- ***Project Background***

The "Heart Failure Prediction" is a dataset from the Kaggle website that contains 299 observations (rows) and 13 attributes (columns) containing physical, medical, and lifestyle data collected at the Faisalabad Institute of Cardiology and the Allied Hospital in Faisalabad (Punjab, Pakistan), during April–December 2015. All the patients were in classes 3 or 4 of the stages of heart failure based on the New York Heart Association (NYHA) classification. On February 3rd, 2020, BMC Medical Informatics and Decision Making, a division of Springer Nature, published this dataset. Springer Nature publishes reputable research to promote open science and the development of new ideas. Their mission is to accelerate solutions to address the world's urgent challenges. Our clients are medical doctors and physicians who want to use our survival model to help their heart failure patients survive based on their clinical records.

- ***Problem Formulation***

Questions/Objectives:

1. What are the most significant attributes of heart failure-related mortality?
2. Will building a mortality prediction model based on the most important features help patients to survive?
3. Does gender make differences in features' importance and the mortality prediction model?

- ***Data Strategy Plan***

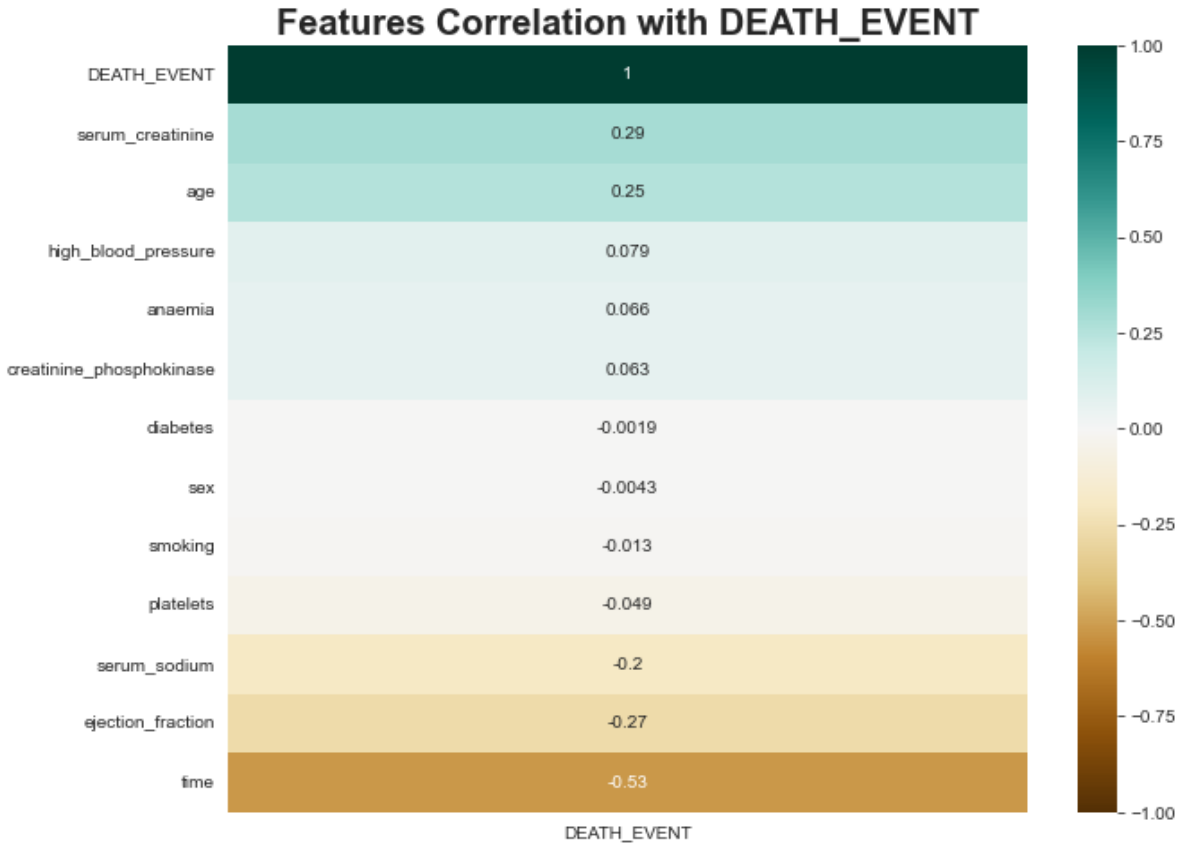
A brief explanation of the features is here: Anaemia (decreases of red blood cells or hemoglobin), high blood pressure (if the patient has hypertension), diabetes (if the patient has diabetes), smoking (if the patient smokes) (0: No, 1: Yes), and sex (0: women, 1: men) features are binary. The following numbers describe things in the blood: age (between 40 and 95 years old), ejection_fraction (percentage of blood leaving), platelets (number of platelets in the blood), serum_ceatinine (amount of creatinine in the blood), serum_sodium (amount of sodium in the blood), ejection_fraction, platelets, platelets, age, and time. The DEATH_EVENT feature states that if the patient died or survived before the end of the follow-up period, it was 130 days on average. It should be our target variable for our analysis. Among these 13 features, 8 are the patients' clinical reports or health statuses, which seem to have more of an effect on death or prevent it.

