PREDICTING HEART DISEASE

An Exploratory Data Analysis

Karthik Badiganti
kbadigan@kent.edu

TABLE OF CONTENTS

1. PROBLEM STATEMENT

This study aims to create a solid machine-learning model that properly predicts the chance of heart disease using a subset of 14 crucial characteristics from the Cleveland database area of the heart disease dataset at the UCI Machine Learning Repository. The objective is to use these characteristics to create a predictive model to help with the early detection and prognosis of various heart diseases. The multidimensional nature of heart illness, which includes a variety of conditions emerging from several root causes such as coronary artery disease, heart rhythm problems, structural heart abnormalities, and heart failure, is the focus of this study.

Through this effort, the aim is to examine the complex relationships in the dataset and find patterns that can greatly improve the accuracy of heart disease prediction. Comprehensive exploratory data analysis (EDA) is used in the project to glean pertinent insights from the data, followed by applying cutting-edge machine learning techniques. The project's results could improve medical practices by giving medical personnel a trustworthy tool to assess a patient's risk of developing heart disease and make educated decisions regarding prevention, early intervention, and treatment options.

Additionally, this study seeks to advance knowledge of heart disease and its underlying causes in addition to aiming to attain high prediction accuracy. The initiative aligns with the larger goal of advancing healthcare technology and enhancing patient outcomes in cardiovascular health by tackling this significant medical challenge utilizing contemporary data analysis and machine learning approaches.

2. DATASET OVERVIEW

The dataset overview relates to the heart disease dataset, which can be accessed by clicking on the following link: Heart Disease Dataset and is taken from the Cleveland database part of the UCI Machine Learning Repository. The dataset, which consists of 76 variables and has been carefully selected to cover a range of heart health-related factors, has been created to aid heart disease prediction and medical research.

The dataset seeks to address this complexity in the context of heart disease, which has several underlying causes. From the original compilation, a particular subset of 14 important attributes has been selected for the project's scope. These characteristics, which include elements like age, sex, blood pressure, cholesterol levels, and the existence of symptoms, aid in a more complex understanding.

The dataset includes binary classification labels indicating whether cardiac disease is present. The ensuing steps, including exploratory data analysis, data preprocessing, feature engineering, and the use of machine learning algorithms, are made possible by this fundamental knowledge. These actions add up to creating a skilled predictive model that can recognize people at risk for heart disease.

2.1 Dataset Description

- Age
- Sex (1 = Male, 0 = Female)
- Chest_Pain_Type: (0 - Asymptomatic, 1 - Atypical angina, 2 - Non-typical Angina, 3 - Typical angina)
- Resting Blood Pressure: blood pressure at rest
- Cholesterol: Cholesterol in the body
- Fasting_Blood_Sugar: Blood Sugar Levels while fasting
- Resting_ECG: (0 - Normal, 1 - Wave abnormality, 2 - Hypertrophy)
- Maximum_Heart_Rate - Maximum heart rate during stress
- Exercise_Angina - angina from exercise
- Old_Peak- oldpeakST depression induced by exercise relative to rest.
- Slope_HR – the slope of the peak ST (2 - Ascending, 1 - flat, 3 - descending)
- No_MV - number of major blood vessels colored by fluoroscopy.
- Thallium - thallium stress (1 - fixed defect, 2 - normal, 3 - reversible defect)
- Heart_Disease_Indicator- (1 - No, 0 - Yes)

	Age <int>	Sex <int>	Chest_Pain_Type <int>	Resting_Blood_Pressure <int>	Cholesterol <int>	Fasting_Blood_Sugar <int>	Resting_ECG <int>	Maximum_Heart_Rate <int>	Exercise_Angina <int>	Old_peak <dbl>
1	63	1	3	145	233	1	0	150	0	2.3
2	37	1	2	130	250	0	1	187	0	3.5
3	41	0	1	130	204	0	0	172	0	1.4
4	56	1	1	120	236	0	1	178	0	0.8
5	57	0	0	120	354	0	1	163	1	0.6
6	57	1	0	140	192	0	1	148	0	0.4

Figure 1

3. DATA PREPARATION

3.1 Checking NA Values

The data is clean and there are no null values in the dataset. It can be seen from the below snippet of figure.

```
colSums(is.na(heart_disease_data))
```

```
                    Age                     Sex         Chest_Pain_Type
                      0                       0                       0
 Resting_Blood_Pressure             Cholesterol     Fasting_Blood_Sugar
                      0                       0                       0
            Resting_ECG      Maximum_Heart_Rate         Exercise_Angina
                      0                       0                       0
              Old_peak                Slope_HR                   No_MV
                      0                       0                       0
              Thallium Heart_Disease_Indicator
                      0                       0
```

Figure 2

3.2 Transforming Values

The data is mostly categorical variables and with few numerical variables. For better visualization of the variables, a few transformations are done as it would be helpful to visualize the variables more effectively. Below are the changes made to the dataset and the transformed data is shown in figure 4.

