Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

COMPSCI 189

Predicting Heart Disease Severity using Machine Learning

Yashas Raman[1*], Raghav Sriram[1]

[1]Department of Information and Computer Science

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract
Heart disease is one of the leading causes of death globally. In this project, we used machine learning to predict the presence and severity of heart disease using clinical patient features from the UCI Heart Disease dataset. We performed both binary classification (disease vs. no disease) and multiclass classification (severity on a 0–4 scale). Three machine learning models—Logistic Regression, Random Forest, and XGBoost—were evaluated using an 80/20 train-test split. While all models achieved strong performance in the binary task, results were notably weaker for multiclass classification, particularly due to class imbalance. Class balancing, performed through oversampling, was attempted but showed limited improvements. We conclude that while ML has potential in predicting heart disease presence, more data and advanced methods are needed for accurate severity classification.

1 Introduction

Cardiovascular disease is a leading cause of death worldwide, responsible for approximately 17.9 million deaths yearly, which accounts for roughly 32% of all global deaths (World Health Organization, 2021). Early and accurate diagnosis of heart disease can significantly reduce death rates and improve outcomes for patients. Traditional diagnostic methods, such as angiography and stress testing, are often invasive and expensive (Fihn et al., 2012). This makes them hard to access at times, specifically in places with low resources. Machine learning offers an alternative method by using clinical data to detect patterns and correlations that may not be apparent to human professionals. Prior studies have shown that machine learning models can effectively detect heart disease from clinical features (Dey et al., 2020), motivating the use of similar techniques in this project.

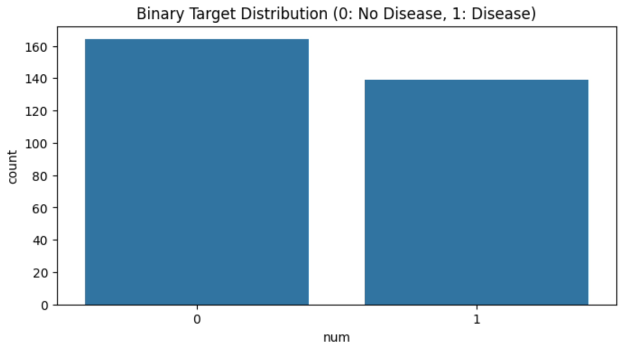
In this project, we aimed to apply machine learning techniques to predict the presence and severity of heart disease using clinical features from the UCI Heart Disease dataset (Dua and Graff, 2019). Specifically, we focused on two classification tasks: a binary classification task to predict whether a patient has heart disease, and a multiclass classification task to estimate the severity of disease on a scale from 0 (no disease) to 4 (most severe). We explored the use of three machine learning algorithms—Logistic Regression, Random Forest, and XGBoost—on both tasks, and we analyzed their performance to evaluate their potential effectiveness in real-world medical applications.

This project investigates the potential of machine learning to enhance clinical diagnoses, especially in situations where time and resources are limited. Additionally, we explored the limitations of using relatively small and imbalanced datasets for disease classification, which represents a real challenge in medical data analysis.

2 Methods

This study used the UCI Heart Disease dataset, which includes 303 patient records and 13 clinical features collected from routine check-ups and cardiovascular assessments. These features include patient demographics (age, sex), clinical symptoms (chest pain type), and test results (resting blood pressure, cholesterol levels, fasting blood sugar, electrocardiogram results, maximum heart rate achieved, ST depression, and thalassemia). The dataset also includes a target variable representing heart disease severity on a scale from 0 (no disease) to 4 (most severe).

For the binary classification task, we relabeled the target variable to indicate the presence or absence of heart disease, where class 0 represented no



heart disease while classes 1 through 4 represented the presence of heart disease in the placement. In Fig. 1, we can see that the binary classification problem is relatively balanced, with a slight majority of patients having no heart disease.

Fig. 1. Distribution of binary classification labels (disease vs. no disease).

.

