Heart Disease Prediction

Machine Learning Approach

Table of Contents

1. **Introduction**

    1.1. Background
    1.2. Problem Statement
    1.3. Objectives

2. **Data Collection and Understanding**

    2.1. Description of the Dataset
    2.2. Data Sources
    2.3. Data Features and Explanation
    2.4. Data Cleaning and Preprocessing

3. **Exploratory Data Analysis (EDA)**

    3.1. Univariate Analysis
    3.2. Bivariate Analysis

4. **Feature Engineering**

    4.1. Scaling and Normalization
    4.2. Encoding Categorical Variables
    4.3. Handling Imbalanced Data
    4.4. Feature Selection and Reduction

5. **Model Building**

    5.1. Model Selection
    5.2. Hyperparameter Tuning Strategies (RandomizedSearchCV, etc.)
    5.3. Model Selection

6. **Conclusion**
7. **Recommendation**
8. **Appendices**

Heart Disease Prediction

1. Introduction

1.1. Background and Motivation

Heart disease is a leading cause of death worldwide. Early detection and prediction of heart disease are crucial for timely intervention and improved patient outcomes. Machine learning techniques have shown promising results in predicting heart disease risk based on individual patient characteristics.

1.2. Problem Statement

This project addresses the problem of predicting heart disease using a machine learning model trained on a dataset of patient information. The goal is to develop a model that accurately classifies individuals as having or not having heart disease.

1.3. Objectives

- To explore and analyze a heart disease dataset.
- To preprocess the data, handle missing values, and engineer relevant features.
- To develop a machine learning model for heart disease prediction.
- To evaluate the performance of the model and compare different algorithms.

2. Data Collection and Understanding

2.1 Description of the Dataset

The dataset used in this project contains information about patients, including various medical attributes such as age, sex, chest pain type, blood pressure, and cholesterol levels. The dataset includes both numerical and categorical features.

- **Data Source:** The dataset used was sourced Data Science Nigeria X Microsoft 2024 AI Bootcamp Qualification Hackathon on Zindi.

- **Data Features:**
    - Age: Age of the patient
    - Sex: Gender of the patient (0 = Female, 1 = Male)
    - cp: Chest pain type (0 = Typical Angina, 1 = Atypical Angina, 2 = Non-Anginal Pain, 3 = Asymptomatic)
    - trestbps: Resting blood pressure (in mm Hg on admission to the hospital)
    - chol: Serum cholesterol in mg/dl
    - fbs: Fasting blood sugar > 120 mg/dl (0 = No, 1 = Yes)

- restecg: Resting electrocardiographic results (0 = Normal, 1 = ST-T Abnormality, 2 = LV Hypertrophy)
- thalach: Maximum heart rate achieved
- exang: Exercise-induced angina (0 = No, 1 = Yes)
- oldpeak: ST depression induced by exercise relative to rest
- slope: Slope of the peak exercise ST segment (0 = Upsloping, 1 = Flat, 2 = Downsloping)
- ca: Number of major vessels (0-3) colored by fluoroscopy
- thal: 0 = Normal; 1 = Fixed defect; 2 = Reversable defect
- target: Heart disease (0 = No, 1 = Yes)

2.4. Data Cleaning and Preprocessing

- **Missing Values:** The dataset happens to be clean but there are 2697 missing values in the target variable (Heart disease)
- **Data Transformation:**
  - Categorical features were encoded using Label Encoding.
  - Age was binned into different age groups ('Young', 'Adult', 'Middle Age', 'Old').

3. Exploratory Data Analysis (EDA)

3.1. Univariate Analysis

The dataset's age distribution was slightly right-skewed, indicating that most patients were middle-aged. Resting blood pressure, on the other hand, followed a relatively normal distribution, though there were a few potential outliers on the higher end. Similarly, cholesterol levels exhibited a fairly normal distribution, with outliers present on the lower and higher extremes.

The maximum heart rate distribution was slightly left-skewed, with a few outliers at the lower end. Regarding gender, there were more male patients than female patients in the dataset. When looking at chest pain type, most patients experienced atypical angina or non-anginal pain, suggesting these were the predominant forms of chest discomfort.

The data on fasting blood sugar showed that the majority of patients had levels below 120 mg/dl while resting electrocardiographic (ECG) results were mostly normal across the patient population. Exercise-induced angina was relatively rare, with most patients not experiencing it. Additionally, the majority of patients had a flat or upsloping ST segment during peak exercise.

