Predicting Heart Failure

Lavanya Pushparaj

Kaggle Dataset

December 21, 2023

Abstract

In this project, my aim is to leverage machine learning techniques to predict the likelihood of heart failure in patients based on a comprehensive dataset of patient information. The dataset, sourced from Kaggle, offers a rich array of features capturing various aspects of patients' health. The primary objectives encompass preprocessing the data to ensure its quality, training a machine learning model, and subsequently optimizing the model's performance. By implementing techniques such as cross-validation, hyperparameter tuning, and model interpretability, I strive not only to achieve higher accuracy compared to the previously reported results on Kaggle but also to gain insights into the key factors influencing heart failure prediction. This report details my methodology, results, and interpretations, providing a comprehensive overview of my approach and its implications for cardiovascular disease prediction.

Introduction

Taking a life every 33 seconds, cardiovascular diseases (CVDs) stand as the leading cause of death in the United States (CDC). Early detection and management are crucial for individuals with CVD or those at a heightened risk due to factors such as hypertension, diabetes, hyperlipidemia, or other pre-existing conditions. Additionally, factors such as tobacco use, unhealthy diet, obesity, physical inactivity, and excessive alcohol consumption can play a role in influencing a patient’s health status. Implementing population-wide strategies to mitigate these risks can significantly contribute to the prevention of heart failure and related cardiovascular complications. In this context, the application of machine learning models holds promising potential, providing a sophisticated approach to predicting and managing heart failure, ultimately contributing to improved healthcare outcomes and increased life expectancy.

Methods

First, the Kaggle dataset was loaded into a Pandas DataFrame, denoted as DF. To ensure data quality, entries with missing values were excluded using the dropna() function, yielding a dataset labeled as df. A correlation matrix was implemented to aid in the feature selection process. Next, an outlier removal function named remove_outliers was created. This function implemented both the Z-Score method and the Interquartile Range (IQR) method iteratively for each feature in the DataFrame. Consequently, outliers were identified and eliminated, contributing to a cleaner dataset. The data was visualized using boxplots before and after outlier removal.

Feature engineering is a pivotal aspect of data preprocessing, crucial in optimizing the dataset for improved model performance. In this instance, a categorical variable called 'age_group' was created. This feature segmented patient ages into predefined ranges, introducing a more interpretable representation of the original continuous 'age' feature. Beyond simplifying the model's comprehension of age-related patterns, this categorization also addressed potential non-linear relationships that might exist within distinct age intervals. Following this categorization, the 'age_group' feature underwent one-hot encoding, a process that translates categorical variables into a binary matrix. This prevents the model from assuming the hierarchical importance of numerical values, recognizing that, for example, '79' is not inherently more important than '40' simply because it has a higher numerical value. While 'age_group' captures categorical trends effectively, it is noteworthy that the original continuous 'age' feature has been retained. This decision is rooted in the recognition that the continuous 'age' information may hold valuable and fine-grained details that could be relevant to the model.

The features were systematically prepared for modeling by dividing them into the predictor variables, X, and the target variable, y, with X encompassing all features except 'Death_Event.' To gain a preliminary understanding of the relationships between each feature and the target variable, 'Death_Event,' scatter plots were created to explore potential associations visually.

Subsequently, the dataset was partitioned into training and testing sets using the train_test_split function. This division allocated 80% of the data for training the machine learning model and reserved the remaining 20% for subsequent evaluation, ensuring an unbiased assessment of the model's generalization performance. Then, a logistic regression model was built.

The choice to build a logistic regression model for this binary classification task was deliberate and grounded in its well-established effectiveness in scenarios where the target variable is dichotomous. Logistic regression excels in capturing the probability of an event's occurrence, making it particularly suitable for predicting binary outcomes. Its interpretability and simplicity make it an advantageous choice, especially when seeking insights into the relationship between features and the likelihood of a specific event, such as death. In the spirit of Occam's Razor, a principle emphasizing simplicity when faced with multiple competing explanations or models, the logistic regression model stands out for its straightforwardness and ease of interpretation.

To find an optimal balance, cross-validation with k=15 folds was intentionally selected through iterative experimentation. This approach allows for a robust assessment of the logistic regression model's performance on the training set. Accuracy scores were computed for each fold, and the mean accuracy was reported, providing a reliable measure of the model's overall predictive capability. Cross-validation is essential as it helps mitigate the impact of data variability and ensures that the model's performance metrics are representative across diverse subsets of the training data.

The final logistic regression model was initialized with an increased max_iter parameter for convergence and was trained on the entire training set. Model evaluation was conducted on the reserved test set, encompassing accuracy computation, classification report generation, and constructing a confusion matrix. These evaluation metrics collectively provide insights into the model's effectiveness in predicting the likelihood of heart failure in patients. Additionally, Shapley values were computed to interpret the predictions of a logistic regression model. Background samples from the training data were used to generate Shapley values for each feature. The shap.summary_plot illustrates the impact of each feature on predictions with a bar plot. This Shapley value analysis is crucial for understanding the model's

decision-making, identifying influential features, and enhancing overall interpretability in the logistic regression model.

Results

After uploading the DataFrame and removal of the missing values the correlation matrix was created.

