# Heart Disease Prediction and Analysis Using the Cleveland Dataset

Jinghui Yu

May 22, 2024

## 1 Main Objective

The primary objective of this analysis is to predict future trends, specifically to determine the presence of heart disease in patients based on various attributes. This falls under the category of prediction.

This analysis is highly beneficial to stakeholders in the healthcare industry. By accurately predicting the presence of heart disease, healthcare providers can implement more effective preventative measures, allocate resources more efficiently, and tailor treatment plans to individual patients. For instance, early identification of patients at risk of heart disease can lead to timely intervention, which is crucial in improving patient outcomes and reducing healthcare costs. Moreover, understanding which attributes are most indicative of heart disease presence can inform future research and development of diagnostic tools. This predictive model can thus significantly enhance patient care and operational efficiency within healthcare institutions.

## 2 Dataset Description

The dataset comprises 76 attributes, though research and experiments typically focus on a subset of 14 attributes. The dataset is sourced from the Cleveland database, which is predominantly used by machine learning researchers. The main goal of this dataset is to predict the presence of heart disease in patients, classified by an integer value ranging from 0 (no presence) to 4. Past experiments have primarily aimed to distinguish between the presence (values 1, 2, 3, 4) and absence (value 0) of heart disease.

To protect patient privacy, names and social security numbers have been replaced with dummy values. The database includes both processed and unprocessed files, with the Cleveland database being the processed one.

## 3 Data Preparation

### 3.1 Exploration

During the initial exploration of the Cleveland dataset, several key observations and potential issues were identified. The dataset contains 76 attributes, but our analysis focuses on the 14 attributes that are most relevant to predicting heart disease. Preliminary analysis revealed some interesting correlations between certain attributes and the presence of heart disease. For example, attributes like age, cholesterol levels, and maximum heart rate appeared to have a notable relationship with heart disease presence. However, data quality issues were also identified, such as missing values and potential outliers that could affect the accuracy of our predictions.

### 3.2 Cleaning

To address these data quality issues, several cleaning steps were undertaken:

- **Handling Missing Values:** Missing values in the dataset were imputed using median values for numerical attributes and the most frequent values for categorical attributes. This method was chosen to minimize bias and maintain the integrity of the data.
- **Removing Duplicates:** The dataset was scanned for duplicate records, which were then removed to ensure that each patient was only represented once.

- **Correcting Errors:** Any identified data entry errors, such as impossible values for attributes (e.g., negative values for age or cholesterol levels), were corrected or removed based on domain knowledge and reasonable assumptions.

### 3.3 Feature Engineering

To enhance the predictive power of our model, several feature engineering techniques were applied:

- **Transforming Categorical Variables:** Categorical variables were transformed into dummy variables using one-hot encoding. This allowed us to include categorical data in our predictive models without introducing bias.
- **Creating Interaction Terms:** Interaction terms were created to capture the combined effects of multiple attributes on the presence of heart disease. For instance, an interaction term between age and cholesterol levels was created to see if the combined effect of these two variables provided more predictive power than each variable individually.
- **Scaling Numerical Features:** Numerical features were scaled to a standard range to ensure that all attributes contributed equally to the model. This was particularly important for algorithms sensitive to the scale of input data, such as logistic regression and neural networks.

These steps ensured that the data was clean, well-prepared, and suitable for building robust predictive models for heart disease detection.

## 4 Model Training and Evaluation

### 4.1 Model Selection

In this analysis, I trained and evaluated three different classifiers to predict the presence of heart disease based on the Cleveland dataset. I began with a simple logistic regression model to establish a baseline for comparison. I then advanced to more complex models, specifically Random Forest and Gradient Boosting Machines, to leverage ensemble techniques for potentially improved performance.

- **Logistic Regression:** Chosen as the baseline model due to its simplicity and interpretability. It helps in understanding the relationship between the features and the target variable.
- **Random Forest:** An ensemble technique that builds multiple decision trees and merges them to get a more accurate and stable prediction. This model is known for handling a large number of features and reducing overfitting.
- **Gradient Boosting Machines (GBM):** Another ensemble method that builds models sequentially, each new model correcting errors made by the previous ones. GBMs are powerful predictors and often achieve high performance on complex datasets.

### 4.2 Evaluation

To ensure consistency in our model comparison, I used the same train-test split method. The dataset was split into 80% training data and 20% testing data. This approach allowed us to directly compare the performance of each model on the same testing set.

### 4.3 Metrics

I evaluated the models based on the following metrics: Accuracy, Precision, Recall and F1 Score, the results are as follow in the table:

| Model | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|
| Logistic Regression | 0.786885 | 0.729812 | 0.786885 | 0.757251 |
| Random Forest | 0.786885 | 0.678349 | 0.786885 | 0.728597 |
| Gradient Boosting | 0.754098 | 0.673302 | 0.754098 | 0.711414 |

# 5 Key Findings and Insights

The analysis identified several key drivers of heart disease and provided insights into their impact on the prediction models.

- **Main Drivers:** Factors such as age, cholesterol levels, maximum heart rate achieved, and the presence of exercise-induced angina were significant predictors in the models.
- **Model Insights:** Logistic Regression highlighted the linear relationships between these features and the likelihood of heart disease, providing clear insights into how each feature impacts the prediction.

# 6 Model Recommendation

After evaluating the models, I recommend the **Logistic Regression** model as the best fit for this project's needs.

The recommendation is based on the following considerations:

- **Performance Metrics:** Logistic Regression and Random Forest both achieved the highest accuracy at 0.786885. However, Logistic Regression had the highest precision and F1-score, indicating better performance in balancing precision and recall.
- **Model Complexity vs. Interpretability:** Logistic Regression, being the simplest and most interpretable model, allows for easier understanding and explanation of the relationships between features and the target variable. This is crucial for stakeholders who need to understand the model's decisions.

The trade-off between model complexity and interpretability is minimal here, as Logistic Regression performs comparably to the more complex Random Forest while providing greater ease of explanation.

# 7 Reference

Janosi,Andras, Steinbrunn,William, Pfisterer,Matthias, and Detrano,Robert. (1988). Heart Disease. UCI Machine Learning Repository. https://doi.org/10.24432/C52P4X.