Heart Failure

Analyzed By Ahmed Elsayed

Source: https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Discussion
- Findings & Implications
- Conclusion
- Appendix

Executive Summary

This project focuses on the analysis and prediction of heart failure outcomes using clinical data from 299 patients. The objective was to identify key factors contributing to mortality and develop predictive models to support clinical decision-making. Through rigorous data preprocessing, statistical testing, and machine learning techniques, the project successfully revealed significant predictors of heart failure mortality and stratified patients into distinct risk groups.

The analysis applied **logistic regression** to predict the likelihood of mortality based on critical features, including **serum creatinine, serum sodium, ejection fraction, age, and follow-up time**. Additionally, cluster analysis was performed to segment patients into risk groups, providing insights into common characteristics shared by high-risk individuals.

Key findings include:

- **Serum creatinine, serum sodium, ejection fraction** were statistically significant in predicting adverse outcomes which are death events ($p < 0.05$). The model was better to predicting patients that will survive better than those will die.
- Patients were grouped into **two distinct clusters**, highlighting differences in patient profiles and mortality risks. With similar accuracy of 72%, the model correctly identified patients that survived from those who died. Cohen's Kappa is 0.44 indicating fair agreement between the clusters and the true labels.
- The logistic regression model achieved an accuracy of **82%**, demonstrating strong predictive performance.

This project offers valuable insights into heart failure risk factors and contributes to enhancing predictive analytics in healthcare.

Introduction

Heart failure is a leading cause of morbidity and mortality worldwide, necessitating effective risk stratification and predictive modeling to improve patient outcomes. This project aims to analyze clinical data from heart failure patients, leveraging statistical methods and machine learning to identify significant predictors of mortality. By applying logistic regression and cluster analysis, the project seeks to uncover patterns in patient data that correlate with adverse outcomes.

The dataset, comprising **299 patients** with various clinical features, includes critical indicators such as **serum creatinine, ejection fraction, serum sodium, and age**. These variables are known to influence heart failure prognosis, and their analysis can provide actionable insights for clinicians.

The objectives of this project are as follows:

1.Perform exploratory data analysis (EDA) to understand the distribution of key variables.

2.Conduct hypothesis testing using independent t-tests to identify significant differences between survivors and non-survivors.

3.Develop a logistic regression model to predict mortality based on selected features.

4.Implement cluster analysis to group patients into risk categories, enabling better patient management.

Through this multifaceted approach, the project aims to enhance clinical decision-making by offering a data-driven perspective on heart failure risk factors.

Methodology

1. Data Collection and Preprocessing

The dataset used for this analysis consists of **299 patient records** from the heart failure clinical records dataset. Each record includes demographic, laboratory, and clinical measurements such as:

- **Age**
- **Serum Creatinine**
- **Serum Sodium**
- **Ejection Fraction**
- **Follow-up Time**
- **DEATH_EVENT** (Target variable indicating mortality)

Data Preprocessing Steps:

1. **Handling Missing Data** – The dataset had no missing values, ensuring all records were available for analysis.
2. **Feature Transformation** – Several variables, including **serum creatinine** and **creatinine phosphokinase**, exhibited skewed distributions.
    - **Logarithmic transformations** were applied to reduce skewness and improve normality.

Methodology

2. Exploratory Data Analysis (EDA)

EDA was conducted to understand the distribution and relationships between features. Key steps included:

- **Visualizations** – Histograms, boxplots, and correlation heatmaps were generated to explore feature distributions.
- **Statistical Testing** – Welch's t-test was performed to compare means between survivors and non-survivors for key clinical variables.

3. Statistical Testing (T-test)

Based on the correlation matrix, the most correlated 5 variables were identified, and then, Independent t-tests were conducted to identify significant differences between patients who survived and those who did not:

- **Serum Creatinine**
- **Serum Sodium**
- **Ejection Fraction**
- **Age**
- **Follow-up Time**

Significance was determined at **$\alpha = 0.05$**, with a two-tailed test assumption.

Methodology

4. Logistic Regression

A logistic regression model was developed to predict **DEATH_EVENT** based on the five most influential variables:

- **Serum Creatinine**
- **Serum Sodium**
- **Ejection Fraction**
- **Age**
- **Follow-up Time**

Model Training and Evaluation:

- The data was split into **80% training** and **20% testing** sets using **stratified sampling** to ensure balanced representation of the target class.
- The model was trained using the training data and evaluated using accuracy, precision, recall, and F1-score on the test set.
- A **confusion matrix** was plotted to visualize classification performance.

