

### TABLE OF CONTENTS

[ABSTRACT 3](#_Toc168603696)

[INTRODUCTION 3](#_Toc168603697)

[GOAL 4](#_Toc168603698)

[METHODS 5](#_Toc168603699)

[RESULTS 8](#_Toc168603700)

[APPENDIX 15](#_Toc168603701)

### ABSTRACT

This project focuses on the comprehensive analysis and modeling of heart failure clinical data. The primary objectives were threefold: predicting ejection fraction using regression techniques, classifying the occurrence of death events, and profiling patient clusters. Various machine learning algorithms were employed including, Random Forest, Ridge and Lasso Regression, Decision Tree, Multivariable Regression (OLS), k-NN Classification, Naïve Bayes Classification, Random Forest Classification, Neural Networks for Regression and Classification, K-Means and Hierarchical Clustering. The models were trained and validated on carefully preprocessed data, ensuring robust and accurate predictions and classifications. The outcomes provide valuable insights into the factors influencing heart failure prognosis and patient profiles.

### INTRODUCTION

Heart failure is a chronic condition characterized by the heart’s inability to pump sufficient blood to meet the body’s needs. It is a significant public health concern, leading to high morbidity and mortality rates. Early prediction and accurate diagnosis are crucial for improving patient outcomes and managing the disease effectively. In recent years, machine learning techniques have shown great promise in medical research, particularly for predictive modeling and patient classification.

This project leverages a variety of machine learning algorithms to analyze a dataset of heart failure clinical records. The dataset includes features such as age, serum sodium levels, creatinine phosphokinase levels, and more, which are used to predict critical outcomes like ejection fraction and death events. Additionally, clustering techniques are applied to profile patients into distinct groups based on their clinical characteristics.

### GOAL

The primary objectives of this project are:

1. Identify Key Predictors of Heart Failure: Utilize machine learning methods to determine the most significant factors affecting ejection fraction and death events in patients.
2. Evaluate the Performance of Different Machine Learning Models: Implement and compare the effectiveness of various regression, classification and clustering algorithms in predicting heart failure outcomes and profiling patients.

By achieving these goals, this project aims to contribute to the understanding and management of heart failure, potential leading to improved patient care and outcomes.

### METHODS

This section details the various steps and techniques employed to predict ejection fraction, classify death events, and profile patient clusters.

***Data Preprocessing***

The dataset was preprocessed to ensure the integrity and consistency of the data. The preprocessing steps included:

1. Handle Missing values: The dataset was checked for missing values, and none were found.
2. Encoding Categorical Variables: The categorical variables ‘diabetes’, ‘high\_blood\_pressure’, ‘smoking’, ‘sex’, ‘anaemia’, and ‘DEATH\_EVENT’ were converted to categorical data types for proper handling during analysis.
3. Renaming Columns: For better readability in graphs and visualizations, the column ‘creatinine\_phosphokinase’ was renamed to ‘creatinine\_pk’.
4. Removing Clinically Improbable Values: Entries with improbably values (likely incorrect data entries) were removed to clean the dataset.
5. Correlation Analysis: The numerical features were examined for correlations.

***Feature Selection***

Random Forest method was employed to identify the most important features for both regression and classification tasks. The importance of each feature was calculated and ranked, allowing for the selection of the most relevant predictors.

***Regression Analysis***

Multiple regression methods were applied to predict the ejection fraction. The selected features for regression included ‘time’, ‘age’, ‘creatinine\_pk’, ‘platelets’, ‘serum\_sodium’, ‘DEATH\_EVENT’, and ‘serum\_creatinine’. The regression methods used were:

1. Random Forest Regression: To capture non-linear relationships in the data.
2. Ridge Regression: A regularized linear regression method to handle multicollinearity.
3. Lasso Regression: Another regularization technique that can shrink coefficients of less important features to zero.
4. Multivariable Regression (OLS): Standard ordinary least squares regression to understand the linear relationships between features and the target variable.
5. Decision Tree Regression: A non-linear regression technique that splits the data into subsets based on feature values.
6. Neural Network Regression: A deep learning approach to model complex relationships in the data.

***Classification Analysis***

Various classification techniques were used to predict death events. The selected features for classification included ‘time’, ‘creatinine\_pk’, ‘platelets’, ‘age’, ‘serum\_creatinine’, ‘serum\_sodium’, and ‘ejection fraction’. The classification methods employed were:

1. k-Nearest Neighbors (k-NN) Classification: An instance-based learning algorithm that classifies a sample based on the majority class of its k-nearest neighbors.
2. Naïve Bayes Classification: A probabilistic classifier based on Bayes’ theorem with strong independence assumptions.
3. Random Forest Classification: An ensemble learning method that constructs multiple decision trees and merges them together to obtain a more accurate and stable prediction.
4. Neural Networks Classification: A deep learning method that uses multiple layers of neurons to learn complex patterns in the data.

