Modelling risk for Coronary Heart Disease

Introduction

Cardiovascular Disease, Coronary Heart Disease (CHD) inclusive, is the leading cause of deaths gloablly. The socioeconomic, mortality and morbidity of CHD emphasises the importance of preventative measures and timely diagnosis. The power of machine learning can effectively be utlisied in clinical medicine and diagnostics. With efficient, innovative and accessible technology, individuals can mitigate their risk for CHD by altering lifestyle and health choices. Research suggests, only 2%-7% of the general population possess no risk factors at all (Björn Dahlöf, 2010). While no exclusive factor has been identified as a direct cause to CHD, the data we have acquired accounts for various risk factors, modifiable and conventional, that are commonly associated to CHD e.g. age, gender and blood glucose levels. The data utilised in this program was retrieved from the Framingham Heart Study, a longitudinal cohort study, that began in 1948 with a sample size of 5,209 individuals. The current study invovles the third generation of participants assessing 15 major risk factors for CHD ("Framingham Heart Study"). Through the utilization of MATLAB's machine learning toolbox, our programme will attempt to model an individuals risk for CHD. The application of machine learning in cardiovascular medicine such as cardiovascular imaging and risk prediction holds substantial potential in healthcare (Mathur, P., Srivastava, S., Xu, X., & Mehta, J. L., 2020). Moreover, the use of models as such could improve efforts in reducing cardiovascular diseases worldwide.

Scrubbing Data

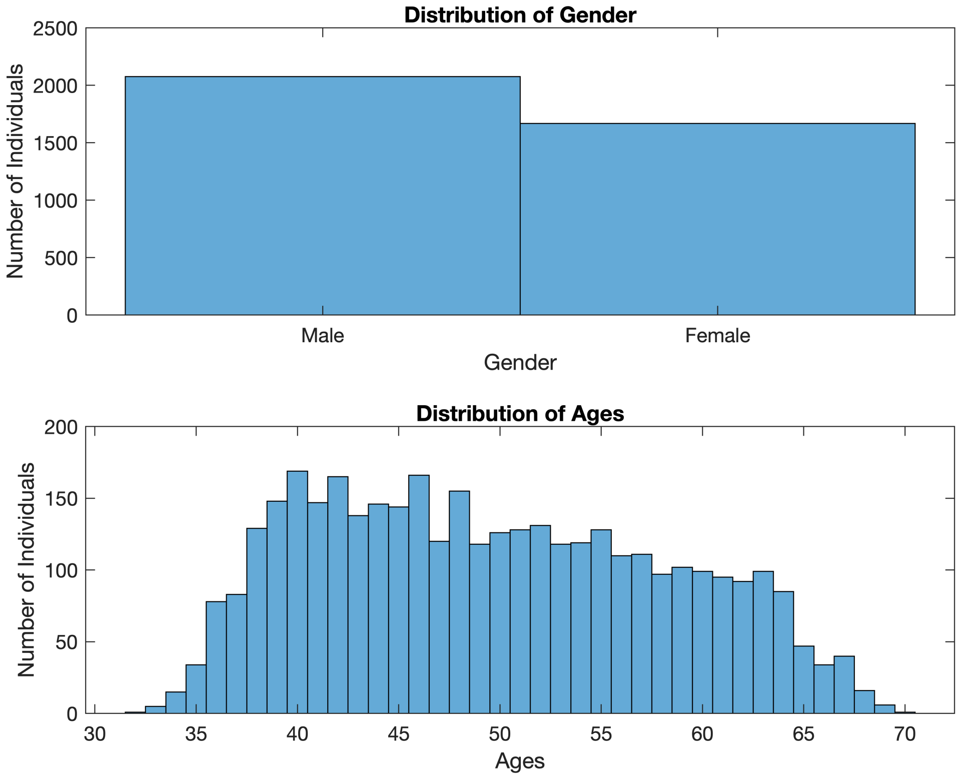
The excel file was initially converted into a raw data structure into MATLAB. Using the rmfield built-in function, the field of education was removed. The data was converted into a matrix; using a for loop, rows that contained NaN values were deleted from the array. The data was then converted into a table, with the following table headers: 'CigsPerDay', 'BPMeds', 'PrevalentStroke', 'PrevalentHypertension', 'Diabetes', 'TotalCholesterol', 'SystolicBP', 'DiabolicBP', 'BMI', 'HeartRate', 'GlucoseLevel' and 'TenYearCHD'.

```
Raw_Data = table2struct(readtable('FinalProjectRawData.xlsx', 'PreserveVariableNames',
Scrub_Data1 = rmfield(Raw_Data, "education");
[StructRow, StructCol] = size(Scrub_Data1);
Mat_Data = cell2mat(struct2cell(Scrub_Data1))';
for i = 1:StructRow
    Mat_Data(any(isnan(Mat_Data),i),:) = [];
end
Table_Data = array2table(Mat_Data,'VariableNames',{'Gender', 'Age', 'Smoker', 'CigsPerD
    'PrevalentHypertension', 'Diabetes', 'TotalCholesterol', 'SystolicBP', 'DiabolicBP'
```

Plotting Histogram

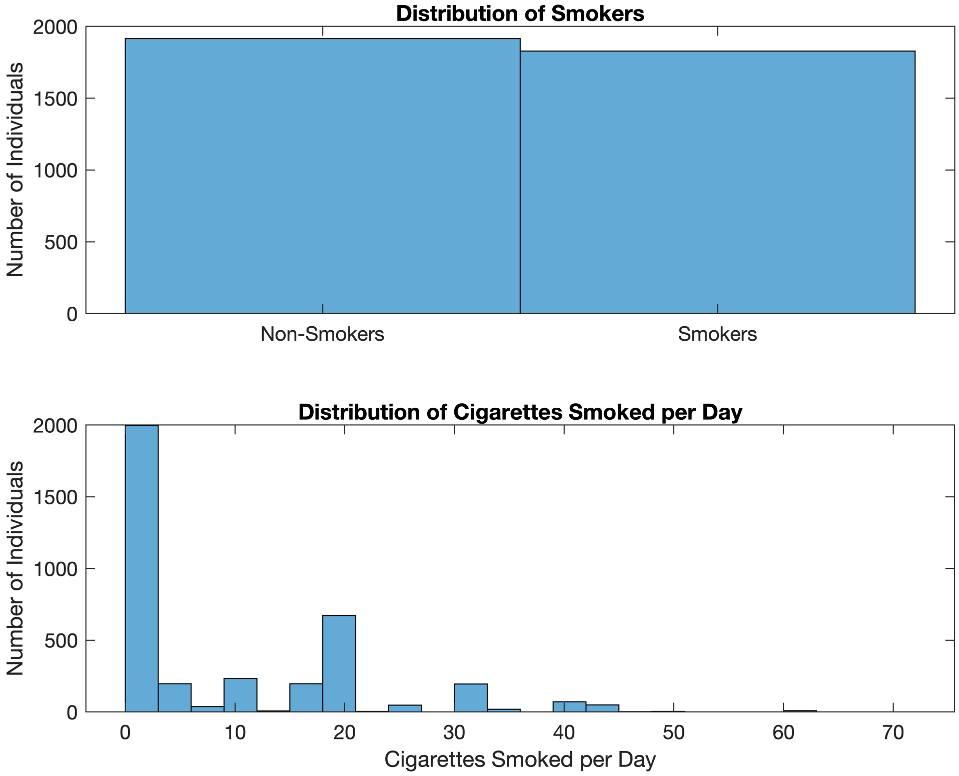
Histograms were used to present the distribution of each variable in our table. The number of individuals for each variable type was plotted. We will propose the following null hypothesis (Ho): All variables affect an individuals risk for CHD.

```
figure(1)
subplot(2,1,1);
histogram(Mat_Data(:,1))
title('Distribution of Gender')
xlabel('Gender')
ylabel('Number of Individuals')
set(gca,'xtick',[0,1],'xticklabel',{'Male';'Female'})
subplot(2,1,2);
histogram(Mat_Data(:,2))
title('Distribution of Ages')
xlabel('Ages')
ylabel('Number of Individuals')
```



