Hilary Henshaw
22 February 2025
DSP 562 Final Project Proposal

Predictive Analysis of Mortality in Heart Failure Patients Using Data Analytics and Visualization Tools

Abstract

This study leverages advanced data analytics and visualization tools to analyze the Heart Failure Clinical Records dataset. By exploring the relationships between clinical features and patient outcomes through platforms such as Jupyter, Tableau, and Bokeh, this research aims to pinpoint key predictors of mortality in heart failure patients. The findings are expected to enhance risk stratification and support the development of personalized treatment plans, ultimately improving patient outcomes.

Introduction

Heart failure is a prevalent condition with high morbidity and mortality rates, posing significant public health challenges. In the United States, approximately 6.7 million adults aged 20 and older are affected by heart failure. In 2022, heart failure was recorded on 457,212 death certificates, accounting for 13.9% of all deaths. The economic burden of heart failure was estimated to be $30.7 billion in 2012, encompassing healthcare services, medications, and lost productivity due to missed workdays. People suffering from heart failure can experience severe symptoms. In some cases, they might require a heart transplant or a mechanical device to assist the heart in pumping blood effectively. Understanding the factors contributing to patient outcomes is crucial for developing effective treatments and improving survival rates. This study focuses on identifying the key clinical features that predict mortality in heart failure patients using various data analytics and visualization tools. The motivation for this research stems from the need for better risk stratification tools and personalized treatment plans to enhance patient care.

Heart failure, also known as congestive heart failure, is a complex syndrome characterized by the heart's inability to pump blood effectively, leading to reduced oxygen supply to tissues and organs. Diagnosis of heart failure relies on symptoms, physical examination findings, and echocardiography. Blood tests and chest X-rays can help identify the underlying cause. Treatment varies based on the severity and specific circumstances of each case. Heart failure symptoms can differ depending on whether the left or right side of the heart is affected. Common symptoms include shortness of breath, excessive fatigue, and swelling in both legs. The severity of heart failure is primarily determined by the ejection fraction and the intensity of the symptoms.

Certain medical conditions can elevate your risk of heart failure, such as coronary artery disease, heart attacks, diabetes, high blood pressure, obesity, and valvular heart disease. Unhealthy behaviors, including smoking, poor diet, lack of physical activity, and excessive alcohol consumption, can also increase the risk, especially for those with existing conditions. Numerous studies have explored the factors influencing heart failure outcomes, including age, comorbidities, and biochemical markers. Previous research has highlighted the importance of variables such as ejection fraction, serum creatinine, and sodium levels in predicting patient mortality. However, there is still a need for comprehensive analyses that incorporate multiple clinical features to improve prediction accuracy.

In this collaborative project, we strive to significantly reduce mortality rates among heart failure patients by providing actionable insights and scalable solutions. Our study's findings hold the potential to shape both clinical practices and policy decisions, particularly in settings where resources are limited, and heart failure-related deaths are prevalent. By harnessing the power of data analytics and visualization, we showcase how data-driven methodologies can be applied to classify and predict heart failure risks effectively. While our analysis focuses on the heart failure dataset, it lays a generalizable foundation for creating predictive tools that can aid clinical decision-making and enhance patient outcomes. This paper will delve into the dataset, methodology, and key findings of our research, illustrating the profound impact of integrating data analytics with real-world health data.

Exploratory Data Analysis

The Heart Failure Clinical Records dataset, sourced from the UCI Machine Learning Repository, consists of 299 observations with 13 attributes: Age, Anemia, Creatinine Phosphokinase, Diabetes, Ejection Fraction, High Blood Pressure, Platelets, Serum Creatinine, Serum Sodium, Sex, Smoking, Time, and Death event. These variables offer a thorough snapshot of heart failure patients, encompassing critical physiological and demographic elements that affect heart failure outcomes.

The data was collected during the follow-up period of patients at the Faisalabad Institute of Cardiology and Allied Hospital in Faisalabad (Punjab, Pakistan) from April to December 2015. All patients had left ventricular systolic dysfunction and a history of heart failure, placing them in classes III or IV according to the New York Heart Association (NYHA) classification. The death event feature, used as the target in the binary classification study, indicates whether the patient died or survived by the end of the follow-up period, which averaged 130 days.