Summary of Data Cleansing and Exploratory Data Analysis

- ***Data Cleansing***

- There are no missing (null or error) observations
- No duplications
- There are outliers in some of the clinical features: creatinine phosphokinase, platelets, serum creatinine, and ejection fraction. We decided to keep those outliers as these parameters are for patients with heart failure, and we assumed having such a high level is possible.

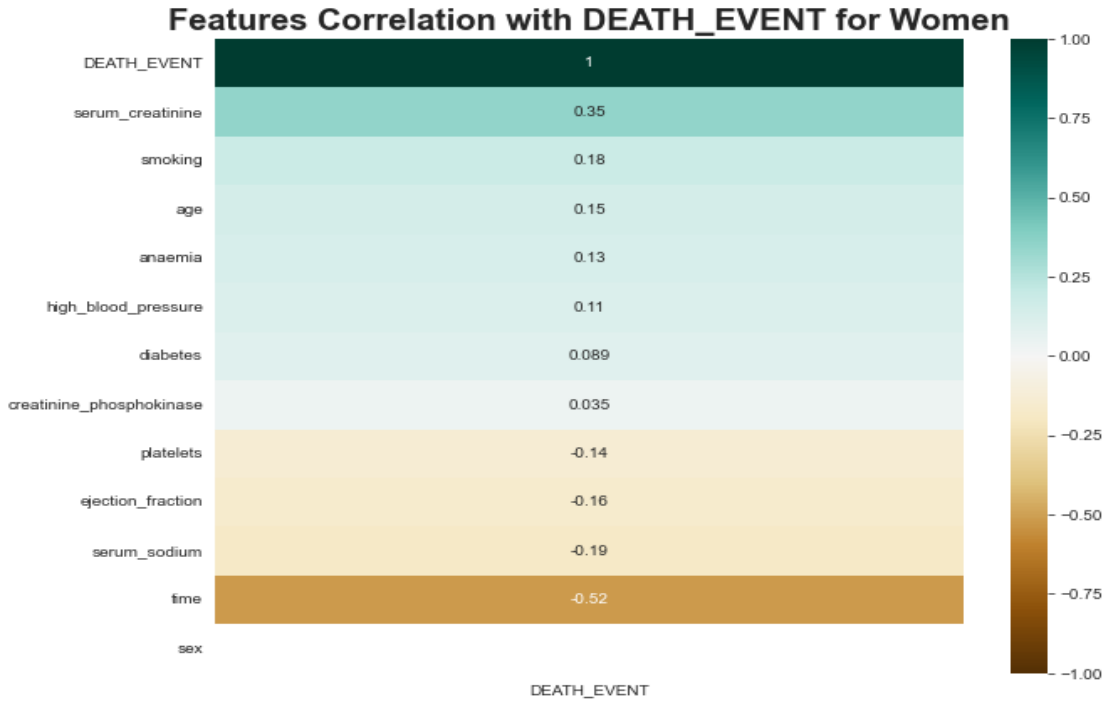
- ***Data Exploratory***



The top 5 correlated features (no matter positive or negative):

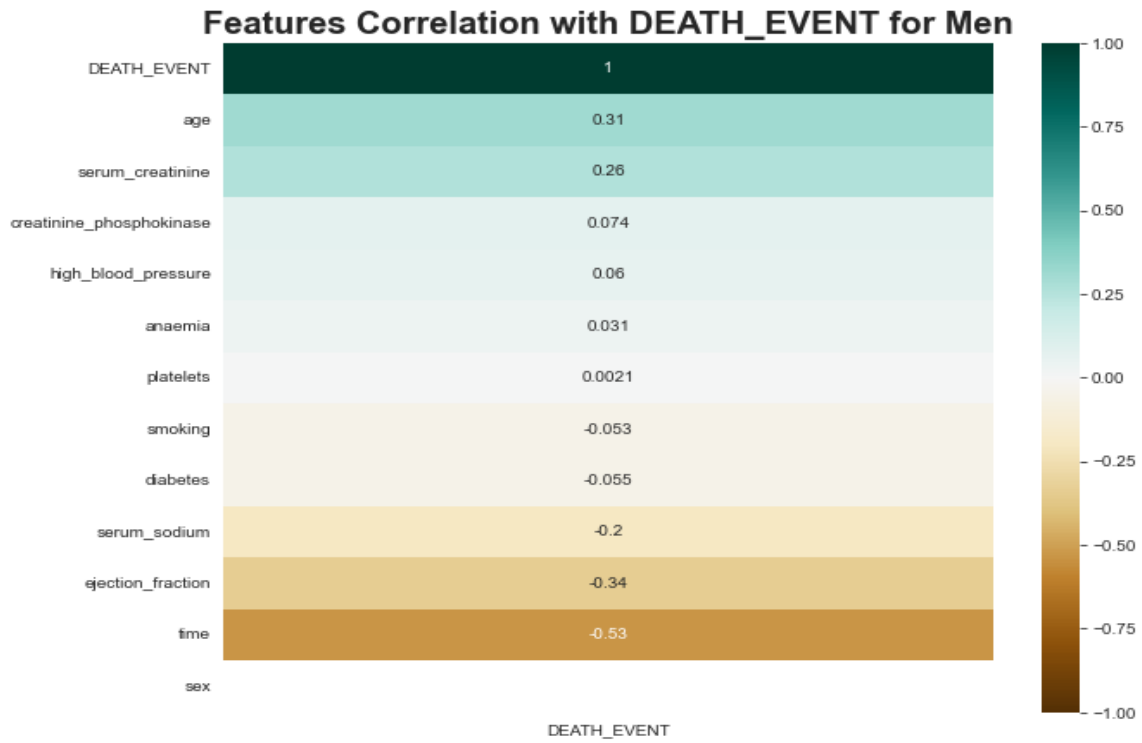
- Time
- Serum_creatinine
- Ejection_fraction
- Age
- Serum_sodium