A significant number of patients had no major vessels colored by fluoroscopy, indicating fewer instances of extensive coronary artery disease. The data on thalassemia showed a mix of patients with different forms of the condition, but no particular type dominated the distribution.

Finally, the target variable—whether a patient had heart disease or not—revealed a class imbalance, with more patients without heart disease than those diagnosed with it.

3.2 Bivariate Analysis

The analysis showed that age was associated with a slightly higher risk of heart disease, as older patients appeared more susceptible. However, cholesterol levels did not display any clear trend to heart disease, as seen from the box plots. Chest pain type, on the other hand, was a significant indicator, with patients experiencing asymptomatic chest pain having a higher likelihood of heart disease. Similarly, those who suffered from exercise-induced angina were more at risk of developing heart disease.

In the multivariate analysis, the combination of chest pain type and maximum heart rate appeared to jointly influence heart disease risk, indicating that these factors may interact to increase the likelihood of the condition.

In summary, the exploratory data analysis pointed to several potential risk factors for heart disease, including age, chest pain type, and exercise-induced angina. It also brought attention to the class imbalance in the target variable, which was addressed during data preprocessing using the SMOTE technique.

4. Feature Engineering

The feature engineering process included several important steps to prepare the data for machine learning models. First, categorical features such as sex, chest pain type, and ST slope were transformed into numerical representations using Label Encoding. This conversion is essential for most machine learning algorithms, as it allows them to effectively process and utilize the information embedded in these categorical variables.

Next, age binning was employed, grouping ages into categories such as "Young," "Adult," "Middle Age," and "Old." This method helps capture potential non-linear relationships between age and heart disease risk, which might be more effective than using raw age values, especially if the relationship between age and heart disease is not strictly linear.

To address class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was used to oversample the minority class—patients with heart disease. This step was crucial in mitigating the class imbalance problem, enabling the models to better identify individuals with heart disease and leading to a more balanced and robust model.

5. Model Building

Several important steps were involved in the model-building and optimization process to ensure the best possible performance.

5.1. Model Selection

A variety of machine learning models were considered to determine the most suitable algorithm for predicting heart disease. XGBoost, Random Forest, and LightGBM, which was later used due to its computational efficiency and scalability, particularly for handling large datasets. Additionally, a Stacking Ensemble method was explored, combining the predictions of models such as LightGBM and XGBoost, with the goal of improving overall accuracy.

5.2. Hyperparameter Tuning

To optimize model performance using hyperparameters. RandomizedSearchCV was employed to efficiently explore the hyperparameters. This technique samples different combinations of hyperparameter values and evaluates their performance through cross-validation.

5.3. Model Comparison

To ensure that LightGBM was the most effective option for heart disease prediction, its performance was rigorously compared with other models. The tuned LightGBM model was benchmarked against Random Forest and XGBoost, both in their default configurations and with optimized hyperparameters. The comparison confirmed that LightGBM consistently achieved the highest accuracy on the dataset, making it the best-performing model for this prediction task.

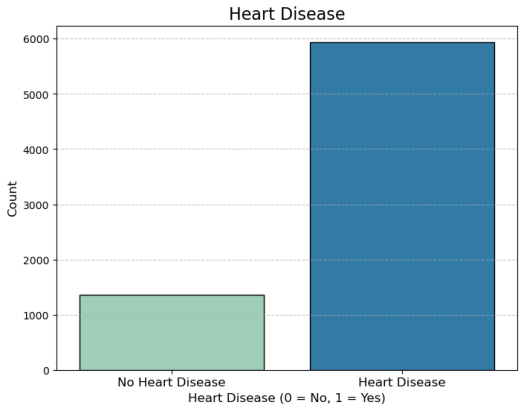
6. Conclusion

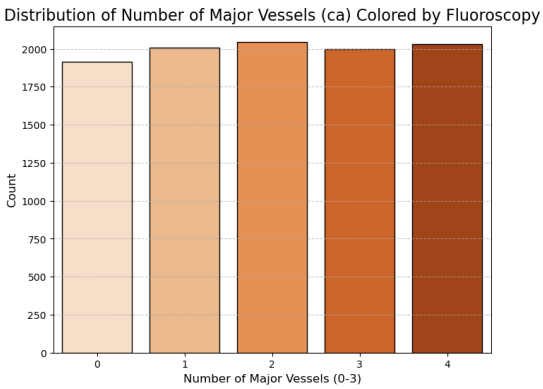
The analysis of the dataset highlighted several key factors associated with heart disease risk. Age, chest pain type, and exercise-induced angina emerged as significant predictors, with older patients and those experiencing asymptomatic chest pain or angina during exercise being more likely to have heart disease. Although cholesterol levels did not show a clear relationship with heart disease, other factors, such as maximum heart rate, appeared to influence the condition when combined with chest pain type. The class imbalance in the target variable, where there were more patients without heart disease, was addressed using the SMOTE technique to ensure balanced model training.

