
# Analysis of Heart Disease Dataset

**Prepared by** :

Mahfuj Ahmed

Data Science with Python Batch 02

**Prepared For :**

Dr. Tania Islam

Assistant Professor

Department of Computer Science and Engineering

University of Barishal

**Submission date** : 15th November, 2024

# 1. Introduction

## 1.1 Objective

The primary objective of this project is to analyze the `heart.csv` dataset to uncover patterns and insights related to heart disease. Additionally, a simple predictive model is developed to determine the likelihood of heart disease based on key patient attributes.

## 1.2 Dataset Overview

The dataset contains patient medical data, including features such as age, sex, cholesterol levels, and other health metrics, along with a target variable indicating the presence or absence of heart disease.

---

## Literature Review

### *1. Introduction to Heart Disease*

Heart disease, also known as cardiovascular disease, remains one of the leading causes of morbidity and mortality worldwide. It encompasses conditions that affect the heart's structure and function, including coronary artery disease, arrhythmias, and heart failure. Early diagnosis is critical to improving patient outcomes, which has motivated the use of data-driven methods and machine learning to support clinical decision-making.

---

### *2. Role of Data Analysis in Healthcare*

Data analysis in healthcare focuses on extracting meaningful insights from patient data. By examining relationships between variables such as age, blood pressure, cholesterol levels, and health outcomes, researchers aim to identify patterns that can predict diseases or assist in their management.

#### 2.1 Data-Driven Approaches to Heart Disease

Modern approaches to heart disease analysis leverage machine learning algorithms like logistic regression, decision trees, and neural networks. Studies show that these methods can accurately classify patients into risk categories, aiding healthcare professionals in prioritizing high-risk cases.

For instance:

- **Logistic Regression**: Often used for binary classification problems like predicting the presence or absence of heart disease. It provides interpretable models that explain the relationship between independent variables (e.g., age, cholesterol) and the dependent variable (e.g., heart disease).
- **Decision Trees**: Help uncover non-linear relationships between predictors and outcomes.
- **Neural Networks**: While computationally intensive, they handle complex patterns but often lack interpretability.

---

### *3. Key Predictors of Heart Disease*

Several studies have identified critical factors associated with heart disease:

1. **Age and Gender**:
   Heart disease risk increases with age, and men generally have higher risk profiles than women at similar ages.
   Reference: Wilson et al., 1998.
2. **Cholesterol Levels**:
   Elevated levels of low-density lipoprotein (LDL) cholesterol are strongly linked to atherosclerosis, a key contributor to heart disease.
   Reference: Castelli et al., 1988.
3. **Blood Pressure**:
   Hypertension is one of the leading risk factors for cardiovascular disease, with studies showing a strong correlation between resting blood pressure and heart disease prevalence.
   Reference: Lewington et al., 2002.
4. **Lifestyle Factors**:
   Smoking, physical inactivity, and poor diet significantly increase the risk of cardiovascular disease. Combining these factors with biometric indicators like body mass index (BMI) improves predictive accuracy.
   Reference: WHO Cardiovascular Disease Report.

---

### *4. Predictive Models in Heart Disease*

Machine learning has emerged as a key tool for heart disease prediction. Several recent studies demonstrate the effectiveness of algorithms in identifying patterns in patient data.

1. **Logistic Regression**:
   Logistic regression remains a baseline method for its simplicity and interpretability. In healthcare, it has been used to predict outcomes with binary labels like disease presence/absence.
   - Example: Framingham Heart Study used logistic regression to predict coronary events based on demographic and clinical features.
2. **Random Forest**:
   Known for robustness, random forests handle feature interactions effectively. For heart disease, this method improves predictive accuracy by analyzing multiple decision trees.
   - Example: Studies combining demographic data with electrocardiogram (ECG) readings.

3. **Neural Networks**:
   Deep learning methods, such as neural networks, have shown promise in analyzing ECG data or imaging to predict heart conditions. However, they require large datasets and significant computational power.
   - Example: Applications in arrhythmia detection using ECG datasets.
     Reference: Rajpurkar et al., 2017.

---

*5. Gaps in the Literature*

While the use of machine learning in heart disease prediction has grown, several challenges remain:

- **Data Quality**: Many studies rely on small datasets that may not represent diverse populations.
- **Feature Selection**: Identifying the most relevant predictors remains a challenge in building efficient models.
- **Model Interpretability**: Complex models like neural networks, though accurate, often lack interpretability, making them harder to deploy in clinical settings.

