University of Strathclyde

Department Of Mathematics and Statistics

MM924: Statistical Modelling and Analysis Project

(Multivariate Analysis Project)

Name of Student: Eugene Osae

Student Number: 202222159

TABLE OF CONTENT

Abstract………………………………………………………………………………..3
Project Questions and Answers……………………………………………………………4
1. Exploratory data analysis………………………………………………………………4
(i) Size, nature, and structure of the data …………………………………………………4
(ii) Summary measure, plots, a potential risk factor for CHD ……………………………..4
(iii) Covariances and correlations of the variables with CHD ……………………………4

2. Logistic Regression ……………………………………………………………………..5
(i) Data splitting and variable selection ……………………………………………..... 5
(ii) Fit of the model, coefficients, and confidence intervals ………………………………5
(iii) Optimal balance point for model's sensitivity and specificity ………………………......6
(iv) Correct classification rate, sensitivity and specificity ………………………………..6

3. Principal Component Analysis …………………………………………………………..6
(i) Principal Component Analysis …………………………………………………………6
(ii) Principal Component Analysis vs Logistic Regression…………………………………...7

References ……………………………………………………………………………………..8
Appendices ……………………………………………………………………………………..9
Appendix 1: Plots ………………………………………………………………………………….9
Appendix 2: Code Output ………………………………………………………………………9-16

1. Abstract

Introduction

Coronary heart disease (CHD) is a major cause of death worldwide. CHD is also known as Ischaemic heart disease or coronary artery disease, and it is mainly caused by blockage of the heart's blood supply by a build-up of fatty substances in the coronary arteries. The aim of this study is to analyse the variations in risk factors associated with CHD and how they affect the possibility of one contracting CHD.

Methods

The dataset used in this project consists of 462 rows and 8 variables which shows the various measurement of all the potential risk factors associated with CHD. The data shows information on whether an individual has CDH. The exploratory data analysis takes into consideration the size, nature and structure of the data, summaries, and visualization to determine the potential risk factors associated with CHD. The model is built using logistic regression using the forward selection method to select the appropriate variables to determine the relationship between the potential risk factors and CHD status. The Principal Component Analysis (PCA) is deployed to investigate the variables for further identification of the effects of each variable on the probability of contracting CHD.

Results

The exploratory data analysis shows that the potential risk factors for CHD comprise the values of high systolic blood pressure, tobacco usage, low-density lipoprotein cholesterol, percentage of body fat, type A behaviour pattern score, body mass index, current alcohol consumption, and age. The sensitivity, specificity, and optimal balance point for the fitted data are estimated. From the PCA sbp, tobacco, ldl, and adiposity indicated 72% of the total variations in the dataset, it is highly likely that obesity and age will also be contributing factors for the diagnoses of CHD. There is consistency across the various analyses which indicates the accuracy of the findings. The findings from the exploratory data analysis, covariance, correlation, and PCA give details of the significance of systolic blood pressure, tobacco, Low-density lipoprotein, adiposity, obesity, alcohol, and age as the potential risk factors of an individual diagnosed with CHD.

Discussion and Conclusion

The results of this project have helped in the determination and comprehension of the potential risk factors of CHD. The logistic regression and principal component analysis indicated insights into the datasets emphasising more on the relationship between the variables and CHD status. Furthermore, the findings emphasise the significance of several risk factors to prevent and manage CHD. To prevent and manage CHD, models must be built to accurately predict the potential risk factors that can lead to the contraction of CHD. Statistical modelling is one of the methods that can be used for the classification and prediction of risk factors related to CHD. In conclusion, the application of exploration data analysis with logistic regression, and principal component analysis has demonstrated that models can be used in the health sector for drug improvement, managing lifestyles and other factors to prevent and manage CHD.

2. Project Questions and Answers

Exploratory Data Analysis
Question 1.

(i)Size, nature, and structure of the dataset

- The size of the dataset is made up of 462 observations i.e. rows and the size is moderate.
- The nature of the dataset consists of variables with information about the potential risk factors for CHD and the response variable CHD shows whether an individual has CHD or does not have CHD, it is a categorical variable, and the remaining variables are continuous in nature.
- The structure of the dataset is a two-dimensional array with rows corresponding to individual rows and columns which shows that there are differences in measurement recorded for each variable. The structure of the dataset makes it convenient for data analysis.

