- ***Project Background***

The “Heart Failure Prediction” is a dataset from Kaggle website that contains 299 observations (rows) and 13 attributes (columns) containing physical, medical, and lifestyle which collected at the Faisalabad Institute of Cardiology and at 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 New York Heart Association (NYHA) classification. This dataset published on February 3rd, 2020, by *BMC Medical Informatics and Decision Making*, which is part of Springer Nature, that publishes trusted research to support the development of new ideas and champion open science. Their mission is accelerating solutions to address the world’s urgent challenges. Our client are medical doctors and physicians who want to use our survival model to help their heart failure patients to survive based on their clinical records.

- ***Problem Formulation***

Questions/Objectives:

1. What are the most important attributes are associated with mortality caused by heart failure?
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***

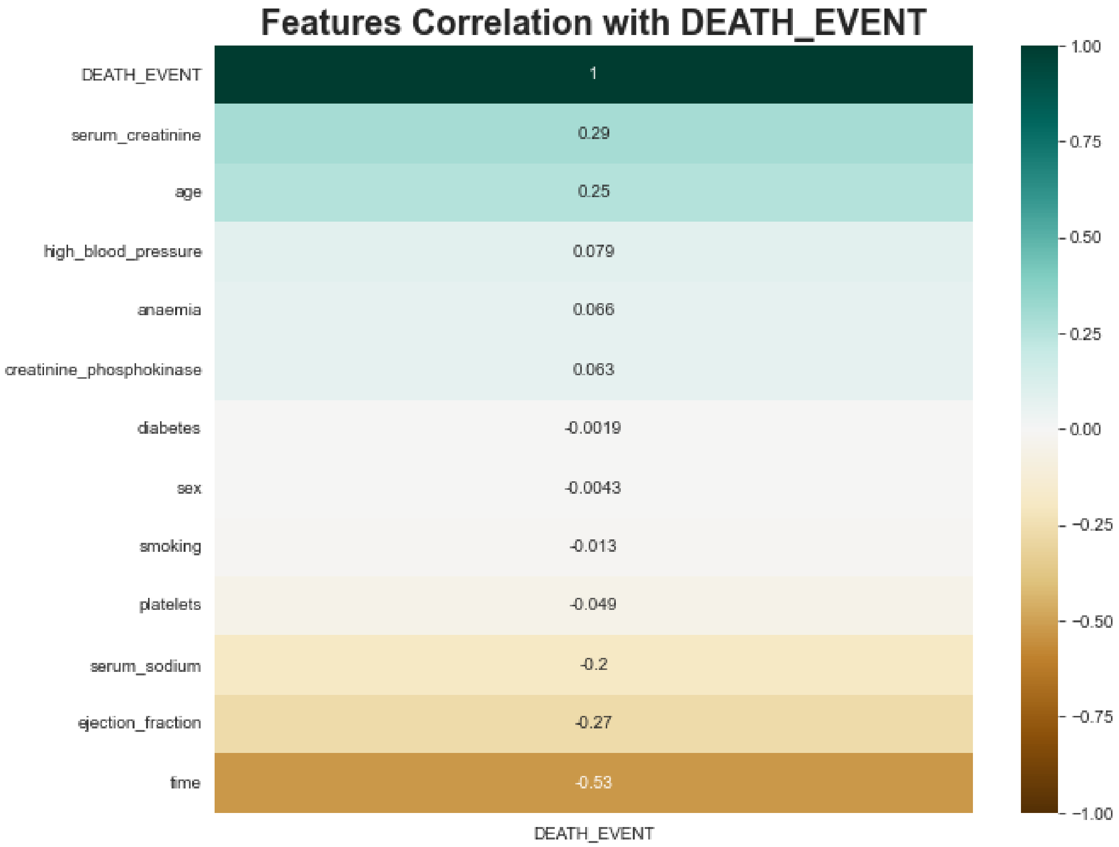
We provide a brief explanation of the features here: Anaemia (decreases of red blood cells or hemoglobin), high blood pressure (if the patient has a hypertension), diabetes (if the patient has diabetes), smoking (if the patient smokes) - (0: No, 1: Yes)- and sex (0: women, 1: men) features are binary. Creatinine_phsphokinase (level of the CPK anzyme in the blood), ejection_fraction (percentage of blood leaving), platelets (platelets in the blood), serum_ceatinine (level of creatinine in the blood), serum_sodium (level of sodium in the blood), age (between 40 and 95 years old), time (follow-up period) are numeric features. The DEATH_EVENT feature, states if the patient died or survived before the end of the follow-up period, that was 130 days on average. It should be our target variable for our analysis. Among these 13 features, 8 of them are the patients’ clinical report or health status which seems to have more effect on death or prevent it.