The age of patients in the dataset ranges from 40 to 95 years, with an average age of 60.83 years and a standard deviation of 11.89 years. The interquartile range (IQR) indicates that 50% of the participants are between 51 and 70 years old, with a median age of 60 years. Ejection fraction values range from 14% to 80%, with a mean of 38.08% and a standard deviation of 11.83%. The median ejection fraction is 38%, and most values fall between 30% and 45%, indicating varying degrees of heart function among the participants.

Creatinine phosphokinase levels range widely, from 23 to 7861 mcg/L, with a mean of 581.84 mcg/L and a standard deviation of 970.29 mcg/L. The median level is 250 mcg/L, and half of the values lie between 116 and 582 mcg/L. Serum creatinine levels vary from 0.50 to 9.40 mg/dL, with a mean of 1.39 mg/dL and a standard deviation of 1.03 mg/dL. The median serum creatinine level is 1.10 mg/dL, with the IQR showing that 50% of observations lie between 0.90 and 1.50 mg/dL.

Serum sodium levels range from 113 to 148 mmol/L, with a mean of 136.63 mmol/L and a standard deviation of 4.41 mmol/L. The median serum sodium level is 137 mmol/L, and most values fall between 134 and 140 mmol/L. The dataset also includes binary variables indicating anemia (0: No, 1: Yes), diabetes (0: No, 1: Yes), high blood pressure (0: No, 1: Yes), sex (0: Female, 1: Male), and smoking status (0: No, 1: Yes).

[1] See appendix

[2] UCI Machine Learning Repository, "Heart Failure Clinical Records" accessed February 14, 2025, https://archive.ics.uci.edu/dataset/519/heart+failure+clinical+records.

[3] See appendix

The target variable, Death event, categorizes patient outcomes into two classes: 0 (No Event) and 1 (Event). During the exploratory data analysis phase, the distribution of death events was thoroughly examined to confirm the dataset's suitability for classification tasks. These descriptive statistics offer valuable insights into the dataset's composition and variability, pinpointing critical predictors like age, ejection fraction, creatinine levels, and blood pressure. These variables are crucial for effectively modeling heart failure outcomes.

Methodology

The Heart Failure Clinical Records dataset was meticulously examined for missing values using isnull(). Remarkably, the dataset contained no missing values, allowing us to retain all 299 entries for a comprehensive analysis. The target variable, Death event, was already in a binary format, making it suitable for our machine learning classification tasks.

The exploratory data analysis (EDA) phase involved a detailed examination of the dataset's distribution of death events to confirm its suitability for classification tasks. Descriptive statistics such as mean, median, standard deviation, and ranges were calculated for all variables to gain an understanding of the data distribution.

Various visualization techniques were employed to explore relationships and patterns within the data. Histograms, box plots, scatter plots, bar charts, and heatmaps were created using libraries such as Matplotlib, Seaborn, and Tableau. These visualizations were instrumental in identifying critical predictors such as age, ejection fraction, creatinine levels, and blood pressure, which are vital for effectively modeling heart failure outcomes. Furthermore, correlation analysis was conducted to identify relationships between different variables, providing insights into potential predictors of the target variable.

Random Forest Classifier was selected for predicting heart failure outcomes due to its ability to handle non-linear relationships, improve classification accuracy, and reduce overfitting by building multiple trees and averaging their predictions. The model was trained on the standardized training dataset and evaluated on the test dataset. To ensure robust model performance and generalizability to unseen data, five-fold stratified cross-validation was employed. This approach ensured proportional representation of the Death event classes in both training and validation subsets during cross-validation. The model's performance was evaluated using metrics such as accuracy, F1-score, and AUC (Area Under the ROC Curve).

The final model was interpreted using feature importance analysis to understand the contribution of each feature. The results were visualized using various plots, including confusion matrices, ROC curves, and feature importance plots, to effectively communicate the findings. These visualizations provided valuable insights into the dataset's structure and variability, highlighting critical predictors for heart failure outcomes.

By applying these data analytics and visualization techniques, the study aimed to explore and interpret the heart failure dataset effectively, assess feature importance, and develop actionable insights for risk assessment in heart failure patients. These insights have potential applications in clinical and real-world settings, contributing to improved patient outcomes and better understanding of heart failure prognosis.

Main Findings

This analysis utilized various visualizations, including histograms, scatter plots, and box plots, to explore the relationships between key variables in heart failure patients. The visualizations focused on ejection fraction, serum creatinine, and death events to uncover significant patterns and trends. It predicts heart failure outcomes using a supervised learning model, focusing on two classes: no death event (0) and death event (1).

