Intermediate Project Report: Heart Failure Analysis

GitHub link: https://github.com/jwilliams2023/heart-failure-dataset.git

1. Introduction & Objective

Heart failure poses a significant public health challenge, contributing to high hospitalization rates and mortality worldwide. Detecting at risk patients early can lead to better clinical decision-making and improved treatment outcomes.

The project examines clinical data from heart failure patients to determine the most influential factors tied to mortality. The study also aims to produce predictive models that classify patient outcomes based on these key features. The objective of this study is to analyze clinical features of patients with heart failure and develop predictive models to classify the risk of mortality (DEATH_EVENT). The analysis includes data preprocessing, exploratory data analysis, EDA, and model evaluation.

Objectives:

- Identify the most significant clinical variables associated with mortality
- Analyze clinical features of patients with data preprocessing, exploratory data analysis (EDA)
- Develop classification models, such as logistic regression, LDA, and decision trees, to predict mortality risk

- Evaluate and compare model effectiveness based on accuracy, precision, recall, and F1-score, along with confusion matrices to assess classification performance and a feature importance analysis to understand the impact of different clinical variables on heart failure mortality

2. Data and Exploratory Data Analysis (EDA)

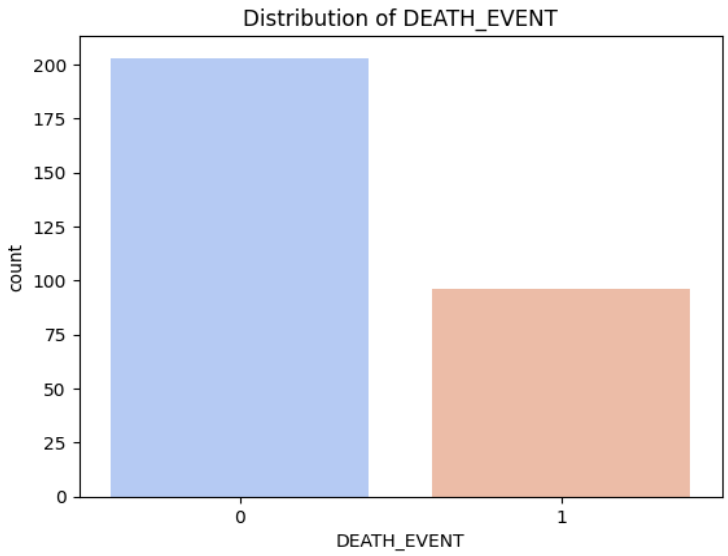
The dataset contains various clinical measurements such as age, ejection fraction, serum creatinine, serum sodium, and time. The target variable is DEATH_EVENT, where 1 indicates mortality and 0 indicates survival.

3. Methodology

Data Preprocessing

- The dataset was loaded using the ucimlrepo package.
- No missing values detected.

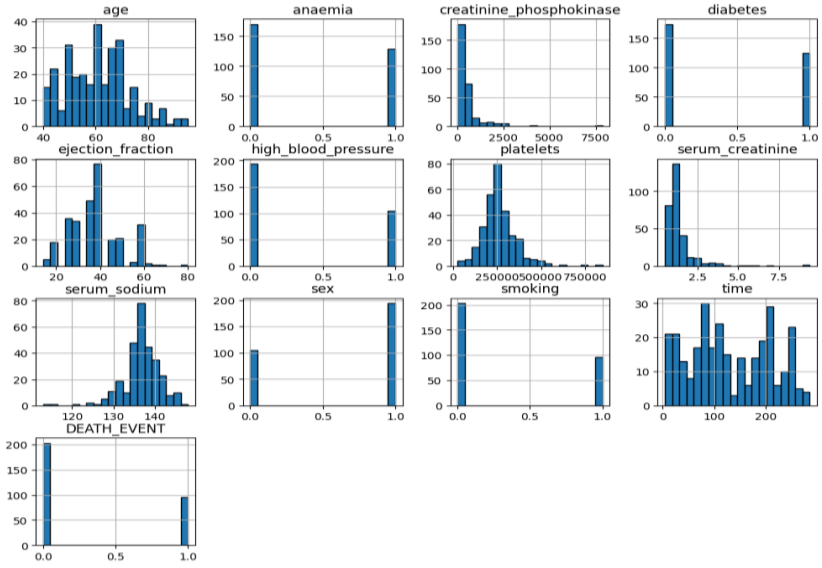
Figure 1: Class distribution of target variable



