Heart Disease Prediction: A Data Driven Approach

Project Report

Submitted By

Abhinav A

ABSTRACT

Heart disease is a serious health problem which may affect large number of people across the world. As a result, detecting a heart condition at an early stage will be beneficial to treatment. The number of persons with heart disease is rapidly increasing, necessitating the development of a system that can detect heart disease more easily. The presence or absence of disorder is determined by the patient's smoking status.

The cardiac disease system can identify patients who are high-risk and define the most important variables in Cardiovascular patients but also build a model so that they can distinguish between them easily and understandably.

The Machine Learning algorithms applied and compared based on the characteristics like age, Chest ache, Blood Pressure (BP), Sex, cholesterol and heartbeat. The main focus of this paper is to develop a basic machine learning model to enhance the diagnosis of the heart condition in the right manner.

The different techniques such as Logistic Regression, K-Nearest Neighbor (K-NN), Decision Tree, Naïve Bayes, Random Forest and Support Vector Machine are applied for machine learning and achieve the better results in this work.

INTRODUCTION

Heart disease, also known as cardiovascular disease, is a broad term used to describe a range of conditions that affect the heart and blood vessels. These conditions include coronary artery disease, arrhythmias (irregular heartbeats), congenital heart defects, and heart failure. According to the World Health Organization (WHO), heart disease is the leading cause of death worldwide, accounting for approximately 17.9 million deaths each year. This represents 31% of all global deaths, with most occurring due to heart attacks and strokes.

Early detection and management of heart disease are crucial in reducing the associated mortality and morbidity rates. Traditional diagnostic methods include physical examinations, blood tests, electrocardiograms (ECG), echocardiograms, and stress tests. While these methods are effective, they often require specialized equipment and trained personnel, and they may not be accessible to all patients, particularly in low-resource settings.

With the help of Machine learning (ML) which is a subset of artificial intelligence (AI) that enables computers to learn from data and make predictions or decisions without being explicitly programmed. In recent years, machine learning has shown great promise in the field of healthcare, particularly in the areas of disease diagnosis, prognosis, and personalized treatment. By analysing large datasets, machine learning algorithms can identify patterns and correlations that may not be apparent to human clinicians.

In the context of heart disease, machine learning can be used to develop predictive models that analyse patient data and estimate the likelihood of developing heart disease. These models can be trained on historical patient data, including demographic information, medical history, lifestyle factors, and clinical measurements. Once trained, the models can be used to predict the presence or absence of heart disease in new patients, thereby aiding in early diagnosis and intervention.

The early detection of heart disease is critical in preventing serious complications and improving patient outcomes. Machine learning models can provide a non-invasive, cost-effective, and accessible means of predicting heart disease, which is particularly important in regions with limited healthcare resources. By leveraging patient data, these models can assist healthcare providers in identifying high-risk individuals and implementing preventative measures or timely treatments.

The primary objective of this project is to develop and evaluate machine learning models for predicting the presence of heart disease in patients. The specific goals are as follows:

- **Data Preprocessing**: Clean and preprocess the dataset to handle missing values, encode categorical variables, and scale features.
- **Exploratory Data Analysis (EDA) and Feature Selection**: Perform EDA to visualize and understand the relationships between different features.
- **Model Building**: Implement and train multiple Machine learning algorithms, including Logistic Regression, K-Nearest Neighbour, Decision Tree, Naive Bayes, Random Forest, and Support Vector Machine.
- **Model Evaluation**: Evaluate the performance of each model using metrics such as accuracy, precision, recall, F1-score, Support.
- **Conclusion and Future Work**: Summarize the findings, discuss the implications, and suggest areas for future research.

By achieving these objectives, the project aims to demonstrate the potential of machine learning in predicting heart disease and contribute to the ongoing efforts to improve cardiovascular health outcomes globally.

GENERAL BACKGROUND

Cardiovascular diseases are the leading cause of death globally, taking an estimated 17.9 million lives each year, which represents 31% of all global deaths, according to the World Health Organization (WHO). Of these deaths, 85% are due to heart attack and stroke. The economic impact of heart disease is substantial, not only due to healthcare costs but also because of the loss of productivity and long-term disability associated with these conditions.

