Heart Failure Mortality Analysis

Dexter Schincke

Bellevue University

DSC 680

Amirfarrokh Iranitalab

23 February 2025

Heart Failure Mortality Analysis

**Business Problem:**

Heart failure is one of the leading causes of mortality worldwide, with early diagnosis playing a pivotal role in improving patient outcomes. By predicting mortality risks based on clinical and demographic data, healthcare providers can prioritize interventions for high-risk patients. This project focuses on developing a predictive model to forecast heart failure-related mortality, aiding healthcare professionals in decision-making.

**Background/History:**

Heart failure affects millions globally and is associated with high morbidity and mortality. Early prediction and identification of patients at high risk for mortality can improve care management, reduce hospitalization rates, and ultimately save lives. Previous studies have demonstrated that clinical features such as ejection fraction, age, and comorbidities (e.g., diabetes and hypertension) contribute to predicting patient outcomes. However, a significant challenge remains in accurately predicting mortality based on these features. Machine learning offers a promising avenue to improve predictive accuracy and assist healthcare providers in delivering personalized care.

**Data Explanation:**

The dataset for this project consists of 299 rows and 14 features, including both clinical and demographic data, with the target variable DEATH\_EVENT indicating whether a patient died during the study period (binary outcome: 1 = deceased, 0 = survived). There are no missing values, and the dataset features both numerical and categorical variables. Key features include:

* Age: Patient’s age.
* Anaemia: Binary variable (0 = no, 1 = yes) indicating anaemia status.
* Creatinine phosphokinase: An enzyme that can indicate heart muscle damage.
* Diabetes: Binary variable (0 = no, 1 = yes) indicating diabetes status.
* Ejection fraction: The percentage of blood pumped out of the heart with each contraction.
* High blood pressure: Binary variable (0 = no, 1 = yes) indicating hypertension.
* Platelets: The number of platelets in the blood, crucial for blood clotting.
* Serum creatinine: Kidney function marker.
* Serum sodium: Sodium levels in the blood.
* Sex: Patient’s gender (0 = female, 1 = male).
* Smoking: Binary variable (0 = no, 1 = yes) indicating smoking status.
* Time: Duration of observation in days.
* Death Event: Binary variable (0 = no, 1 = yes).

**Methods:**

The dataset used in this study contains various clinical and demographic features, with the primary objective being the prediction of mortality in heart failure patients. A correlation matrix was computed to understand the relationships among the features and the target variable, DEATH\_EVENT (which indicates mortality). The correlation matrix revealed several notable findings. Age was found to have a moderate positive correlation with mortality, with a coefficient of 0.2538, suggesting that older age is associated with a higher likelihood of death. Serum creatinine exhibited a stronger positive correlation with mortality, having a coefficient of 0.2943, which indicates that elevated creatinine levels are linked to an increased risk of mortality. On the other hand, ejection fraction displayed a negative correlation with mortality, with a coefficient of -0.2686, indicating that lower ejection fractions are associated with higher mortality. These findings are essential for the predictive modeling process, as they suggest that certain clinical features such as age, serum creatinine, and ejection fraction may significantly influence mortality outcomes.

The distribution of death events in the dataset reveals a significant imbalance, with approximately 200 individuals having no mortality (DEATH\_EVENT = 0) and 100 individuals experiencing mortality (DEATH\_EVENT = 1). This imbalance must be addressed during the modeling phase to prevent the model from being biased toward the majority class.

A graph of a number of blue squares

Description automatically generated

An analysis of the sex variable and its relationship with mortality shows that males tend to have a higher proportion of mortality cases. Specifically, among females (sex = 0), approximately 70 individuals survived, while 35 females died. In contrast, for males (sex = 1), around 130 males survived, and 60 males died. This suggests that gender may influence mortality rates, which could be a key factor to include in the predictive models.

A graph of a number of people

Description automatically generated with medium confidence

Similarly, the relationship between smoking and mortality reveals that non-smokers (smoking = 0) have a higher survival rate. Among non-smokers, approximately 138 individuals survived, and 65 individuals died. In contrast, 65 smokers survived, and 30 smokers died. This suggests that smoking status is a potential risk factor for mortality and should be considered in the modeling process.

A graph of smoking vs mortality

Description automatically generated

The age distribution of the dataset shows that there are certain age groups with higher frequencies. Notably, there are clusters of individuals at the ages of 50 years, 60 years, 65 years, and 70 years, with approximately 30 individuals at age 50, 40 individuals at age 60, 30 at age 65, and 35 at age 70. This indicates that age is a critical factor in the analysis of mortality and may influence the predictive modeling results.

A graph of age distribution

Description automatically generated

A box plot comparing age and DEATH\_EVENT provides additional insights into the relationship between age and mortality. For individuals who did not die (DEATH\_EVENT = 0), the median age is approximately 60, with the upper quartile around 65 and the lower quartile around 50. For individuals who died (DEATH\_EVENT = 1), the median age is around 65, with the upper quartile around 75 and the lower quartile around 55. This indicates that older individuals are more likely to experience mortality, reinforcing the findings from the correlation analysis.