Interpretation:

- Most patients in the dataset survived, indicating a class imbalance in the target variable.

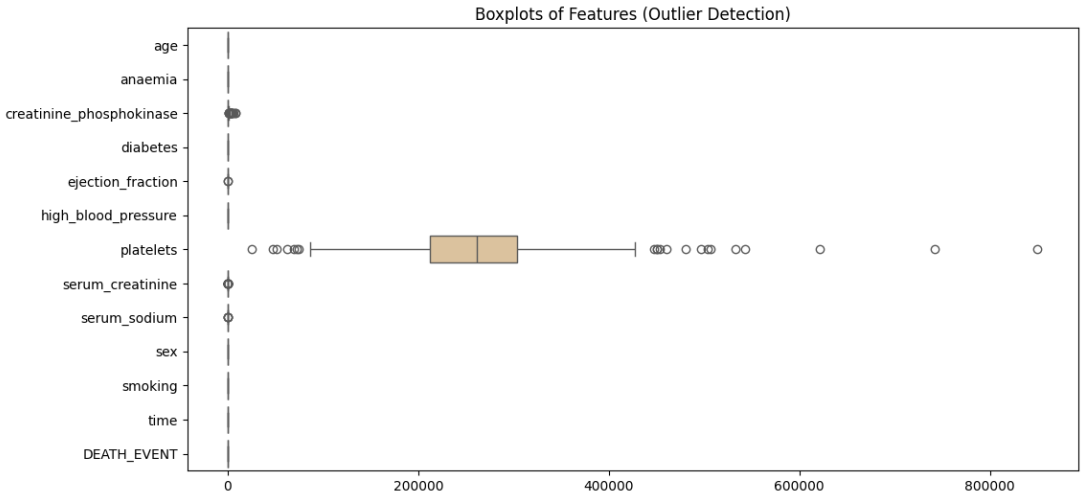
Figure 2: Histograms for feature distributions



Interpretation:

- Several features like serum creatinine and creatinine phosphokinase are right skewed, while variables such as age, ejection fraction, and platelets show more normal distributions.
- Binary features including anemia, diabetes, smoking, and sex are clearly separated between 0 and 1, reflecting binary nature.

