FBS over 120 (*categorical*): Fasting blood sugar test results are over 120 mg/dl (0 = False, 1 = True)
- EKG results (*categorical*): Resting electrocardiographic results (0 = normal, 1 = having ST-T wave abnormality, 2 = Showing Probable or Definite Left Ventricular Hypertrophy by Estes' Criteria)

Stress Test Data

- Exercise Angina (*categorical*): Patient experiences exercise-induced angina (0 = False, 1 = True) Max HR (*numeric*): The maximum heart rate levels achieved during exercise testing in bpm
- Slope of ST (*categorical*): Slope of the peak exercise ST segment (1 = Upsloping, 2 = Flat, 3 = Downsloping)
- ST depression (*numeric*): The ST(Stress Test) depression on an Electrocardiogram induced by exercise relative to rest in mm
- Thallium (*categorical*): Thallium Stress test finding (3 = Normal, 6 = Fixed defect, 7 = Reversible defect)

In our analysis we took a subset of six of these variables focusing on clinical data.

These six variables were: * Age * Sex * BP (Blood Pressure) * Cholesterol * Max HR (Maximum Heart Rate) * Chest Pain Type

Exploratory Data Analysis (EDA)

A preliminary EDA will be done on the variables that were decided to be analysed. Histograms, notched boxplots and Q-Q plots for age, blood pressure, cholesterol levels, and maximum heart rate will be made to analyse the distribution of these variables. The covariance matrix and correlogram of the predictor values will be used to identify any possible relationships between them. Anderson-Darling tests for normality will be used to determine if assumptions of normality on the distribution of our predictor variables, that are required for various analytical techniques such as linear discriminant analysis, are met. If assumptions for normality are not met, transformations will need to be applied to the data, then tested again to see if the transformed data more closely resembles a normal distribution.

Linear Discriminant Analysis

To answer our research questions, we decided to use linear discriminant analysis to predict the presence of heart disease using the numerical variables analyzed in the EDA, alongside sex and chest pain type as factors of interest. Since we have two distinct levels of heart disease, presence and absence, and these are known for all observations, we deemed linear discriminant analysis a possible avenue for prediction analysis. Linear discriminant analysis requires that covariance matrices are equal, and that data is normally distributed - these will be tested to see if the assumptions for linear discriminant analysis hold. Cluster analysis was deemed not useful for answering our questions, as we saw that technique better used for classifying data when the true value of the class, i.e. whether heart disease was present or not, was unknown to us.

Bayesian Classification

It was unknown whether or not the difference in prior probabilities of being assigned heart disease, obtained from the data to be $\approx 44.4\%$ and $\approx 55.6\%$ for presence and absence respectively, would affect the results of our linear discriminant analysis. To be cautious, we decided we will perform a Bayesian classification analysis, assuming that data for each predictor belongs to the Gaussian Normal distribution. This technique accounts for the difference in prior probabilities, and does not use strictly linear discriminants, and would yield potentially different results to the linear discriminant analysis.

Interaction

To answer our second research question, if there are any differences in heart disease prediction accuracy due to differences in sex, we induced 4 classes of heart disease and sex interaction. Namely, these are males with heart disease, females with heart disease, males without heart disease, and females without heart disease. We will then use both linear discriminant analysis and Bayesian classifiers assuming predictor values belong to the Gaussian normal distribution, to see whether heart disease prediction differs between sex, i.e. if models are more accurate at classifying one sex than the other.

Comparisons

For both research questions, the effectiveness of linear discriminant analysis and Bayesian classifiers will be compared using confusion matrices on a test sample set of the data, partition matrix plots, and relevant statistics obtained from the confusion matrices. These relevant statistics include sensitivity, or true positive rate, specificity, or true negative rate, overall and balanced accuracies, and if these models are significantly better predictors of heart disease than randomly classifying values based on prior probabilities, i.e. the No Information Rate. The model with the higher accuracies, and which makes the least mistakes, will be chosen as the preferred between the two approaches of linear discriminant analysis and Bayesian classification.

Exploratory Data Analysis

Histograms, notched box plots, and Q-Q plots will be generated from the heart disease data, considering particularly the numerical variables Age, Blood pressure, Cholesterol, and Maximum heart rate. This will be done to visualise the distribution of the data, and check for if the data is Gaussian normally distributed. These plots will be found in Figures 2, 3, and 4 respectively.

Histograms

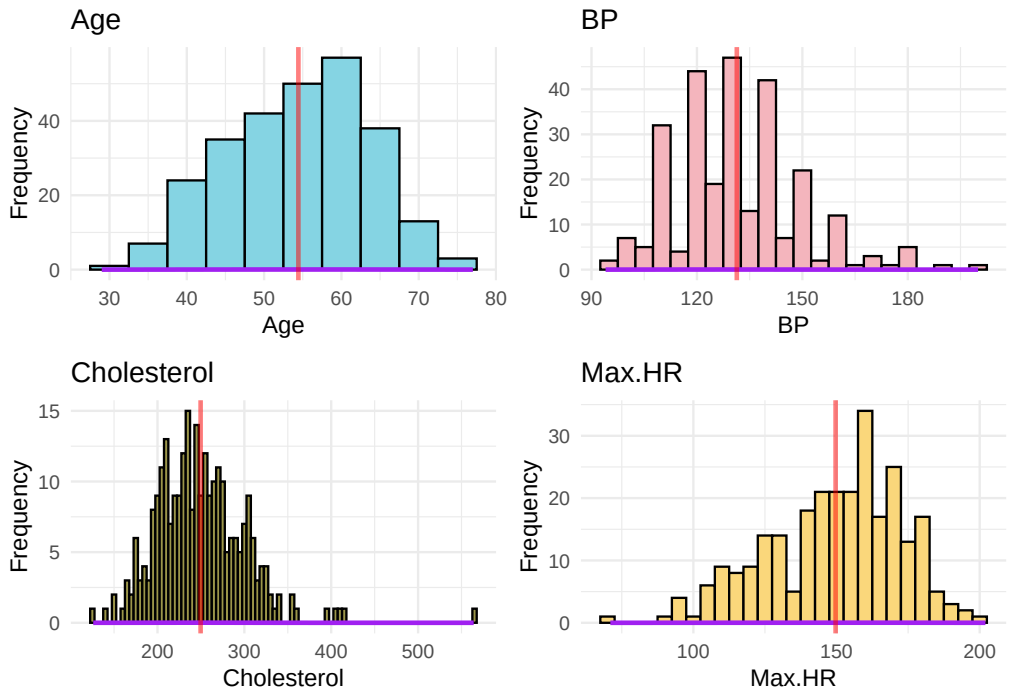


Figure 2: Histograms for the numerical variables of the heart disease dataset.

From Figure 2, the histograms for age appears to be the most normal with the least significant outliers, whereas cholesterol, maximum heart rate, and blood pressure have distributions that are visually not normally distributed, with significant outliers.

The blood pressure histogram has many peaks at the 10 mmHg marks. This is a result of "zero end-digit preference", a phenomenon where blood pressure readings are rounded to the nearest ten. This is known to occur especially around treatment cutoffs (120 mmHg is considered 'elevated', 130mmHg is 'stage 1 hypertension', and 140mmHg is 'stage 2 hypertension') [5]. 150 of the 270 observations, approximately 56%, are given as a multiple of 10. Most blood pressure indicators give guidelines to measure to the nearest 2 mmHg. The potential rounding of data indicates a lack of normality in the data since many observations may have been rounded up or down significantly.

