Title: Heart Disease Diagnostic Analysis

Report

---

1. Introduction

Heart disease remains one of the leading causes of death worldwide. Early diagnosis and accurate prediction can significantly improve patient outcomes. This project focuses on analyzing a dataset related to heart disease and building a predictive model to diagnose the condition. The analysis aims to uncover patterns in the data and develop a model that can assist in early detection of heart disease.

---

2. Datasets Used

The dataset utilized for this analysis is Heart Disease data.csv. It contains various attributes related to heart disease diagnosis, including:

- **Age:** Patient's age.
- **Sex:** Gender of the patient.
- **cp:** Chest pain type.
- **trestbps:** Resting blood pressure.
- **chol:** Serum cholesterol level.
- **fbs:** Fasting blood sugar.
- **restecg:** Resting electrocardiographic results.
- **thalach:** Maximum heart rate achieved.
- **exang:** Exercise induced angina.
- **oldpeak:** Depression induced by exercise.
- **slope:** Slope of the peak exercise ST segment.
- **ca:** Number of major vessels colored by fluoroscopy.
- **thal:** Thalassemia type.
- **target:** Presence or absence of heart disease.

---

3. Objectives

The primary objectives of this project are:

- To explore and preprocess the dataset for analysis.
- To build and evaluate different classification models for heart disease prediction.
- To identify key features influencing heart disease.

- To visualize the data and model performance to provide actionable insights.

---

4. Methodology

The methodology followed in this project includes:

- **Data Collection:** Load and review the dataset to understand its structure and contents.
- **Data Cleaning:** Address missing values, duplicates, and ensure data consistency.
- **Feature Engineering:** Separate features and target variable, and apply transformations such as standardization and one-hot encoding.
- **Exploratory Data Analysis (EDA):** Conduct visual and statistical analysis to understand data distributions and relationships.
- **Model Training:** Train multiple classification models, including Logistic Regression, Decision Tree, and Random Forest.
- **Model Evaluation:** Assess model performance using metrics such as accuracy, confusion matrix, ROC-AUC score, and feature importance.
- **Hyperparameter Tuning:** Optimize the Random Forest model using Grid Search to improve performance.

---

5. Key Findings

- **Data Insights:** Key features such as age, cholesterol levels, and maximum heart rate significantly impact the prediction of heart disease.
- **Model Performance:** The Random Forest Classifier provided the best performance in terms of accuracy and ROC-AUC score compared to other models.
- **Feature Importance:** Certain features, including age and maximum heart rate, were identified as having the highest impact on model predictions.

---

6. Visualizations

- **Distribution of Heart Disease by Age:** A histogram showing the prevalence of heart disease across different age groups.
- **Heart Disease Rates by Gender:** A count plot illustrating the distribution of heart disease among different genders.
- **Correlation Heatmap:** A heatmap indicating the correlation between various features in the dataset.
- **Pairplot:** A pairplot showing relationships between pairs of features with respect to the target variable.
- **Confusion Matrix:** Visual representation of the model's performance in predicting heart disease.

- **ROC-AUC Curve:** A plot illustrating the model's ability to distinguish between positive and negative cases of heart disease.

---

7. Conclusion

The project successfully demonstrated the ability to predict heart disease using machine learning techniques. The Random Forest Classifier proved to be the most effective model, with strong performance metrics. The analysis highlighted important features and provided valuable visualizations to understand the dataset and model outcomes. Future work could explore additional models and feature engineering approaches to further enhance prediction accuracy.

---

8. Acknowledgments

I would like to express my gratitude to:

- **Data Providers:** For supplying the dataset used in this analysis.
- **Mentors and Colleagues:** For their guidance and support throughout the project.
- **Tools and Libraries:** The Python libraries, including pandas, scikit-learn, seaborn, and matplotlib, which were instrumental in data processing, visualization, and model development.