Data Science in Health Project Documentation Report

Overview

This data science project aims to analyze health data collected from individuals to predict the likelihood of developing coronary heart disease (CHD) within ten years. Insights produced in this project could save countless lives and provide tremendous benefit in the medical space. The project utilizes logistic regression to build a predictive model based on various health indicators.

This was not the initial project I originally intended to do; it was a mobile health dataset from body sensors containing over ten thousand data points. Because of my computer, it took forever to run and it became evident that I had to switch plans. During that time, I thought about a disease that runs in my family and decided to use it as inspiration for this project.

Literature Review of Previous Works

Nishat, M., Ahmed, S., Hasan, M. M., Ali, M. H., Saha, R., & Mahmud, M. (2021). Performance Evaluation and Comparative Analysis of Different Machine Learning Algorithms in Predicting Cardiovascular Disease.

The previous study utilized various machine learning methods to predict CHD. Utilizing data from the University of California, Irvine repository, twelve algorithms were assessed using default hyperparameters, grid search cross-validation, and random search cross-validation methods. Both accuracy and computational time were measured, with hard and soft voting ensemble classifiers achieving 92% accuracy. Adaboost algorithm demonstrated superior precision and specificity compared to ensemble classifiers. The analysis extensively compares algorithm performance across multiple metrics including accuracy, precision, sensitivity, specificity, F1 score, and ROC-AUC.

Even though there were many models that intrigued me, I decided to personally use logistic regression because of its simplicity, efficiency, and clinical acceptance. Attempts of data analysis on the dataset posted on kaggle showed the following:

- Men are more likely to get heart disease than women. As people get older, smoke more cigarettes, or have higher blood pressure, their chances of getting heart disease also go up.

- Having higher total cholesterol doesn't seem to make much difference in the chance of getting heart disease. This might be because the cholesterol test includes both good and bad cholesterol. Glucose levels also don't have a big impact on the chance of getting heart disease, only a tiny bit.
- The model we used predicted heart disease correctly 88% of the time. It's better at saying who doesn't have heart disease than who does.

Methodology

I will be experimenting with min-max normalization and using the top absolute valued correlated variables associated with the variable 'TenYearCHD' which describes whether a subject has cardiovascular disease or not. The top correlated values were selected based on their absolute values.The objective of the project is to play around with the data and double check the claims mentioned in previous analysis.

Preliminaries

Loading Dataset and Packages

The project utilizes several R packages for data manipulation, visualization, and model training. Key packages include `caret`, `pROC`, `ggplot2`, and `dplyr`. The dataset is loaded using the `read.csv()` function from the `foreign` package. R studio was used as the main code editor to compile the project and its resources.

Data Cleaning and Preparation

Handling Missing Values

Missing values in the dataset are removed using the `na.omit()` function to ensure data integrity and consistency. Without removing the null values, the dataset's dimensions would have been problematic in the analysis and training phase.

Descriptive Statistics

Descriptive statistics such as mean, median, minimum, maximum, standard deviation, and quartiles are calculated for numeric variables in the dataset to gain insights into the distribution of health indicators. A function was created to compare male and female subjects as well as the average statistics of subjects with CHD and without CHD.

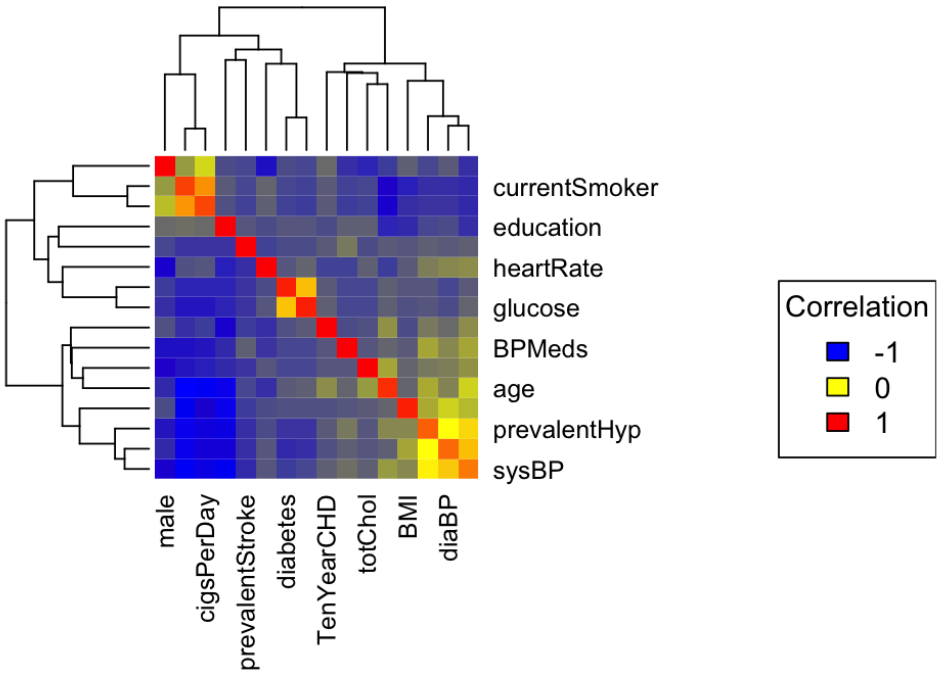
	Feature <chr>	CHD_Mean <dbl>	No_CHD_Mean <dbl>	Difference <dbl>
age	age	54.278276	48.708938	5.5693381
education	education	1.827648	2.007099	-0.1794509
cigsPerDay	cigsPerDay	10.488330	8.758632	1.7296985
totChol	totChol	246.350090	235.169732	11.1803576
sysBP	sysBP	143.981149	130.280736	13.7004133
diaBP	diaBP	87.157989	82.148919	5.0090702
BMI	BMI	26.569838	25.642975	0.9268633
heartRate	heartRate	76.310592	75.626331	0.6842614
glucose	glucose	88.732496	80.620200	8.1122954

