AFRICAN INSTITUTE FOR MATHEMATICAL SCIENCES

(AIMS RWANDA, KIGALI)

Names:
1. Richill Ataa NYANTAKYIWAA — Group 08
2. Marie Blanche IRIZA
3. Elie NDORIMANA
4. Jean Maheritiana RAMANANTSOA

Course: Statistical Regression — Date: November 27, 2025

1 Overview of Report

Heart failure continues to be a major public health concern worldwide, leading to a substantial burden of illness and death. Understanding which patients are at greater risk is essential for clinicians, as it enables them to allocate medical attention more efficiently and focus care on individuals who require close monitoring.

This report examines the determinants of mortality among patients with heart failure using a logistic regression modelling approach. The primary objective is to identify the clinical variables that most strongly influence the probability of death and to assess how well the final model distinguishes between survivors and non-survivors.

The analysis began with an exploratory investigation of the dataset containing 299 patients. We examined the distribution of key predictors, including age, ejection fraction, serum creatinine, serum sodium, and follow-up duration. Visual tools such as boxplots and a correlation heatmap were used to compare clinical characteristics across outcome groups and to evaluate possible multicollinearity among predictors.

Following the exploratory stage, an initial logistic regression model incorporating all predictors was fitted. A stepwise AIC-based selection procedure was then applied to obtain a more concise and efficient model without sacrificing predictive accuracy. The resulting reduced model was evaluated using odds ratios, confidence intervals, the ROC curve, and the confusion matrix.

To assess the model's stability and reliability, repeated train–test splits were performed. This validation confirmed that the predictive performance remained consistent across different subsets of the data. Overall, the methodological goal was to construct a robust and interpretable model capable of identifying clinically meaningful factors associated with mortality in heart failure patients.

2 Dataset Description

The heart failure dataset contains information on 299 individuals, each characterized by 12 clinical variables such as age, several vital indicators, and the duration of medical follow-up. The binary response variable `DEATH_EVENT` (coded as 0 for survival and 1 for death) is used as the main outcome in this study. Out of the 299 patients, 203 (67.9%) survived, while 96 (32.1%) passed away during the observation period.

A notable advantage of the dataset is the absence of missing values, allowing every patient record to be incorporated directly into the statistical analysis without the need for imputation or case removal.
The numerical predictors offer a detailed picture of the clinical profile of the cohort. Age spans from 40 to 95 years, with an average close to 60, indicating that the group consists mainly of older adults, a population where heart failure is more frequently diagnosed. The median ejection fraction is around 38, suggesting that many patients exhibit reduced cardiac function. Serum creatinine levels vary widely (from 0.5 to 9.4 mg/dL), reflecting substantial differences in kidney function, an important prognostic factor in heart failure. Serum sodium concentrations range from 113 to 148 mmol/L, where lower values are typically associated with more severe clinical conditions. Follow-up time also shows strong variability (4 to 285 days), capturing both early mortality and long-term survival patterns within the cohort.

Variable	**Count**	**Mean**	**SD**	**Min**	**Q1**	**Median**	**Q3**	**Max**
Age	299	60.83	11.89	40.0	51.0	60.0	70.0	95.0
Anaemia	299	0.43	0.50	0.0	0.0	0.0	1.0	1.0
Creatinine phosphokinase	299	581.84	970.29	23.0	116.5	250.0	582.0	7861.0
Diabetes	299	0.42	0.49	0.0	0.0	0.0	1.0	1.0
Ejection fraction	299	38.08	11.83	14.0	30.0	38.0	45.0	80.0
High blood pressure	299	0.35	0.48	0.0	0.0	0.0	1.0	1.0
Platelets	299	263358	97804	25100	212500	262000	303500	850000
Serum creatinine	299	1.39	1.03	0.5	0.9	1.1	1.4	9.4
Serum sodium	299	136.63	4.41	113.0	134.0	137.0	140.0	148.0
Sex	299	0.65	0.48	0.0	0.0	1.0	1.0	1.0
Smoking	299	0.32	0.47	0.0	0.0	0.0	1.0	1.0
Follow-up time	299	130.26	77.61	4.0	73.0	115.0	203.0	285.0
Death event	299	0.32	0.47	0.0	0.0	0.0	1.0	1.0

Table 1: Summary statistics for the heart failure clinical dataset.