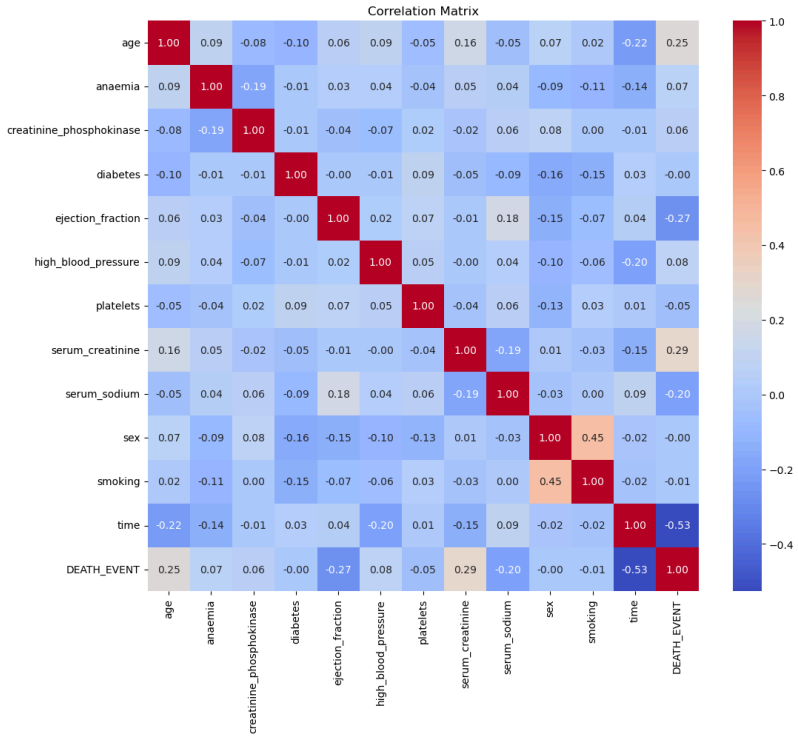


Fig 1: Correlation Matrix

Next, the outliers were removed, and the before and after were visualized using boxplots.