- Transforming Numerical Variable: Fasting Blood Sugar to categorical variable.
- Giving names to the existing categories in each column based on the dataset descriptions.

```
Rows: 303
Columns: 14
$ Age                    <int> 63, 37, 41, 56, 57, 57, 56, 44, 52, 57, 54…
$ Sex                    <fct> FEMALE, FEMALE, FEMALE, FEMALE, FEMALE, FE…
$ Chest_Pain_Type        <fct> Typical Angina, Typical Angina, Typical An…
$ Resting_Blood_Pressure <int> 145, 130, 130, 120, 120, 140, 140, 120, 17…
$ Cholesterol            <int> 233, 250, 204, 236, 354, 192, 294, 263, 19…
$ Fasting_Blood_Sugar    <fct> <=120, <=120, <=120, <=120, <=120, <=120, …
$ Resting_ECG            <fct> Hypertrophy, Hypertrophy, Hypertrophy, Hyp…
$ Maximum_Heart_Rate     <int> 150, 187, 172, 178, 163, 148, 153, 173, 16…
$ Exercise_Angina        <fct> No, No, No, No, No, No, No, No, No, No, No…
$ Old_peak               <dbl> 2.3, 3.5, 1.4, 0.8, 0.6, 0.4, 1.3, 0.0, 0.…
$ Slope_HR               <fct> downsloping, downsloping, upsloping, upslo…
$ No_MV                  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
$ Thallium               <fct> fixed defect, normal, normal, normal, norm…
$ Heart Disease Indicator <fct> Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, Ye…
```

Figure 3

The below graph shows the frequencies of all the categorical variables present in the data set. It can be observed that chest_pain_type and thallium columns have more than 3 categories in the dataset.

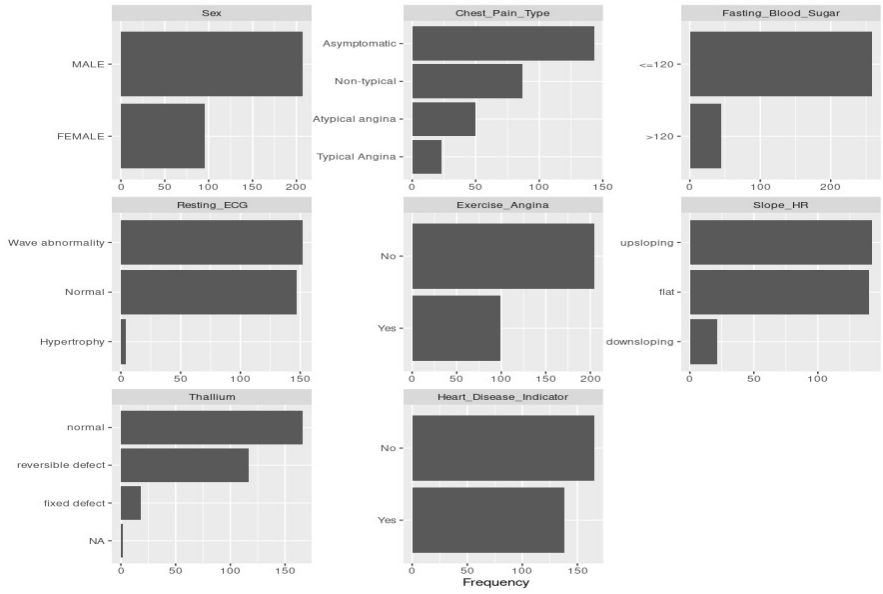


Figure 4

4. EXPLORATORY DATA ANALYSIS

Utilizing a prominent density plot tool made it possible to create a grid of plots for a more in-depth analysis as it allowed for the examination of variables and their effects on the outcome (Heart_Disease_Indicator).