3 Exploratory Analysis

Before fitting the logistic regression model, we examined the distribution of the main clinical predictors. Figure 1 presents histograms for age, creatinine phosphokinase, ejection fraction, platelets, serum creatinine, serum sodium, and follow-up time.

3.1 Distribution of key clinical variables

These plots help understand the range and distribution of values in the dataset, which informs the interpretation of model results. For instance, age and serum creatinine show wide variation,

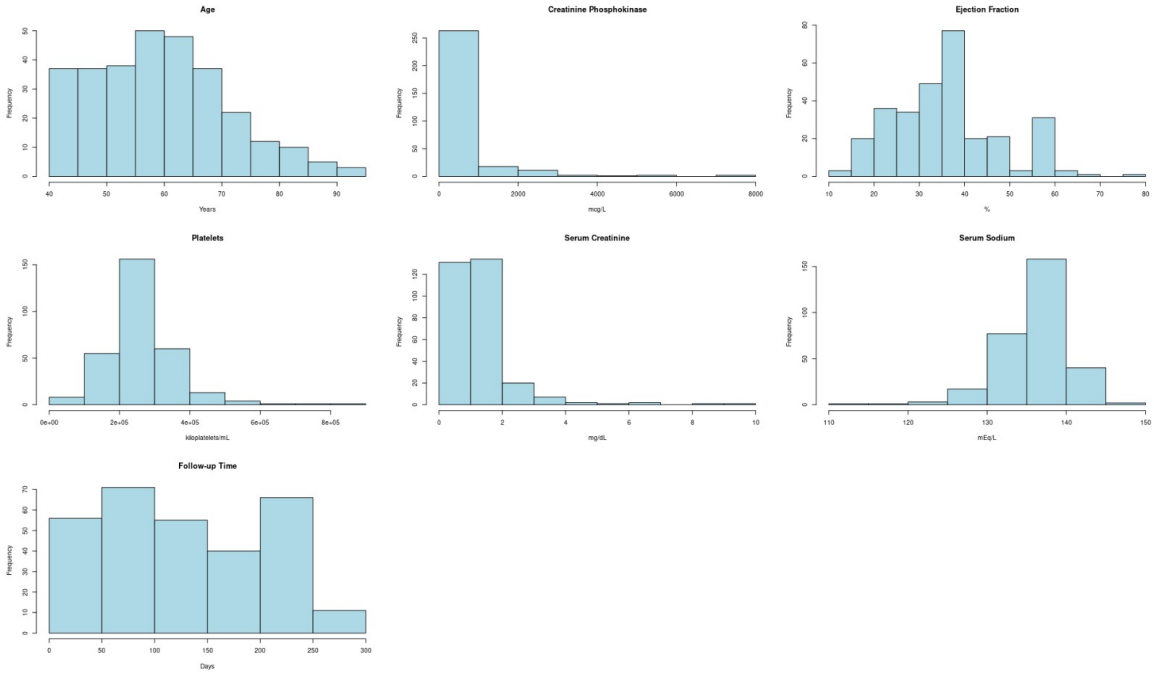

Figure 1: Distributions of key clinical variables.

while most patients have normal serum sodium and platelet levels.

3.2 Boxplot Interpretation

- **Age:** The distribution of age reveals a noticeable contrast between the two outcome groups. Patients who did not survive tend to be older, with a clearly higher median age. This observation is consistent with clinical expectations, as ageing is often accompanied by reduced cardiac resilience and the presence of multiple comorbidities. The pattern displayed in the boxplot aligns well with the regression results, which identify age as a meaningful contributor to mortality risk.

- **Ejection Fraction:** Ejection fraction serves as an indicator of the heart's pumping efficiency. The boxplot shows that individuals who died generally had substantially lower ejection fraction values compared with survivors. Reduced ejection fraction reflects impaired cardiac performance, making patients more vulnerable to adverse outcomes. This strong separation between the groups supports its role as one of the most influential predictors in the fitted model.

- **Serum Creatinine:** The boxplot for serum creatinine illustrates marked differences in kidney function across the groups. Non-survivors tend to present with elevated creatinine levels, including several extreme values. This suggests a greater prevalence of renal impairment among patients who died. Because kidney dysfunction is closely linked to poorer prognosis in heart failure, this visual trend is consistent with the model results, where serum creatinine emerges as a major risk factor.