An interesting thing here is the feature's correlation with DEATH_EVENT changes for women and men.



The top 5 correlated features for women (no matter positive or negative):

- Time
- Serum_creatinine
- Serum_sodium
- Smoking
- Ejection_fraction



The top 5 correlated features for men (no matter positive or negative):

- Time
- Ejection_fraction
- Age
- Serum_creatinine
- Serum_sodium
- Serum_phosphokinase

Interesting insight from heatmaps:

- **Smoking** has a higher positive correlation with death among women than men:
    - 75% of smoked women died
    - 27.6% of smoked men died
- **Time** is the most negatively correlated feature with death
    - 51% of fatalities happened before 50 follow-up days for all patients

Percentage of patients who died with the abnormality level of important features:

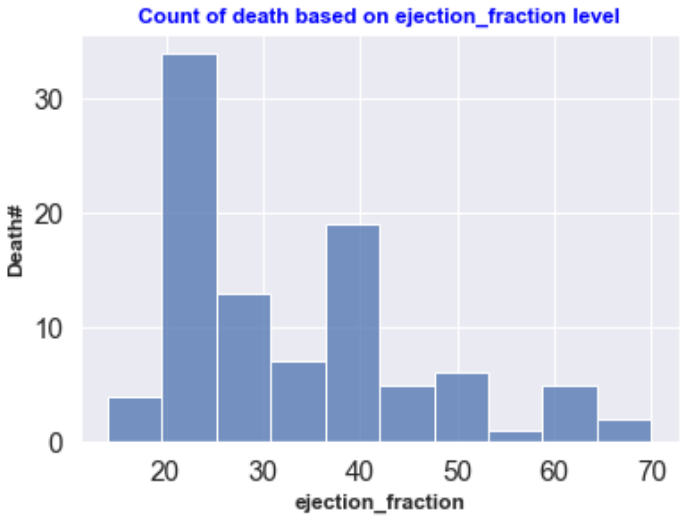
Ejection_fraction: 85.4% (the most important feature after time). This makes sense in the real world since the heart is not capable of pumping more blood to other body parts at a lower ejection fraction.

Creatinine_phosphokinae: 80%

Serum_creatinine: 70.6% for women and 30.65% for men (more correlated feature with death among women than men)

Serum_sodium**:** 43.75%

Platelets: 16.67%



Normal range for ejection fraction:

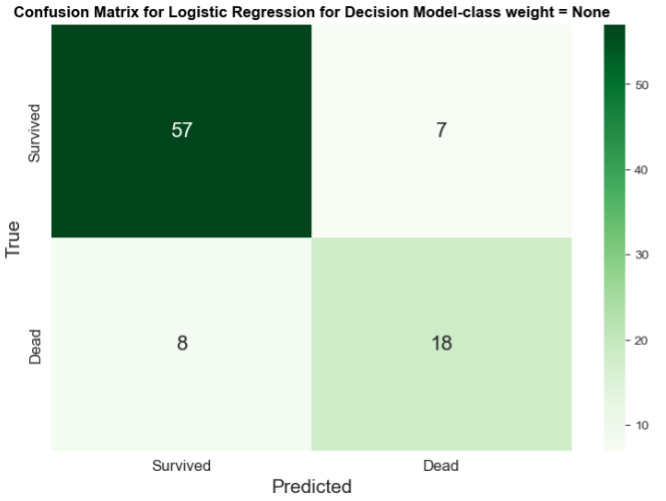
50%~75%

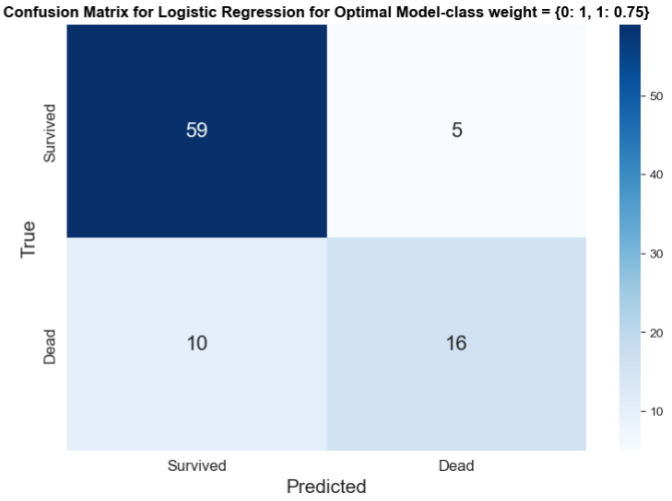
- ***Analysis and Business insight***

The methods used here to analyze the Heart Failure Prediction Dataset are Logistic Regression, Decision Tree, and Random Forest classifiers.