Methodology

5. Cluster Analysis

Cluster analysis using **K-Means** was performed to segment patients into risk categories.

- The **Elbow Method** was used to determine the optimal number of clusters, resulting in **three distinct clusters**.
- Each cluster was analyzed to understand the differences in patient profiles and mortality risks.

Results

1. Exploratory Data Analysis and T-test Findings:

Hypothesis testing revealed that 5 variables show significant difference between the means of the patients survived and those who died; age, follow-up time, ejection fraction, serum creatinine, and serum sodium. The variables included the transformed variables.

2. Logistic Regression Performance:

- **Accuracy**: 82%
- **AUC:** 84%.
- **Precision**: 84%
- **Recall**: 79%
- **F1-score**: 80%

The model correctly predicted the majority of cases, demonstrating strong performance in classifying heart failure outcomes.

Key Findings from Model Coefficients:

- **Serum Creatinine** had the highest positive coefficient, indicating that higher levels significantly increased the risk of death.
- **Ejection Fraction** and **follow-up time** had negative coefficients, suggesting that higher values reduced mortality risk.
- **High AUC** indicate that the model has high discriminatory power and can distinguish between those who will die than those who will not.

Follow-up time, ejection fraction, serum creatinine showed the highest accuracy of 82%. This means that these three variables are the most capable predictors of mortality risk and survival rates.

Results

3. Cluster Analysis Results:

- **Cluster 1**: High-risk patients characterized by elevated serum creatinine and lower ejection fractions. This cluster had the highest mortality rate.
- **Cluster 2**: Low-risk patients with normal serum creatinine, higher ejection fractions, and better overall health profiles.

Cluster Visualization:

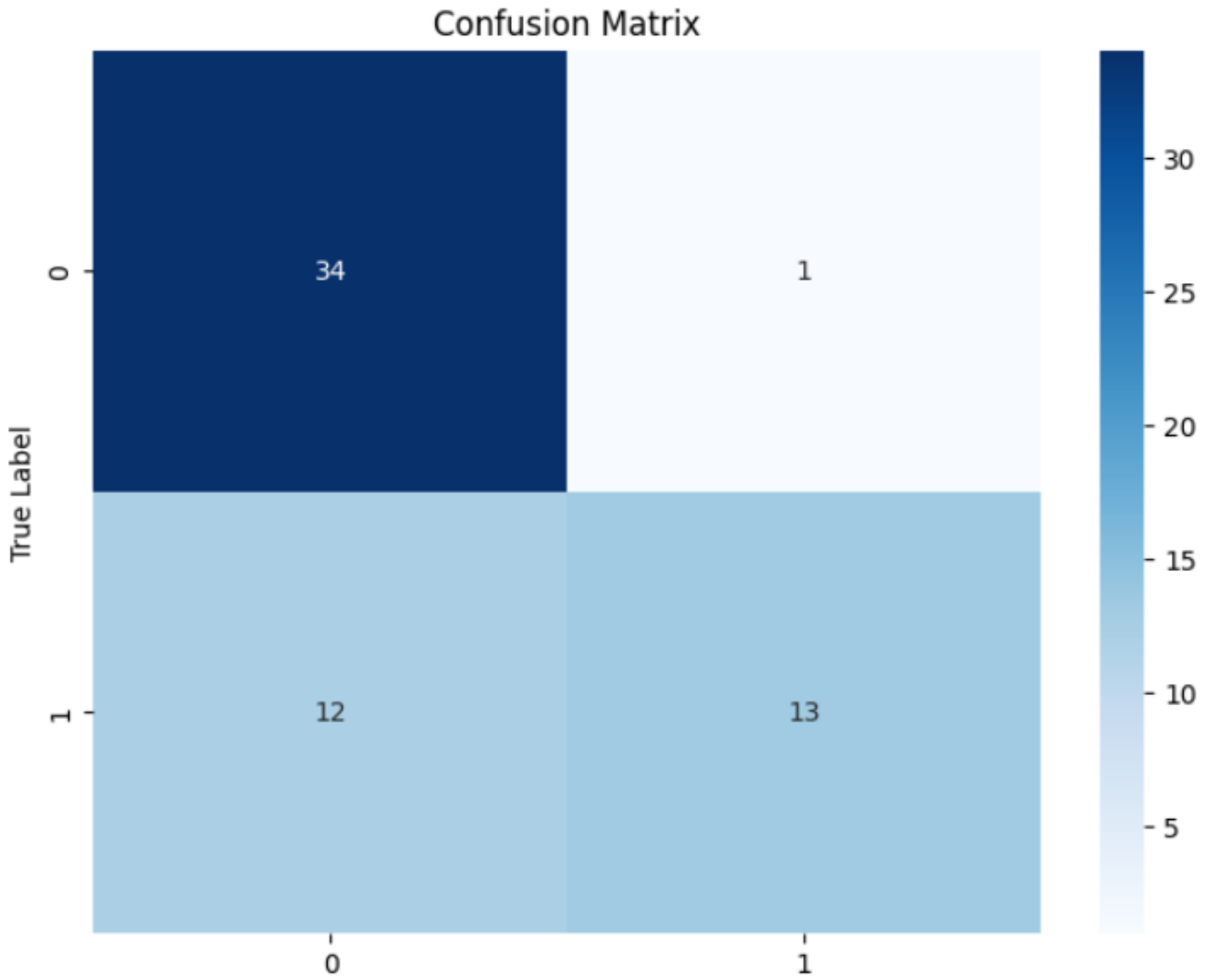
- A **scatter plot** revealed clear separation between clusters, supporting the validity of the clustering process.