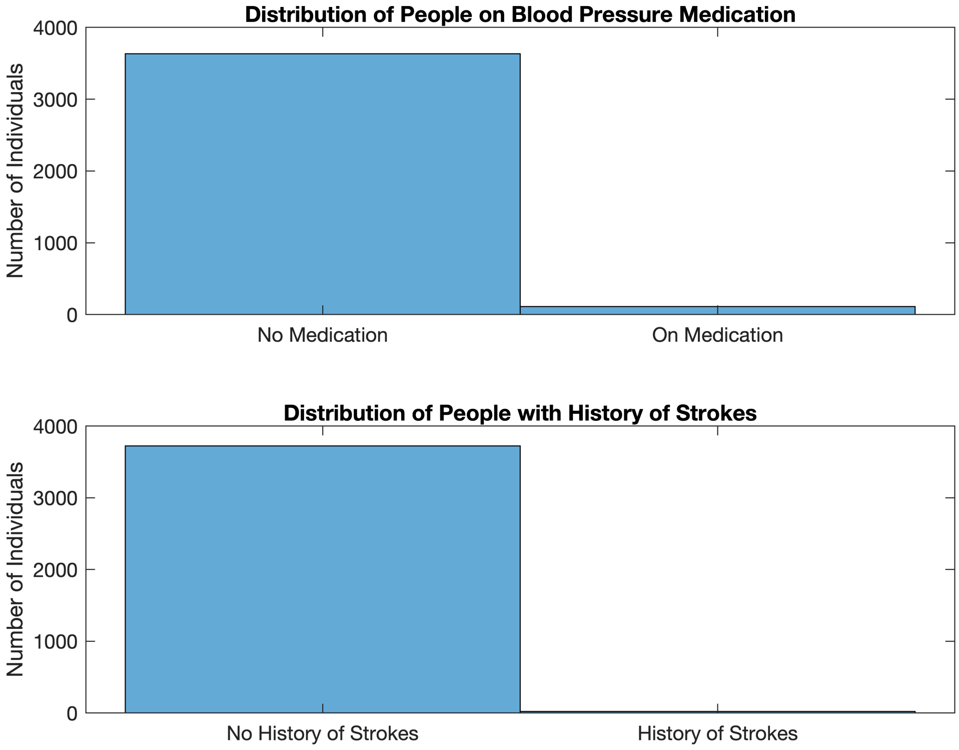
```
clf;
figure(2)
subplot(2,1,1);
histogram(Mat_Data(:,3))
title('Distribution of Smokers')
ylabel('Number of Individuals')
set(gca,'xtick',[0,1],'xticklabel',{'Non-Smokers';'Smokers'})
subplot(2,1,2);
histogram(Mat_Data(:,4))
title('Distribution of Cigarettes Smoked per Day')
xlabel('Cigarettes Smoked per Day')
```

```
ylabel('Number of Individuals')
```



```
clf;

figure(3)
subplot(2,1,1);
histogram(Mat_Data(:,5))
title('Distribution of People on Blood Pressure Medication')
ylabel('Number of Individuals')
set(gca,'xtick',[0,1],'xticklabel',{'No Medication';'On Medication'})
subplot(2,1,2);
histogram(Mat_Data(:,6))
title('Distribution of People with History of Strokes')
ylabel('Number of Individuals')
set(gca,'xtick',[0,1],'xticklabel',{'No History of Strokes';'History of Strokes'})
```

