Heart Disease Dataset Analysis Report

Introduction

This report aims to examine a heart disease dataset by analysing its variables, identifying correlations, and building a predictive model. Heart disease remains one of the leading causes of mortality worldwide, and early detection can save lives. By analysing clinical and demographic features associated with heart disease, this report intends to uncover significant patterns and correlations. Following exploratory data analysis (EDA), we develop and evaluate a logistic regression model to predict the likelihood of heart disease. This analysis supports understanding which variables most contribute to heart disease risk and how machine learning models can be applied in the medical field for predictive purposes.

Dataset Overview

The dataset comprises a variety of clinical measurements and demographic information, each potentially influential in determining heart disease risk. The key variables in this dataset include patient age, gender, cholesterol levels, resting electrocardiogram results, and other health indicators. To provide an overview, a summary table lists each variable alongside a brief description. These variables allow us to explore associations with heart disease and inform model training later in the analysis. Each variable's role and characteristics are essential for understanding the predictive capacity of our logistic regression model.

Table of all variables and description

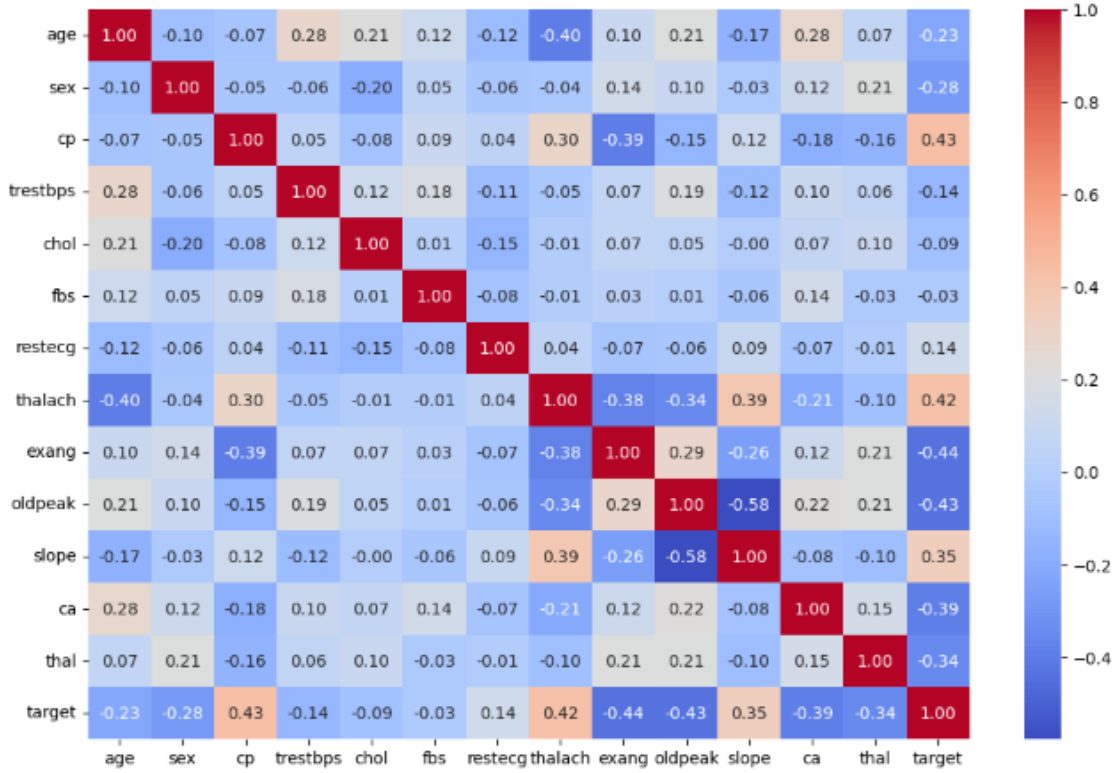
Variable Name	Description
age	Age of the patient (years)
sex	Gender of the patient (0 = female, 1 = male)
cp	Chest pain type (0-3)
trestbps	Resting blood pressure (mm Hg)
chol	Serum cholesterol (mg/dl)
fbs	Fasting blood sugar > 120 mg/dl (1 = true; 0 = false)
restecg	Resting electrocardiographic results (0-2)
thalach	Maximum heart rate achieved
exang	Exercise induced angina (1 = yes; 0 = no)

oldpeak	ST depression induced by exercise relative to rest
slope	Slope of the peak exercise ST segment (0-2)
ca	Number of major vessels (0-3) colored by fluoroscopy
thal	Thalassemia (1 = normal; 2 = fixed defect; 3 = reversable defect; 0 = unknown)
target	Diagnosis of heart disease (1 = presence; 0 = absence)