(ii) Summary measures, plots, and potential risk factors for CHD

The average systolic blood pressure is 138.30 mmHg, with a range from 101 mmHg to 218 mmHg. The average yearly tobacco usage is 3.64kg with a maximum of 31.20kg and a minimum of 0kg. The average of Low-density lipoprotein cholesterol level is 4.74 mmol/L, with a maximum level of 0.98 mmol/L and a minimum level of 15.33 mmol/L. The average percentage of body fat is 25.41%, with a maximum of 6.74% and a minimum of 42.49%. The average type-A behaviour pattern score is 53.1, ranging from a minimum of 13 to a maximum of 78. The average body mass index is 26.04 with a range of 14.7 to 46.58. The average current alcoholic consumption is 17.04 litres, ranging from 0 litres to 147.19 litres. The average age is 42.82 years, with a minimum age of 15 and maximum age of 64 years old. 160 individuals were diagnosed with CHD.

The boxplot in Appendix 1, figure 1 compares the range of potential risk factors for the retrospective sample of males with a high risk of heart disease. In the comparison, individuals with systolic blood pressure, yearly tobacco usage, low-density lipoprotein Cholesterol, percentage of body fat, type-A behaviour pattern, body mass index, current alcohol and age values recorded the highest diagnosis of CHD. The comparison of the median also showed that these variables are slightly higher for those with CHD than those without CHD. This shows that all the variables are potential risk factors for CHD when the data was visualized using the boxplot. The variables in the dataset also showed that there is the presence of outliers and further analysis can be carried out to investigate whether there is an association between the variables and CHD.

(iii) Covariance and Correlations of the Variables with CHD

From Appendix 2, table 1 and 2 indicates the estimates of the covariances and correlations between the variable with CHD status. The interpretation is shown in the tables below:

Variable	**Relationship with CHD Status Interpretation**
sbp	Positive Low Covariance- Weak Relationship
tobacco	Positive Low Covariance- Weak Relationship
ldl	Positive Low Covariance- Weak Relationship
Adiposity	Positive Low Covariance- Weak Relationship
Typea	Negative Low Covariance- Weak Relationship
Obesity	Positive Low Covariance- Weak Relationship
Alcohol	Positive Low Covariance- Weak Relationship
Age	Positive Low Covariance- Weak Relationship

Covariance of the variables with CHD

Variable	Relationship with CHD Status Interpretation
sbp	Positive weak
tobacco	Positive moderate
ldl	Positive moderate
Adiposity	Positive moderate
Typea	Positive weak
Obesity	Positive weak
Alcohol	Positive weak
Age	Positive moderate

Correlation of the variables with CHD

The comparison of the relationship between the variables with CHD status using the covariance and correlation indicates that tobacco, ldl, adiposity and age variables showed a strong relationship with CHD status, and it is the most appropriate variables in the dataset to fit the model.

Question 2. Logistic Regression
(i.) Data splitting and variable selection

The full model is fitted, and the backward selection method is applied. The backward selection method is applied by eliminating the variables with high p-values until all remaining variables have a significant p-value, this method is applied because the number of variables is very low and it is more efficient than the forward selection method. The variables that remained in the final model are tobacco, Low-density lipoprotein, Type A behaviour pattern score and Age because they have lower p-values. The data is split into training and testing sets with the training set having 2/3 and testing set 1/3 of the dataset.

Train data:
False Positives (CHD=0 incorrectly predicted as CHD=1): 25
False Negatives (CHD=1 incorrectly predicted as CHD=0): 9
True Positives (CHD=1 correctly predicted as CHD=1): 22
True Negatives (CHD=0 correctly predicted as CHD=0): 83

Test set:
False Positives (CHD=0 incorrectly predicted as CHD=1): 31
False Negatives (CHD=1 incorrectly predicted as CHD=0): 66
True Positives (CHD=1 correctly predicted as CHD=1): 47
True Negatives (CHD=0 correctly predicted as CHD=0): 179

(ii) Assess the fit of the final model and interpret the coefficients and confidence intervals for the estimates.

The results of the final model show that it is a good fit since the deviance residuals are low and the Akaike Information Criterion values are also low, and the estimates of the coefficients have a lower p-value. The estimates of the coefficients of the variables indicate that the model variables have a great effect on the likelihood of an individual being diagnosed with CHD. The confidence interval shows that there is a strong relationship between higher values of this variable and is more likely to cause CHD. This shows that 95% are confident that these variables in the model have a greater effect on the likelihood of an individual getting CHD.