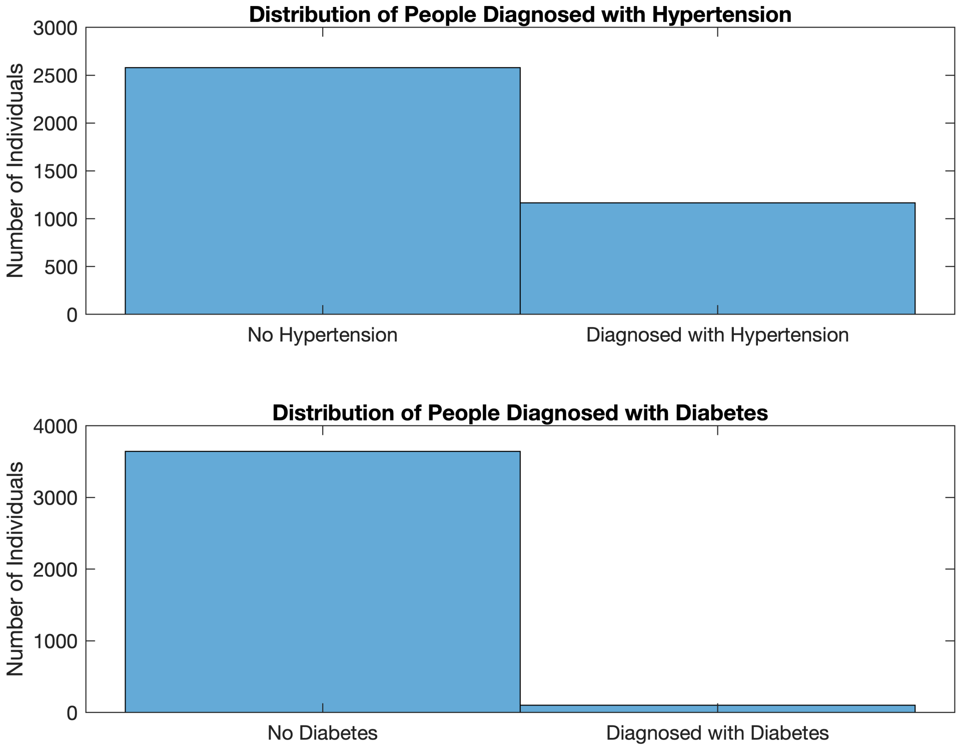


Distribution of People on Blood Pressure Medication

Distribution of People with History of Strokes

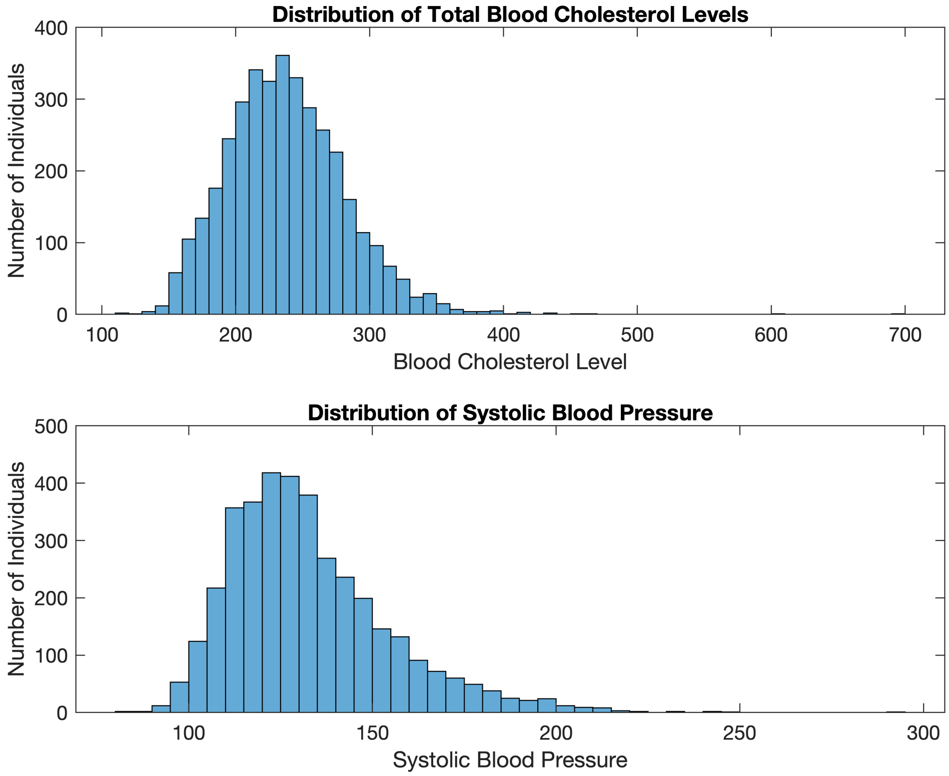
```
clf;

figure(4)
subplot(2,1,1);
histogram(Mat_Data(:,7))
title('Distribution of People Diagnosed with Hypertension')
ylabel('Number of Individuals')
set(gca,'xtick',[0,1],'xticklabel',{'No Hypertension';'Diagnosed with Hypertension'})
subplot(2,1,2);
histogram(Mat_Data(:,8))
title('Distribution of People Diagnosed with Diabetes')
ylabel('Number of Individuals')
set(gca,'xtick',[0,1],'xticklabel',{'No Diabetes';'Diagnosed with Diabetes'})
```



```
clf;

figure(5)
subplot(2,1,1);
histogram(Mat_Data(:,9))
title('Distribution of Total Blood Cholesterol Levels')
xlabel('Blood Cholesterol Level')
ylabel('Number of Individuals')
subplot(2,1,2);
histogram(Mat_Data(:,10))
title('Distribution of Systolic Blood Pressure')
xlabel('Systolic Blood Pressure')
ylabel('Number of Individuals')
```

