Project 3: Machine Learning for Mortality Prediction in Heart Failure Patients

Xiaocong Xuan (xx2438)

December 15, 2024

1 Abstract

This project explores the use of machine learning models to predict mortality in heart failure patients based on structured clinical and demographic data. Three models: Logistic Regression, Random Forest, and a Neural Network, were developed and evaluated using ROC-AUC and Positive Predictive Value as key performance metrics. Among these, the Random Forest model delivered the best results, achieving the highest predictive accuracy and identifying the most significant predictors of mortality. These findings highlight the potential of machine learning, particularly ensemble methods like Random Forest, to provide accurate and interpretable predictions that can support clinical decision-making and improve patient care.

2 Introduction

Heart failure is a complex and challenging clinical condition in which the heart cannot pump blood effectively, leading to inadequate delivery of oxygen and nutrients to the body's tissues. It is a major global health issue, accounting for significant morbidity and mortality. Patient outcomes in heart failure vary widely, influenced by a combination of clinical and demographic factors. Traditional prediction tools, such as the Seattle Heart Failure Model (SHFM) and the Meta-Analysis Global Group in Chronic Heart Failure (MAGGIC) risk score, have been widely used to estimate patient prognosis. These models rely on predefined statistical frameworks and often assume linear relationships among variables. While effective to some extent, they may lack the precision required to address the complex, nonlinear dynamics present in diverse patient populations.

Machine learning offers a promising alternative, enabling the analysis of large, complex datasets to uncover patterns and interactions that traditional models cannot capture. By integrating a broad range of variables, such as age, sex, comorbidities, biomarker levels, and socioeconomic factors, machine learning models can provide more accurate and personalized risk assessments. These advanced tools have the potential to support clinicians in stratifying patient risk more effectively and tailoring treatment strategies to improve outcomes.

To assess the value of machine learning in this context, established models like SHFM and MAGGIC serve as benchmarks, providing a foundation for comparison. This study seeks to determine whether machine learning can outperform these traditional methods in predicting mortality and identifying the most significant predictors of patient outcomes.

The research question at the core of this study is: How can machine learning models effectively predict mortality in heart failure patients, and which clinical and demographic features play the most significant role in determining patient outcomes? By addressing this question, the study aims to advance the development of innovative, data-driven tools for improving heart failure management and patient care.

3 Data Processing

The dataset used in this study comes from the data set of heart failure clinical records, which contains medical records for 299 heart failure patients collected at the Faisalabad Institute of Cardiology and the Allied Hospital in Faisalabad, Punjab, Pakistan. It includes 13 clinical and demographic characteristics such as age, gender, serum sodium, ejection fraction, creatinine levels, and mortality status. The dataset was provided in '.csv' format and processed using Python, with all steps carefully documented.

To evaluate the performance of machine learning models, the dataset was split into training and test sets using an 80:20 ratio. The split was performed using the 'train_test_split' function from the 'sklearn' library with a fixed random state to ensure reproducibility. To preserve the integrity of the model evaluation, we ensured there was no overlap between the two sets. A detailed analysis of key variables, including age, ejection fraction, and comorbidities, confirmed that both sets had similar distributions, making them representative of real-world heart failure populations.

The dataset captures a range of patient characteristics. Age varies between 40 and 95 years, while clinical indicators such as serum creatinine and serum sodium show considerable variability, with creatinine levels ranging from 0.5 to 9.4 mg/dL and sodium levels between 114 and 148 mEq/L. The ejection fraction, a critical measure of heart function, ranges from 14% to 80%. Categorical variables such as anemia, diabetes, smoking, high blood pressure, and gender provide valuable information on comorbid conditions and demographic profiles. Patients were monitored over a follow-up period ranging from 4 to 285 days, with the target variable recording mortality as a binary outcome: 0 for survival and 1 for death. This combination of clinical and demographic data creates a comprehensive and diverse representation of heart failure patients, essential for building robust predictive models.

One of the key challenges with the dataset was the imbalance in the target variable. Only 32% of patients experienced a mortality event (class 1), while the remaining 68% survived (class 0). Without intervention, this imbalance could cause models to favor the majority class, leading to poor sensitivity in identifying mortality events. To address this, we applied the Synthetic Minority Oversampling Technique (SMOTE) to the training set, generating synthetic samples to balance the class distribution. Additionally, for models like logistic regression and tree-based classifiers, class weights were adjusted to penalize misclassification of the minority class. These approaches helped ensure that the models could detect mortality events more effectively.