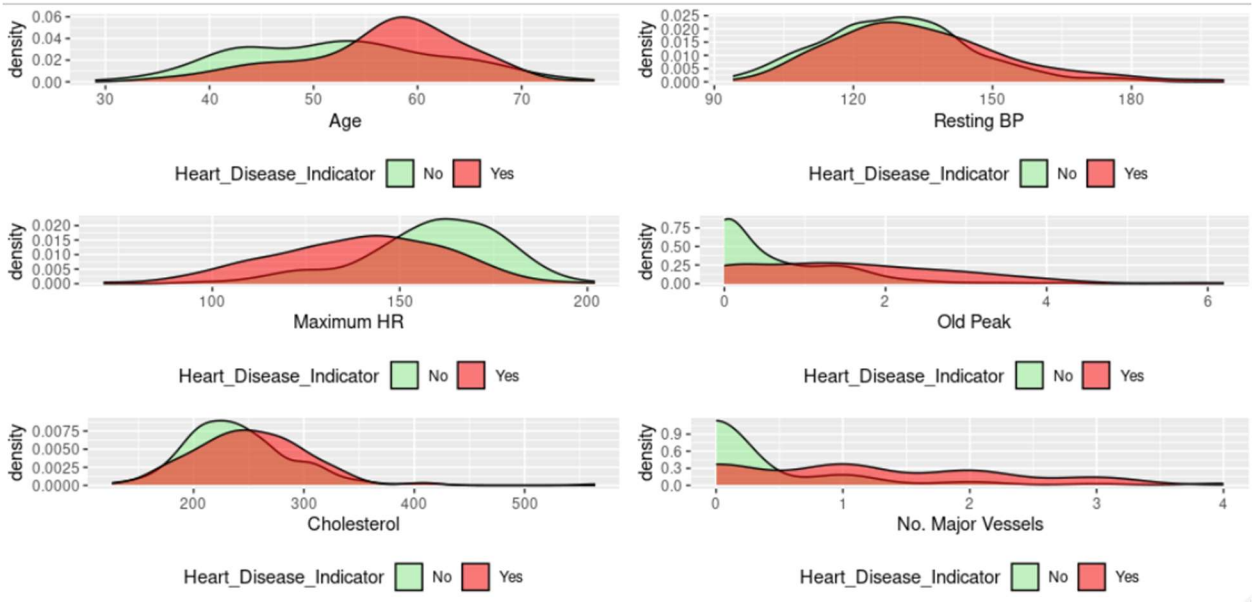


Figure 5

4.1 Age vs Heart Disease Indicator

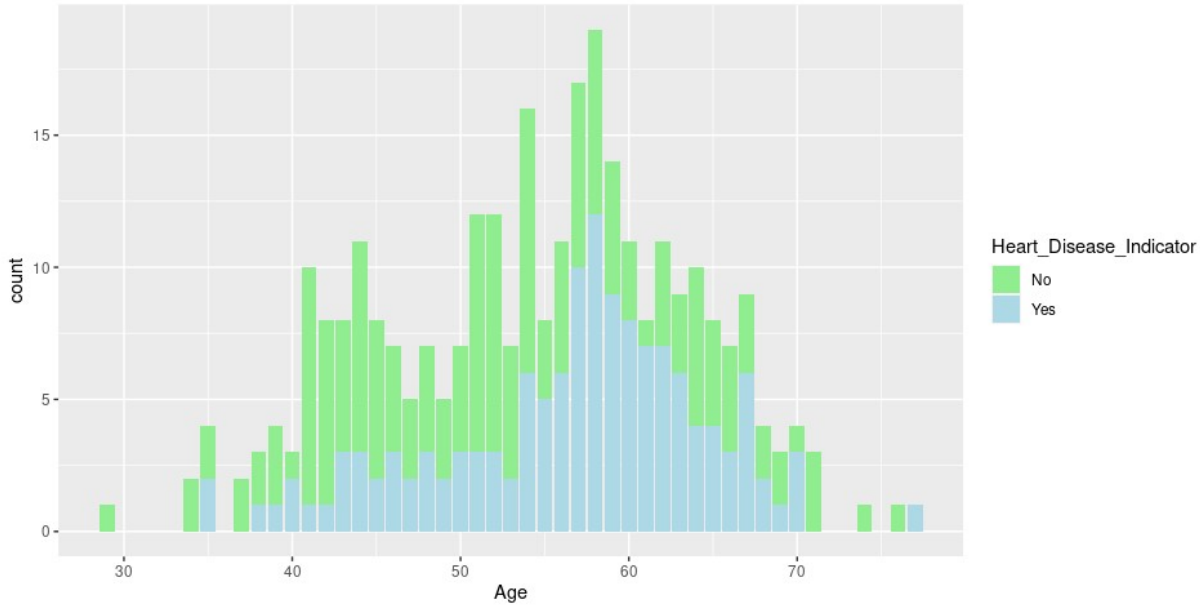


Figure 6

When evaluating the age bar plot, the density plot and disease indicator intersection reveal an unusual finding. Notably, the statistics imply that the presence of cardiac disease may not always be determined solely by age. The biggest percentage of heart disease cases and non-cases are in people between 55 and 65. This complex overlap suggests that relying merely on age as a discrete predictor of the disease's occurrence may need to be corrected, highlighting the need for a more thorough approach to understanding the varied nature of heart disease's predictors.

4.2 Chest Pain Type vs Heart Disease Indicator

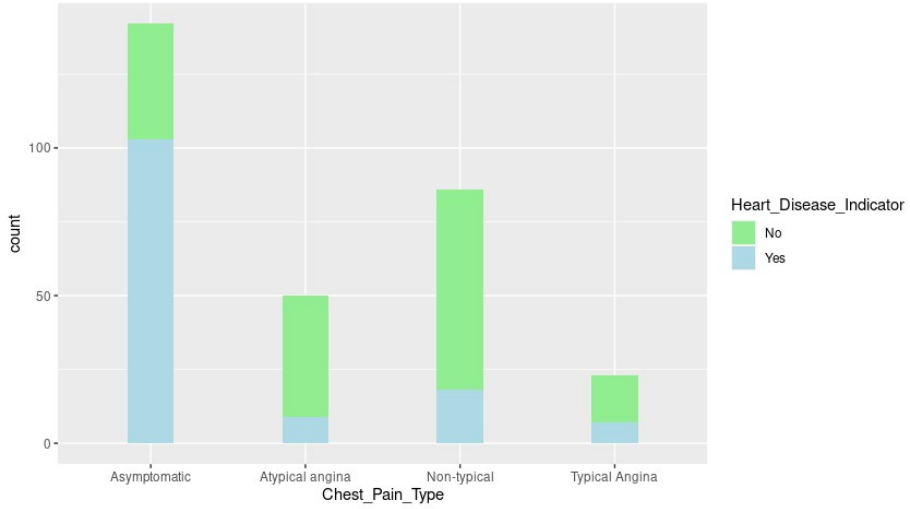


Figure 7

In chest pain, most people also suffer an asymptomatic condition, which denotes either the lack of significant symptoms during a heart attack or the presence of minor ones. Blood flow to a particular heart area is momentarily blocked in this asymptomatic form. The second most frequent occurrence is labeled as non-typical pain, defined as chest discomfort that resembles chest pain but lacks its distinct severity. The remaining two types of symptoms are allocated proportionally lesser amounts.

4.3 Sex vs Heart Disease Indicator

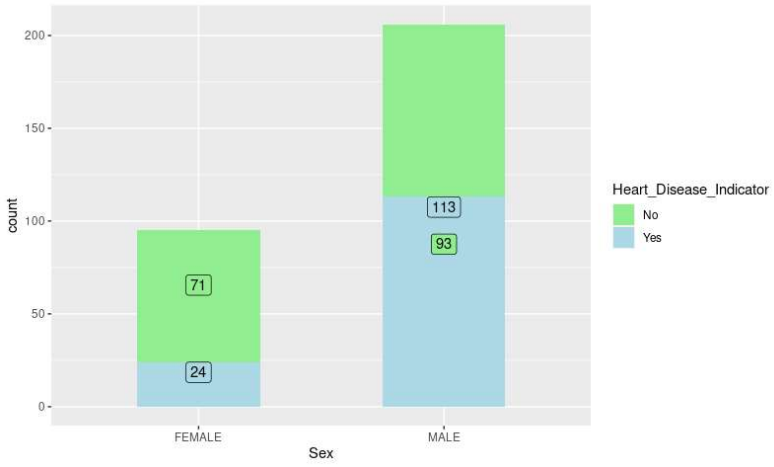


Figure 8

The disease indicator shows about a 50/50 ratio of males, which is balanced. However, the ratio drastically changes for females, where the sickness affects about 25%, while males experience a larger prevalence—where the affected rate reaches 50%. This striking difference in disease prevalence between the sexes emphasizes the possibility that gender may play a significant role in determining vulnerability to heart disease.