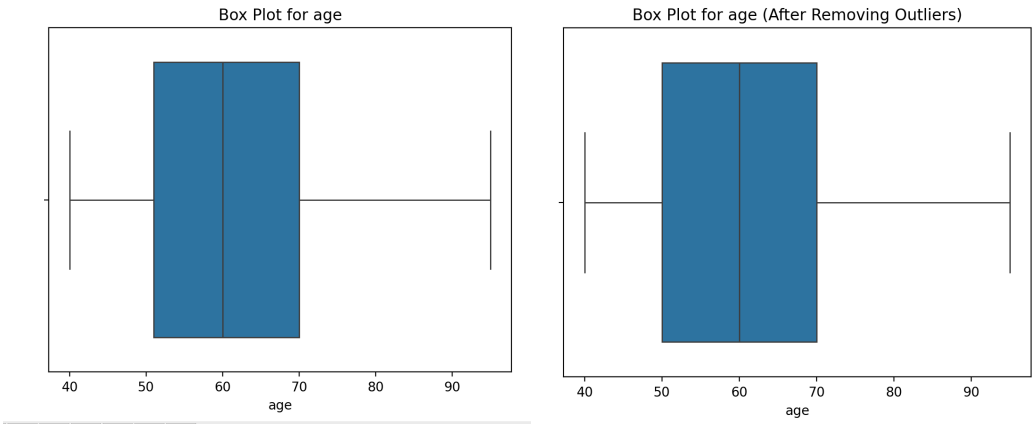


Fig 2: Box Plots of Age

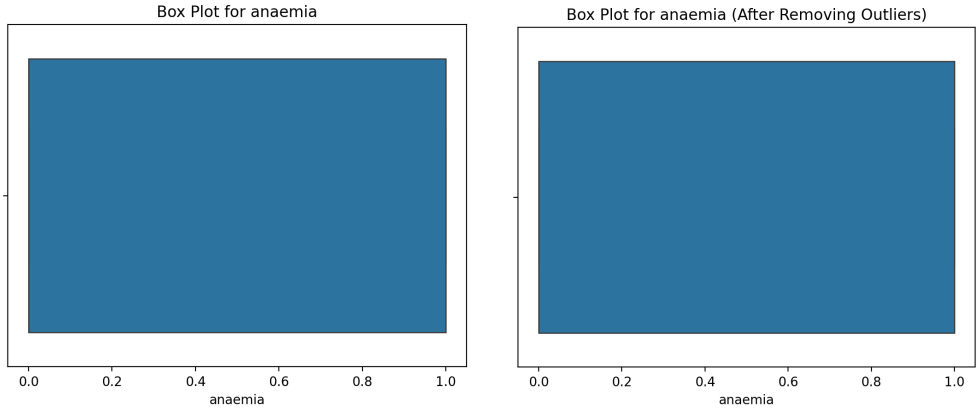


Fig 3: Box Plots of Anemia

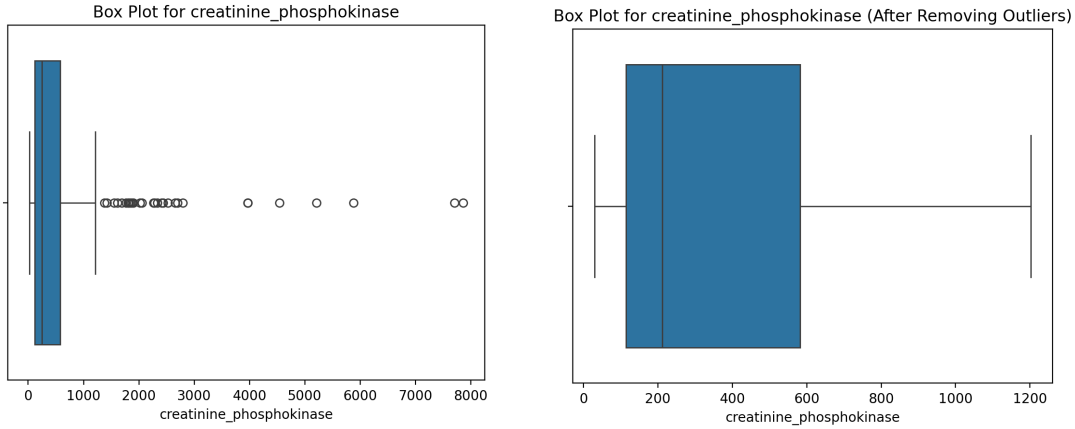


Fig 4: Box Plots of Creatinine Phosphokinase

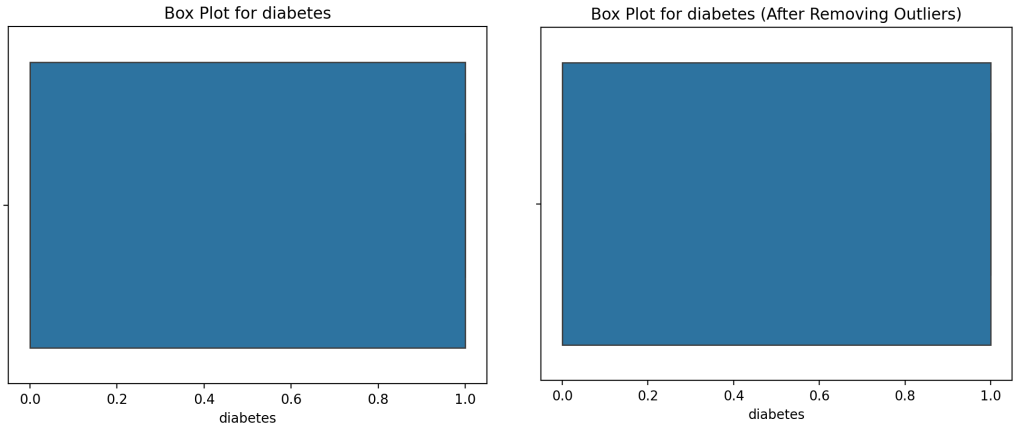