Summary of Data Cleansing and Exploratory Data Analysis

- ***Data Cleansing***

- Not any 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 high level are possible.

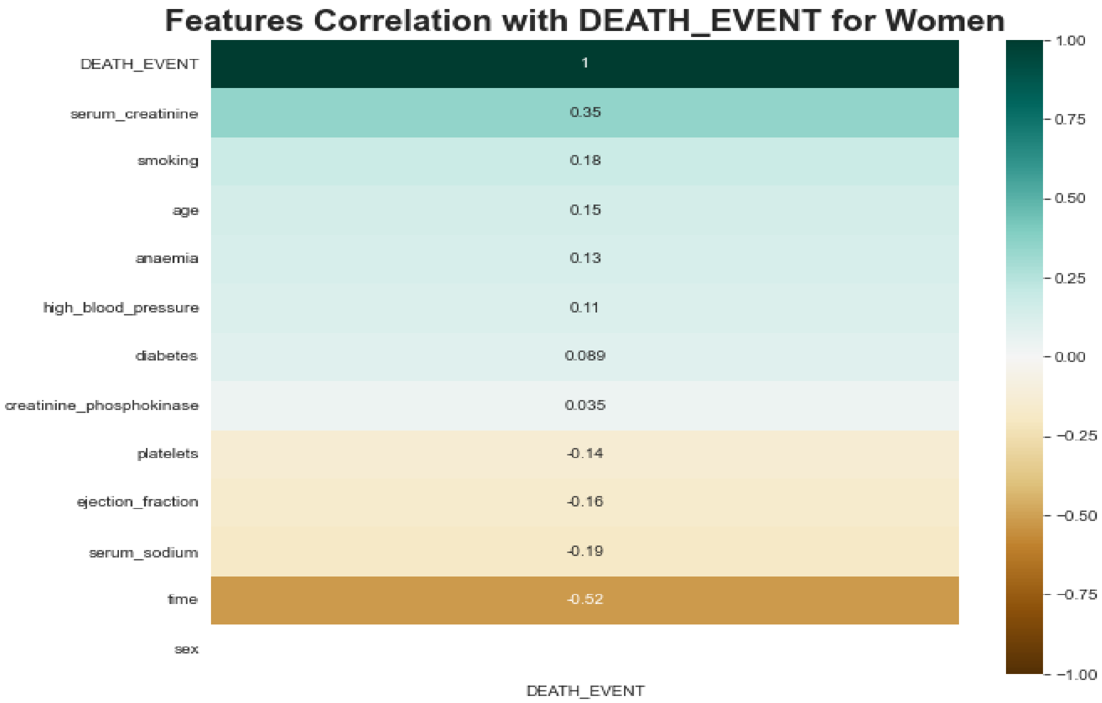
- ***Data Exploratory***



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

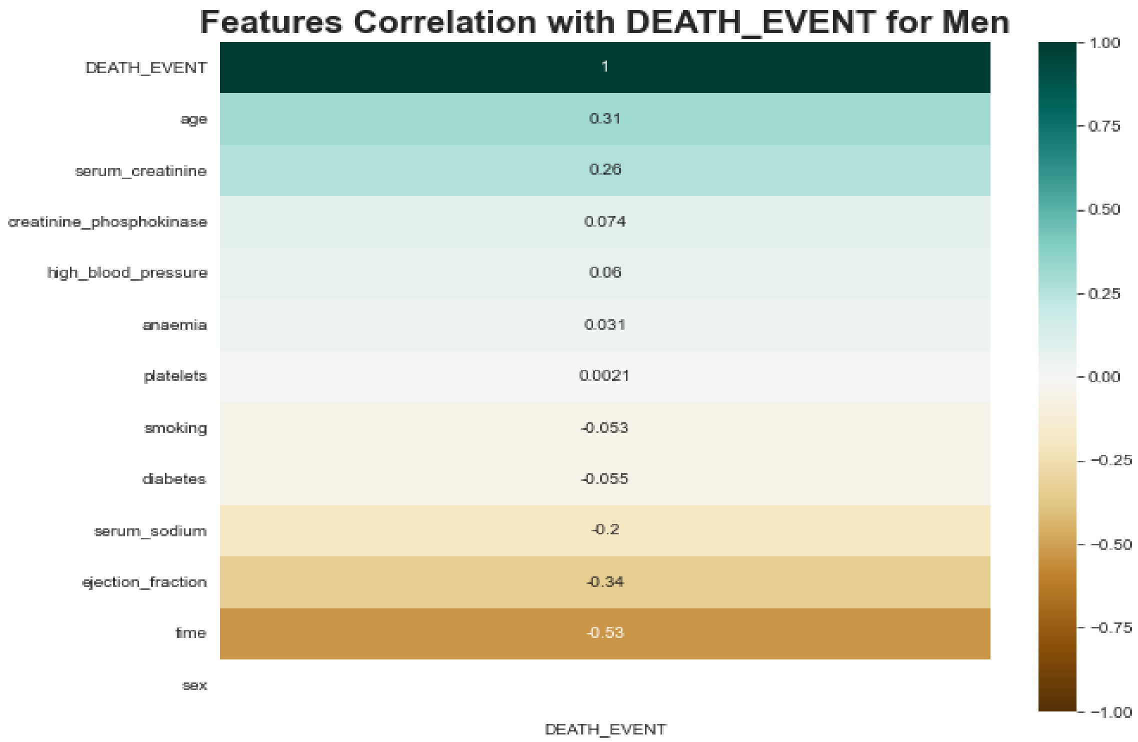
- Time
- Serum_creatinine
- Ejection_fraction
- Age
- Serum_sodium

There is an interesting thing here which is the features correlation with DEATH_EVENT changes for women and men.