(iii) Optimal Balance point for model's sensitivity and specificity
The sensitivity shows the proportion of CHD cases that are correctly identified by the model while the specificity also shows that the proportion of individuals without CHD cases are correctly identified by the model. For a mode to be classified as a good model, the sensitivity and specificity should be large, but this is not possible because one increases while the other decreases. This is balanced using the ROC curve as shown in Appendix 1, figure 2. The optimal balance point value from the ROC curve is 0.283 which shows that there is a decrease in false negatives and an increase in true positives in the test dataset and an increase in false negatives and an increased in true positives in the train datasets.

(iv) Correct Classification rate, sensitivity, and specificity
The correct classification rate for the training dataset and testing dataset is approximately 77% with the sensitivity and specificity of the testing datasets of approximately 62% and 85% respectively**.**

Question 3.
(i) Principal Component Analysis

From the observation of the correlation and covariance matrices, the variances are measured using different units and there are variations in the values of the variances, the correlation matrix is then applied for the PCA. It is observed that the relationship between the variables, such as adiposity is high (72%) and that of obesity and adiposity is also high (72%), etc.

- The correlation of sbp is positive with tobacco, ldl, adiposity, obesity, alcohol and age and negative correlation with typea.
- The correlation of tobacco is positive with sbp, ldl, adiposity, obesity, alcohol and age and negative with typea.
- The correlation of ldl is positive with sbp, tobacco, adiposity, typea, obesity and age and negative with alcohol.
- The correlation of adiposity is positive with sbp, tobacco, ldl, obesity, alcohol and age and negative with typea.
- The correlation of typea is positive with ldl, obesity and alcohol and negative with sbp, tobacco, adiposity, and age.
- The correlation of obesity is positive with all the variables.
- The correlation of alcohol is positive with all the variables except ldl.
- The correlation of age is positive for all the variables except typea.

From the cumulative proportion of the variance, it is observed that 45% of the total variance is described by the first component while 95% is described by the first four components. Few components represent a small proportion of the variance, and these four components show enough representation of the datasets. From Kaiser's criterion in the plot in Appendix 1, figure 3, the first three components show a variance that is higher than 1%. Since the first four components are 95%, in total variation, the first four components are displayed on the plot.

- The contrast between sbp, tobacco, ldl, adiposity, obesity, age versus typea values is PC1
- The contrast between ldl, typea, obesity versus sbp, tobacco and alcohol values is PC2.
- The contrast between age versus typea and alcohol values is PC3.
- The contrast between tobacco, ldl, typea, age versus sbp, obesity, alcohol is PC4.

A detailed interpretation of the various components and the relationship between the components and original datasets with their PC scores is as follows:

- PC1 showed a positive score with higher values for sbp, tobacco, ldl, adiposity, obesity and age with the indication that individuals with positive scores on PC1 are more likely to develop CHD.
- PC2 showed a positive score with higher values for ldl, typea and obesity with the indication that individuals with positive scores on PC2 are more likely to develop CHD.

- PC3 showed a positive score with higher values for age with the indication that individuals with positive scores on PC3 are more likely to develop CHD.
- PC4 indicate showed a positive score with higher values for tobacco, ldl, typea, and age with the indication that individuals with positive scores on PC4 are more likely to develop CHD.

The PCA showed that individuals with high values of sbp, tobacco, ldl, adiposity, typea, obesity and age are more likely to be diagnosed with CHD. Furthermore, the PCA is interpreted by using the biplot as shown in Appendix 1, figure 4. The interpretation of the biplot is shown below:

For example, 450 is strong on alcohol, unlike 424

- For example, 458 strong on sbp, unlike 220
- The variables typea and adiposity showed very little correlation
- The variables obesity and ldl showed very little correlation
- The variables age and adiposity showed very little correlation
- The variables sbp, age and adiposity are negatively correlated with PC1

(ii) Comparison of Principal Component Analysis with Logistic Regression

- The observation made from the variables that contribute to the likelihood of an individual being diagnosed with CHD is common in both Logistic regression and PCA with variables tobacco, ldl, typea, and age variables.

- PCA is focused on interpreting the relationships between the variables while logistic regression focused on the relationships between each variable and the response variable.

- From the result indicated in both analyses, it is better to use logistic regression than to use PCA because the variables are smaller in number as indicated by the dataset. Furthermore, PCA provides more insights into the dataset when used with logistic regression specifically to show the accuracy of the analysis.

- It is best to use PCA in exploratory data analysis while logistic regression is best used for the prediction of CHD.

3. References

NHS inform, Coronary heart disease available at: https:// www.nhsinform.scot/illness-and-blood-vesses/conditions/coronary-heart-disease.