Box plots

Next, box plots should be generated to analyse the distribution of data for each variable further.

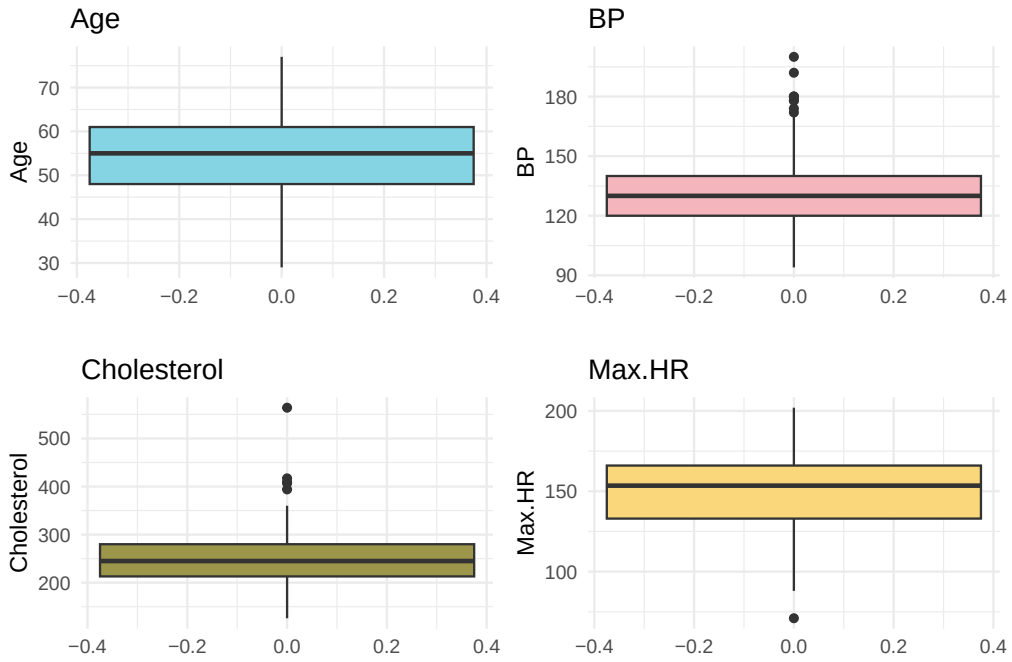


Figure 3: Notched box plots for the numerical variables of the heart disease dataset.

The box plots from Figure 3 shows that blood pressure and cholesterol values have noted outliers to the higher end of values, where maximum heart rate has one outlier on the lower end. The outlier for maximum heart rate could be due to a patient already receiving treatment for heart disease, and being on beta blockers, a commonly prescribed drug to lower heart rate and reduce the risk of heart attack [6].

Normal Q-Q plots

Next, normal Q-Q plots should be generated for each variable to check for the normality of residuals.

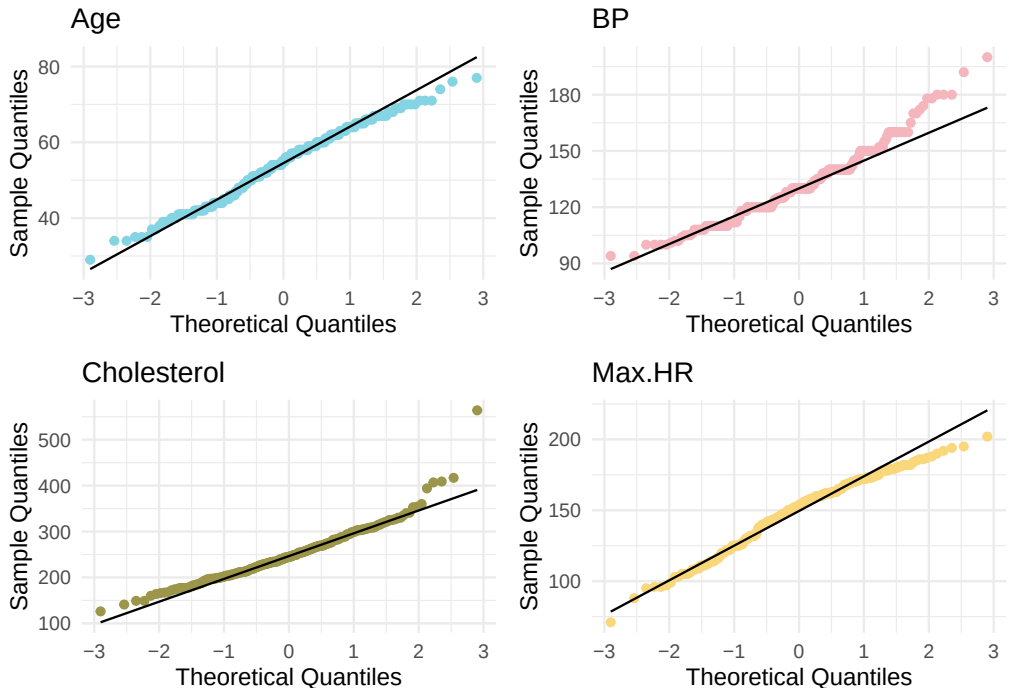


Figure 4: Normal Q-Q plots for the numerical variables of the heart disease dataset.

The normal Q-Q plots from Figure 4 for the variables suggest normality in residuals for all variables may be present. The only deviations from the Q-Q line are at the high ends data where outliers where noted to be present in cholesterol, maximum heart rate, and blood pressure.

Summary statistics

Summary statistics, the covariance matrix, and the correlation matrix are provided below in Tables 1, 2, and 3 respectively.

Table 1: Exploratory data analysis summary for the numerical variables of heart disease data.

	Age	BP	Cholesterol	Max.HR
Sample size	270	270	270	270
Minimum	29.00	94.0	126.0	71.0
1st quartile	48.00	120.0	213.0	133.0
Median	55.00	130.0	245.0	153.5
Mean	54.43	131.3	249.7	149.7
3rd quartile	61.00	140.0	280.0	166.0
Maximum	77.00	200.0	564.0	202.0
Skewness	-0.1627	0.7186	1.1771	-0.5248
Kurtosis	2.4431	3.884	7.7833	2.8767

Table 2: Covariance matrix of the numerical variables of the heart disease dataset.

	Age	BP	Cholesterol	Max.HR
Age	82.97509	44.42639	103.60545	-84.87472
BP	44.42639	319.03705	159.73119	-16.19343
Cholesterol	103.60545	159.73119	2671.46711	-22.43734
Max.HR	-84.87472	-16.19343	-22.43734	536.65043

Table 3: Covariance matrix of the numerical variables of the heart disease dataset.

	Age	BP	Cholesterol	Max.HR
Age	1.0000000	0.2730528	0.2200563	-0.4022154
BP	0.2730528	1.0000000	0.1730192	-0.0391357
Cholesterol	0.2200563	0.1730192	1.0000000	-0.0187392
Max.HR	-0.4022154	-0.0391357	-0.0187392	1.0000000

A correlogram is generated as a visualisation of the correlation matrix between variables and presented in Figure 5.