1- **Logistic Regression**

The target (dependent) variable is DEATH_EVENT, which is a binary variable with a value of 1 when death happens and 0 when it doesn't happen. The logistic regression machine learning model predicts





whether the patient with heart failure dies or survives, followed by understanding the influence of significant factors that truly affect the patient. All methods start by setting the data into X and y datasets, then splitting the dataset into training and test data, creating the model, fitting the model, training the model, and creating a prediction. The prediction model here is a mortality prediction that wants to predict mortality with the minimum faults. Logistic regression uses scaled data as it works better than raw data. However, scaled data doesn't work well with decision trees and random forest classifiers.

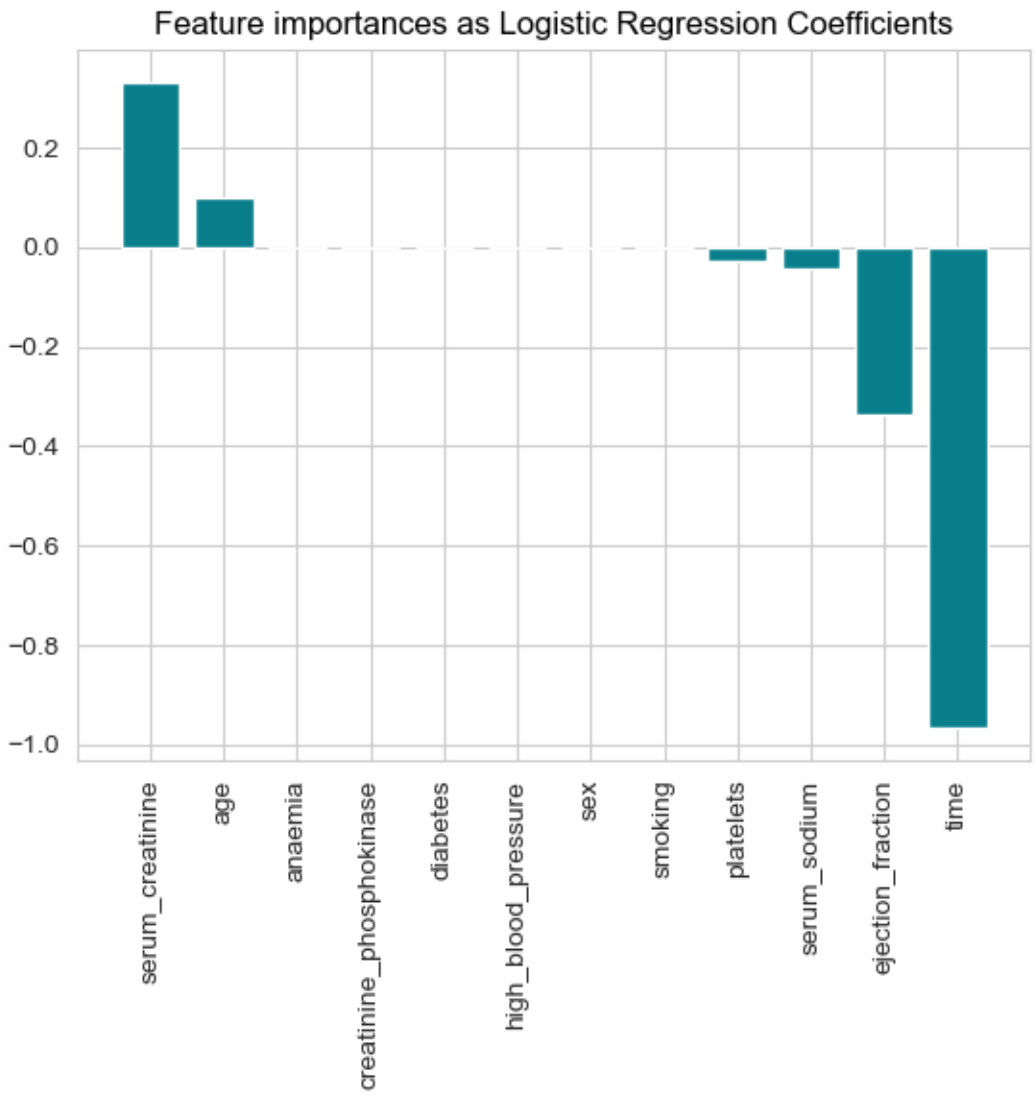
Accuracy Score: 0.8333
Recall Score: 0.6923
Precision Score: 0.7200
ROC AUC Score: 0.8636

