Heart Disease Prediction Using Logistic Regression: A Comprehensive Approach for Early Diagnosis

Abstract

Heart disease remains one of the leading causes of mortality worldwide, emphasizing the importance of early diagnosis and prediction. This study utilizes Logistic Regression, a widely used classification algorithm, to predict the likelihood of heart disease. The model was trained and evaluated using a publicly available dataset, achieving an accuracy of 85.12% on the training data and 81.96% on the test data. This paper presents the methodology, experimental work, and results, highlighting Logistic Regression's capability to provide reliable predictions in healthcare applications.

Introduction

Cardiovascular diseases (CVD) are a significant global health burden, accounting for a large proportion of deaths annually. Predicting heart disease risk at an early stage can lead to timely intervention and treatment, reducing the likelihood of severe outcomes. Traditionally, medical practitioners rely on risk factors such as age, cholesterol levels, and blood pressure to diagnose heart disease. However, machine learning algorithms can provide more accurate and data-driven predictions by analyzing complex interactions between variables. Logistic Regression is a powerful yet interpretable machine learning algorithm suitable for binary classification problems, such as heart disease prediction. This study explores the use of Logistic Regression in predicting heart disease and evaluates its effectiveness using real-world clinical data.

Related Works

Several machine learning algorithms have been applied to predict heart disease, including Decision Trees, Random Forest, Support Vector Machines, and Neural Networks. Logistic Regression, due to its simplicity and effectiveness, has been a popular choice in medical studies. Previous research has demonstrated that Logistic Regression performs well in predicting binary outcomes such as heart disease presence or absence, especially when datasets contain both continuous and categorical variables. Studies like "Heart Disease Prediction Using Logistic Regression and Neural Networks" and "A Review of Machine Learning Techniques for Heart Disease Diagnosis" highlight its efficacy in producing accurate predictions with minimal computational cost. Compared to more complex models like Neural Networks, Logistic Regression offers better interpretability, making it ideal for use in healthcare.

Algorithm

Logistic Regression

Logistic Regression is a statistical method used for binary classification problems. It models the probability that a given input belongs to a particular class (e.g., presence or absence of heart disease). The model calculates a weighted sum of the input features and passes the result through a logistic function to produce a probability between 0 and 1. If the predicted probability exceeds a threshold (typically 0.5), the output is classified as 1 (positive), otherwise 0 (negative). The logistic function is defined as:

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}}$$

where:

θ represents the weights (coefficients) learned by the model,

x is the input feature vector.

The model is trained using maximum likelihood estimation to minimize the difference between predicted and actual outcomes.

Methodology

Data Collection: The dataset used for this study is the "Heart Disease" dataset, containing clinical data on patients such as age, sex, cholesterol levels, blood pressure, and more. The target variable indicates whether the patient has heart disease (1) or not (0).

Data Preprocessing: The dataset was cleaned by handling missing values, encoding categorical variables (such as sex and chest pain type), and scaling numerical features like cholesterol and age. The dataset was then split into training (80%) and testing (20%) sets to evaluate model performance.

Model Training: Logistic Regression was used as the classification algorithm. The model was trained on the training data using the sklearn library. The weights were learned by minimizing the binary cross-entropy loss function. Hyperparameters such as regularization (L2 penalty) were tuned using cross-validation to avoid overfitting.

Model Evaluation: The trained Logistic Regression model was evaluated on both the training and testing datasets using accuracy, precision, recall, and F1-score as performance metrics. The accuracy on the training data was 85.12%, while on the test data, it achieved 81.96%.

Experimental Work

The experimental setup involved the following key steps:

Data Analysis: Exploratory Data Analysis (EDA) was conducted to understand the distribution of the features and their correlation with heart disease. Visualizations such as histograms and correlation matrices were used to explore patterns and relationships.

Model Training and Validation: Logistic Regression was implemented using the LogisticRegression class from the sklearn library. The data was split using the train_test_split function, with 80% used for training and 20% for testing. Cross-validation was performed to tune the regularization parameter (C).

Feature Importance: The coefficients learned by the Logistic Regression model were analyzed to determine the importance of each feature in predicting heart disease. Factors such as cholesterol level, age, and maximum heart rate were found to be significant predictors.

Results

The Logistic Regression model achieved the following results:

Accuracy on Training Data: 85.12%

Accuracy on Test Data: 81.96%

Additional performance metrics:

Precision: The model demonstrated good precision, indicating a low false-positive rate.

Recall: Recall was also high, meaning the model was able to detect most cases of heart disease.

F1-Score: The balance between precision and recall was captured by the F1-Score, further confirming the model's reliability.

The model's performance indicates that Logistic Regression is a suitable method for predicting heart disease, with an acceptable trade-off between simplicity and accuracy.

Conclusion

This study demonstrated the effectiveness of Logistic Regression in predicting heart disease using clinical data. The model achieved an accuracy of 85.12% on the training data and 81.96% on the test data, highlighting its capability to provide accurate predictions. The simplicity and interpretability of Logistic Regression make it a valuable tool for healthcare professionals, as it allows them to identify important risk factors and make data-driven decisions. Future work could involve comparing Logistic Regression with more complex models like Support Vector Machines or Neural Networks to further enhance prediction accuracy.

References

- Alizadehsani, R., et al. (2018). "A Review of Machine Learning Techniques for Heart Disease Diagnosis and Prediction." Journal of Medical Systems, 42(7), 1-13.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R. Springer.
- Khosravi, A., et al. (2020). "Heart Disease Prediction Using Data Mining Techniques." International Journal of Advanced Computer Science and Applications, 11(6), 23-28.
- Pedregosa, F., et al. (2011). "Scikit-learn: Machine Learning in Python." Journal of Machine Learning Research, 12, 2825-2830.