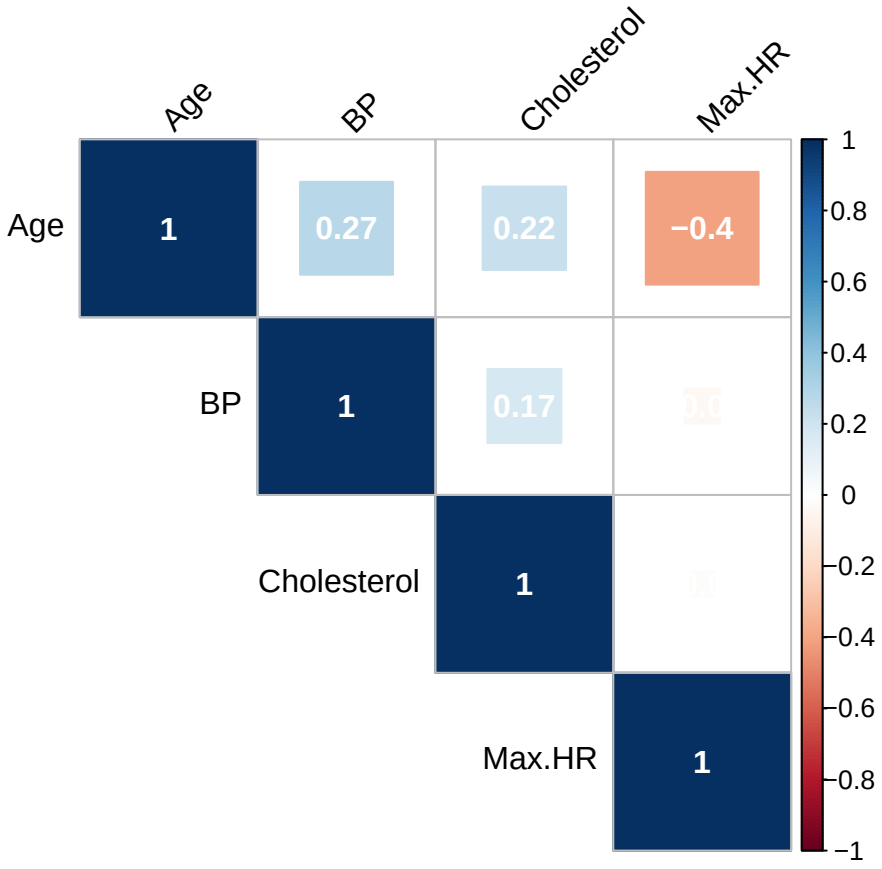


Figure 5: Correlogram of numerical variables in the heart disease dataset.

The only notable correlation is with age and maximum heart rate, with a moderate negative correlation of

-0.4. This is expected as it is known that as a person ages, the condition of the heart and therefore its maximum capacity for pumping blood, decreases [5].

Assumptions testing and transformations

Next the data should be tested for normality using the Anderson-Darling test. The results will be found in Table 4.

Table 4: Anderson-Darling test p-values for normality.

Age	BP	Cholesterol	Max.HR
0.0060111	8.3e-06	0.0008372	1.05e-05

There is evidence to reject normality for all numerical variables by the Anderson-Darling test, at the $\alpha = 0.05$ significance level. The data can be log-transformed to see if the distribution of the log-transformed data is normal, and more Anderson-Darling tests can be performed. The results for the log-transformed data's Anderson-Darling tests can be found in 5.

Table 5: Anderson-Darling test p-values for normality on log-transformed data.

Age	BP	Cholesterol	Max.HR
2.8e-06	0.0042294	0.5529058	0

Transforming cholesterol and blood pressure give better p-values for testing whether or not the distribution of the log-transformed data is normal, so log-transformations to cholesterol and blood pressure should be applied.

Heart Disease Prediction on Clinical Variables

The first question we seek to answer is if we can predict heart disease using clinical variables. The variables that will be used are: Age, sex, maximum heart rate, cholesterol level, blood pressure, and type of chest pain.

A pairs plot will be generated to visualise potential differences in the distributions of the aforementioned variables that are numeric, i.e. age, blood pressure, cholesterol, and maximum heart rate, depending on heart disease presence or absence. The plot can be found in Figure 6.

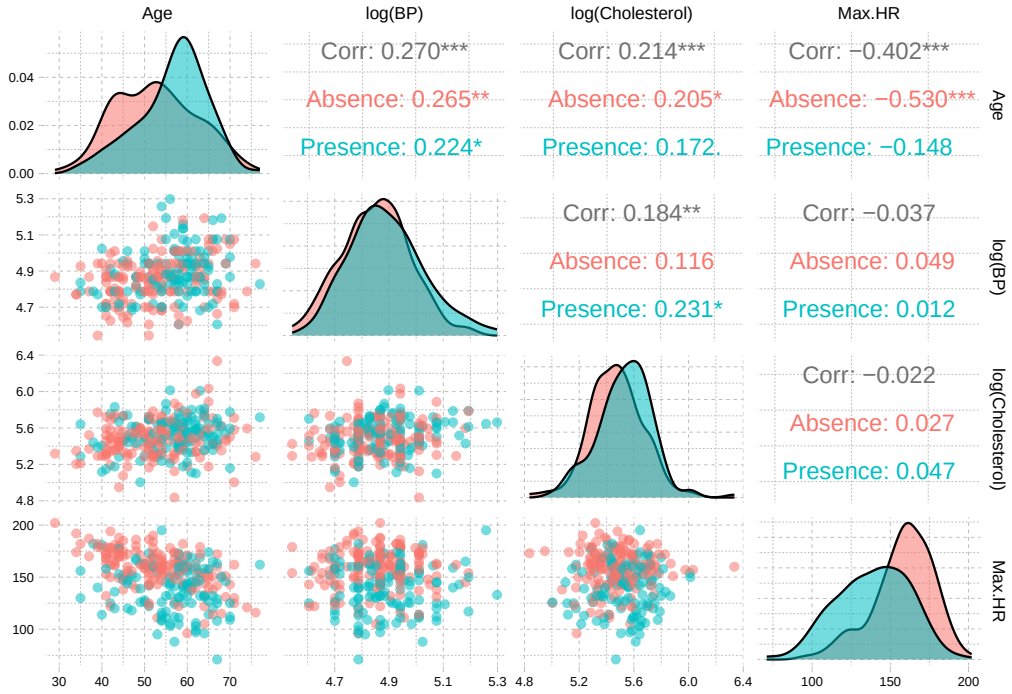


Figure 6: Pairs plot for the continuous numeric variables.

From the pairs plot it is seen that the distribution of age and maximum heart rate between patients with and without heart disease differ significantly. The age plot shows that heart disease is more prevalent in older people. The maximum heart rate plot shows that the maximum heart rate for those with heart disease is more normally distributed, and tends to be lower, than those without heart disease. The plots for the log-transformed cholesterol and log-transformed blood pressure are similar in shape, with slight shifts to higher values for those with heart disease. From these plots there is some evidence that classes of heart disease may be separable due to differences in distribution between predictor values.

Linear discriminant analysis

A linear discriminant analysis will be performed on the heart disease dataset in order to predict the heart disease variable using age, sex, chest pain type, the log-transformed blood pressure and cholesterol variables, and maximum heart rate. Test and training datasets will be obtained by randomly sampling the data with replacement, with 80% and 20% probabilities respectively.

Results

Model fitting: The results of the linear discriminant analysis can be seen in Figure 7 and Figure 8. In the figures, overlap in the plots indicates that those data points could be assigned either with heart disease, or assigned to not have heart disease, meaning misclassification errors are possible. Both figures represent the same data, where one is a histogram of each class on its own axis, and the other is a density plot where both classes' data is on the same axis.