4.4 Cholesterol with Age

Let's analyze cholesterol levels in more detail. These levels show a nearly equal distribution of males and females, with certain female fractions showing elevated values. These high cholesterol levels are not the disease's primary cause, either. On the other hand, when looking at how cholesterol affects heart disease in men, an interesting finding is made. Males account for a sizable share of cases where heart disease is linked to cholesterol-related factors. This emphasizes the possibility that cholesterol levels may have a gender-specific impact on the onset of heart disease.

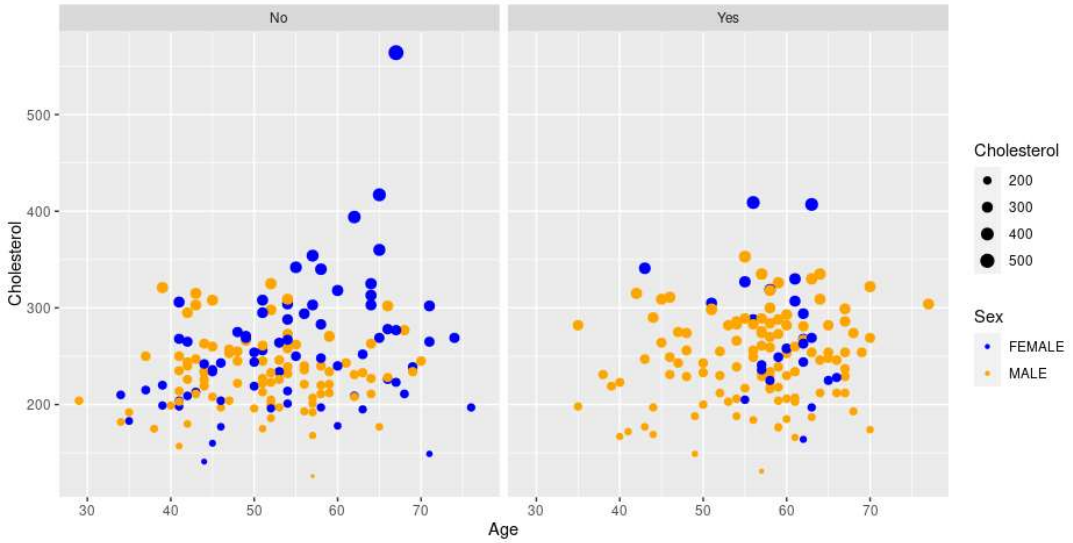


Figure 9

4.5 Age vs Maximum Heart Rate

The variation in maximum heart rate among age groups is the next issue. Maximum heart rates vary widely, with some people in their 30s showing rates equivalent to those in their 70s. This difference is shown even more when both male and female populations are considered. People with heart rates between 130 and 150 BPM seem more vulnerable to the disease's symptoms. Intriguingly, there are instances where people have maximum heart rates higher than 200 BPM yet don't show any symptoms of the condition. This fascinating finding highlights the intricate relationship between maximum heart rate and the prevalence of cardiac disease and implies additional risk variables are involved in predicting disease susceptibility.