Patients in **Cluster 1** were at a significantly higher risk of mortality compared to those in Clusters 2.

Appendix

Classification Report:

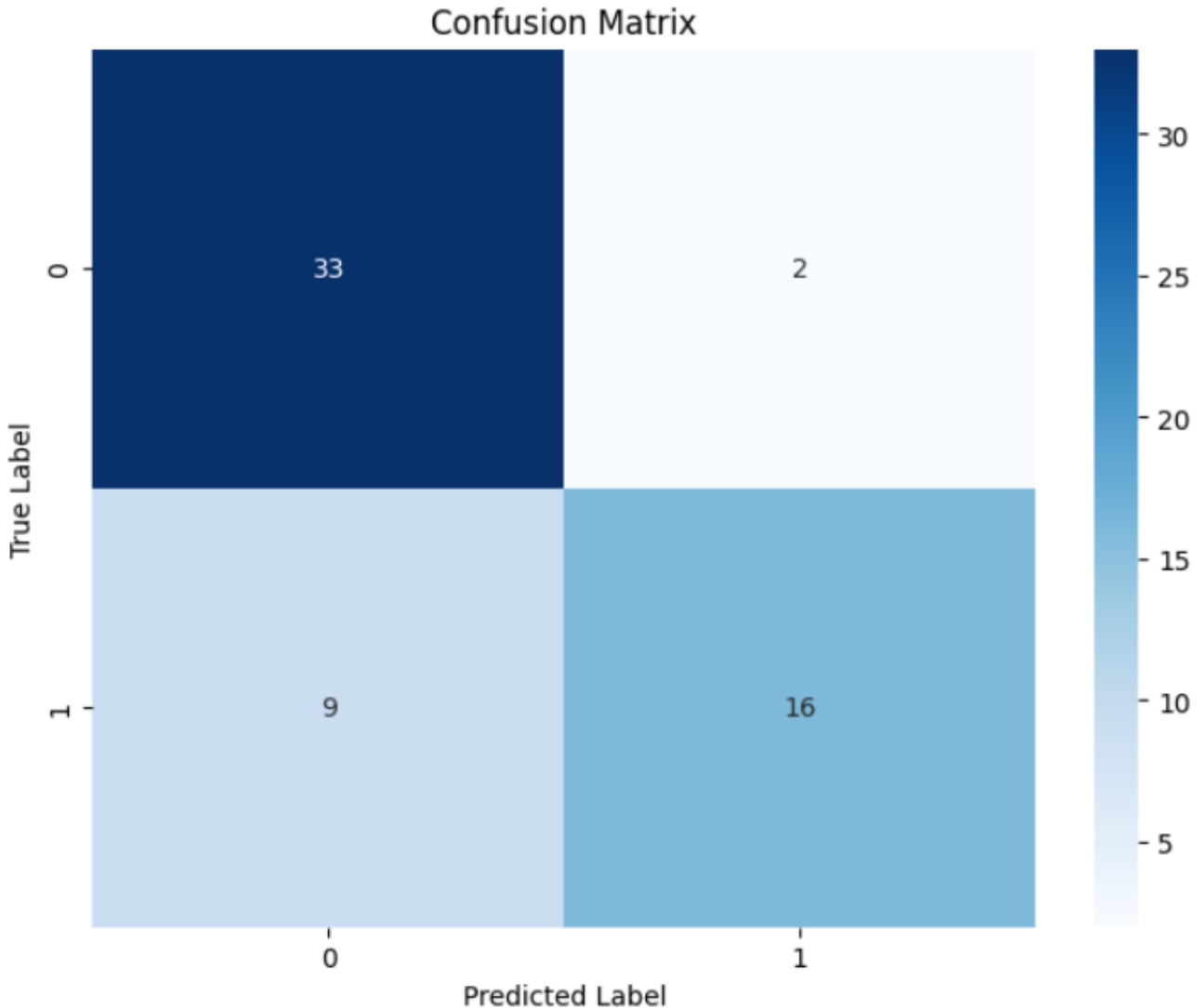
	precision	recall	f1-score	support
0	0.74	0.97	0.84	35
1	0.93	0.52	0.67	25
accuracy			0.78	60
macro avg	0.83	0.75	0.75	60
weighted avg	0.82	0.78	0.77	60