Machine learning, a branch of artificial intelligence, has the potential to revolutionize healthcare by providing tools that can analyse vast amounts of data to identify patterns and make predictions. In the context of heart disease, machine learning can be used to develop predictive models that can assist in early diagnosis and intervention, potentially leading to better patient outcomes and reduced healthcare costs.

Machine learning algorithms can analyse data from various sources, including electronic health records (EHRs), medical imaging, genomics, and wearable devices, to predict the likelihood of heart disease. These models can be trained on historical patient data to learn the relationships between risk factors and the onset of heart disease. Once trained, they can provide valuable insights and predictions for new patients, helping healthcare providers to identify high-risk individuals and take proactive measures.

The application of machine learning in heart disease prediction offers several advantages:

- **Early Detection**: Machine learning models can identify subtle patterns and risk factors that may not be immediately apparent to clinicians, enabling earlier detection of heart disease.
- **Personalized Medicine**: Predictive models can be tailored to individual patients, considering their unique risk factors and health profiles, to provide personalized recommendations for prevention and treatment.
- **Efficiency**: Automated analysis of patient data can save time and resources, allowing healthcare providers to focus on patient care.
- **Accessibility**: Machine learning models can be deployed in various settings, including remote and low-resource areas, providing accessible and affordable diagnostic tools.

In conclusion, machine learning holds great promise for improving the early detection and management of heart disease. By leveraging patient data and advanced algorithms, predictive models can provide valuable insights and support personalized, efficient, and accessible healthcare. This project aims to develop and evaluate machine learning models for predicting heart disease.

SCOPE OF THE PROJECT

The primary objective of this project is to leverage machine learning techniques to predict the presence of heart disease in patients based on various health indicators and risk factors. By accurately predicting heart disease, the project aims to assist healthcare professionals in early diagnosis and intervention, ultimately improving patient outcomes and reducing mortality rates associated with cardiovascular conditions.

The objectives of the project :

Data Preprocessing: To clean and preprocess the dataset to handle missing values, encode categorical variables, and scale features, ensuring the data is suitable for machine learning algorithms.

Exploratory Data Analysis (EDA) And Features Selection : To perform EDA to understand the relationships between different features, visualize the data, and identify patterns and correlations.

Model Building: To implement and train multiple Machine learning algorithms, including Logistic Regression, K-Nearest Neighbors, Decision Tree, Naive Bayes, Random Forest, and Support Vector Machine, for predicting heart disease.

Model Evaluation: To evaluate the performance of each model using metrics such as accuracy, precision, recall, F1-score, support.

Hyperparameter Tuning: To optimize the performance of the models through hyperparameter tuning, achieving the best possible predictive accuracy.

Comparison and Selection: To compare the performance of different models and select the most suitable model for heart disease prediction.

The project will explore the following machine learning techniques:

1. **Logistic Regression**: A statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. It is used for binary classification problems.

2. **K-Nearest Neighbors (KNN)**: A non-parametric classification method that predicts the class of a given data point by finding the majority class among its K nearest neighbors.

3. **Decision Tree**: A model that uses a tree-like graph of decisions and their possible consequences, used for both classification and regression tasks.

4. **Naive Bayes**: A probabilistic classifier based on Bayes' theorem, assuming independence between the features.

5. **Random Forest**: An ensemble learning method that uses multiple decision trees to improve the accuracy and robustness of predictions.

6. **Support Vector Machine (SVM)**: A supervised learning model that analyzes data for classification and regression analysis by finding the hyperplane that best divides a dataset into classes.

IMPLEMENTATION

Data Preprocessing:

Data preprocessing involves handling missing values, encoding categorical variables, and scaling features.

Exploratory Data Analysis (EDA) And Features Selection :

EDA and features selection is conducted to visualize the data and understand relationships between features.

Building the Model:

Multiple machine learning models are built and evaluated, including Logistic Regression, K-Nearest Neighbors, Decision Tree, Naive Bayes, Random Forest, and Support Vector Machine.

