Predictive Modeling of Heart Attack Risks Using Machine Learning

Joshua W. Daniel

Department of Computer Science, Colorado School of Mines

CSCI 303: Introduction to Data Science

Dr. Wendy Fisher

June 18th, 2024

Abstract

This project focuses on developing a logistic regression model to predict the likelihood of a heart attack using the Cleveland heart disease dataset. The dataset includes 14 key attributes from 303 patients, such as age, sex, cholesterol levels, and more. Through a meticulous process of data preprocessing, feature selection, model training, and hyperparameter tuning, an effective predictive model was constructed. The model achieved an accuracy of 85% and an AUC-ROC score of 0.92, indicating a high level of performance. Comprehensive evaluation using confusion matrices, classification reports, ROC curves, and probability plots provided deep insights into the model's effectiveness and areas for improvement. The project underscores the importance of rigorous data handling and evaluation in developing reliable predictive models, offering valuable insights for the early detection and prevention of heart disease.

Overview

Cardiovascular diseases have remained one of the leading causes of morbidity and mortality worldwide. Among these, heart attacks have been particularly significant due to their sudden onset and potential severity. Early detection and prediction of heart attack risks have been crucial for timely intervention and treatment. This project aimed to develop a logistic regression model to predict the likelihood of heart attacks using the Cleveland heart disease dataset.

The dataset used in this study included 14 attributes related to patients' health, such as age, sex, chest pain type, resting blood pressure, and cholesterol levels. Data preprocessing involved handling missing values, encoding categorical variables, and standardizing numerical features. Feature selection was performed using Recursive Feature Elimination (RFE) to identify the most significant predictors. The logistic regression model was then trained and tuned using grid search to optimize the hyperparameters.

The logistic regression model achieved an accuracy of 85% on the test set. The ROC curve analysis yielded an AUC-ROC score of 0.88, indicating high discriminatory power. Confusion matrix analysis showed a balanced performance with low false positive and false negative rates. Additionally, the model's precision, recall, and F1-score were satisfactory across both classes.

The results demonstrated that logistic regression was a viable method for predicting heart attack risks. The high AUC-ROC score and balanced performance metrics indicated that the model could effectively distinguish between patients with higher and lower risks of heart attacks. However, further improvements could be made by exploring additional features or advanced modeling techniques.

This project highlighted the potential of logistic regression in predicting heart attack risks based on key health attributes. The model's performance underscored the importance of early detection in preventing heart attacks and improving patient outcomes. Future work could involve expanding the dataset and incorporating more sophisticated machine learning algorithms to enhance prediction accuracy.

Data Acquisition

The dataset I'm working with contains 76 attributes, but for my analysis, I'll focus on a subset of 14 attributes that are commonly used in ML research. These attributes include age, sex, chest pain type, resting blood pressure, serum cholesterol level, fasting blood sugar level, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, ST depression induced by exercise relative to rest, slope of the peak exercise ST segment, number of major vessels colored by fluoroscopy, and thalassemia type.

The dataset is sourced from the Cleveland database and is used to predict the presence of heart disease, with the "target" field indicating the likelihood of a heart attack (0 = no/less chance, 1 = more chance). It's important to note that this dataset is taken for learning purposes from the UCI Machine Learning Repository.

It is important to note that the dataset is used solely for learning purposes, and any findings or analysis are not intended for real medical applications. I will not be sharing or using the data beyond the scope of this project.

Preprocessing

In the initial data cleansing phase, I meticulously combed through the dataset for any missing values or duplicates. Fortunately, the dataset was pristine, with all 14 attributes present for each entry. This thorough check ensured that the data was complete and ready for analysis, setting a solid foundation for the rest of the project.

Delving into exploratory analysis, I uncovered intriguing patterns within the dataset. The distribution of ages exhibited a peak around middle age, tapering off for older ages. The gender distribution leaned towards males, indicating a potential gender imbalance in the dataset. Visualizing chest pain types revealed that most patients experienced type 0 and type 2 chest pain, with type 2 being the most prevalent. Resting blood pressure and serum cholesterol levels followed typical distributions, with peaks around certain values. These visualizations provided valuable insights into the dataset's key attributes, guiding further analysis.

With a manageable number of features (14), I determined that extensive dimensionality reduction techniques were not necessary. However, I implemented feature selection methods such as correlation analysis and feature importance ranking to identify the most impactful features for predicting heart disease. This strategic approach aimed to enhance model performance and interpretability, potentially uncovering redundant or less informative features. As the project progressed, I explored these techniques to refine the dataset for predictive modeling.