Accuracy Score: 0.8333
Recall Score: 0.6154
Precision Score: 0.7619
ROC AUC Score: 0.8642

Decision model works better for the expected prediction. From the confusion matrix, with this model, we predict 18 times correctly for dead and just 8 times wrong out of all 26 deaths, while the model predicts 57 times right and 7 times wrong among 64 survived. Then:

Mortality Prediction Model Using Logistic Regression:

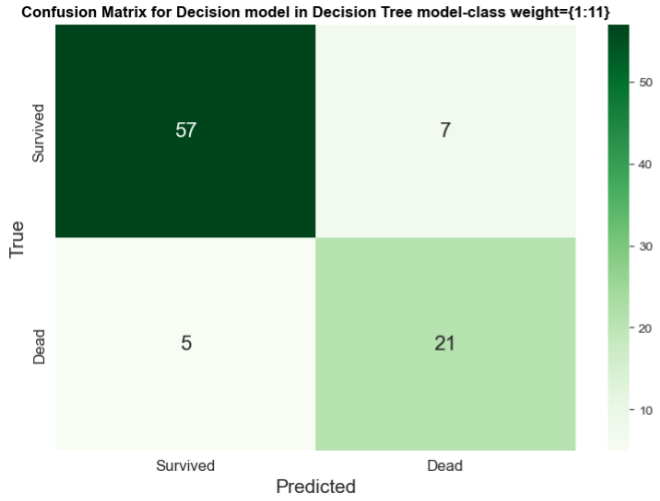
- Is 83% accurate in classifying the patients as dead or survived (Accuracy score 0.833)
- Predicted 69% correctly for predicting mortality (Recall Score of 0.692)
- Out of all mortality predictions, 72% is genuine mortality (Precision Score of 0.72)
- is performing 86.36% well (ROC AUC Score of 0.8636)

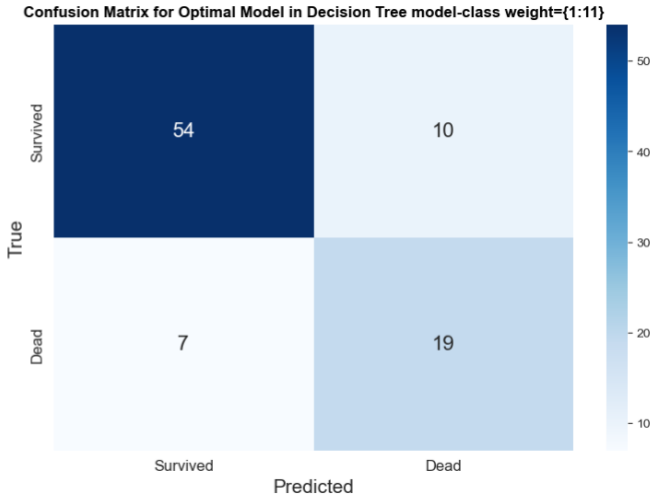


The larger the coefficient (in both positive and negative directions), the more influence it has on a prediction.

2- Decision Tree classifier

Decision trees are a popular method for predictive modeling because they are effective and easy to understand. The model wants to make a good mortality prediction but not a lousy survival prediction. That's why we chose the decision model on the left from below.