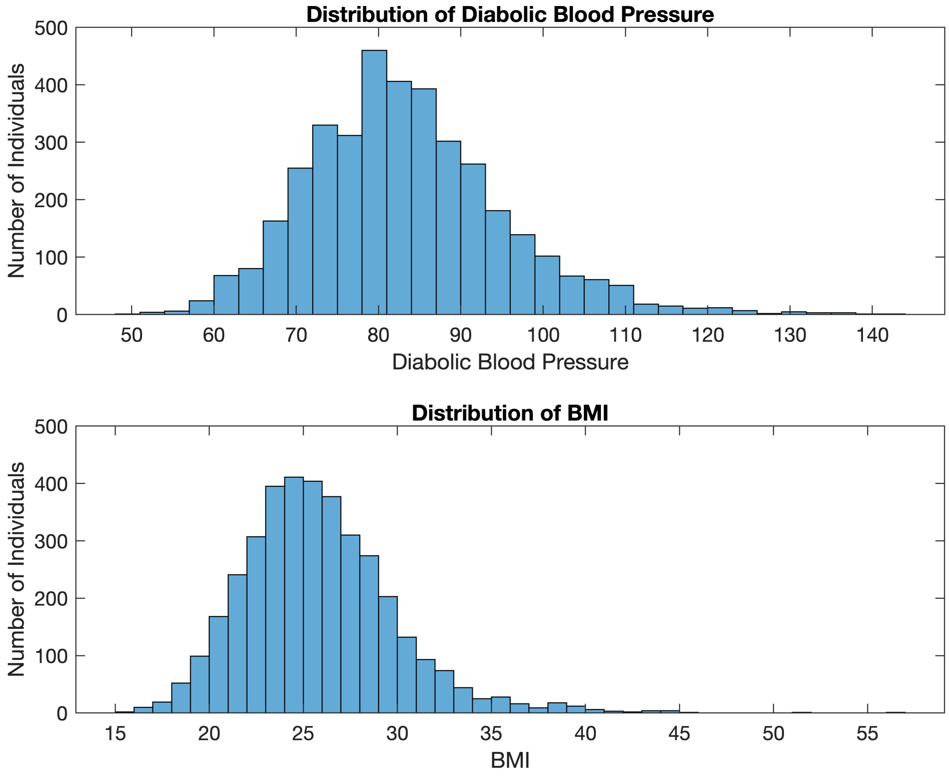


Distribution of Total Blood Cholesterol Levels

Distribution of Systolic Blood Pressure

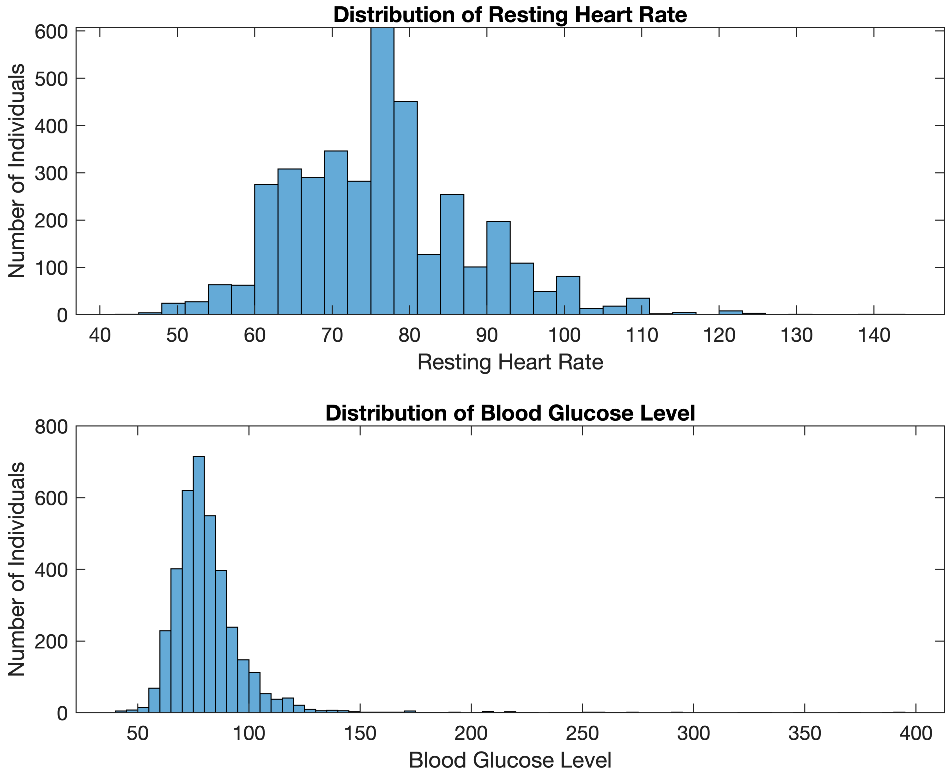
```
clf;

figure(6)
subplot(2,1,1)
histogram(Mat_Data(:,11))
title('Distribution of Diabolic Blood Pressure')
xlabel('Diabolic Blood Pressure')
ylabel('Number of Individuals')
subplot(2,1,2);
histogram(Mat_Data(:,12))
title('Distribution of BMI')
xlabel('BMI')
ylabel('Number of Individuals')
```



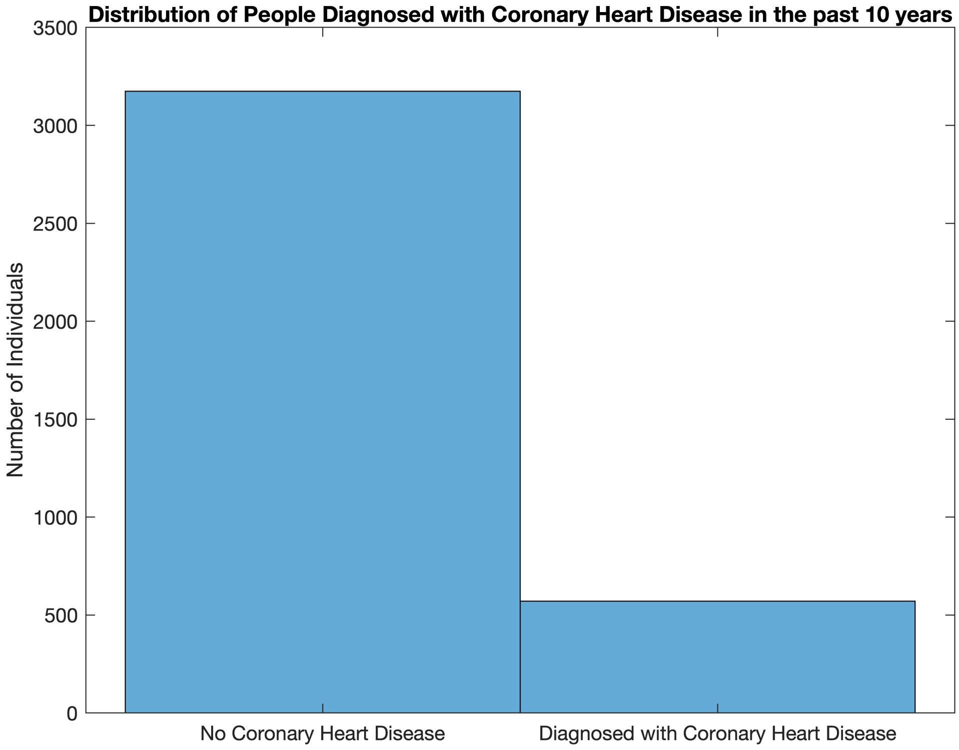
```
clf;

figure(7)
subplot(2,1,1);
histogram(Mat_Data(:,13))
title('Distribution of Resting Heart Rate')
xlabel('Resting Heart Rate')
ylabel('Number of Individuals')
subplot(2,1,2);
histogram(Mat_Data(:,14))
title('Distribution of Blood Glucose Level')
xlabel('Blood Glucose Level')
ylabel('Number of Individuals')
```



```
clf;

figure(8)
histogram(Mat_Data(:,15))
title('Distribution of People Diagnosed with Coronary Heart Disease in the past 10 year
ylabel('Number of Individuals')
set(gca,'xtick',[0,1],'xticklabel',{'No Coronary Heart Disease';'Diagnosed with Coronar
```