The box plot of ejection fraction versus DEATH\_EVENT shows that non-survivors (DEATH\_EVENT = 0) tend to have a higher median ejection fraction, with a median value around 38. The upper quartile is approximately 45, and the lower quartile is around 35. In contrast, individuals who died (DEATH\_EVENT = 1) have a lower median ejection fraction of approximately 30, with the upper quartile around 38 and the lower quartile around 25. This suggests that lower ejection fractions are strongly associated with higher mortality, which highlights the clinical importance of this feature.

**Analysis:**

After completing the exploratory data analysis (EDA) and preparing the dataset, various machine learning models were applied to predict DEATH\_EVENT (mortality). The models considered in this study include Logistic Regression, Random Forest, and XGBoost. These models were evaluated based on their ability to predict mortality accurately and their performance on several evaluation metrics.

The Logistic Regression model achieved an accuracy of 0.8000, indicating that 80% of predictions were correct. However, its recall was relatively low at 0.5600, meaning that the model missed around 44% of the actual mortality cases. The precision of the Logistic Regression model was high at 0.9333, meaning that when the model predicted mortality, it was correct approximately 93% of the time. The F1-score, which balances both precision and recall, was 0.7000.

The Random Forest model performed slightly worse than Logistic Regression, with an accuracy of 0.7500. The precision was 0.8571, and the recall was lower at 0.4800, meaning that it missed 52% of the true positive mortality cases. The F1-score was 0.6154, indicating a moderate balance between precision and recall.

The XGBoost model showed the most promising results in terms of recall, achieving a recall of 0.6000, meaning it correctly identified 60% of the mortality cases. Its accuracy was 0.7667, and its precision was 0.7895. The F1-score was 0.6818, indicating a solid balance between precision and recall.

Among the models tested, XGBoost performed the best in terms of recall, which is crucial when predicting mortality, as it minimizes the risk of false negatives. Despite its slightly lower precision than Logistic Regression, XGBoost achieved the highest F1-score, indicating a better balance between precision and recall. Therefore, XGBoost is considered the most effective model for predicting mortality in this dataset.

**Conclusion:**

The analysis of the heart failure dataset and subsequent model evaluations highlighted several key factors influencing mortality in heart failure patients. The correlation analysis revealed significant relationships between clinical features such as age, serum creatinine, and ejection fraction with the mortality outcome. Age was positively correlated with mortality, as expected, with older individuals being at greater risk. Additionally, serum creatinine levels and ejection fraction were important predictors, with lower ejection fractions being associated with a higher likelihood of death. The categorical features, such as sex and smoking, further emphasized the impact of demographic and lifestyle factors, with males and smokers showing higher mortality rates.

Three machine learning models were employed—Logistic Regression, Random Forest, and XGBoost—to predict mortality. Logistic Regression performed well in terms of accuracy (80%) but had a relatively low recall (56%), indicating that it struggled to detect mortality cases. The Random Forest model also demonstrated a solid precision of 85.7%, but its recall was lower at 48%, suggesting that it missed a significant portion of the mortality cases. XGBoost outperformed the others with the highest recall of 60%, making it the most effective model for identifying high-risk patients. This higher recall is crucial in a clinical setting where false negatives (missed mortality cases) could lead to inadequate intervention.

The findings highlight the importance of using models that prioritize recall in imbalanced datasets, especially in healthcare scenarios where early detection of high-risk patients can save lives. XGBoost, with its higher recall and balanced performance, proved to be the most reliable model for mortality prediction in this context. Future improvements could involve further tuning the models, addressing data imbalance with techniques like SMOTE, or exploring other advanced models to enhance performance even further.

A blue squares with white text

Description automatically generated

Overall, the analysis underscores the utility of machine learning in healthcare applications, particularly in predicting mortality risks for heart failure patients. The insights from this study can aid healthcare professionals in making more informed decisions, ultimately improving patient outcomes through early detection and timely intervention.

**Assumptions:**

In conducting this analysis, several assumptions were made regarding the data and modeling process. Firstly, it was assumed that the data provided by the dataset was complete and accurately reflected the clinical conditions of heart failure patients. There was also the assumption that the relationships between the features and the target variable (mortality) were linear or approximately linear, particularly for the machine learning models used, such as Logistic Regression. Another assumption was that the features, though anonymized, were relevant and sufficient to predict mortality, which limited the exploration of additional factors that might have affected outcomes. Finally, it was assumed that all patients in the dataset had received similar levels of care and treatment, though no information on this was included in the dataset.

