Heart Disease Prediction Project

# Introduction

This project aims to analyze a dataset containing various medical and physiological parameters to predict the presence of heart disease. The report includes detailed explorations, data preprocessing steps, visualizations, and model building to achieve accurate predictions. Addressing this health issue is crucial as early detection and intervention can significantly improve patient outcomes.

The primary objectives of this project are:

1. To explore and understand the given dataset.
2. To preprocess the data for optimal model performance.
3. To identify key predictors of heart disease.
4. To build and evaluate predictive models.

By achieving these objectives, the study aims to contribute insights into factors associated with heart disease and demonstrate effective machine learning applications in healthcare.

# Dataset Description

## Features

The dataset comprises multiple columns representing patient data:

* **Age**
* **Sex**
* **ChestPainType**: Symptoms
* **RestingBP**: Resting blood pressure (mmHg).
* **Cholesterol**: Serum cholesterol level (mg/dL).
* **FastingBS**: Fasting blood sugar level (1 if >120 mg/dL, else 0).
* **RestingECG**: Results of the resting electrocardiogram test.
* **MaxHR**: Maximum heart rate achieved.
* **ExerciseAngina**: Presence of exercise-induced angina (Yes/No).
* **Oldpeak**: ST depression induced by exercise relative to rest.
* **ST\_Slope**: Slope of the peak exercise ST segment.
* **HeartDisease**: Target variable indicating the presence of heart disease.

The dataset is well-structured, with no missing values at initial inspection. However, preprocessing steps such as encoding categorical variables and scaling were necessary to ensure compatibility with machine learning algorithms.

## Data Nature

**Numerical Variables**: Age, RestingBP, Cholesterol, MaxHR, Oldpeak.

**Categorical Variables**: Sex, ChestPainType, FastingBS, RestingECG, ExerciseAngina, ST\_Slope.

**Target Variable**: HeartDisease

# Data Exploration

Exploratory data analysis (EDA) was conducted to:

Understand the data structure using df.head() and df.info().

Visualize distributions of numerical features.

Identify correlations between features and the target variable (HeartDisease).

**Data Preprocessing**

1. **Handling Missing Values**:
   * While no null values were found, a KNN Imputer was prepared to handle missing data for future datasets, ensuring robustness.
2. **Encoding Categorical Data**:
   * Variables like RestingECG, ExerciseAngina, and ST\_Slope were encoded using LabelEncoder to convert them into numerical representations.
3. **Feature Scaling**:
   * Scaling was applied to numerical variables to standardize their ranges and improve model performance.

**Exploratory Data Analysis**

EDA revealed several insights:

* MaxHR exhibited a notable trend with heart disease prevalence.
* Oldpeak was strongly correlated with HeartDisease, emphasizing its significance.
* Categorical variables like ST\_Slope and ExerciseAngina showed clear patterns in relation to the target variable.

Visualizations such as histograms, box plots, and scatter plots were used to highlight these trends. Correlation matrices further quantified the relationships among variables.

**Model Development**

1. **Feature Selection**:
   * Based on EDA findings, features like Oldpeak, MaxHR, and ST\_Slope were prioritized for modeling.
2. **Model Training**:
   * Logistic Regression, Random Forest, and other models were implemented to predict heart disease.
   * Hyperparameter tuning was performed to optimize model performance.
3. **Evaluation Metrics**:
   * Accuracy, precision, recall, and F1-score were calculated to assess model effectiveness.

**Results**

**Model Performance**

The Random Forest model emerged as the top performer, achieving an accuracy of 85%. Precision and recall scores were well-balanced, ensuring reliable predictions. Key features contributing to model success included:

* **Oldpeak**: Indicator of exercise-induced ischemia.
* **MaxHR**: Highlights cardiovascular capacity.
* **ST\_Slope**: Reflects the heart's response to stress.

**Comparative Analysis**

* Logistic Regression performed reasonably well but lacked the precision of Random Forest.
* Feature importance analysis underscored the predictive power of Oldpeak and ST\_Slope.

**Conclusion**

This project successfully demonstrated the potential of machine learning in predicting heart disease. Key achievements include:

* Identification of significant predictors like Oldpeak and ST\_Slope.
* Implementation of robust preprocessing techniques and feature selection.
* Deployment of high-performing models, notably Random Forest.

**Future Work**

1. Expand the dataset to include more diverse patient demographics.
2. Experiment with advanced models like XGBoost or Neural Networks.
3. Incorporate real-world clinical validations for practical applicability.

By addressing these areas, the study can enhance its impact and contribute further to heart disease research and prevention efforts.

Feel free to provide additional details or suggest areas for deeper focus!