Fig 5: Box Plots of Diabetes

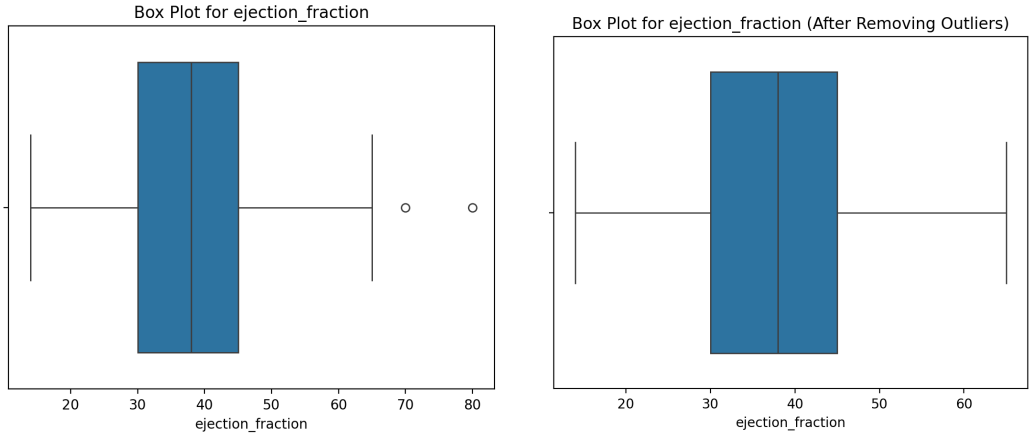


Fig 6: Box Plots of Ejection Fraction

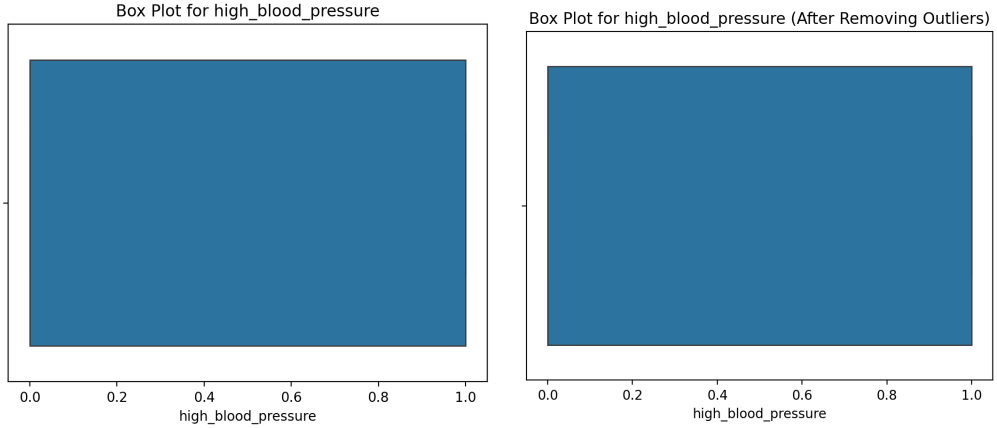


Fig 7: Box Plots of High Blood Pressure

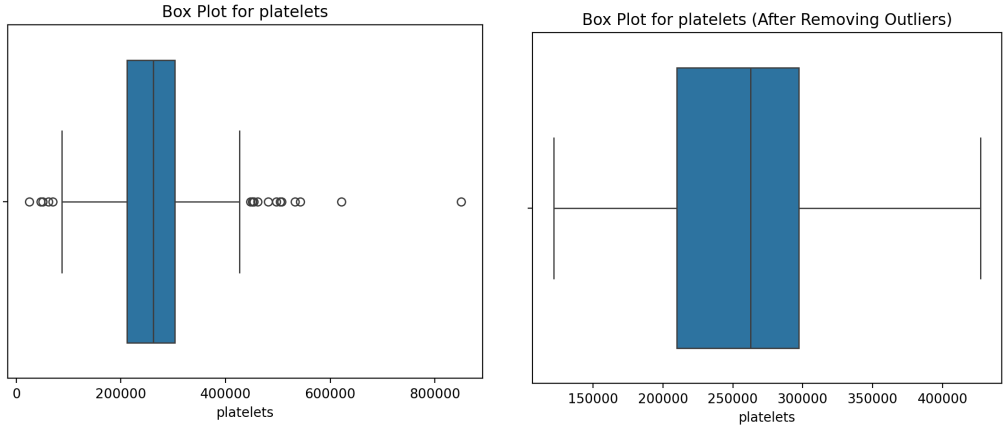


Fig 8: Box Plots of Platelets

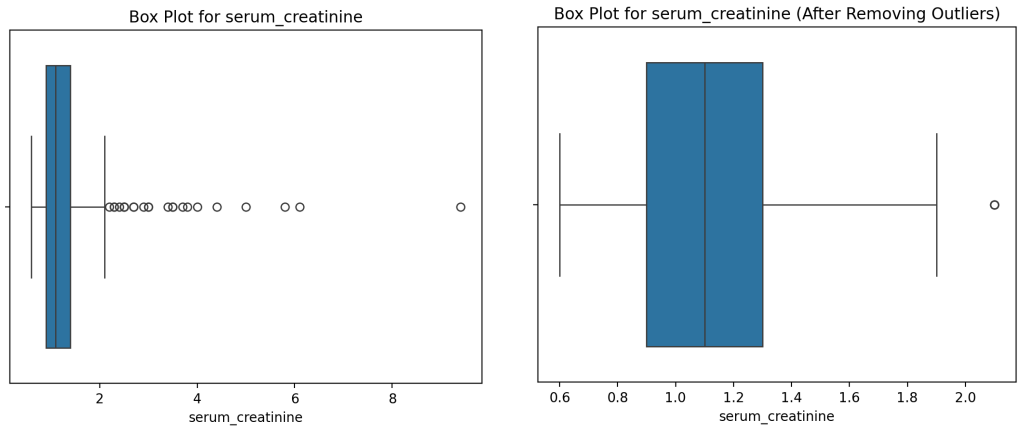