International journal of statistics and data science, coronary heart disease prediction using binary logistic regression based on principal component analysis, University of Islam Indonesia available at: https: // journal.uii.ac.id/ ENTHUSIASTIC/ article/download/ 23130/13610 (1 April 2022)

University of Strathclyde. (2023) Week 6: Week6 video 6.2 Multivariate data and big data.pptx', MM924: Statistical Modelling and Analysis. University of Strathclyde. 18 March. (1 July 2023)

University of Strathclyde. (2023). 'Week 6: Week6_video_6.6_Looking at associations- 1 continuous and 1 categorical variable- part 1.pptx', MM924: Statistical Modelling and Analysis. University of Strathclyde. 18 March. (2 July 2023).

University of Strathclyde. (2023). 'Week 7: Model Building', MM924: Statistical Modelling and Analysis. (2 July 2023)

University of Strathclyde. (2023). 'Week 7: Model Comparisons', MM924: Statistical Modelling and Analysis. University of Strathclyde. (3 July 2023).

University of Strathclyde. (2023). 'Week 7: Classification and Data Splitting', MM924: Statistical Modelling and Analysis. University of Strathclyde. (3 July 2023).

University of Strathclyde. (2023). 'Week 7: Model Prediction', MM924: Statistical Modelling and Analysis.
University of Strathclyde. (July 2023).

University of Strathclyde. (2023). 'Week 7: Model Performance', MM924: Statistical Modelling and Analysis. University of Strathclyde. (3 July 2023)

University of Strathclyde. (2023). 'Week 10: Principal Component Analysis, MM924: Statistical Modelling and Analysis. University of Strathclyde (3 July 2023).