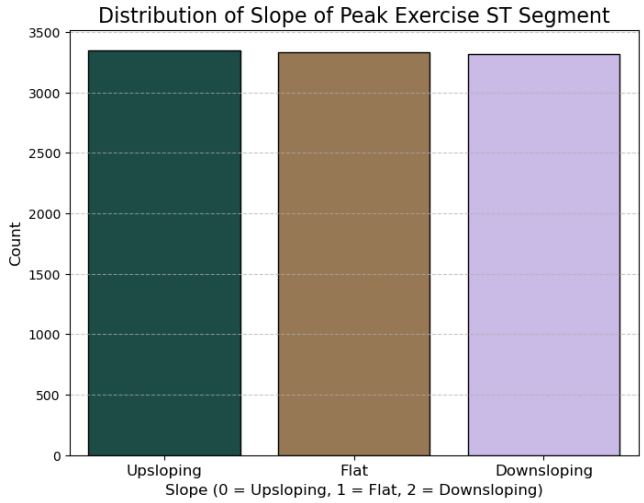
7. Recommendations

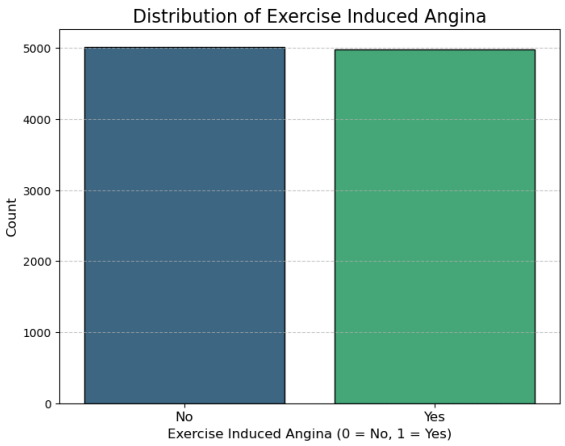
Emphasizing the significant predictors—age, chest pain type, and exercise-induced angina—will likely result in more accurate and targeted interventions for heart disease risk. Further investigation into other potential interactions between variables, such as maximum heart rate and chest pain type, could uncover additional insights.