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

- Time
- Serum_creatinine
- Serum_sodium
- Smoking
- Ejection_fraction



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

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

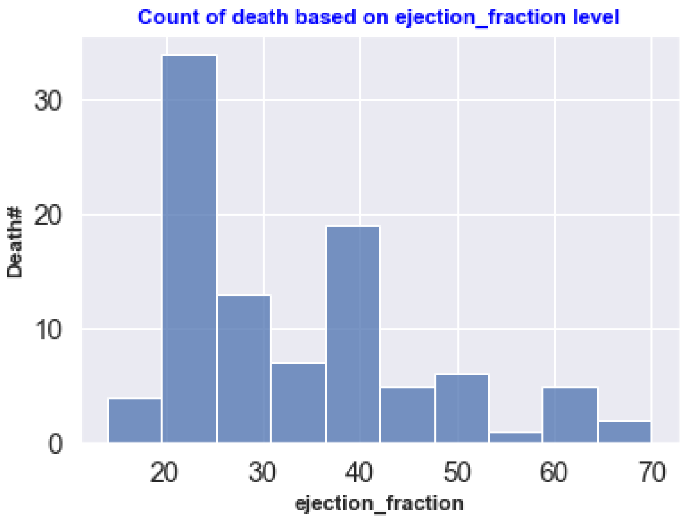
Ejection_fraction: 85.4% (most important feature after time). It makes sense in the real world since heart is not capable of pumping more blood to other body parts at lower ejection fraction.

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:

