Survival Prediction in Heart Failure Using Machine Learning

(A Data Science Project Report)

# Introduction

Heart failure is a major global health issue, and early risk prediction is crucial to improve patient outcomes. This project focuses on predicting the survival status of patients with heart failure using real-world medical data from 5000 individuals. We formulated this as a binary classification task where the target is whether the patient died during follow-up or not. We explored several machine learning models including Logistic Regression, Decision Tree, and Random Forest. Our goal was not just to achieve high accuracy, but to understand the role of medical factors like age, serum creatinine, blood pressure, and ejection fraction in predicting outcomes. This project involved extensive exploratory data analysis (EDA), statistical testing, and model evaluation using metrics like accuracy, recall, and AUC-ROC to ensure reliable predictions.

# Dataset Overview

The dataset used in this project contains medical records of 5000 patients who were diagnosed with heart failure. Each record includes 12 clinical features and one binary target variable named DEATH\_EVENT, indicating whether the patient died during the follow-up period.

The dataset contains 8 numerical and 5 categorical variables. A brief structure of the dataset is as follows:

* Total Observations: 500
* Target Variable: DEATH\_EVENT (1 = died, 0 = survived)
* Numerical Variables:

1. age – age of the patient in years
2. creatinine\_phosphokinase – leve of creatinine phosphokinase enzyme
3. ejection\_fraction – percentage of blood pumped in each turn.
4. platelets – platelets count of the patient
5. serum\_creatinine – level of serum creatinine enzyme
6. serum\_sodium – level of serum sodium enzyme
7. time – follow up time in days

* Categorical Variables:

1. anaemia – anaemic status (1 = yes, 0 = no)
2. diabetes – diabetic status (1 = yes, 0 = no)
3. high\_blood\_pressure – blood pressure status (high = 1, normal = 0)
4. sex – male = 1, female = 0
5. smoking – smoking status (1 = yes, 0 = no)

# Methodology

To predict patient survival and extract clinical insights, the project followed the steps below:

### Data Understanding & Cleaning

1. Loaded the dataset and verified its structure.
2. Checked for null values, data types.
3. Found the dataset clean and free from missing values.

### Univariate Analysis

1. Explored the distribution of each variable individually using histograms, KDE plots, and boxplots for numerical features.
2. Used count plots for categorical variables.
3. Applied log transformation to skewed variables like CPK and serum creatinine for better modeling performance.

### Bivariate Analysis

Analyzed the relationship between each feature and the target variable (DEATH\_EVENT) using visual tools boxplots, KDE, grouped bar charts.

Generated interpretations that highlighted which variables might be influencing patient outcomes.

### Statistical Testing

Applied Mann-Whitney U test for numerical features and Chi-Square test for categorical ones to check statistical dependence on DEATH\_EVENT.

This helped to confirm whether observed differences were statistically significant or occurred by chance.

### Multivariate Analysis

Checked for multicollinearity using correlation heatmap and Variance Inflation Factor (VIF).

Ensured that predictors provide independent signals to the mode.

### Model Building

Built and evaluated three classification models:

* Logistic Regression (baseline)
* Decision Tree
* Random Forest

Evaluated models using accuracy, precision, recall, F1-score, and AUC.

### Model Comparison & Selection

Compared model performances and selected Random Forest as the best-performing model due to its high accuracy and AUC.

# Data Preparation

Before training models, the dataset was carefully preprocessed to ensure quality and suitability for machine learning. The following key steps were performed:

* Handling Missing Values: No missing values were found in the dataset, so no imputation was needed.
* Outlier Treatment: For skewed numerical features like creatinine\_phosphokinase and serum\_creatinine, log transformation was applied for better understanding, but used original form to preserve clinical interpretability.
* Encoding Categorical Variables: Binary categorical variables (sex, smoking, anaemia, etc.) were already in 0/1 format, so no additional encoding was required.
* Train-Test Split: The dataset was split into training (80%) and testing (20%) subsets using stratified sampling to maintain the class distribution of the target variable (DEATH\_EVENT).
* Scaling: StandardScaler was applied only to numerical features for models like Logistic Regression to ensure fair coefficient estimation. Tree-based models do not require scaling.

# Model Building and Evaluation

To predict patient survival (DEATH\_EVENT), we trained and evaluated three supervised classification models:

### Logistic Regression

* Logistic regression was used as a baseline linear model.
* It achieved a training accuracy of 85.4% and testing accuracy of 82.9%.
* AUC score: 0.90, indicating good class separation.
* Recall (death cases): 0.70, meaning 70% of actual death cases were correctly identified.
* This model was interpretable and performed decently despite class imbalance.

### Random Forest Classifier

* An ensemble-based model that builds multiple decision trees.
* Achieved training accuracy: 99.95% and testing accuracy: 99.1%.
* AUC score: 0.9998, with precision, recall, and F1-scores ~0.99.
* Extremely high performance and generalization with minimal overfitting.
* Best-performing model overall.

### Decision Tree Classifier

A simpler tree-based model with:

* Training Accuracy: 99.95%
* Testing Accuracy: 98.4%
* High precision and recall for both classes.
* Slight risk of overfitting but still delivered strong performance.

# Conclusion and Key Insights

This project focused on predicting patient survival in heart failure cases using machine learning classification techniques. Through detailed univariate, bivariate, and multivariate analysis, we discovered several key risk factors influencing death events — such as age, serum creatinine, ejection fraction, follow-up time, and blood pressure.

Three models were trained and compared. While Logistic Regression offered strong baseline interpretability with decent recall for death prediction, Random Forest stood out with 99.1% accuracy, 0.9998 AUC, and exceptional balance between precision and recall. Decision Tree also performed well but showed slight overfitting.

### Key insights:

* High serum creatinine and low ejection fraction were strong predictors of death, aligning well with medical expectations.
* Patients with shorter follow-up time had significantly higher mortality, making it a sensitive survival indicator.
* Middle-aged patients (55–65) had better survival chances, while deaths were more spread across older age groups.
* Anaemia and high blood pressure showed moderate association with death risk, but diabetes and smoking did not show strong patterns.