Introduction

Heart disease, also known as cardiovascular disease, refers to conditions that affect the heart and blood vessels. According to the Centers for Disease Control and Prevention (CDC), an estimated 121.5 million adults in the United States have cardiovascular disease (CVD), which includes heart disease. This represents approximately 48% of the adult population. This shows the importance of understanding the causes and risk factors associated with heart disease to promote early detection and prevention.

In this project, I will develop a classification model to predict whether a patient is likely to have heart disease based on key health factors such as blood sugar levels, exercise habits, cholesterol, blood pressure, and other cardiovascular indicators.

Introduction to the data

The dataset is sourced from Kaggle called "Heart Failure Prediction". This dataset contains medical and lifestyle related features that can predict heart disease. It compiles 5 heart disease datasets from Cleveland, Hungarian, Switzerland, Long Beach VA, and Stalog (Heart) Data Set. There are a total of 918 rows and 12 columns in the dataset.

The columns in the dataset include:

- Age: age of the patient
- Sex: sex of the patient
- ChestPainType: chest pain type (TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic)
- RestingBP: resting blood pressure (mm Hg)
- Cholesterol: serum cholesterol (mm/dl)
- FastingBS: fasting blood sugar (1: if FastingBS > 120 mg/dl, 0: otherwise)

- RestingECG: resting electrocardiogram results (Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria])
- MaxHR: maximum heart rate achieved (Numeric value between 60 and 202)
- ExerciseAngina: exercise-induced angina (Y: Yes, N: No])
- Oldpeak: oldpeak = ST (Numeric value measured in depression)
- ST_Slope: the slope of the peak exercise ST segment (Up: upsloping, Flat: flat, Down: downsloping)
- HeartDisease: output class (1: heart disease, 0: Normal)

