FINAL PROJECT REPORT

STATISTICAL, VISUALIZATION & PREDICTIVE ANALYSIS OF HEART FAILURE DATASET

GROUP 09 - MEMBERS

VISHAY PAKA (G01408948)
ADITHI KALLEM (G01407396)
ROHITH PANJALA (G01409071)

MENTOR: DR. ISURU DASSANAYAKE

Abstract – Nowadays, Heart failure is a serious health condition that is affecting millions of people worldwide and became and huge concern in health care system. The heart's inability to pump enough blood to meet the body's demand is leading to mortality worldwide, with an estimated 17 million deaths attributed to cardiovascular diseases annually. Monitoring changes in clinical features of a person is crucial, and analyzing trends among the features can help assess the impact of those features on heart failure. The electronic medical records provide a useful information that can be analyzed using machine learning techniques to identify correlations and patterns that may not be immediately evident to healthcare professionals. By analyzing patient data, machine learning can provide accurate predictions of patient survival and identify crucial features in medical records that are relevant to heart failure diagnosis and treatment. Therefore, this study aims to investigate the potential of machine learning algorithms in predicting patient survival and identifying key features in electronic medical records to improve heart failure management.[1]

I. INTRODUCTION

The World Health Organization (WHO) claims that the top cause of death worldwide is cardiovascular disease. The prevalence of heart failure is anticipated to rise as the population ages and the number of risk factors like hypertension, diabetes, and obesity increases. People with cardiovascular disease (due to the presence of one or more risk factors such as high blood pressure, diabetes, hypertension etc.) need early detection and management wherein a machine learning model can be of great help. Heart failure care is difficult and necessitates a multidisciplinary strategy that includes medicine, device therapy, and lifestyle changes. The morbidity and mortality rates related to heart failure remain high despite breakthroughs in treatment. The dataset selected for analysis provides valuable insights into the impact of the clinical features on heart failure.[1]

DATASET

The dataset is taken from UCI Machine Learning Repository which is a healthcare dataset. This dataset contains 299 rows with 12 columns which are useful for the prediction of heart failure. The structure of the data is analyzed, and cleaning is done for platelets and age columns.[2]

Tools & Languages Used: R Programming.

Research Questions:

By utilizing the trends obtained by applying statistical and predictive analytic methods to the dataset, this paper will answer the following research questions:

1. Can we predict death of a patient during the follow-up period using the dataset and the patient's clinical features. Which machine learning algorithm plays a vital role in this prediction and which algorithm gives the best accuracy preventing overfitting and underfitting?
2. Can we identify the most important clinical features in predicting the mortality of heart failure patients, and how do they contribute to the prediction accuracy?
3. If a patient of age 78 admitted in hospital and stayed more than 70 days and the level of serum creatinine in his blood is 1.5(mg/dL). The percentage of the blood leaving his heart at each contraction is 42. Will the patient be alive or dead? With which model you can interpret the result in a meaningful way?
4. As the dataset has only 299 observations, can we collect more data from an API which prevents our model from underfitting or is it possible to use random normal distributions to stimulate data based on original data to do predictions? If, so how does model behave in terms of accuracy?
5. Is there a significant association between smoking status and death rate among patients with cardiovascular disease? If not, is there any other categorical variable which have association with the death rate among patients? Can the above two questions be answered by performing a chi-square test?
6. Can we predict the effect of age, ejection fraction, serum creatinine level, and other clinical features on the patient's follow-up period?

II. METHODOLOGY

The data contains information on the person's age, having anemia or not, level of the CPK enzyme in the blood (mcg/L), has diabetes or not, Percentage of blood leaving the heart at each contraction, has hyper tension or not, Platelets in the blood (kilo platelets/mL), Level of serum creatinine in the blood (mg/dL), Level of serum sodium in the blood (mEq/L), gender, smokes or not, follow-up period (days), and the person deceased during the follow-up period or not.

The data preprocessing has been done before several statistical and visualization analysis have been performed on the dataset and drawn conclusions on whether the data is skewed or not and if it's skewed transformation of the variables is done by scaling the continuous variables. Some of the data cleaning has been done on the dataset. The column of age one observation as age 66.67, which stands out of the rest, so rounding off as we have less observations for prediction. Also, platelets column has scientific notation for two records. So, formatting it into numeric as a pre-processing step.

Then the dataset has been fitted with different classification algorithms such as logistic regression, k-nearest neighborhood, decision tree, random forest algorithm for both scaled and unscaled data. The Logistic model has been fitted with only significant variables and the results are compared with the full model. As the observations in the dataset are less, simulation of the data using random normal distributions is done and trained the model to see how the model behaves. Then concluded with all the model’s performances with there metrics and what features play important role in predicting the survival of an individual. The libraries that are used in this research are dplyr, tidyr, ggplot2, caret, psych, tree, rpart, rattle, randomForest, lattice, car, ggsci, gridExtra, class.

III. RESULTS

STATISTICAL ANALYSIS AND VISUALIZATIONS

The summary of the dataset with mean, median, minimum, and maximum values of all the variables is studied. The data distribution and whether the dataset is balanced or not can be interpreted from the summary.

The summary tells that total death events taken place during the follow up period are 96 and the patients who survived are 203. By observing all these statistical summaries of the predictors and response variables. We can conclude that the data set is slightly imbalanced as there are more records of the patients who survived compared to the people who died because of the heart failure but have less difference with patients having anaemia or not, diabetes or not, high blood pressure or not, smoking or not. In general, out of 7.2 billion population only less proportionate of people will die with heart failure condition. So, visualization of box plots, histograms, and density plots on each variable with response variable can be observed in Fig. 1 & 2.

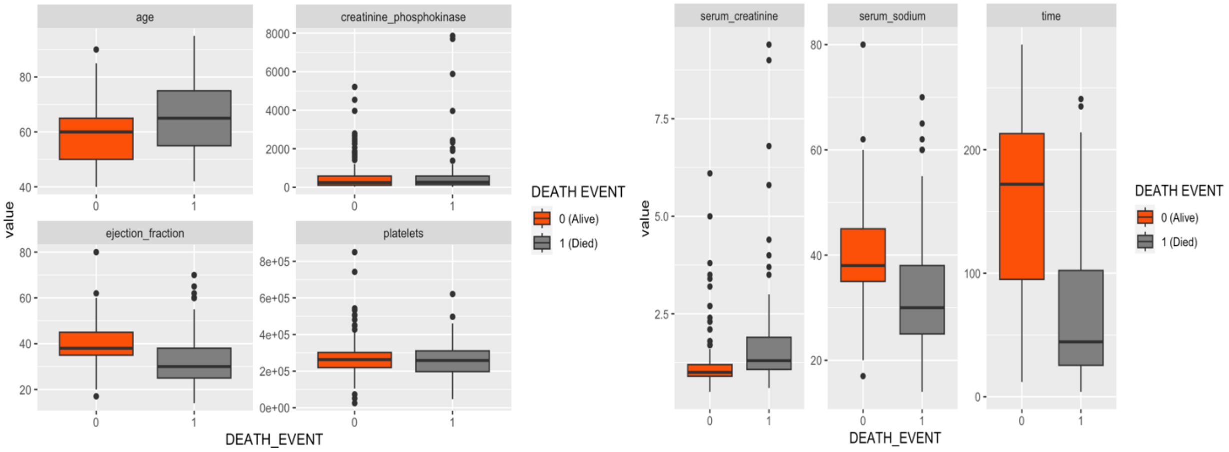


Fig. 1 & 2: Box plot representation of different predictors and death event.

From the above Fig. 1 & 2, box plot has been done on death event with various predictors. We can observe that ejection fraction and serum sodium have same kind of box plot with same scale and same type of data distribution. Also, creatinine phosphokinase and serum creatinine have data skewed to one side resulting in many outliers followed by platelets. The remaining predictors are normally distributed among the response variable and seems to not cause any further problem.

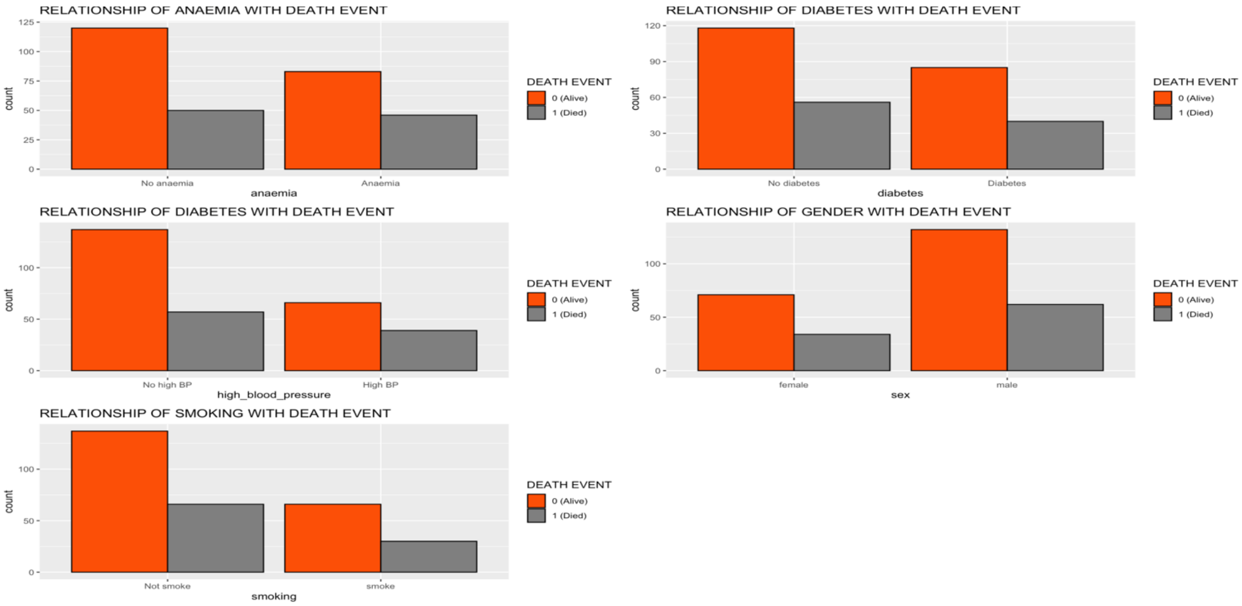


Fig. 3: Bar chart of various categorical predictors on death event.