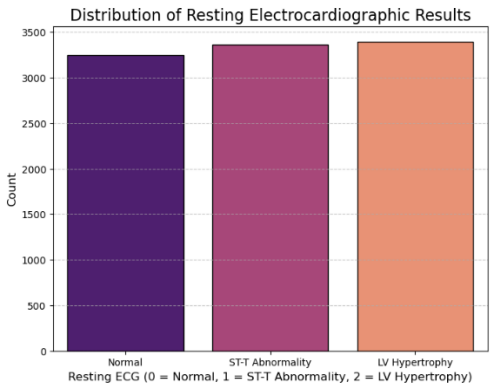
Appendix

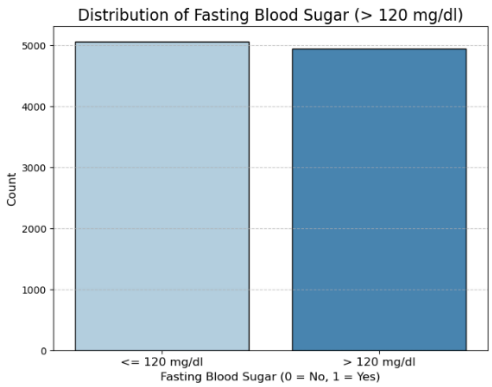


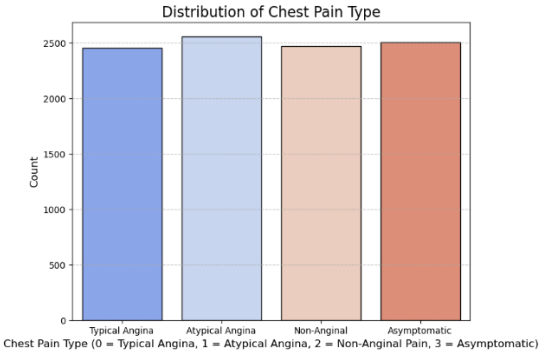


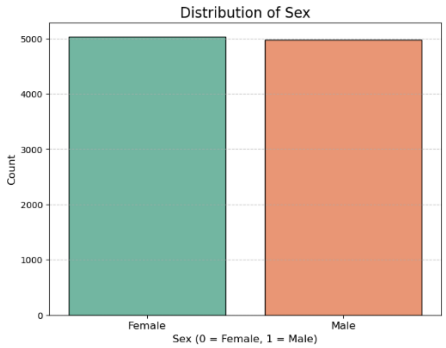


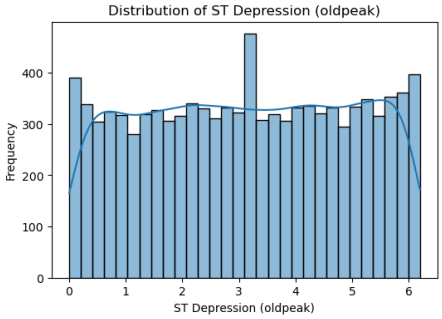


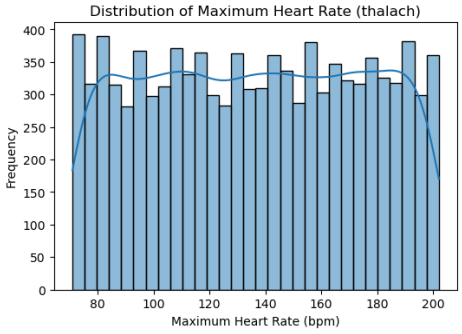


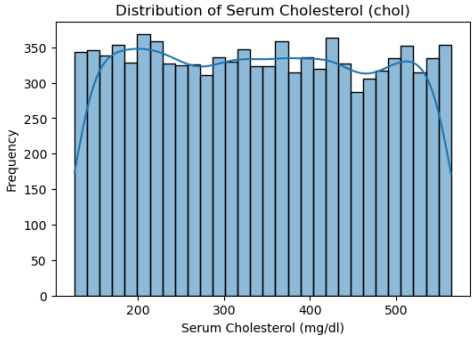


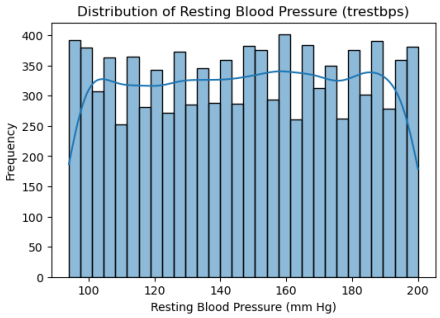

Distribution of Chest Pain Type
Count
Typical Angina
Atypical Angina
Non-Anginal
Asymptomatic
Chest Pain Type (0 = Typical Angina, 1 = Atypical Angina, 2 = Non-Anginal Pain, 3 = Asymptomatic)

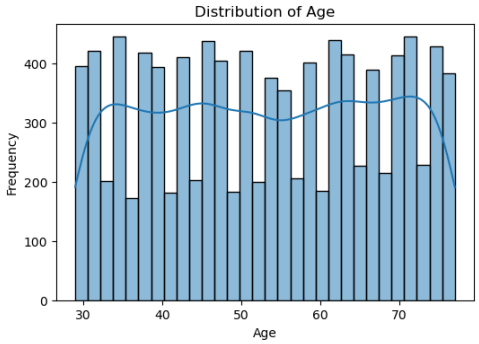

Distribution of Sex
Count
Female
Male
Sex (0 = Female, 1 = Male)