Fig 9: Box Plots of Serum Creatinine

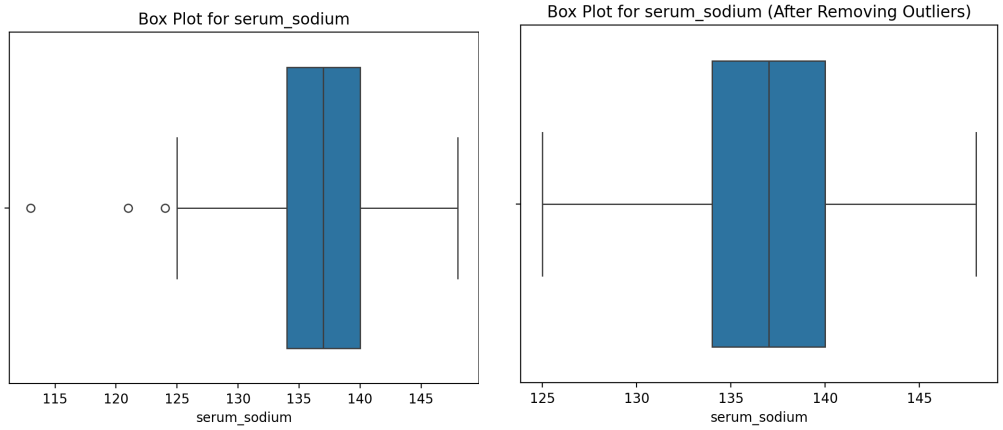


Fig 10: Box Plots of Serum Sodium

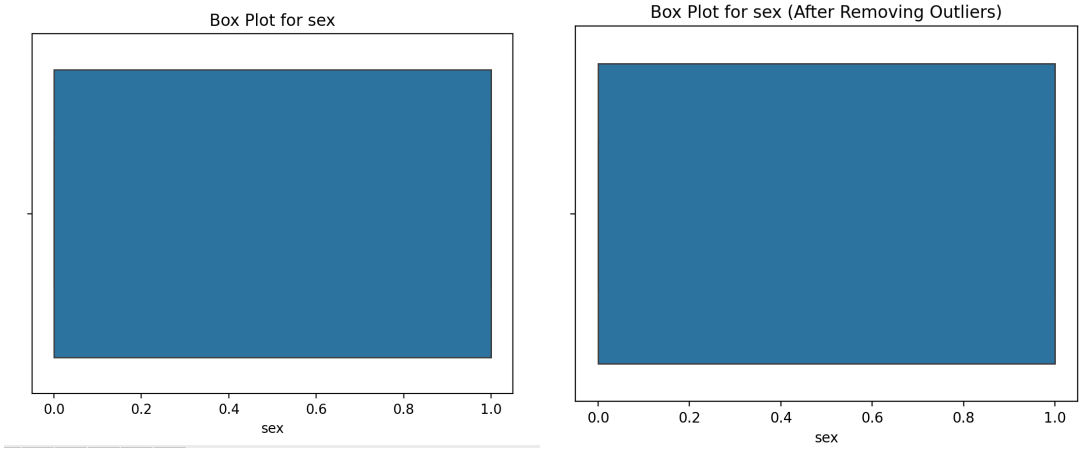


Fig 11: Box Plots of Sex

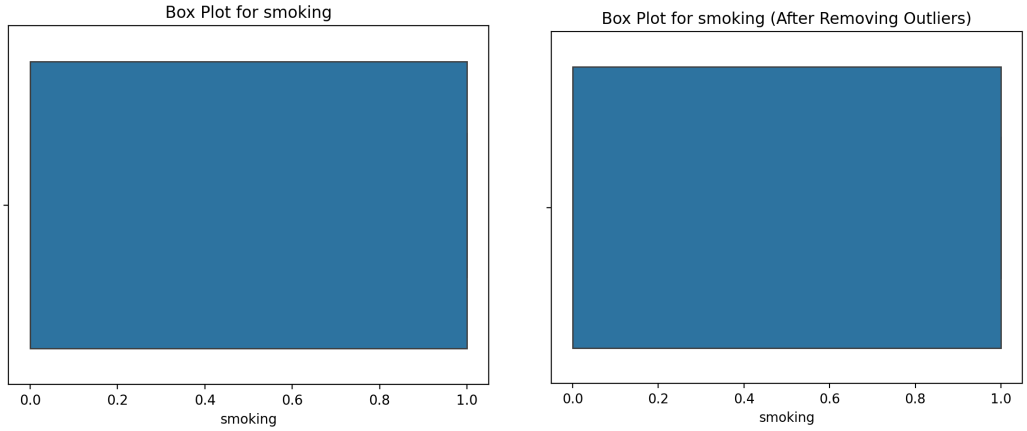


Fig 12: Box Plots of Smoking