The primary model evaluated was a classifier whose performance was measured by accuracy, confusion matrix, classification report, and ROC AUC score. The classifier achieved an overall accuracy of 82.2%. The confusion matrix showed that the model correctly classified 58 instances of no death event (0) and 16 instances of death event (1). However, it incorrectly classified 13 instances of death event (1) as no death event (0) and 3 instances of no death event (0) as death event (1).

The ROC AUC score of 0.8836 indicates a good discrimination ability of the model between the two classes. The classifier demonstrated strong performance in predicting no death event (0) cases with high precision and recall. However, it faced challenges in predicting death event (1) cases, with a lower recall of 0.55, indicating some difficulty in correctly identifying all death event instances. The strong ROC AUC score suggests that the model can effectively distinguish between the two classes.

Overall, this model shows promise in identifying heart failure outcomes, particularly in predicting no death event cases. Effective early identification of death event cases can lead to timely interventions, potentially reducing adverse outcomes and improving patient care.

Key Trends and Patterns

Ejection Fraction: The boxplot showed that the majority of patients had an ejection fraction between 40% and 60%. There was a noticeable decrease in the number of patients with ejection fractions below 40%, indicating that lower ejection fractions are less common.

Ejection Fraction by Death Event: The box plot indicated that patients who experienced a death event had significantly lower ejection fractions compared to those who survived. The median ejection fraction for deceased patients was around 35%, whereas for surviving patients, it was around 50%.

Serum Creatinine vs. Ejection Fraction: The scatter plot revealed no significant correlation between serum creatinine levels and ejection fraction. The data points did not show any clear pattern or relationship between these two variables.

Serum Creatinine: The boxplot showed that the majority of patients with high Serum Creatinine have a high death event.

Patients with lower ejection fractions and higher serum creatinine levels were more likely to experience death events. This suggests that these two variables are critical indicators of heart failure severity.

Discussion/Conclusion

This study aimed to analyze the factors contributing to mortality in heart failure patients using data analytics and visualization tools. Through the exploration of the Heart Failure Clinical Records dataset, we identified key clinical features that significantly influence patient outcomes, such as ejection fraction, serum creatinine levels, and age. The combination of these features was then used to predict the likelihood of mortality using machine learning techniques.

One of the most notable findings was the strong association between lower ejection fractions and increased mortality risk. The boxplot analysis revealed that deceased patients had significantly lower ejection fractions compared to those who survived. This aligns with existing medical literature that identifies reduced ejection fraction as a critical indicator of heart failure severity. Ejection fraction, a measure of the heart's pumping ability, is essential in assessing the prognosis of heart failure patients, with lower values typically indicating worse outcomes.

Serum creatinine levels also emerged as a crucial predictor of mortality, with patients exhibiting higher creatinine levels having an increased likelihood of death events. Elevated serum creatinine often signals kidney dysfunction, which is a common complication of heart failure and has been linked to worse patient outcomes. This underscores the interconnectedness of organ systems, where kidney impairment can exacerbate the progression of heart failure, ultimately affecting survival rates.

The Random Forest Classifier used in the study demonstrated high performance, achieving an accuracy of 82.2%, which is promising for clinical decision-making. The ROC AUC score of 0.8836 indicates that the model is effective at distinguishing between patients who will survive and those who will experience a fatal event. However, while the model was successful in identifying patients who survived (high precision and recall for "no death event"), it showed some challenges in predicting "death event" cases, as indicated by the lower recall for this class. This suggests that while the model is strong overall, it could benefit from further tuning, particularly for improving sensitivity in identifying high-risk patients who are more likely to experience a death event.

The visualizations provided critical insights into the distribution and relationships of various clinical variables. Histograms, box plots, and scatter plots helped identify trends, such as the higher frequency of death events among patients with lower ejection fractions and higher serum creatinine levels. These visual tools proved invaluable in guiding the analysis and confirming the relevance of certain features, such as ejection fraction and serum creatinine, in predicting mortality risk.