From the above Fig. 3 bar chart is plotted between various categorical variables with the response variable i.e., death event. We expected the death rate ratio to be more with the people who have anaemia but from the visualization we found that the data distribution for anaemia is different, we have more percentage of deaths in no anaemia than anaemia. We then checked for all the other categorical predictors; it is the same case over here as well. Let's check the histogram and density plots for certain variables.

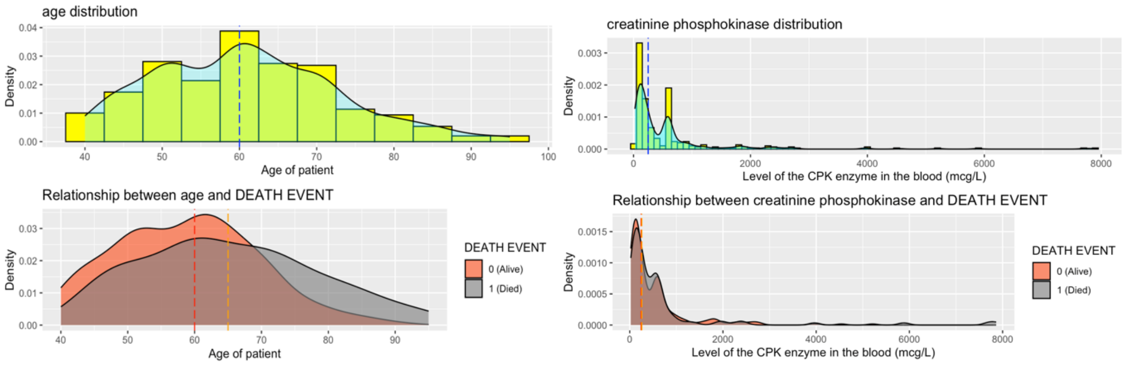


Fig. 4 & 5: Histogram and Density charts of age and creatinine predictors on death event.

From Fig. 4 it can be observed that the age of the patients was highest around 60 years and you can observe that, younger the age, the density plot of survival is high and as the age increases density plot of death is more. After the age 70 the density plot is reversed. Whereas in Fig. 5 we can interpret that the distribution is heavily skewed to one side and the occurrences are more around the range of 20 -250, and we need to calculate the skewness and scale it accordingly. This ensures us to prevent any one variable from dominating the analysis simply because it has larger values.

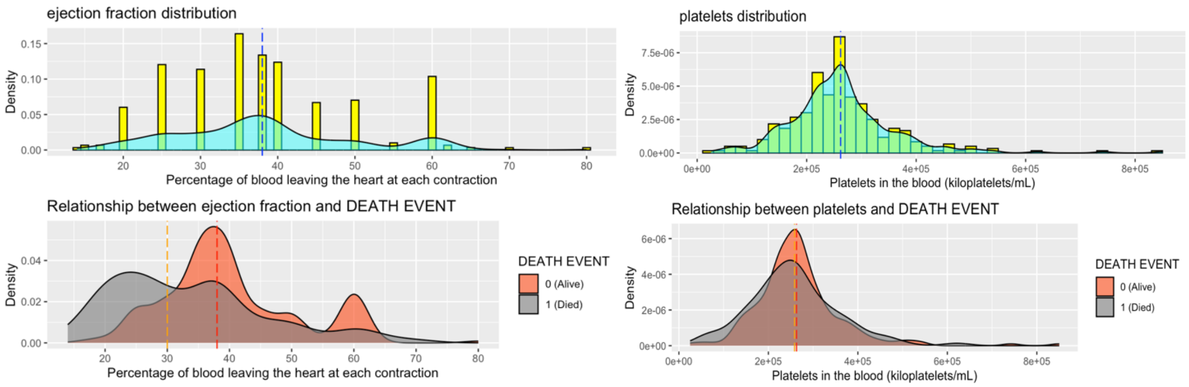


Fig. 6 & 7: Histogram and Density charts of ejection fraction and platelets predictors on death event.

The distribution is discrete and ejection fraction for a normal healthy person will be around 50-60 and even if it raises also, there will be no problem. But if the ejection fraction falls to 30 and below there is a higher chances of heart failure. The same trend we can observe from the Fig. 6. The insights we can get from Fig. 7 are, the distribution is symmetric, and survivals have the highest platelets.

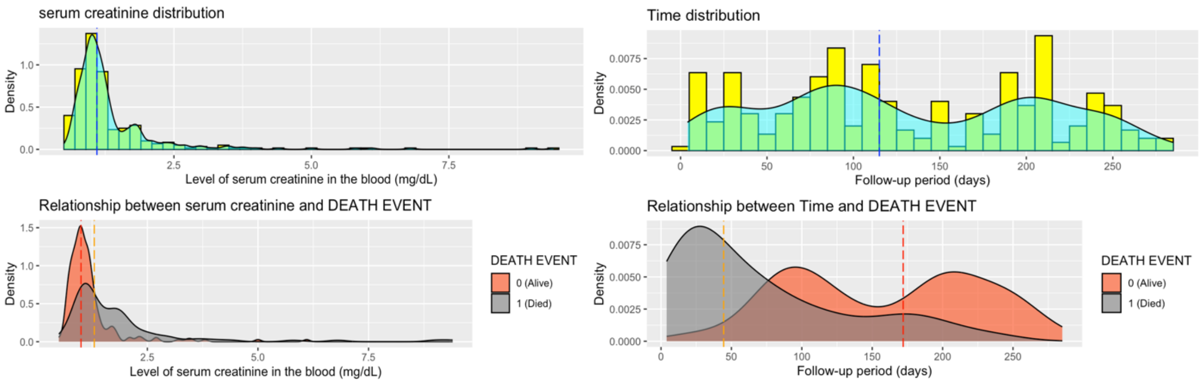
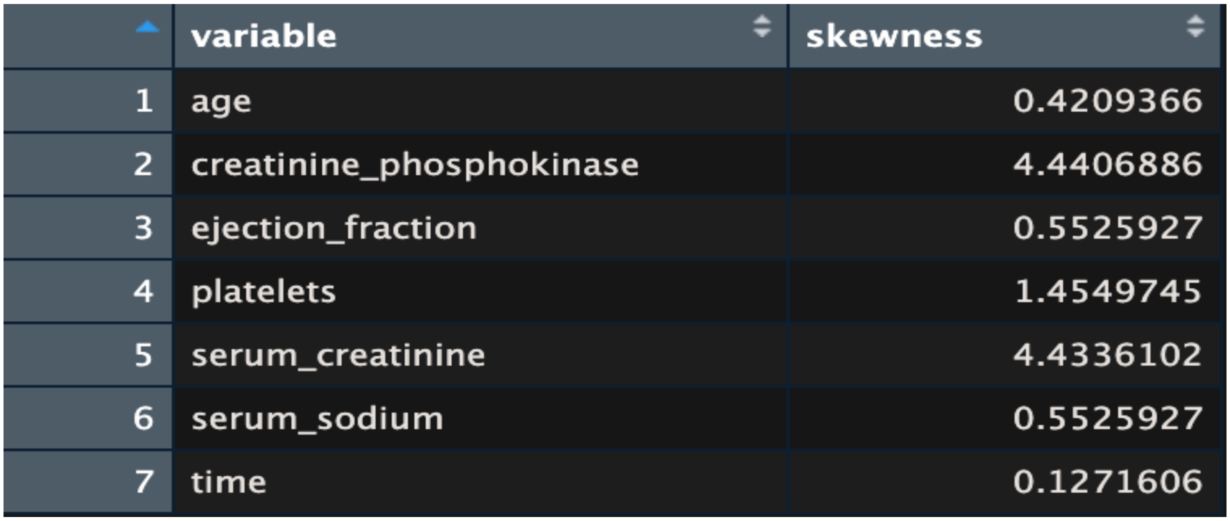


Fig. 8 & 9: Histogram and density charts of serum creatinine and time predictors on death event.

The insights we get from above Fig. 8 for serum creatinine distribution are, the distribution is highly skewed and will be scaled later in this research, for survivals the value is around the median and the patients who have 1.5mg/dL have high risk of heart failure. Whereas from Fig. 9 time has equal distribution. The patients who have more follow-up days are likely to have higher chances of survival and the patients who have less than 60 follow-up days have higher chances of death. So, the more you follow-up on you regular health check-up's the more you have probability to survive.

From the above density and histogram plots, we can observe the data is moderately skewed, it may violate the assumptions of the model, leading to biased results or inaccurate predictions. Now let's calculate the skewness of the dataset and observe which has zero, negative and positive skewness and transform them.

Standardizing a variable means transforming it so that it has a mean of 0 and a standard deviation of 1. This is also known as z-score normalization [3]. It's important to note that standardizing categorical variables or binary variables (such as anaemia, diabetes, high blood pressure, sex, smoking) doesn't make sense, as they only take on values of 0 or 1. Therefore, it is important to carefully consider which variables are continuous and which are categorical, and only standardize the continuous variables in order to avoid any potential loss of information or interpretation problems.



	variable	skewness
1	age	0.4209366
2	creatinine_phosphokinase	4.4406886
3	ejection_fraction	0.5525927
4	platelets	1.4549745
5	serum_creatinine	4.4336102
6	serum_sodium	0.5525927
7	time	0.1271606

Fig. 10: skewness calculations of continuous predictors.