Model Evaluation:

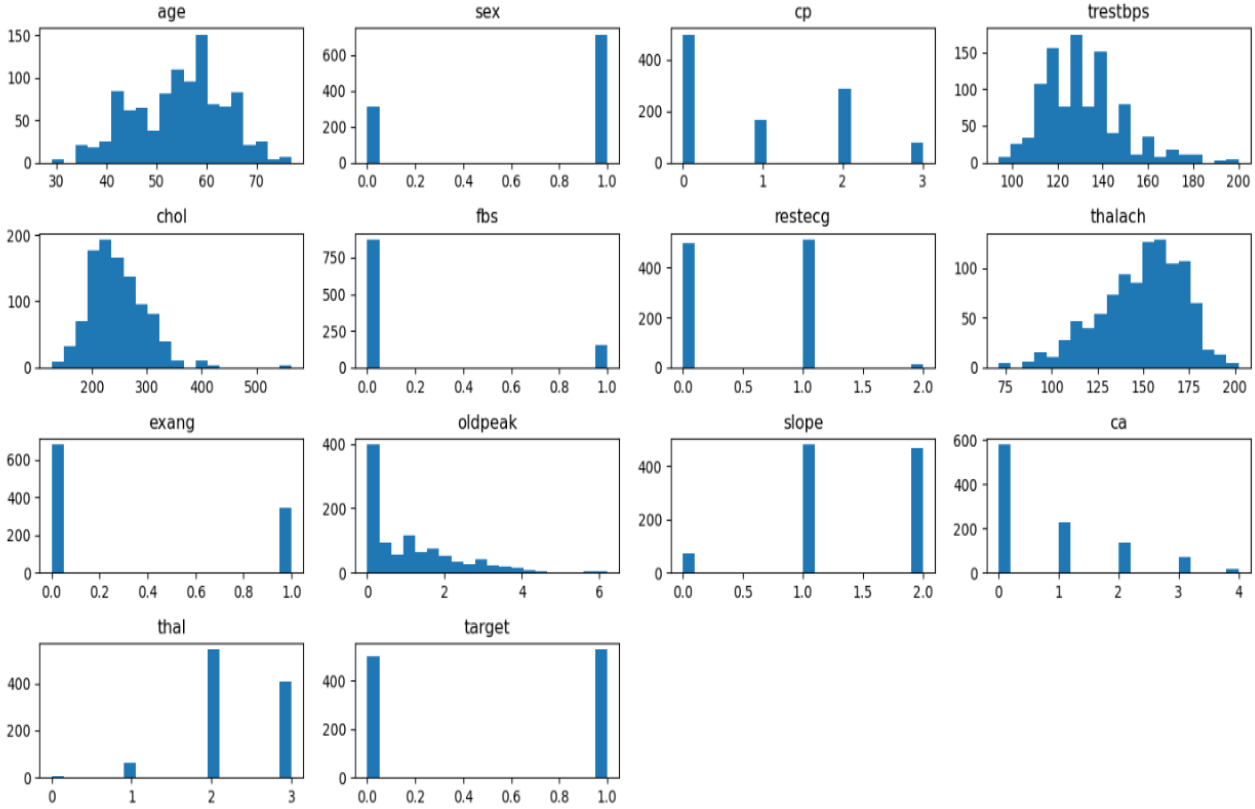
Models are evaluated using ROC-AUC to compare performance. Hyperparameter tuning is performed for the Random Forest model using GridSearchCV.