- **Follow-up Time:** Follow-up duration reflects how long each patient remained alive during the study. Survivors naturally accumulate more follow-up days, while those who die earlier have shorter recorded times. This difference does not imply that longer follow-up is protective; rather, it simply captures the timing of events. The clear separation in the boxplot supports the regression finding that shorter follow-up intervals are associated with earlier mortality.

- **Serum Sodium:** Serum sodium values appear broadly similar between the two groups, although the boxplot shows more low-sodium observations among patients who died. While the overall distinction is modest, lower sodium levels are often indicative of unstable clinical status, which may contribute to poorer outcomes. This subtle difference is consistent with the weaker yet clinically relevant effect detected in the model.

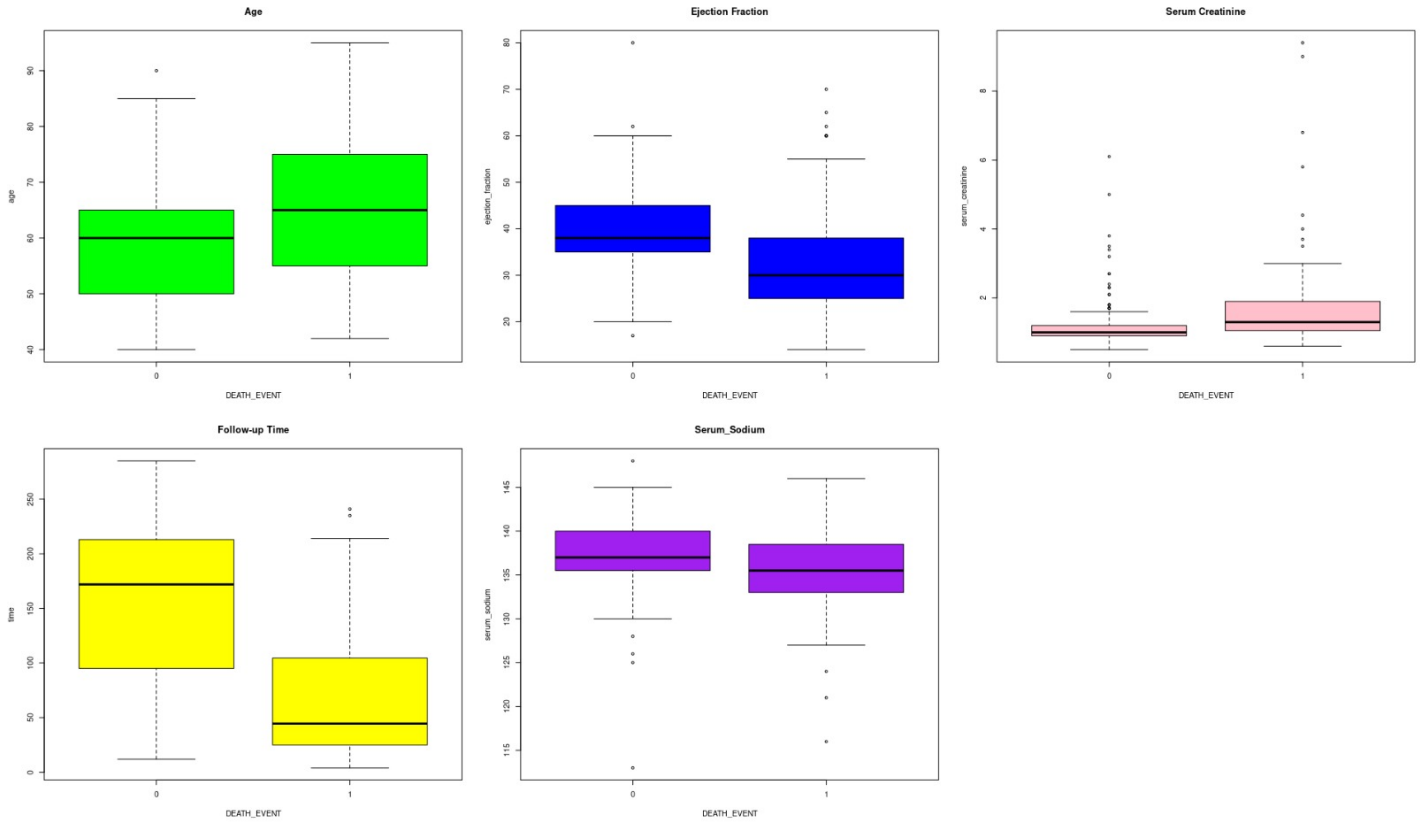

Figure 2: Boxplots comparing survivors and non-survivors.