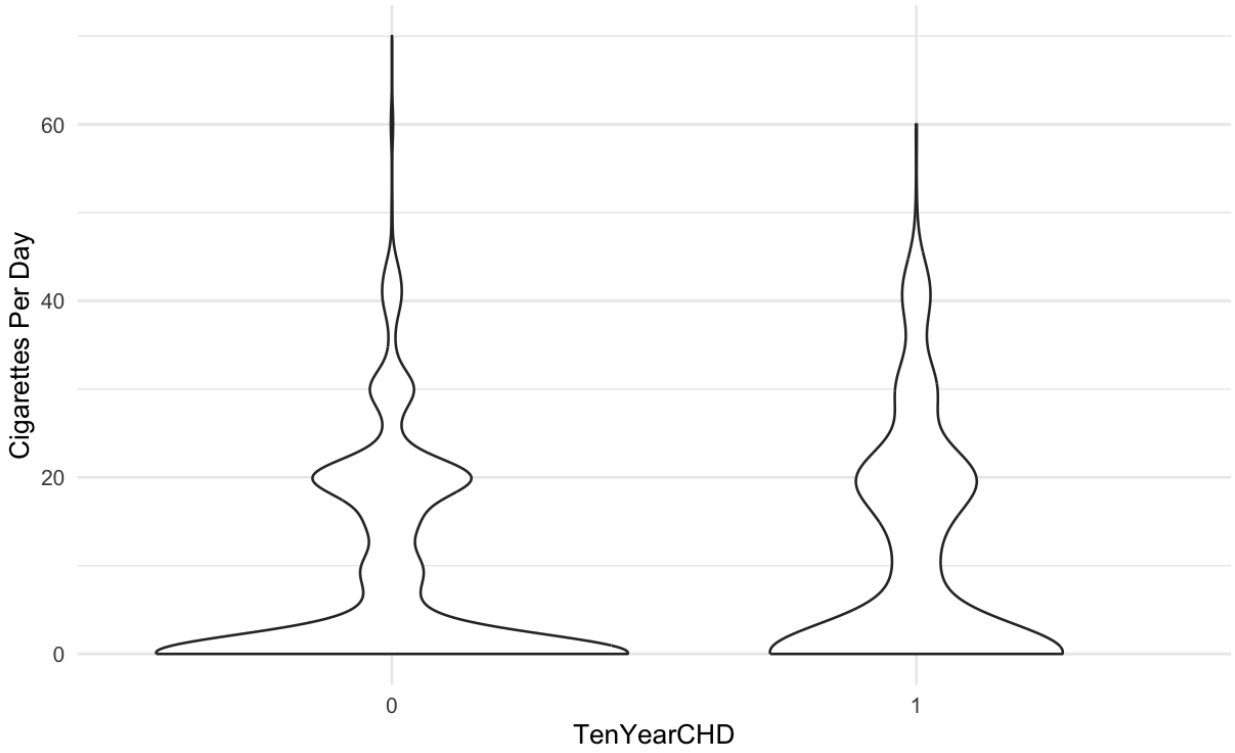
9 rows

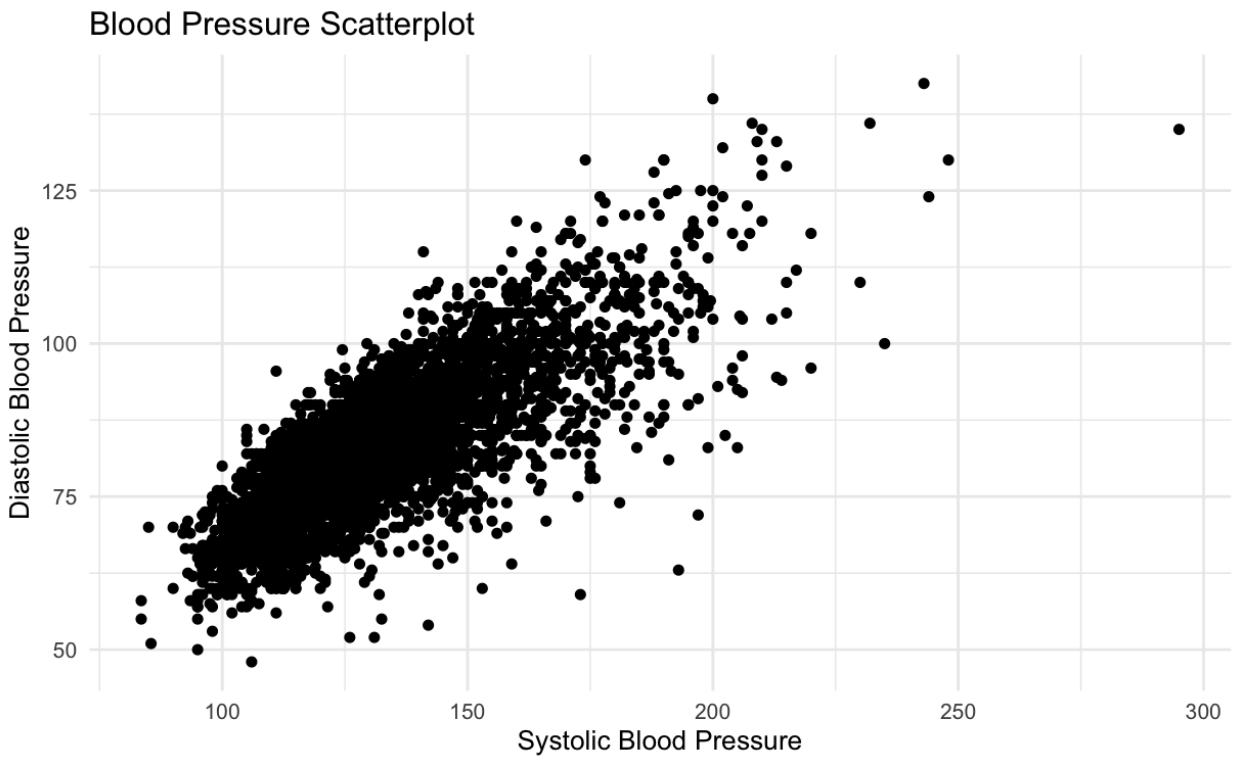
I had problems with this function because I had misplaced the variables ‘clean_data’ and ‘data’ causing the difference not to be shown.