**Limitations:**

This analysis faced several limitations due to the nature of the dataset and the scope of the project. One of the primary limitations was the relatively small sample size of 299 observations, which might not fully represent the entire population of heart failure patients. The dataset also lacked detailed information on treatment plans, comorbidities, and other health factors, which could have provided a more comprehensive view of the mortality risk. Additionally, some features, like creatinine phosphokinase or serum sodium, had weak correlations with mortality, which suggests that other unobserved factors might play a more significant role in predicting death than what was captured. Moreover, the dataset’s binary classification of mortality could be limiting since more granular outcomes (e.g., severity of complications or stages of heart failure) were not included.

**Challenges:**  
 The primary challenge encountered in this project was the class imbalance in the target variable, where the number of patients who survived (death event = 0) was significantly higher than those who did not survive (death event = 1). This imbalance made it difficult for models to predict mortality accurately, as they were biased toward predicting survival. Despite using different machine learning algorithms, including Logistic Regression, Random Forest, and XGBoost, achieving a high recall for the mortality class remained challenging. Additionally, the model selection process involved trade-offs between precision and recall, as optimizing one often led to a decrease in the other.

**Future Uses:**

The model developed in this analysis has several potential future applications in healthcare. It could be integrated into clinical decision support systems to assist healthcare professionals in identifying patients at high risk of mortality due to heart failure. This tool could help in prioritizing treatment plans and resources, especially in emergency and intensive care settings. Furthermore, the model could be expanded to include additional features, such as medications, genetic data, or clinical interventions, to improve the prediction accuracy and applicability in real-world scenarios. A promising area for future research would be to apply this analysis to other chronic diseases, such as diabetes or kidney failure, where similar predictive models could enhance early detection and intervention.

**Recommendations:**  
 Based on the analysis and model performance, several recommendations can be made for improving mortality prediction in heart failure patients. First, it is recommended to focus on improving data quality by gathering more comprehensive datasets that include clinical treatment histories, patient monitoring, and lifestyle data, which could enrich the predictive power of the models. Second, addressing the class imbalance through resampling techniques, such as Synthetic Minority Over-sampling Technique (SMOTE), could improve the model's recall without sacrificing precision. Additionally, exploring ensemble models and advanced techniques, such as deep learning, may yield more robust performance. Finally, it is recommended to conduct cross-validation on larger and more diverse datasets to ensure the model's generalizability.

**Implementation Plan:**  
 To implement this model in a clinical setting, a phased approach is recommended. The first phase would involve integrating the predictive model into an existing clinical decision support system as a pilot program, targeting hospitals or healthcare centers specializing in heart failure care. During this phase, the model’s predictions would be reviewed by healthcare professionals to validate its usefulness and accuracy in real-world settings. Afterward, in the second phase, the model could be refined based on user feedback, and additional features could be added to enhance its performance. The final phase would involve scaling the model to more healthcare institutions, with continuous monitoring to ensure its effectiveness and adapt to new data.

**Ethical Assessment:**

The ethical implications of using machine learning models in healthcare are crucial and need to be carefully considered. One of the primary ethical concerns is data privacy, especially considering that patient data is often sensitive and can include health conditions, treatment details, and personal identifiers. It is essential to ensure that any data used is anonymized and handled following strict data protection regulations such as HIPAA in the United States or GDPR in the European Union. Moreover, model bias is another significant ethical concern. If the model is not trained on diverse datasets that represent all demographics equally, it could lead to biased predictions that disproportionately affect certain groups. Finally, the implementation of this model in clinical practice should be done with transparency, ensuring that healthcare professionals understand how predictions are made and can make informed decisions based on the model’s output rather than relying solely on it.

**Potential Questions:**

1. What was the primary objective of your heart failure mortality analysis, and why is it important for healthcare professionals?
2. How did you address the class imbalance in your dataset, and what effect did it have on the performance of your models?
3. Which features were most influential in predicting mortality, and how did you determine their importance?
4. Why did you choose Logistic Regression, Random Forest, and XGBoost as the machine learning algorithms for this analysis?
5. How did you evaluate the effectiveness of each model, and which one performed best overall?
6. What challenges did you encounter during data preprocessing, and how did you overcome them?
7. What role did exploratory data analysis (EDA) play in shaping your modeling approach?
8. How can the findings from your analysis be used to inform healthcare practices or improve patient outcomes?
9. What are the limitations of your model, and how could these be addressed in future work or through further data collection?
10. How did you ensure the ethical use of patient data throughout the project, and what considerations should be made when applying the model in real-world scenarios?

References

**Projects/Papers:**

Mustanger. (2022, December 1). heart failure: Eda and prediction. Kaggle.

<https://www.kaggle.com/code/eisgandar/heart-failure-eda-and-prediction#3%7CLOAD->DATASET

RICCIARDI, C. (2025, January 22). Heart failure prediction. Kaggle. https://www.kaggle.com/code/anacpricciardi/heart-failure-prediction

**Datasets:**

Davide Chicco, Giuseppe Jurman: Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Medical Informatics and Decision Making 20, 16 (2020).

<https://www.kaggle.com/code/eisgandar/heart-failure-eda-and-prediction#3%7CLOAD->DATASET