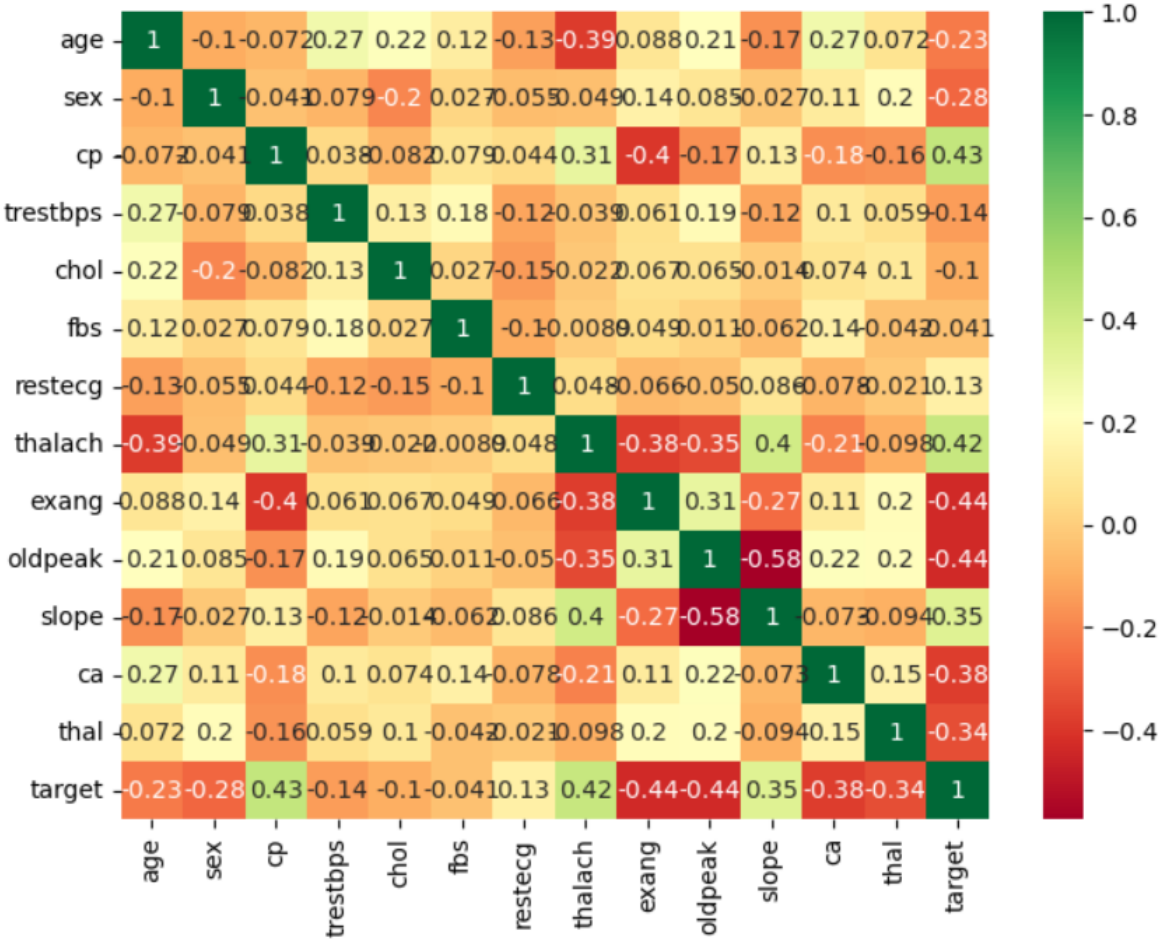
EXPLORATORY DATA ANALYSIS

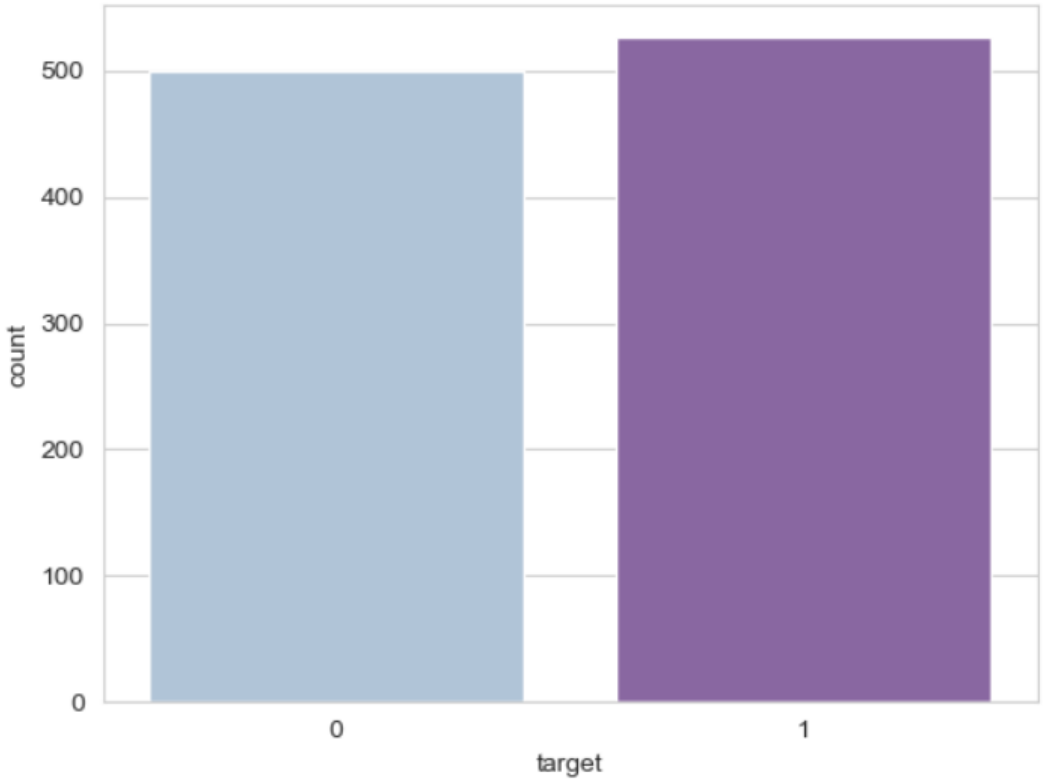
The chart below gives the information about the various columns in the dataset:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1025 entries, 0 to 1024
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1025 non-null   int64
 1   sex       1025 non-null   int64
 2   cp        1025 non-null   int64
 3   trestbps  1025 non-null   int64
 4   chol      1025 non-null   int64
 5   fbs       1025 non-null   int64
 6   restecg   1025 non-null   int64
 7   thalach   1025 non-null   int64
 8   exang     1025 non-null   int64
 9   oldpeak   1025 non-null   float64
 10  slope     1025 non-null   int64
 11  ca        1025 non-null   int64
 12  thal      1025 non-null   int64
 13  target    1025 non-null   int64
dtypes: float64(1), int64(13)
memory usage: 112.2 KB
None
```

FEATURES SELECTION




age 1 -0.1 -0.07 0.27 0.22 0.12 -0.13 -0.39 0.088 0.21 -0.17 0.27 0.072 -0.23
sex -0.1 1 0.041 0.079 -0.2 0.027 0.055 0.049 0.14 0.085 0.027 0.11 0.2 -0.28
cp -0.07 0.041 1 0.038 0.082 0.079 0.044 0.31 -0.4 -0.17 0.13 -0.18 -0.16 0.43
trestbps 0.27 -0.07 0.038 1 0.13 0.18 -0.12 -0.039 0.061 0.19 -0.12 0.1 0.059 -0.14