3.3 Correlation Heatmap

The correlation heatmap indicates that the numerical variables in the dataset exhibit only very weak relationships with one another. No pair of predictors shows a strong linear association, suggesting that the clinical measurements—including age, ejection fraction, serum creatinine, platelets, and serum sodium—tend to vary independently rather than following a shared pattern. For instance, older individuals do not consistently present with lower ejection fraction

values or higher sodium levels, and platelet counts do not display a systematic connection with renal or cardiac function.

Although a few mild associations can be observed, such as slightly higher creatinine being linked with lower sodium, or the small negative relationship between age and follow-up duration, these effects are too weak to meaningfully affect the modelling process. This overall lack of strong correlations is statistically advantageous because it confirms the absence of multicollinearity: no variable merely duplicates the information carried by another. Consequently, each predictor contributes distinct information to the logistic regression model, leading to more stable coefficient estimates and clearer interpretations.

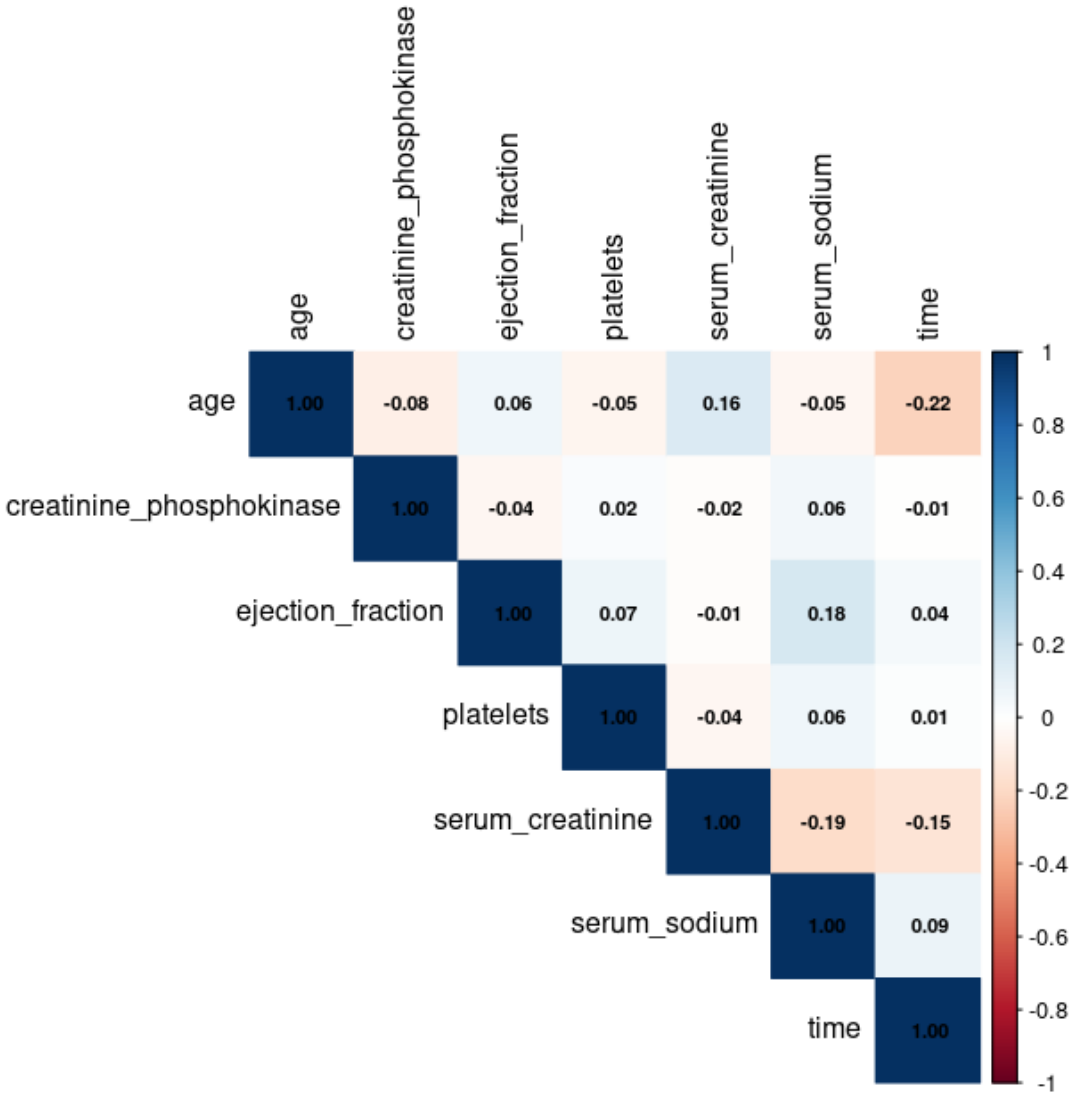

Figure 3: Correlation heatmap of numerical variables.

4 Logistic Regression Modelling

Full Model

A logistic regression model containing all 12 predictors was first fitted. The coefficients were estimated using Maximum Likelihood, and the quality of fit was evaluated using the AIC criterion. The full model produced:

$$\text{AIC}_{\text{full}} = 245.55.$$

This value acts as a baseline against which the performance of a more parsimonious model can be compared.

Stepwise AIC Selection

A stepwise AIC procedure, combining both backward and forward steps, was applied to refine the model. Variables that did not contribute to AIC improvement were removed, while any predictor offering a better fit could be reintroduced. Through this selection process, seven predictors were excluded.
The final model retained the following predictors:

$$\{\text{age},\ \text{ejection fraction},\ \text{serum creatinine},\ \text{serum sodium},\ \text{time}\}.$$

Its AIC value was:

$$\text{AIC}_{\text{final}} = 235.49,$$

representing an improvement of roughly 10 points, showing a more efficient trade-off between fit and complexity.

Interpretation