Appendix 1

Exploratory Data Analysis

(ii) Summary measures and plot of potential risk factors

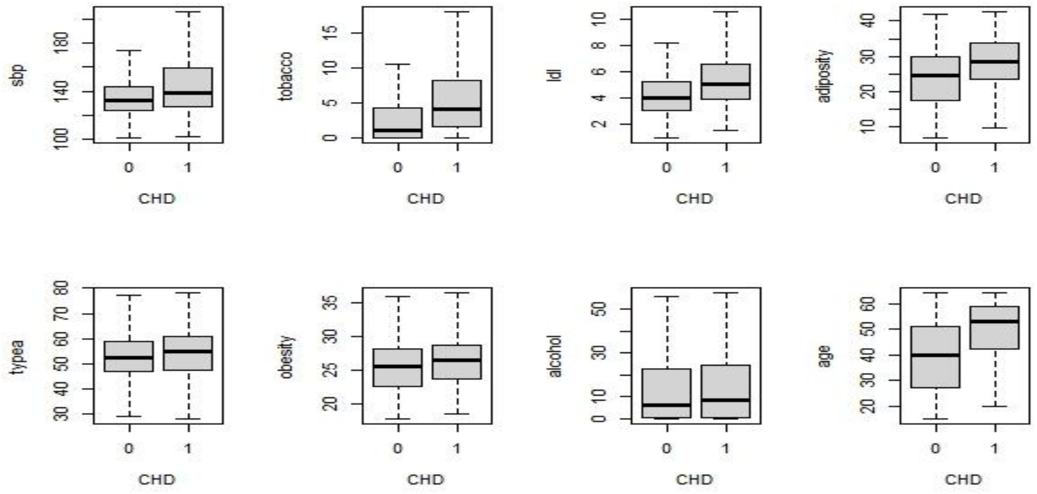

Figure 1. Boxplot of variables against CHD

- **Logistic Regression**

(iii) Optimal balance point for sensitivity and specificity

ROC curve for Training and Testing dataset

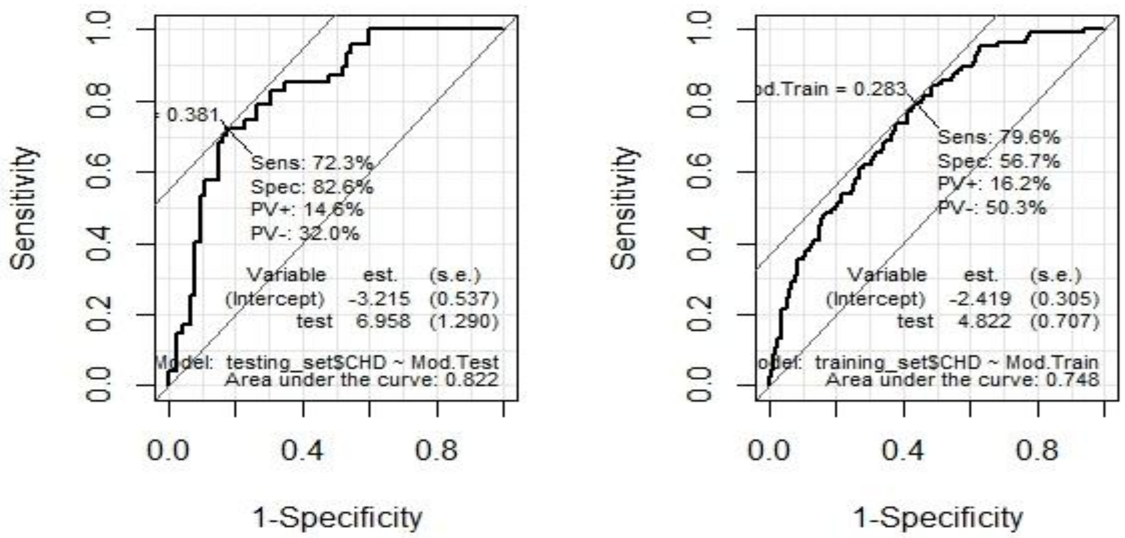

Figure 2. Sensitivity and Specificity of Train and Test data

3. Principal Component Analysis

(i) Plot of Principal Components of CHD

Plot of Principal Components of CHD

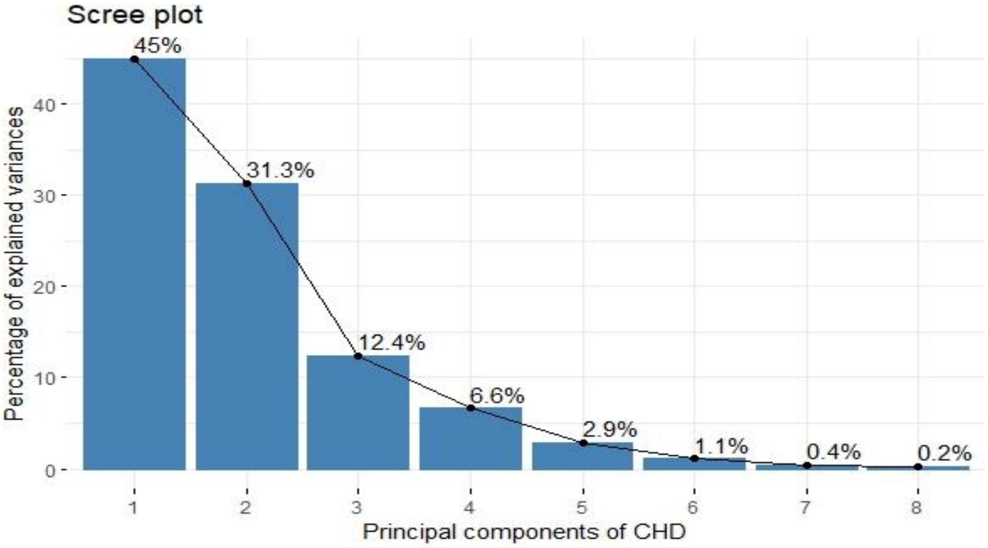

Figure 3. Principal components against variances

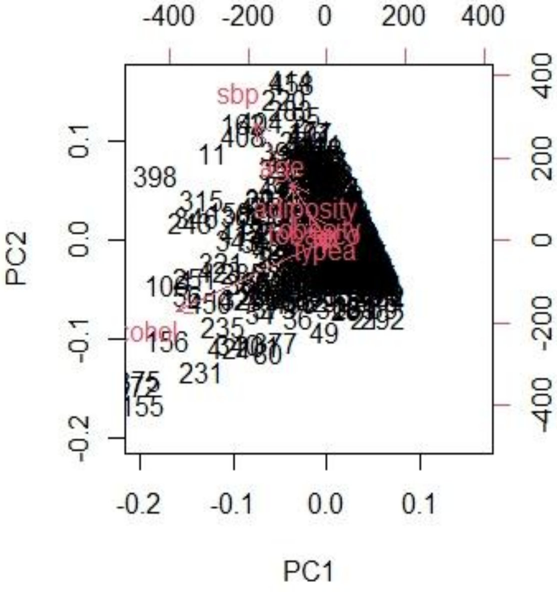

Figure 4 : PCA Biplot

.

