SDS 291 Final Project: Topic and Data Selection

Miya Dang, Mia Tran, Alua Birgebayeva

2025-04-28

Data

The dataset we are using for the final project is the Heart Failure Clinical Records Dataset, originally sourced from the UCI Machine Learning Repository (https://archive.ics.uci.edu/dataset/519/heart+failure+clinical+records). It contains the medical records of 299 patients who experienced heart failure, collected during their follow-up period. Each observation is a patient profile, with 13 variables measuring their clinical records such as blood test results, whether or not they have diabetes or hypertension, as well as their age and sex.

Research Question/Purpose

Our project aims to answer the following research question: **Which clinical characteristics are most predictive of death among patients with heart failure?**

We will identify important predictors that may offer insights into patient outcomes after heart failure. Then, using multiple logistic regression, we will model the probability of death during the follow-up period based on clinical data.

Response Variable

Our response variable is DEATH_EVENT, a binary categorical variable with 2 levels:

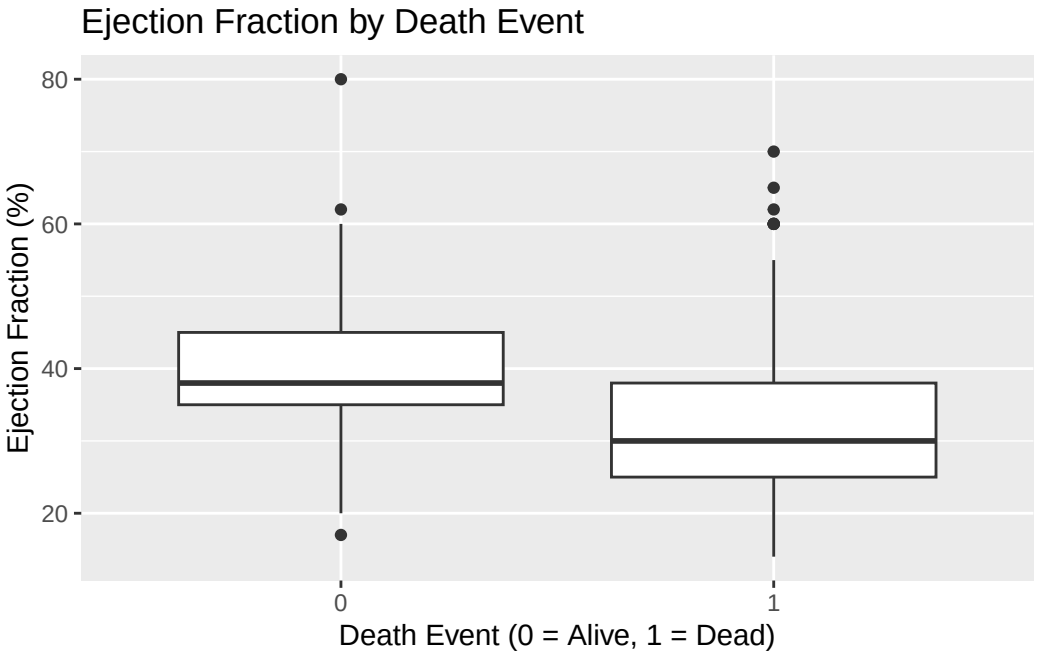
- 0 = Patient survived during the follow-up period (considered "failure" when modelling).
- 1 = Patient died during the follow-up period (considered "success" when modelling).

Explanatory Variables

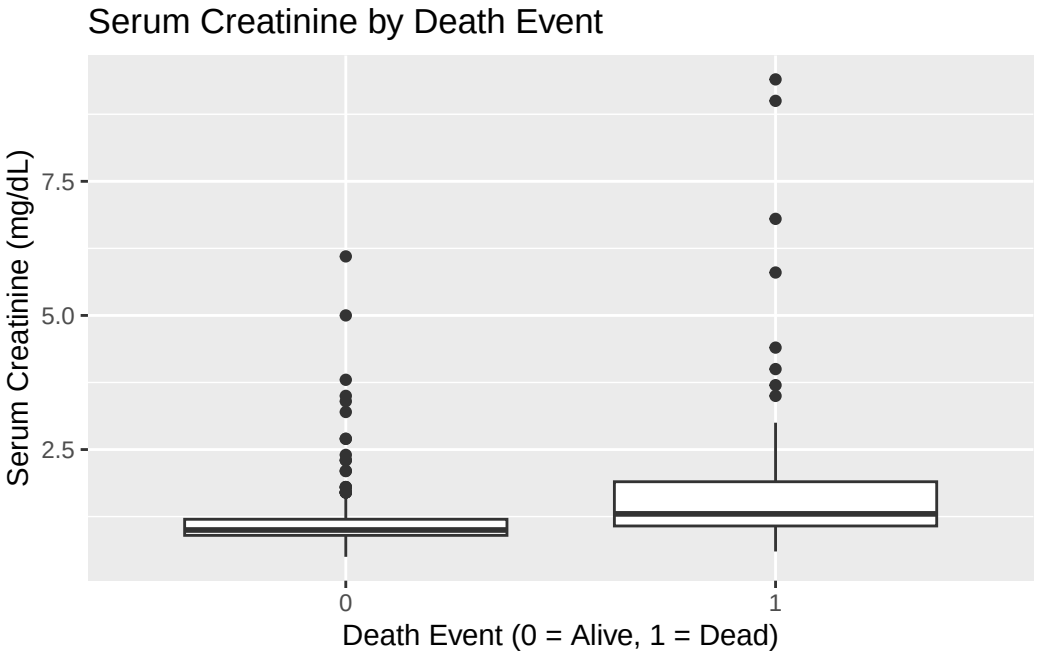
We plan to include all 12 variables other than DEATH_EVENT as explanatory variables to include in the richest possible model .

Variable	Type	Description	Units/Levels
age	Numeric	Age of the patient	Years (40–95)
anaemia	Categorical	Presence of anemia (decrease of red blood cells or hemoglobin)	0 = No, 1 = Yes
creatinine_phosphokinase	Numeric	Level of CPK enzyme in blood (indicator of muscle damage)	mcg/L (23–7861)
diabetes	Categorical	Presence of diabetes	0 = No, 1 = Yes
ejection_fraction	Numeric	Percentage of blood leaving the heart at each contraction	% (14–80)
high_blood_pressure	Categorical	Presence of hypertension (high blood pressure)	0 = No, 1 = Yes
platelets	Numeric	Amount of platelets in the blood	kiloplatelets/mL (~25,000–850,000)
serum_creatinine	Numeric	Level of serum creatinine in blood (kidney function indicator)	mg/dL (0.5–9.4)
serum_sodium	Numeric	Level of serum sodium in blood	mEq/L (113–148)
sex	Categorical	Gender of patient	0 = Woman, 1 = Man
smoking	Categorical	Smoking status of patient	0 = No, 1 = Yes
time	Numeric	Length of follow-up period	Days (4–285)

Exploratory Visualizations



The first boxplot shows that ejection fraction was generally lower among patients who died compared to those who survived, suggesting reduced heart pumping function as a key predictor of mortality.



The second boxplot shows that serum creatinine levels were higher among patients who died, indicating that impaired kidney function is associated with worse outcomes in heart failure patients.