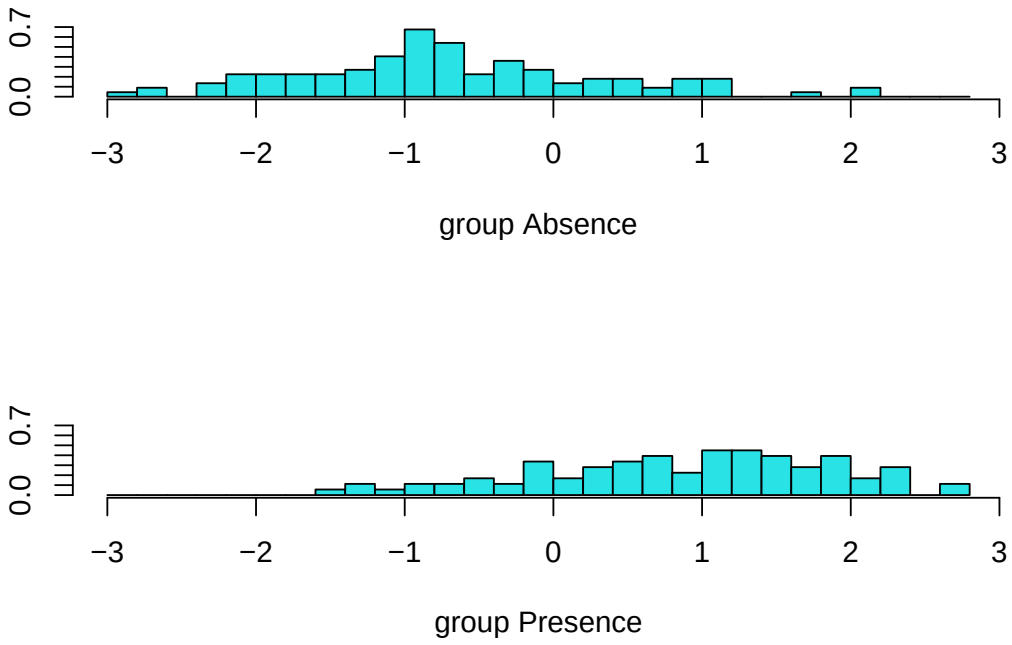


Figure 7: Histogram of assigned class of heart disease.

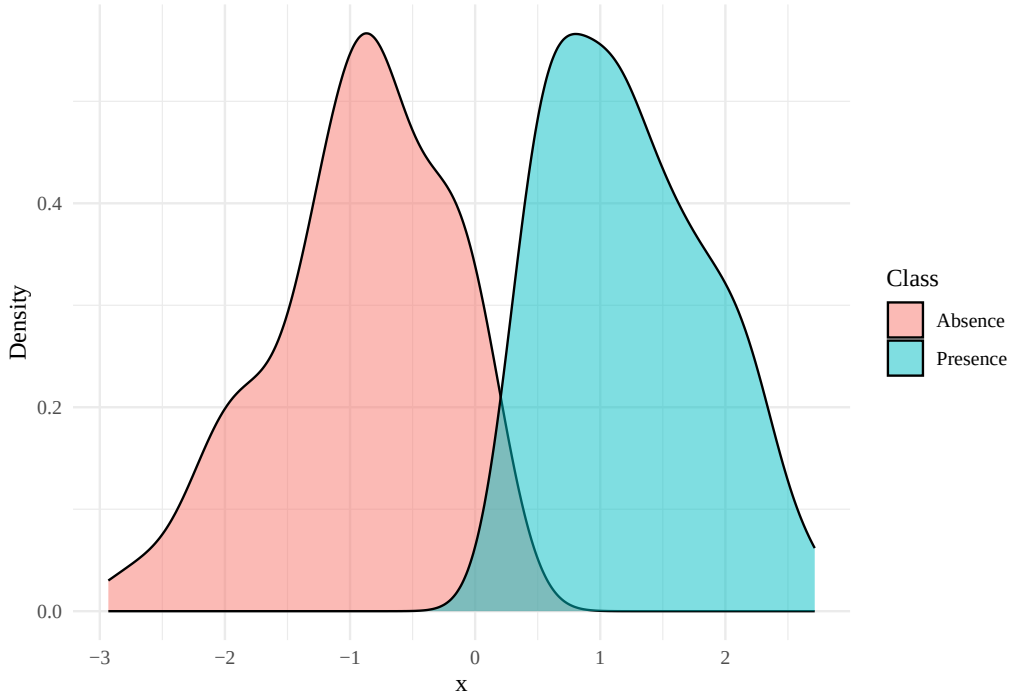


Figure 8: Density plot of assigned class of heart disease.

From the figures it is clearly seen that there is overlap in predicted values for the data, meaning misclassification of patients' heart disease status is present and possible in the model.

An assumption for performing linear discriminant analysis is that covariance matrices of variables, between the data where heart disease is present, and the data where heart disease is absent, are similar. The test for covariance equality is found in Table 6.

Table 6: Test results for comparing covariance matrices for absence and presence of heart disease for equality.

test_statistic	pvalue
3.583	0.07649

The covariance equality test is performed at the $\alpha = 0.05$ significance level, with null hypothesis that covariance matrices are equal. The associated p-value for this test is $p = 0.07649$. Since $p > \alpha$, there is insufficient evidence to reject that covariance matrices are equal. The assumption required for linear discriminant analysis thus holds.

Prediction: Statistics for the misclassifications can be obtained via confusion matrices. Using the test data, we can report a realistic confusion matrix for the model, which is presented in Table 7.

Table 7: Confusion matrix for predicting heart disease using linear discriminant analysis.

	Absence	Presence
Absence	36	7
Presence	3	24

The linear discriminant model has wrongly assigned 7 of 43 patients without heart disease as having heart disease, and 3 of 27 patients with heart disease as not having heart disease, errors of 22.58% and 7.69% respectively.

More detailed statistics concerning the confusion matrix are presented in Table 9 and Table 8.

Table 8: Overall statistics for predicting heart disease using linear discriminant analysis.

	x
Accuracy	85.71%
Kappa	70.66%
AccuracyLower	75.29%
AccuracyUpper	92.93%
AccuracyNull	55.71%
AccuracyPValue	0.00%
McnemarPValue	34.28%

The linear discriminant model has an overall accuracy of $85.71\% \pm 10.42\%$. The accuracy of the model is tested against the No Information Rate, which is the largest proportion of any class in the model, i.e. P(*Absence*), at the $\alpha = 5\%$ significance level. The p-value for the significance of this model with respect to the No Information Rate of 55.71% is $p \approx 0.00\%$ at the $\alpha = 5\%$ significance level, meaning the linear discriminant analysis is highly significant in its prediction power for heart disease over the no-information model.

Table 9: Prediction statistics for predicting heart disease using linear discriminant analysis.

	x
Sensitivity	77.42%
Specificity	92.31%
Pos Pred Value	88.89%
Neg Pred Value	83.72%
Precision	88.89%
Recall	77.42%
F1	82.76%
Prevalence	44.29%
Detection Rate	34.29%
Detection Prevalence	38.57%
Balanced Accuracy	84.86%

The model has sensitivity of 77.42%, specificity of 92.31%, and a balanced accuracy of 84.86%. This means it can predict someone with heart disease as having it with 77.42% certainty, and someone without heart disease as not having it with 92.31% certainty. Equivalently, this means the model has a false positive rate of 22.58% and a false negative rate of 7.69%. In terms of the research question, this means the linear discriminant analysis fails to detect heart disease in patients that do have heart disease 7.69% of the time.