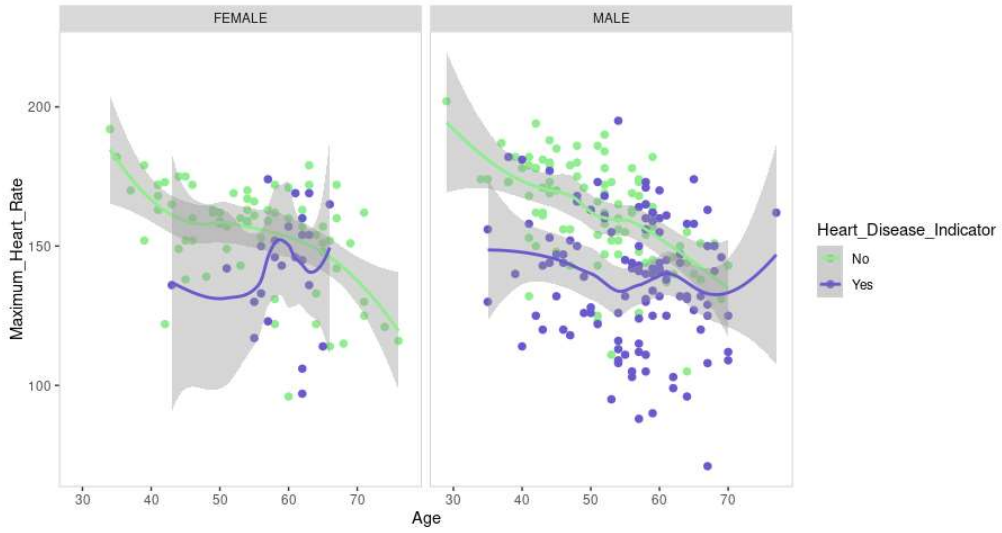


Figure 10

4.6 Resting Blood Pressure vs Chest Pain Type

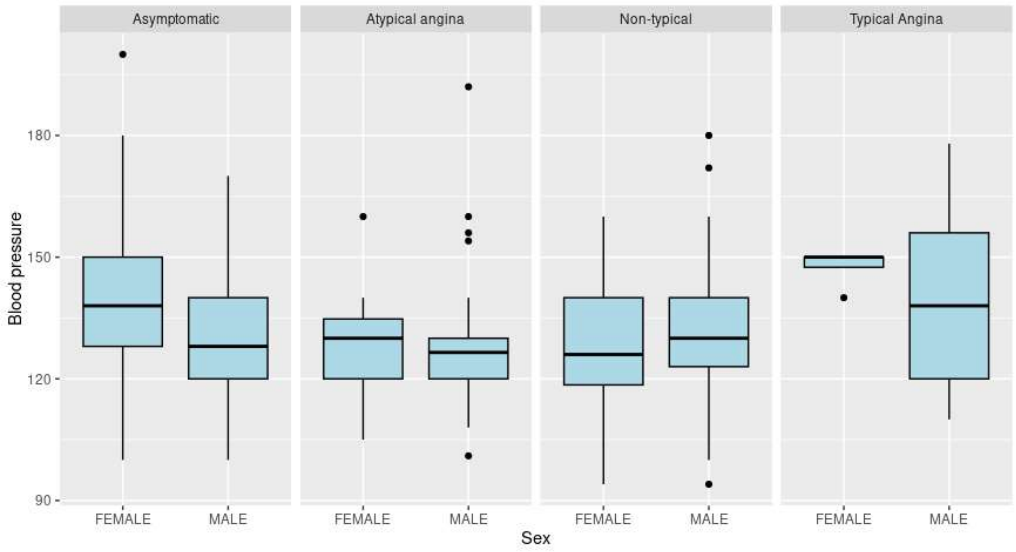


Figure 11

The next one deals with the variation in maximal heart rate between age groups. There is a wide range in maximal heart rate values, with some people in their 30s showing rates like those in their 70s. This discrepancy is shown even more when both male and female populations are considered. Notably, the condition appears more frequently in people with heart rates between 130 and 150 BPM. It's interesting to note that some people have maximum heart rates higher than 200 BPM without displaying any symptoms of the condition. The complex relationship between maximal heart rate and the occurrence of cardiac disease is highlighted by this intriguing result, which raises the possibility that additional risk variables are involved in predicting disease susceptibility.

4.7 Heart Rate Slope vs Heart Disease Indicator

Many patients with the condition exhibit a flat slope, as opposed to those who show a rising slope, which is revealed by looking at the slope variable. This difference in slope profiles suggests a possible relationship between the vulnerability to heart disease and the slope properties of some cardiac features. In-depth research into the physiological foundations of these slope patterns may reveal vital insights into the complex mechanisms triggering the disease's beginning.

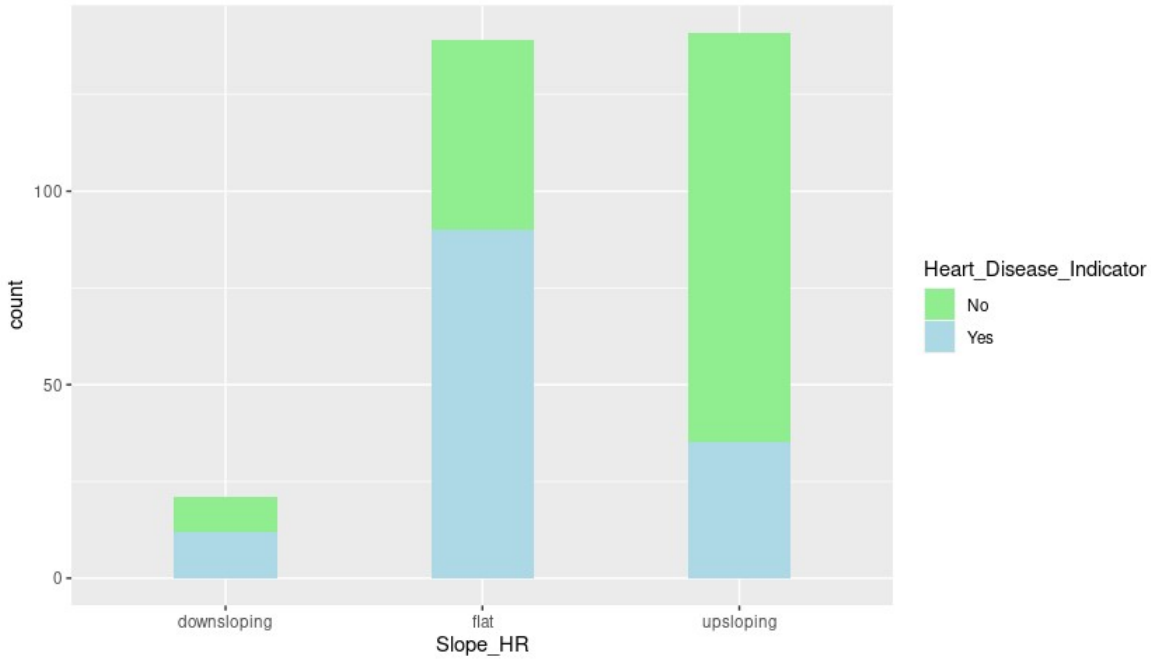


Figure 12

4.8 Thallium vs Heart Disease Indicator

According to the thallium stress test results, a striking pattern appears, people who show a reversible impairment in the thallium stress test typically have signs of cardiac disease. This finding points to a substantial correlation between a reversible abnormality and the risk of receiving a cardiac disease diagnosis. Thorium stress testing's diagnostic value for identifying people at risk for heart disease may be improved with further investigation into the physiological processes contributing to such test outcomes.