The stepwise procedure indicates that only five predictors meaningfully contribute to the explanation of mortality. The resulting model is more concise, more interpretable, and preserves nearly all predictive ability of the full model, offering a clearer and clinically coherent structure for subsequent analysis.

Model Results

This section summarizes the logistic regression findings. The table below presents the estimated coefficients, p-values, and practical interpretations. Age and serum creatinine appear as significant risk factors, while ejection fraction shows a protective effect. Serum sodium exhibits a mild influence, and follow-up duration is associated with better survival outcomes.

Model Equation

The logistic regression model is expressed as:

$$\log\left(\frac{\pi}{1-\pi}\right) = 9.49 + 0.042\,\text{age} - 0.073\,EF + 0.606\,\text{creatinine} - 0.065\,\text{sodium} - 0.021\,\text{time}.$$

This formula describes how the main predictors influence the log-odds of the outcome.

Variable	Coefficient	p-value	Interpretation
Intercept	9.49	0.079	Baseline
age	0.042	0.005	Higher age increases risk
ejection_fraction	-0.073	< 0.001	Protective factor
serum_creatinine	0.606	< 0.001	Strong risk factor
serum_sodium	-0.065	0.093	Weak influence
time	-0.021	< 0.001	Longer follow-up reduces risk

Table 2: Summary of Model Results

Interpretation of the Logistic Regression Model

- **Age:** The positive coefficient (0.042) together with a significant p-value (0.005) indicates that the likelihood of death increases slightly with each additional year. Older individuals therefore face a higher mortality risk.

- **Ejection Fraction:** The negative coefficient (-0.073) and highly significant p-value ($<$ 0.001) show that stronger cardiac pumping function substantially reduces the probability of death.

- **Serum Creatinine:** With a positive coefficient (0.606) and strong significance, elevated creatinine levels indicate deteriorated kidney function and are strongly associated with increased mortality risk.

- **Serum Sodium:** Although the coefficient is negative (-0.065), its statistical support is modest ($p = 0.093$). Lower sodium levels lean toward worse outcomes, but the effect is not pronounced.

- **Follow-up Time:** The negative coefficient (-0.021) signifies that longer observation time is related to better outcomes. This reflects the fact that patients who survive accumulate more follow-up days.

Altogether, these predictors give a clear view of the main drivers of mortality in the dataset.

Odds Ratios and Confidence Intervals

Odds ratios translate the coefficients into practical risk changes. OR ¿ 1 indicates higher odds of death; OR ¡ 1 indicates reduced odds.