A partition matrix is generated considering all numerical variables as before. Red points represent that the observation has been misclassified, and black points represent correctly classified observations. The partition plot can be found in Figure 9.

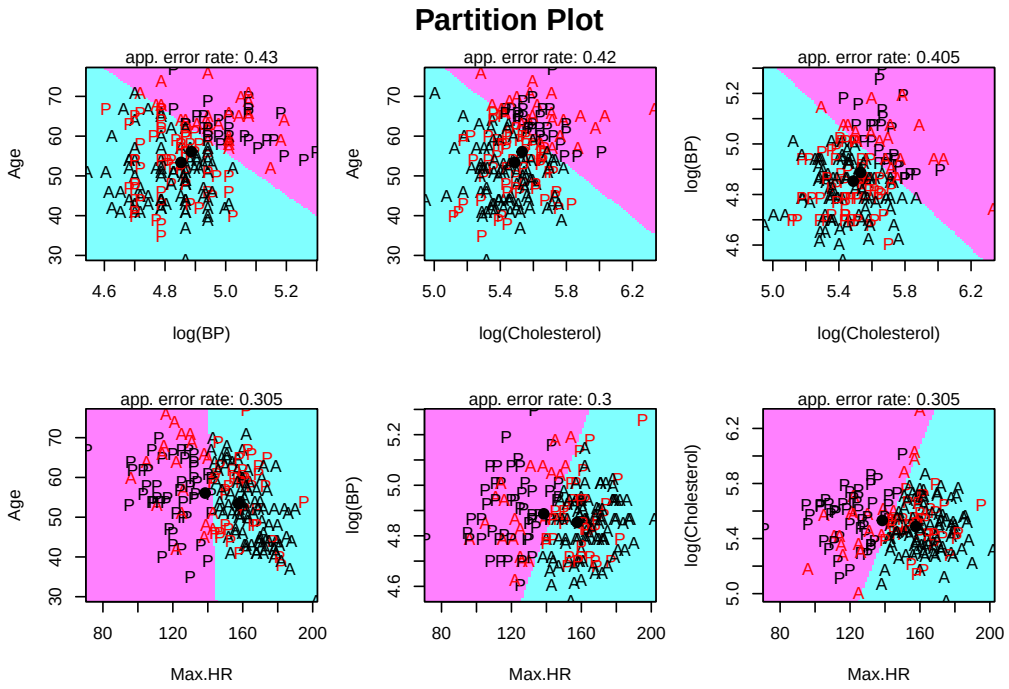


Figure 9: Partition plot of numerical variables used in the linear discriminant analysis.

The approximate error rate for classification is highest at 43% between blood pressure and age, and lowest at 29.5% between maximum heart rate and cholesterol. This gives a range of classification error between

$29.5\% - 43\%$ between the numerical variables of the linear discriminant model.

Bayesian classification

The prior probabilities of having heart disease and not having heart disease are different in the data, with approximately 45% of patients having heart disease. Linear discriminant analysis where classes are unbalanced could mean that large amounts of data is misclassified by the linear discriminant. To account for this, a Bayesian classification can be used instead, as this accounts for the prior probability differences. It will be assumed that the data belongs to the Gaussian normal distribution.

From the data the prior probabilities of an observation belonging to either distribution are given as P(Absence) $= \frac{111}{200} = 0.555$ and P(Presence) $= \frac{89}{200} = 0.445$.

Results

Model fitting: Bayesian classification can be applied to the test data for the variables age, sex, chest pain type, blood pressure, cholesterol level, and maximum heart rate, the same as in the linear discriminant analysis. Its performance will be compared to that of the linear discriminant analysis. Tables of results can be found in Table 10, Table 11, and Table 12.

Table 10: Confusion matrix for predicting heart disease using the Bayesian classification method.

	Absence	Presence
Absence	32	9
Presence	7	22

Prediction: The Bayesian classification method wrongly assigned 7 of 39 patients without heart disease as having heart disease, and misclassified 24 of 31 patients with heart disease as not having it, errors of 17.95% and 29.03% respectively. These error values are different than those of the linear discriminant analysis, which had errors of 22.58% and 7.69%. Notably, the Bayesian classification has less error when identifying absence of heart disease than the linear discriminant analysis, but much more error when identifying heart disease in patients, approximately 3.7 times worse.

Table 11: Overall statistics for predicting heart disease using the Bayesian classification method.

	x
Accuracy	77.14%
Kappa	53.37%
AccuracyLower	65.55%
AccuracyUpper	86.33%
AccuracyNull	55.71%
AccuracyPValue	0.02%
McnemarPValue	80.26%

The Bayesian classification has an overall accuracy of $77.14\% \pm 11.59\%$. The accuracy of this model is tested against the null hypothesis that the model is no better than the largest proportion of classes, the No Information Rate. This hypothesis is tested at the $\alpha = 5\%$ level, and the corresponding p-value is 0.02%. Since $p < \alpha$, there is sufficient evidence to suggest the model is better at predicting levels of heart disease than the no information model. This means the Bayesian classification method is significant in its prediction power for heart disease by this metric.

Table 12: Prediction statistics for predicting heart disease using the Bayesian classification method

	x
Sensitivity	70.97%
Specificity	82.05%
Pos Pred Value	75.86%
Neg Pred Value	78.05%
Precision	75.86%
Recall	70.97%
F1	73.33%
Prevalence	44.29%
Detection Rate	31.43%
Detection Prevalence	41.43%
Balanced Accuracy	76.51%

The Bayesian classification has sensitivity of 70.97%, specificity of 82.05%, and a balanced accuracy of 76.51% for predicting the presence of heart disease. Equivalently, this means the model has a false positive rate of 29.03% and a false negative rate of 17.95%. In terms of the research question, this means the Bayesian classification assuming the data came from the Gaussian normal distribution does not detect heart disease in patients that have it 17.95% of the time.

A partition matrix can also be generated for this data on its numerical variables.

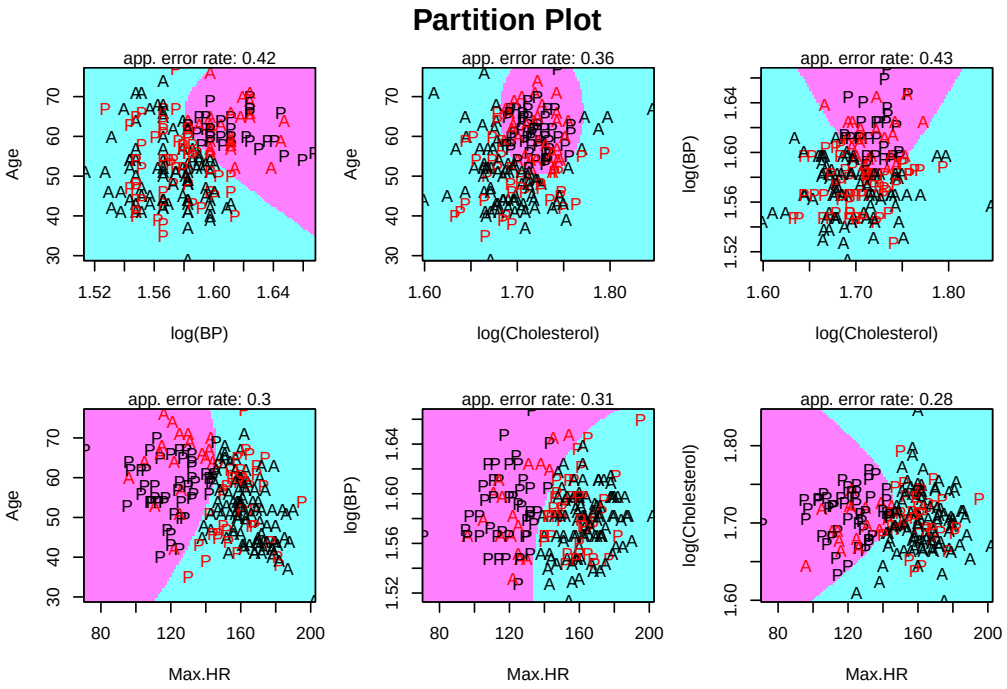


Figure 10: Partition plot of the numerical variables used in the Bayesian classification.