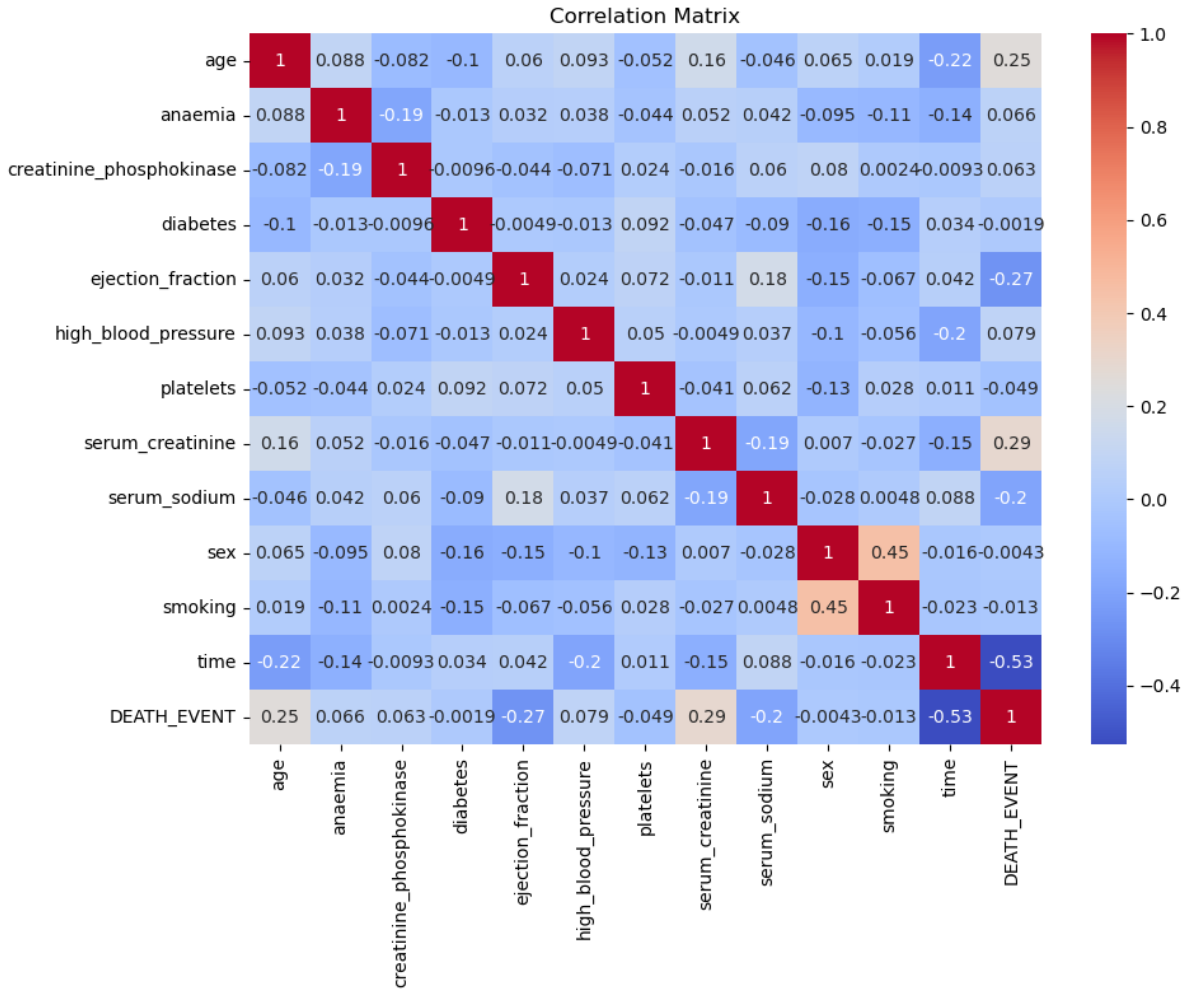
Figure 3: Boxplot of features for outlier detection



Interpretation:

- Several features, especially creatinine phosphokinase and platelets, have extreme outliers, which could impact model performance.

Figure 4: Correlation matrix showing relationships between the variables and mortality.



Interpretation:

- Time at -0.53 and ejection fraction at -0.27 have the strongest negative correlation with death, meaning longer follow ups and higher ejection fractions are associated with survival.
- Serum creatinine at 0.29 and age at 0.25 have the strongest positive correlation with death, indicating that higher creatinine levels and older age increase the risk of mortality.
- Most other features show weak correlations, suggesting they may have less predictive power.