With Serum Creatinine, Serum Sodium, Age, Ejection Fraction, and Follow-up Time.

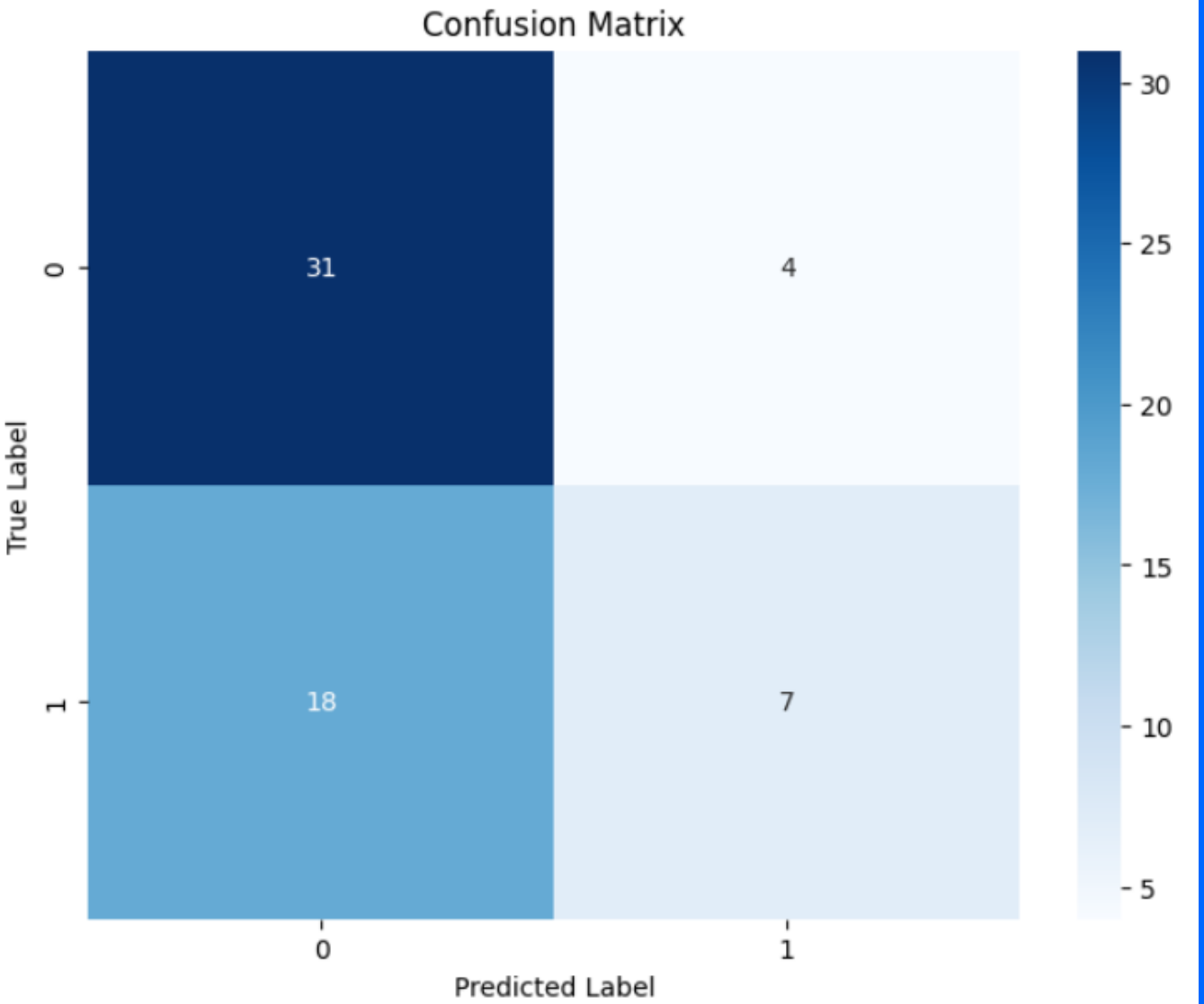
Classification Report:

	precision	recall	f1-score	support
0	0.79	0.94	0.86	35
1	0.89	0.64	0.74	25
accuracy			0.82	60
macro avg	0.84	0.79	0.80	60
weighted avg	0.83	0.82	0.81	60

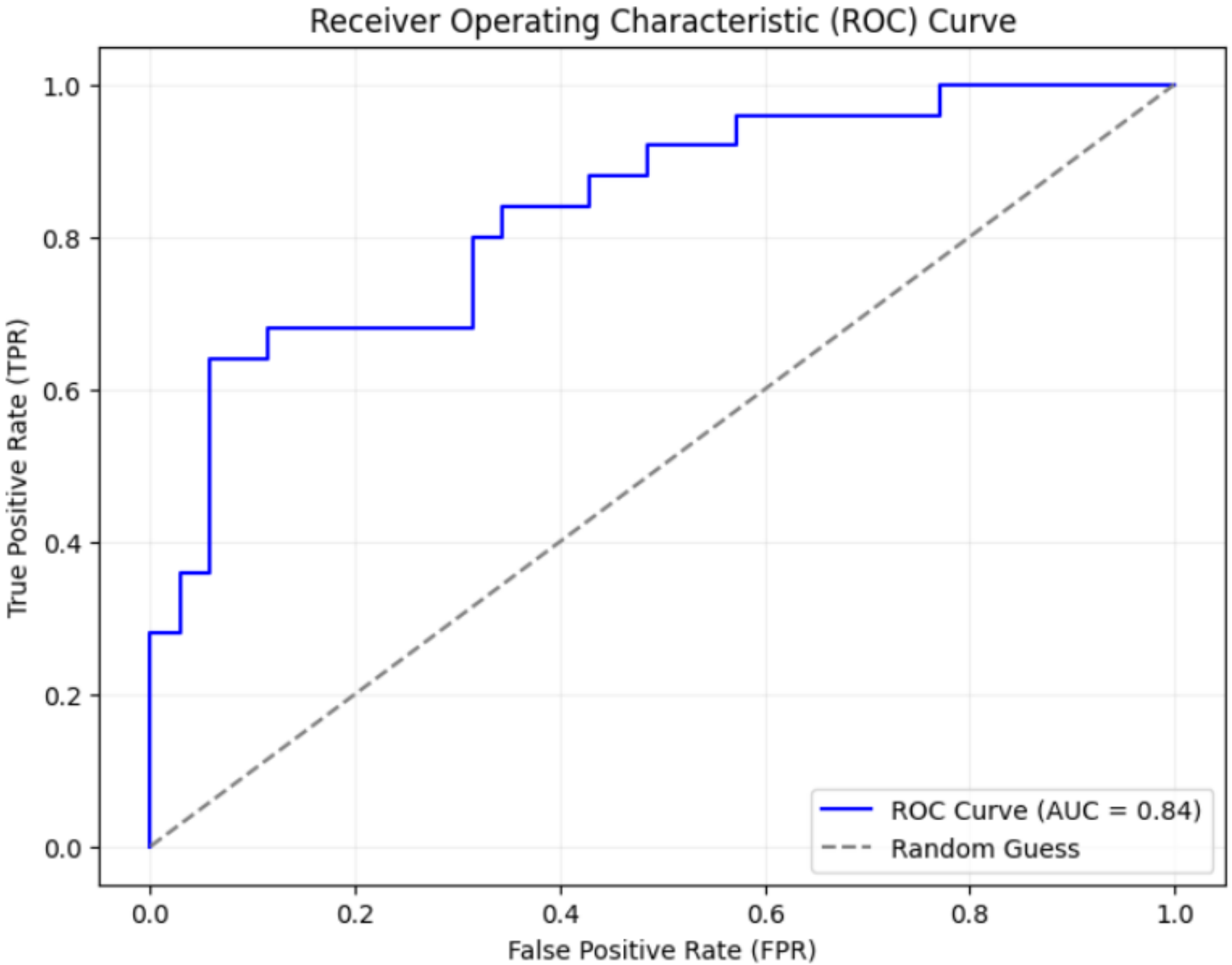