Appendix 2

Que 1. Covariances and Correlations for the relationship with CHD status

```
> print(Cov.Matrix, digits=3)
```

```
                  CHD.data.sbp CHD.data.tobacco CHD.data.ldl
CHD.data.sbp          6.89e-05         2.11e-04     5.76e-05
CHD.data.tobacco      2.11e-04         1.42e-02     1.22e-03
CHD.data.ldl          5.76e-05         1.22e-03     1.38e-03
CHD.data.adiposity    8.11e-05         1.06e-03     4.67e-04
CHD.data.typea       -7.69e-06         3.62e-06     1.32e-05
CHD.data.obesity      2.55e-05         2.28e-04     1.64e-04
CHD.data.alcohol      2.27e-04         5.47e-03     8.30e-05
CHD.data.age          9.35e-05         1.97e-03     4.22e-04
                   CHD.data.adiposity CHD.data.typea
CHD.data.sbp                 8.11e-05      -7.69e-06
CHD.data.tobacco             1.06e-03       3.62e-06
CHD.data.ldl                 4.67e-04       1.32e-05
CHD.data.adiposity           7.45e-04      -1.84e-05
CHD.data.typea              -1.84e-05       2.03e-04
CHD.data.obesity             2.25e-04       1.04e-05
CHD.data.alcohol             6.71e-04       5.34e-05
CHD.data.age                 5.34e-04      -3.02e-05
                   CHD.data.obesity CHD.data.alcohol CHD.data.age
CHD.data.sbp               2.55e-05         2.27e-04     9.35e-05
CHD.data.tobacco           2.28e-04         5.47e-03     1.97e-03
CHD.data.ldl               1.64e-04         8.30e-05     4.22e-04
CHD.data.adiposity         2.25e-04         6.71e-04     5.34e-04
CHD.data.typea             1.04e-05         5.34e-05    -3.02e-05
CHD.data.obesity           1.37e-04         2.24e-04     1.25e-04
CHD.data.alcohol           2.24e-04         2.69e-02     9.11e-04
CHD.data.age               1.25e-04         9.11e-04     8.85e-04
```

Table 1. Covariances of variables with CHD status

```
> print(Corr.Matrix, digits = 3)
                   CHD.data.sbp CHD.data.tobacco CHD.data.ldl
CHD.data.sbp              1.000          0.21344       0.1868
CHD.data.tobacco          0.213          1.00000       0.2748
CHD.data.ldl              0.187          0.27482       1.0000
CHD.data.adiposity        0.358          0.32653       0.4609
CHD.data.typea           -0.065          0.00214       0.0249
CHD.data.obesity          0.262          0.16361       0.3757
CHD.data.alcohol          0.167          0.28050       0.0136
CHD.data.age              0.379          0.55615       0.3824
                   CHD.data.adiposity CHD.data.typea
CHD.data.sbp                   0.3581       -0.06496
CHD.data.tobacco               0.3265        0.00214
CHD.data.ldl                   0.4609        0.02489
CHD.data.adiposity             1.0000       -0.04743
CHD.data.typea                -0.0474        1.00000
CHD.data.obesity               0.7028        0.06250
CHD.data.alcohol               0.1502        0.02285
CHD.data.age                   0.6577       -0.07122
                   CHD.data.obesity CHD.data.alcohol CHD.data.age
CHD.data.sbp           0.2625       0.1668     0.3786
CHD.data.tobacco         0.1636       0.2805     0.5562
CHD.data.ldl          0.3757       0.0136     0.3824
CHD.data.adiposity       0.7028       0.1502     0.6577
CHD.data.typea         0.0625       0.0229     -0.0712
CHD.data.obesity        1.0000        0.1166     0.3594
CHD.data.alcohol        0.1166        1.0000     0.1867
CHD.data.age           0.3594       0.1867     1.0000
```

Table 2. Correlations of variables with CHD status

Que 2. Logistics Regression and Data Splitting

(i) Data splitting and variable selection

```
> Conf.matrix.tr

         0   1
  FALSE 179  66
  TRUE   31  47
```

```
> Conf.matrix.te

        0  1
  FALSE 83 25
  TRUE   9 22
```