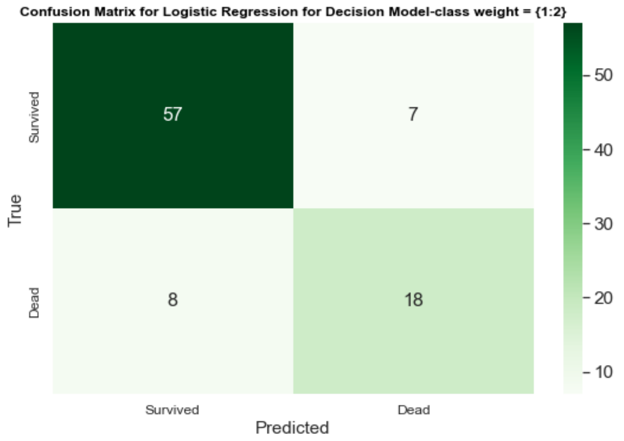
50%~75%

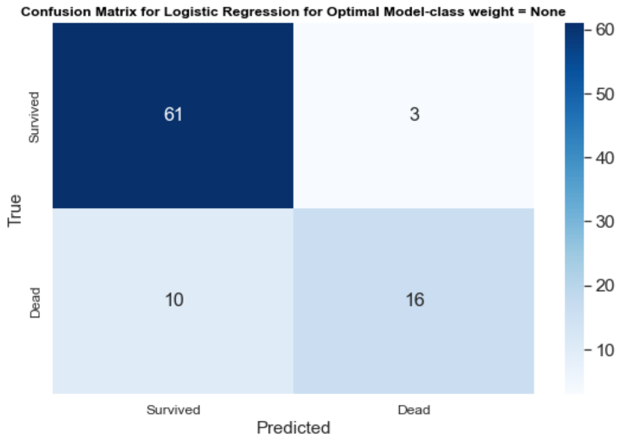
- ***Analysis and business insight***

The methods were used here for the analysis 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 value of 1 when death happens and 0 when it doesn't happen. Logistic regression machine learning model predicts whether if the patient with heart failure dies or survives, followed by understanding the influence of significant factors that truly affects. All methods start by setting the data into X and y datasets, then, splitting the dataset into training and test data, create the model, fit the model, train the model, and create prediction. The prediction model here is a mortality prediction which wants to predict mortality with the minimum faults. Logistic regression uses scaled data as it works better than raw data for it. But scaled data don't work well with decision tree and forest classifiers.





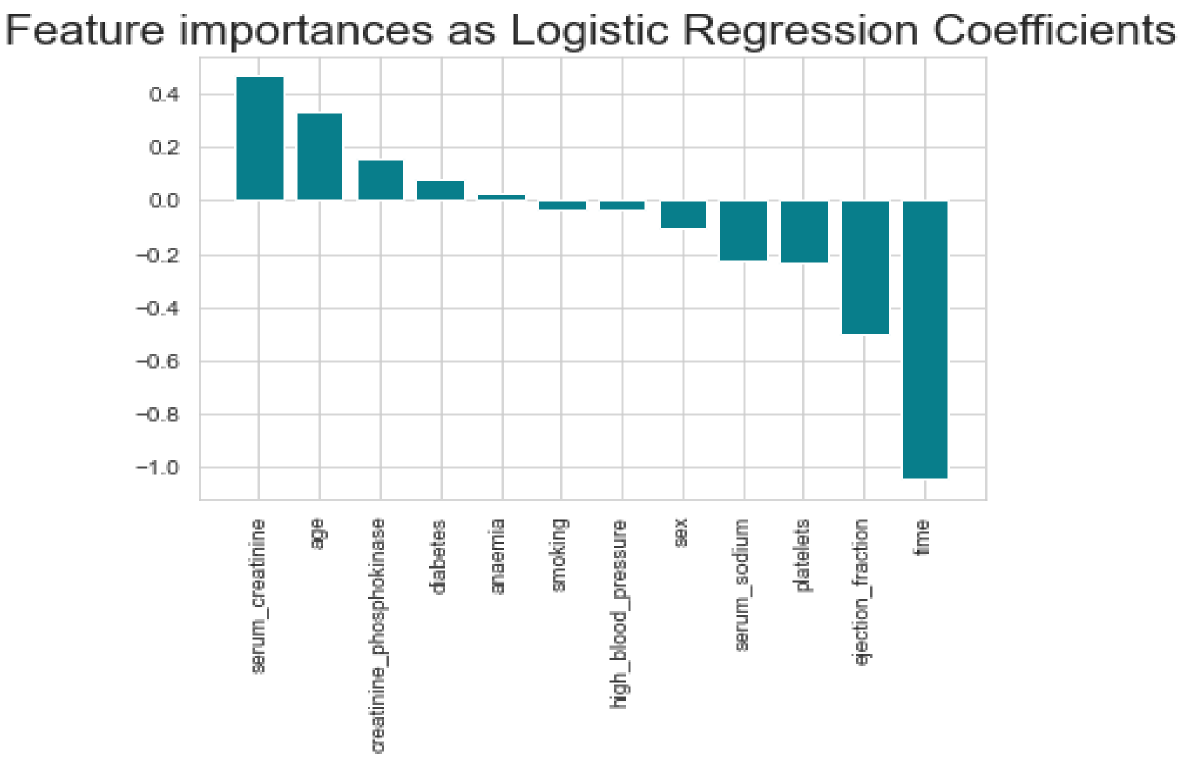
Accuracy Score: 0.8333
Recall Score: 0.6923
Precision Score: 0.7200
ROC AUC Score: 0.8636