Fig. 10 shows that the variables creatinine phosphokinase, serum creatinine and possibly platelets have a high degree of skewness, which could potentially affect the performance of some statistical models. The performance of the model might improve if the values are scaled. Let's compare both scaled and un-scaled data below by fitting into different models.

Before training and testing the models on scaled and unscaled dataset, let’s observe the correlation analysis of the dataset.

CORRELATION ANALYSIS ON THE DATASET

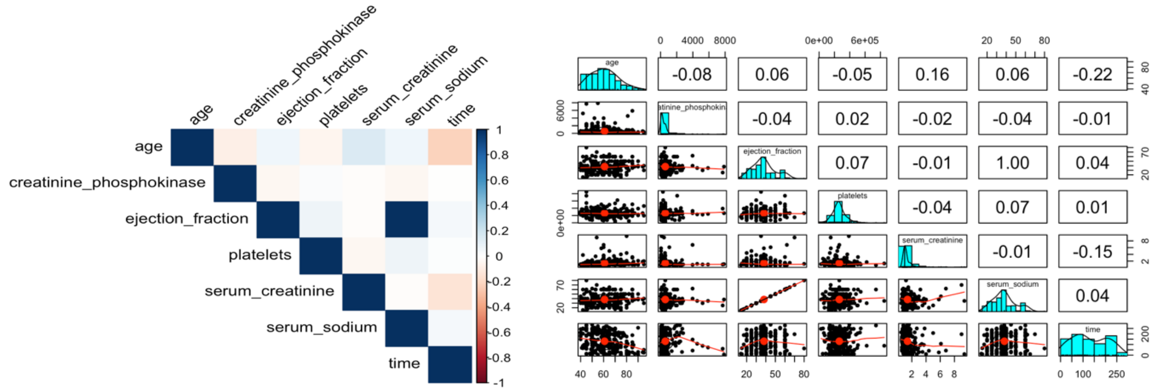


Fig. 11 & 12: Correlation Matrix on the dataset.

The correlation has been calculated for predictors in the dataset. Fig. 11 shows that the correlation between serum sodium and ejection fraction is 1, which indicates that they are perfectly linearly related and increase or decrease together. Having correlation 1 means they are perfect positive and from the Fig. 12, we can see a linear line between serum sodium and ejection fraction. Let's see how it affects our models when fitted later. This is the reason we have seen same kind of box plots which can be observed from the Fig. 1 & 2.

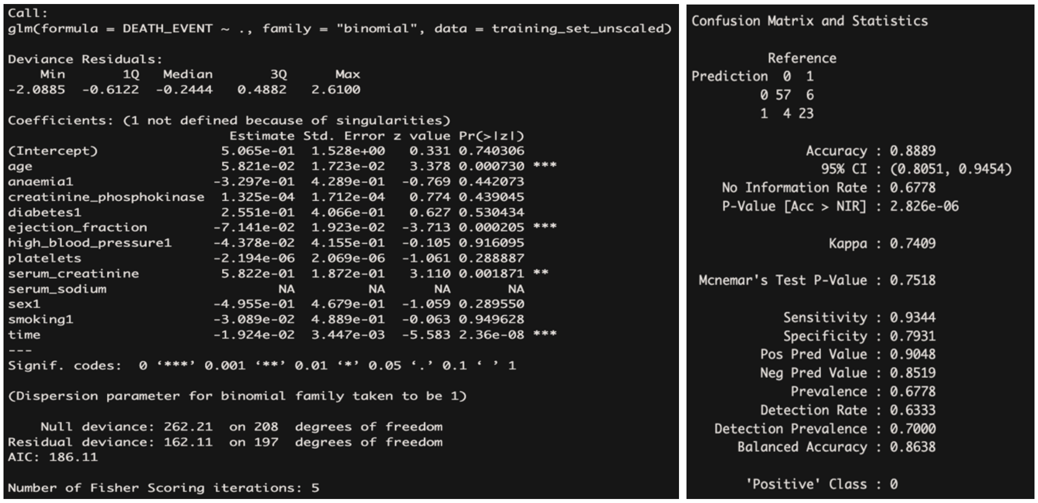
Other than this the only two variables which have moderate correlation is time and age, they show negative correlation of -0.22. No, other variables have good correlation.

FITTING THE MODELS AND ANALYZING THEIR METRICS

A different type of **classification** algorithms has been fitted to the dataset and their performances have been noted. The models used in this paper to answer the research questions are Logistic Regression, Decision tree, Random Forest, and K-Nearest Neighbors Algorithms.

The dataset has been split into 70% of training data and 30% of testing data for both the datasets (scaled and unscaled). The metrics of both the datasets have been noted when fitted to different classification problems and tabled at the conclusion of the paper.

MODEL 1: LOGISTIC REGRESSION - ALL VARIABLES (UNSCALED)



```
Call:
glm(formula = DEATH_EVENT ~ ., family = "binomial", data = training_set_unscaled)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0885  -0.6122  -0.2444   0.4882   2.6100

Coefficients: (1 not defined because of singularities)
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)               5.065e-01  1.528e+00   0.331 0.740306
age                       5.821e-02  1.723e-02   3.378 0.000730 ***
anaemia1                 -3.297e-01  4.289e-01  -0.769 0.442073
creatinine_phosphokinase  1.325e-04  1.712e-04   0.774 0.439045
diabetes1                 2.551e-01  4.066e-01   0.627 0.530434
ejection_fraction        -7.141e-02  1.923e-02  -3.713 0.000205 ***
high_blood_pressure1     -4.378e-02  4.155e-01  -0.105 0.916095
platelets                -2.194e-06  2.069e-06  -1.061 0.288887
serum_creatinine          5.822e-01  1.872e-01   3.110 0.001871 **
serum_sodium                     NA         NA      NA       NA
sex1                     -4.955e-01  4.679e-01  -1.059 0.289550
smoking1                 -3.089e-02  4.889e-01  -0.063 0.949628
time                     -1.924e-02  3.447e-03  -5.583 2.36e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 262.21  on 208  degrees of freedom
Residual deviance: 162.11  on 197  degrees of freedom
AIC: 186.11

Number of Fisher Scoring iterations: 5
```

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 57  6
         1  4 23

               Accuracy : 0.8889
                 95% CI : (0.8051, 0.9454)
    No Information Rate : 0.6778
    P-Value [Acc > NIR] : 2.826e-06

                  Kappa : 0.7409

 Mcnemar's Test P-Value : 0.7518

            Sensitivity : 0.9344
            Specificity : 0.7931
         Pos Pred Value : 0.9048
         Neg Pred Value : 0.8519
             Prevalence : 0.6778
         Detection Rate : 0.6333
   Detection Prevalence : 0.7000
      Balanced Accuracy : 0.8638

       'Positive' Class : 0
```

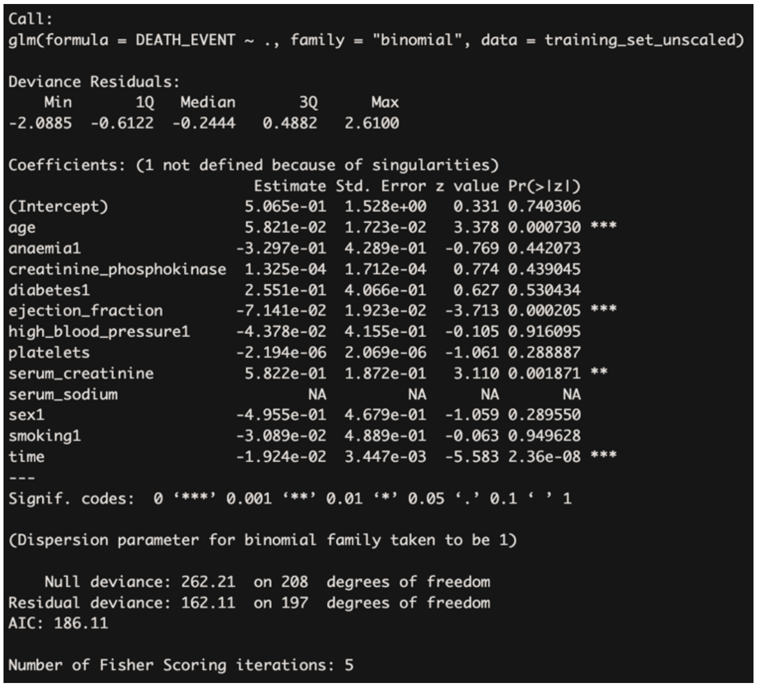
Fig. 13: Summary statistics and confusion matrix for logistic regression – All variables (unscaled).

The logistic regression has fitted to whole dataset of an unscaled data with death event as a response variable. We got an accuracy of 88.9% when logistic regression is done on the un-scaled data and a true positive rate (recall) of 93.44%. From Fig. 13 it can be noted the coefficients (age, ejection fraction, serum creatinine, time) have asterisks next to their p-values, indicating that they are statistically significant at certain significance levels ($p < 0.05$), while others are not. The coefficients with the variables which are insignificant have the p value greater than 0.05 meaning they have weak evidence against the null hypothesis, and we do not reject null hypothesis. The lower p-values are considered more statistically significant and are typically interpreted as having a stronger association with the response variable (Death Event). The others with no asterisk are insignificant variables and we can remove them for further models.

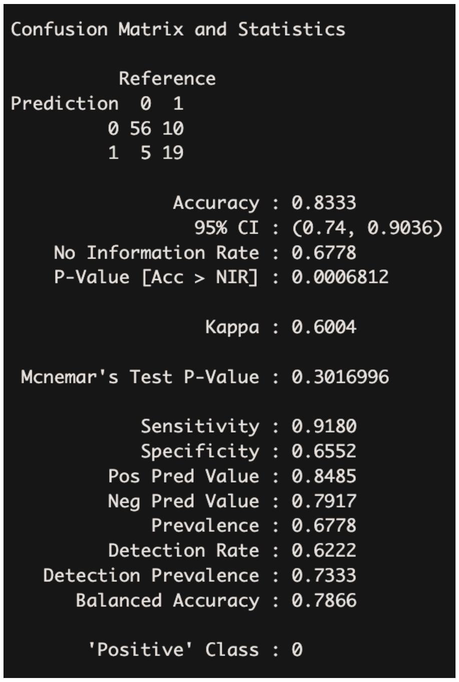
We can also see serum sodium has NA values in the estimate column, standard error column, and p-value column. Which means the variable is not included in the model and the variable was likely dropped during the selection process. We observed that it is perfect positive with ejection fraction predictor. The main reason for the model to not select serum sodium is, this variable is highly correlated with the other predictor which is already included in the model.

MODEL 2: LOGISTIC REGRESSION - ALL VARIABLES (SCALED)

Now the scaled data has been fitted with logistic regression to see how it performs and dose it improves the metrics of the model. As we can see Fig. 14 shows that the un-scaled data when fitted for the logistic model, we got an accuracy of 88.9% and when fitted with the scaled data we got an accuracy of 83.3%. It is possible that scaling the data may have reduced the accuracy of the logistic regression model compared to the un-scaled data. This could be due to a few reasons, one is outliers, scaling the data can sometimes amplify the effect of outliers, which can negatively impact the accuracy of the model. The other reason could be if there are non-linear relationships between the predictors and the response, scaling the data may reduce the ability of the model to capture these relationships. In general, it is good to experiment both scaled and unscaled data and choose the model with good metrics. Here in our case, we choose un-scaled data fitted for logistic model.