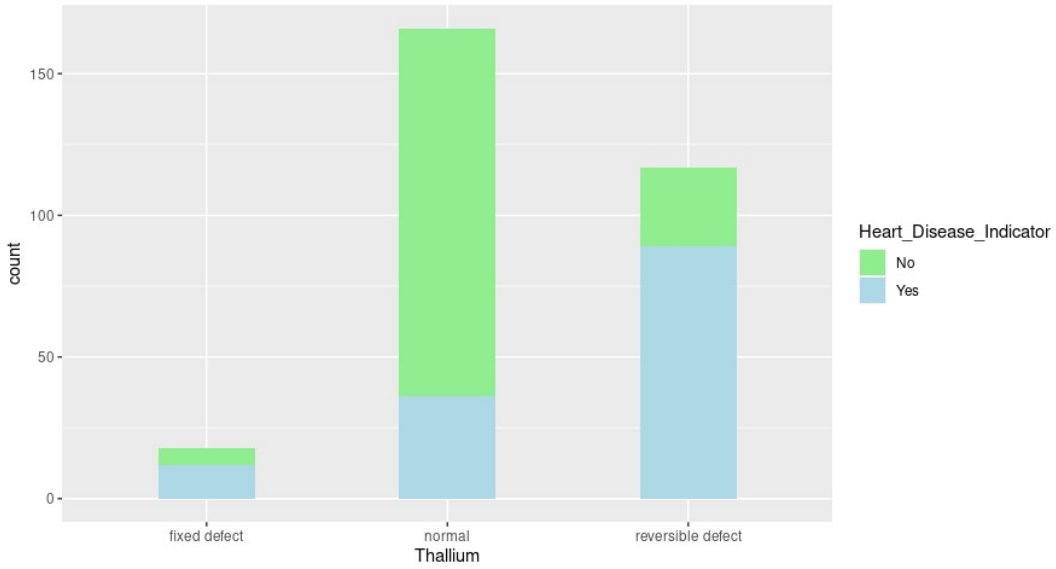


Figure 13

4.8 Correlation plot & PCA

Detailed correlations between various variables are shown in the table below, along with the results of a principal component analysis on a selection of the variables. This combined presentation provides insightful information about how these components interact, highlighting potential patterns and dependencies that may help us comprehend the dynamics of the dataset. Researchers can learn important details about the underlying structure of the data by looking at both the correlation matrix and the outcomes of principal component analysis, thereby paving the way for more informed analyses and interpretations.

.

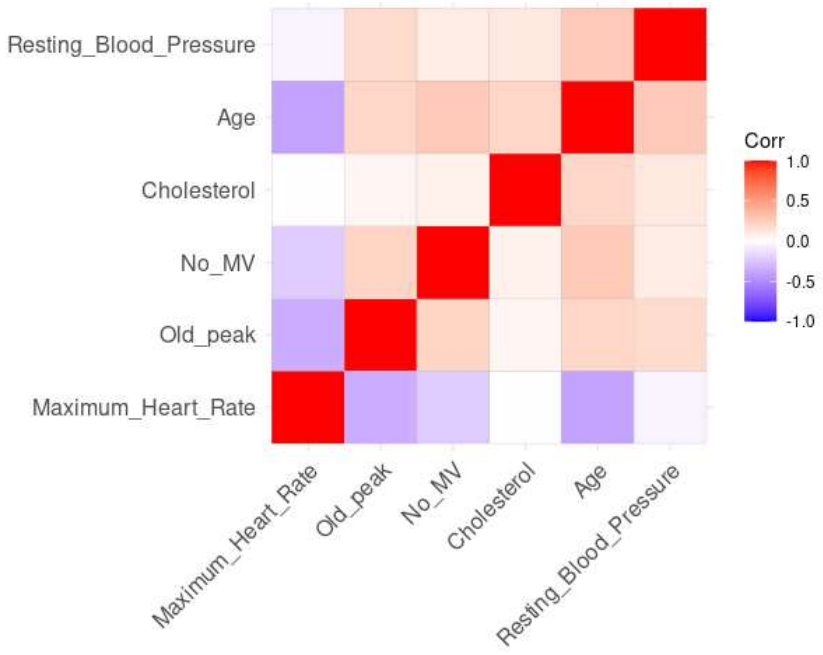


Figure 14

```
Importance of components:
                          PC1     PC2      PC3     PC4     PC5     PC6
Standard deviation     52.0067 23.2756 17.51948 7.66468 1.10593 0.93430
Proportion of Variance  0.7483  0.1499  0.08492 0.01625 0.00034 0.00024
Cumulative Proportion   0.7483  0.8982  0.98317 0.99942 0.99976 1.00000
```

5. DATA PRE-PROCESSING

5.1 Feature Selection

After finishing the data analysis stage, we are prepared to preprocess the data for model construction. We will examine a variable importance plot, which displays the importance and arrangement of columns, to direct our feature selection. To maximize the usefulness and accuracy of our predictive model, we can use this information to rank the most important qualities.

