Introduction

Heart Disease Prediction Using Logistic Regression

Heart disease remains one of the leading causes of death worldwide, making early detection and prediction crucial for effective treatment and prevention. Machine learning has proven to be a powerful tool in healthcare like predicting heart disease. One such model that can be used is logistic regression that is used to model the probability of a binary outcome such as the presence or absence of heart disease—based on input features. In this project we will explore how Logistic regression to predict the likelihood of heart disease in patients and the accuracy of the module.

Project Objective

The primary objective of this project is to develop a machine learning model using logistic regression to find the probability and outcome of heart disease with accuracy. By using the data from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts we'll predict whether the patient has 10-year risk of future coronary heart disease (CHD). Then we'll train and test the data to predict the number of patients affected by CHD and in the end, we'll evaluate the accuracy of the predicated data of the patients.

Key Features of the Project

1) Data Preparation and cleaning

- The dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether the patient has 10-year risk of future coronary heart disease (CHD). The dataset provides the patients information. It includes over 4,000 records and 15 attributes.
- Then **remove any rows** with missing values (NaN) from the DataFrame.
- Now prints the count of unique values in the TenYearCHD column which likely indicates whether a patient has heart disease.

2) Splitting the Dataset into Test and Train Sets

- Scale the features to have a mean of 0 and standard deviation of 1 using StandardScaler.
- Scaling is important for many machine learning models, especially when the features have different units or magnitudes.
- Training set (70% of data)
- Test set (30% of data)

3) Exploratory Data Analysis of Heart Disease Dataset

- Exploratory Data Analysis (EDA) is the step where we analyse a dataset to summarize its main characteristics and discover patterns, trends or anomalies. In

this section, we perform EDA on the heart disease dataset to understand and gain insights into the dataset before building a predictive model for heart disease.

• This creates a count plot using Seaborn. It visualizes the distribution of the values in the TenYearCHD column showing how many individuals have heart disease (1) vs. how many don't (0).

• The count plot shows a high imbalance in the dataset where the majority of individuals (over 3000) do not have heart disease (label 0) while only a small number (around 500) have heart disease (label 1).

4) Fitting Logistic Regression Model for Heart Disease Prediction

• logreg = LogisticRegression(): This creates an instance of the LogisticRegression model.

• logreg.fit(): This trains the logistic regression model using the training data.

• logreg.predict(): This uses the trained logistic regression model to make predictions on the test set.

Implementation Details

Tools and Technologies:

- Programming Language: Python

- Libraries and Frameworks:
  - Pandas and NumPy for data manipulation and analysis.
  - Scikit-learn for implementing machine learning algorithms.
  - Statsmodel for statistical modelling for fitting logistic regression.
  - Matplotlib and Seaborn for data visualization.

Workflow:

Data cleaning:

- Removed null values ana NaN.
- Conducted exploratory data analysis (EDA) to understand the dataset's structure and distribution.

Model Training:

- Compared multiple machine learning models to identify the best-performing Algorithm.
- Used Statsmodel for statistical modelling for fitting logistic regression.

Testing and Training:

- Testing and training the data is more important in machine learning.
- Scaling also plays an important role in machine learning for testing and training.

Challenges and Solutions

1. Challenge:
   Empty Values

Solution: Prepare your data before using it in the machine learning model to get the prediction right.

2. Challenge:
   Scaling

Solution: Scaling is important in machine learning when the when the features have different units or magnitudes.

3. Challenge:
   Accuracy

Solution: Used regularization techniques like L2 regularization and limited the complexity of algorithms.

Results and Insights

The machine learning model achieved an accuracy of approximately 84% on the testing dataset. Logistic Regression emerged as the most effective algorithm due to its simplicity and strong performance. The project demonstrated how a systematic approach to data preprocessing and model selection that could significantly enhance the probability and outcome of heart disease with accuracy.