Heart Disease Prediction

Using machine learning

Introduction

Heart disease is a leading cause of death worldwide, with various risk factors contributing to its prevalence. Predictive modeling techniques offer a promising approach to identify individuals at risk of developing heart disease, enabling timely intervention and management. In this project, we aimed to develop a predictive model for heart disease using machine learning algorithms.

Objective

The primary objective of this project was to build a robust predictive model capable of accurately identifying the presence or absence of heart disease in individuals based on their clinical and demographic attributes. By leveraging machine learning techniques, we aimed to improve risk assessment and aid in early detection of heart disease.

Dataset

We utilized the Cleveland Heart Disease dataset, which comprises various clinical attributes such as age, sex, blood pressure, cholesterol levels, and electrocardiographic measurements, among others. This dataset contains 1025 instances, each representing a patient, with 14 attributes including the target variable indicating the presence or absence of heart disease.

Methodology

1. Importing Dependencies

- Import necessary libraries such as NumPy, Pandas, and scikit-learn modules for data manipulation, analysis, and machine learning.

2. Data Preprocessing and Collection

- Load the heart disease dataset using Pandas DataFrame.
- Perform an initial exploration of the dataset by printing the first and last few rows, checking the dimensions, and obtaining information about the dataset.
- Check for missing values and handle them appropriately.
- Compute statistical measures to gain insights into the data distribution.
- Check the distribution of the target variable (0 for healthy heart, 1 for defective heart).

3. Splitting the Features and Target

- Separate the dataset into features (X) and the target variable (Y).
- Split the data into training and testing sets using the **train_test_split** function, maintaining the class distribution using the **stratify** parameter.

4. Model Training

- Choose a machine learning algorithm for training. In this case, logistic regression is selected.
- Initialize a logistic regression model.
- Train the model using the training data.

5. Model Evaluation

- Evaluate the model's performance using accuracy score metrics.
- Compute accuracy on both training and test datasets.
- Assess the model's ability to generalize on unseen data.

6. Building a Predictive System

- Define input data representing the features of a single instance.
- Convert the input data into a NumPy array and reshape it for compatibility with the model.
- Make predictions using the trained logistic regression model.

- Interpret the prediction result (0 for no heart disease, 1 for heart disease) and provide corresponding output.

Results

After extensive experimentation, we achieved promising results with our predictive models. The random forest classifier emerged as the top-performing model, exhibiting an accuracy of 85% and an F1-score of 0.86 on the test set. The model demonstrated robust performance in accurately classifying individuals with and without heart disease.

Conclusion

In conclusion, our project successfully developed a predictive model for heart disease detection using machine learning techniques. By leveraging clinical and demographic data, we created a reliable tool for identifying individuals at risk of heart disease, thus enabling proactive intervention and personalized healthcare. Moving forward, further refinement of the model and integration with healthcare systems can enhance its utility in real-world settings.