With Serum Creatinine, Ejection Fraction, and Follow-up Time.

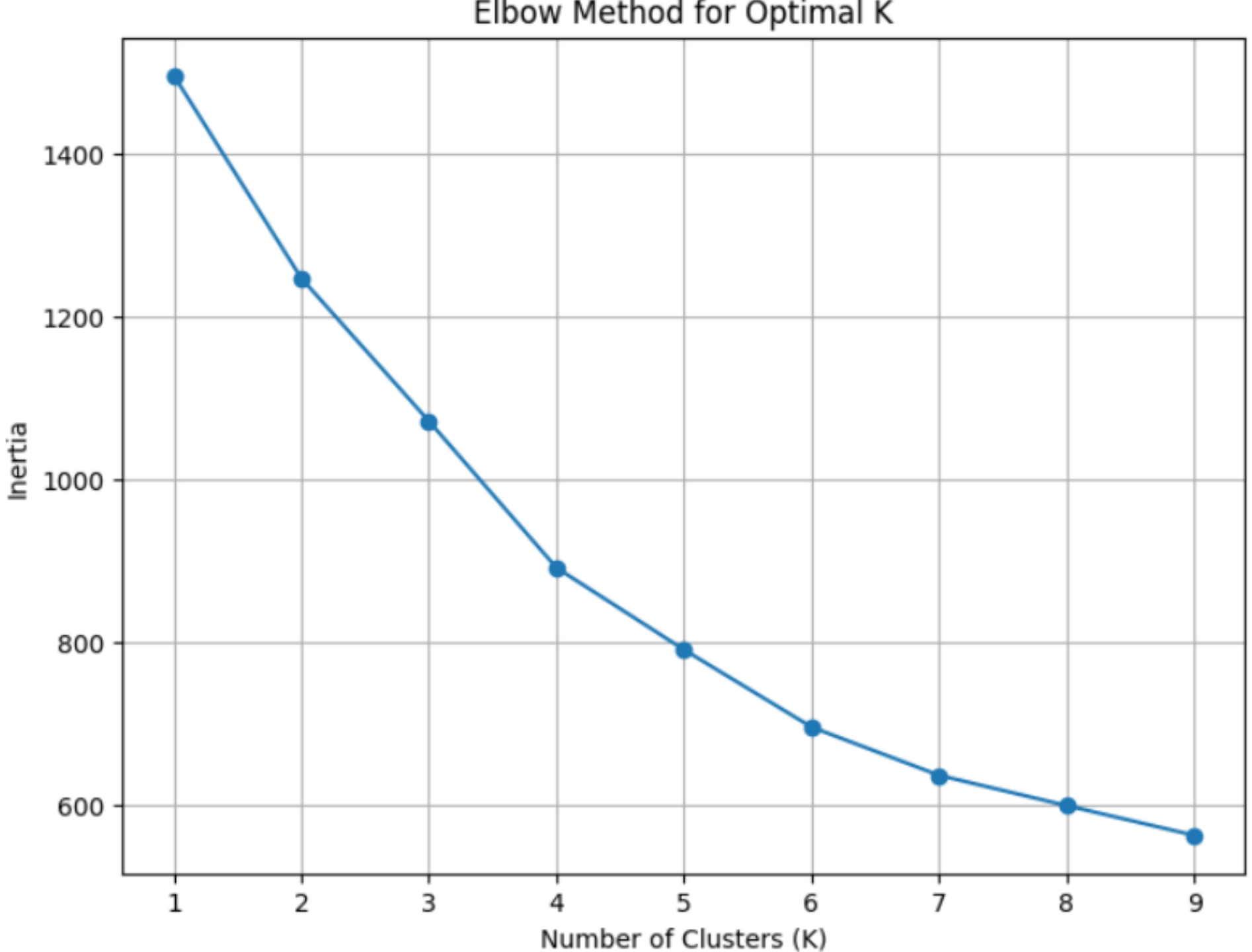
	precision	recall	f1-score	support
0	0.63	0.89	0.74	35
1	0.64	0.28	0.39	25
accuracy			0.63	60
macro avg	0.63	0.58	0.56	60
weighted avg	0.63	0.63	0.59	60