```
Call:
glm(formula = DEATH_EVENT ~ ., family = "binomial", data = training_set_unscaled)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0885  -0.6122  -0.2444   0.4882   2.6100

Coefficients: (1 not defined because of singularities)
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)             5.065e-01  1.528e+00   0.331 0.740306
age                     5.821e-02  1.723e-02   3.378 0.000730 ***
anaemia1               -3.297e-01  4.289e-01  -0.769 0.442073
creatinine_phosphokinase  1.325e-04  1.712e-04   0.774 0.439045
diabetes1               2.551e-01  4.066e-01   0.627 0.530434
ejection_fraction      -7.141e-02  1.923e-02  -3.713 0.000205 ***
high_blood_pressure1   -4.378e-02  4.155e-01  -0.105 0.916095
platelets              -2.194e-06  2.069e-06  -1.061 0.288887
serum_creatinine        5.822e-01  1.872e-01   3.110 0.001871 **
serum_sodium                   NA         NA      NA       NA
sex1                   -4.955e-01  4.679e-01  -1.059 0.289550
smoking1               -3.089e-02  4.889e-01  -0.063 0.949628
time                   -1.924e-02  3.447e-03  -5.583 2.36e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 262.21  on 208  degrees of freedom
Residual deviance: 162.11  on 197  degrees of freedom
AIC: 186.11

Number of Fisher Scoring iterations: 5
```



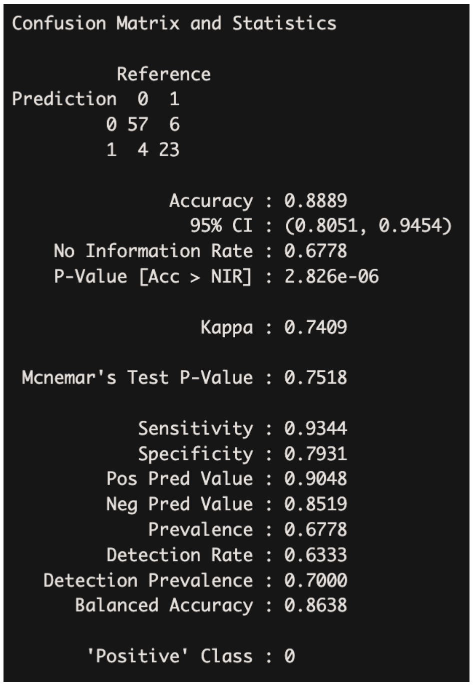
```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 56 10
         1  5 19

               Accuracy : 0.8333
                 95% CI : (0.74, 0.9036)
    No Information Rate : 0.6778
    P-Value [Acc > NIR] : 0.0006812

                  Kappa : 0.6004

 Mcnemar's Test P-Value : 0.3016996

            Sensitivity : 0.9180
            Specificity : 0.6552
         Pos Pred Value : 0.8485
         Neg Pred Value : 0.7917
             Prevalence : 0.6778
         Detection Rate : 0.6222
   Detection Prevalence : 0.7333
      Balanced Accuracy : 0.7866

       'Positive' Class : 0
```

Fig. 14: Summary statistics and confusion matrix for logistic regression – All variables (scaled).

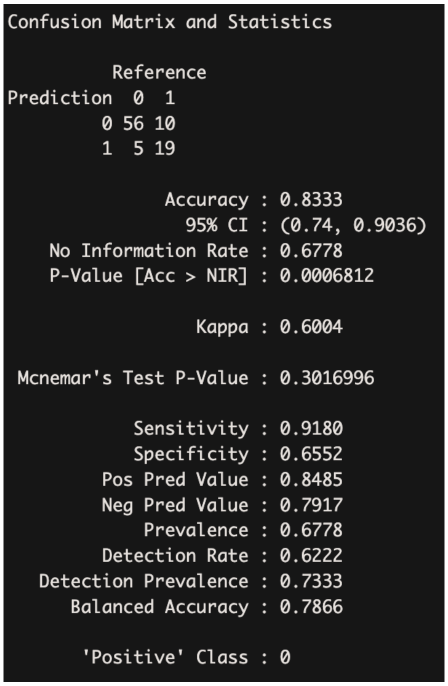
Since, we had perfect positive correlation between serum sodium, ejection fraction. Let’s try removing the serum sodium and observe how it performs on scaled and unscaled data.

MODEL 3: LOGISTIC REGRESSION – REMOVING SERUM SODIM VARIABLE

UNSCALED **SCALED**



```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 57  6
         1  4 23

               Accuracy : 0.8889
                 95% CI : (0.8051, 0.9454)
    No Information Rate : 0.6778
    P-Value [Acc > NIR] : 2.826e-06

                  Kappa : 0.7409

 Mcnemar's Test P-Value : 0.7518

            Sensitivity : 0.9344
            Specificity : 0.7931
         Pos Pred Value : 0.9048
         Neg Pred Value : 0.8519
             Prevalence : 0.6778
         Detection Rate : 0.6333
   Detection Prevalence : 0.7000
      Balanced Accuracy : 0.8638

       'Positive' Class : 0
```

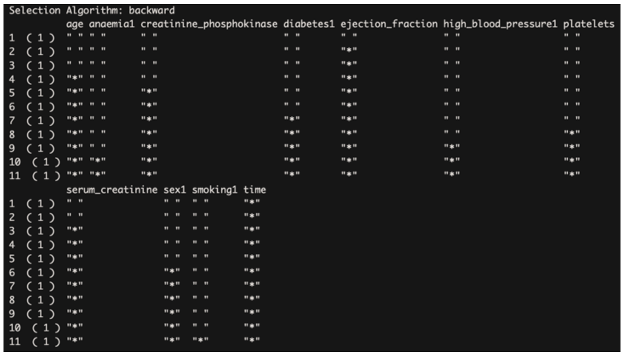


```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 56 10
         1  5 19

               Accuracy : 0.8333
                 95% CI : (0.74, 0.9036)
    No Information Rate : 0.6778
    P-Value [Acc > NIR] : 0.0006812

                  Kappa : 0.6004

 Mcnemar's Test P-Value : 0.3016996

            Sensitivity : 0.9180
            Specificity : 0.6552
         Pos Pred Value : 0.8485
         Neg Pred Value : 0.7917
             Prevalence : 0.6778
         Detection Rate : 0.6222
   Detection Prevalence : 0.7333
      Balanced Accuracy : 0.7866

       'Positive' Class : 0
```

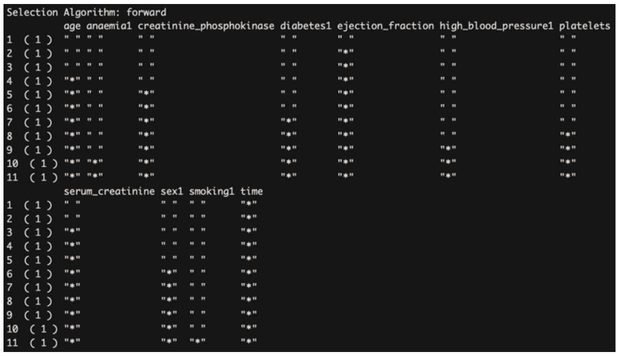
Fig. 15: confusion matrix for logistic regression – Removing serum sodium variable (unscaled & scaled).

From Fig. 15, we got the same confusion matrix for the full model and when serum sodium is removed from both scaled and un-scaled data. The accuracies remain the same i.e., 88.89% and 83.3% respectively.

Now let's find the significant predictors and fit the model and compare the accuracies for both scaled and unscaled data and compare the model with the above models. From the summary of model 1 and 2 we got to know that time, age, ejection fraction, and serum creatinine are significant, let’s check the same with backward approach and forward approach.