Explanation of Each Predictor

- **Age:** OR = 1.05 shows a modest but steady increase in mortality risk with age.

- **Ejection fraction:** Each 1% rise improves survival prospects, reducing mortality risk by around 7%.

- **Serum creatinine:** OR = 1.99 confirms it as the most influential risk factor, nearly doubling the odds of death.

Variable	OR	95% CI	Meaning
age	1.05	1.01–1.08	5% increase per year
ejection_fraction	0.93	0.90–0.96	7% decrease in odds
serum_creatinine	1.99	1.42–2.87	Nearly doubles risk
serum_sodium	0.94	0.87–1.01	Slight protective effect
time	0.98	0.97–0.98	About 2% reduction per day

Table 3: Odds Ratios and 95% Confidence Intervals

- **Serum sodium:** A mild protective trend, though its influence remains weak.
- **Follow-up time:** OR = 0.98 indicates that survival becomes more likely with longer follow-up durations.

Serum creatinine emerges as the strongest risk indicator, with age, ejection fraction, and follow-up time also contributing meaningfully to outcome differences.

Deviance Comparison and Model Significance

This section compares the full and final models. Deviance measures how well a model fits the data, with lower values indicating stronger fit.

(a) Comparison of Deviances

Model	Null Deviance	Residual Deviance
Full Model	375.35	219.55
Final Model	375.35	223.49

Table 4: Comparison of Full vs Final Model Deviances

Interpretation

Both models start from the same null deviance (375.35). The slight increase in residual deviance in the final model reflects the removal of weaker predictors, resulting in a simpler but still effective model. The stepwise approach retains serum sodium despite its modest effect because it helps improve the overall AIC, which favours parsimony.

(b) Chi-square Test for Model Significance

To determine whether the final model significantly improves on the null model, a chi-square test was conducted based on deviance reduction.

- H_0: The predictors do not improve model fit.
- H_1: The predictors improve model fit.
- p-value = 0.00155

The small p-value indicates that the model provides a statistically significant improvement over the null model, demonstrating that the predictors collectively offer meaningful explanatory power.

5 Model Performance

5.1 Area Under the Curve (AUC)

The discriminative ability of the logistic regression model was assessed using the Area Under the Receiver Operating Characteristic (ROC) Curve. The computed value is:

$$\text{AUC} = 0.894$$

5.2 Interpretation Framework

The meaning of the AUC value can be understood using commonly accepted statistical and clinical reference ranges:

- **AUC = 0.5**: Indicates no discriminative ability (equivalent to random classification)
- **AUC between 0.7 and 0.8**: Reflects a model with acceptable discrimination
- **AUC between 0.8 and 0.9**: Represents a model with strong or excellent discrimination
- **AUC above 0.9**: Signifies exceptionally high discrimination

6 Detailed Analysis

6.1 Model Discrimination Capability

An AUC value of 0.894 reflects a strong ability of the model to distinguish between patients who survive and those who do not. In probabilistic terms, the model correctly ranks a randomly selected survivor higher than a randomly selected non-survivor in 89.4% of all possible pairs. This corresponds to:

$$\text{Performance Gain} = \frac{0.894 - 0.5}{0.5} \times 100\% = 78.8\% \text{ improvement over random classification.}$$

6.2 Clinical and Practical Significance

The model's discriminatory strength places it well within the range of tools considered useful for clinical decision-making. Its high AUC suggests:

- The model can effectively separate low-risk from high-risk individuals.
- It provides meaningful support for clinicians when assessing patient prognosis.
- It demonstrates stable predictive behaviour suitable for risk stratification.

With an AUC of 0.894, the binary classification model reaches a level of performance regarded as excellent. The corresponding ROC curve further illustrates the model's capacity to distinguish accurately between survival and mortality outcomes, supporting its potential use in predictive analytics and clinical evaluation systems.