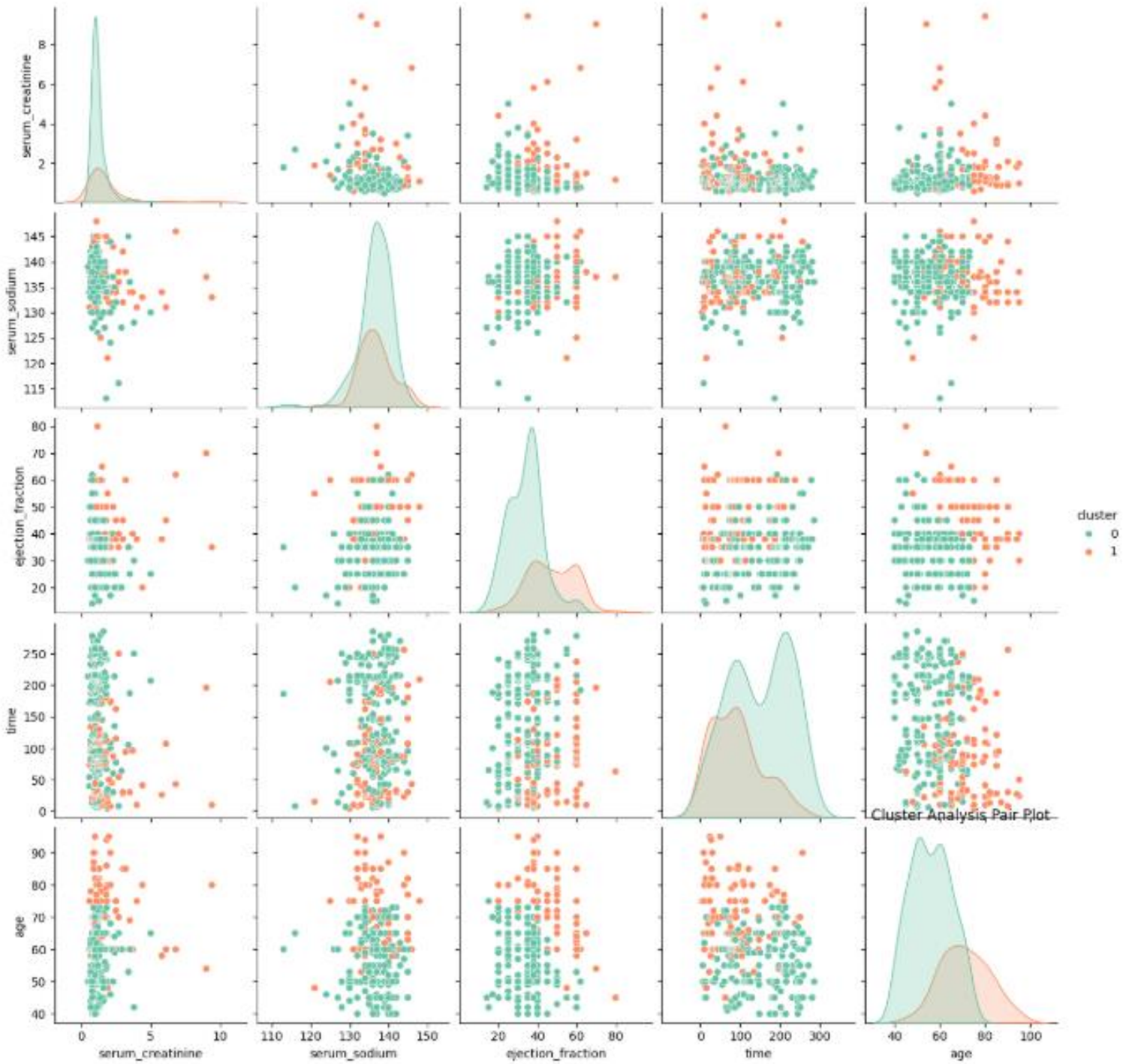
With Serum Creatinine, and Ejection Fraction.