The approximate error rate for classification is highest at 43.5% between the log of blood pressure and log of cholesterol, and lowest at 28.0% between maximum heart rate and the log cholesterol. This gives a range of classification error between $28.0\% - 43.5\%$ between the numerical variables of the Bayesian classification, slightly better than that of the linear discriminant analysis.

Comparison of approaches

Important summary statistics that have been previously calculated are collated in Table 13.

Table 13: Summary of important statistics for model comparison for heart disease prediction.

	LDA	Bayesian
Sensitivity	77.42%	70.97%
False Positive Rate	22.58%	29.03%
Specificity	92.31%	82.05%
False Negative Rate	7.69%	17.95
Overall Accuracy	85.71%	77.14%
Balanced Accuracy	84.86%	76.51%

The linear discriminant analysis has lower false positive rates, lower false negative rates, higher overall accuracy, and higher balanced accuracy compared to the Bayesian classification assuming Gaussian normal distribution approach. By all metrics obtained from confusion matrices, the linear discriminant analysis provides a more accurate and more useful model for predicting the presence of heart disease in patients considering their age, sex, maximum heart rate, the log-transformation of their cholesterol and blood pressure readings, and their type of chest pain. It correctly classifies patients with and without heart disease more often than the Bayesian classification approach does, as it makes less errors.

The difference in performance of each model could be attributed to a number of possibilities. The Bayesian classification we used assumed data belonged to the Gaussian normal distribution, which according to the Anderson-Darling tests was only likely for the log-transformed cholesterol value. It also assumed independence of predictor variables, which may have been violated by covariances between predictors in the data. The linear discriminant analysis however does not assume independence between predictor values, instead assuming that covariances matrices are equal. Linear discriminant analysis like Bayesian classification, assumes normality of data, so the disparity between results of the models is likely to be due to the lack of independence between predictor variables. This is evidenced by correlation between predictors being prevalent as seen in Figure 6.

Heart Disease Prediction Accounting for Sex

After performing analysis using LDA and Bayesian to predict Heart Disease, we did some further research on our data to see if there could be a better way to predict heart disease. We noticed that there were many discrepancies between male and female characteristics and symptoms associated with heart disease. This led us to question whether we could more accurately predict heart disease if we induce four classes by using sex interacting with Heart Disease presence. These four classes were:

fp – female and heart disease present mp – male and heart disease present FA – female and heart disease absent MA – male and heart disease absent.

This allowed us to see if our model had difficulties predicting heart disease for a certain sex.

Results:

```
## [1] "0.Absence"  "1.Absence"  "0.Presence" "1.Presence"
```

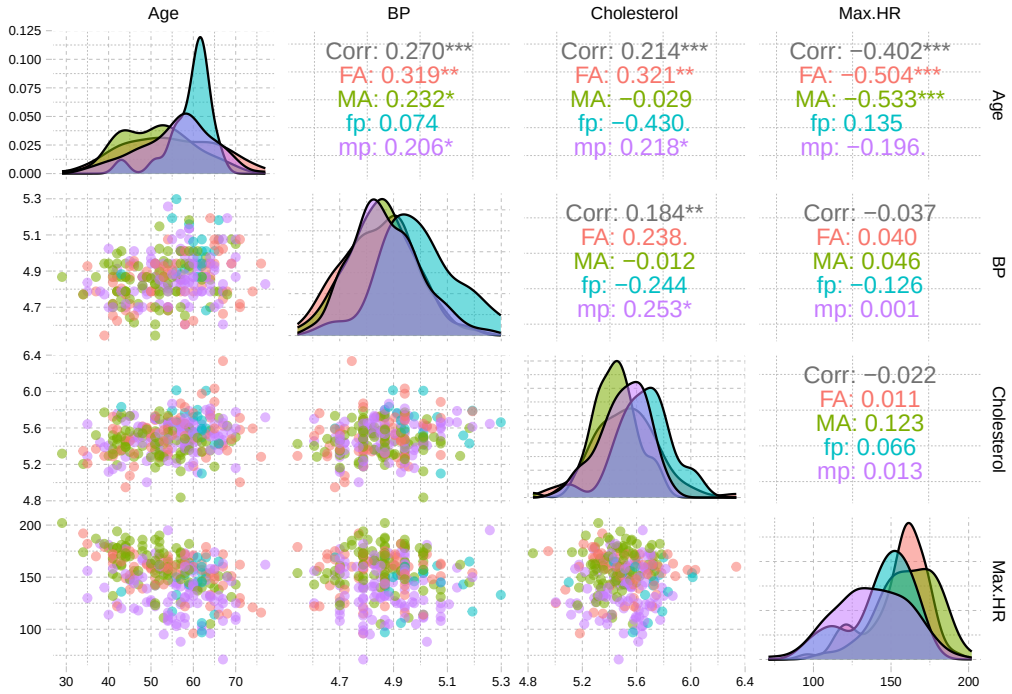


Figure 11: Figure showing Pairs Plot of the four induced classes.

```
## Call:
## lda(Interaction ~ Age + Chest.pain.type + log(BP) + log(Cholesterol) +
##     Max.HR, data = HD_Train)
##
## Prior probabilities of groups:
##         FA         MA         fp         mp
## 0.24691358 0.31481481 0.09259259 0.34567901
##
## Group means:
##         Age Chest.pain.type2 Chest.pain.type3 Chest.pain.type4  log(BP)
## FA 56.02500       0.17500000        0.4250000        0.3250000 4.850360
## MA 49.66667       0.21568627        0.4117647        0.2549020 4.852964
## fp 58.40000       0.06666667        0.0000000        0.9333333 5.014925
## mp 56.37500       0.07142857        0.1607143        0.7142857 4.864139
##    log(Cholesterol)   Max.HR
## FA         5.542285 153.9500
## MA         5.453713 162.1373
## fp         5.666068 143.9333
## mp         5.491281 137.0179
##
## Coefficients of linear discriminants:
##                            LD1          LD2          LD3
## Age                 0.02148906 -0.006866984  0.093597871
## Chest.pain.type2    0.52943538  0.928007222  0.510056058
## Chest.pain.type3    0.21753853  0.536951302  0.662536788
## Chest.pain.type4    1.81578957  0.778844269 -0.244046869
## log(BP)             2.60993511  4.654711729 -3.214179747
## log(Cholesterol)    1.19048863  2.161743180  2.335368562
## Max.HR             -0.01890306  0.029340275  0.006133466
```

```
##
## Proportion of trace:
##    LD1    LD2    LD3
## 0.7063 0.1911 0.1026
```