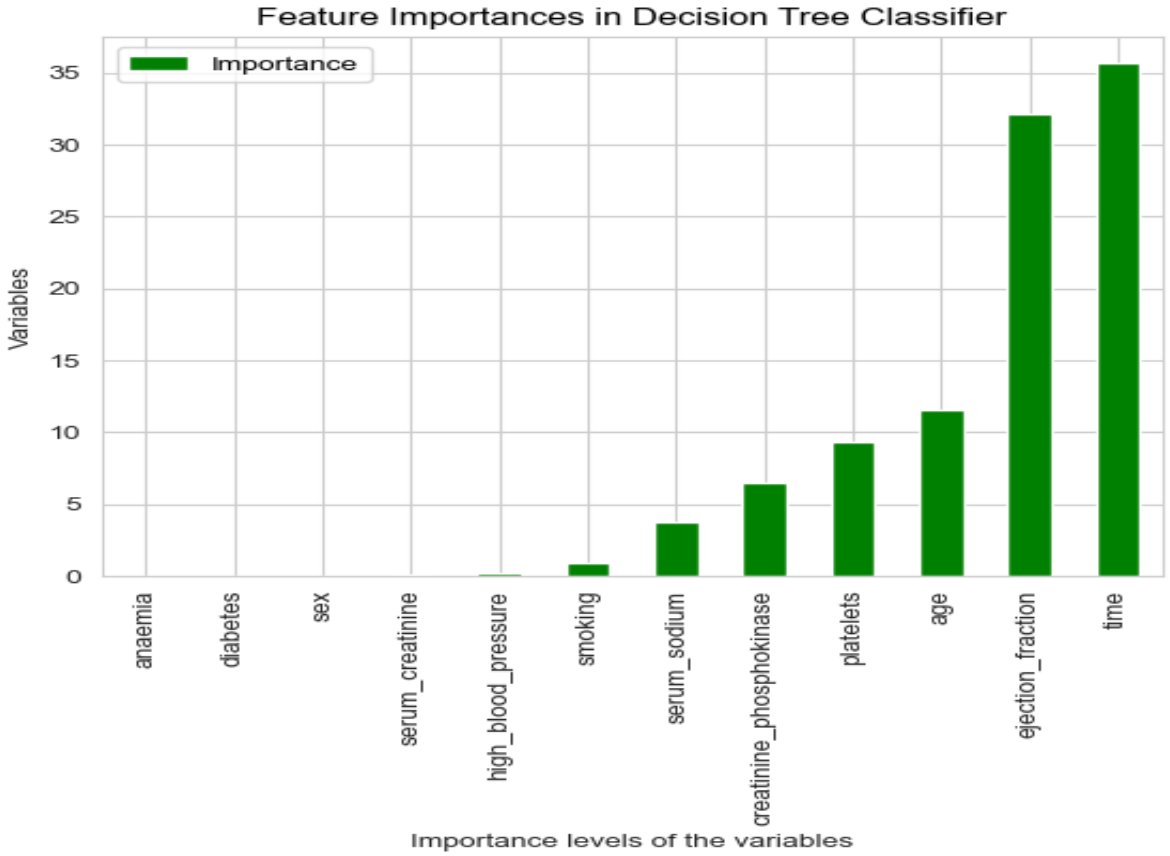


Accuracy Score: 0.8667
Recall Score: 0.8077
Precision Score: 0.7500
ROC AUC Score: 0.8338

Accuracy Score: 0.8111
Recall Score: 0.7380
Precision Score: 0.6552
ROC AUC Score: 0.7728

Mortality Prediction Model with Decision Tree:

- Is 86.7% accurate to classifying the patients as dead or survived (**Accuracy Score** of 0.867)
- Predicted 80.8% correctly for predicting mortality (**Recall Score** of 0.808)
- Out of all mortality prediction 75% is truly mortality (**Precision Score** of 0.75)
- is performing 83.4% well (**ROC AUC Score** of 0.834)

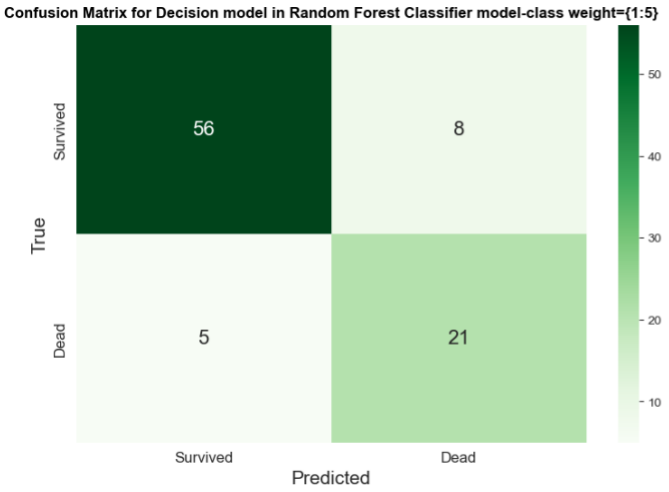


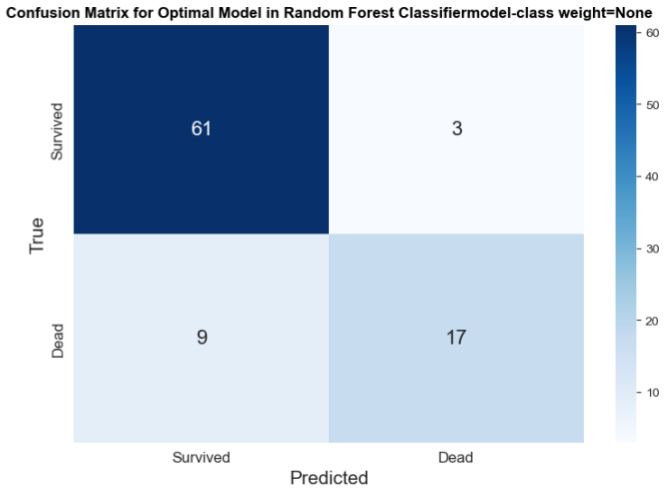
The above table shows the importance levels of the features, which are time, platelets, creatinine_phosphokinase, ejection_fraction, serume_creatinine, age, and serum_sodium, respectively. Compared to the logistic regression feature importance, there is one more feature that has an influence in death, which is creatinine_phosphokinase.

Note: The dataset contains many outliers for creatinine_phosphokinase (CPK). Total CPK normal values are 10 to 120 micrograms per liter (mcg/L). About 74.2% of all patients in the dataset have a higher CPK level than normal. A higher level of CPK is a crucial factor in heart attacks. 80% of all dead patients had a higher CPK level than normal.