```
Selection Algorithm: backward
          age anaemia1 creatinine_phosphokinase diabetes1 ejection_fraction high_blood_pressure1 platelets
1  ( 1 )  " " " "      " "                      " "       " "               " "                  " "
2  ( 1 )  " " " "      " "                      " "       "*"               " "                  " "
3  ( 1 )  " " " "      " "                      " "       "*"               " "                  " "
4  ( 1 )  "*" " "      " "                      " "       "*"               " "                  " "
5  ( 1 )  "*" " "      "*"                      " "       "*"               " "                  " "
6  ( 1 )  "*" " "      "*"                      " "       "*"               " "                  " "
7  ( 1 )  "*" " "      "*"                      "*"       "*"               " "                  " "
8  ( 1 )  "*" " "      "*"                      "*"       "*"               " "                  "*"
9  ( 1 )  "*" " "      "*"                      "*"       "*"               "*"                  "*"
10  ( 1 ) "*" "*"      "*"                      "*"       "*"               "*"                  "*"
11  ( 1 ) "*" "*"      "*"                      "*"       "*"               "*"                  "*"
          serum_creatinine sex1 smoking1 time
1  ( 1 )  " "              " "  " "      "*"
2  ( 1 )  " "              " "  " "      "*"
3  ( 1 )  "*"              " "  " "      "*"
4  ( 1 )  "*"              " "  " "      "*"
5  ( 1 )  "*"              " "  " "      "*"
6  ( 1 )  "*"              "*"  " "      "*"
7  ( 1 )  "*"              "*"  " "      "*"
8  ( 1 )  "*"              "*"  " "      "*"
9  ( 1 )  "*"              "*"  " "      "*"
10  ( 1 ) "*"              "*"  " "      "*"
11  ( 1 ) "*"              "*"  "*"      "*"
```



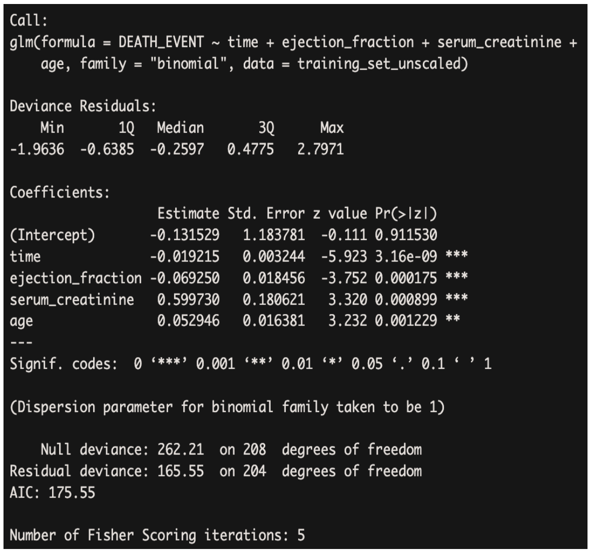
```
Selection Algorithm: forward
          age anaemia1 creatinine_phosphokinase diabetes1 ejection_fraction high_blood_pressure1 platelets
1  ( 1 )  " " " "      " "                      " "       " "               " "                  " "
2  ( 1 )  " " " "      " "                      " "       "*"               " "                  " "
3  ( 1 )  " " " "      " "                      " "       "*"               " "                  " "
4  ( 1 )  "*" " "      " "                      " "       "*"               " "                  " "
5  ( 1 )  "*" " "      "*"                      " "       "*"               " "                  " "
6  ( 1 )  "*" " "      "*"                      " "       "*"               " "                  " "
7  ( 1 )  "*" " "      "*"                      "*"       "*"               " "                  " "
8  ( 1 )  "*" " "      "*"                      "*"       "*"               " "                  "*"
9  ( 1 )  "*" " "      "*"                      "*"       "*"               "*"                  "*"
10  ( 1 ) "*" "*"      "*"                      "*"       "*"               "*"                  "*"
11  ( 1 ) "*" "*"      "*"                      "*"       "*"               "*"                  "*"
          serum_creatinine sex1 smoking1 time
1  ( 1 )  " "              " "  " "      "*"
2  ( 1 )  " "              " "  " "      "*"
3  ( 1 )  "*"              " "  " "      "*"
4  ( 1 )  "*"              " "  " "      "*"
5  ( 1 )  "*"              " "  " "      "*"
6  ( 1 )  "*"              "*"  " "      "*"
7  ( 1 )  "*"              "*"  " "      "*"
8  ( 1 )  "*"              "*"  " "      "*"
9  ( 1 )  "*"              "*"  " "      "*"
10  ( 1 ) "*"              "*"  " "      "*"
11  ( 1 ) "*"              "*"  "*"      "*"
```

Fig. 16: Backward and Forward variable selection techniques.

From the Fig. 16 we can find that time, age, ejection fraction, and serum creatinine are significant, by the number of asterisks the variable has, which tells how many time that a variable is present in different models. Let’s fit a logistic model with these predictors and compare the performances.

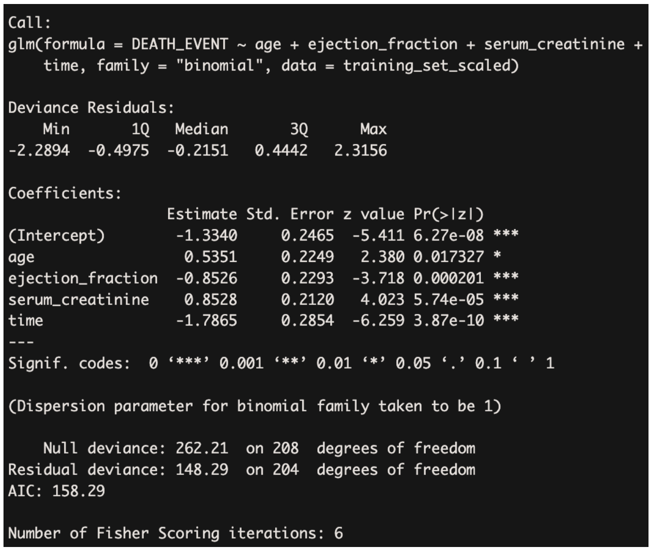
MODEL 4: LOGISTIC REGRESSION – ONLY SIGNIFICANT VARIABLES

UNSCALED

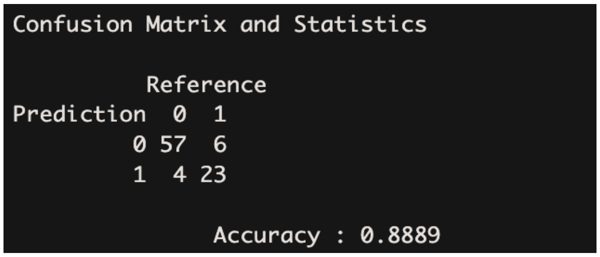


```
Call:
glm(formula = DEATH_EVENT ~ time + ejection_fraction + serum_creatinine +
    age, family = "binomial", data = training_set_unscaled)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9636  -0.6385  -0.2597   0.4775   2.7971

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -0.131529   1.183781  -0.111 0.911530
time              -0.019215   0.003244  -5.923 3.16e-09 ***
ejection_fraction -0.069250   0.018456  -3.752 0.000175 ***
serum_creatinine   0.599730   0.180621   3.320 0.000899 ***
age                0.052946   0.016381   3.232 0.001229 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 262.21  on 208  degrees of freedom
Residual deviance: 165.55  on 204  degrees of freedom
AIC: 175.55

Number of Fisher Scoring iterations: 5
```

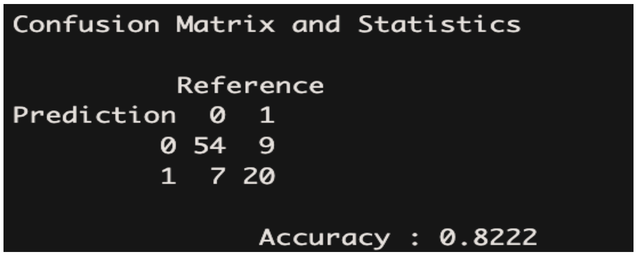
SCALED



```
Call:
glm(formula = DEATH_EVENT ~ age + ejection_fraction + serum_creatinine +
    time, family = "binomial", data = training_set_scaled)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2894  -0.4975  -0.2151   0.4442   2.3156

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.3340     0.2465  -5.411 6.27e-08 ***
age                 0.5351     0.2249   2.380 0.017327 *
ejection_fraction  -0.8526     0.2293  -3.718 0.000201 ***
serum_creatinine    0.8528     0.2120   4.023 5.74e-05 ***
time               -1.7865     0.2854  -6.259 3.87e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 262.21  on 208  degrees of freedom
Residual deviance: 148.29  on 204  degrees of freedom
AIC: 158.29