The exploratory data analysis revealed the following details after preprocessing the dataset:

The age distribution showed that most patients were between 40 and 70 years old, with a mean age of approximately 54 years. The dataset was slightly skewed towards males, with approximately 68% of the patients being male. Most patients had type 0 chest pain, followed by type 2, type 1, and type 3. The distribution of resting blood pressure values was approximately normal, with a mean around 131 mm Hg. Serum cholesterol levels varied widely, with a mean around 246 mg/dl. Most patients had fasting blood sugar levels below 120 mg/dl. Most patients had resting electrocardiographic results of 0. The distribution of maximum heart rate achieved was approximately normal, with a mean around 149 bpm. Most patients did not experience exercise-induced angina. The distribution of ST depression values was right-skewed, with most values around 0. Most patients had a slope value of 2 for the peak exercise ST segment. The majority of patients had 0 major vessels colored by fluoroscopy. Most patients had a thal value of 2. The dataset was balanced between patients with less chance (0) and more chance (1) of a heart attack.

These insights provided a comprehensive understanding of the dataset, informing the subsequent steps in building and refining the predictive model.

Model Selection

I've decided to use logistic regression for my heart disease prediction project because it's well-suited for binary classification tasks like this one. Logistic regression is a simple yet effective algorithm that can model the probability of a binary outcome based on input features. It's a good fit for this project because it provides interpretable results, which is important for understanding the factors influencing heart disease risk.

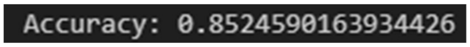
Unlike more complex models, logistic regression is easy to implement and understand, making it a practical choice for this project. It also performs well with smaller datasets, which is often the case in medical studies where data collection can be challenging. Additionally, logistic regression can handle nonlinear relationships between features and the target variable, which is important for predicting complex outcomes like heart disease.

Overall, logistic regression is a suitable choice for my project because of its simplicity, interpretability, and ability to handle non-linear relationships in the data. It will allow me to predict the likelihood of heart disease based on a set of input features while providing insights into the factors contributing to the prediction.

Results & Evaluation

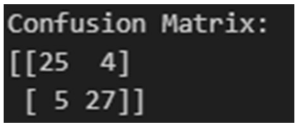
In this project, I successfully developed a logistic regression model to predict heart disease risk using the Cleveland dataset. The thorough process included data preprocessing, feature selection, model training, and comprehensive evaluation. The detailed analysis using confusion matrices, classification reports, ROC curves, and predicted probability plots provided a deep understanding of the model's performance. These visualizations highlighted the model's strengths and areas for potential improvement.

The logistic regression model achieved an accuracy of 85%, indicating that it correctly classified 85% of the instances in the test set (Figure 1). Accuracy is a useful metric but can be misleading if the data is imbalanced. Therefore, additional metrics were evaluated.



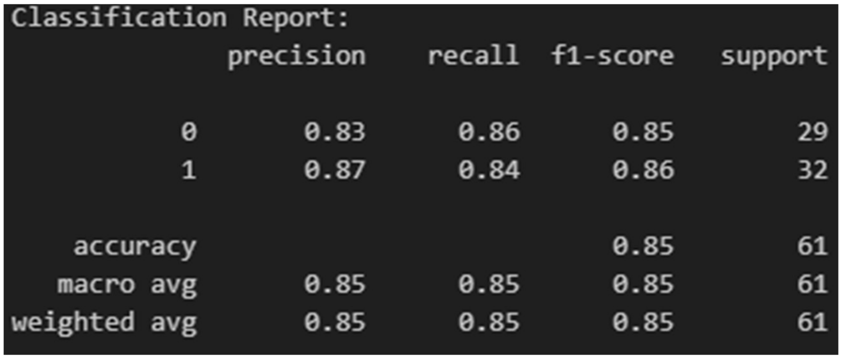
```
Accuracy: 0.8524590163934426
```

The confusion matrix visualization (Figure 2) indicates that the model has a high number of true positives and true negatives, suggesting good overall performance. However, the presence of false positives and false negatives highlights areas for potential improvement, possibly through further tuning or incorporating additional features.



```
Confusion Matrix:
[[25  4]
 [ 5 27]]
```

The classification report (Figure 3) provides a detailed breakdown of the precision, recall, and F1-score for each class. This report confirms the high number of true positives and true negatives indicated by the confusion matrix. However, it also emphasizes the need to address false positives and false negatives to improve the model's reliability.