***Clustering Analysis***

Clustering techniques were applied to profile patients based on their clinical features. The methods used were:

1. K-Means Clustering: A method that partitions the data into k distinct clusters based on feature similarities.
2. Hierarchical Clustering: An agglomerative clustering method that builds a hierarchy of clusters by progressively merging or splitting them.

### RESULTS

***Regression Analysis***

The goal of the regression analysis was to predict the ejection fraction of heart failure patients using various regression models. The performance of each model was evaluated using Mean Squared Error (MSE) and R² Score on both the validation and test sets.

1. **Random Forest Regression**

Validation set - Mean Squared Error: 183.44 & R² Score: -0.02

Test Set - Mean Squared Error: 122.30 & R² Score: -0.04

1. **Ridge Regression**

Validation set - Mean Squared Error: 183.12 & R² Score: -0.02

Test Set - Mean Squared Error: 105.78 & R² Score: 0.1

1. **Lasso Regression**

Validation set - Mean Squared Error: 187.90 & R² Score: -0.05

Test Set - Mean Squared Error: 110.48 & R² Score: 0.06

1. **Multivariable Regression (OLS)**

Validation set - Mean Squared Error: 187.48 & R² Score: -0.04

Test Set - Mean Squared Error: 100.23 & R² Score: -0.04

1. **Decision Tree Regression**

Validation set - Mean Squared Error: 309.30 & R² Score: -0.72

Test Set - Mean Squared Error: 198.57 & R² Score: -0.69

1. **Neural Network Regression**

Validation set - Mean Squared Error: 213.49 & R² Score: -0.19

Test Set - Mean Squared Error: 150.78 & R² Score: -0.28

The regression models generally performed poorly in predicting ejection fraction, as indicated by the high MSE values and negative R² scores for most models. The Ridge Regression model showed a slight improvement with a positive R² score on the test set, but overall, the models did not effectively capture the variability in ejection fraction. The high MSE values suggest that the models’ predictions were not close to the actual values, indicating limited practical applicability for this specific regression task.

***Classification Analysis***

The classification analysis aimed to predict the occurrence of death events using various classification models. The performance of each model was evaluated using F1 Score, precision, recall, and accuracy on the test set (Class 0 = No Death Event, Class 1 = Yes Death Event).

1. **k-Nearest Neighbors Classification**

Test set:

F1 Score: 0.57, Accuracy: 0.77, Precision 0.76 (class 0) & 0.81 (class 1),

Recall: 0.95 (class 0) & 0.43 (class 1),

Chi-square Statistic: 16.53, p-value = 0.00

1. **Naïve Bayes Classification**

Test set:

F1 Score: 0.78, Accuracy: 0.85, Precision 0.88 (class 0) & 0.79 (class 1),

Recall: 0.89 (class 0) & 0.77 (class 1),

Chi-square Statistic 35.77, p-value: 0.00

1. **Random Forest Classification**

Test set:

F1 Score: 0.81, Accuracy: 0.87, Precision 0.88 (class 0) & 0.77 (class 1),

Recall: 0.93 (class 0) & 0.77 (class 1),

Chi-square Statistic 41.35, p-value: 0.00

1. **Random Forest Classification**

Test set:

F1 Score: 0.71, Accuracy: 0.80, Precision 0.84 (class 0) & 0.72 (class 1),

Recall: 0.86 (class 0) & 0.70 (class 1),

Chi-square Statistic 29.76, p-value: 0.00

The classification models showed varying degrees of performance, with the Random Forest Model achieving the highest performance with an accuracy of 0.87 and a F1 score of 0.81, indicating it was the most effective model for predicting death events. The Naïve Bayes and Neural Networks models also performed well with accuracies of 0.85 and F1 scores of 0.78, showing balanced precision and recall. The k-NN model, while having decent accuracy of 0.77, struggled with recall for class 1, resulting in a lower F1 score of 0.57. Therefore, Random Forest is recommended for its superior performance, with Naïve Bayes and Neural Networks as strong alternatives.

***Clustering Analysis***

The clustering analysis aimed to group heart failure patients into distinct clusters based on clinical features, using both K-Means and Hierarchical Clustering methods.