To prepare the data for modeling, necessary transformations were applied. Continuous variables such as age, creatinine, and ejection fraction were standardized using the 'StandardScaler' to bring them to a uniform scale. Categorical features, including gender and smoking status, were one-hot encoded to make them suitable for machine learning algorithms. Importantly, these pre-processing steps were applied only to the training set and subsequently replicated on the test set to maintain independence between the two cohorts.

Furthermore, the representativeness of the dataset was assessed by analyzing cohort characteristics. The average patient age of approximately 60 years aligns with clinical expectations for heart failure populations. Key indicators like serum creatinine levels and ejection fraction also reflect real-world trends, lending further confidence to the dataset's reliability and generalizability. Hence, the data processing steps: splitting the dataset, addressing class imbalance, and applying transformations, were carefully designed to maintain the integrity and representativeness of the data. This rigorous pipeline laid the groundwork for developing machine learning models capable of accurately predicting mortality in heart failure patients.

4 Model Performance

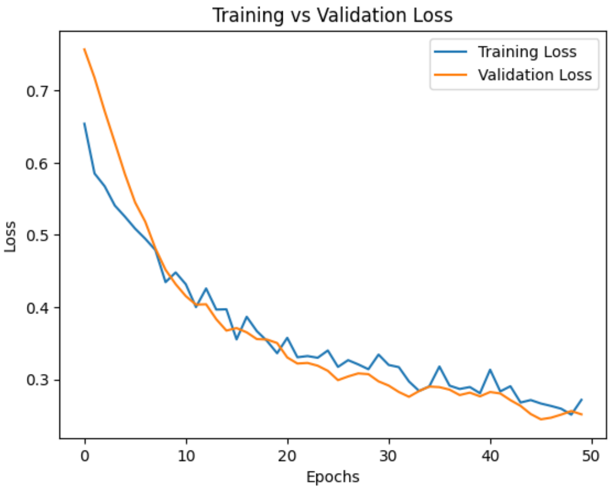
In this study, we evaluated three machine learning models: Logistic Regression, Random Forest Classifier, and a Neural Network, to predict mortality in heart failure patients. The models were assessed based on their performance and clinical utility using key metrics such as ROC-AUC and Positive Predictive Value (PPV). These metrics were carefully chosen: ROC-AUC measures the model's ability to discriminate between mortality and survival, while PPV highlights the reliability of positive predictions, which is particularly critical for identifying high-risk patients.

The Logistic Regression model was used as the baseline for comparison due to its simplicity and interpretability. The model was built using the 'LogisticRegression' class from the 'sklearn' library, with max_iter=1000 to ensure convergence. The training was performed on the SMOTE-resampled dataset, and predictions were generated on the test set. The model achieved an ROC-AUC score of 0.86, a PPV of 0.58, and an accuracy of 81.67%. While the model provided a solid starting point, its performance was limited, especially in handling the minority class. The convergence warning also indicated the need for further optimization or a more sophisticated model.

To improve predictive performance, we implemented a Random Forest Classifier using the 'RandomForestClassifier' from 'sklearn'. The ensemble method was chosen for its ability to handle non-linear relationships and feature interactions effectively. Hyperparameters such as the number of trees (n_estimators) and the random state were initialized for consistency. After training on the SMOTE-resampled training set, the model achieved a ROC-AUC score of 0.93, a PPV of 0.77, and an accuracy of 88.33% on the test set, demonstrating both algorithmic strength and clinical relevance.

Lastly, to capture more complex, non-linear relationships, we developed a Neural Network model using TensorFlow/Keras. The network architecture consisted of two hidden layers with

64 and 32 neurons, respectively, using ReLU activation functions, along with dropout layers to mitigate overfitting. A final sigmoid activation layer was used for binary classification. The input data, standardized using ‘StandardScaler’, was fed into the network and trained for 50 epochs with a batch size of 32. Early stopping and learning rate reduction techniques were applied to optimize the training process. The model achieved a ROC-AUC score of 0.87, a PPV of 0.56, and an accuracy of 80%, but its performance fell short of the Random Forest model, particularly in terms of precision. The training vs. validation loss plot (shown below) confirms that the model converged effectively, with both losses decreasing steadily over epochs, validating the learning process. It is likely that with a larger dataset, the Neural Network’s performance could improve further.



Overall, the Random Forest Classifier emerged as the best-performing model, achieving statistically significant improvements over the Logistic Regression and the Neural Network model. Its superior ROC-AUC and PPV scores, coupled with its ability to highlight key clinical predictors, make it the most reliable and clinically relevant model for this task.