Number of Fisher Scoring iterations: 6
```



```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 57  6
         1  4 23

               Accuracy : 0.8889
```



```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 54  9
         1  7 20

               Accuracy : 0.8222
```

Fig. 17: Summary and confusion matrix for logistic regression – only significant variables (unscaled & scaled).

Fig. 17, explains that we got the same confusion matrix and the calculations of the model's performance for full variable model and model with only significant predictors. So, the model with 4 significant variables fitted on unscaled data is selected as the best model with an accuracy of 88.89% and recall of 93.44%.

MODEL 5: K-NEAREST NEIGHBORS

Let’s see how the model behaves for all continuous variables and for the significant variables (age, time, ejection fraction, serum creatinine).

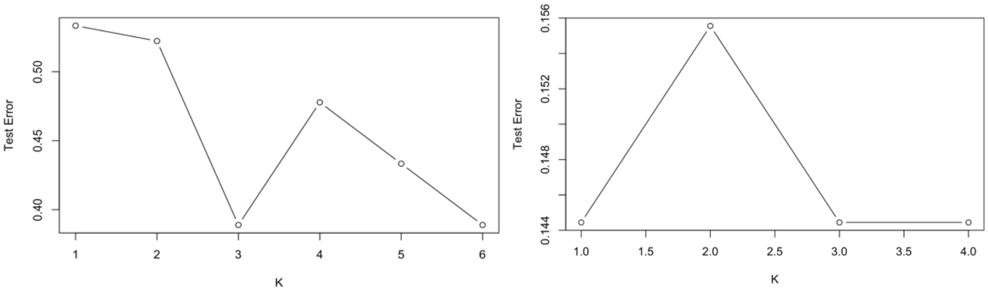
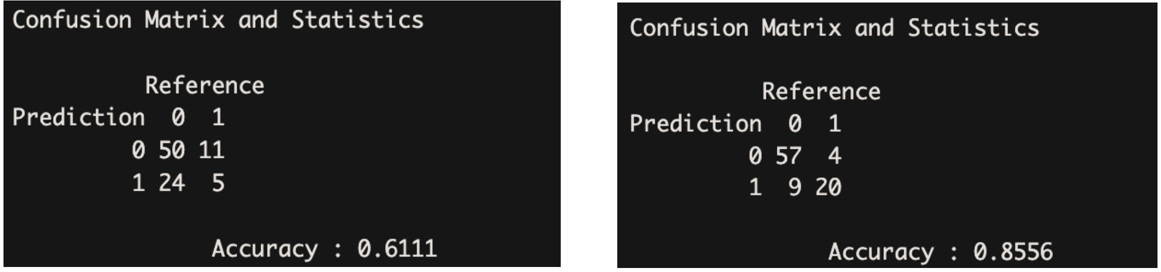


Fig. 18: Finding the best k-value for unscaled and scaled data.

The number of nearest neighbors came out as 3 for all continuous variables and for the significant variables which can be seen in Fig. 18. KNN is now trained and tested with all continuous variables and only significant variables with k=3 respectively.

ALL VARIABLES **SIGNIFICANT (4 VARIABLES)**



```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 50 11
         1 24  5

               Accuracy : 0.6111
```

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 57  4
         1  9 20

               Accuracy : 0.8556
```

Fig. 19: Confusion matrix for KNN (All variables & Significant variables).

We can observe the KNN algorithm performs well on significant data compared with all variables from Fig. 19. So, the model with significant data is selected as best model in KNN with 85.56% accuracy and we will compare it with other models later in this research.

MODEL 6: DECISION TREES

Decision tree algorithm is implemented on both the unscaled and scaled datasets to build a classification model to predict the response i.e., death event based on the set of input features. The decision tree algorithm splits the dataset into smaller subsets based on the most important feature at each step until each terminal node has fewer than some minimum number of observations.

UNSCALED **SCALED**

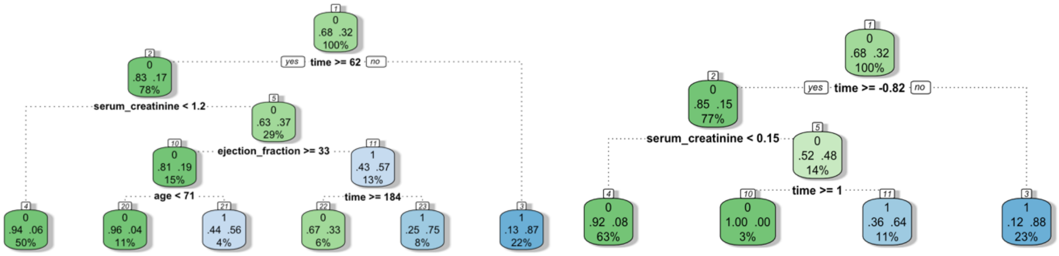
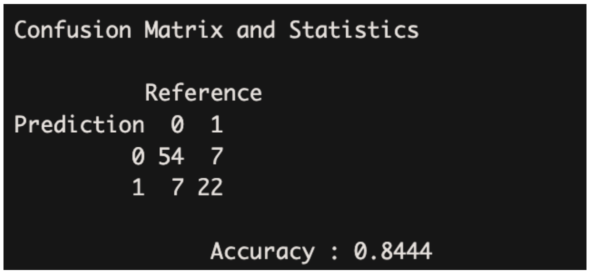


Fig. 20: Decision trees for unscaled and scaled data using rpart package.

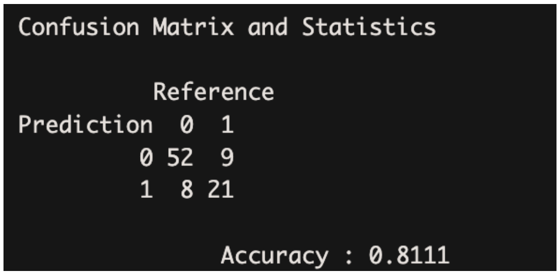
Fig. 20 shows the decision trees for both unscaled and scaled data. We can observe from both the trees that the time, serum creatinine, ejection fraction, and age are the most significant variables which we have done using the variable selection techniques and logistic regression. Let's check it with random forest algorithm using importance function later in the research. The third research question in the research paper can be answered with help of this tree plot.

UNSCALED



```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 54  7
         1  7 22

               Accuracy : 0.8444
```

SCALED



```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 52  9
         1  8 21

               Accuracy : 0.8111
```

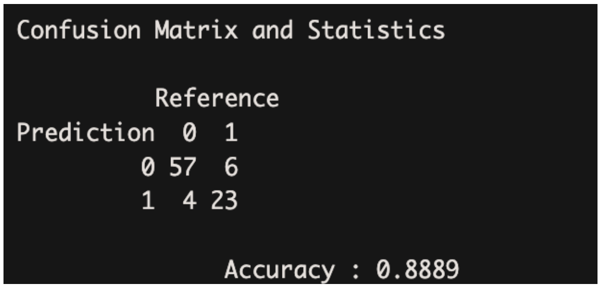
Fig. 21: Confusion matrix for Decision tree (unscaled & scaled).

From Fig. 21, we can observe the Decision tree algorithm performs well on un-scaled data compared to scaled data. So, the model with un-scaled data is selected as best model in Decision tree with 84.4 accuracy and will compare it with other models later in this research. Now let's implement the better version of the decision tree i.e., random forest which builds multiple decision trees and combines them to make predictions. It provides feature importance score, which can help to identify the most important predictors for predicting heart failure.

MODEL 7: RANDOM FOREST ALGORITHM

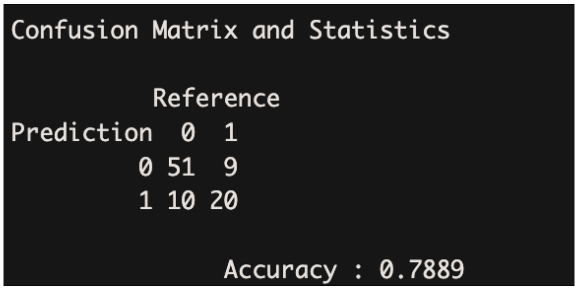
Before fitting the model, the mtry parameter is calculated with out-of-bag error rate and the lower error rate mtry is chosen. It is more recommended than decision trees for Classification tasks to improve the accuracy and reduce the overfitting. Let's fit for both unscaled and scaled data and observe the performances.

UNSCALED



```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 57  6
         1  4 23

               Accuracy : 0.8889
```

SCALED



```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 51  9
         1 10 20

               Accuracy : 0.7889
```

Fig. 22: Confusion matrix for Random Forest (unscaled & scaled).

From Fig. 22, we can observe the random forest algorithm performs well on un-scaled data compared to scaled data. So, the model with un-scaled data is selected as best model in random forest with 88.89 accuracy and will compare it with other models below later in the research.

The importance variables obtained from the random forest model are shown in Fig. 23 below.

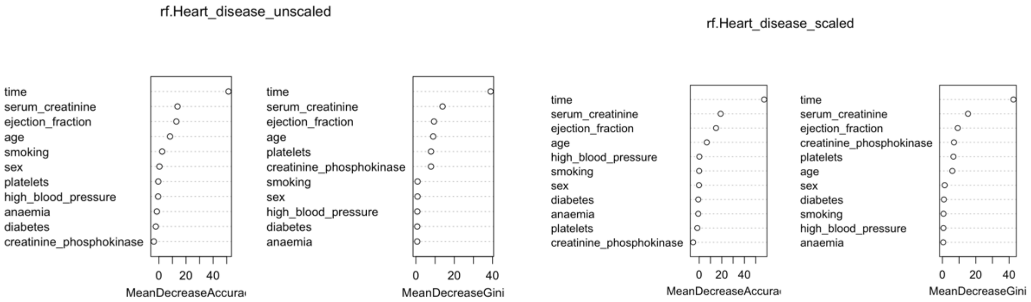
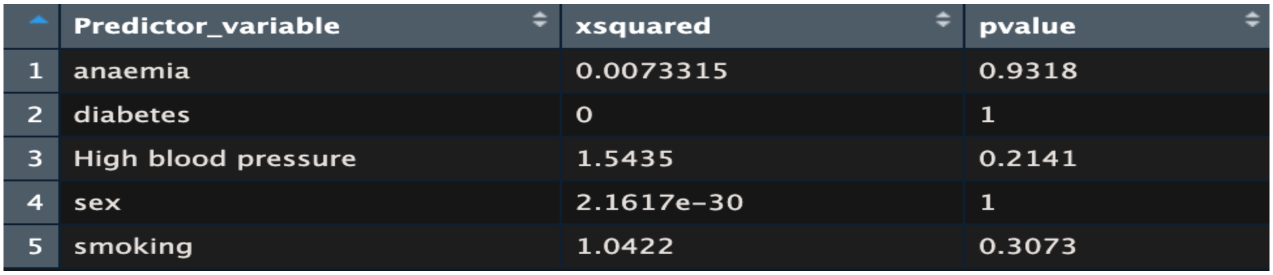


Fig. 23: Important variables from random forest.