Accuracy Score: 0.8556
Recall Score: 0.6154
Precision Score: 0.8421
ROC AUC Score: 0.8696

- 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:

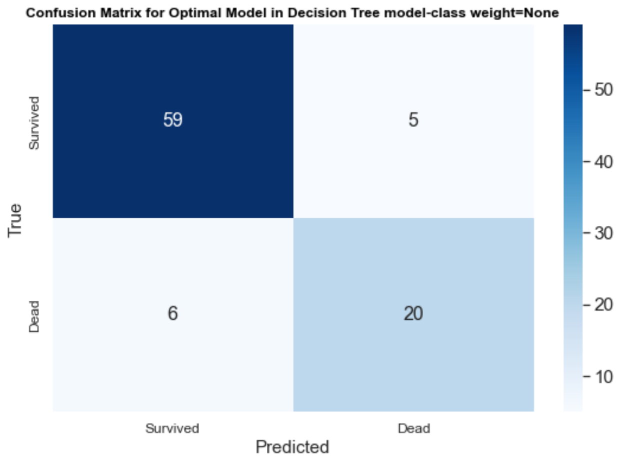
- Is 83% accurate to classifying the patients as dead or survived (**Accuracy Score** of 0.833)
- Predicted 69% correctly for predicting mortality (**Recall Score** of 0.692)
- Out of all mortality prediction 72% is truly mortality (**Precision Score** of 0.72)
- is performing 86.36% well (**ROC AUC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

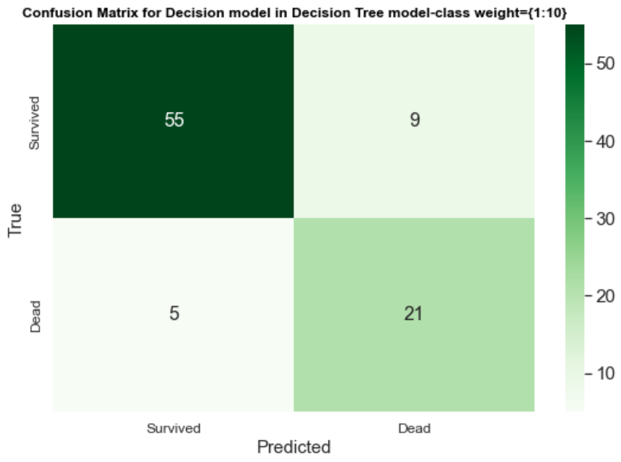


the larger the coefficient is (in both positive and negative direction), the more influence it has on a prediction.

2- Decision Tree classifier

Decision tree is one of the methods of choice for predictive modeling because it is very effective and also easy to understand.