3- Random Forest Classifier

It can perform both regression and classification tasks. A random forest produces good predictions that can be understood easily.

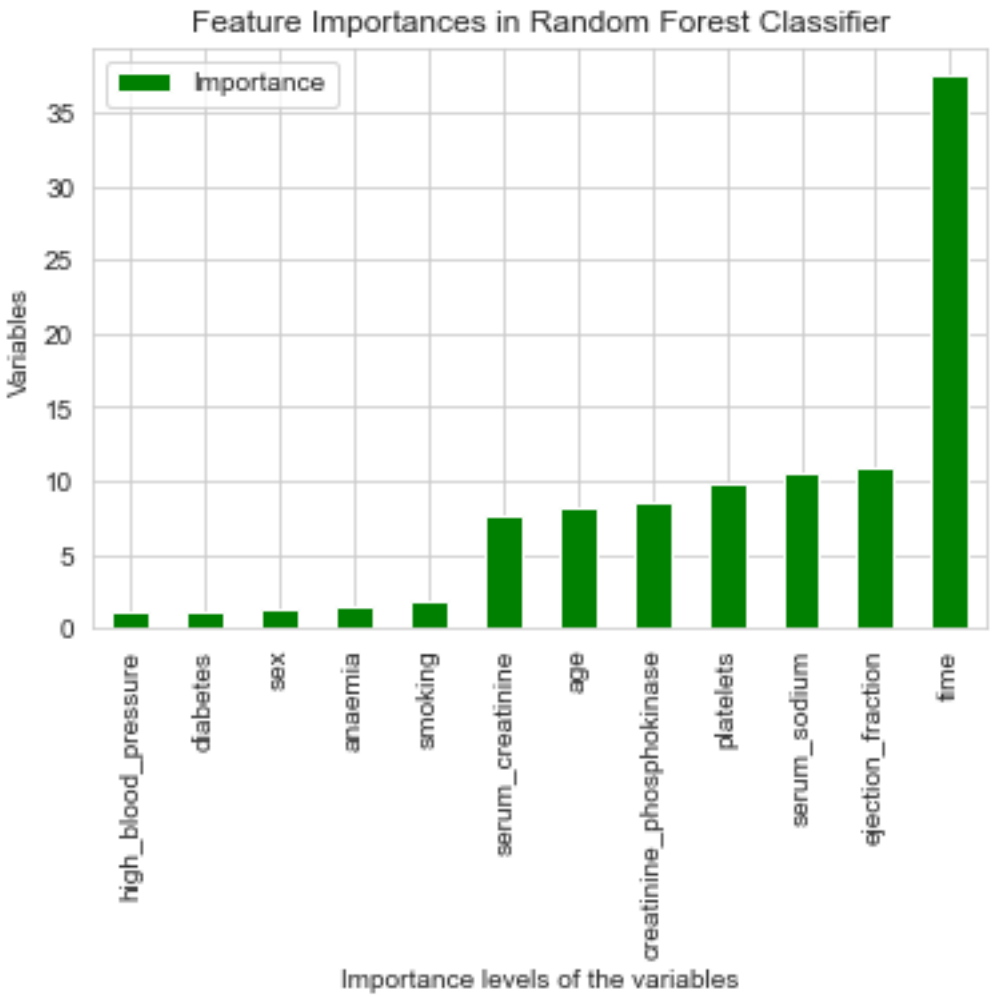




Accuracy Score: 0.8556
Recall Score: 0.8077
Precision Score: 0.7241
ROC AUC Score: 0.8960

Accuracy Score: 0.8667
Recall Score: 0.6538
Precision Score: 0.8500
ROC AUC Score: 0.9008

Between Optimal and decision models, we chose the decision model by sacrificing a little bit of accuracy score to get a higher recall score, which resulted in better mortality prediction.



Mortality Prediction Model Random Forest:

- Is 85.6% accurate in classifying the patients as dead or survived (Accuracy Score of 0.856)
- Predicted 80.8% correctly for predicting mortality (Recall Score of 0.808)
- Out of all mortality predictions, 72.4% is truly mortality (Precision Score of 0.724)
- is performing 89.6% well (ROC AUC Score of 0.896)

The above chart shows all the features have an influence on death, but 5 features, such as smoking, anemia, sex, diabetes, and high pressure, have less impact than the other 7 features.