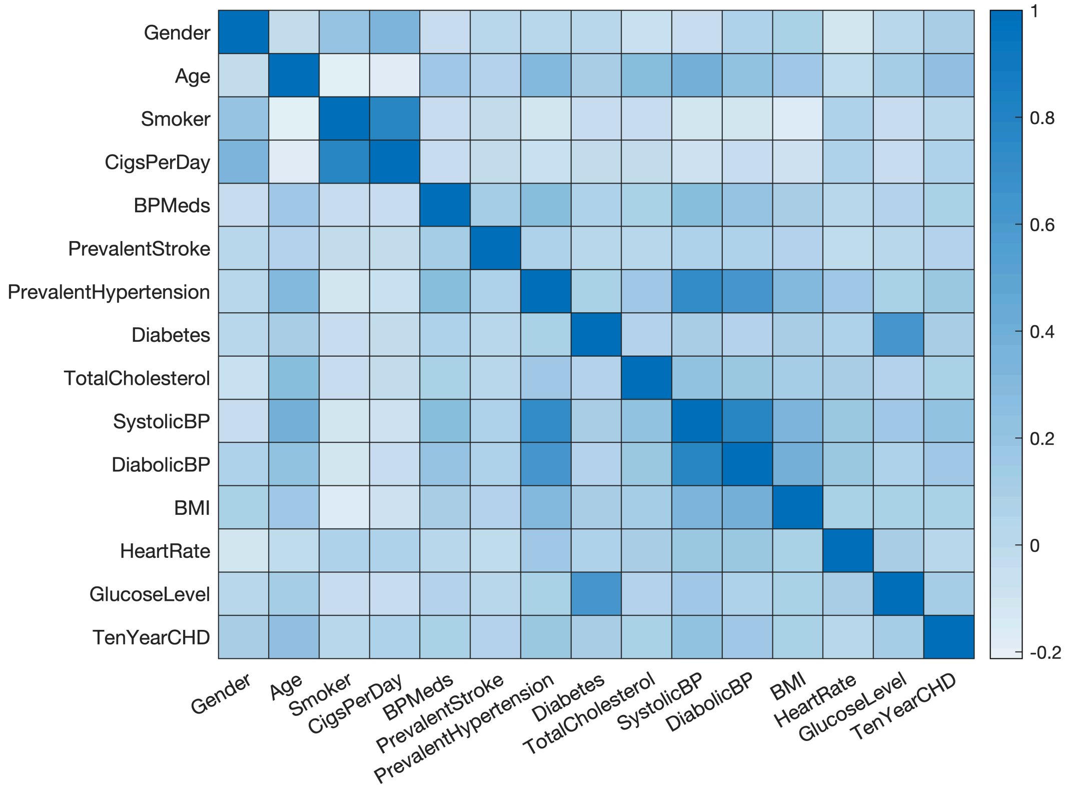
```
clf;
```

Plotting Heatmap to Visualize Correlation between all Variables

The heatmap displays the correlation between each variable that affects CHD in the table. The darker blues represent highly correlated variables that should not be paired together in the same model. These correlated features add noise and inaccuracies to the model.

```
[Mat_Data_Rows, Mat_Data_Cols] = size(Mat_Data);

Corr_Mat_Data = corr(Mat_Data);
Heatmap_Mat_Data = heatmap(Corr_Mat_Data);
Field_Names = fieldnames(Table_Data);
for k = 1:Mat_Data_Cols
    Heatmap_Mat_Data_Labels(k) = Field_Names(k);
end
Heatmap_Mat_Data.XDisplayLabels = Heatmap_Mat_Data_Labels;
Heatmap_Mat_Data.YDisplayLabels = Heatmap_Mat_Data_Labels;
```



Plotting Heatmap to Visualize Correlation between significant Variables

Using the built-in function mnrfit, we calculated the p-value for each variable in our table to determine whether to reject or accept our null hypothesis. Through a for loop, we will evaluate each variable; if the p-value is greater than 0.05, we will reject our null hypothsis and state that the variable is statistically insignificant in affecting CHD. This will allow us to reduce the noise in our data, increasing the accuracy of our model. The following heatmap shows respective correlation values for the significant variables.

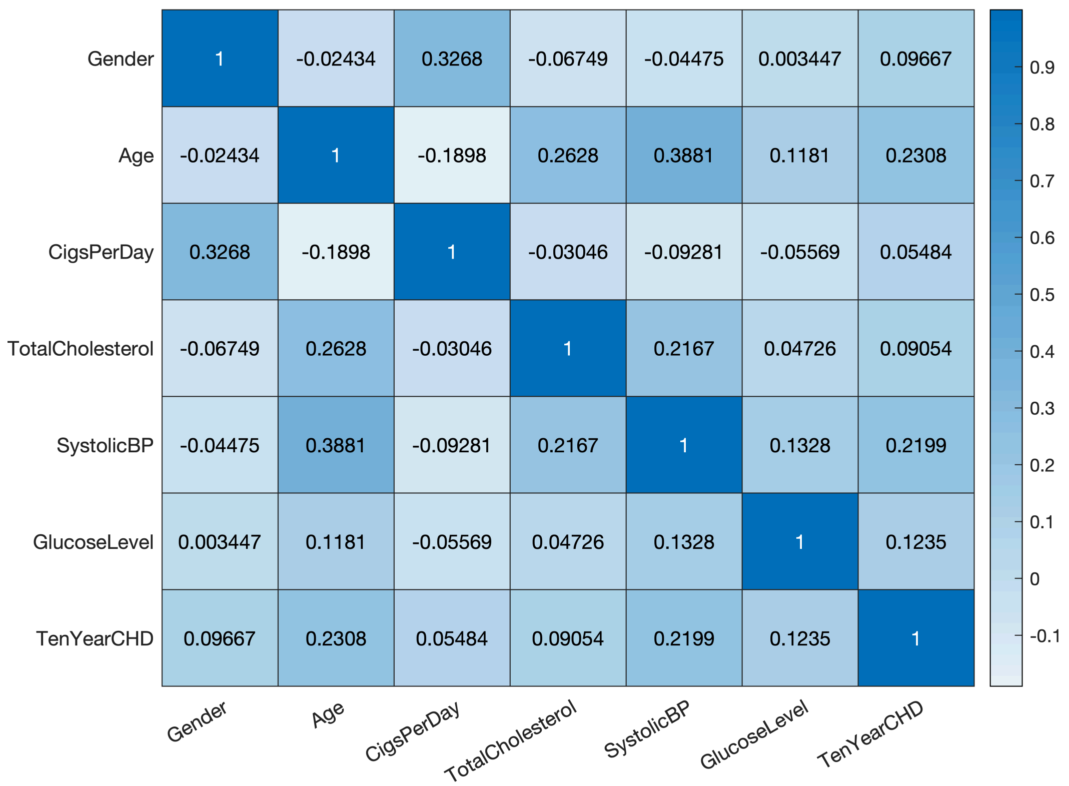
```
Y = categorical(Table_Data.TenYearCHD);
X = [Table_Data.Gender, Table_Data.Age, Table_Data.Smoker, Table_Data.CigsPerDay, Table
    Table_Data.PrevalentHypertension, Table_Data.Diabetes, Table_Data.TotalCholesterol,
   \ Table_Data.BMI, Table_Data.HeartRate, Table_Data.GlucoseLevel];
[Coeff_Values,dev,stats] = mnrfit(X,Y,'model','hierarchical');
P_Values = stats.p;

counter = 1;
for j = 1:14
    if P_Values(j+1) < 0.05
      Scrub_Data2(:,counter) = Mat_Data(:,j);
      Coeff_Values_Scrub_Data2(counter) = abs(Coeff_Values(j+1));
      counter = counter + 1;
    end
end

Scrub_Data2(:,counter) = Mat_Data(:,Mat_Data_Cols);
```

```
Final_Table_Data = array2table(Scrub_Data2,'VariableNames',{'Gender', 'Age', 'CigsPerDa
    'GlucoseLevel', 'TenYearCHD'});

Corr_Scrub_Data2 = corr(Scrub_Data2);
Heatmap_Scrub_Data2 = heatmap(Corr_Scrub_Data2);
Field_Names = fieldnames(Final_Table_Data);
for k = 1:counter
    Heatmap_Scrub_Data2_Labels(k) = Field_Names(k);
end
Heatmap_Scrub_Data2.XDisplayLabels = Heatmap_Scrub_Data2_Labels;
Heatmap_Scrub_Data2.YDisplayLabels = Heatmap_Scrub_Data2_Labels;
```



```
Heatmap_Scrub_Data2_Labels2 = Heatmap_Scrub_Data2_Labels;
Heatmap_Scrub_Data2_Labels2(counter) = [];
clf;

for r = 1:counter-1
    Percentage_Mat(r) = (exp(Coeff_Values_Scrub_Data2(r)) - 1) * 100;
end
```

Regression coefficents

Using the regression coefficient values, we created a table of percentage values for each variable. These values represent the change in the response variable, CHD, for one unit of change in a specific predictor variable i.e. Gender; the other predictor variables i.e. Age, CigsPerDay, TotalCholestrol, SystolicBP and