For the multiclass classification task, we kept the original target labels from the dataset. However, as we can see in Fig. 2, the distribution of these classes is highly imbalanced. The overwhelming majority of patients belong in class 0, with all of the individual disease classes being severely underrepresented.

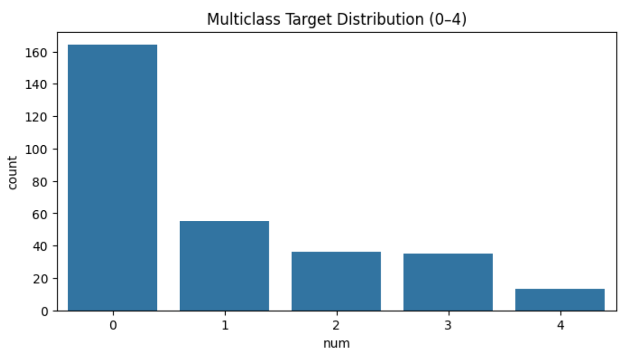


Fig. 2. Distribution of heart disease severity classification labels (0-4).

Before training the models, several preprocessing steps were applied. Missing values were imputed using the mean for all features. Continuous variables, were scaled using z-score normalization, to ensure that all features were on the same scale when used by our models.

To assess relationships between features, we computed a correlation matrix, as seen in Fig. 3, which helped identify associations between features and the target variable. Notably, the target variable (num) is positively correlated with features such as cp (chest pain type), ca (number of major vessels), and thal (thalassemia). On the other hand, thalach (maximum heart rate achieved) shows a negative correlation with num, indicating that patients with heart disease tend to have lower maximum heart rates. The heatmap suggests that these features are potentially important for predicting heart disease

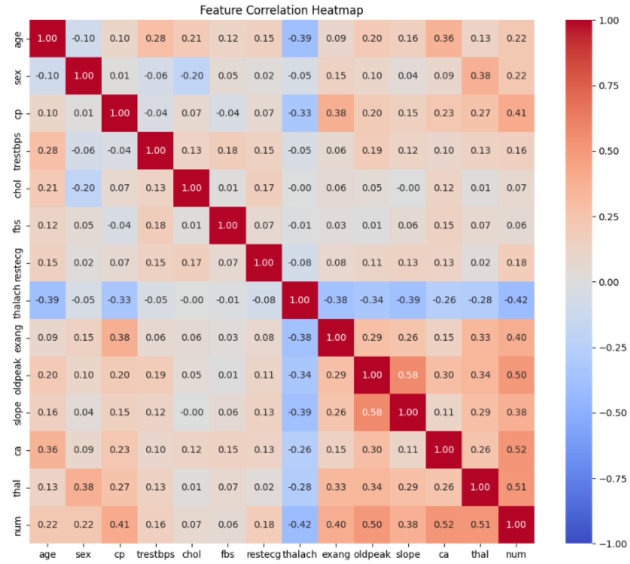


Fig. 3. Correlation heatmap of features in the dataset.

We implemented three machine learning models—Logistic Regression, Random Forest, and XGBoost—on both classification tasks. The dataset was split into 80% training and 20% testing sets. For the multiclass task, we trained the models on the original imbalanced dataset, as well as a balanced dataset obtained after performing oversampling on the original data, in an attempt to see the effects of class imbalance on the results.

3 Results

3.1 Binary Classification

All three models were first evaluated on the binary classification task, where the goal was to predict whether a patient had any form of heart disease. As we can see in Fig. 4, all models performed well, with extremely high accuracies Random Forest achieved the highest accuracy at 90%, followed closely by Logistic Regression at 89%, and XGBoost at 87%. This suggests that multiple different models, of significantly varying complexity, can all perform effecively on binary classification using clinical data.