Results 1. Confusion Matrices for Training and Testing datasets

(ii) Asses the fit of the final model and interpret the coefficients and confidence intervals for the estimates

```
> summary(z6)

Call:
glm(formula = CHD == "1" ~ tobacco + ldl + typea + age, family = binomial,
    data = training_set)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9545  -0.9139  -0.5101   1.0431   2.3495

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.82305    0.99864  -4.830 1.37e-06 ***
tobacco      0.08433    0.03010   2.802  0.00508 **
ldl          0.13032    0.05842   2.231  0.02569 *
typea        0.02085    0.01361   1.531  0.12565
age          0.04678    0.01137   4.114 3.89e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 418.19  on 322  degrees of freedom
Residual deviance: 357.66  on 318  degrees of freedom
AIC: 367.66

Number of Fisher Scoring iterations: 4
```

Results 2. Summary of Final Model

```
> # Confidence Interval of the estimates
> cbind(Mymodel$coefficients,confint.default(Mymodel))
                                2.5 %      97.5 %
(Intercept) -4.82304831 -6.780349732 -2.86574688
tobacco      0.08433392  0.025346625  0.14332122
ldl          0.13032372  0.015827594  0.24481984
typea        0.02084970 -0.005833146  0.04753254
age          0.04677710  0.024490139  0.06906407
> exp(cbind(Mymodel$coefficients,confint.default(Mymodel)))
                              2.5 %     97.5 %
(Intercept) 0.008042235 0.001135878 0.05694059
tobacco     1.087992139 1.025670582 1.15410046
ldl         1.139197102 1.015953514 1.27739116
typea       1.021068572 0.994183834 1.04868033
age         1.047888411 1.024792486 1.07150485
```

Result 3. Confidence intervals for the estimates

(iv)

```
> Conf.matrix1

         0   1
  FALSE 156  49
  TRUE   54  64

> Conf.matrix2

        0  1
  FALSE 78 18
  TRUE  14 29
```

Result 4. Confusion Matrices for Training and Testing Dataset

```
 > Correct.Class.Rate
 [1] 76.97842
```

```
> Sensitivity
[1] 61.70213

> Specificity
[1] 84.78261
```

Result 5. Correct Classification rate, Sensitivity and Specificity

Que 3. Principal Component Analysis

(i) Principal Component Analysis

```
Corr.Matrix<-round(cor(Mydata[,-1]),2)
> Corr.Matrix
            sbp tobacco   ldl adiposity typea obesity alcohol   age
sbp        1.00    0.21  0.16      0.36 -0.06    0.24    0.14  0.39
tobacco    0.21    1.00  0.16      0.29 -0.01    0.12    0.20  0.45
ldl        0.16    0.16  1.00      0.44  0.04    0.33   -0.03  0.31
adiposity  0.36    0.29  0.44      1.00 -0.04    0.72    0.10  0.63
typea     -0.06   -0.01  0.04     -0.04  1.00    0.07    0.04 -0.10
obesity    0.24    0.12  0.33      0.72  0.07    1.00    0.05  0.29
alcohol    0.14    0.20 -0.03      0.10  0.04    0.05    1.00  0.10
age        0.39    0.45  0.31      0.63 -0.10    0.29    0.10  1.00

>

Cov.Matrix<- round(var(Mydata[,-1]),2)
> Cov.Matrix
             sbp tobacco   ldl adiposity  typea obesity alcohol
sbp       420.10   19.98  6.72     56.85 -11.56   20.56   70.30
tobacco    19.98   21.10  1.51     10.24  -0.66    2.41   22.58
ldl         6.72    1.51  4.29      7.10   0.90    2.88   -1.69
adiposity  56.85   10.24  7.10     60.54  -3.30   23.49   19.11
typea     -11.56   -0.66  0.90     -3.30  96.38    3.06    9.49
obesity    20.56    2.41  2.88     23.49   3.06   17.76    5.32
alcohol    70.30   22.58 -1.69     19.11   9.49    5.32  599.32
age       116.41   30.22  9.43     71.15 -14.72   17.96   36.17
             age
sbp       116.41
tobacco    30.22
ldl         9.43
adiposity  71.15
typea     -14.72
obesity    17.96
alcohol    36.17
age       213.42
```

