

# Heart Failure Logistic Regression Analysis

GROUP 8

Elie NDORIMANA

Marie Blanche IRIZA

Richill Ataa NYANTAKYIWAA

Jean Maheritiana RAMANANTSOA

November 26, 2025

# Presentation Outline

- 1 Introduction
- 2 Dataset Description
- 3 Exploratory Data Analysis
- 4 Model Development Process
- 5 Deviance Test and Model Performance
- 6 Model Validation
- 7 Clinical Implications
- 8 Conclusions

**Heart failure** is a major public health concern worldwide.

Understanding which factors predict mortality can help:

- Identify high-risk patients early
- Guide treatment decisions
- Allocate healthcare resources effectively

**Our approach:** Use logistic regression to build a simple, interpretable prediction model from clinical data.

**Dataset:** 299 heart failure patients from UCI Repository [1]

## Outcome Variable:

- DEATH\_EVENT: 1 = died, 0 = survived
- Class balance: 203 survived (68%), 96 died (32%)

## Predictors (12 variables):

- Demographics: age, sex
- Medical history: anaemia, diabetes, high blood pressure, smoking
- Lab values: creatinine phosphokinase, ejection fraction, platelets, serum creatinine, serum sodium
- Follow-up: time (days)

**Good news:** No missing values!

# Exploratory Data Analysis

## What we looked at:

- Distributions of continuous variables by outcome
- Box plots for key predictors
- Correlation matrix for multicollinearity check

## Key observations:

- Patients who died had higher serum creatinine
- Lower ejection fraction associated with death
- Shorter follow-up time for deceased patients
- Aged people tend to die more than young adults.
- All the key predictors are approximately normal except Serum Sodium which is right skewed.

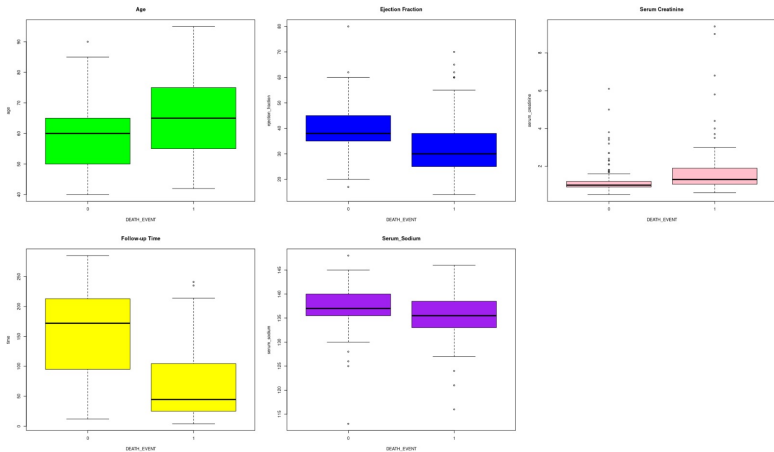


Figure 1: Box plots for key predictors.

# Correlation Analysis

**Finding:** No strong correlations between predictors ( $|r| < 0.22$ ). Meaning:

- No multicollinearity problems
- Each predictor provides unique information

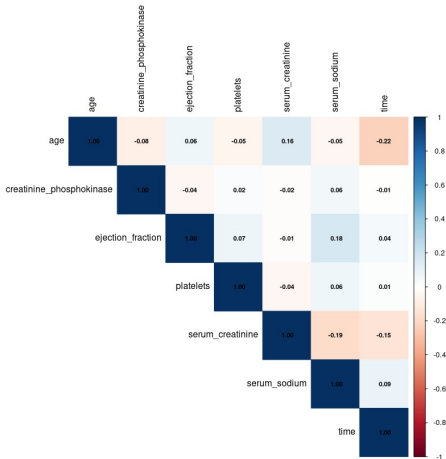


Figure 2: Correlation matrix for multicollinearity check.

# Model Building Process

## Step 1: Full Model

- Started with all 12 predictors
- Used Maximum Likelihood Estimation (MLE)
- $AIC = 245.55$

## Step 2: Stepwise Selection

- Used AIC criterion for variable selection
- Automatic backward/forward procedure
- Removed 7 non-significant variables

## Final Model (5 predictors):

- age, ejection\_fraction, serum\_creatinine, serum\_sodium, time
- AIC improved to 235.49 (optimal)

# Model Results

Variable	Coefficient	p-value	Interpretation
Intercept	9.49	0.079	Baseline
age	0.042	0.005	Risk increases with age
ejection_fraction	-0.073	< 0.001	Protective factor
serum_creatinine	0.606	< 0.001	Strong risk factor
serum_sodium	-0.065	0.093	Weak but important
time	-0.021	< 0.001	Longer follow-up = survival

Table 1: Summary of Model Results

## Model equation:

$$\log\left(\frac{\pi}{1-\pi}\right) = 9.49 + 0.042age - 0.073EF + 0.685creatinine - 0.065sodium - 0.021time \quad (1)$$