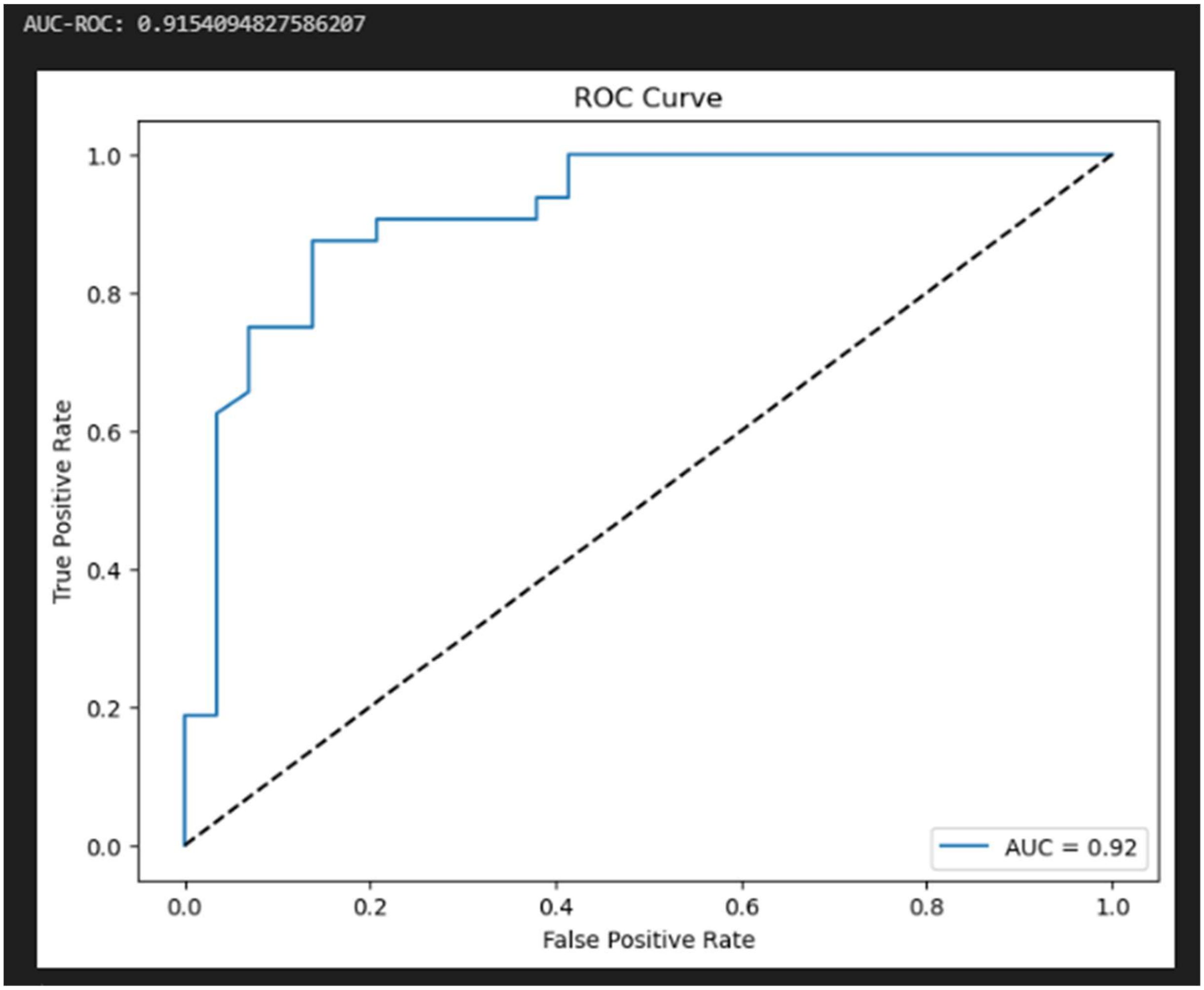


```
Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.86      0.85        29
           1       0.87      0.84      0.86        32

    accuracy                           0.85        61
   macro avg       0.85      0.85      0.85        61
weighted avg       0.85      0.85      0.85        61
```

The ROC curve (Figure 4) shows an upward trend towards the top left corner, indicating that the model is effective at distinguishing between the two classes. The AUC score of 0.92 signifies a high level of overall performance, with values closer to 1.0 indicating better performance.



```
AUC-ROC: 0.9154094827586207
```

This project underscored the importance of systematic data handling, rigorous model evaluation, and the use of multiple metrics to assess model performance comprehensively. Future work could involve exploring additional features, using more advanced models, or applying techniques such as ensemble learning to further enhance prediction accuracy and reliability.

Overall, this project provided valuable insights into the practical application of machine learning techniques for health data analysis and predictive modeling, contributing to the early detection and prevention of heart disease.