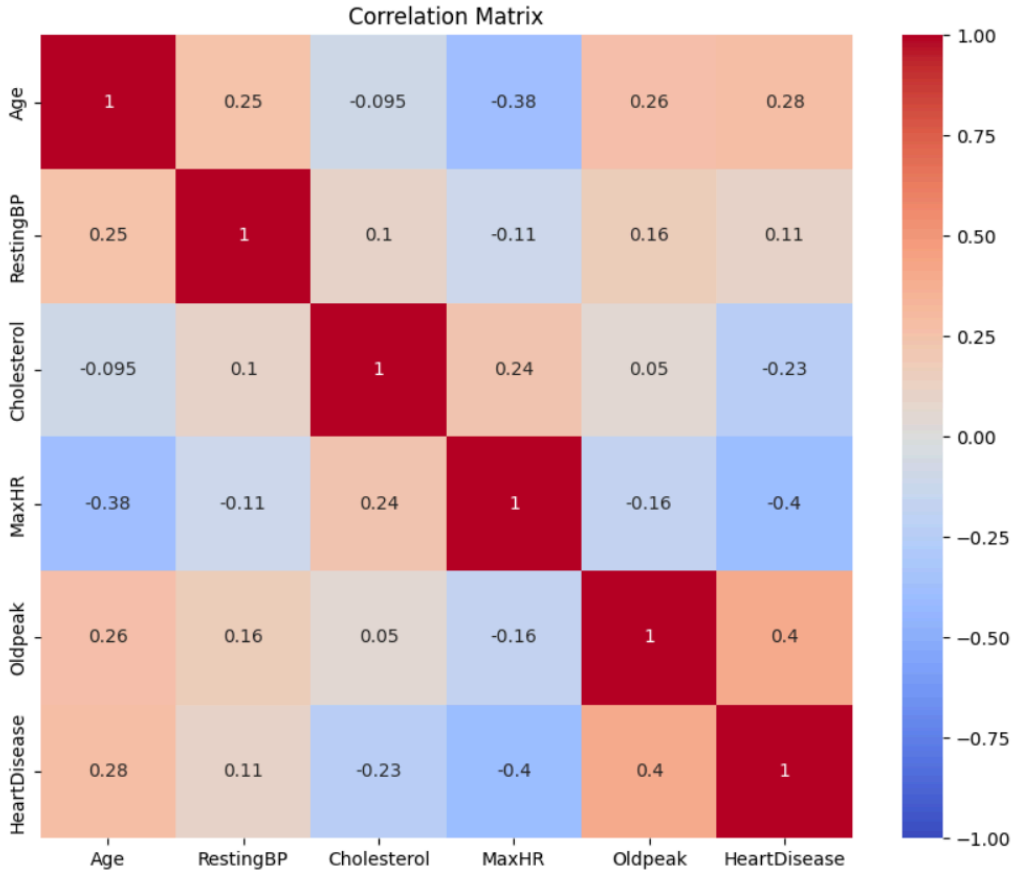
Pre-processing

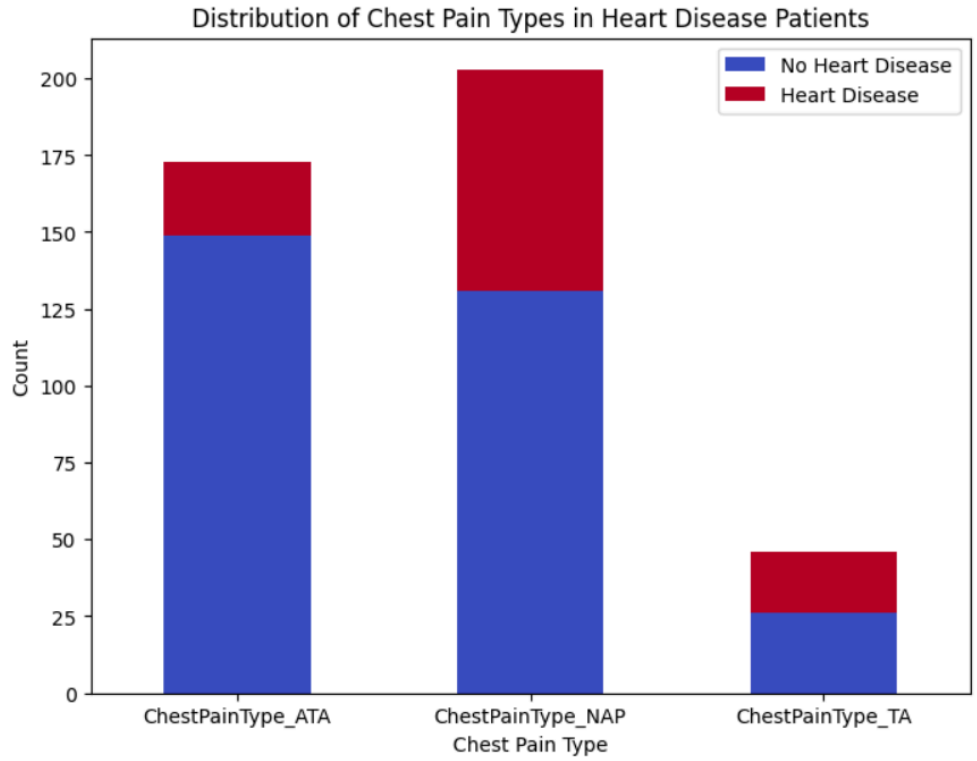
I did several steps to clean the dataset. First, I checked for null and duplicate values in the dataset. Since there were no missing values or duplicate rows, I proceeded with the next steps. Which is converting binary and multi-class categorical variables into numerical format.
Label encoding applied to features:

- Sex: M(Male) = 1 and F(Female) = 0
- ExerciseAngina: Y (Yes) = 1 and N (No) = 0

Then I used one-hot encoding for variables like ChestPainType, RestingECG, and ST_Slope. By doing so I created separate binary columns for each category. I also used the drop_first=True parameter to avoid multicollinearity. After these steps the dataset was ready to use for analysis and modeling.

Data Understanding/Visualization


Correlation Matrix
Age
RestingBP
Cholesterol
MaxHR
Oldpeak
HeartDisease
1
0.25
-0.095
-0.38
0.26
0.28
0.25
1
0.1
-0.11
0.16
0.11
-0.095
0.1
1
0.24
0.05
-0.23
-0.38
-0.11
0.24
1
-0.16
-0.4
0.26
0.16
0.05
-0.16
1
0.4
0.28
0.11
-0.23
-0.4
0.4
1
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00


Distribution of Chest Pain Types in Heart Disease Patients
No Heart Disease
Heart Disease
Count
200
175
150
125
100
75
50
25
0
ChestPainType_ATA
ChestPainType_NAP
ChestPainType_TA
Chest Pain Type

To understand the dataset I made a heatmap. I observed MaxHR and Oldpeak seem to be the most linked factors to heart disease. Cholesterol and RestingBP have weak correlations with heart disease, meaning they might not be the best predictors. Age has a slight correlation, indicating older people may have a higher risk of heart disease.