Accuracy Score: 0.8778
Recall Score: 0.7692
Precision Score: 0.8000

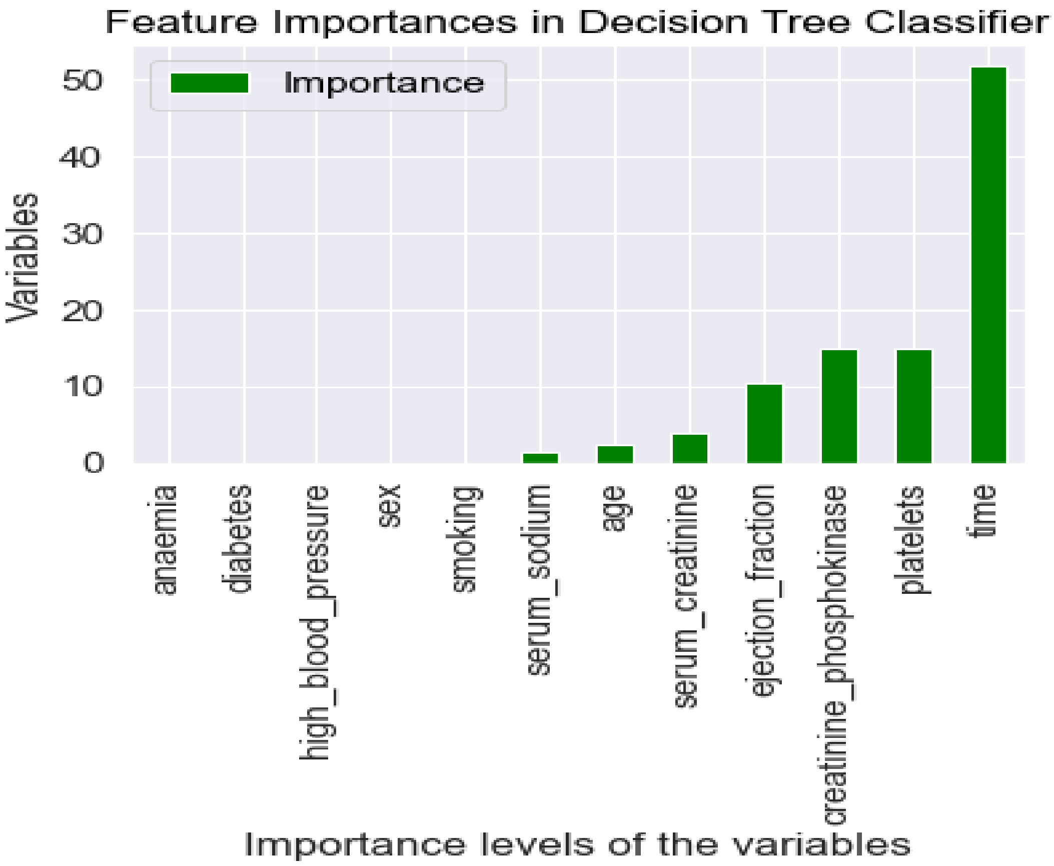
Accuracy Score: 0.8333
Recall Score: 0.9231
Precision Score: 0.6486

ROC AUC Score: 0.8224 ROC AUC Score: 0.8353

The model wants to give a good prediction on mortality but in the other hand it doesn’t want to predict survival badly. That’s why we choose the decision model from above in the right.

Mortality Prediction Model Random Forest:

- Is 85.5% accurate to classifying the patients as dead or survived (**Accuracy Score** of 0.833)
- Predicted 80.8% correctly for predicting mortality (**Recall Score** of 0.692)
- Out of all mortality prediction 72.4% is truly mortality (**Precision Score** of 0.72)
- is performing 89.6% well (**ROC AUC Score** of 0.896)

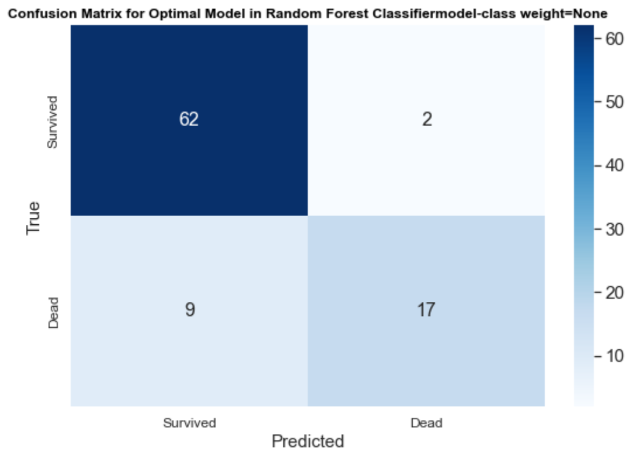


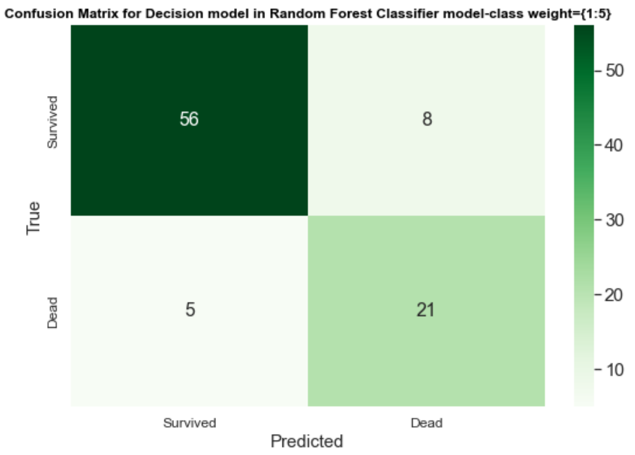
Above table shows the importance levels of the features which are time, platelets, creatinine_phosphokinase, ejection_fraction, serume_creatinine, age, and serum_sodium respectively. Compared to logistic regression feature importance, there is one more feature that has influence in death which is creatinine_phosphokinase