The dataset's variables each serve a unique function in understanding heart disease. For example, age is a straightforward numerical variable, as heart disease risk typically increases with age. Gender is categorical, where males and females may have different risk profiles. Cholesterol level is a continuous variable reflecting potential blockage in arteries, often associated with heart disease. Resting electrocardiogram results represent categorical data on electrical heart activity, crucial for detecting arrhythmias or other abnormalities. Through the analysis of these variables, we can better understand how each contributes to overall heart disease risk.

Exploratory Data Analysis (EDA)

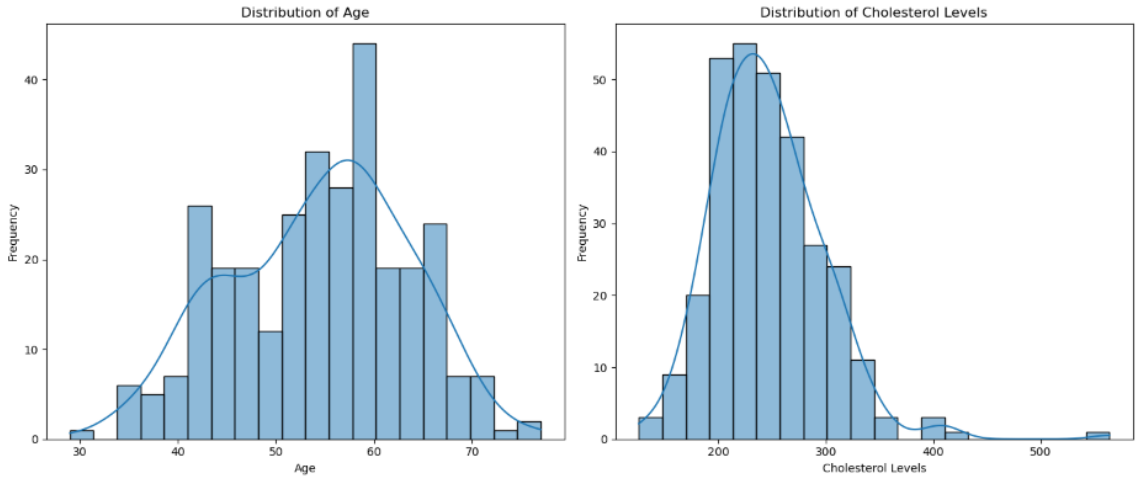
The exploratory data analysis stage involves identifying key patterns within the dataset. First, we conduct a correlation analysis, displaying the relationships among variables via a heatmap.



This heatmap helps identify strong correlations among features, including those related to cholesterol and other health indicators. Notably, key variables such as chest pain type, exercise-induced angina, oldpeak, and slope appear to play crucial roles in predicting heart disease. Understanding these correlations can guide the selection of significant features for predictive models, shedding light on which factors are most associated with heart disease risk in this dataset. For instance, "cp" (chest pain type) and "thalach" (maximum heart rate achieved) show a moderately strong negative correlation of -0.39, suggesting that individuals with higher chest pain levels may tend to have lower peak heart rates during exercise. Similarly, "ca" (the number of major vessels colored by fluoroscopy) shows moderate correlations with "thal" (around 0.51) and "exang" (0.46), suggesting some shared diagnostic or physiological relationships among these factors.

Some variables, such as "fbs" and "restecg" exhibit weak correlations with other features, implying they may have a more limited role in predicting heart disease risk in this dataset. These insights from the correlation analysis can inform feature selection for predictive modeling, focusing on variables that demonstrate stronger associations with the target outcome.

Next, we examine the distribution of age and cholesterol levels by creating histograms.