Figure 5: Boxplot of age vs. DEATH_EVENT

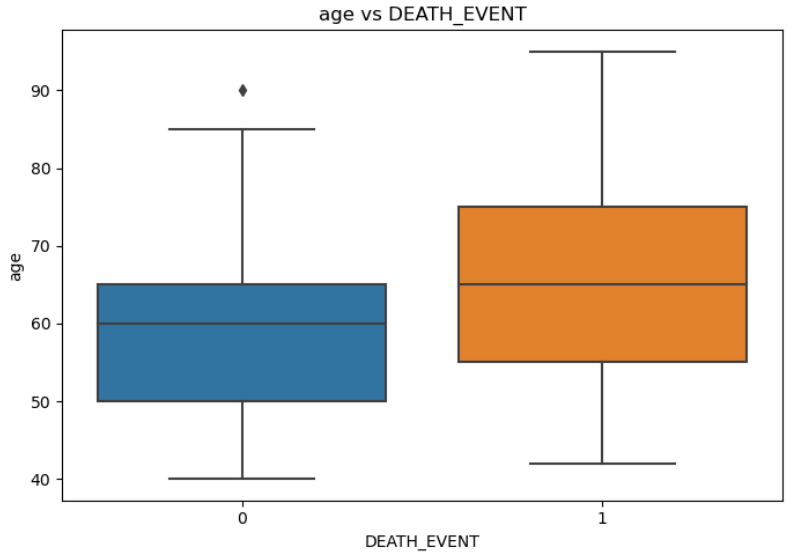


Figure 6: Boxplot of ejection fraction vs. DEATH_EVENT

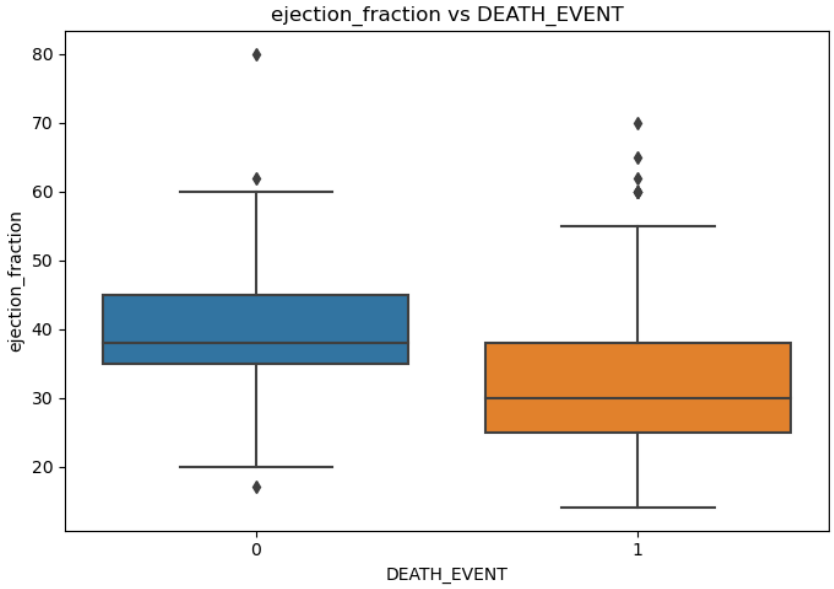


Figure 7: Boxplot of serum creatinine vs. DEATH_EVENT

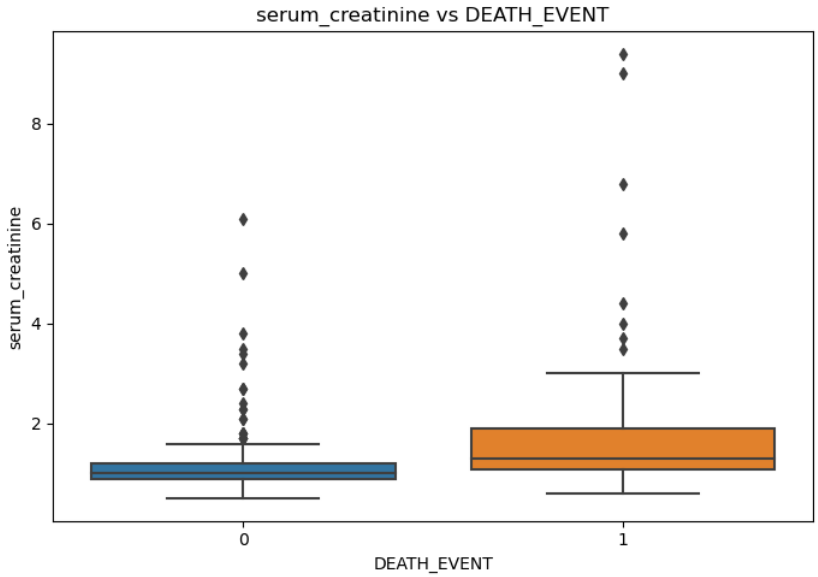