As expected, we got time, serum creatinine, ejection fraction and age as first four significant predictors, which we confirmed from decision tree and logistic models. So, from the variable selection techniques, logistic model, top 4 predictors from decision tree and from random forest second research question can be answered.

CHI-SQUARE TEST

Now, as per our research question let's perform chi-square test, it is a statistical hypothesis test that is used to determine whether there is a significant association between two categorical variables. The x-squared value indicates the degree of association and the higher value of x-squared indicates a stronger association between the predictor variable and the target variable.[4] From Fig. 24, we can observe that for smoking and DEATH_EVENT, we have a p value of 0.3 which is very much more than 0.05 and so we don't reject the null hypothesis. These variables are not significantly associated. This is the same for all the other variables showed below, we don't have good association with any of the variables. Every variable has their p-value >0.05. If the x-squared value is high, we can reject the null hypothesis and can state they are associated. But, in this case, we have all the x-squared values less and the p-value is also greater than 0.05. So, there is no association between the variables.



	Predictor_variable	xsquared	pvalue
1	anaemia	0.0073315	0.9318
2	diabetes	0	1
3	High blood pressure	1.5435	0.2141
4	sex	2.1617e-30	1
5	smoking	1.0422	0.3073

Fig. 24: Chi-square test between categorical predictors and death event.

PREDICTION OF TIME TO DEATH BASED ON ALL CLINICAL FEATUERS

We have fitted a linear model having time as the predictor variable and have got an adjusted R-squared value of 0.086. We have got a very low value with which we can interpret that the time to death cannot be predicted. We got a residual standard error of 74.2 which shows the average difference between observed values and predicted values. We can find decent correlation with age, serum creatinine, anaemia, high blood pressure but still we have a low accuracy as well. With this, we can conclude that the time to death cannot be predicted with clinical features.

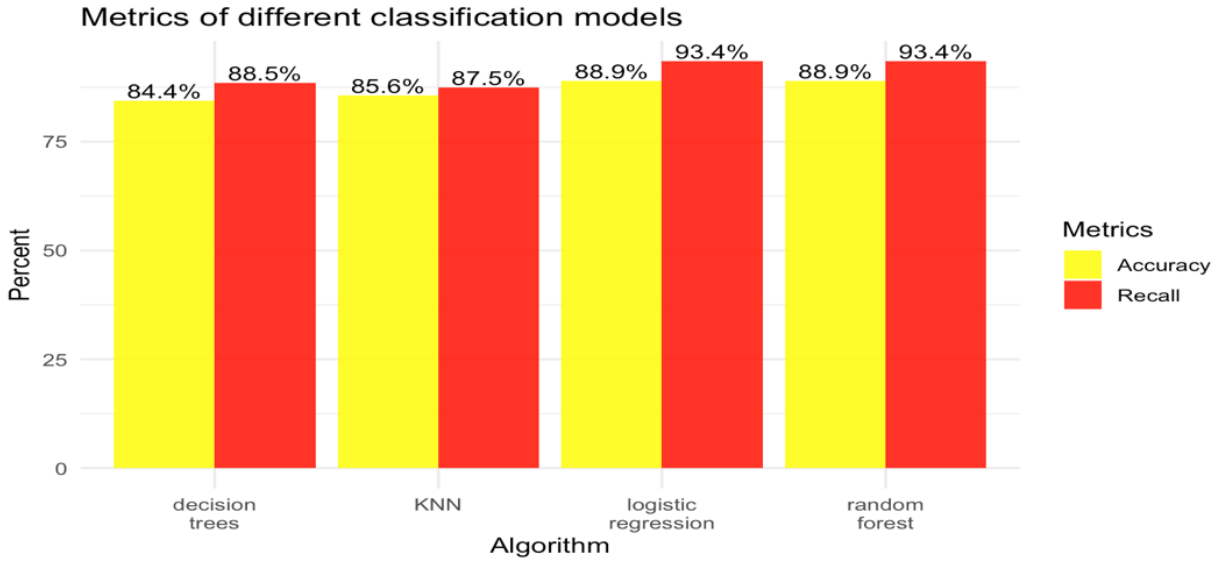
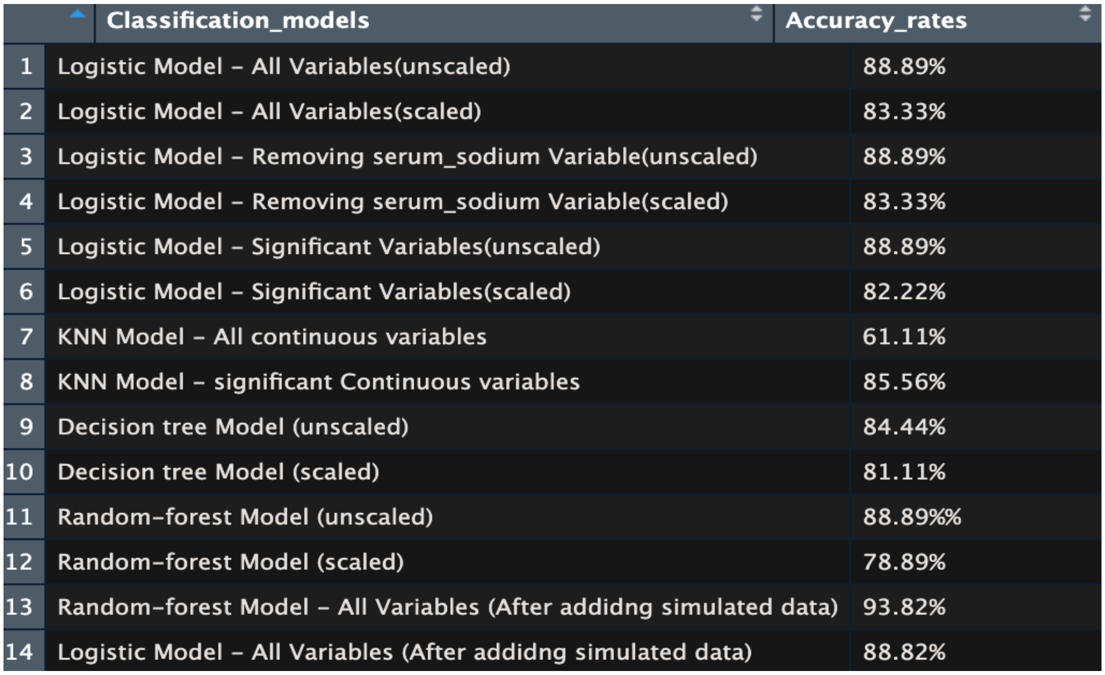


Fig. 25: Accuracy and recall comparison on different classification models

From the fig. 25, We can see that the random forest and logistic regression show similar recall and accuracy and they are the highest having an accuracy of 88.9% followed by the KNN with accuracy of 85.6% and the model that gave low accuracy compared to other three classification models is decision tree model with an accuracy of 84.4%. Let's check whether we have any changes with the accuracy by adding simulated data from existing data using mean and standard deviation on logistic and random forest models as they have better accuracy rates compared with other two models. And see, how these models behave when simulated data is fitted.

SIMULATION OF THE DATA BASED ON EXISTING DATA

Using random normal distributions with mean and standard deviation values taken from the original heart disease dataset. We will create a new dataset of 1400 observations with the same columns as the original, but with different values generated by the random distributions on existing data. The new dataset can be used for testing or training machine learning models without having to collect new data. The simulated data is sliced into 80% training and fitted with the best model we concluded which is random forest followed by logistic regression. From the Fig. 28 we can observe random forest model is fitted with 93.82% accuracy on simulated data and logistic with 88.82%. The accuracy is increased with the simulated data for random forest algorithm.[5]



	Classification_models	Accuracy_rates
1	Logistic Model – All Variables(unscaled)	88.89%
2	Logistic Model – All Variables(scaled)	83.33%
3	Logistic Model – Removing serum_sodium Variable(unscaled)	88.89%
4	Logistic Model – Removing serum_sodium Variable(scaled)	83.33%
5	Logistic Model – Significant Variables(unscaled)	88.89%
6	Logistic Model – Significant Variables(scaled)	82.22%
7	KNN Model – All continuous variables	61.11%
8	KNN Model – significant Continuous variables	85.56%
9	Decision tree Model (unscaled)	84.44%
10	Decision tree Model (scaled)	81.11%
11	Random-forest Model (unscaled)	88.89%%
12	Random-forest Model (scaled)	78.89%
13	Random-forest Model – All Variables (After addidng simulated data)	93.82%
14	Logistic Model – All Variables (After addidng simulated data)	88.82%

Fig. 26: Accuracy of all the models fitted in this research.

From the Fig. 26 we can see the tabulated accuracies of all the models with scaled, unscaled, and simulated data. The random forest model is the best fit for the unscaled dataset before simulating. Also, we can see that after adding the simulated data to the original data, random forest performs very well and gets an accuracy of 93.82% which is better than the previous selected model.

IV. CONCLUSION

To conclude, we can predict death of an individual affected with heart failure by accepting certain inputs from the patient with the help of classification algorithms in which, we got random forest model as the best fit for the dataset. With this algorithm we did not face any overfitting and underfitting issues. We used various methods to select the significant variables. The methods are forward and backward selection techniques, logistic regression, decision tree and random forest algorithms. We got the same significant variables with all the methods that are age, ejection fraction, serum creatinine, time.