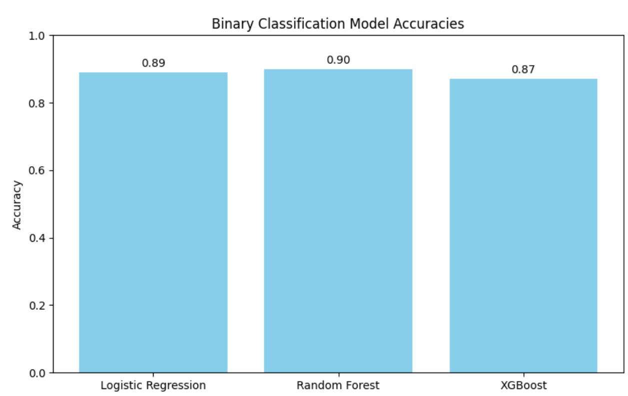


Fig. 4. Binary classification accuracy of Logistic Regression, Random Forest, and XGBoost models.

To better understand which features contributed most to model performance, we analyzed feature importances in the Random Forest model, as it had the highest accuracy. As we can see in Fig. 5, cp (chest pain type), oldpeak (ST Depression), thalach (maximum heart rate), and ca (number of vessels) were the top predictors of the presence of heart disease. These results correspond with the trends seen in the feature correlation heatmap, showing that our model prioritized the features that had the highest correlation with the target variable, which was expected.

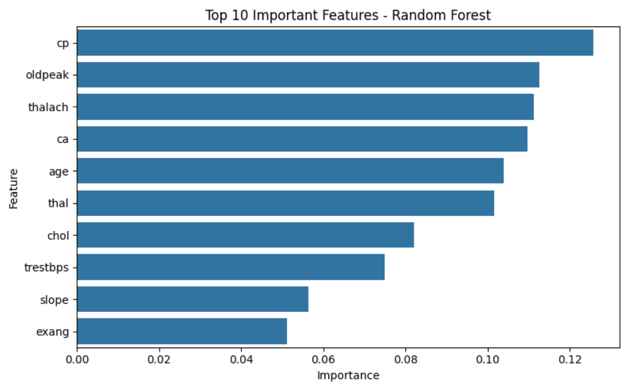


Fig. 5. Top 10 most important features in the Random Forest model for binary classification.

3.2 Multiclass Classification

All three models performed significantly worse in the multiclass classification task, where the goal was to predict the severity of heart disease on a scale from 0 (no disease) to 4 (most severe). As we can see in Fig. 6, Logistic Regression achieved the highest accuracy at 54%, followed by Random Forest at 49% and XGBoost at 48%. All three accuracies were significantly lower than the binary classification task, indicating that all models struggled with the increased complexity of multiclass classification.

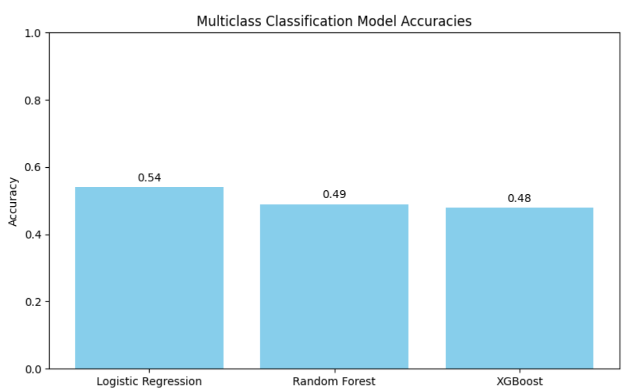


Fig. 6. Multiclass classification accuracy of Logistic Regression, Random Forest, and XGBoost models.

One likely cause of the low performance was class imbalance in the dataset, as the majority of samples belonged to class 0. To address this, we applied oversampling to the training set, producing the balanced class distribution shown in Fig. 7. After oversampling, all five severity classes were equally represented, which aimed to give the models more data points representing classes 1, 2, 3, and 4 of heart disease severity.