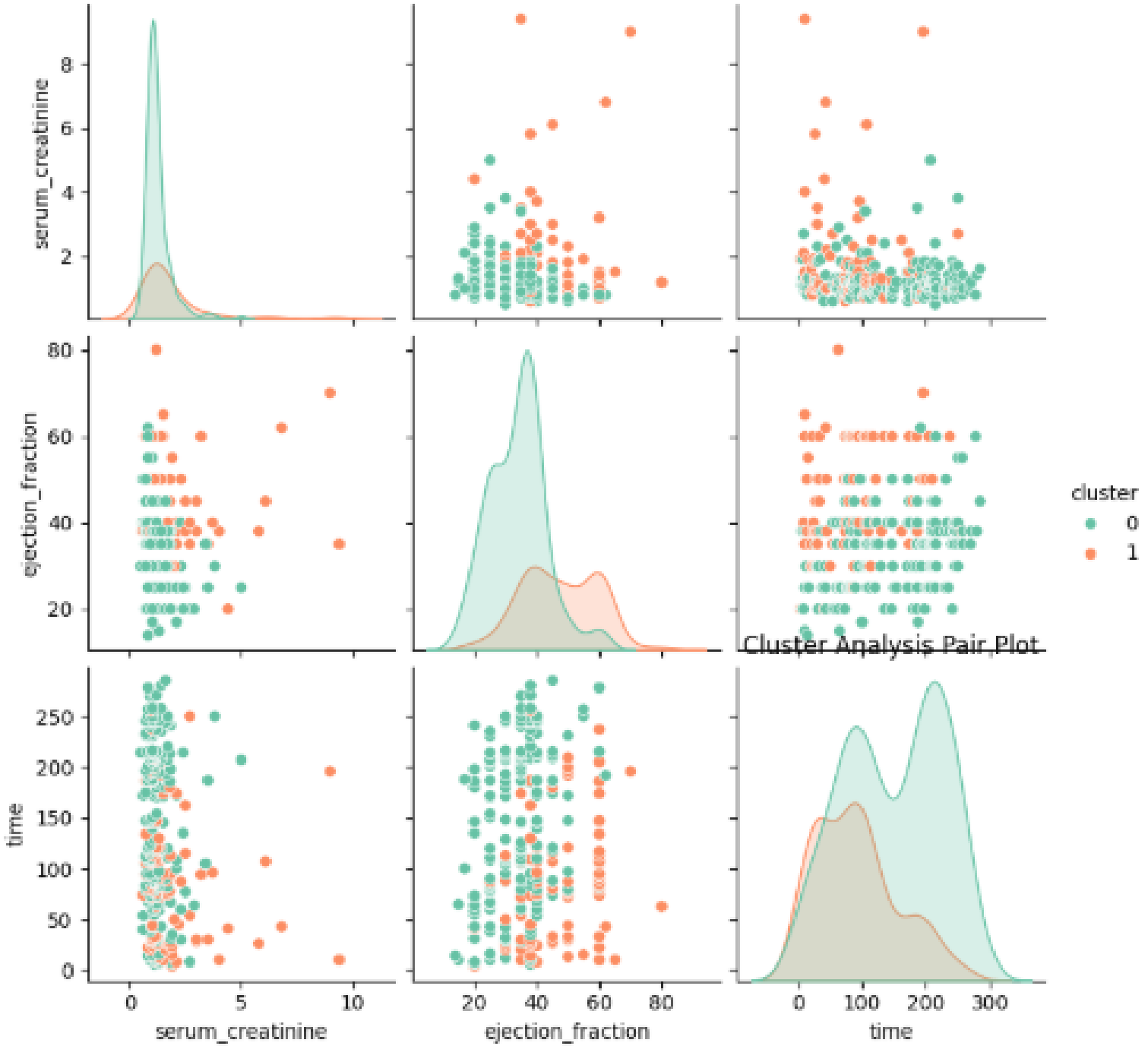
ROC Curve and AUC



Elbow Method



Pair Plot of Clusters



Pair Plot of Clusters

Thank you

- Ahmed Elsayed
- +39 392 766 6298
- a7madv4d2@gmail.com
- https://github.com/a7madv4d2
- https://www.linkedin.com/in/ahmed-elsayed-2a8208239