1. **K-Means Clustering**

K-Means clustering was performed with 4 clusters, using features such as time, creatinine phosphokinase, platelets, age, serum creatinine, serum sodium, and ejection fraction.

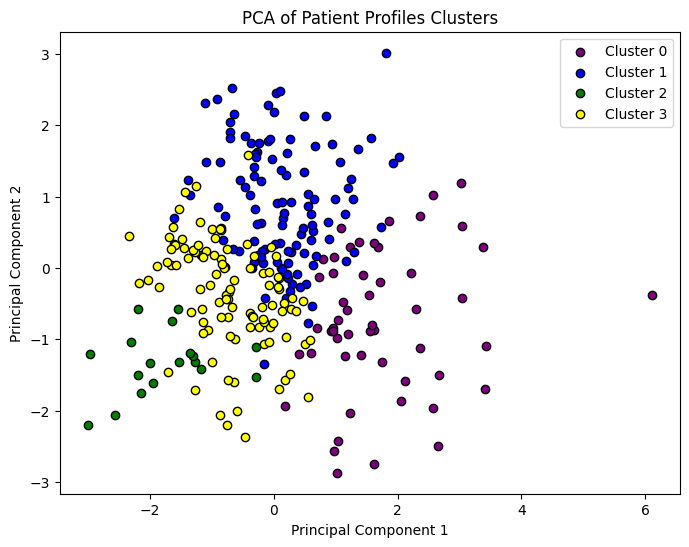
Cluster Descriptions:

 Cluster 0 (Purple): Older patients with shorter survival times, lower platelet counts, higher serum creatinine, significantly lower serum sodium, and lower ejection fraction. Count: 53 patients.

 Cluster 1 (Blue): Patients with shorter survival times, average platelet counts, slightly higher age, lower serum creatinine, higher serum sodium, and ejection fraction. Count: 112 patients.

 Cluster 2 (Green): Patients with longer survival times and significantly higher creatinine phosphokinase levels, lower platelet counts, slightly younger age, near-average ejection fraction and serum creatinine levels. Count: 21 patients.

 Cluster 3 (Yellow): Patients with the longest survival times, average creatinine phosphokinase levels, slightly lower platelet counts and ejection fraction, younger age, near-average serum creatinine and sodium levels. Count: 101 patients.



1. **Hierarchical Clustering**

Hierarchical clustering was also performed with 4 clusters using the Ward method.

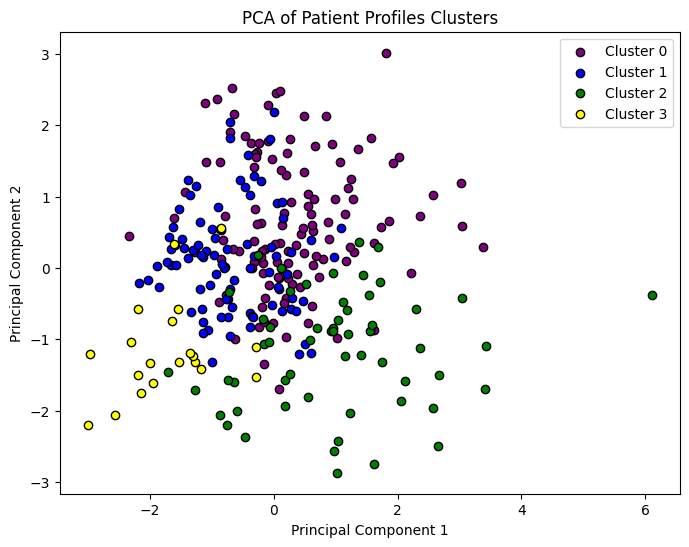
The cluster descriptions remain the same as the clusters were formed using similar characteristics as in K-Means Clustering.

Cluster 0 (Purple): 121 patients.

Cluster 1 (Blue): 89 patients.