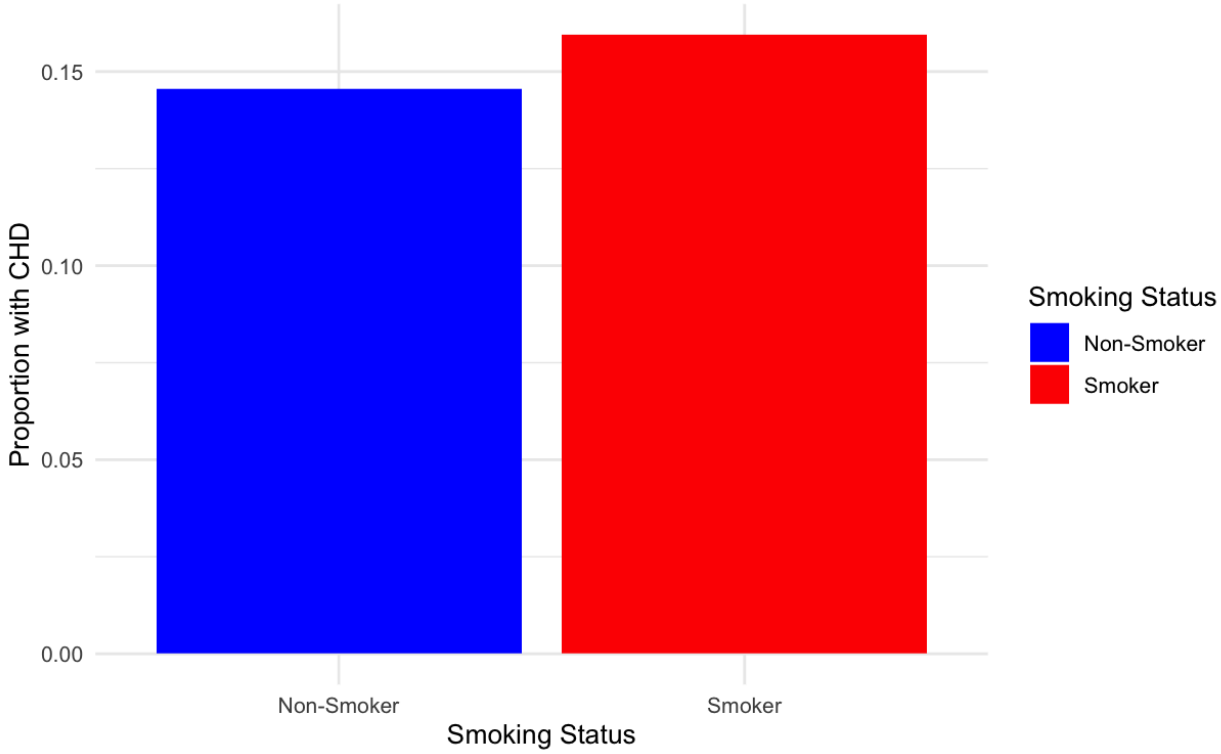
Data Visualization

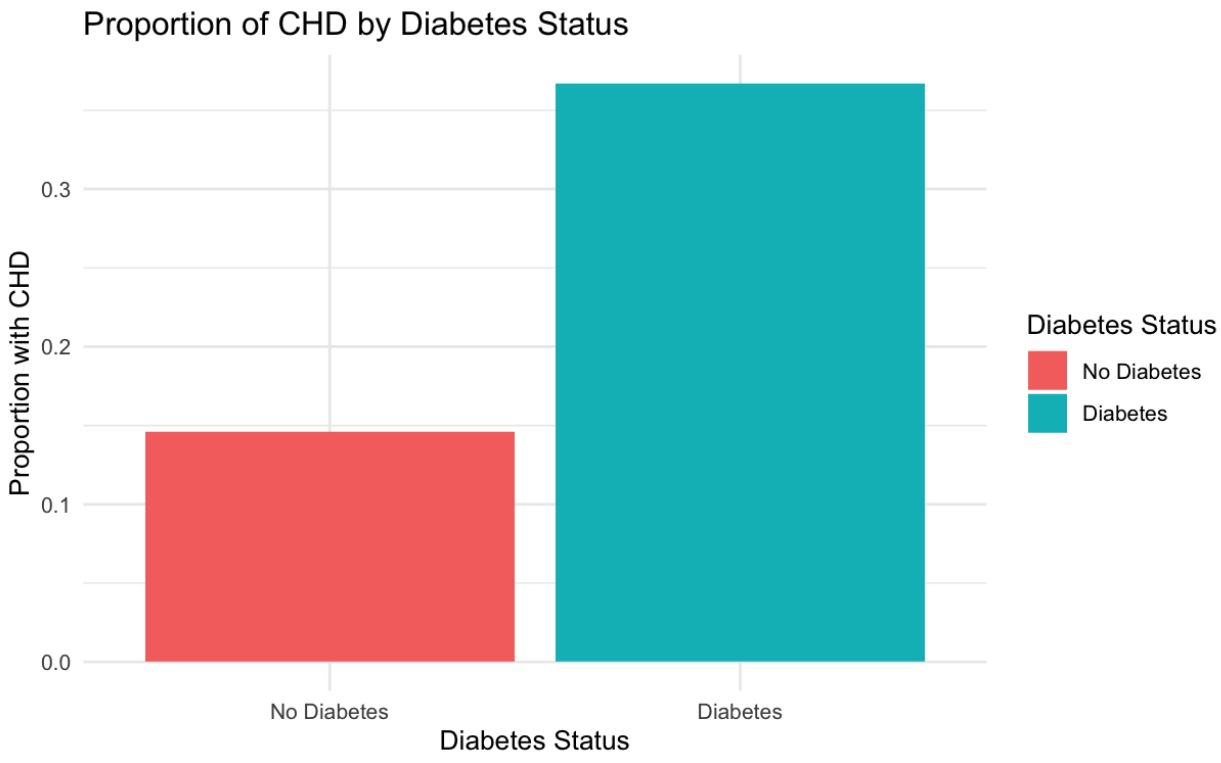
The project includes various data visualizations to explore relationships between different health variables and their impact on the likelihood of developing CHD. Visualizations include scatter plots, bar charts, and violin plots.