Note: In the dataset there are a lot of 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 higher CPK level than normal. Higher level of CPK is a crucial factor for heart attacks. 80% of all dead patients had 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.


Confusion Matrix for Optimal Model in Random Forest Classifiermodel-class weight=None


Confusion Matrix for Decision model in Random Forest Classifier model-class weight={1:5}

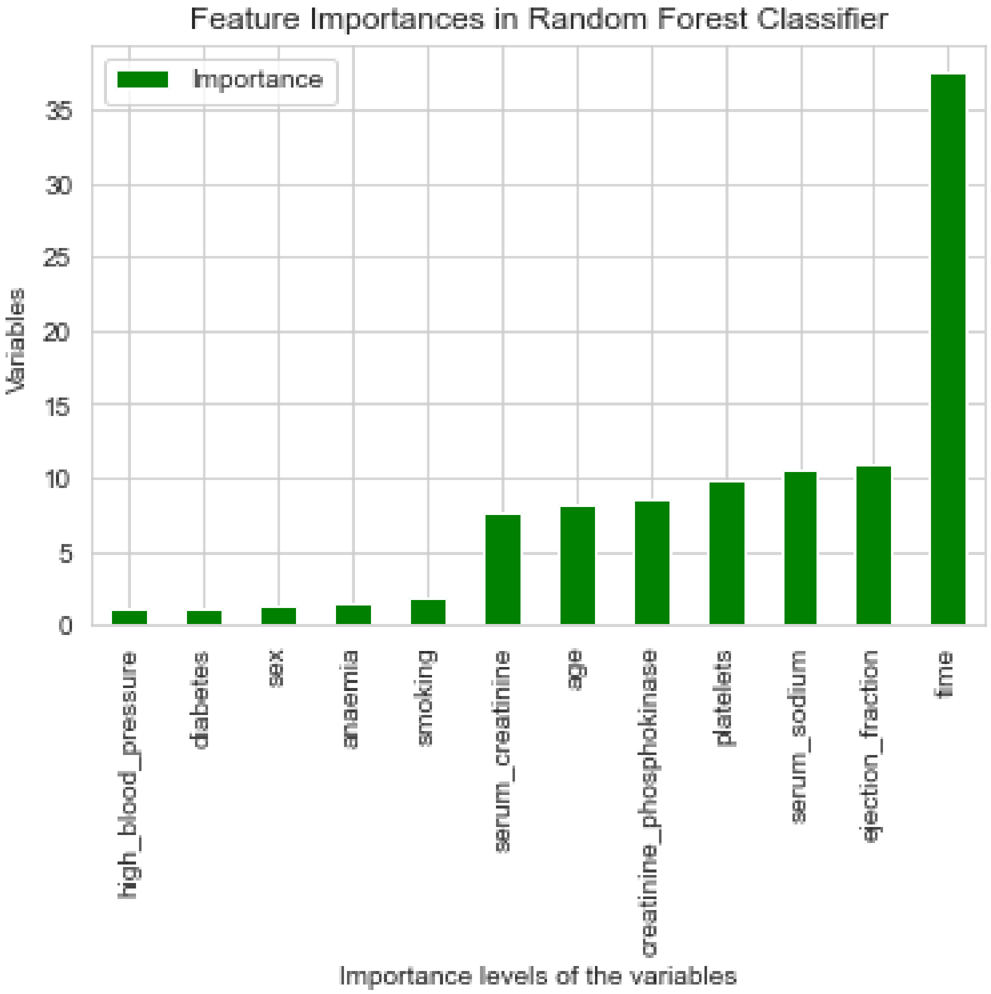
Accuracy Score: 0.8889
Recall Score: 0.6538
Precision Score: 0.9444
ROC AUC Score: 0.905

