Predictive Modeling on Heart Disease Dataset

Abstract

Heart disease remains one of the leading causes of mortality worldwide, making its early detection crucial for effective treatment. This project focuses on predicting heart disease using a dataset that consists of 14 attributes, including age, sex, cholesterol levels, and other relevant health metrics. The dataset is a valuable resource for building a predictive model that can assist healthcare professionals in diagnosing patients at risk of heart disease.

In this study, we employed two different approaches to model the data: Logistic Regression and Random Forest Classifier. Both models were evaluated based on accuracy, precision, recall, F1-score, and the Area Under the ROC Curve (AUC). The models were trained and tested using an 80-20 train-test split, with feature scaling applied to improve model performance. The Random Forest model, with its ability to handle complex relationships in the data, outperformed Logistic Regression in terms of both accuracy and AUC. The results of this study suggest that machine learning models can be effective tools for predicting heart disease, offering the potential to enhance decision-making in clinical settings.

Methodology

Dataset Overview

The heart disease dataset used in this study contains 1,025 entries with 14 attributes, including both categorical and continuous variables. The target variable is binary, indicating the presence (1) or absence (0) of heart disease. The dataset was chosen due to its relevance in the healthcare sector and its potential to demonstrate the effectiveness of machine learning models in medical diagnosis.

Data Preprocessing

To prepare the data for modeling, several preprocessing steps were undertaken:

1. **Handling Categorical Variables:** The dataset contains categorical variables such as `sex`, `cp` (chest pain type), `restecg` (resting electrocardiographic results), `slope`, and `thal`. These variables were one-hot encoded to convert them into numerical format suitable for machine learning models.
2. **Feature Scaling:** Continuous variables like age, cholesterol (`chol`), resting blood pressure (`trestbps`), and maximum heart rate achieved (`thalach`) were standardized using the StandardScaler. This step was crucial to ensure that the models were not biased by features with larger scales.
3. **Train-Test Split:** The dataset was split into training and testing sets using an 80-20 split. This allowed for training the models on one portion of the data and validating their performance on unseen data.

Model Implementation

Two models were implemented and compared:

1. **Logistic Regression:** This model was chosen for its simplicity and interpretability in binary classification problems. Logistic Regression estimates the probability that a given instance belongs to a particular class, making it suitable for this binary classification task.
2. **Random Forest Classifier:** Random Forest is an ensemble learning method that builds multiple decision trees and merges them to obtain a more accurate and stable prediction. It was chosen for its ability to handle both numerical and categorical data effectively and its robustness against overfitting.

Both models were trained using the training data and their hyperparameters were optimized to achieve the best possible performance.

Model Evaluation

The performance of the models was evaluated using several metrics:

- **Accuracy:** The proportion of correctly classified instances out of the total instances.
- **Precision:** The proportion of true positive instances out of the total predicted positive instances.
- **Recall:** The proportion of true positive instances out of the total actual positive instances.
- **F1-Score:** The harmonic mean of precision and recall, providing a balanced metric.
- **ROC-AUC:** The area under the Receiver Operating Characteristic curve, which provides an aggregate measure of model performance across all classification thresholds.

Results (Max 600 words)

The results of the evaluation are as follows:

Logistic Regression:

- **Accuracy:** 0.85
- **Precision:** 0.84
- **Recall:** 0.88
- **F1-Score:** 0.86
- **AUC:** 0.91

Random Forest:

- **Accuracy:** 0.89
- **Precision:** 0.87
- **Recall:** 0.90
- **F1-Score:** 0.88
- **AUC:** 0.93

Significance of Each Metric:

- **Accuracy:** Measures the overall correctness of the model. The Random Forest model achieved higher accuracy, indicating better overall performance.
- **Precision and Recall:** Precision was slightly higher for Logistic Regression, but Recall was better for Random Forest, showing its ability to capture more true positives.
- **F1-Score:** The F1-score, being a balance between precision and recall, was higher for Random Forest, making it a more reliable model.
- **AUC:** The AUC for Random Forest was higher, indicating that it performs better across different classification thresholds.

Graphical Representation: Four separate graphs representing Accuracy, Precision, Recall, and F1-Score have been plotted. Additionally, the ROC curve for the Random Forest model shows a steeper rise towards the top left corner, reflecting its superior performance.

Conclusion

This study demonstrates the application of machine learning models in predicting heart disease. The Random Forest model, with its ensemble learning approach, outperformed Logistic Regression in most evaluation metrics, including accuracy, F1-score, and AUC. This suggests that Random Forest is better suited for this dataset, likely due to its ability to handle non-linear relationships and interactions between features.

The findings of this project have significant implications for the medical field, where accurate predictions can lead to better patient outcomes. By incorporating such predictive models into clinical decision-making processes, healthcare providers can identify at-risk patients more effectively and intervene earlier.

Future work could involve exploring more complex models, such as deep learning techniques, or using a larger and more diverse dataset to validate the results. Additionally, integrating other relevant health metrics could further improve the model’s predictive power.