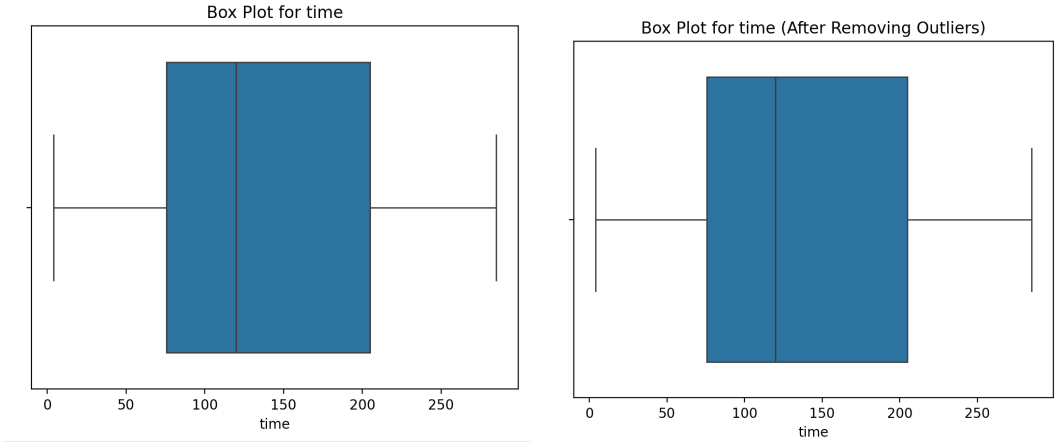


Fig 13: Box Plots of Time

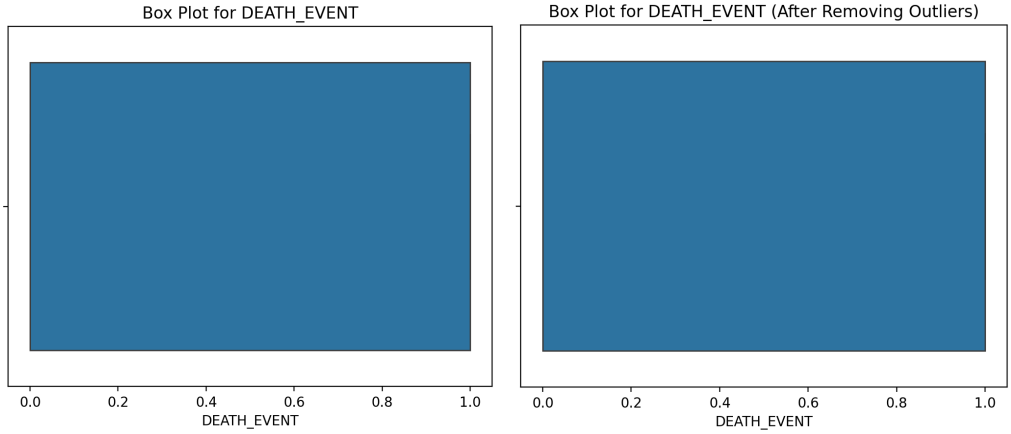
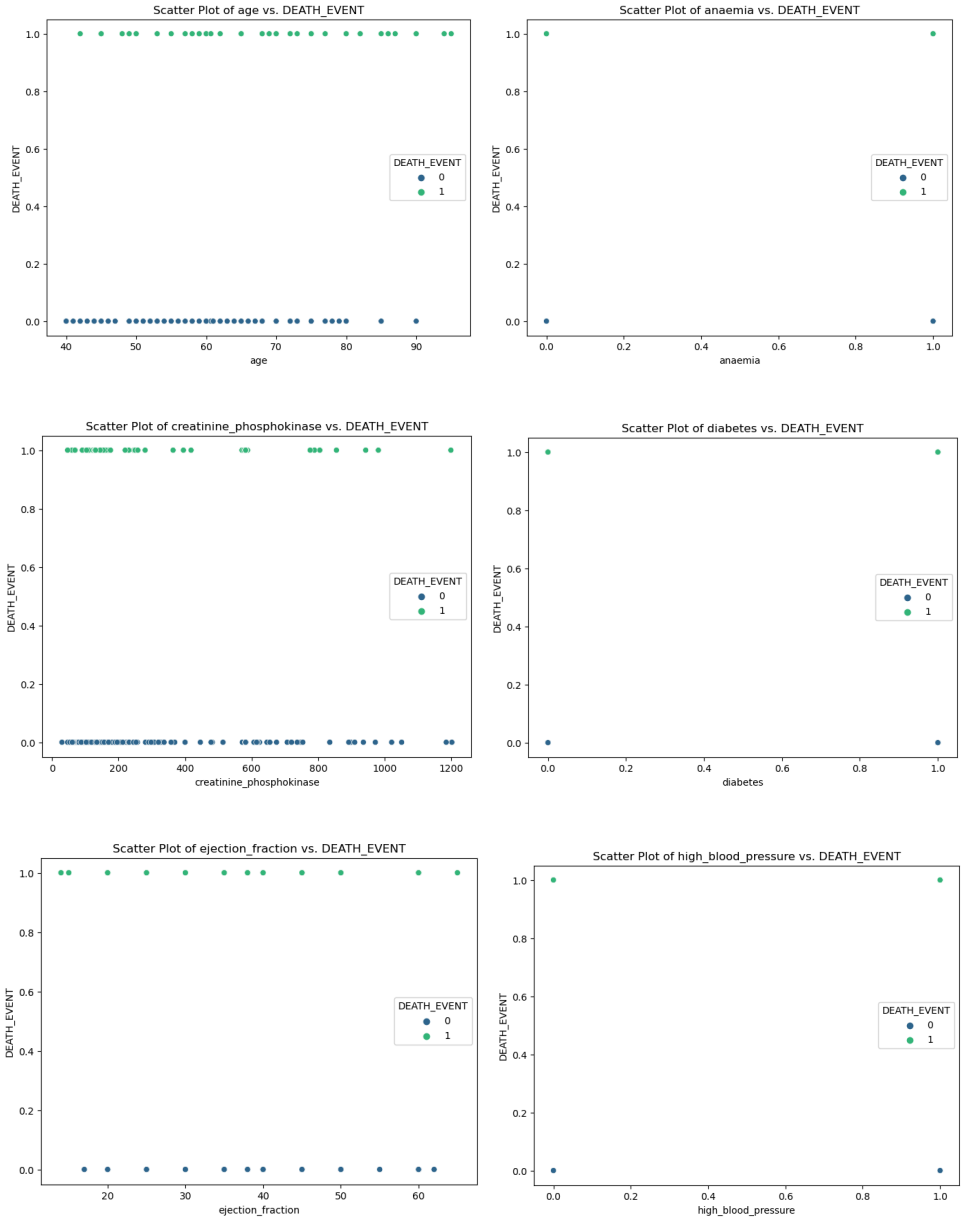


Fig 14: Box Plots of Death Event

Using this, scatter plots were created to visualize the correlation between each feature and the target value.



