Heart Failure - EDA project Report

Mansi Bansal

1 Introduction

Heart failure is a critical health condition affecting millions worldwide. This report analyzes the **Heart Failure Clinical Records Dataset** to uncover key statistical inferences regarding the relationships between patient characteristics and death events.

2 Dataset Overview

The dataset contains clinical records with the following key features:

- **Age**: Age of the patient.
- **Sex**: Gender of the patient.
- **Anaemia**: Presence of anaemia (Yes/No).
- **Diabetes**: Presence of diabetes (Yes/No).
- **High Blood Pressure**: Presence of high blood pressure (Yes/No).
- **Ejection Fraction**: Percentage of blood leaving the heart each time it contracts.
- **Platelets**: Platelet count.
- **Serum Creatinine**: Levels of creatinine in the blood.
- **Serum Sodium**: Sodium levels in the blood.
- **Smoking**: Smoking status (Yes/No).
- **Death Event**: Indicates whether the patient died (Yes/No).

3 Methodology

3.1 Data Cleaning and Preparation

The dataset was cleaned by converting categorical variables into meaningful labels. Unnecessary columns, such as 'time', were removed. Duplicate entries were checked and handled appropriately.

3.2 Exploratory Data Analysis (EDA)

The analysis utilized various statistical and graphical methods to explore the data:

- **Descriptive Statistics**: Summary statistics provided insights into the central tendency and dispersion of numerical variables.
- **Pairplots**: Visualized relationships between multiple features with respect to death events and sex.
- **Bar Charts**: Displayed frequencies of categorical variables, allowing for quick assessments of distributions.
- **Heatmaps**: Illustrated correlations between categorical variables, particularly between sex and death events.
- **Regression Analysis**: Explored relationships between continuous variables, such as age and platelet counts.

4 Statistical Inferences

4.1 Descriptive Statistics

The average age of patients was approximately **60 years**, suggesting that heart failure primarily affects older adults. The summary statistics indicated varying degrees of health indicators, such as ejection fraction and serum creatinine levels, which could inform potential risk assessments.

4.2 Categorical Variable Analysis

- **Frequencies**:
  - The dataset showed an almost equal distribution of males and females, with slight male predominance.
  - **Death Event Frequencies**: A substantial percentage of patients experienced death events, indicating a high mortality rate associated with heart failure.
- **Risk Factor Analysis**:
  - **Smoking**: A significant correlation was observed where smokers had higher mortality rates than non-smokers, underscoring smoking as a critical risk factor for heart failure.
  - **High Blood Pressure**: The relationship between high blood pressure and age suggested that older individuals are more likely to suffer from high blood pressure, leading to higher death rates.

4.3 Correlation Analysis

The heatmap indicated a correlation between **sex** and **death events**, suggesting that gender may play a role in mortality risk. The analysis of age against platelet counts indicated a potentially nonlinear relationship, warranting further investigation into underlying biological mechanisms.

4.4 Regression Analysis

The regression plots suggested that as age increases, platelet counts may fluctuate, although the relationship was not strictly linear, indicating that other factors may also influence platelet levels.

5 Conclusions

The analysis of the heart failure clinical records dataset leads to several key conclusions:

- **Age as a Significant Risk Factor**: Older age is associated with higher mortality rates, emphasizing the need for targeted interventions in older populations.
- **Impact of Smoking**: Smoking significantly increases the risk of death in heart failure patients, reinforcing the need for smoking cessation initiatives.
- **High Blood Pressure and Diabetes**: These conditions emerge as critical factors affecting patient outcomes, warranting regular screening and management strategies.

6 Recommendations

- **Preventive Healthcare**: Implement screening and preventive measures for high blood pressure and diabetes, particularly in older adults and smokers.
- **Smoking Cessation Programs**: Develop and promote programs aimed at reducing smoking prevalence among heart failure patients.
- **Further Research**: Additional studies should investigate the interplay of various factors influencing heart failure outcomes to develop comprehensive treatment strategies.

7 Limitations

The analysis has several limitations:

- The dataset may not include all relevant variables affecting heart failure outcomes, such as genetic predispositions or treatment histories.
- The cross-sectional nature of the data limits causal inferences.

8 Future Work

Future research could involve:

- Integrating broader datasets that include lifestyle factors and treatment responses.
- Employing advanced statistical models or machine learning techniques to predict outcomes based on a wider range of health indicators.

9 Acknowledgments

The dataset is sourced from Kaggle. Documentation for Pandas, Matplotlib, and Seaborn was referenced throughout the project.