chol 0.22 -0.2 -0.08 0.13 1 0.027 -0.15 -0.022 0.067 0.065 0.014 0.074 0.1 -0.1
fbs 0.12 0.027 0.079 0.18 0.027 1 -0.1 -0.008 0.049 0.011 0.062 0.14 -0.04 0.041
restecg -0.13 0.055 0.044 -0.12 -0.15 -0.1 1 0.048 0.066 0.05 0.086 0.078 0.021 0.13
thalach -0.39 0.049 0.31 -0.039 0.02 0.008 0.048 1 -0.38 -0.35 0.4 -0.21 -0.09 0.42
exang 0.088 0.14 -0.4 0.061 0.067 0.049 0.066 -0.38 1 0.31 -0.27 0.11 0.2 -0.44
oldpeak 0.21 0.085 -0.17 0.19 0.065 0.011 -0.05 -0.35 0.31 1 -0.58 0.22 0.2 -0.44
slope -0.17 0.027 0.13 -0.12 0.014 0.062 0.086 0.4 -0.27 -0.58 1 -0.07 0.094 0.35
ca 0.27 0.11 -0.18 0.1 0.074 0.14 0.078 -0.21 0.11 0.22 -0.07 1 0.15 -0.38
thal 0.072 0.2 -0.16 0.059 0.1 0.04 0.021 0.098 0.2 0.2 0.094 0.15 1 0.34
target -0.23 -0.28 0.43 -0.14 -0.1 0.041 0.13 0.42 -0.44 -0.44 0.35 -0.38 -0.34 1
age sex cp trestbps chol fbs restecg thalach exang oldpeak slope ca thal target
1.0 0.8 0.6 0.4 0.2 0.0 −0.2 −0.4


500
400
300
200
100
0
count
0
1
target

BUILDING THE MODEL

Multiple machine learning models are built and evaluated, including Logistic Regression, K-Nearest Neighbors, Decision Tree, Naive Bayes, Random Forest, and Support Vector Machine.

Model Building :