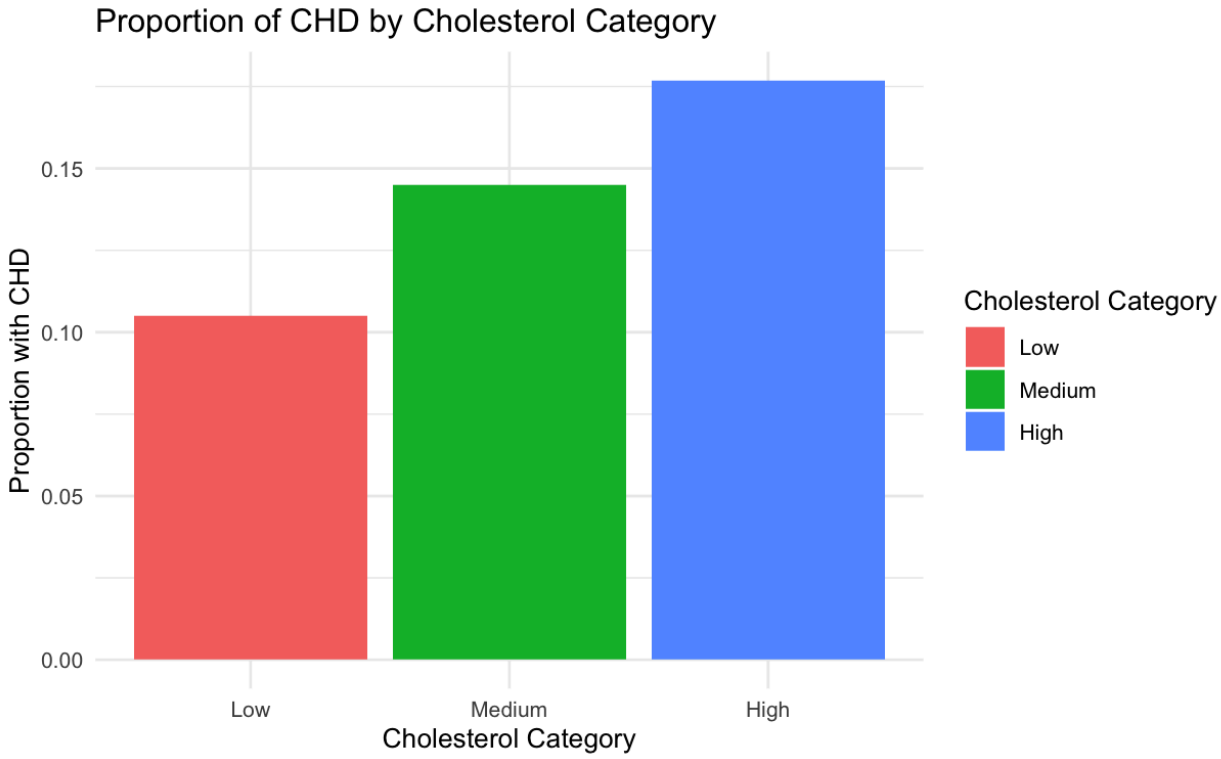

currentSmoker
education
heartRate
glucose
BPMeds
age
prevalentHyp
sysBP
male
cigsPerDay
prevalentStroke
diabetes
TenYearCHD
totChol
BMI
diaBP
Correlation
-1
0
1

