Heart Failure Data Analysis: A Clinical Records Study

---

1. Introduction

Heart failure is a critical condition affecting millions globally. In this project, we analyze clinical records of heart failure patients to explore the relationship between patient characteristics and heart failure outcomes. The objective is to better understand the contributing factors to patient mortality using exploratory data analysis (EDA) and basic statistical techniques.

---

2. Objectives

- Conduct a thorough data cleaning and preparation process.
- Perform exploratory data analysis to understand key features and their distributions.
- Identify relationships and correlations between features and heart failure outcomes.
- Use visualizations to better understand these relationships, especially with the target variable, **DEATH_EVENT**.

---

3. Dataset Overview

The dataset consists of 299 patients, each represented by 13 features. These features capture a range of clinical metrics and demographic data, summarized as follows:

- **Age**: Age of the patient (in years).
- **Anaemia**: Whether the patient has anaemia (1 = Yes, 0 = No).
- **Creatinine Phosphokinase (CPK)**: Level of the CPK enzyme in the blood.
- **Diabetes**: Whether the patient has diabetes (1 = Yes, 0 = No).
- **Ejection Fraction**: Percentage of blood leaving the heart at each contraction (heart pumping efficiency).
- **High Blood Pressure**: Whether the patient has high blood pressure (1 = Yes, 0 = No).
- **Platelets**: Platelet count in the blood.
- **Serum Creatinine**: Level of creatinine in the blood, an indicator of kidney function.
- **Serum Sodium**: Level of sodium in the blood.
- **Sex**: Patient's gender (0 = Female, 1 = Male).
- **Smoking**: Whether the patient smokes (1 = Yes, 0 = No).
- **Time**: Duration of patient follow-up (in days).
- **DEATH_EVENT**: Indicates if the patient died during the follow-up period (1 = Yes, 0 = No).

---

4. Data Cleaning and Preparation

Before performing the analysis, several data cleaning steps were conducted:

- **Handling Missing Values**: The dataset was checked for missing or null values. None were found, indicating a complete dataset.
- **Unnecessary Columns**: The `time` column, which represents the follow-up duration, was removed since it is not directly used in our current analysis.
- **Categorical Encoding**: Categorical variables such as `anaemia`, `diabetes`, `high_blood_pressure`, `sex`, `smoking`, and `DEATH_EVENT` were already encoded as 0 (No) and 1 (Yes), so no further encoding was necessary.

---

5. Exploratory Data Analysis (EDA)

5.1 Descriptive Statistics

We began by obtaining summary statistics for the dataset:

- The average age of patients was **60.8 years**, with a minimum age of **40** and a maximum age of **95**.
- The mean **ejection fraction** was around **38%**, indicating that many patients had compromised heart function.
- Patients had a wide range of **creatinine phosphokinase** levels (23 to 7861 U/L), with a median value around 250 U/L.
- **Serum creatinine** levels ranged from **0.5** to **9.4 mg/dL**, with elevated levels often indicating kidney impairment.

5.2 Distribution of Features

Several key distributions were plotted to better understand the data:

- **Age Distribution**: Most patients were in their 60s, with a slight peak around ages 60–70.
- **Ejection Fraction**: The ejection fraction was skewed towards lower values, suggesting that a significant number of patients had poor heart function.
- **Platelets**: Platelet counts were fairly evenly distributed, though there was a concentration around 250,000.

5.3 Gender and Death Event Analysis

We explored how **gender** correlates with the **DEATH_EVENT**:

- **Death Event by Gender**: There was a near-equal split between males and females in the dataset. However, more male patients experienced a **DEATH_EVENT** than females.
- **Death Event Rate**: Out of the 299 patients, **32%** experienced a **DEATH_EVENT**.

---

6. Visualizations

6.1 Pairplot of the Dataset with DEATH_EVENT

A **pairplot** was created to visualize relationships between all the numerical features with the target variable (DEATH_EVENT). It shows how different features (like age, ejection fraction, and serum creatinine) relate to patient outcomes. Features like **ejection fraction** and **serum creatinine** show visible patterns, with patients who died generally having lower ejection fractions and higher serum creatinine levels.

6.2 Bar Plot of Sex Distribution

A bar plot was generated to show the distribution of patients by **sex**:

- **Males** were slightly overrepresented, making up approximately 65% of the dataset.
- This gender imbalance may affect the interpretation of some relationships.

6.3 Death Event by Sex

We visualized the death rate by sex:

- A higher proportion of male patients experienced a **DEATH_EVENT** compared to female patients, which suggests a possible connection between gender and heart failure outcomes.

6.4 Smoking and Death Event

A bar plot comparing **smoking** status with **DEATH_EVENT** showed that smokers had a slightly higher rate of death compared to non-smokers.

6.5 High Blood Pressure and Age

A **catplot** was created to visualize the relationship between **high blood pressure** and **age**. It showed that:

- Older patients (70+ years) tended to have higher rates of high blood pressure.
- Interestingly, younger patients with high blood pressure also had a notable presence in the dataset, which could indicate risk factors independent of age.

---

7. Key Findings

1. **Ejection Fraction**: Patients with lower ejection fractions were more likely to experience a **DEATH_EVENT**. This aligns with medical understanding, where poor heart function is a critical factor in heart failure.
2. **Serum Creatinine**: Higher levels of serum creatinine were associated with increased mortality, indicating that kidney function plays a role in patient outcomes.
3. **Age**: While older patients were more likely to experience adverse outcomes, death was not exclusive to the elderly. Patients across a wide age range (40–95) were affected, making age an important but not definitive factor.
4. **Smoking**: Smokers had a slightly higher mortality rate than non-smokers, suggesting that smoking could be a contributing risk factor for heart failure outcomes.
5. **Gender**: Males had a higher rate of death than females, indicating potential gender differences in heart failure progression or treatment response.

---

8. Conclusion

This analysis explored the relationships between patient characteristics and heart failure outcomes. Key factors like **ejection fraction**, **serum creatinine**, and **age** showed strong associations with the likelihood of death. Insights from the analysis can help guide clinical decision-making, providing early warnings about patients who may be at higher risk of adverse outcomes.

Further work could include deeper statistical analysis and predictive modeling to refine the understanding of how these factors contribute to heart failure mortality. Nevertheless, this initial exploration provides valuable insights into the dataset.