Result5. Covariance and Correlation Matrices

```
PCA.Mydata <- prcomp(Mydata[,-1], scale=FALSE)
> print(PCA.Mydata, digits=2)   # Compute the Standard Deviations
Standard deviations (1, .., p=8):
[1] 25.4 21.2 13.3  9.8  6.4  4.0  2.5  1.8
```

```
Rotation (n x k) = (8 x 8):
              PC1     PC2     PC3      PC4       PC5      PC6      PC7
sbp       -0.41458  0.774  0.4773 -0.0036  0.03754 -0.0042  0.0029
tobacco   -0.05683  0.042 -0.1153  0.0230  0.08934  0.9862 -0.0133
ldl       -0.00642  0.024 -0.0393  0.0262 -0.08869  0.0223  0.0687
adiposity -0.09659  0.164 -0.2894  0.0836 -0.78734  0.0271  0.4872
typea     -0.00045 -0.055  0.0893  0.9903  0.08258 -0.0177  0.0319
obesity   -0.03085  0.056 -0.0727  0.0785 -0.48570  0.0180 -0.8650
alcohol   -0.87851 -0.476  0.0082 -0.0277 -0.00088 -0.0286 -0.0024
age       -0.20687  0.373 -0.8126  0.0651  0.34659 -0.1570 -0.0923
              PC8
sbp        0.00051
tobacco   -0.01981
ldl        0.99199
adiposity -0.12309
typea     -0.01566
obesity    0.00958
alcohol    0.00784
age       -0.00355
```

Result 6. Rotation Matrix

```
> sum(PCA.Mydata$sdev^2)
[1] 1432.906

# Percentage of Variance
> round((PCA.Mydata$sdev^2)/ sum(PCA.Mydata$sdev^2)*100,2)
[1] 45.00 31.31 12.42  6.64  2.86  1.11  0.42  0.23
> round(cumsum((PCA.Mydata$sdev^2)/ sum(PCA.Mydata$sdev^2)*100),2)
[1]  45.00  76.31  88.73  95.37  98.23  99.34  99.77 100.00
```

```
 summary(PCA.Mydata)
Importance of components:
                        PC1      PC2      PC3      PC4      PC5      PC6
Standard deviation     25.39 21.1813 13.3411 9.75528 6.40345 3.98972
Proportion of Variance  0.45  0.3131  0.1242 0.06641 0.02862 0.01111
Cumulative Proportion   0.45  0.7631  0.8873 0.95369 0.98231 0.99342
                          PC7      PC8
Standard deviation     2.46646 1.82973
Proportion of Variance 0.00425 0.00234
Cumulative Proportion  0.99766 1.00000
```

Result 7: PCA Summary for Dataset

```
> round(cor(Mydata[,-1],PCA.Mydata.Pr[,1:4]),2)
            PC1   PC2   PC3   PC4
sbp       -0.51  0.80  0.31  0.00
tobacco   -0.31  0.19 -0.33  0.05
ldl       -0.08  0.25 -0.25  0.12
```

```
adiposity -0.32  0.45 -0.50  0.10
typea      0.00 -0.12  0.12  0.98
obesity   -0.19  0.28 -0.23  0.18
alcohol   -0.91 -0.41  0.00 -0.01
age       -0.36  0.54 -0.74  0.04
```

Result 8. Correlation of Variables with PCs

```
 head(PCA.Mydata.Pr)
            PC1        PC2        PC3        PC4       PC5       PC6
 [1,] -81.537531 -17.765019   2.885674 -5.8015894  6.416292  4.440353
 [2,]   6.447149  19.475008 -14.347381  3.9846369  3.189050 -6.241799
 [3,]  18.846155  -7.064247 -14.250668  0.2585229 -6.860110 -3.308823
 [4,] -24.243226  29.460564  -1.945432  0.2444004 -6.358088  1.613032
 [5,] -35.674507 -19.823447  -7.928035  6.5197777  1.672812  7.634303
 [6,]   3.372561  -1.005607  -7.857105 10.4337827 -9.468796  2.555543
            PC7        PC8
 [1,] -1.6318779  1.7623910
 [2,] -2.6088016 -0.8389953
 [3,]  0.2764259 -2.1042766
 [4,] -0.3265895  0.1346637
 [5,]  0.5230044 -1.5369282
 [6,]  1.3311178  0.2074338
```

```
> round(head(PCA.Mydata.Pr),2)
       PC1    PC2    PC3   PC4   PC5   PC6   PC7   PC8
[1,] -81.54 -17.77   2.89 -5.80  6.42  4.44 -1.63  1.76
[2,]   6.45  19.48 -14.35  3.98  3.19 -6.24 -2.61 -0.84
[3,]  18.85  -7.06 -14.25  0.26 -6.86 -3.31  0.28 -2.10
[4,] -24.24  29.46  -1.95  0.24 -6.36  1.61 -0.33  0.13
[5,] -35.67 -19.82  -7.93  6.52  1.67  7.63  0.52 -1.54
[6,]   3.37  -1.01  -7.86 10.43 -9.47  2.56  1.33  0.21
```