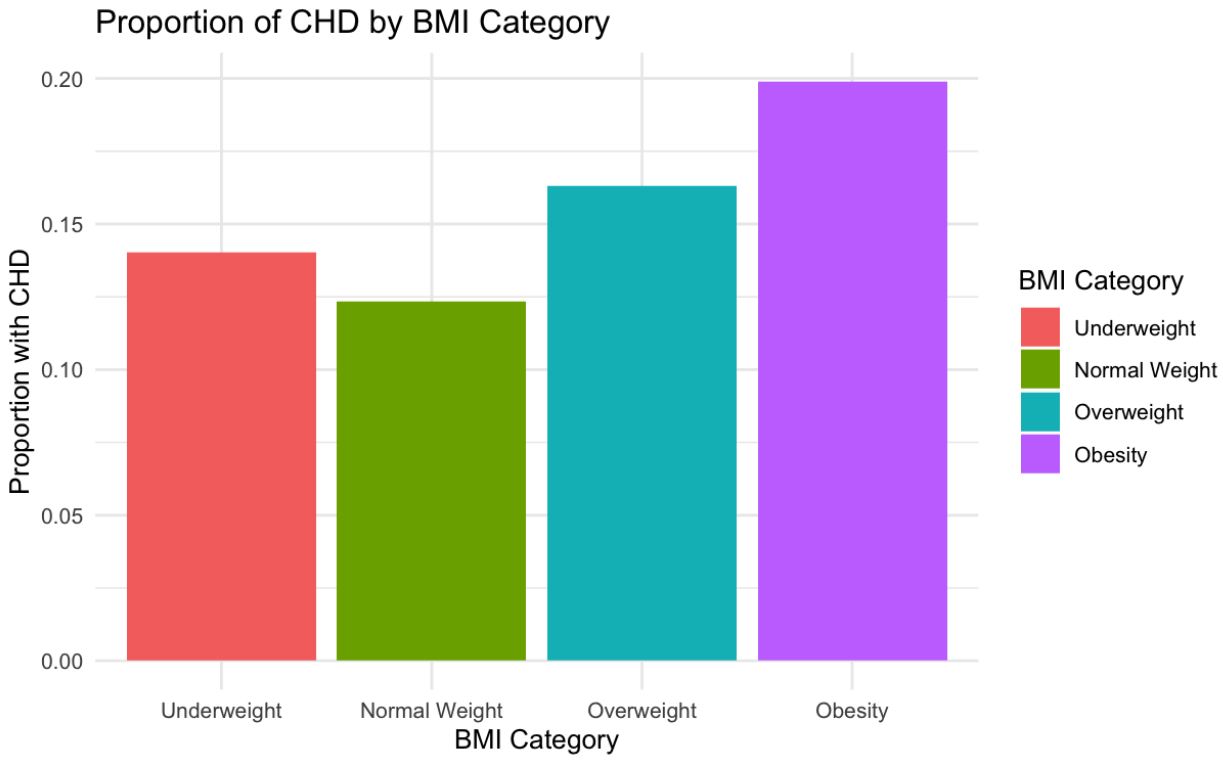

Cigarettes Per Day
60
40
20
0
0
1
TenYearCHD

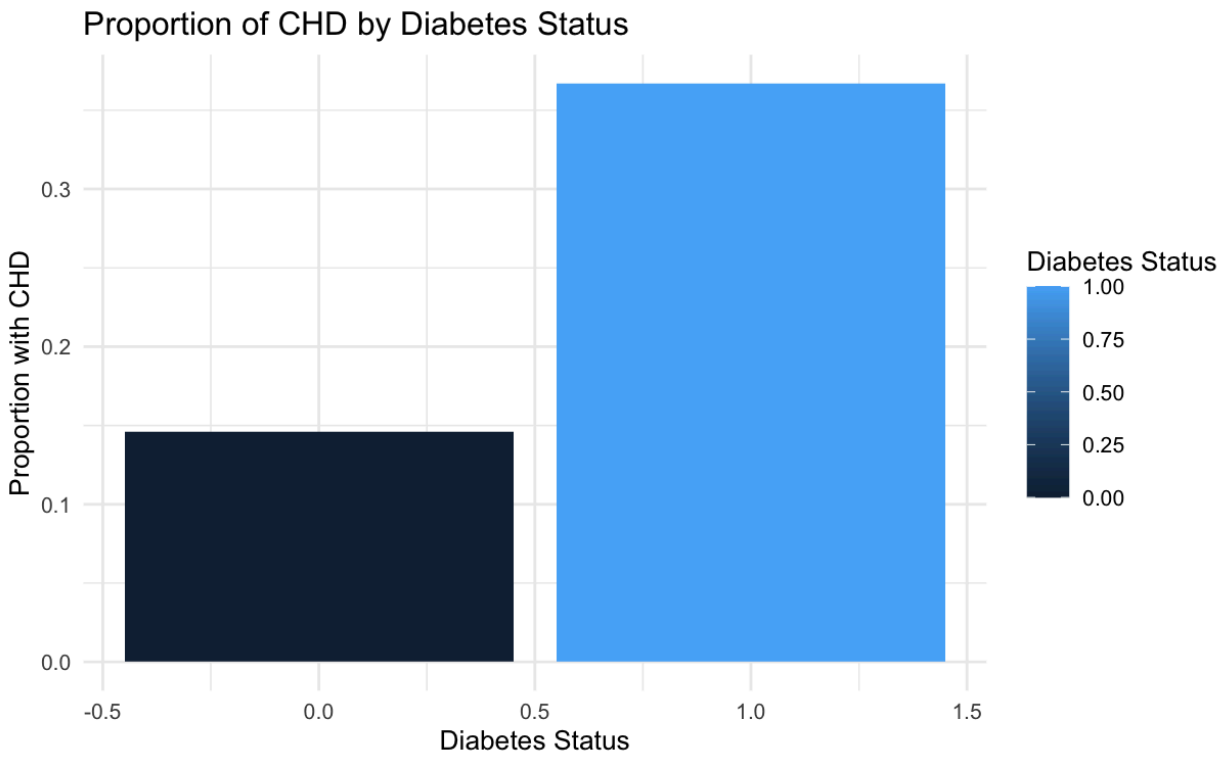

Blood Pressure Scatterplot
Diastolic Blood Pressure
125
100
75
50
100
150
200
250
300
Systolic Blood Pressure

