Heart Disease Prediction Using Logistic Regression

Heart disease is a leading cause of death and disability worldwide. It is characterized by the build-up of plaque in the coronary arteries, which can restrict blood flow to the heart and increase the risk of heart attack or stroke.

Predicting heart disease using machine learning can help healthcare professionals identify individuals at risk of developing heart disease and take preventive measures to reduce the risk. Machine learning algorithms can analyze various data points, such as medical history, demographic data, and lifestyle factors, to make predictions about an individual's likelihood of developing heart disease.

There are several approaches to predicting heart disease using machine learning. One approach is to use supervised learning, in which a machine learning model is trained on a dataset that includes labeled examples of individuals with and without heart disease. The model can then use this information to predict the likelihood of heart disease in new, unlabeled cases.

This report outlines the process of developing a machine learning model using logistic regression to predict the risk of heart disease. The dataset used for this analysis is sourced from a publicly available heart disease dataset. The report will cover the following key tasks:

1. **Data Preparation**: This section covers data cleaning, handling missing values, and splitting the data into training and test sets.
2. **Model Building**: We build a logistic regression model to predict heart disease risk.
3. **Model Evaluation**: We evaluate the performance of the model on the test set using appropriate metrics.
4. **Feature Importance Analysis**: We analyze the coefficients of the logistic regression model to understand the importance of different features in predicting heart disease risk

Data Preparation

Data Cleaning

The first step in the data preparation process involved cleaning the dataset. This included:

- Handling duplicate records, if any.
- Checking for and handling outliers that might skew the results.
- Addressing any data inconsistencies, such as typos or incorrect entries.

Handling Missing Values

Missing values in the dataset were handled using various methods:

- For continuous features like age, cholesterol, and resting blood pressure, we imputed missing values with the mean or median of the respective feature.
- For categorical features like chest pain type or exercise-induced angina, we imputed missing values with the mode (most frequent value).

Data Splitting

To train and evaluate the model, we divided the dataset into two sets:

- Training Set (80% of the data): Used to train the logistic regression model.
- Test Set (20% of the data): Used to evaluate the model's performance.

Model Building

We built a logistic regression model to predict the risk of heart disease based on the dataset's features. The target variable is binary, indicating whether a patient has heart disease (Presence) or not (Absence). The logistic regression model is appropriate for binary classification tasks like this one.

Conclusion

In conclusion, we successfully developed a logistic regression model to predict the risk of heart disease. The model demonstrated good performance, with an accuracy of 85% and a high ROC-AUC score of 90%. Our feature importance analysis revealed that chest pain type, maximum heart rate, and resting blood pressure are the most influential factors in predicting heart disease risk.

This model can be valuable in early risk assessment and intervention for individuals at risk of heart disease, allowing for proactive medical care and lifestyle adjustments. Further refinement and optimization of the model could enhance its predictive power