I also made a bar chart to show the distribution of chest pain types in heart disease patients. I observed Typical Angina (TA), is strongly associated with heart disease. Atypical Angina (ATA) is commonly seen in people without heart disease, making it a weaker predictor. Non-Anginal Pain (NAP) is more frequently observed in heart disease patients compared to ATA.

Modeling

For this project, I used Logistic Regression and Random Forest to predict heart disease based on patient data. Logistic Regression is a supervised machine learning algorithm that accomplishes binary classification tasks by predicting the probability of an outcome, event, or observation. It calculates a weighted sum of input features and passes it through a sigmoid function to output a probability between 0 and 1. If the probability is above a chosen threshold, the instance is classified as class 1 or otherwise 0. Some benefits of this model are that it is easy to interpret and fast. Some disadvantages of this model are that it constructs linear boundaries, and is not effective for complex relationships in data. I chose Logistic Regression as a baseline model because it is interpretable and useful for understanding feature importance in predicting heart disease.

Random Forest which is a machine learning algorithm that uses many decision trees to make predictions. It creates multiple decision trees from random subsets of the training data. Each tree makes a prediction, and the final output is determined by classification or regression. Some

benefits of this model are it can handle nonlinear relationships and interactions between features, and it can handle large datasets. Some disadvantages of this model are it is less interpretable than Logistic Regression and requires more computational power. I chose Random Forest because it can handle more complex relationships in the data and can provide more accuracy than the Logistic Regression model.

Evaluation

The metrics I used to access the model are accuracy, precision, recall, F1-score, and the confusion matrix. These metrics provide a comprehensive evaluation of the model's performance.

Logistic regression results:

```
Logistic Regression Accuracy: 0.8532608695652174
              precision    recall  f1-score   support

           0       0.80      0.87      0.83        77
           1       0.90      0.84      0.87       107

    accuracy                           0.85       184
   macro avg       0.85      0.86      0.85       184
weighted avg       0.86      0.85      0.85       184

Confusion Matrix:
 [[67 10]
 [17 90]]
```

Logistic regression model performed well, but it misclassified 17 false negatives, which means it failed to identify 17 patients with heart disease.

Random Forest results:

```
Random Forest Accuracy: 0.8586956521739131
              precision    recall  f1-score   support

           0       0.82      0.84      0.83        77
           1       0.89      0.87      0.88       107

    accuracy                           0.86       184
   macro avg       0.85      0.86      0.86       184
weighted avg       0.86      0.86      0.86       184

Confusion Matrix:
[[65 12]
 [14 93]]
```

Random Forest also had a good accuracy score and was slightly higher than the Logistic regression model. It misclassified 14 false negatives, which means it failed to identify 14 patients with heart disease. Both models performed well, with Random Forest having slightly better accuracy and recall than Logistic Regression.

Storytelling

Through this project I learned how different models perform in predicting heart disease and the importance of evaluating errors. Logistic Regression provided a clear understanding of which factors contribute to heart disease risk, while Random Forest improved prediction accuracy by looking at complex patterns in the data. My question was about predicting heart disease using patient data, and the results from the models show that both Logistic Regression and Random Forest have good performance, Random Forest being slightly better. However, since both models still misclassified some cases, additional research may be needed to enhance predictions.

Impact Section

The positive impact of this project is that it can help individuals understand what factors could cause heart disease. This will help them make healthy lifestyle choices and maintain better

health. A potential negative impact is that some important factors, such as family history, may not have been considered, which could lead to inaccurate results.

References

https://www.nm.org/healthbeat/healthy-tips/heart-facts-infographic

https://www.heart.org/en/news/2019/01/31/cardiovascular-diseases-affect-nearly-half-of-american-adults-statistics-show

https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-logistic-regression/#:~:text=Logistic%20regression%20is%20a%20supervised%20machine%20learning%20algorithm%20that%20accomplishes,1%2C%20or%20true%2Ffalse.

https://www.ibm.com/think/topics/random-forest#:~:text=Random%20forest%20is%20a%20commonly,both%20classification%20and%20regression%20problems.