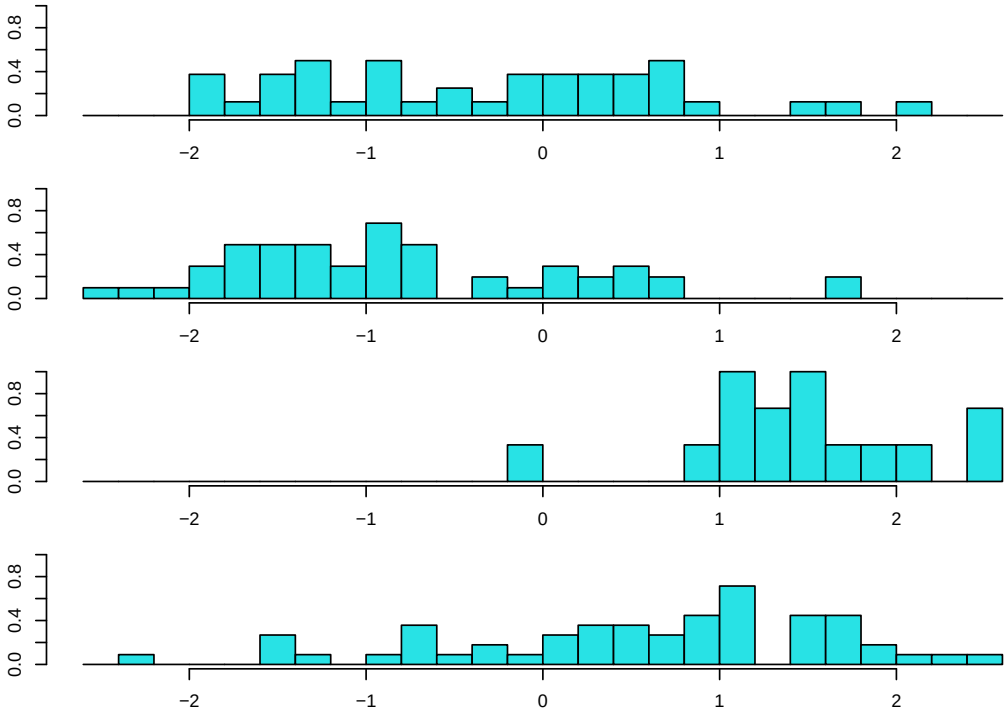


Figure 12: Histogram of the classifications from our LDA model.

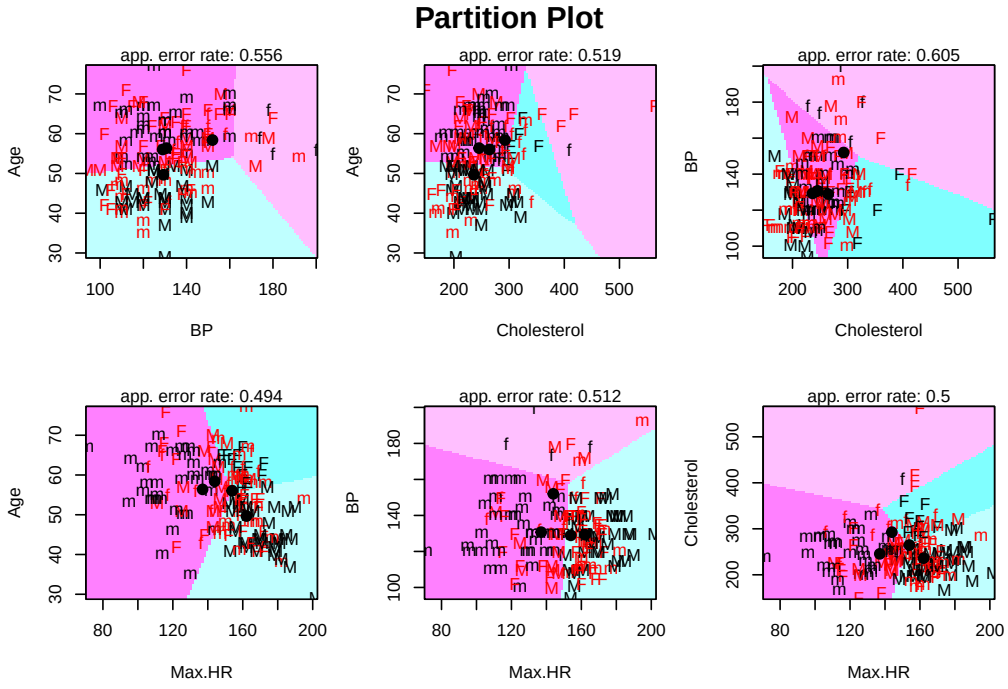


Figure 13: Partition Plot of the LDA classification.

Our pairs plot, Figure 11 shows the general distribution of our four classes. This uses the log values for cholesterol and BP. We see that there does seem to be differences between our four classes, specifically in the max HR variable. This indicates that this separation may help with our analysis.

```
##              Actual
## realisticpred FA MA fp mp
##            FA  8  7  1  8
##            MA 15 18  0  5
##            fp  0  0  2  5
##            mp  4  7  2 26
```

Table 14: Confusion matrix for predicting Heart Disease with interaction between Sex and Heart Disease variables using LDA.

	FA	MA	fp	mp
FA	8	7	1	8
MA	15	18	0	5
fp	0	0	2	5
mp	4	7	2	26

Table 15: Overall statistics for predicting heart disease using interactions between Sex and Heart Disease.

	x
Accuracy	50.00%
Kappa	27.54%
AccuracyLower	40.22%
AccuracyUpper	59.78%
AccuracyNull	40.74%
AccuracyPValue	3.22%
McnemarPValue	NA

```
##           Sensitivity Specificity Pos Pred Value Neg Pred Value Precision
## Class: FA   0.2962963   0.8024691      0.3333333      0.7738095 0.3333333
## Class: MA   0.5625000   0.7368421      0.4736842      0.8000000 0.4736842
## Class: fp   0.4000000   0.9514563      0.2857143      0.9702970 0.2857143
## Class: mp   0.5909091   0.7968750      0.6666667      0.7391304 0.6666667
##              Recall        F1 Prevalence Detection Rate Detection Prevalence
## Class: FA 0.2962963 0.3137255  0.2500000     0.07407407           0.22222222
## Class: MA 0.5625000 0.5142857  0.2962963     0.16666667           0.35185185
## Class: fp 0.4000000 0.3333333  0.0462963     0.01851852           0.06481481
## Class: mp 0.5909091 0.6265060  0.4074074     0.24074074           0.36111111
##           Balanced Accuracy
## Class: FA         0.5493827
## Class: MA         0.6496711
## Class: fp         0.6757282
## Class: mp         0.6938920
```

