Heart Disease Prediction Using Machine Learning

Objective

The objective of this project is to develop a machine learning model that predicts the presence of heart disease based on several medical attributes. This model aims to assist healthcare professionals in identifying patients at risk of heart disease, enabling timely intervention.

Dataset Used

The dataset used for this project is the "Heart Disease Dataset," which contains 13 features related to patient health metrics such as age, sex, chest pain type (cp), resting blood pressure (trestbps), cholesterol levels (chol), fasting blood sugar (fbs), maximum heart rate achieved (thalach), and others. The target variable (target) indicates whether a patient has heart disease (1) or not (0).

Key Features:

- **age**: Age of the patient
- **sex**: Gender (1 = male, 0 = female)
- **cp**: Chest pain type (categorical: 0–3)
- **trestbps**: Resting blood pressure
- **chol**: Serum cholesterol in mg/dl
- **fbs**: Fasting blood sugar (>120 mg/dl; 1 = true, 0 = false)
- **restecg**: Resting electrocardiographic results
- **thalach**: Maximum heart rate achieved
- **exang**: Exercise-induced angina (1 = yes, 0 = no)
- **oldpeak**: ST depression induced by exercise relative to rest
- **slope**: Slope of the peak exercise ST segment
- **ca**: Number of major vessels colored by fluoroscopy
- **thal**: Thalassemia (categorical)
- **target**: Presence of heart disease (1 = disease, 0 = no disease)

Model Chosen

The project uses a Logistic Regression as the primary machine learning model. Random Forest was selected for its ability to handle both categorical and numerical data effectively and its robustness against overfitting.

Implementation Steps:

1. **Data Preprocessing:**
    - Handle missing values.
    - Normalize numerical features.
    - Encode categorical variables using one-hot encoding.

2. **Model Training:**
    - Split the dataset into training and testing subsets.
    - Train the Random Forest Classifier using the training data.
3. **Evaluation:**
    - Evaluate model performance using metrics like accuracy, precision, recall, and F1-score.
4. **Visualization:**
    - Generate feature importance plots and confusion matrices.

Performance Metrics

The model's performance was evaluated using the following metrics:

- **Accuracy:** Measures overall correctness of predictions.
- **Precision:** Proportion of true positive predictions among all positive predictions.
- **Recall:** Proportion of true positive predictions among all actual positives.
- **F1-score:** Harmonic mean of precision and recall.

Results:

- Accuracy: 85%
- Precision: 82%
- Recall: 88%
- F1-score: 85%

Challenges & Learnings

Challenges:

1. Imbalanced dataset with fewer positive cases compared to negative cases.
2. Selecting optimal hyperparameters for the Random Forest model to avoid overfitting.
3. Handling categorical features effectively during preprocessing.

Learnings:

1. Importance of feature engineering in improving model performance.
2. Visualization techniques for understanding feature importance and model behavior.
3. The trade-off between precision and recall in medical datasets where false negatives can have severe consequences.