Capstone project-3 Exploratory Data Analysis(EDA) Python

Heart failure data analysis

Introduction:

The Heart Failure dataset contains clinical records of patients, which can be used to predict heart failure events. The dataset includes various features such as age, anaemia, diabetes, high blood pressure, and more.

Dataset link: https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data

Modules used:

1.Numpy-Numerical operations

2.Pandas-Data manipulation & cleaning

3.Matplotlib & Seaborn-Visualization

Data Loading and Initial Exploration:

We start by loading the dataset and examining the first and last few rows to get an initial understanding of the data.

Data Cleaning:

We check for any missing or null values in the dataset. If any are found, we handle them by either dropping the rows or filling them with appropriate values.

Data Mapping:

For better readability, we map the categorical values to more understandable labels. For example, we map `0` and `1` to `No` and `Yes` for features like anaemia, diabetes, high blood pressure, smoking, and death event. Similarly, we map `0` and `1` to `Female` and `Male` for the sex feature.

Dataset Information:

We use the `info()` method to get a concise summary of the dataset, including the number of non-null entries and the data types of each column.

Removing Unneeded Data:

We remove the `time` column as it is not needed for our analysis.

Descriptive Statistics:

We use the `describe()` method to get a summary of the central tendency, dispersion, and shape of the dataset’s distribution.

Dataset Shape:

We check the shape of the dataset to understand the number of rows and columns.

Value Counts:

We use the `value_counts()` method to find the distribution of values for specific columns such as sex, high blood pressure, diabetes, smoking, and death event.

Visualizations:

Modules used- Matplotlib & Seaborn

- **Age Distribution**: The age distribution plot shows the spread of ages among the patients, helping us understand the age range most affected by heart failure.
- **Gender Distribution**: The gender distribution plot reveals the proportion of male and female patients.
- **High Blood Pressure, Diabetes, and Smoking Distributions**: These plots show the prevalence of high blood pressure, diabetes, and smoking among the patients.
- **Death Event Distribution**: This plot indicates the number of patients who experienced a death event.
- **Correlation Heatmap**: The heatmap helps identify the relationships between different features, highlighting which factors are more correlated with heart failure.
- **Box Plot of Age vs. Death Event**: This plot shows the age distribution for patients who experienced a death event versus those who did not.
- **Pair Plot**: The pair plot provides a comprehensive view of the relationships between multiple features, colored by the death event.

Insights:

The exploratory data analysis (EDA) of the Heart Failure dataset provides several key insights:

1. **Age Distribution**: The majority of patients fall within a specific age range, indicating that heart failure is more prevalent among certain age groups. This can help target preventive measures and treatments more effectively.
2. **Gender Distribution**: The dataset shows a higher number of male patients compared to female patients. This could suggest a gender disparity in heart failure cases, which might warrant further investigation.
3. **High Blood Pressure, Diabetes, and Smoking**: A significant portion of the patients have high blood pressure, diabetes, or are smokers. These factors are known risk factors for heart failure, highlighting the importance of managing these conditions to prevent heart failure.

4. **Death Event Distribution**: The distribution of death events provides a stark reminder of the severity of heart failure. Understanding the factors that contribute to these events can help in developing better treatment protocols.
5. **Correlation Heatmap**: The heatmap reveals strong correlations between certain features, such as age and death event, which can be crucial for predictive modeling. Identifying these relationships helps in understanding the underlying factors contributing to heart failure.
6. **Box Plot of Age vs. Death Event**: The box plot shows that older patients are more likely to experience a death event, emphasizing the need for targeted interventions for older age groups.
7. **Pair Plot**: The pair plot provides a comprehensive view of the relationships between multiple features, helping to identify patterns and potential predictors of heart failure.

Conclusion:

The analysis highlights the importance of age, gender, high blood pressure, diabetes, and smoking in understanding heart failure. These insights can guide healthcare professionals in developing targeted prevention and treatment strategies. By focusing on these key factors, we can improve patient outcomes and reduce the incidence of heart failure.