Heart Disease Prediction Model

Overview

This project aims to predict the likelihood of heart disease in patients using a dataset containing various medical predictor variables and one target variable, target, which indicates whether the patient has heart disease (1) or not (0). The project involves data preprocessing, model training using Logistic Regression, and evaluation of the model's performance.

Dataset

The dataset used in this project contains 1025 rows and 14 columns. The columns include:

- **age**: Age of the patient
- **sex**: Gender of the patient (1 = male, 0 = female)
- **cp**: Chest pain type (0-3)
- **trestbps**: Resting blood pressure (mm Hg)
- **chol**: Serum cholesterol (mg/dl)
- **fbs**: Fasting blood sugar > 120 mg/dl (1 = true, 0 = false)
- **restecg**: Resting electrocardiographic results (0-2)
- **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 (0-3) colored by fluoroscopy
- **thal**: Thalassemia (3 = normal, 6 = fixed defect, 7 = reversible defect)
- **target**: Target variable (0 = no heart disease, 1 = heart disease)

Project Steps

1. Importing Libraries

The necessary libraries for data processing, model training, and evaluation are imported:

- **NumPy**: For numerical operations.
- **Pandas**: For data manipulation and analysis.
- **StandardScaler**: For standardizing the dataset.
- **train_test_split**: For splitting the dataset into training and testing sets.
- **LogisticRegression**: For training the Logistic Regression model.
- **accuracy_score**: For evaluating the model's accuracy.

2. Data Collection and Processing

The dataset is loaded into a Pandas DataFrame, and basic statistical measures are analyzed to understand the data distribution.

3. Data Preprocessing

The dataset is checked for missing values, and the target variable is analyzed to understand the distribution of heart disease cases.

4. Splitting Features and Target

The dataset is split into features (X) and labels (Y). The features are standardized to ensure that all variables contribute equally to the model.

5. Splitting Data into Training and Testing Sets

The dataset is split into training and testing sets, with 80% of the data used for training and 20% for testing.

6. Model Training

A Logistic Regression model is trained on the training data.

7. Model Evaluation

The model's accuracy is evaluated on both the training and testing datasets to ensure it generalizes well to unseen data.

8. Making Predictions

The trained model is used to make predictions on new data. The input data is standardized before making the prediction.

Results

- **Training Data Accuracy**: Approximately 85.85%
- **Test Data Accuracy**: Approximately 80.49%

The model performs well on both the training and test datasets, indicating that it generalizes well to unseen data.

Conclusion

This project demonstrates the use of Logistic Regression for predicting heart disease based on medical data. The model achieves a reasonable accuracy and can be further improved by tuning hyperparameters or using more advanced techniques.

How to Run the Code

1. Clone the repository.
2. Install the required libraries using pip install -r requirements.txt.
3. Run the Jupyter notebook Heart_disease_prediction.ipynb.

Requirements

- Python 3.x
- NumPy
- Pandas
- Scikit-learn

Acknowledgments

- The dataset is sourced from Kaggle.