Contents

1. Title Page

- **Project Title:** Heart Disease Prediction using Machine Learning
- **Author:** Minhaj Asghar
- **Date:** 09/12/2024

---

2. Abstract

The project aims to predict heart disease in individuals based on medical data using machine learning algorithms. The dataset contains various medical attributes such as age, cholesterol levels, resting blood pressure, and more. Several machine learning models were implemented, including Logistic Regression, Random Forest, and Support Vector Machine, to evaluate their performance in predicting heart disease. The project also involves preprocessing steps such as handling missing data, encoding categorical variables, and scaling numerical features.

---

3. Introduction

Heart disease is one of the leading causes of death worldwide. Early detection can significantly improve treatment outcomes. This project seeks to predict whether a person is likely to have heart disease based on a set of medical features. By applying machine learning techniques, we aim to build an effective model that can assist in the early diagnosis of heart conditions.

---

4. Dataset Description

The dataset used in this project contains medical records from patients. Each row represents a patient with several attributes, and the target variable indicates the presence or absence of heart disease. The dataset includes:

- **Age:** Patient's age
- **Trestbps:** Resting blood pressure
- **Chol:** Serum cholesterol level
- **Thalach:** Maximum heart rate
- **Oldpeak:** Depression induced by exercise
- **Sex, Cp, Fbs, Restecg, Exang, Slope, Ca, Thal:** Various categorical features (such as sex, chest pain type, and exercise-induced angina)
- **Num:** Target variable (1 for heart disease, 0 for no disease)

---

5. Data Preprocessing

Before training the machine learning models, several preprocessing steps were applied to the dataset:

1. **Missing Value Handling:** The dataset initially contained missing values represented by '?', which were replaced with NaN and then imputed using mean (for numerical features) and most frequent (for categorical features).
2. **Feature Encoding:** Categorical features were encoded using One-Hot Encoding to transform them into numerical data that can be used by machine learning algorithms.
3. **Feature Scaling:** Numerical features such as age, blood pressure, and cholesterol were scaled using StandardScaler to normalize the data, ensuring that all features have a similar scale.

4. **Train-Test Split:** The dataset was split into 80% for training and 20% for testing using the train_test_split function, ensuring the models were trained and tested on separate data.

---

6. Exploratory Data Analysis (EDA)

Various visualizations were used to understand the dataset better:

- **Target Variable Distribution:** A count plot was used to show the distribution of the target variable, highlighting the balance (or imbalance) between the classes (heart disease present vs. not present).
- **Correlation Matrix:** A heatmap was generated to examine the correlation between different numerical features. This helps to identify any strong relationships that may be important for model training.

---

7. Machine Learning Models

Three machine learning algorithms were used to predict the presence of heart disease:

- **Logistic Regression:** A linear model used for binary classification.
- **Random Forest:** An ensemble learning method that builds multiple decision trees to improve accuracy.
- **Support Vector Machine (SVM):** A classification model that finds the optimal hyperplane to separate different classes.

Each model was trained on the preprocessed data and evaluated on the test set.

---

8. Model Evaluation

The performance of each model was evaluated using several metrics:

- **Accuracy:** The percentage of correct predictions made by the model.
- **Precision, Recall, and F1-Score:** These metrics were used to evaluate the performance of the models, especially in dealing with class imbalances.
- **Confusion Matrix:** Used to summarize the performance by showing the number of true positives, false positives, true negatives, and false negatives.

The classification report for each model provided detailed insights into how well each model predicted both classes (heart disease and no heart disease).

10. Conclusion

Based on the performance evaluation, the most suitable model for heart disease prediction was selected. The results indicate which features are most predictive of heart disease and highlight the importance of using machine learning for early diagnosis. The project demonstrates the feasibility of automating heart disease prediction, which can significantly aid in clinical decision-making.

11. Future Work

Future improvements may include:

- Hyperparameter tuning to optimize the model performance.
- Exploring more advanced models, such as ensemble methods or deep learning.

- Collecting additional data to further improve the accuracy of the models.