Objective for Classification Modeling

- Finding the best and most accurate model for our mission
- Minimize the number of wrong predictions on survived (FN) (As not wanting to miss patients who need help)
- Maximize the number of correct predictions on dead (TP)
- Minimize the number of wrong predictions on dead (FP) (As not wanting to increase patients 'stress

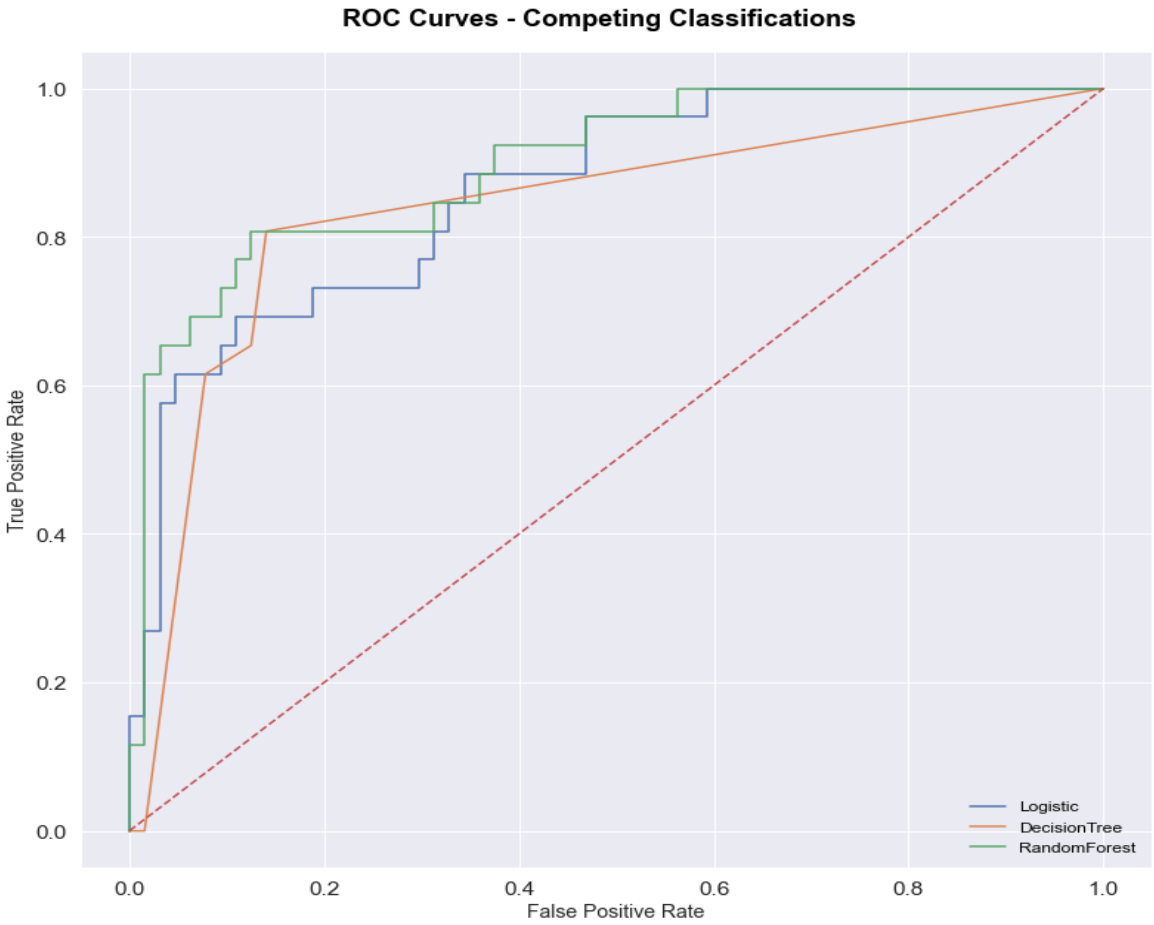
Conclusion:

The best predictor is random forest, as it works best out of the other classifiers, with an average recall of 81% and an average precision of 72%. It correctly predicts about 86% mortality and 80% survival, which is better than the other models. Its F1 score is 0.76, which is higher than the other 2 models. This is concluded from the table below.

actual	Model	pred_survived	pred_dead	Score	Recall	Precision	ROC AUC	F1
Survived	Logistic	57	7	0.8333	0.6923	0.7200	0.8960	0.7059
Dead	Logistic	8	18	0.8333	0.6923	0.7200	0.8960	0.7059
Survived	DecisionTree	55	9	0.8444	0.8077	0.7000	0.8636	0.7500
Dead	DecisionTree	5	21	0.8444	0.8077	0.7000	0.8636	0.7500
Survived	RandomForest	56	8	0.8556	0.8077	0.7241	0.8371	0.7636
Dead	RandomForest	5	21	0.8556	0.8077	0.7241	0.8371	0.7636

After applying Logistic Regression, Decision Tree, and Random Forest methods and analyzing the dataset based on the objectives, the project ended up with a mortality prediction model using Random Forest.

Also, by using the Sklearn method of roc_curve() and computing the roc auc score, we can plot the ROC curves for the 3 algorithms. The plot shows that the AUC for the Random Forest ROC curve is higher than that for the Decision Tree and Logistic curve. Therefore, Random Forest did better in classifying the positive class in the dataset.



Final Summary:

Doctors can use the Random Forest mortality prediction model for their patients to help them by:

- Increasing/improvingejection_fraction,
- reducing the level of creatinine_phosphokinaseand serume_creatinine
- increasing serum_sodiumlevel
- normalize platelets
- asked smoked women patients to quit smoking as 75% of smoked women (4 out of 3) died.
- start treatments to normalize the level of the above factors as soon as possible to survive their patients (as the follow-up period is the most negatively correlated feature).