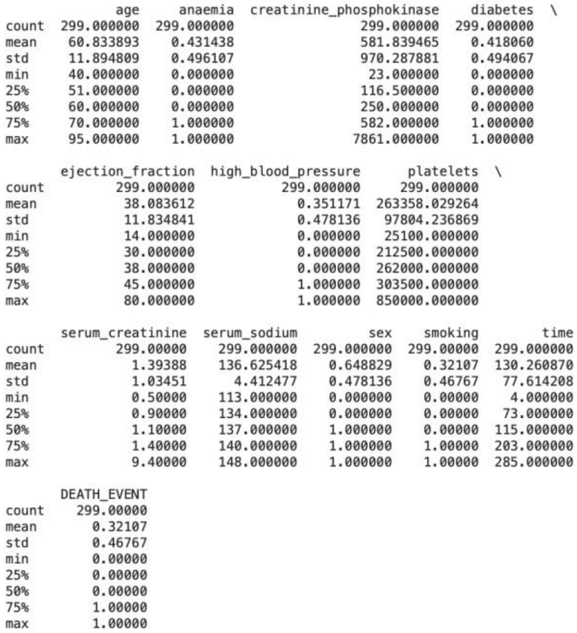
The findings from this analysis have important implications for clinical practice. By identifying the key predictors of mortality in heart failure patients, healthcare providers can better stratify patient risk, enabling earlier interventions and more personalized treatment plans. For instance, patients with low ejection fractions and elevated serum creatinine levels could be closely monitored and provided with targeted therapies to improve their chances of survival.

[3] See appendix
[4] See appendix
[5] See appendix
[6] See appendix

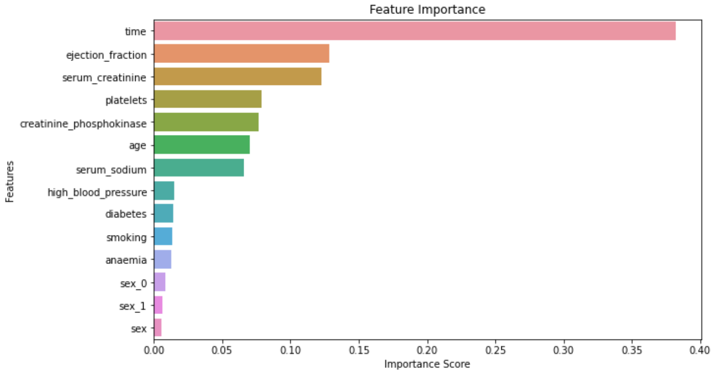
Finally, this research demonstrates the potential of leveraging data analytics and machine learning to improve the prediction of mortality outcomes in heart failure patients. The combination of robust data visualization and predictive modeling offers a powerful tool for clinicians to make more informed decisions, ultimately contributing to better patient outcomes. Future work could involve exploring other machine learning algorithms and expanding the dataset to further enhance prediction accuracy and generalizability. Additionally, integrating real-time clinical data into these predictive models could make them even more actionable in clinical settings, ultimately transforming the approach to heart failure management.