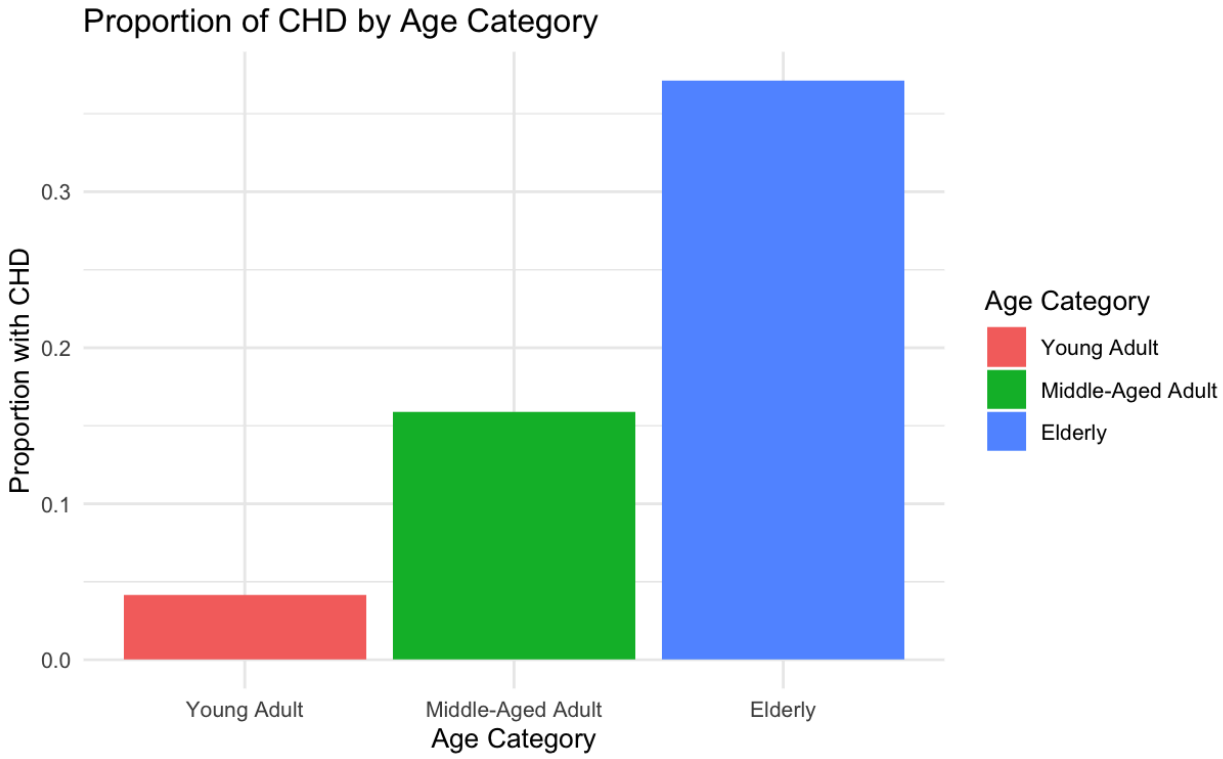

Proportion with CHD
0.15
0.10
0.05
0.00
Non-Smoker
Smoker
Smoking Status
Smoking Status
Non-Smoker
Smoker


Proportion of CHD by Diabetes Status
Proportion with CHD
0.0
0.1
0.2
0.3
No Diabetes
Diabetes
Diabetes Status
Diabetes Status
No Diabetes
Diabetes


Proportion of CHD by Cholesterol Category
Proportion with CHD
0.00
0.05
0.10
0.15
Low
Medium
High
Cholesterol Category
Cholesterol Category
Low
Medium
High


Proportion of CHD by BMI Category
0.20
0.15
0.10
0.05
0.00
Proportion with CHD
Underweight
Normal Weight
Overweight
Obesity
BMI Category
BMI Category
Underweight
Normal Weight
Overweight
Obesity


Proportion of CHD by Diabetes Status
Proportion with CHD
0.0
0.1
0.2
0.3
-0.5
0.0
0.5
1.0
1.5
Diabetes Status
Diabetes Status
1.00
0.75
0.50
0.25
0.00


Proportion of CHD by Age Category
Proportion with CHD
0.0
0.1
0.2
0.3
Young Adult
Middle-Aged Adult
Elderly
Age Category
Age Category
Young Adult
Middle-Aged Adult
Elderly

Data Normalization

Numeric variables in the dataset are normalized using min-max scaling to ensure uniformity and prevent any single variable from dominating the model due to differences in scale. Min-max normalization was used to ensure ranges of zero to one since ROC utilizes probability.

Model Training

Logistic Regression

Logistic regression is employed as the primary machine learning model for predicting the probability of developing CHD. The `glm()` function is used to train the logistic regression model, and evaluation metrics such as accuracy, precision, recall, specificity, F1 score, and Matthews correlation coefficient (MCC) are computed to assess model performance.

I had a tremendously difficult time with a different library during the training process with cross validation, so I decided to take a step back and use the basic `glm()` function instead. It was an error due to multiplications of incorrect object dimensions.

```
Error in dimnames(out) <- *vtmp* : length of 'dimnames' [2] not equal to array extent
```

The formula that was used for the logistic regression were the highly correlated values with the variable TenYearCHD.

$$TenYearCHD \sim male + age + sysBP + prevalentHyp + diaBP + glucose + diabetes"$$

Evaluation Metrics

Model performance is evaluated using various metrics, including accuracy, precision, recall, specificity, F1 score, and MCC. These metrics provide insights into the predictive power and reliability of the logistic regression model in identifying individuals at risk of developing CHD.