GlucoseLevel, are kept constant in the model. We can infer from this table that the odds for males suffering from CHD is 78.9946% higher than females, whilst for age, an increase in one year increases an individuals risk by 6.5757%. Furthermore, for every one additional cigarette smoked or one unit increase in Systolic Blood Pressure, increases an individuals risk by 1.8115% and 1.5514% respectively. However, changes in total blood cholestrol and glucose levels showed insignificant change.

```
Evaluation_Table = table(Coeff_Values_Scrub_Data2', Percentage_Mat', 'RowNames', reshap
```

```
Evaluation_Table = 6×2 table
```

	Coefficient Values	Percentage Values (%)
1 Gender	0.5822	78.9946
2 Age	0.0637	6.5757
3 CigsPerDay	0.0180	1.8115
4 TotalCholesterol	0.0023	0.2304
5 SystolicBP	0.0154	1.5514
6 GlucoseLevel	0.0075	0.7510

```
    'VariableNames', {'Coefficient Values', 'Percentage Values (%)'})
```

Machine Learning

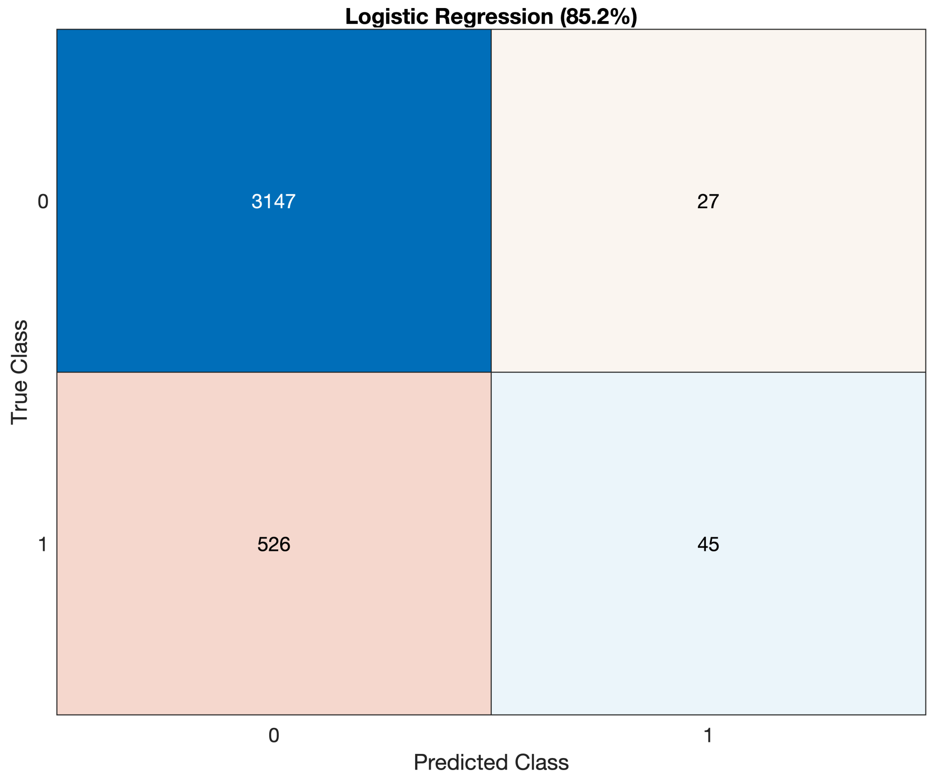
We used the classification learner tool in the machine learning toolbox. The data set was used to measure the accuracy of the model during the development procedure. We decided to choose the logistic regression algorithm to predict our variable which is categorical. After several iterations of our training set, the following models were formed:

```
[Mat_Data_Rows Mat_Data_Cols] = size(Scrub_Data2);
P = 0.90;
idx = randperm(Mat_Data_Rows);
Training_Data = Scrub_Data2(idx(1:round(P*Mat_Data_Rows)),:);
Test_Data = array2table(Scrub_Data2(idx(round(P*Mat_Data_Rows)+1:end),:),'VariableNames
    'TotalCholesterol', 'SystolicBP', 'GlucoseLevel', 'TenYearCHD'});
```

Logistic regression using all features

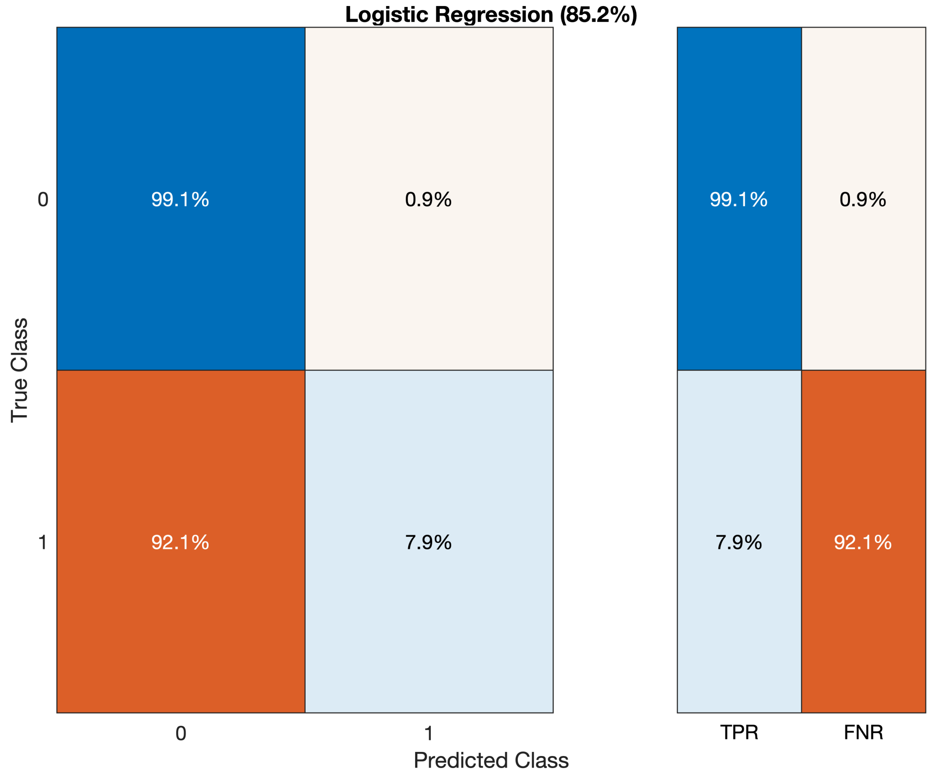
WWe obtained a confusion matrix for the number of individuals that are expected to have CHD. The confusion matrix shows that our preliminary models misclassify 92.1% of CHD cases as non-CHD cases while classifying 0.9% of non-CHD cases as CHD cases. We fail to detect more than 90% of CHD patients, a situation that can cause detrimental effects in medical practice.

```
Fig1 = openfig('LR1_Mat1.fig');
F = figure;
set(Fig1.Children,'Parent',F)
```