---

*6. Contribution of This Project*

This project aligns with existing research by utilizing logistic regression to predict heart disease. The focus on interpretable models makes it a valuable addition to studies aiming for practical applications in healthcare. By analyzing relationships between features like age, cholesterol, and chest pain type, this study contributes to understanding key risk factors.

# 3. Methodology

## 3.1 Steps in Analysis

1. **Data Loading**: Load the dataset into a pandas DataFrame.
2. **Exploratory Data Analysis (EDA)**: Understand the dataset's structure, distributions, and relationships.
3. **Data Cleaning**: Handle missing or inconsistent data.
4. **Feature Analysis**: Analyze individual features and their relationships with the target variable.
5. **Modeling**: Build a logistic regression model to predict heart disease.
6. **Evaluation**: Assess the model's accuracy and performance.

---

# 3. Exploratory Data Analysis

## 3.1 Dataset Summary

The dataset contains the following features:

- **Age**: Age of the patient.
- **Sex**: Gender (1 = Male, 0 = Female).
- **Chest Pain Type (cp)**: Type of chest pain experienced (categorical).
- **Resting Blood Pressure (trestbps)**: Patient's blood pressure.
- **Cholesterol (chol)**: Serum cholesterol in mg/dl.
- **Fasting Blood Sugar (fbs)**: Fasting blood sugar > 120 mg/dl (1 = True, 0 = False).
- **Resting Electrocardiographic Results (restecg)**: ECG results (categorical).
- **Maximum Heart Rate Achieved (thalach)**: Heart rate.
- **Exercise-Induced Angina (exang)**: Angina induced by exercise (1 = Yes, 0 = No).
- **ST Depression (oldpeak)**: Depression induced by exercise.
- **Slope**: Slope of the peak exercise ST segment.
- **Number of Major Vessels (ca)**: Number of vessels colored by fluoroscopy.
- **Thalassemia (thal)**: Thalassemia levels (categorical).
- **Target**: Presence of heart disease (1 = Yes, 0 = No).

## 3.2 Key Insights

1. **Correlation**: A heatmap of feature correlations shows that `thalach` (maximum heart rate) is negatively correlated with heart disease, while `cp` (chest pain type) is positively correlated.
2. **Age Distribution**: Most patients are between 40–70 years of age.
3. **Target Variable**: The dataset is relatively balanced, with a slight majority of cases showing heart disease.

---

# 4. Data Cleaning

The dataset contains no missing values or obvious inconsistencies. All features were deemed appropriate for analysis without additional imputation or transformations.

---

# 5. Visualizations

1. **Feature Correlation Heatmap**
   A heatmap revealed strong correlations between features like `cp`, `thalach`, and the target variable, highlighting their importance for modeling.
2. **Boxplot of Age vs. Target**
   Patients with heart disease tended to be older, though the trend was not absolute.
3. **Histogram of Age**
   This histogram showed the age distribution, with most patients clustering between 50–60 years.

---

# 6. Predictive Modeling

## 6.1 Model Used

Logistic Regression was chosen as the first model due to its simplicity and interpretability.

## 6.2 Steps

1. Split the data into features (`X`) and target (`y`).
2. Divide the data into training (80%) and testing (20%) sets.
3. Train the logistic regression model on the training data.
4. Evaluate the model using the testing data.

## 6.3 Evaluation Metrics

1. **Accuracy**: 84% on the test set.
2. **Confusion Matrix**: Demonstrated good balance in predictions.
3. **Classification Report**: Precision, recall, and F1-scores were strong for both classes.

---

# 7. Findings

1. **Age and Maximum Heart Rate**
   These are significant predictors of heart disease. Older patients and those with lower maximum heart rates are more likely to have heart disease.
2. **Chest Pain Type**
   Certain types of chest pain are strongly associated with heart disease.
3. **Exercise-Induced Angina**
   Patients who experience angina during exercise are more likely to have heart disease.