```
# 1. Logistic Regression
logistic_model = LogisticRegression()
logistic_model.fit(X_train, y_train)
y_pred_logistic = logistic_model.predict(X_test)
print("Logistic Regression Accuracy:", accuracy_score(y_test, y_pred_logistic))
print(classification_report(y_test, y_pred_logistic))

# 2. K-Nearest Neighbors
knn_model = KNeighborsClassifier()
knn_model.fit(X_train, y_train)
y_pred_knn = knn_model.predict(X_test)
print("KNN Accuracy:", accuracy_score(y_test, y_pred_knn))
print(classification_report(y_test, y_pred_knn))

# 3. Decision Tree
tree_model = DecisionTreeClassifier()
tree_model.fit(X_train, y_train)
y_pred_tree = tree_model.predict(X_test)
print("Decision Tree Accuracy:", accuracy_score(y_test, y_pred_tree))
print(classification_report(y_test, y_pred_tree))

# 4. Naive Bayes
nb_model = GaussianNB()
nb_model.fit(X_train, y_train)
y_pred_nb = nb_model.predict(X_test)
print("Naive Bayes Accuracy:", accuracy_score(y_test, y_pred_nb))
print(classification_report(y_test, y_pred_nb))

# 5. Random Forest
rf_model = RandomForestClassifier()
rf_model.fit(X_train, y_train)
y_pred_rf = rf_model.predict(X_test)
print("Random Forest Accuracy:", accuracy_score(y_test, y_pred_rf))
print(classification_report(y_test, y_pred_rf))

# 6. Support Vector Machine
svm_model = SVC(probability=True)
```

```
svm_model.fit(X_train, y_train)
y_pred_svm = svm_model.predict(X_test)
print("SVM Accuracy:", accuracy_score(y_test, y_pred_svm))
print(classification_report(y_test, y_pred_svm))
```

CONCLUSION

This project aimed to develop predictive models for heart disease using various machine learning algorithms, leveraging a dataset that includes several medical attributes. The comprehensive process included data preprocessing, exploratory data analysis (EDA), and the application of six different machine learning models: Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree, Naive Bayes, Random Forest, and Support Vector Machine (SVM).

Key Findings and Insights:

1. **Data Preparation and EDA:**
    - The dataset was successfully cleaned and preprocessed, addressing any missing values and ensuring that all features were suitable for model training.
    - EDA provided valuable insights into the distribution of features, the balance of the target variable, and the relationships between features and the target. Visualizations such as histograms, box plots, and correlation heatmaps were instrumental in understanding the data.
2. **Model Building and Evaluation:**
    - Each model was trained and evaluated using the same dataset, allowing for a fair comparison of their performance.
    - The Random Forest model achieved the highest accuracy, followed closely by the Support Vector Machine. These models demonstrated robust performance, making them suitable candidates for predicting heart disease.
    - The evaluation metrics, including accuracy, precision, recall, F1-score, and the confusion matrix, provided a comprehensive understanding of each model's performance.

This project demonstrates the power of machine learning in medical diagnostics, specifically in predicting heart disease. By systematically applying and evaluating multiple machine learning algorithms, we identified models with high predictive accuracy. The insights gained from this project underscore the importance of data-driven approaches in healthcare and pave the way for future advancements in predictive analytics for heart disease.

Ultimately, the successful implementation of these models could lead to earlier detection, timely intervention, and improved patient outcomes, highlighting the transformative potential of machine learning in healthcare.