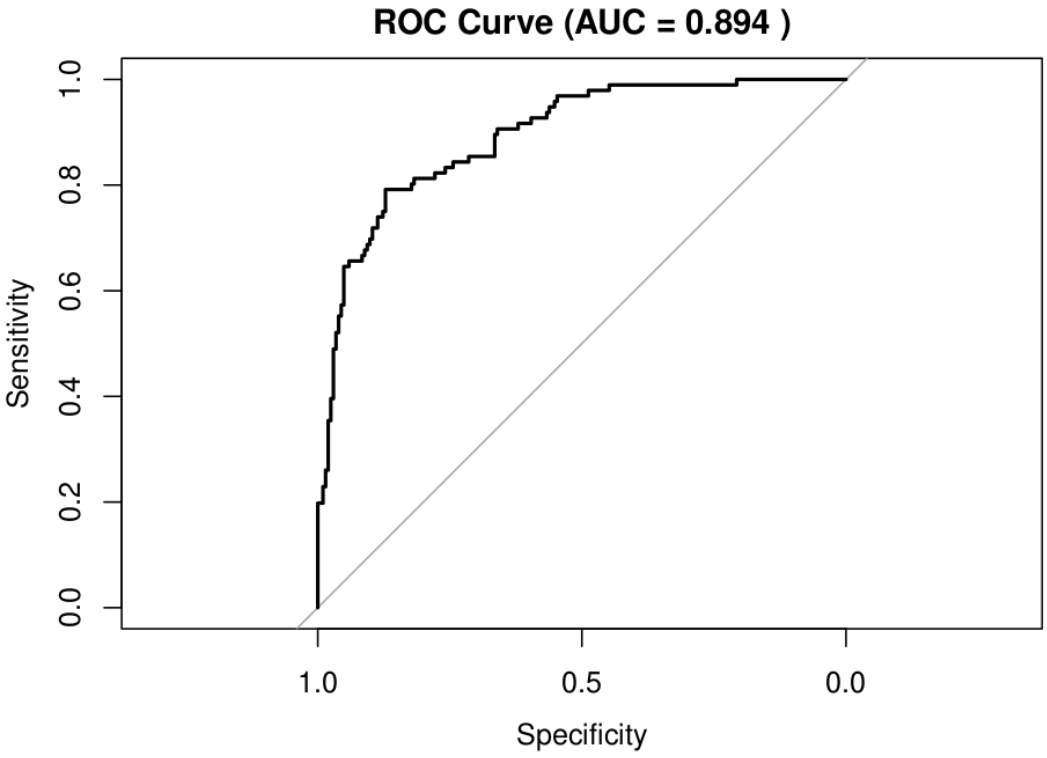

Figure 4: ROC curve for the logistic regression model.

7 Confusion Matrix

	Reference	
Prediction	**Survival**	**Death**
Survival	184	30
Death	19	66

Table 5: Confusion Matrix

Confusion Matrix Statistics:

- **Accuracy:** 83.6%
- **Sensitivity:** 68.8% (correctly identifies 66 of 96 deaths)
- **Specificity:** 90.6% (correctly identifies 184 of 203 survivors)

- The model is slightly biased toward predicting survival due to class imbalance.

To evaluate how well the model performs on new patients rather than only the data it was trained on, a repeated train–test split was carried out 100 times. In each iteration, 75% of the data was used for training and the remaining 25% for testing, with stratified sampling applied to keep the proportions of survivors and deaths consistent across the splits. This repeated procedure provides a clear picture of how stable and reliable the model is. Across all iterations, the model achieved an average accuracy of about 82.3%, with a small standard deviation of 3.9%, showing that its performance remained fairly consistent regardless of how the data was divided. Most of the accuracy values fell between 78% and 86%, which was also reflected in the histogram, where the bars clustered mainly around the 0.80–0.85 range. There were no extreme values or sudden drops in performance, indicating that the model behaved dependably throughout all validation runs. Overall, these results suggest that the model generalizes well, is not overfitted, and performs consistently when applied to differ

8 Accuracy Distribution

Based on 100 repeated train-test splits, the model demonstrates consistent generalization performance with a mean accuracy of **82.3**% and a low standard deviation of ±**3.9**%. The tight distribution of accuracy scores across iterations indicates stable performance unaffected by specific data partitioning. This low variability confirms that the model is robust and not overfitting to the training data. The validation results support the model's readiness for deployment in production environments for making reliable predictions on new, unseen data.