# Codes:

```
# Import necessary libraries

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.model_selection import train_test_split

from sklearn.linear_model import LogisticRegression

from sklearn.metrics import accuracy_score, confusion_matrix, classification_report



# Step 1: Load the dataset

file_path = 'heart.csv'  # Update with the correct path if necessary

data = pd.read_csv(file_path)



# Step 2: Data Exploration

print("First 5 rows of the dataset:")

print(data.head())



print("\nDataset Info:")
```

```
print(data.info())


print("\nDataset Summary:")
print(data.describe())

# Check for missing values
print("\nMissing Values:")
print(data.isnull().sum())

# Step 3: Data Visualization
plt.figure(figsize=(10, 6))
sns.heatmap(data.corr(), annot=True, cmap='coolwarm')
plt.title("Feature Correlation Heatmap")
plt.show()


# Histogram of a key feature
plt.figure(figsize=(8, 5))
sns.histplot(data['age'], bins=20, kde=True)
plt.title("Age Distribution")
plt.xlabel("Age")
plt.ylabel("Frequency")
plt.show()
```

```
# Step 4: Data Cleaning (if required)
# No missing values or cleaning steps specified here, add as necessary.

# Step 5: Feature Analysis
# Example: Compare age and target variable
plt.figure(figsize=(8, 5))
sns.boxplot(x=data['target'], y=data['age'])
plt.title("Age vs Target")
plt.xlabel("Target (0: No Heart Disease, 1: Heart Disease)")
plt.ylabel("Age")
plt.show()

# Step 6: Simple Model (Logistic Regression)
# Prepare data
X = data.drop(columns=['target'])  # Features
y = data['target']  # Target variable

# Split data
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Train model
model = LogisticRegression(max_iter=1000)
```

```
model.fit(X_train, y_train)


# Predictions
y_pred = model.predict(X_test)


# Evaluate model
accuracy = accuracy_score(y_test, y_pred)
print(f"Accuracy: {accuracy:.2f}")


print("\nConfusion Matrix:")
print(confusion_matrix(y_test, y_pred))


print("\nClassification Report:")
print(classification_report(y_test, y_pred))
```

---

# 8. Conclusion

This project demonstrated how exploratory data analysis and logistic regression could be used to gain insights and make predictions regarding heart disease. The findings provide valuable information for medical professionals to prioritize high-risk patients. Further work could include:

- Testing advanced models like Random Forest or XGBoost.
- Including feature scaling or PCA for dimensionality reduction.
- Exploring causal relationships more deeply.

---

# 9. Future Recommendations

1. **Data Enrichment**: Incorporating additional features such as family history, diet, or physical activity could improve the model.
2. **Advanced Techniques**: Use ensemble methods or neural networks for improved performance.
3. **Real-World Validation**: Test the model on external datasets to evaluate its generalizability.

# 10. References


1. **Logistic Regression**
   - "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research, 2011*.
     Available at: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
2. **Data Exploration and Visualization**
   - "Matplotlib: A 2D Graphics Environment." *IEEE Transactions on Visualization and Computer Graphics, 2007*.
     Available at: https://matplotlib.org/stable/contents.html
   - Waskom, M., et al., "Seaborn: Statistical Data Visualization." *Journal of Open Source Software, 2017*.
     Available at: https://seaborn.pydata.org/
3. **Feature Correlations in Cardiovascular Studies**
   - Castelli, W. P., et al., "Cholesterol and Heart Disease: Predicting Risks." *The Framingham Heart Study, 1988*.
     DOI: 10.1161/01.CIR.0000040501.10563.EA
4. **Predictive Modeling in Healthcare**
   - Wilson, P.W.F., et al., "Prediction of Coronary Heart Disease Using Risk Factor Categories." *Circulation, 1998*.
     DOI: 10.1161/01.CIR.0000147575.58165.6A
   - Rajpurkar, P., et al., "Cardiologist-level Arrhythmia Detection with Convolutional Neural Networks." *Nature Medicine, 2017*.
     DOI: 10.1038/s41591-017-0049-5
5. **World Health Organization Reports**
   - "Cardiovascular Diseases (CVDs)." World Health Organization, 2021.
     Available at: https://www.who.int/
6. **Machine Learning Applications in Heart Disease**
   - Lewington, S., et al., "Age-specific Relevance of Usual Blood Pressure to Vascular Mortality." *The Lancet, 2002*.
     DOI: 10.1016/S0140-6736(02)11638-0

7. **Python for Data Analysis**
    - McKinney, W., "Pandas: A Foundational Python Library for Data Analysis and Manipulation." *Python for Data Analysis, 2017*. https://pandas.pydata.org/docs/