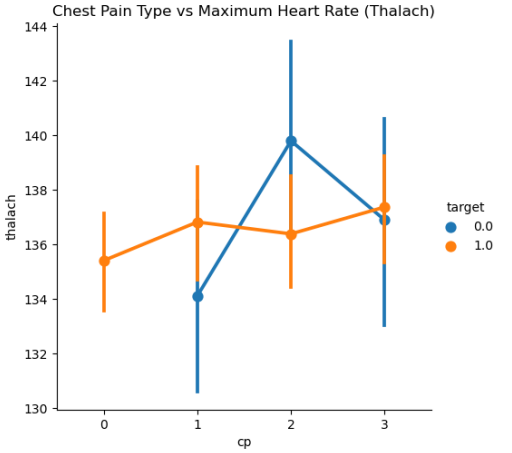

Distribution of ST Depression (oldpeak)
Frequency
ST Depression (oldpeak)

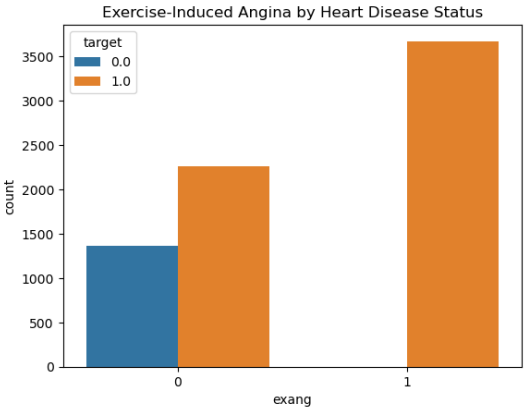

Distribution of Maximum Heart Rate (thalach)
Frequency
Maximum Heart Rate (bpm)

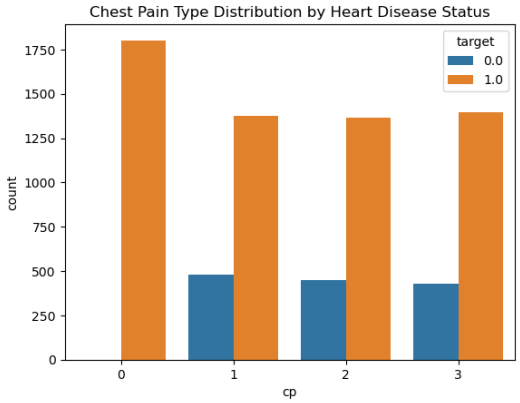

Distribution of Serum Cholesterol (chol)
Frequency
Serum Cholesterol (mg/dl)

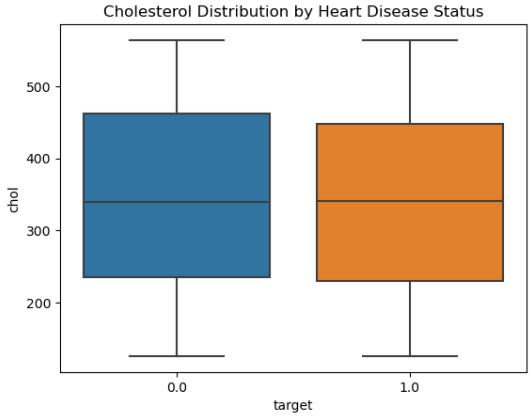

Distribution of Resting Blood Pressure (trestbps)
Frequency
Resting Blood Pressure (mm Hg)

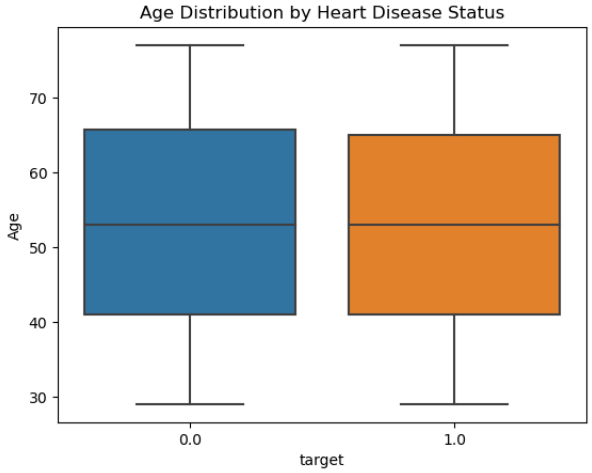

Distribution of Age
Frequency
Age


Chest Pain Type vs Maximum Heart Rate (Thalach)
thalach
cp
target
0.0
1.0


Exercise-Induced Angina by Heart Disease Status
count
exang
target
0.0
1.0


Chest Pain Type Distribution by Heart Disease Status
count
cp
target
0.0
1.0


Cholesterol Distribution by Heart Disease Status
chol
target


Age Distribution by Heart Disease Status
Age
target