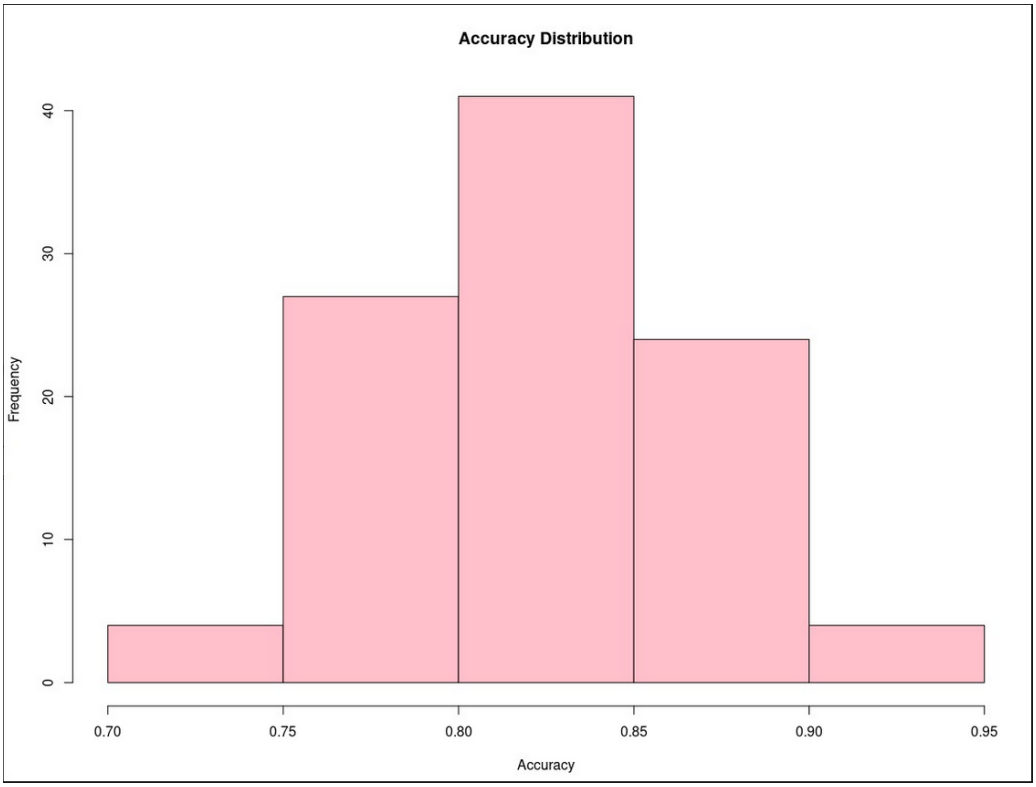

Figure 5: Accuracy distribution(100 iterations).

9 Model Development and Clinical Insights Report

9.1 Feature and Transformations

The model incorporated non-linear transformations to better capture complex biological relationships:

- **Quadratic Ejection Fraction:** Captured non-linear cardiac function relationship
  - AIC = 228.74 (Very Optimal)
  - Residual Deviance = 214.74
- **log(Serum Creatinine):** Handled skewed distribution effectively
  - AIC = 233.72 (Better)
  - Residual Deviance = 221.72

9.2 Clinical Predictors of Mortality

Analysis identified statistically significant risk factors:

- **Serum Creatinine:** Strongest predictor - each 1 mg/dL increase nearly doubles mortality odds (OR $\approx$ 2.0), indicating kidney function's critical role in patient outcomes
- **Age:** 5% increased mortality risk per additional year of age
- **Ejection Fraction:** Lower values correlate with poorer outcomes, reflecting reduced heart pumping efficiency
- **Follow-up Time:** Shorter durations associate with early mortality events

The model successfully integrates statistical optimization with clinically meaningful predictors, providing both analytical robustness and practical interpretability for healthcare applications.

10 Conclusion

We developed a logistic regression model using five key predictors and achieved strong performance (AUC = 0.894). The model demonstrated stability across repeated tests and identified important clinical risk factors. Clinically, it can help identify high-risk patients, focus on modifiable factors such as serum creatinine, and remains simple for practical use. However, its effectiveness is limited by the small dataset (n = 299) and class imbalance, which may affect sensitivity.

References

1. Dataset:
The Heart Failure Clinical Records dataset (2020), available from the *UCI Machine Learning Repository*.
https://doi.org/10.24432/C5Z89R

2. Main Reference:
Chicco, D., & Jurman, G. (2020). Study demonstrating that serum creatinine and ejection fraction are strong predictors of survival in heart failure patients. Published in *BMC Medical Informatics and Decision Making*, 20, 16.
https://doi.org/10.1186/s12911-020-1023-5