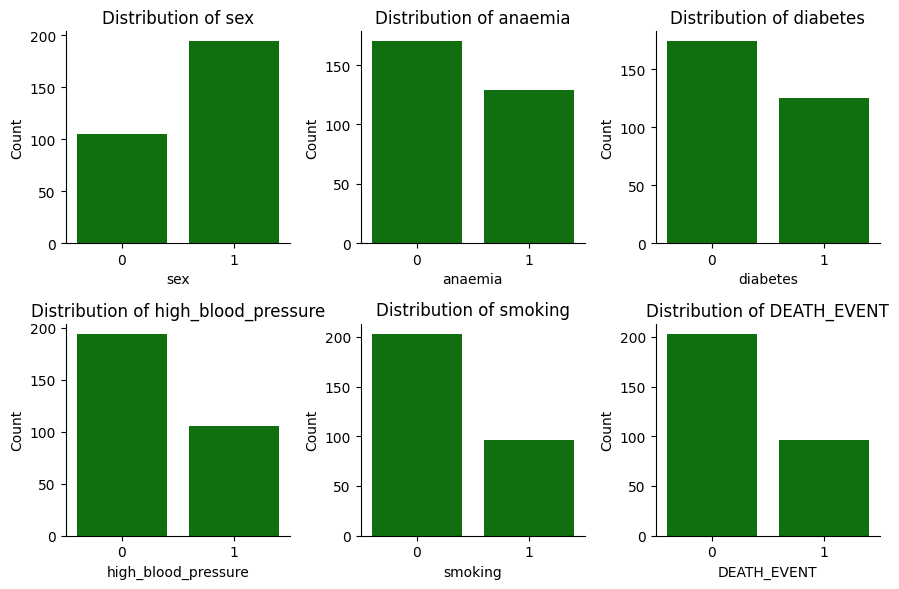
Cluster 2 (Green): 57 patients.

Cluster 3 (Yellow): 20 patients.



The K-Means clustering method identified four distinct patient profiles, primarily differentiated by survival times, platelet counts, serum creatinine levels, serum sodium levels and ejection fraction. Hierarchical clustering also identified four distinct groups, showing a consistent pattern with the K-Means results in terms of survival times and clinical feature distribution. The visualization techniques, including pair plots and PCA plots, provided clear insights into the distinct separation and characteristics of the clusters.

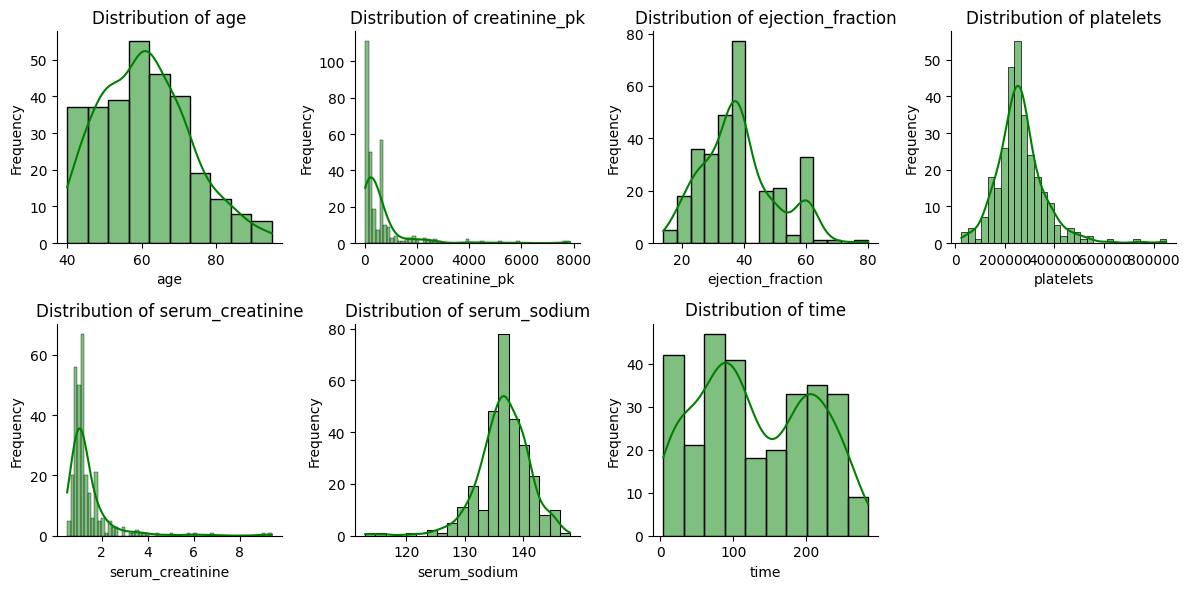
### APPENDIX

***The categorical variables’ distribution*** shows counts and proportions for each category.

Number of patients who survived:

Yes: 203, No: 96

***Histograms for numerical variables*** display their range, central tendency, and spread.



Key statistics of ejection fraction:

Count: 299, Mean: 38.08, Std: 11.83,

Min: 14.00, 25%: 30.00, 50%: 38.00, 75%: 45.00, Max: 80.00

The correlation matrix heatmap shows a moderate negative correlation between age and time (-0.23), as well as serum sodium and serum creatinine (-0.23), and a smaller but positive correlation between age and serum creatinine (0.17).