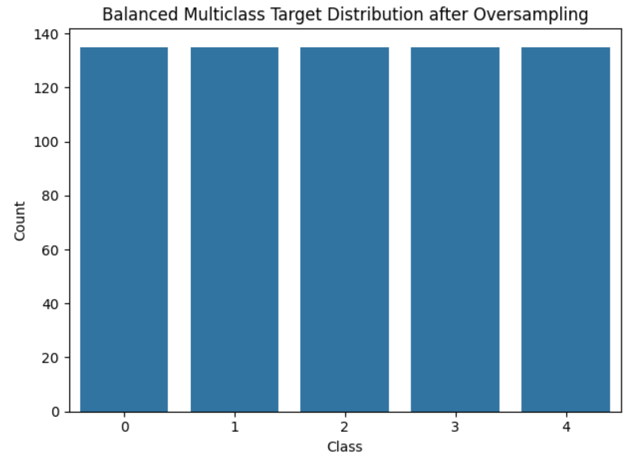


Fig. 7. Balanced class distribution for heart disease severity after oversampling

Despite balancing, only minor performance changes were observed. As shown in Fig. 8, Random Forest's accuracy improved marginally to 54%, while Logistic Regression dropped slightly to 46% and XGBoost remained at 48%. Although class balancing improved data representation, it was not very effective in improving model performance. This suggests that the multiclass prediction of heart disease severity is a complex task, and likely requires more sophisticated data or methods to properly handle.

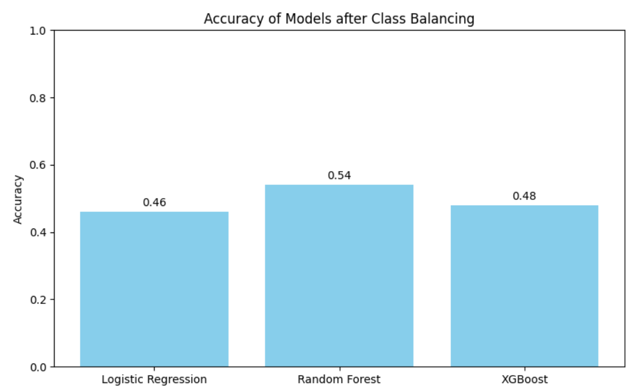


Fig. 8. Accuracy of models on multiclass classification after balancing the dataset.

3.3 Discussion

These findings demonstrate that while machine learning models are highly capable of detecting the presence of heart disease, predicting its severity remains much more challenging. Even after balancing the dataset, the models were unable to significantly improve their multiclass classification performance, suggesting that the current dataset may not contain enough information to distinguish between differences in disease severity. This may potentially be due to overlaps in the clinical data of patients across different severity classes. However, attempting the same techniques on an initially balanced dataset may still yield better results.

In future work, another possible direction is to include additional features which may better differentiate between stages of heart disease. Furthermore, applying more advanced modeling techniques such as neural networks or ensemble stacking could help improve multiclass performance.

Ultimately, achieving high performance in multiclass prediction may require a combination of better data and more robust models.

Acknowledgements

I would like to thank Professor Zhang and the COMPSCI 189 staff for giving me the opportunity to dive into my interest of bioinformatics and conduct this project, as well as for providing help and guidance throughout the entirety of this quarter.

References

Dey, D., Slomka, P.J., Leeson, P. et al. (2020) Artificial Intelligence in Cardiovascular Imaging: JACC State-of-the-Art Review. J Am Coll Cardiol, 76(11), 1318–1335.

Dua, D. and Graff, C. (2019) UCI Machine Learning Repository. University of California, Irvine. Available at: http://archive.ics.uci.edu/ml

Fihn, S.D., Gardin, J.M., Abrams, J. et al. (2012) 2012 ACCF/AHA/ACP/AATS/PCNA/SCAI/STS guideline for the diagnosis and management of patients with stable ischemic heart disease. Circulation, 126(25), e354–e471.

World Health Organization. (2021) Cardiovascular diseases (CVDs). WHO. Available at: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)