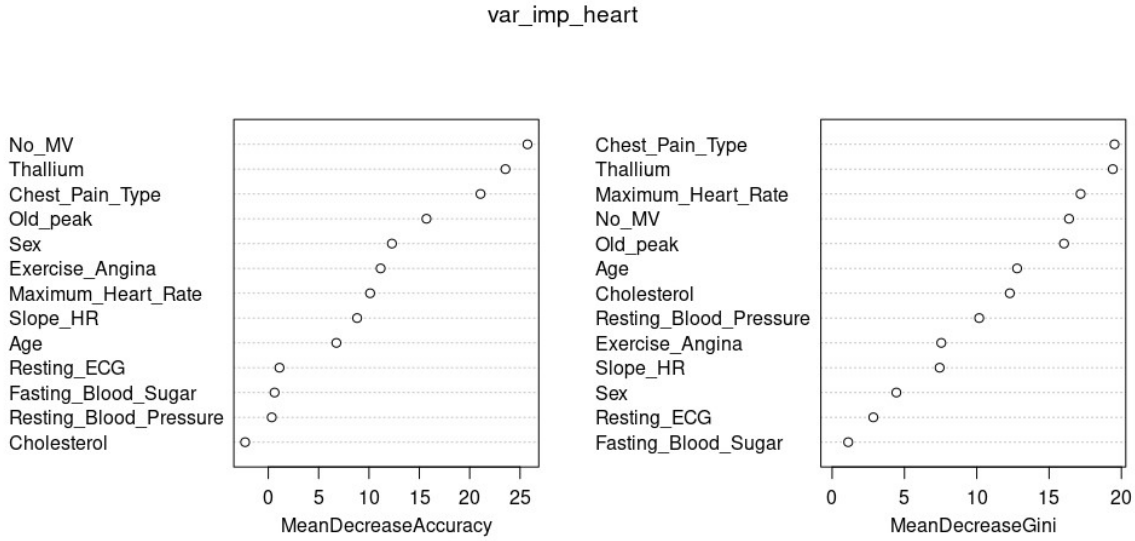


Figure 15

As we can see from the plot Fast_Blood_Sugar, Cholesterol, resting_ECG and Resting_Blood_Pressure don't contribute much to our model so we will exclude it from our dataset while predicting accuracy.

5.2 Splitting Data to Train and Test

The next stage is partitioning the data into training and testing sets for the model training process. We will use a split ratio of 75 to 25, as the data is very low splitting with 90-10 or 80-20 resulted in very less test data.

6. MODEL APPROACH

Since it is a classification problem, several modelling techniques have been used and accuracy has been used as a performance metric.

6.1 Logistic Regression

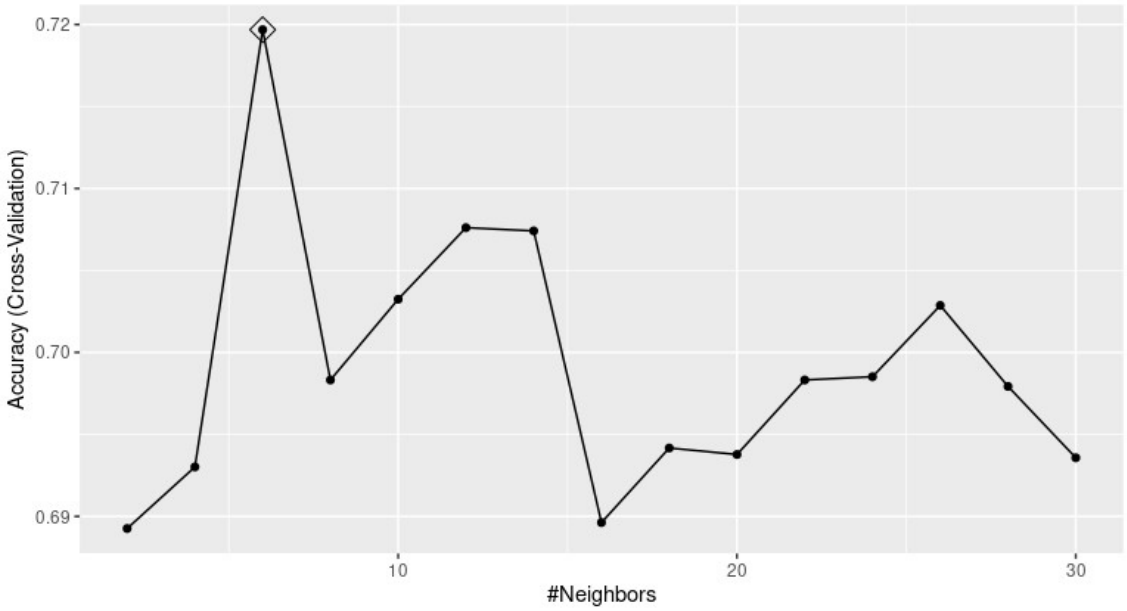
A logistic regression approach began modeling with a preset seed value of 0811. The model was trained on the training dataset (H_train) using all available predictors. The training method was "glm," and a control parameter (ctrl) was implemented to control the training procedure. The trained logistic regression model (glm_train) was then used to produce predictions on the testing dataset (H_test) using the predict function.

A confusion matrix (glm_CM) was built to assess the model's performance. The evaluation of several performance parameters, such as overall accuracy, sensitivity (true positive rate), and specificity (true negative rate), was made possible by this matrix. The logistic regression model's obtained accuracy was roughly 0.8933 (89.33%). This initial accuracy score shows that the logistic regression model is off to a promising start. To determine the best strategy for forecasting heart disease results, the analysis will also explore other models.

Method <chr>	Accuracy <dbl>
Logistic Regression	0.8933333

6.2 KNN

Optimal K value for performing KNN is 6.



By performing KNN, the model had much less accuracy when compared to the Logistic Regression model. It has achieved an accuracy of 68%.

Method <chr>	Accuracy <dbl>
Logistic Regression	0.8933333
KNN	0.6800000

6.3 Decision Trees

Decision trees has achieved 76% accuracy which is slightly better than decision trees.

Method <chr>	Accuracy <dbl>
Logistic Regression	0.8933333
KNN	0.6800000
Decision Trees	0.7600000

6.4 Random Forest

Below is the variable importance plot showing the effect on target variable of each feature. After performing random forest algorithm, it yielded slightly greater accuracy than decision trees but less than the logistic regression model. The model has given 81.3% accuracy.