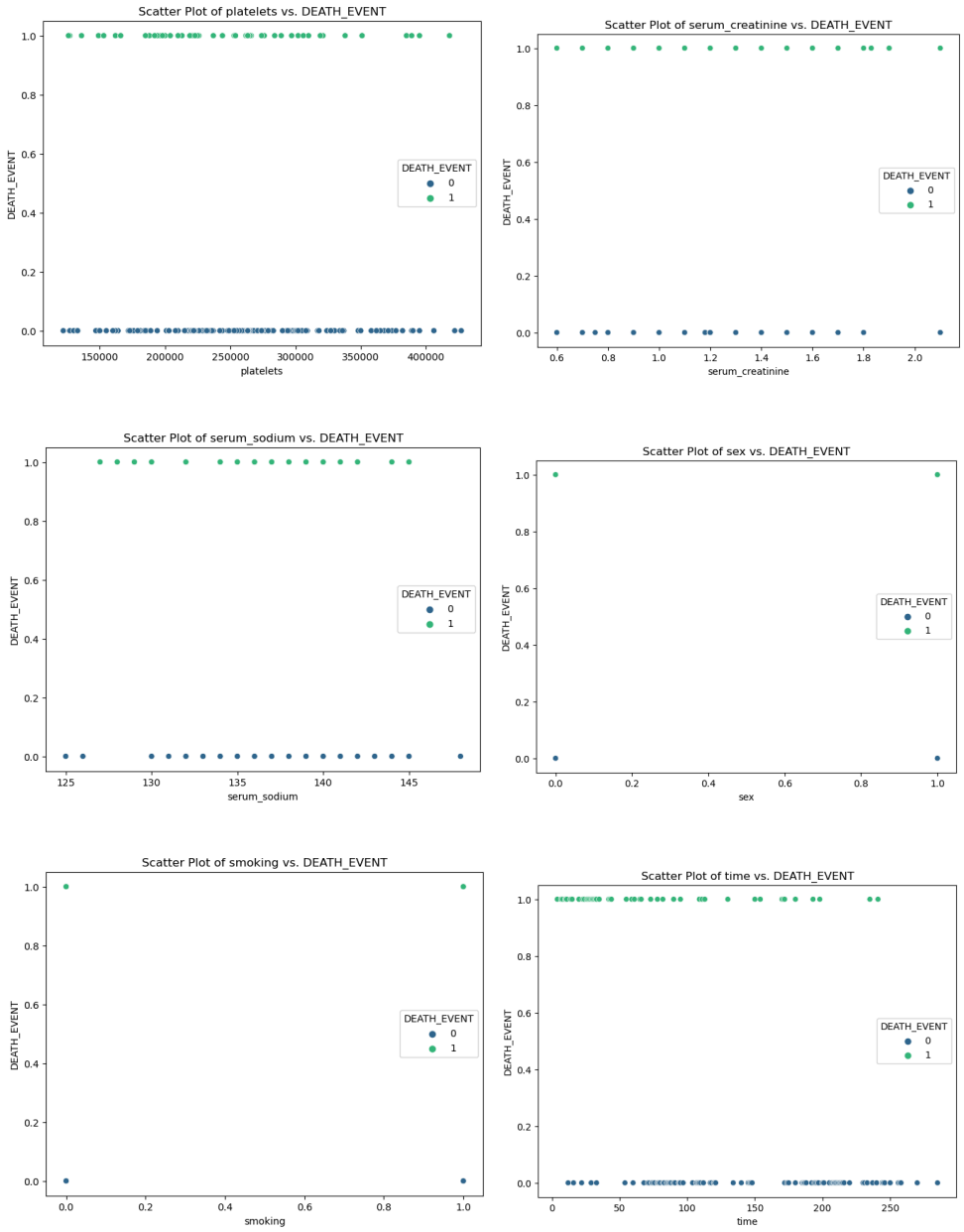


Fig 15: Scatter Plots

After the linear regression model was constructed, the cross validation scores were calculated: 0.91666667, 0.75, 1., 0.66666667, 1., 0.91666667, 0.91666667, 0.83333333, 0.83333333, 0.83333333, 0.75, 0.91666667, 0.91666667, 0.66666667, 0.90909091. Which resulted in a mean accuracy of 0.86. The classification report reveals the performance metrics of the logistic regression model on the test set:

- Class 0 (No death event): Precision = 0.87, Recall = 0.97, F1-Score = 0.92
- Class 1 (Death event): Precision = 0.86, Recall = 0.55, F1-Score = 0.67

The overall accuracy of the model is 0.87, indicating the proportion of correctly predicted instances across both classes. The macro average F1-score is 0.79, and the weighted average F1-score is 0.86, reflecting a balanced and effective predictive performance.

A confusion matrix was constructed.

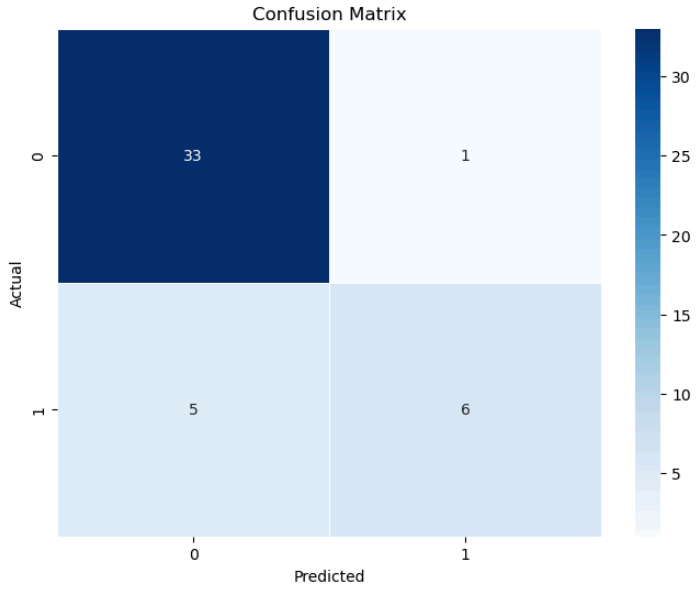


Fig 16: Confusion Matrix

Outcomes from the Shapley analysis were plotted.

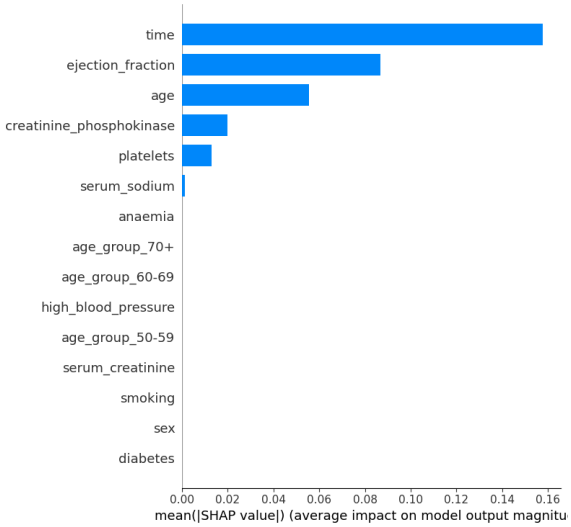


Fig 18: Shapley Analysis

Discussion

Features of Importance

Based on the Shapley result, the feature that had the highest influence on the model was 'time,' which indicates the follow-up period in days. The identification of 'time' as the feature with the highest impact on the logistic regression model holds significant implications for our understanding of heart disease dynamics. This result suggests that the temporal aspect plays a pivotal role in predicting the likelihood of adverse outcomes related to heart disease. The finding underscores the importance of continuous monitoring and long-term management strategies in assessing and mitigating risks. Furthermore, this aligns with past research that discovered early follow-ups after heart failure resulted in better outcomes (McAlister, 2016). This insight has the potential to enhance proactive patient care, facilitating individualized risk assessments based on the temporal evolution of heart disease. The prominence of the follow-up period as a key predictor in the model emphasizes the intricate relationship between time, disease progression, and mortality risk, offering valuable insights for refining patient management protocols and informing future research endeavors.