The simulated data showed a higher accuracy of 93.82 with the random forest model. This shows that the accuracy improved with the simulated data when compared to the dataset. By comparing the logistic regression and random forest models for the simulated data, we found out that the accuracy for the random forest model increased while the accuracy of logistic model decreased. By observing the chi-square results, we found out that there is no association between any of the categorical variables with the death event.

From the linear regression, we found out that we did not not have any proper significance between any of the variables and follow-up period. We got very low adjusted R-squared value of 0.086.

V. IMPLICATIONS AND FUTURE WORK

The implications of this research are significant, as they highlight the importance of lifestyle modifications and interventions in reducing the mortality rate among patients with cardiovascular disease. Healthcare professionals can explicitly target these risk factors in their treatment plans and treatments by identifying the variables that are most closely linked to mortality. The dataset could be further examined to find patient subgroups with particular heart failure risk factors. The creation of focused therapies to lower the risk of heart failure in these populations may result from this.

The future work for this research would be collecting more data possible from an API linked with any hospitals or from any external resources. Feature engineering can be performed to extract additional features from the existing data which can be more significant for prediction of heart diseases. For this we need to study good amount of data and need to have many predictors related to heart and their functioning. The other extension for this research can be implementing this predictive analysis on a predictive tool or on application which can be used by hospitals, so that they can predict how long will a patient be staying for higher chances of his survival. This can give them the clear idea of vacancies of the beds and the treatments they need to do for a patient.

REFERENCES CITED

[1] W. H. Organization, "Cardiovascular diseases (CVDs)," [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[2] UCI, "Heart failure clinical records Data Set," [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records.

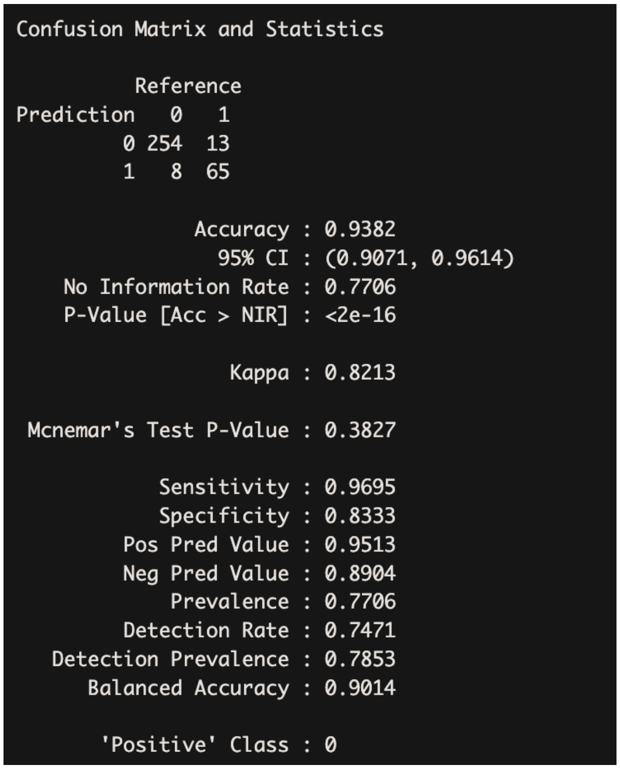
[3] Investopedia, "Skewness: Positively and Negatively Skewed Defined With Formula," [Online]. Available: https://www.investopedia.com/terms/s/skewness.asp.

[4] D. Flair, "Chi-Square Test in R," [Online]. Available: https://data-flair.training/blogs/chi-square-test-in-r/.

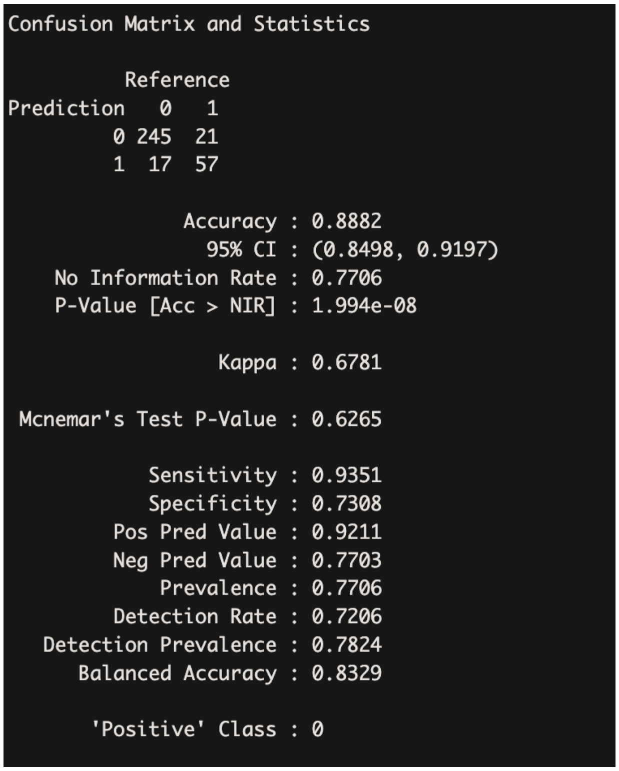
[5] A. Muldoon, "Getting started simulating data in R," [Online]. Available: https://aosmith.rbind.io/2018/08/29/getting-started-simulating-data/#.

APPENDIX

PERFOMANCES OF MODELS AFTER ADDING SIMULATED DATA



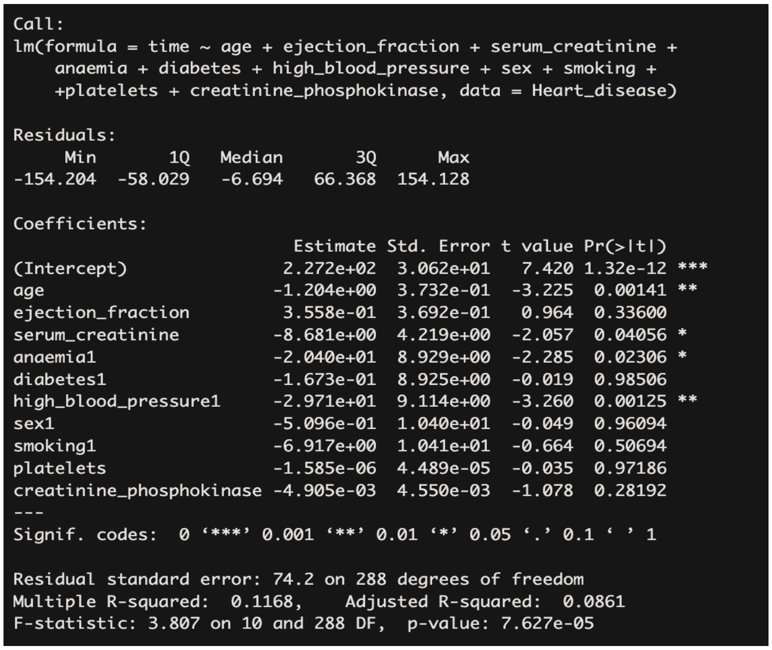
```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 254  13
         1   8  65

               Accuracy : 0.9382
                 95% CI : (0.9071, 0.9614)
    No Information Rate : 0.7706
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.8213

 Mcnemar's Test P-Value : 0.3827

            Sensitivity : 0.9695
            Specificity : 0.8333
         Pos Pred Value : 0.9513
         Neg Pred Value : 0.8904
             Prevalence : 0.7706
         Detection Rate : 0.7471
   Detection Prevalence : 0.7853
      Balanced Accuracy : 0.9014

       'Positive' Class : 0
```



```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 245  21
         1  17  57

               Accuracy : 0.8882
                 95% CI : (0.8498, 0.9197)
    No Information Rate : 0.7706
    P-Value [Acc > NIR] : 1.994e-08

                  Kappa : 0.6781

 Mcnemar's Test P-Value : 0.6265

            Sensitivity : 0.9351
            Specificity : 0.7308
         Pos Pred Value : 0.9211
         Neg Pred Value : 0.7703
             Prevalence : 0.7706
         Detection Rate : 0.7206
   Detection Prevalence : 0.7824
      Balanced Accuracy : 0.8329

       'Positive' Class : 0
```

Confusion matrix of random forest and logistic regression on simulated data.

RESULTS OF MULTIPLE LINEAR REGRESSION ON TIME (FOLLOW-UP DAYS)



```
Call:
lm(formula = time ~ age + ejection_fraction + serum_creatinine +
    anaemia + diabetes + high_blood_pressure + sex + smoking +
    +platelets + creatinine_phosphokinase, data = Heart_disease)

Residuals:
     Min       1Q   Median       3Q      Max
-154.204  -58.029   -6.694   66.368  154.128

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              2.272e+02  3.062e+01   7.420 1.32e-12 ***
age                     -1.204e+00  3.732e-01  -3.225  0.00141 **
ejection_fraction        3.558e-01  3.692e-01   0.964  0.33600
serum_creatinine        -8.681e+00  4.219e+00  -2.057  0.04056 *
anaemia1                -2.040e+01  8.929e+00  -2.285  0.02306 *
diabetes1               -1.673e-01  8.925e+00  -0.019  0.98506
high_blood_pressure1    -2.971e+01  9.114e+00  -3.260  0.00125 **
sex1                    -5.096e-01  1.040e+01  -0.049  0.96094
smoking1                -6.917e+00  1.041e+01  -0.664  0.50694
platelets               -1.585e-06  4.489e-05  -0.035  0.97186
creatinine_phosphokinase -4.905e-03  4.550e-03  -1.078  0.28192
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 74.2 on 288 degrees of freedom
Multiple R-squared:  0.1168,    Adjusted R-squared:  0.0861
F-statistic: 3.807 on 10 and 288 DF,  p-value: 7.627e-05
```