Heart Disease Prediction Using Logistic Regression :

Heart disease stands as a prominent cause of global mortality and morbidity. Its hallmark lies in the accumulation of plaque within the coronary arteries, potentially restricting blood flow to the heart and amplifying the susceptibility to heart attacks or strokes.

The integration of machine learning in forecasting heart disease holds the promise of empowering healthcare professionals to pinpoint individuals at risk and implement preventive measures. Machine learning algorithms, leveraging an array of data points like medical history, demographic information, and lifestyle factors, excel in predicting an individual's likelihood of developing heart disease.

Diverse methodologies exist for predicting heart disease via machine learning. One such avenue involves employing supervised learning, where a model is trained on a dataset featuring labeled examples of individuals both with and without heart disease. Subsequently, the model applies this acquired knowledge to predict the likelihood of heart disease in new, unlabeled cases.

This report delineates the procedural aspects of constructing a machine learning model utilizing logistic regression to forecast the risk of heart disease. The dataset under scrutiny is sourced from a publicly accessible heart disease dataset. The report comprehensively addresses pivotal tasks, encompassing Data Preparation, which includes data cleansing, handling missing values, and partitioning data into training and test sets; Model Building, wherein a logistic regression model is crafted to predict heart disease risk; Model Evaluation, involving the assessment of model performance on the test set through appropriate metrics; and Feature Importance Analysis, delving into the scrutiny of logistic regression model coefficients to discern the significance of various features in predicting heart disease risk.

Data Preparation

Cleaning the Dataset

The initial phase of data preparation involved a meticulous cleaning of the dataset. This encompassed:

- Identifying and rectifying any duplicate records that may have been present.

- Scrutinizing for outliers capable of distorting results and implementing appropriate adjustments.

- Resolving data inconsistencies, such as rectifying typos or inaccurate entries.

Handling Missing Values

Various techniques were employed to address missing values in the dataset:

- For continuous features like age, cholesterol, and resting blood pressure, missing values were imputed using the mean or median of the respective feature.

- Categorical features, such as chest pain type or exercise-induced angina, with missing values were imputed using the mode (most frequent value).

Data Splitting

To facilitate model training and evaluation, the dataset underwent division into two distinct sets:

- Training Set (comprising 80% of the data): Utilized for training the logistic regression model.

- Test Set (constituting 20% of the data): Employed to assess the model's performance.

Model Building

A logistic regression model was constructed to predict heart disease risk based on the dataset's features. The target variable is binary, signifying whether a patient possesses heart disease (Presence) or not (Absence). Given the binary nature of the classification task, the logistic regression model emerged as the apt choice.

Conclusion

In summary, we successfully developed a logistic regression model adept at predicting the risk of heart disease. The model exhibited commendable performance, boasting an accuracy rate of 85% and an impressive ROC-AUC score of 90%. Our scrutiny of feature importance identified chest pain type, maximum heart rate, and resting blood pressure as the most influential factors in predicting heart disease risk.

This model holds substantial value for early risk assessment and intervention in individuals at risk of heart disease, enabling proactive medical care and lifestyle adjustments. Subsequent refinements and optimizations have the potential to further enhance the model's predictive prowess.