Accuracy Score: 0.855
Recall Score: 0.8077
Precision Score: 0.7241
ROC AUC Score: 0.8960

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

Mortality Prediction Model Random Forest:

- Is 85.5% accurate to classifying the patients as dead or survived (**Accuracy Score** of 0.855)
- Predicted 80.8% correctly for predicting mortality (**Recall Score** of 0.8077)
- Out of all mortality prediction 72.4% is truly mortality (**Precision Score** of 0.72)
- is performing 89.6% well (**ROC AUC Score** of 0.896)



Above chart shows all the features have influence in death but 5 features as smoking, anaemia, sex, diabetes, and high_blood_pressure have less influence 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 really need help)
- Maximize the number of right predictions on dead (TP)
- Minimize the number of wrong predictions on dead (FP) (As not wanting to increase patients ‘stress

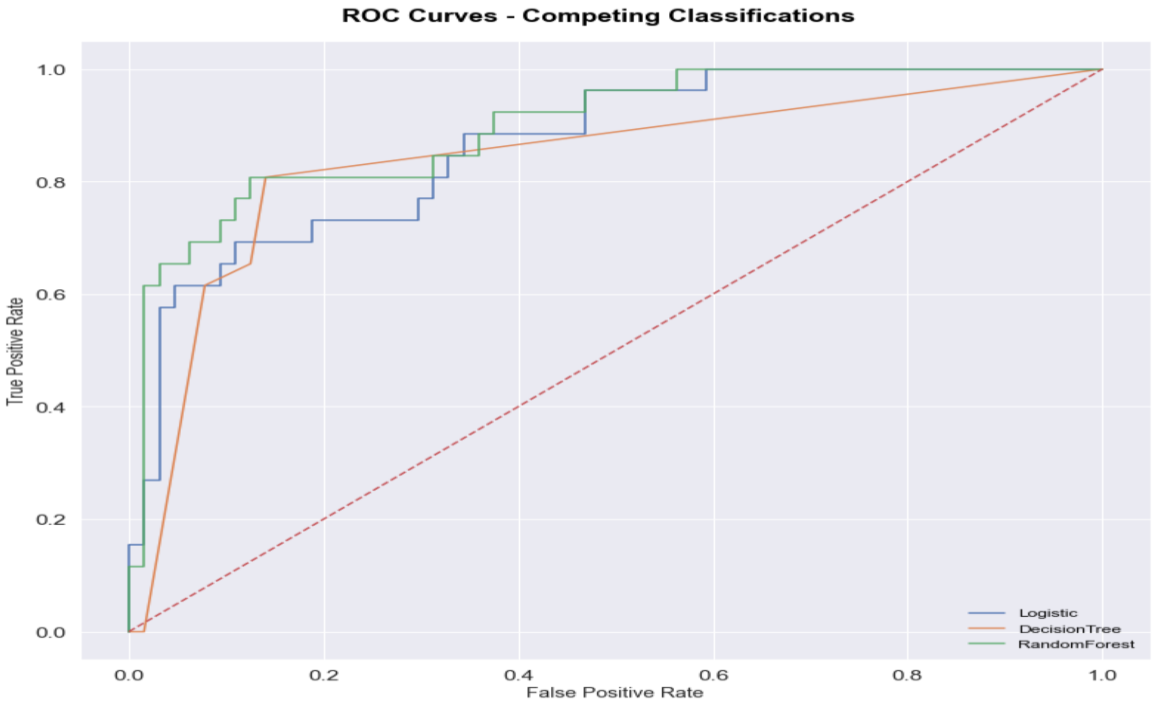
Conclusion:

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

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 analyze the dataset, based on the objectives, the project ended up with the mortality prediction model using Random Forest.

Also, by using Sklearn method of roc_curve() and computing roc auc score, we can plot the ROC curves for the 3 algorithms. It is evident from the plot that the AUC for the Random Forest ROC curve is higher than that for the Decision Tree and Logistic curve. Therefore, we can say that Random Forest did a better job of 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/improving ejection_fraction,
- reducing the level of creatinine_phosphokinase and 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 level of the above factors as soon as possible to survive their patients (as the follow-up period is the most negative correlated feature).