Accuracy: 0.84

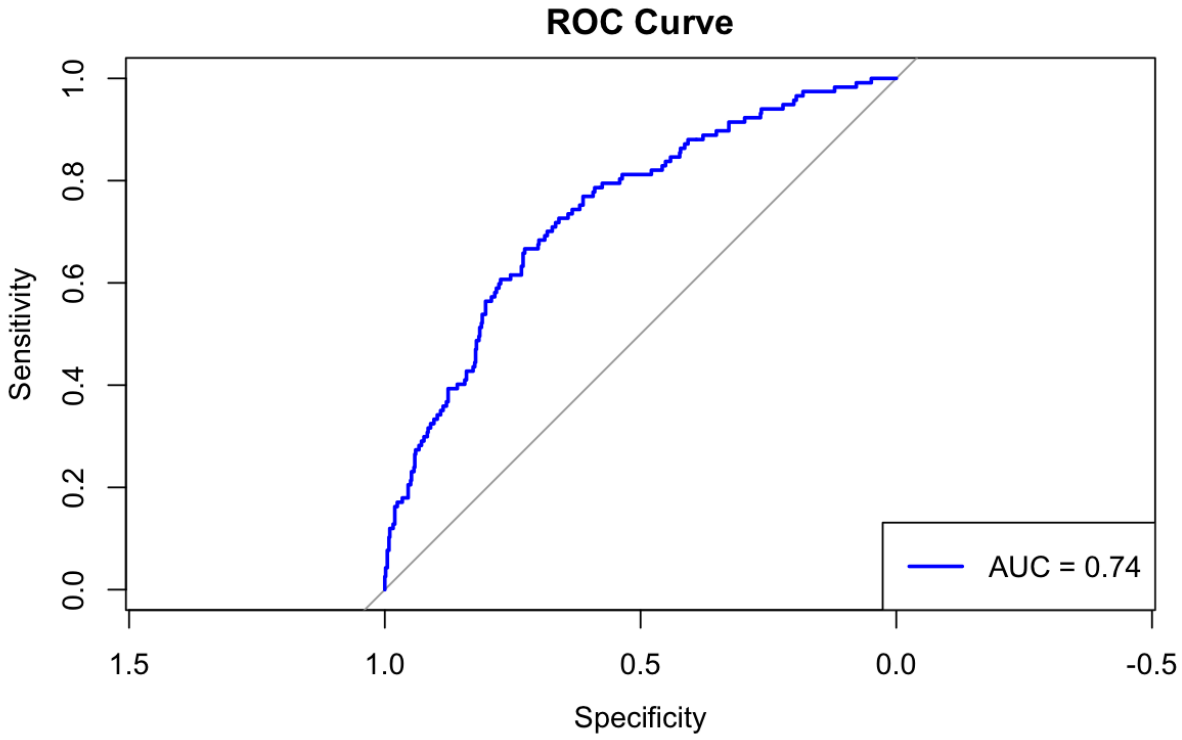
Precision: 0.85

Recall: 0.99

Specificity: 0.05

F1 Score: 0.91

Matthews Correlation Coefficient: 0.15



```r
summary(model)
```

```
Call:
glm(formula = formula, family = binomial, data = train_data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -4.1131     0.2542 -16.180  < 2e-16 ***
male           0.6544     0.1125   5.815 6.07e-09 ***
age            2.2340     0.2679   8.338  < 2e-16 ***
sysBP          3.3312     0.8811   3.781 0.000156 ***
prevalentHyp   0.3302     0.1542   2.142 0.032229 *
diaBP         -0.6968     0.6577  -1.059 0.289396
glucose        1.7789     0.8747   2.034 0.041986 *
diabetes       0.2024     0.3608   0.561 0.574903
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2477.2  on 2924  degrees of freedom
Residual deviance: 2220.8  on 2917  degrees of freedom
AIC: 2236.8

Number of Fisher Scoring iterations: 5
```

References

https://www.datacamp.com/tutorial/logistic-regression-R
https://georgrspinner.github.io/amds/
http://www.who.int/mediacentre/factsheets/fs317/en/
https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset/data
https://www.researchgate.net/profile/Mirza-Nishat/publication/351706115_Performance_Evaluation_and_Comparative_Analysis_of_Different_Machine_Learning_Algorithms_in_Predicting_Cardiovascular_Disease/links/60a54974a6fdcc3f30c9eed1/Performance-Evaluation-and-Comparative-Analysis-of-Different-Machine-Learning-Algorithms-in-Predicting-Cardiovascular-Disease.pdf