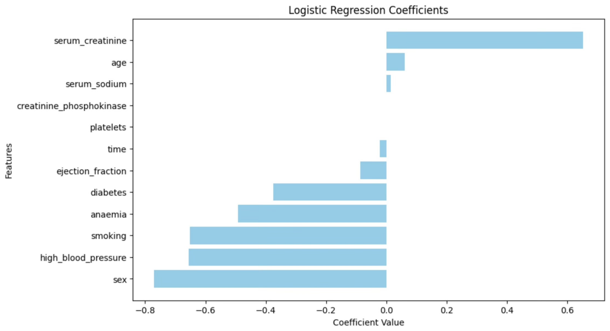
The complete code used for data processing and analysis is available in the following GitHub repository:[https://github.com/xx23438/STATGR5243-Project3].

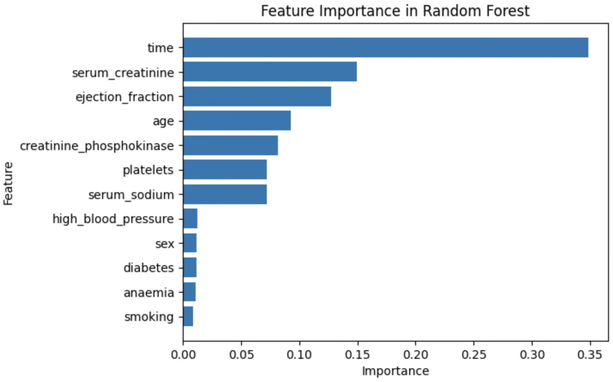
5 Discussion

The analysis of Logistic Regression coefficients and Random Forest feature importance underscores the critical predictors influencing mortality. In the Logistic Regression model, the coefficients reveal both the direction and magnitude of each feature’s impact on mortality. For instance, serum creatinine displays a strong positive coefficient, indicating that higher levels increase the likelihood of mortality, which aligns with its known clinical role as a marker of kidney dysfunction. Conversely, features like sex and high blood pressure have notable negative coefficients, suggesting they are associated with a lower risk of mortality.

In comparison, the Random Forest feature importance analysis identifies time, serum creatinine, and ejection fraction as the most influential predictors. These findings are consistent with established clinical markers of disease progression and severity in heart failure, where longer follow-up periods (time), worsening kidney function (serum creatinine), and reduced heart function (ejection fraction) are key indicators of adverse outcomes.

The close alignment between these results and clinical expectations not only enhances the credibility of the models but also validates their ability to capture meaningful relationships within the data. This reinforces confidence in the models' performance and their potential applicability in real-world clinical decision-making.





Model interpretability is essential, particularly in clinical decision-making, where transparency and trust are paramount. Logistic Regression, with its clear and interpretable coefficients, allows clinicians to easily understand the relationship between each predictor and the outcome. This transparency makes it straightforward to explain the model's decisions at the individual case level. On the other hand, while Random Forest does not provide direct interpretability, feature importance analysis offers valuable insights into which variables contribute most to the predictions. For cases requiring even deeper interpretability, tools such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) can be used to explain individual predictions, bridging the gap between complex models and practical clinical use. This level of interpretability is vital for fostering clinician confidence and ensuring these models are adopted in real-world settings.

The reliability and robustness of the models must also be considered, particularly when the underlying data distribution shifts. Random Forest models are naturally resilient to nonlinear relationships and slight variations in data but may still struggle with unseen clinical scenarios if they deviate significantly from the training distribution. Logistic Regression, while interpretable, assumes linear relationships between features and the outcome, making it less flexible when dealing with more complex or evolving patterns in patient data. To maintain robustness, it is crucial to retrain the models on updated datasets regularly and validate their performance over time. In addition, techniques such as cross-validation and continuous performance monitoring can help detect performance degradation and ensure model reliability as patient demographics or clinical conditions evolve. By combining interpretability with robust validation practices, these models can provide reliable and actionable insights for improving patient care in dynamic clinical environments.

6 Conclusion

This project applied machine learning models to predict mortality in heart failure patients using structured clinical and demographic data. The data processing pipeline addressed class imbalance with SMOTE and ensured robust model evaluation through careful data splitting and feature transformation. Logistic Regression, Random Forest, and Neural Network models were developed and evaluated using ROC-AUC and Positive Predictive Value (PPV) as key performance metrics. The Random Forest model emerged as the best performer, with time, serum creatinine, and ejection fraction identified as the most influential predictors.

The findings highlight that machine learning models, particularly Random Forest, can deliver both strong predictive performance and interpretability, making them suitable for clinical decision-making. While Logistic Regression offered transparency, Random Forest's ability to capture complex relationships proved superior. Future research will focus on validating these models on larger datasets and enhancing their robustness to evolving clinical scenarios, ensuring their reliability and utility in improving patient outcomes.