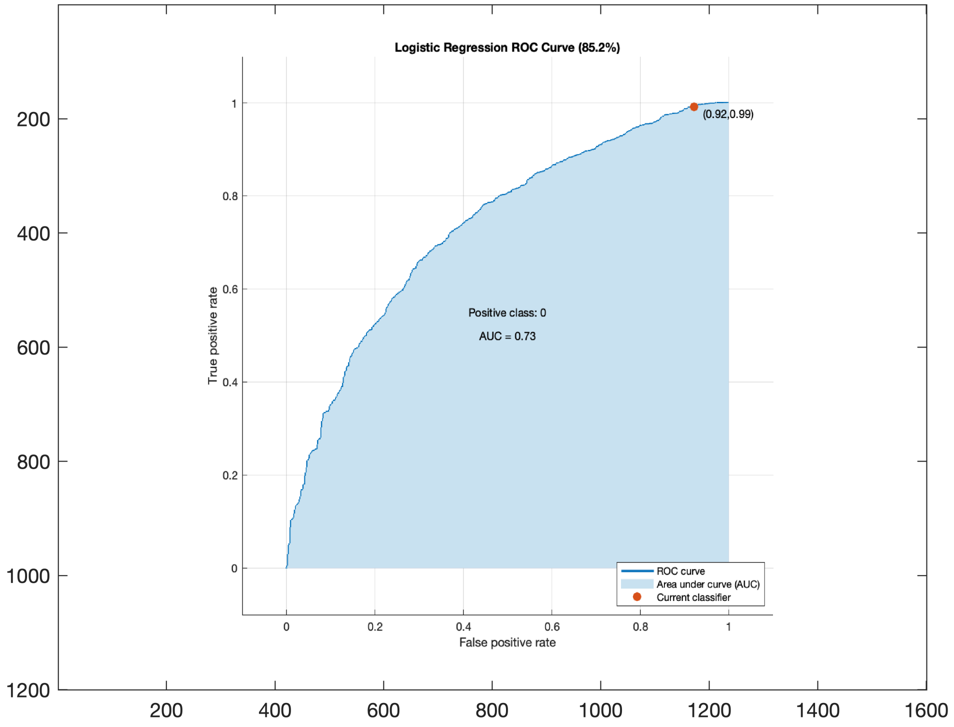


	0	1
0	3147	27
1	526	45

```
Fig2 = openfig('LR1_Mat2.fig');
F = figure;
set(Fig2.Children,'Parent',F)
```



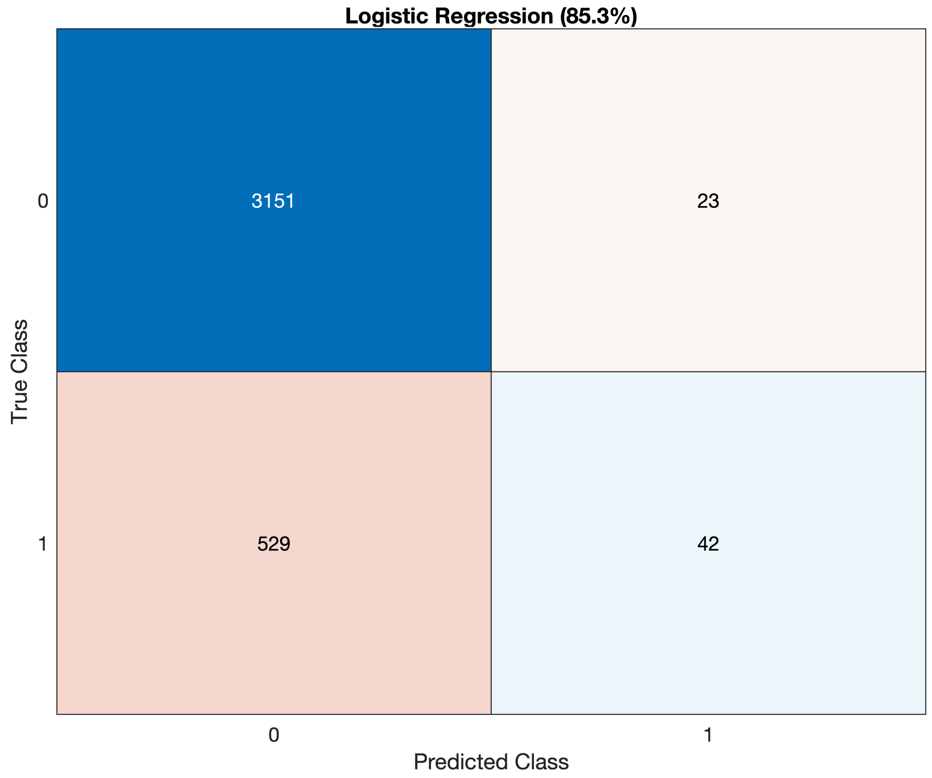
```
Image1 = imread('LR1_ROC.png');
image(Image1);
```



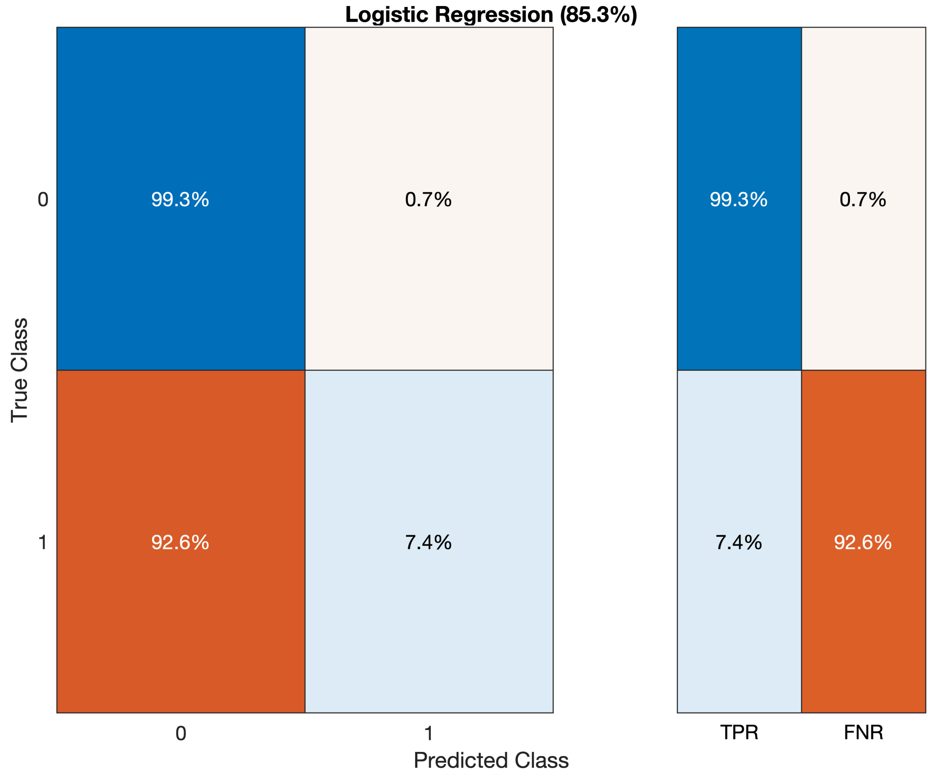
Logistic regression using selective features

This logistic regression only takes into account features that have a p-value < 0.05, supporting our null hypothesis. Through scrubbing data in the earlier sections of our code, our second model compared (85.3%) to the raw data model (85.2%) is more accurate. However, the percentage of misclassification of CHD cases as non-CHD cases is higher than the first model. The confusion matrix shows that our preliminary models misclassify 92.6% of CHD cases as non-CHD cases while classifying 0.7% of non-CHD cases as CHD cases. We still fail to detect more than 90% of CHD patients.

