Heart Disease Data Analysis and Prediction using ML

Aditi Bhatia
aditi.bhatia@ucdconnect.ie

Heart disease is one of the leading causes of death worldwide. The ability to predict heart disease early using clinical data can greatly improve patient outcomes. In this project, exploratory data analysis was done and machine learning techniques were employed to build models for heart disease prediction, aiming to identify which factors are most significant in determining whether a patient is at risk of heart disease.

Dataset: https://archive.ics.uci.edu/dataset/45/heart+disease

Data Overview

The dataset consists of 14 attributes, including:

- **Age**: Age in years.
- **Sex**: 0 = Female, 1 = Male.
- **Chest Pain Type (cp)**: Encoded as 1 (Typical Angina), 2 (Atypical Angina), 3 (Non-Anginal Pain), and 4 (Asymptomatic).
- **Cholesterol (chol)**: Serum cholesterol in mg/dl.
- **Fasting Blood Sugar (fbs)**: 1 if fasting blood sugar > 120 mg/dl, 0 otherwise.
- **Thalassemia (thal)**: 3 = Normal, 6 = Fixed Defect, 7 = Reversible Defect.
- **Diagnosis (num)**: 1 = Presence of heart disease, 0 = No heart disease.

Data Cleaning

Handling Missing Values

The dataset had some missing values in fields such as the ca (number of major vessels coloured by fluoroscopy) and thalassemia columns. These missing values were addressed using the following approaches:

- For categorical features like thalassemia, missing values were replaced with the most frequent category to minimise data distortion.
- For numerical features with missing values, such as ca, the median value was imputed to reduce the impact of extreme values.

Categorical Encoding

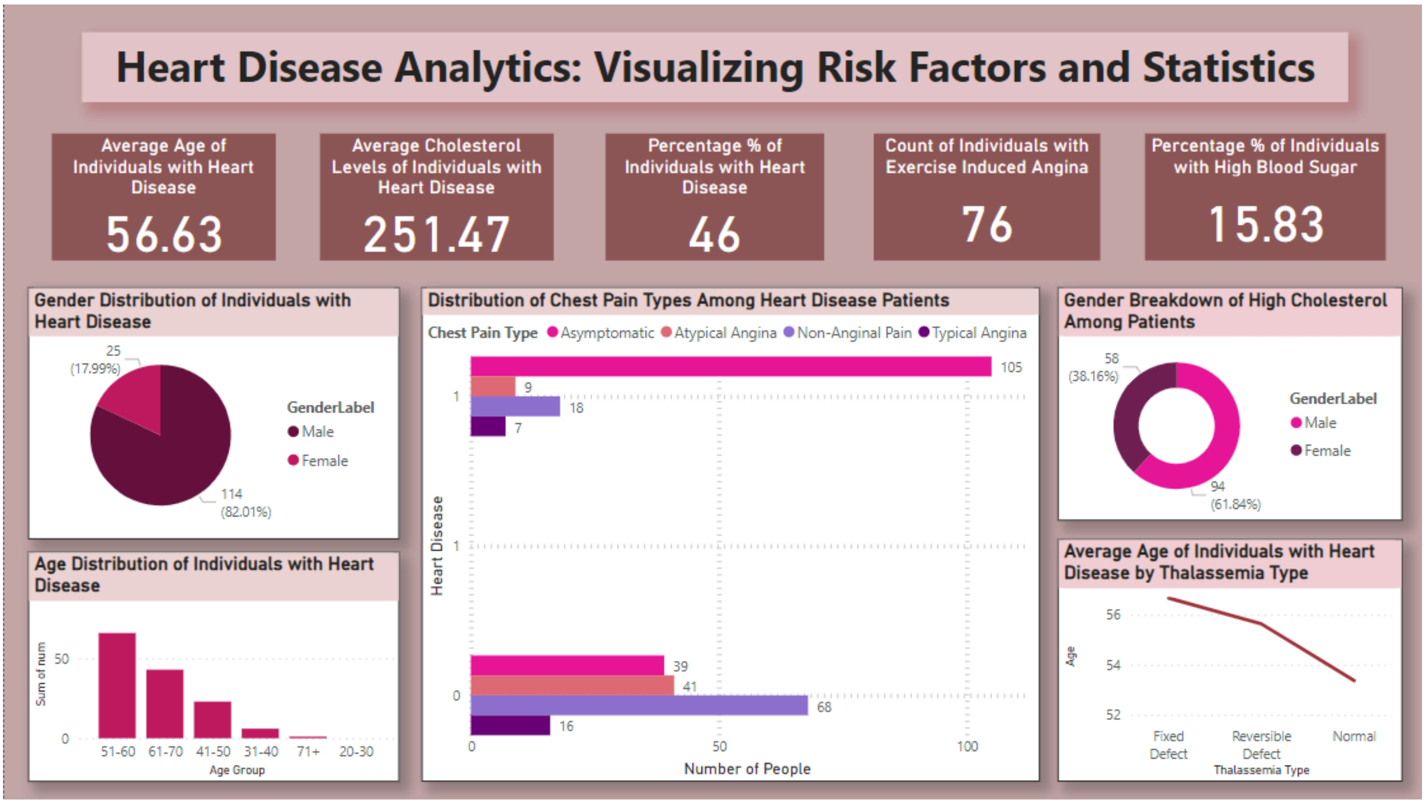
Several attributes in the dataset were categorical but encoded as numerical values:

- **Sex**: Encoded as 0 (Female) and 1 (Male). These were relabeled with their corresponding gender names for clearer visualisation in the dashboard.
- **Chest Pain Type (cp)**: Encoded as 1, 2, 3, and 4, representing different types of chest pain. These were relabeled to provide meaningful descriptions such as "Typical Angina" and "Atypical Angina."
- **Thalassemia (thal)**: Encoded as 3 (Normal), 6 (Fixed Defect), and 7 (Reversible Defect), and appropriately labelled to aid interpretation.