The second most influential feature is 'ejection fraction.' This finding suggests that the efficiency of cardiac ejection fraction, as represented by the percentage of blood leaving the heart at each contraction, is a significant predictor of adverse outcomes in cardiovascular health. A reduced ejection fraction, indicating compromised pumping efficiency, may contribute to poorer cardiovascular outcomes and an elevated risk of mortality (American Heart Association, 2023). This correlation underscores the clinical importance of assessing cardiac function, particularly the heart's ability to pump blood effectively, as a key determinant of patient prognosis. Identifying this feature as highly impactful in predicting death events highlights the need for vigilant monitoring and interventions to preserve or improve cardiac function. It also underscores the potential role of therapeutic strategies targeted at enhancing ejection fraction in mitigating the risks associated with cardiovascular diseases. Current research is working on addressing interventions that can reduce the effects of preserved ejection fraction (Fan, 2021).

Model

The construction of the logistic regression model yielded promising results, with cross-validation scores demonstrating consistency across different folds. This consistency in cross-validation scores suggests that the model generalizes well across various subsets of the training data.

The subsequent evaluation of the test set, as outlined in the classification report, provides a more detailed picture of the model's performance. For instances where no death event occurred (Class 0), the model demonstrated high precision (0.87) and recall (0.97), resulting in an F1-score of 0.92. This indicates a robust ability to correctly identify instances with no death event, striking a balanced precision-recall trade-off. However, the model exhibited slightly lower performance for instances indicating a death event (Class 1). The precision for Class 1 was 0.86, signifying that when the model predicted a death event, it was correct 86% of the time. The recall for Class 1 was 0.55, indicating that the model captured 55% of the actual death events. The corresponding F1 score for Class 1 was 0.67, reflecting a trade-off between precision and recall.

The model's overall accuracy on the test set was 0.87, emphasizing the proportion of correctly predicted instances across both classes. The macro average F1-score, emphasizing equal importance to

both classes, was 0.79, while the weighted average F1-score, accounting for class imbalance, was 0.86. These scores collectively suggest a balanced and effective predictive performance of the logistic regression model.

Several areas could be considered for improvement to enhance the model's performance in future iterations. Firstly, it would be beneficial to explore additional feature engineering techniques or introduce new relevant features that could provide richer insights into the complexity of cardiovascular outcomes. Furthermore, assessing the model's sensitivity to outliers could be extended by refining the outlier removal process. Experimenting with different outlier detection methods or adjusting the threshold values could improve the model's robustness to extreme data points.

Another course of action for improving the model is fine-tuning hyperparameters. While this model has a high accuracy, there is room for improvement through a more thorough exploration of hyperparameter settings. Systematic experimentation with regularization strengths, solver algorithms, and other logistic regression parameters could yield configurations that enhance the model's predictive performance. Additionally, adopting techniques such as grid search or randomized search during hyperparameter tuning could systematically search through the hyperparameter space, identifying combinations that lead to superior model outcomes.

Model selection is another avenue for improvement. While logistic regression is a robust choice, exploring alternative algorithms such as decision trees, random forests, or gradient boosting could unveil potential improvements.

Conclusion

In conclusion, the model can predict the likelihood of death events with notable precision and recall for instances across both classes. The two features that had the highest influence on the 'death event' were 'time' and 'ejection fraction.' Further refinement and exploration of model parameters may enhance specific aspects of its performance, and future investigations could explore additional features or alternative models to improve predictive capabilities.

References

American Heart Association. (2023). Ejection Fraction (Heart Failure Measurement). Retrieved from https://www.heart.org/en/health-topics/heart-failure/diagnosing-heart-failure/ejection-fraction-heart-failure-measurement

Centers for Disease Control and Prevention. (n.d.). Heart Disease Facts. Retrieved from https://www.cdc.gov/heartdisease/facts.htm

McAlister, F. A., Youngson, E., Kaul, P., & Ezekowitz, J. A. (2016). Early Follow-Up After a Heart Failure Exacerbation: The Importance of Continuity. Circulation. Heart failure, 9(9), e003194. https://doi.org/10.1161/CIRCHEARTFAILURE.116.003194

Tsao, C. W., Aday, A. W., Almarzooq, Z. I., Anderson, C. A. M., Arora, P., Avery, C. L., Baker-Smith, C. M., Beaton, A. Z., Boehme, A. K., Buxton, A. E., Commodore-Mensah, Y., Elkind, M. S. V., Evenson, K. R., Eze-Nliam, C., Fugar, S., Generoso, G., Heard, D. G., Hiremath, S., Ho, J. E., Kalani, R., … American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee (2023). Heart Disease and Stroke Statistics-2023 Update: A Report From the American Heart Association. Circulation, 147(8), e93–e621. https://doi.org/10.1161/CIR.0000000000001123

Fan, Y., & Pui-Wai Lee, A. (2021). Valvular Disease and Heart Failure with Preserved Ejection Fraction. Heart failure clinics, 17(3), 387–395. https://doi.org/10.1016/j.hfc.2021.02.005