Figure 8: Boxplot of serum sodium vs. DEATH_EVENT

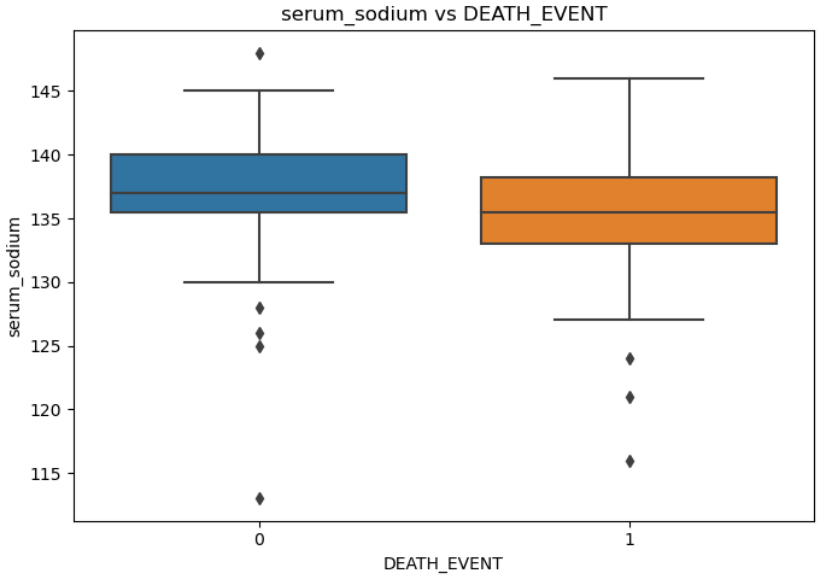
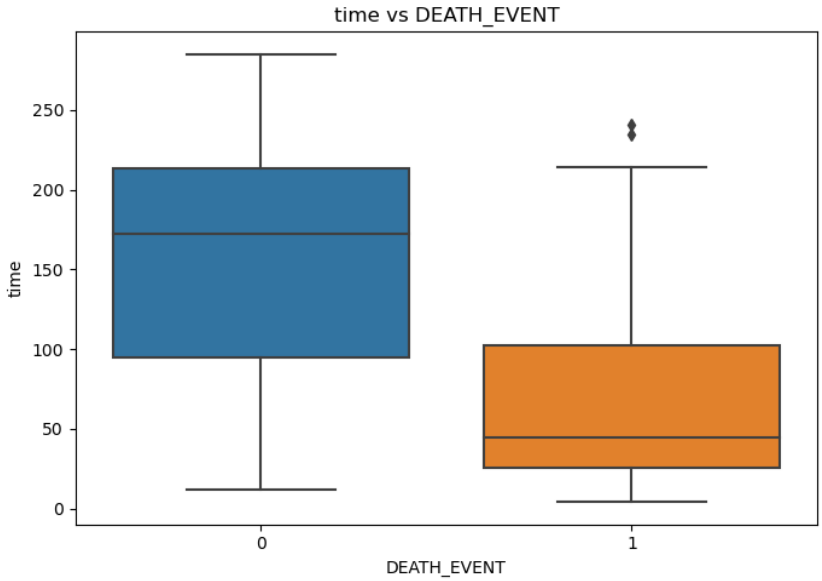


Figure 9: Boxplot of time vs. DEATH_EVENT



- Data was split into training and testing sets (80-20 split).