Inferences

From the visualisations and data analysis, several important inferences were drawn:

- **Age**: Heart disease is most prevalent among individuals aged 50-60.
- **Gender**: Men are more likely to suffer from heart disease than women.
- **Chest Pain**: Typical angina is the most common symptom among individuals with heart disease.
- **Blood Sugar**: A significant portion of heart disease patients also have elevated fasting blood sugar, indicating a relationship between diabetes and heart disease.
- **Thalassemia**: Individuals with thalassemia, particularly with fixed or reversible defects, are at a higher risk of developing heart disease as they age.



Heart Disease Analytics: Visualizing Risk Factors and Statistics

Average Age of Individuals with Heart Disease: 56.63

Average Cholesterol Levels of Individuals with Heart Disease: 251.47

Percentage % of Individuals with Heart Disease: 46

Count of Individuals with Exercise Induced Angina: 76

Percentage % of Individuals with High Blood Sugar: 15.83

Gender Distribution of Individuals with Heart Disease

Distribution of Chest Pain Types Among Heart Disease Patients

Chest Pain Type ● Asymptomatic ● Atypical Angina ● Non-Anginal Pain ● Typical Angina
105
9
18
7
39
41
68
16
1
1
0
Heart Disease
0
50
100
Number of People

Gender Breakdown of High Cholesterol Among Patients

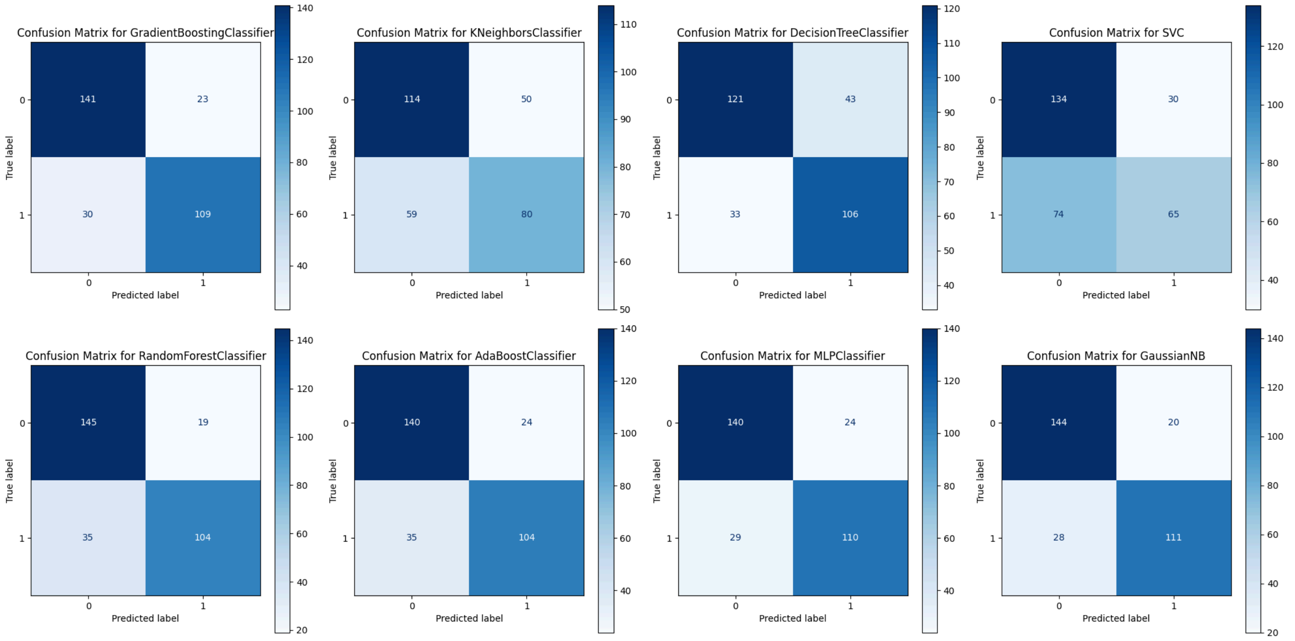
Age Distribution of Individuals with Heart Disease

Average Age of Individuals with Heart Disease by Thalassemia Type

To explore the predictive power of machine learning models on heart disease data, a range of algorithms was selected, including both simple models like **Gaussian Naive Bayes** and more complex ensemble methods such as **Gradient Boosting**.

```
                     Model  F1 Score  Accuracy   ROC AUC  Precision
                GaussianNB  0.841123  0.841584  0.838305   0.841851
             MLPClassifier  0.824794  0.825083  0.822513   0.824959
GradientBoostingClassifier  0.824651  0.825083  0.821964   0.825110
    RandomForestClassifier  0.820485  0.821782  0.816174   0.823893
        AdaBoostClassifier  0.804429  0.805281  0.800930   0.805734
    DecisionTreeClassifier  0.749586  0.749175  0.750197   0.751627
                       SVC  0.644795  0.656766  0.642350   0.662571
      KNeighborsClassifier  0.639050  0.640264  0.635331   0.638970
```

The **Gaussian Naive Bayes** model performed the best overall, achieving the highest F1 score of 0.841 and accuracy of 84.16%, making it the most effective model for predicting heart disease in this dataset.



Confusion Matrix for GradientBoostingClassifier

Confusion Matrix for KNeighborsClassifier

Confusion Matrix for DecisionTreeClassifier

Confusion Matrix for SVC

Confusion Matrix for RandomForestClassifier

Confusion Matrix for AdaBoostClassifier

Confusion Matrix for MLPClassifier

Confusion Matrix for GaussianNB