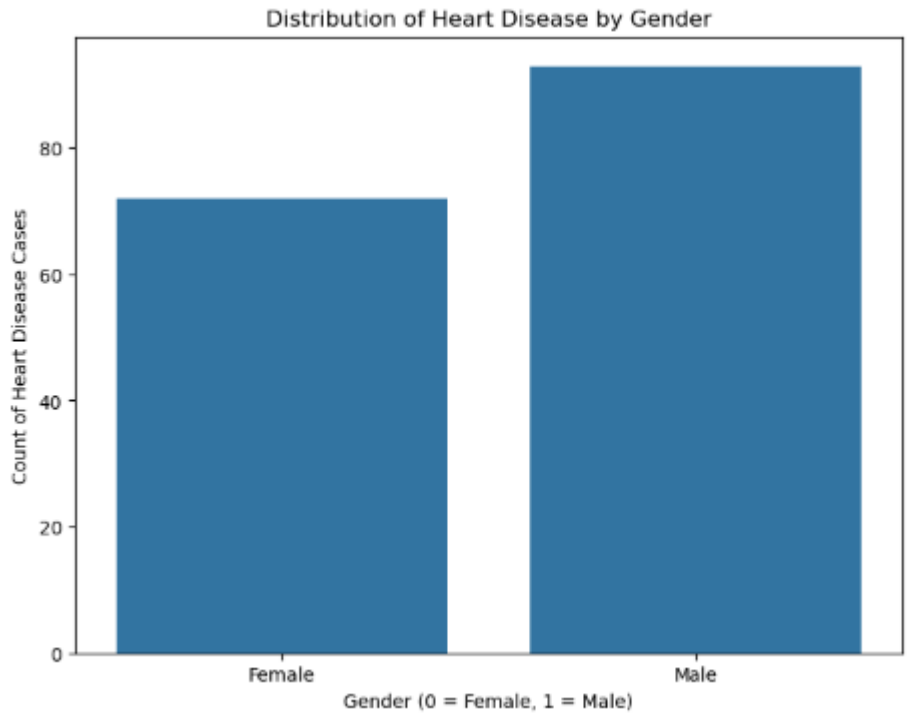


These visualizations reveal the prevalence of heart disease within specific age groups and cholesterol levels, showing which age ranges or cholesterol thresholds are more commonly affected.

The age distribution histogram displays the distribution of ages within the dataset, illustrating a fairly symmetrical, bell-shaped curve centred around the age range of 50 to 60 years. The highest frequency of individuals is in the age group around 60, indicating a higher prevalence of middle-aged and older adults in this dataset. The distribution shows that a large number of participants are between the ages of 40 and 70, with only a fewer individuals under 40 or over 70. This age range concentration is often relevant in heart disease studies, as heart disease risk generally increases with age. The smooth curve overlaying the histogram suggests that the age distribution approximates a normal distribution, although it appears slightly skewed towards older age groups. This age-related insight is important, as it may influence the heart disease prediction model, given that age is a known risk factor.

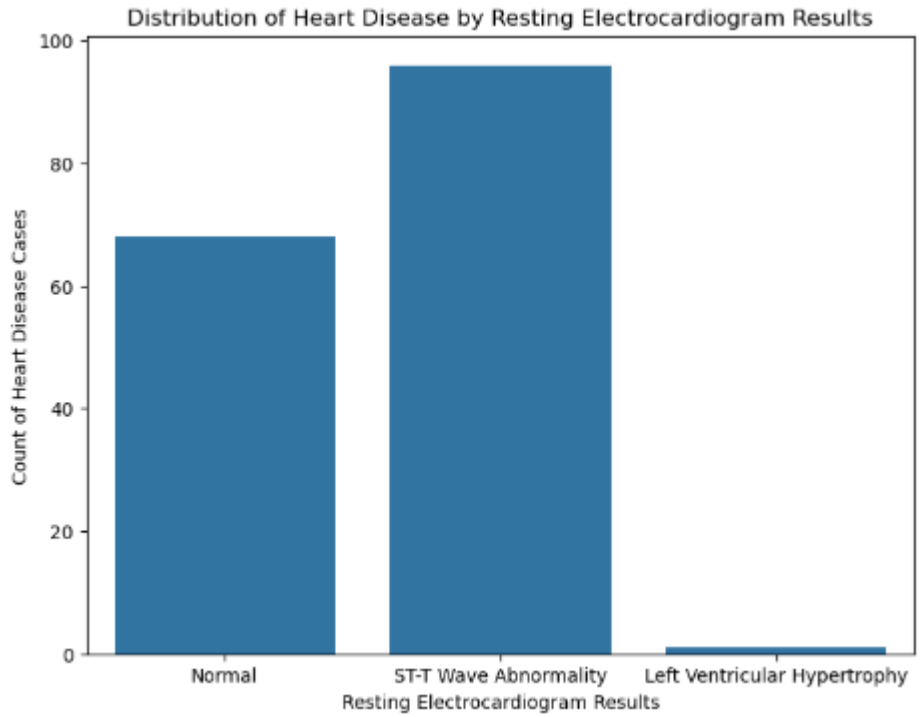
The cholesterol distribution histogram on the right displays the distribution of cholesterol levels, which appears positively skewed. A large concentration is observed at levels between 200 and 300 mg/dL, with a peak around 250 mg/dL. However, there are a few individuals with cholesterol levels exceeding 400 mg/dL, indicating high variability in cholesterol levels within the sample. Elevated cholesterol is a known risk factor for heart disease, and the skewed distribution suggests that some participants may be at a higher risk due to elevated cholesterol levels.