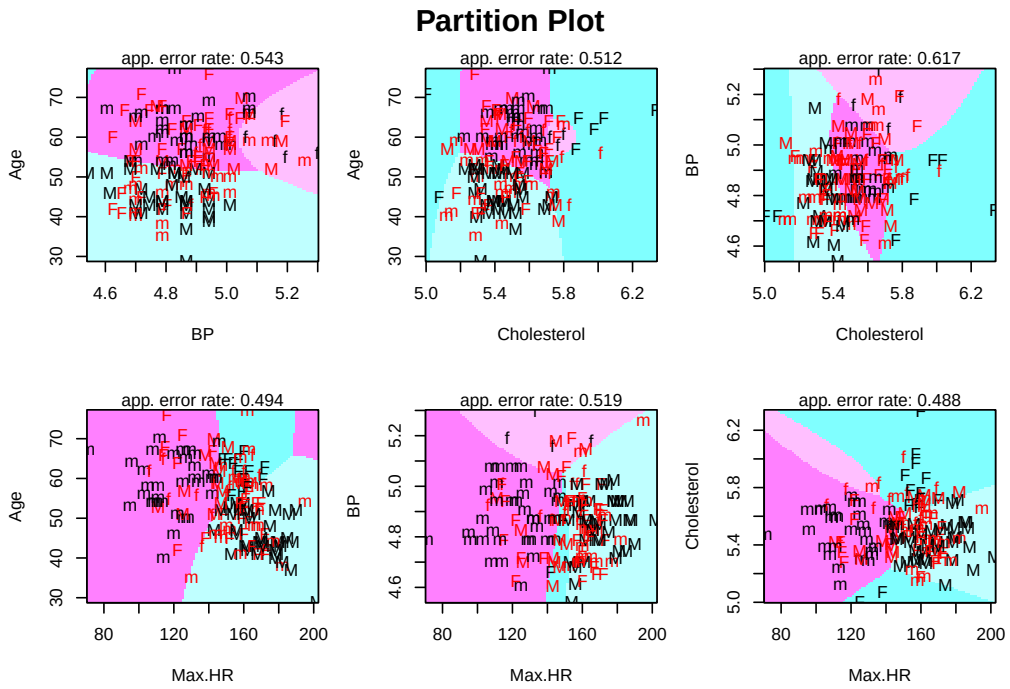


Figure 14: Partition Plot of the Bayesian Method of Classification.

Table 16: Confusion matrix for predicting heart disease using the Bayesian classification method.

	FA	MA	fp	mp
FA	6	5	1	5
MA	14	20	0	9
fp	0	0	2	6
mp	7	7	2	24

Table 17: Overall statistics for predicting heart disease using the Bayesian classification method.

	x
Accuracy	48.15%
Kappa	24.67%
AccuracyLower	38.43%
AccuracyUpper	57.97%
AccuracyNull	40.74%
AccuracyPValue	7.17%
McnemarPValue	NA

```
##           Sensitivity Specificity Pos Pred Value Neg Pred Value Precision
## Class: FA   0.2222222   0.8641975      0.3529412      0.7692308 0.3529412
## Class: MA   0.6250000   0.6973684      0.4651163      0.8153846 0.4651163
## Class: fp   0.4000000   0.9417476      0.2500000      0.9700000 0.2500000
## Class: mp   0.5454545   0.7500000      0.6000000      0.7058824 0.6000000
##              Recall        F1 Prevalence Detection Rate Detection Prevalence
```

```
## Class: FA 0.2222222 0.2727273  0.2500000     0.05555556           0.15740741
## Class: MA 0.6250000 0.5333333  0.2962963     0.18518519           0.39814815
## Class: fp 0.4000000 0.3076923  0.0462963     0.01851852           0.07407407
## Class: mp 0.5454545 0.5714286  0.4074074     0.22222222           0.37037037
##           Balanced Accuracy
## Class: FA         0.5432099
## Class: MA         0.6611842
## Class: fp         0.6708738
## Class: mp         0.6477273
```

Conclusion

It is important to note that our dataset was not balanced, and we had more males in our dataset, and more people without heart disease in our dataset. This meant our four classes were not equal in size. We also had different prior probabilities, which is why we used Bayesian prediction methods as well as LDA.

Interestingly, our accuracy of predicting heart disease decreased significantly after inducing the four classes. This was the case for both LDA and Bayesian.

For the LDA, we saw that we had an overall accuracy of 50% (40.22%-59.68% CI). This means our model correctly classified 50% of the data. Table 14 shows us the confusion Matrix and Table 15 shows the Accuracy Stats.

We can see in Figure 12 a Histogram of the classified observations via LDA. The overlap indicates potentially incorrectly classified observations. We have quite a bit of overlap which is consistent with the 50% accuracy gathered.

In Figure 13 it is a Partition Plot for the numerical variables of our dataset. We see that there is error rates which are high, from about 0.5-0.6.

The sensitivity and specificity of the output allows us the look further into the separate classes. We want to be able to accurately predict having heart disease, so the sensitivity value is more valuable to our research question. We see our model is much better at predicting correctly if males have heart disease than females. Both female classes have sensitivity of below 40%, which is very low, indicating our model is not accurate. However for specificity this is the opposite, with our model being better at accurately classifying females into the correct class, than males. All the values of specificity are quite high, however female present (fp), is 95.15%, which is very high. It is important to note that there is far less females with heart disease that were tested in our data set.

For the Bayesian prediction model, we had a low overall accuracy of 48.15% (38.43%-57.97% CI). Table 16 shows the confusion matrix for this model. Table 17 shows the overall statistics.

Figure 14 shows the partition plot for the Bayes method of classification. We again seeing high error rates from 0.48-0.62. This indicates our model is not great at classifying Heart Disease Presence.

We had similar patterns with the specificity and similarity of the model. Our sensitivity for males increased for MA, and decreased for mp. For females, it decreased for FA and stayed the same for fp. We see these changes due to the nature of Bayesian, which takes into consideration the prior probabilities, which were significantly different between the groups. For specificity, both male groups decreased. This indicates that this model is worse with incorrectly classifying observations to their correct group.

The LDA again had higher accuracy than the Bayesian classifier. They were both better at predicting heart disease in males. This could be due to our dataset having more males, giving it more opportunity for training. Although it may be more likely be due to inherent differences between male and females.

These clear differences with the accuracy between female and male groups may indicate that there is discrepencies between sex and heart disease characteristics. We may need different predictors for females, such as weight to increase the accuracy of our prediction of heart disease. This is consistent with literature on heart disease. We see that females are diagnosed with heart disease later in life, and have different symptoms.

Although this disease is often though of as a 'man disease' this is not the case. With it often actually being more common in females, and them having a higher mortality rate.

Bibliography

[1] Ministry of Health 2021. Custom requested mortality dataset provided to Heart Foundation from the NZ Mortality Collection. August 2021

[2] Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J. J., Sandhu, S., Guppy, K. H., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American journal of cardiology*, 64(5), 304–310. https://doi.org/10.1016/0002-9149(89)90524-9

[3] Bobbio, M., Detrano, R., Shandling, A. H., Ellestad, M. H., Clark, J., Brezden, O., Abecia, A., & Martinez-Caro, D. (1992). Clinical Assessment of the Probability of Coronary Artery Disease: Judgmental Bias from Personal Knowledge. *Medical Decision Making*, 12(3), 197–203. https://doi.org/10.1177/0272989X9201200305

[4] Janosi,Andras, Steinbrunn,William, Pfisterer,Matthias, and Detrano,Robert. (1988). Heart Disease. UCI Machine Learning Repository. https://doi.org/10.24432/C52P4X.

[5] National Center for Chronic Disease Prevention and Health Promotion, Division for Heart Disease and Stroke Prevention (2023). https://www.cdc.gov/bloodpressure/facts.htm

[6] The National Heart Foundation of New Zealand, Beta Blockers (2023). https://www.heartfoundation.org.nz/your-heart/heart-treatments/medications/beta-blockers