Model Selection

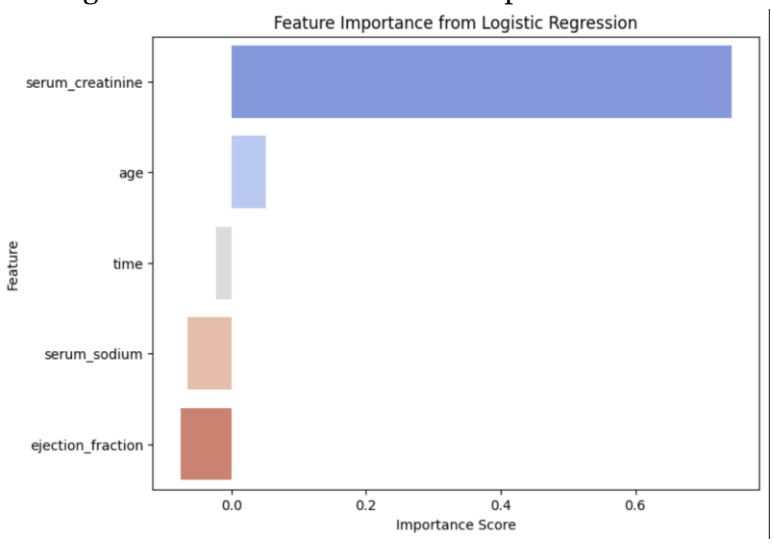
Three classification models were implemented:

- **Logistic Regression**
- **Linear Discriminant Analysis (LDA)**
- **Decision Tree**

Feature Importance

Positive values indicate a positive association with mortality, while negative values suggest a protective effect.

Figure 10: Bar Chart of Feature Importance Scores



Interpretation:

- **Serum Creatinine** has the strongest positive association, meaning higher values significantly increase mortality risk.
- **Time** has a minimal negative impact, indicating that patients who survived had slightly longer follow-up periods.
- **Ejection Fraction** also negatively correlates heavily, meaning lower values contribute to increased mortality risk.

4. Modeling Approaches and Results

- **Logistic Regression** showed the highest accuracy but lower recall, indicating it is good at predicting survival but may miss some mortality cases.
- **LDA** performed similarly to logistic regression with slightly lower accuracy.
- **Decision Tree** had the lowest performance.