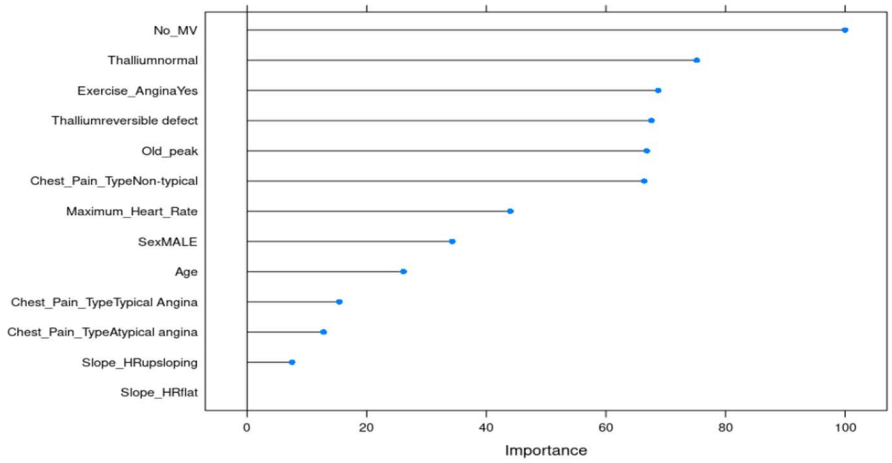


Figure 16

Method <chr>	Accuracy <dbl>
Logistic Regression	0.8933333
KNN	0.6800000
Decision Trees	0.7600000
Random Forest	0.8133333

6.5 Adaboost

Boosting algorithm 'Adaboost' is performed. But this has given a very less accuracy when compared to Random Forest model.

Method <chr>	Accuracy <dbl>
Logistic Regression	0.8933333
KNN	0.6800000
Decision Trees	0.7600000
Random Forest	0.8133333
Ada Boost	0.7733333

7. RESULTS AND INSIGHTS

7.1 Results

	Logistic_Regression	KNN	Regression_Trees	Random_Forest	Ada_boost
Sensitivity	0.8529412	0.6470588	0.7941176	0.7647059	0.7941176
Specificity	0.9268293	0.7073171	0.7317073	0.8536585	0.7560976
Pos Pred Value	0.9062500	0.6470588	0.7105263	0.8125000	0.7297297
Neg Pred Value	0.8837209	0.7073171	0.8108108	0.8139535	0.8157895
Precision	0.9062500	0.6470588	0.7105263	0.8125000	0.7297297
Recall	0.8529412	0.6470588	0.7941176	0.7647059	0.7941176
F1	0.8787879	0.6470588	0.7500000	0.7878788	0.7605634
Prevalence	0.4533333	0.4533333	0.4533333	0.4533333	0.4533333
Detection Rate	0.3866667	0.2933333	0.3600000	0.3466667	0.3600000
Detection Prevalence	0.4266667	0.4533333	0.5066667	0.4266667	0.4933333
Balanced Accuracy	0.8898852	0.6771879	0.7629125	0.8091822	0.7751076

7.2 Insights

Sensitivity (True Positive Rate): Logistic Regression has the highest sensitivity of approximately 85%, indicating that it correctly identifies a substantial proportion of individuals with heart disease. Random Forest and AdaBoost models also show good sensitivity scores, around 76%, and 79%, respectively.

Specificity (True Negative Rate): Logistic Regression and Random Forest have relatively higher specificity values, suggesting their ability to identify individuals without heart disease correctly. This is crucial to minimize false positives.

Positive Predictive Value: Positive predictions made by Logistic Regression and Random Forest are more likely to be accurate because of their larger positive predictive values.

Negative Predictive Value: Random Forest and AdaBoost models have relatively higher negative predictive values, indicating their ability to predict negative cases accurately.

Balanced Accuracy: Logistic Regression exhibits the highest balanced accuracy, considering both sensitivity and specificity, making it a strong performer in overall classification accuracy.

Prevalence and Detection Rates: Prevalence represents the proportion of positive cases, and detection rate indicates the proportion of correctly predicted positive cases. These values are similar across models.

Detection Prevalence: The proportion of predicted positive cases out of all predictions is highest for the Regression Trees model.

In conclusion, the logistic regression and random forest models have demonstrated the best performance among the models considered. Their balanced accuracy, sensitivity, specificity, and positive predictive value are noteworthy. We may better grasp the models' strengths and potential weaknesses with the help of these findings, which can be very important when making judgments to anticipate heart disease.

7.3 Conclusion

The main goal of this study was to effectively diagnose people with heart illnesses by utilizing the Cleveland heart disease dataset. Through thorough exploratory data analysis, we learned more about the different dataset characteristics essential for forecasting the prevalence of the disease. To increase the model's effectiveness, we also found a few parameters that had little effect on the results and removed them.

To maximize prediction accuracy, a variety of machine learning models were created. The models that were most effective in making accurate predictions were the Random Forest and Logistic Regression models. The K-Nearest Neighbors (KNN) model, on the other hand, produced fewer desirable outcomes. Although the accuracy was considered satisfactory, the sensitivity and specificity scores remained below the intended 90% level. Given the available dataset, this is a cause for caution, but the results are still valid.

To maximize prediction accuracy, a variety of machine learning models were created. The models that were most effective in making accurate predictions were the Random Forest and

Logistic Regression models. The K-Nearest Neighbors (KNN) model, on the other hand, produced fewer desirable outcomes. Although the accuracy was considered satisfactory, the sensitivity and specificity scores remained below the intended 90% level. Given the available dataset, this is a cause for caution, but the results are still valid.

8 REFERENCES

https://www.heartandstroke.ca/heart-disease/what-is-heart-disease/types-of-heart-disease

https://archive.ics.uci.edu/ml/datasets/heart+disease

https://uc-r.github.io/regression_trees

https://towardsdatascience.com/random-forest-in-r-f66adf80ec9

http://finzi.psych.upenn.edu/R/library/caret/html/sensitivity.html