Following this, a bar plot illustrates the distribution of heart disease cases by gender, providing insights into gender-specific trends.



The analysis of heart disease distribution by gender reveals that men have a notably higher incidence of heart disease compared to women in this dataset. This trend aligns with established research indicating that men, especially as they age, are more prone to heart-related health issues than women. Factors such as lifestyle choices, including higher rates of smoking and physical inactivity among men, may contribute to this disparity. Additionally, physiological differences play a role; for instance, women often experience protective effects from hormones like estrogen before menopause, which can reduce certain cardiovascular risks.

Additionally, we explore the resting electrocardiogram results, comparing the frequency of heart disease across different types of ECG readings, highlighting any notable patterns.



The distribution of individuals with heart disease based on resting electrocardiogram (ECG) results reveals significant insights into cardiac health. Among the ECG findings, ST-T wave abnormalities are the most prevalent, indicating a higher risk of heart disease within this group. These abnormalities can reflect underlying issues. Following this category, the ECG results classified as normal account for a notable proportion of the cases, suggesting that while these individuals may present with heart disease, their resting ECG does not show any immediate signs of distress or dysfunction. This underscores the complexity of heart disease, as patients may still be at risk despite normal ECG readings.

Model Performance Evaluation

The logistic regression model developed for predicting heart disease has demonstrated a commendable accuracy of 87%, indicating its effectiveness in separating between individuals with and without heart disease based on the analysed input variables. The confusion matrix reveals that the model correctly predicted the absence of heart disease in 27 instances, while there were 2 false positives, where heart disease was incorrectly predicted. Additionally, the model had 6 false negatives, failing to identify heart disease in patients who actually had it, and successfully identified 26 true positives. This matrix highlights the model's strong capability to accurately classify the majority of cases.

A detailed classification report further elucidates the model's performance metrics. The precision for predicting heart disease (1) is 93%, meaning that when the model predicts heart disease, it is correct 93% of the time. Conversely, the precision for non-heart disease cases (0)

stands at 82%. Recall, which measures the model's ability to identify all relevant instances, is 81% for heart disease (1), indicating that the model correctly identifies 81% of actual positive cases, while recall for non-heart disease (0) is notably higher at 93%. The F1-score, representing the harmonic mean of precision and recall, reflects a strong balance between the two metrics, with both classes achieving an F1-score of approximately 87%.

These results underscore the robust performance of the logistic regression model in predicting heart disease, reinforcing the insights gained from exploratory data analysis and demonstrating its practical application in clinical settings. The balance between precision and recall indicates that the model not only achieves a high level of accuracy but also reliably identifies patients at risk of heart disease while minimizing false positives. This makes the model a valuable tool for early detection and intervention, potentially leading to improved patient outcomes in the realm of cardiovascular health.

Conclusion

In summary, this analysis of the heart disease dataset has provided valuable insights into the clinical and demographic factors associated with heart disease risk. By examining key variables such as age, gender, cholesterol levels, and resting electrocardiogram results, we have identified significant patterns and correlations that inform our understanding of heart health. The exploratory data analysis revealed that age and cholesterol levels play critical roles in predicting heart disease, with a clear age distribution indicating a higher prevalence among middle-aged and older individuals. Furthermore, the analysis of gender differences highlighted a marked disparity, with men exhibiting a higher incidence of heart disease compared to women, consistent with existing literature on cardiovascular health.

The correlation analysis further emphasized the importance of specific features, such as chest pain type and maximum heart rate achieved, in predicting heart disease outcomes. This informed the development of a logistic regression model, which serves as a powerful tool for predicting the likelihood of heart disease based on these variables. The model can help healthcare professionals identify at-risk patients more effectively and support early intervention strategies.

Overall, this report underscores the potential of leveraging machine learning techniques in the medical field to enhance predictive capabilities and improve patient outcomes. As heart disease remains a leading cause of mortality worldwide, ongoing research and analysis of relevant datasets are essential for advancing our understanding and management of this critical health issue.