Figure 11: Bar chart of model accuracy scores

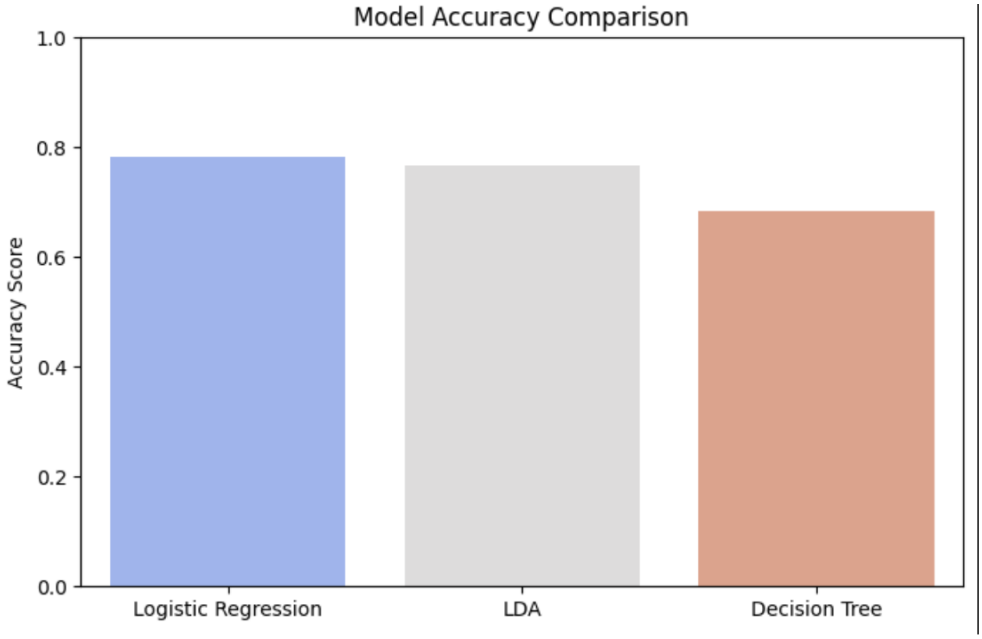


Figure 12: Confusion matrices for each model

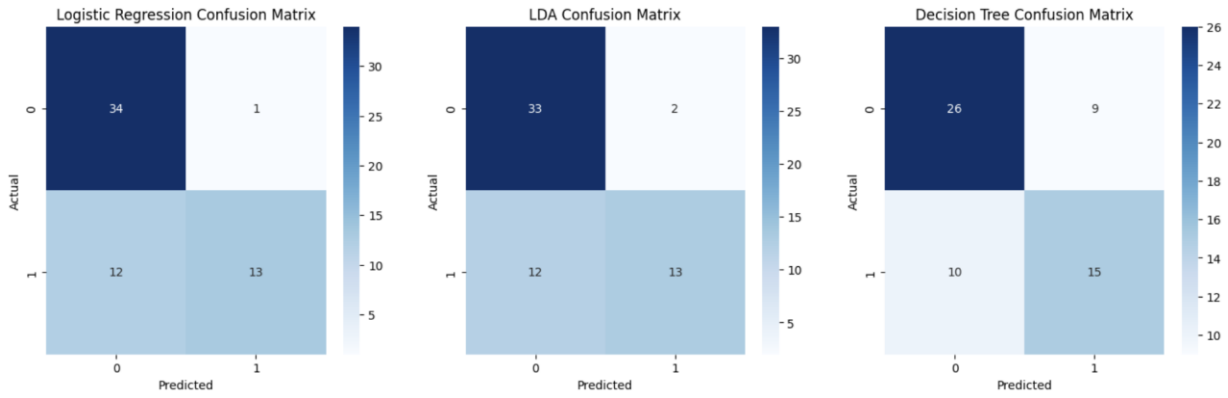
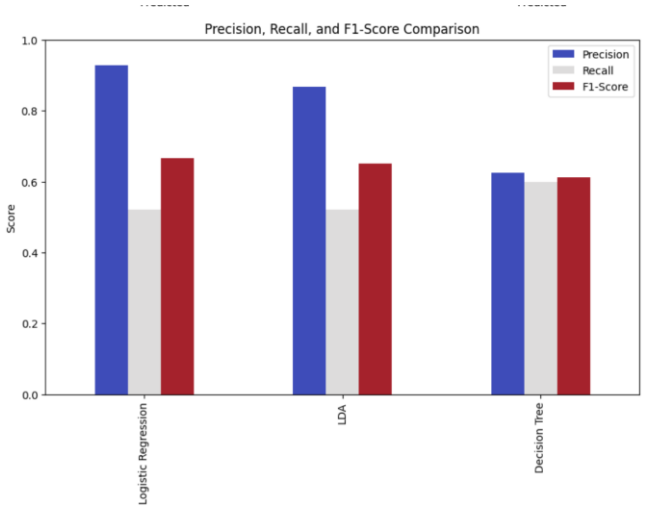


Figure 12: Bar chart of model precision, recall, and F1-scores



Interpretation:

- **Logistic Regression** achieved the highest accuracy and precision but had lower recall, meaning it predicts survival well but misses some mortality cases.
- **Linear Discriminant Analysis (LDA)**: Performed similarly to Logistic Regression with slightly lower accuracy and recall, making it a strong alternative.
- **Decision Tree**: Had the lowest accuracy and higher misclassifications, likely due to overfitting, making it the weakest model for this dataset.

5. Conclusions

This analysis highlights the importance of the features serum creatinine and ejection fraction in mortality rates, with higher creatinine levels increasing mortality risk and lower ejection fraction values correlating with worse outcomes. Among the models, logistic regression performed the best, balancing accuracy and precision, while LDA was a strong alternative. The Decision Tree model struggled with overfitting and had lower predictive power.