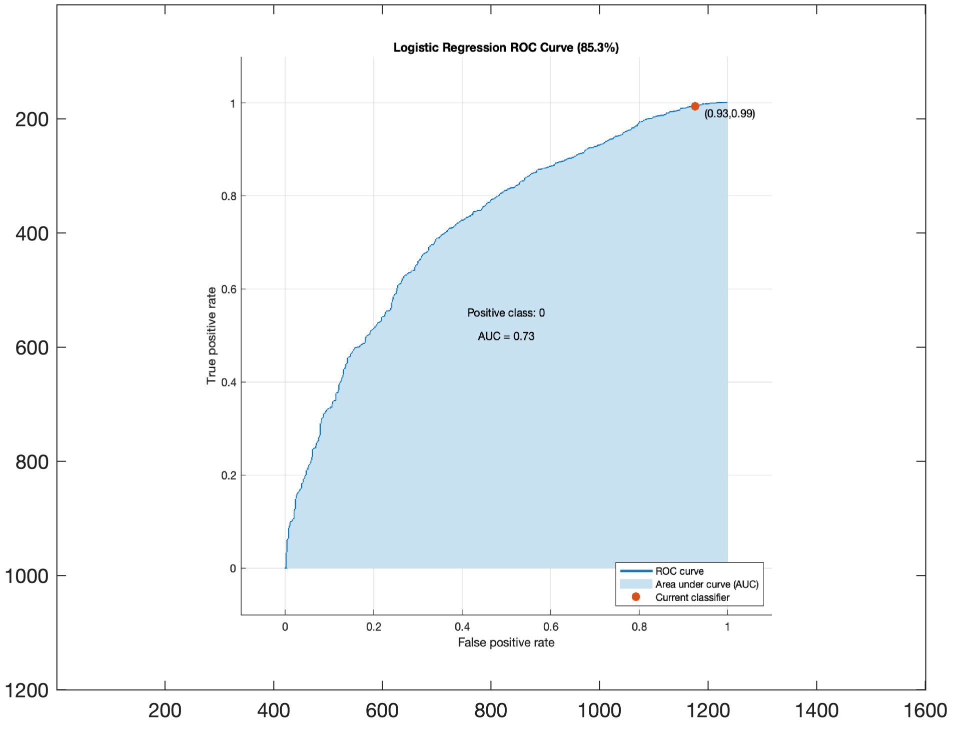
```
Fig4 = openfig('LR2_Mat1.fig');
F = figure;
set(Fig4.Children,'Parent',F)
```



```
Fig5 = openfig('LR2_Mat2.fig');
F = figure;
set(Fig5.Children,'Parent',F)
```



```
Image2 = imread('LR2_ROC.png');
image(Image2);
```



Model Testing

We used the built-in trainedModel function and the test data to test our model in this Live Script. We then found the accuracy of the model and evaluated the results produced.

```
Results_Data = trainedModel.predictFcn(Test_Data);
Correct = 0;
T1P1 = 0;
T0P0 = 0;
T1P0 = 0;
T0P1 = 0;
Length_Results_Data = length(Results_Data);
for a = 1:Length_Results_Data
    if Results_Data(a) == 0 && Test_Data.TenYearCHD(a) == 0
        T0P0 = T0P0 + 1;
        Correct = Correct + 1;
    elseif Results_Data(a) == 1 && Test_Data.TenYearCHD(a) == 0
        T0P1 = T0P1 + 1;
    elseif Results_Data(a) == 0 && Test_Data.TenYearCHD(a) == 1
        T1P0 = T1P0 + 1;
    elseif Results_Data(a) == 1 && Test_Data.TenYearCHD(a) == 1
        T1P1 = T1P1 + 1;
        Correct = Correct + 1;
    end
end

Result_Mat = [Correct, T1P1, T0P0, Length_Results_Data - Correct, T0P1, T1P0];
Percentage_Result_Mat = Result_Mat/Length_Results_Data*100;

Result_Table = table(Result_Mat', Percentage_Result_Mat', 'RowNames', {'Correct', 'Prec
```

```
Result_Table = 6×2 table
```

	Number of Predictions	Percentage Values (%)
1 Correct	309	82.6203
2 Predicted CHD	5	1.3369
3 Predicted NO CHD	304	81.2834
4 Incorrect	65	17.3797
5 False Positive	2	0.5348
6 False Negative	63	16.8449

```
    'False Positive', 'False Negative'}, 'VariableNames', {'Number of Predictions', 'Pe
```

Conclusion

After scrubbing our raw database our model accuracy was 85.2%. Referring to the ROC curve, the area under the curve (AUC) is 0.73, which is considered within the acceptable range of 0.7-0.8. However, our false-positive rate at this AUC value is 0.92 suggesting that our model lacks precision. Therefore, we can conclude that the models failed to effectively classify members of the training dataset concerning CHD. To improve our model, we could acquire a greater sample size, ensuring the difference between CHD and non-CHD causes are negligible.

However, after Model Testing, the accuracy of the model is 82.6203% which is lower than both of the previous models, the percentage of false negatives were 16.8449%. While this value is significantly lower than 92.6%, the percentage is still high. However, by iterating more variables we can improve the accuracy of this model.

Study Evaluation

Even though our model was 85.2% accurate, it can not be used in medical practice as our misclassified results of CHD cases for both models were higher than 90%. This was perhaps due to our sample size for CHD cases being drastically smaller than the sample size for non-CHD cases. Comparatively, having a smaller sample size for the CHD cases decreases the reliability of our results. This has skewed our model. Furthermore, we could have used an alternative approach to scrubbing our data instead of using p-values to assure that those variables were statistically insignificant. Moreover, there are more attributes then mentioned in the dataset that affect an individual’s risk of CHD, limiting our model.

Citations

Björn Dahlöf, (2010). Cardiovascular Disease Risk Factors: Epidemiology and Risk Assessment, The American Journal of Cardiology, Volume 105, Issue 1, Supplement, Pages 3A-9A, ISSN 0002-9149, https://doi.org/10.1016/j.amjcard.2009.10.007.

"Framingham Heart Study". *Framinghamheartstudy.Org*, 2020, https://framinghamheartstudy.org/.

Mathur, P., Srivastava, S., Xu, X., & Mehta, J. L. (2020). Artificial Intelligence, Machine Learning, and Cardiovascular Disease. *Clinical Medicine Insights. Cardiology*, *14*, 1179546820927404. https://doi.org/10.1177/1179546820927404