# Odd Ratio and Confidence Interval

- $OR = e^{\beta}$  tells how odds of death change per unit increase
- $OR = 1 \rightarrow$  no effect death risk
- $OR > 1 \rightarrow$  increases death risk
- $OR < 1 \rightarrow$  decreases death risk (protective)

Variable	OR	95% CI	Meaning
age	1.05	1.01–1.08	5% higher odds per year
ejection_fraction	0.93	0.90–0.96	7% lower odds
serum_creatinine	1.99	1.42–2.87	Nearly doubles odds
serum_sodium	0.94	0.87–1.01	Slightly lower odds per unit
time	0.98	0.97–0.98	2% lower odds per day

Table 2: Odds Ratios and 95% Confidence Intervals

**Most important predictor:** Serum creatinine (kidney function)

# Deviance Comparison and Model Significance

Model	Null Deviance	Residual Deviance
Full Model	375.35	219.55
Final Model	<b>375.35</b>	<b>223.49</b>

Table 3: Comparison of Full vs Final Model Deviances

**Interpretation:** The final model residual deviance is higher than the full model due to AIC retention of the *serum\_sodium* variable which reduces the value of AIC.

## Chi-square test:

$H_0$  : The model is not covariate (null)

$H_1$  : The model is covariate (final model)

p-value = 0.00155, reject null hypothesis.

**Conclusion:** Our model is indeed a good fit.

# Model Performance

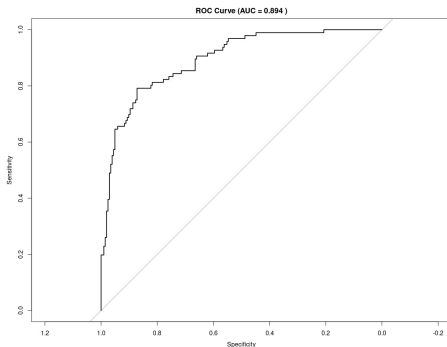


Figure 3: Model Discrimination Check.

## Interpretation:

- Area Under the Curve (AUC) = 0.894
- Model can perfectly distinguish between survival and death.
- Much better than random (0.5) guess.

# Confusion Matrix

Prediction	Reference	
	Survival	Death
Survival	184	30
Death	19	66

Table 4: Confusion Matrix

## Confusion Matrix Statistics:

- Accuracy: 83.6%
- Sensitivity: 68.8% (catches 66 of 96 deaths)
- Specificity: 90.6% (catches 184 of 203 survivors)
- Model slightly biased toward predicting survival due to class imbalance.

# Model Validation

Validation to check if model works on new data

**Method:** Train/Test Split (repeated 100 times)

- 75% training, 25% testing
- Used stratified sampling to maintain class balance
- Each iteration: fit model on train, test on holdout

**Results:**

- Mean accuracy: 82.3%
- Standard deviation: 3.9%
- Consistent performance.

# Accuracy Distribution

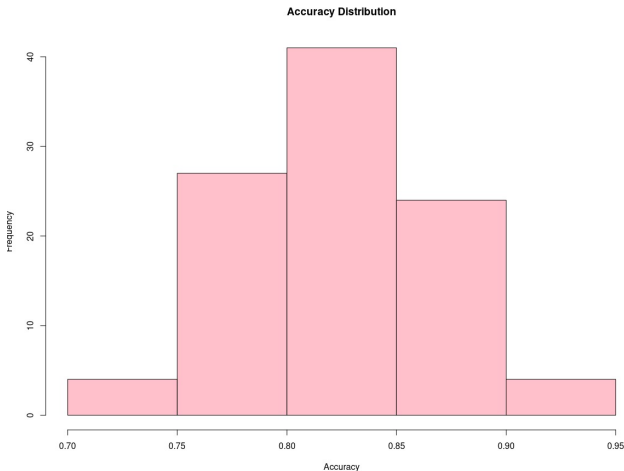


Figure 4: Accuracy Distribution over 100 iterations.

## Quadratic Ejection Fraction:

- Captures non-linear effects, mortality may accelerate at very low ejection fractions with a p-value of 0.002
- AIC=228.74 (Very Optimnal)
- Residual Deviance=214.74

## Logarithm of serum creatinine:

- Handles its skewed distribution with a p-value close to 0
- AIC=233.72 (Better than our final model)
- Residual Deviance=221.72

# Conclusions

## What we accomplished:

- Built a logistic regression model with 5 key predictors
- Achieved excellent discrimination ( $AUC = 0.894$ )
- Validated model stability through repeated testing
- Identified clinically meaningful risk factors

## Clinical usefulness:

- Can identify high-risk patients for monitoring
- Focus on modifiable factors (serum creatinine)
- Simple model and easy to use in practice

## Limitations:

- Small dataset ( $n=299$ )
- Class imbalance affects sensitivity

## 1. Dataset:

Heart Failure Clinical Records Dataset. (2020). UCI Machine Learning Repository. <https://doi.org/10.24432/C5Z89R>

## 2. Primary Source:

Chicco, D., & Jurman, G. (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Medical Informatics and Decision Making*, 20, 16. <https://doi.org/10.1186/s12911-020-1023-5>

# Thank You!