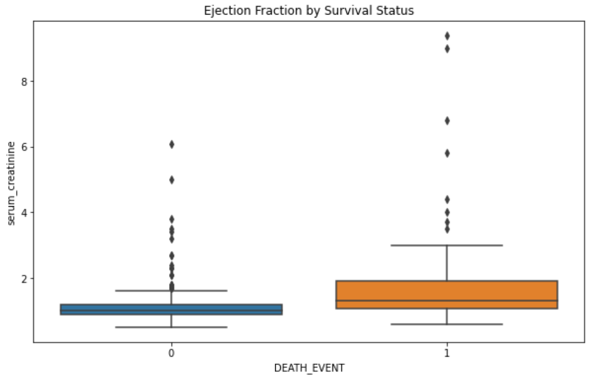
Appendix

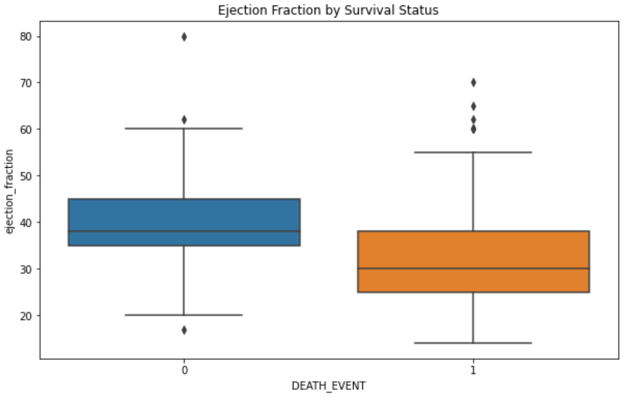
Descriptive Statistics



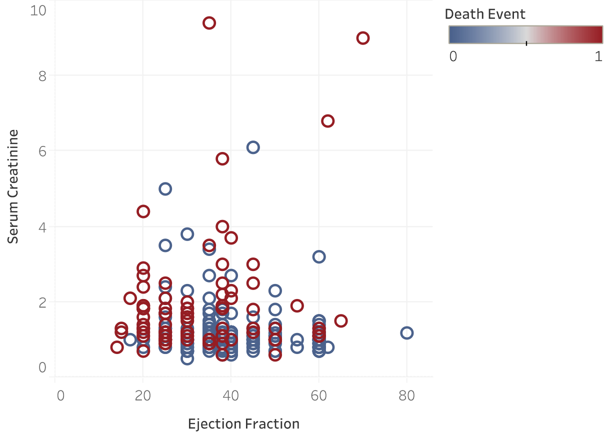
```
             age     anaemia  creatinine_phosphokinase    diabetes  \
count  299.000000  299.000000                299.000000  299.000000
mean    60.833893    0.431438                581.839465    0.418060
std     11.894809    0.496107                970.287881    0.494067
min     40.000000    0.000000                 23.000000    0.000000
25%     51.000000    0.000000                116.500000    0.000000
50%     60.000000    0.000000                250.000000    0.000000
75%     70.000000    1.000000                582.000000    1.000000
max     95.000000    1.000000               7861.000000    1.000000

       ejection_fraction  high_blood_pressure      platelets  \
count         299.000000           299.000000     299.000000
mean           38.083612             0.351171  263358.029264
std            11.834841             0.478136   97804.236869
min            14.000000             0.000000   25100.000000
25%            30.000000             0.000000  212500.000000
50%            38.000000             0.000000  262000.000000
75%            45.000000             1.000000  303500.000000
max            80.000000             1.000000  850000.000000

       serum_creatinine  serum_sodium         sex    smoking        time  \
count         299.00000    299.000000  299.000000  299.00000  299.000000
mean            1.39388    136.625418    0.648829    0.32107  130.260870
std             1.03451      4.412477    0.478136    0.46767   77.614208
min             0.50000    113.000000    0.000000    0.00000    4.000000
25%             0.90000    134.000000    0.000000    0.00000   73.000000
50%             1.10000    137.000000    1.000000    0.00000  115.000000
75%             1.40000    140.000000    1.000000    1.00000  203.000000
max             9.40000    148.000000    1.000000    1.00000  285.000000

       DEATH_EVENT
count    299.00000
mean       0.32107
std        0.46767
min        0.00000
25%        0.00000
50%        0.00000
75%        1.00000
max        1.00000
```





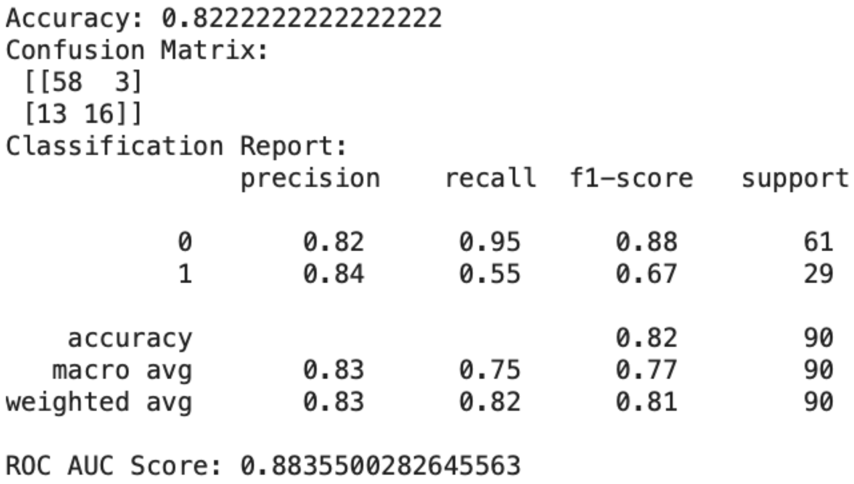


Heart Failure Analysis Over Time:
Ejection Fraction and Serum Creatinine



Ejection Fraction vs. Serum Creatinine. Color shows Death Event.

Random Forest Classifier



```
Accuracy: 0.8222222222222222
Confusion Matrix:
 [[58  3]
 [13 16]]
Classification Report:
               precision    recall  f1-score   support

           0       0.82      0.95      0.88        61
           1       0.84      0.55      0.67        29

    accuracy                           0.82        90
   macro avg       0.83      0.75      0.77        90
weighted avg       0.83      0.82      0.81        90

ROC AUC Score: 0.8835500282645563
```