Introduction –

The heart failure clinical records dataset contains medical records of 299 patients who had a heart failure and the data was gathered during their follow up period which contains 13 attributes such as Age, anaemia, High blood Pressure, Diabetes, Ejection Fraction, platelets, Sex, Serum Creatinine, Serum Sodium, Smoking and time (which was the follow up period in days), and the death event which tells us if a patient was deceased during the follow up period. All the values in our data set was integers and most of them were binary which made our explanatory analysis part less complicated since there was no need to convert any categorical values to build our models. In this report, we will begin by exploring our data set visually which will illustrate all the attributes responsible for a death event. We implemented machine learning models and data mining strategies such as classification trees and Stepwise Logistic Regression that can detect and predict heart attacks using the patient's medical record.

Explanatory Data Analysis -

The data sets were split into 80% training set and 20 % testing sets. Once our models were built, Cross validation was performed to compare the methods and we compared the performance and accuracy of all the models. Our EDA is done using python 3 and our modelling is done with Rstudio.

We can begin our Explanatory Data Analysis to gain insights on the correlation of various factors that would cause a heart attack/ death. This will help us to analyse our data before making any assumptions for our models. Our data set contains 299 rows and 13 Columns, and we have 3 types of float and 10 integer values and no categorical values to convert as mentioned earlier.

Missing values in our data set – None

Duplicated values in our data set – None

Next, we obtained the mean values to determine the central value of all our attributes and they are as follows –

Table 1. Mean distribution of every variable in the dataset

Attributes	Mean
age	60.833893
anaemia	0.431438
creatinine_ phosphokinase	581.839465
diabetes	0.418060
ejection_fraction	38.083612
high_blood_pressure	0.351171
platelets	263358.029264
serum_creatinine	1.39388
serum_sodium	136.625418
sex	0.648829
smoking	0.32107
time	130.260870
DEATH_EVENT	0.32107

Now that we have our mean values, we can visualize our data to find the relationship between all the variables with Death_event. This analysis will separate each attribute individually and explore it's relationship with our target variable Death_ event. Followed by that it will help us to understand this process and make it less complicated to find out the relationship between the variables with help of a correlation matrix.

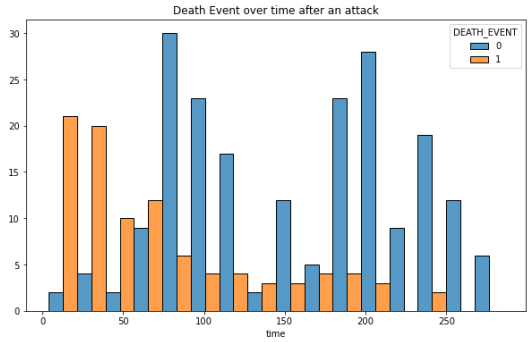


Figure 1. Death event over time in days.

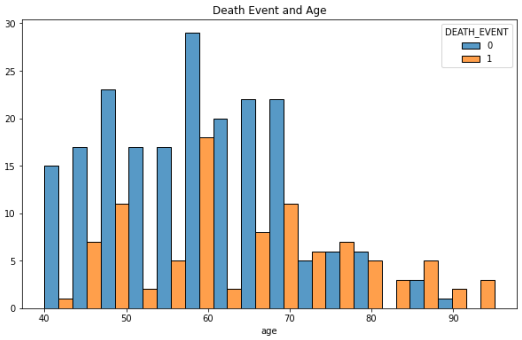


Figure 2. Death event with age.

In (figure.1), the death event over time after a heart failure is plotted on a graph and we can tell that the chances of a patient dying is relatively high right after a heart failure (first two months period) and later on the chances of them surviving is high. In the above chart, 0 is alive and 1 is death. On the other hand, we checked the survival chances with age and (figure.2) illustrates that as a patient gets more and more older the survival rate is very low for patients and it becomes important for them to take necessary steps to prevent heart failure.

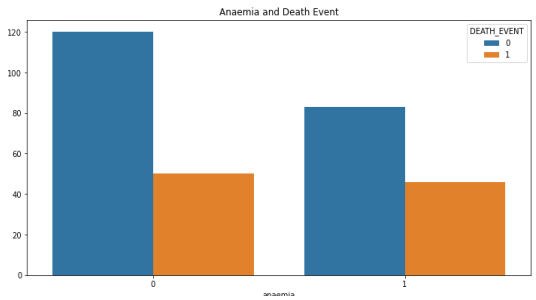


Figure 3.anaemia and death event

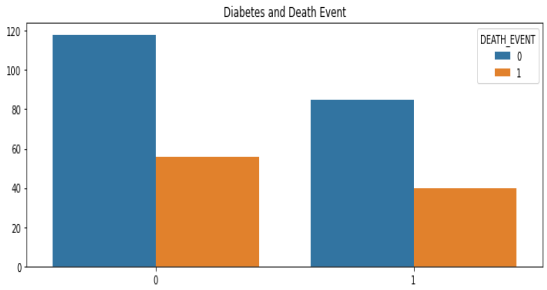


Figure 4.Diabetes and death event

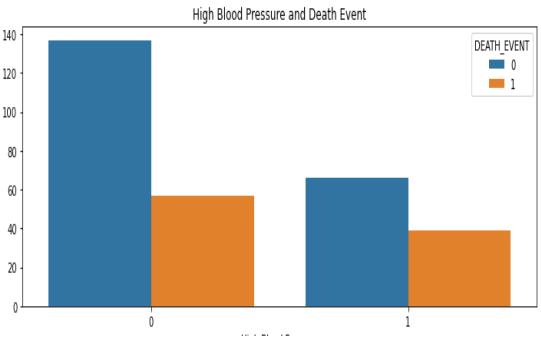


Figure 6. High blood pressure and death event

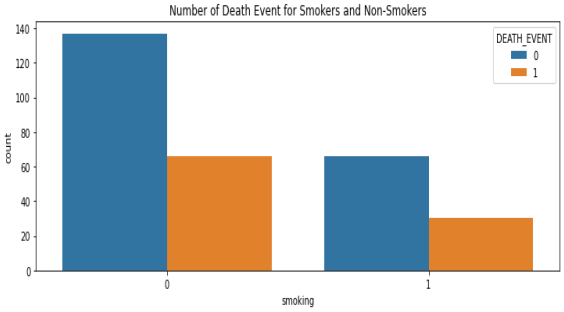


Figure 5. smoking and death event

In all the figures above 0 represents a patient alive and 1 represents a patient deceased. Starting off with (figure 3), people with (level is below 40) and without anaemia (above 40), it says that increase in anaemia contributes to a patient dying after the heart failure regardless of whether or not they have low haemoglobin. In (figure 4) there were more people who did not have diabetes died (death level is between 40 and 60) compared to the ones that actually had diabetes (Death level is below 40). Similarly, in (figure 5) it illustrates that the death rate is almost double in non- smokers compared to smokers. In (figure 6), the death event was higher for people who did not have high blood pressure compared to the ones that did not have high blood pressure. From the above all, we can conclude that death event takes place regardless of a patient having a high level of any attributes like smoking high blood pressure, anaemia and diabetes. These seem to be bad classifiers.

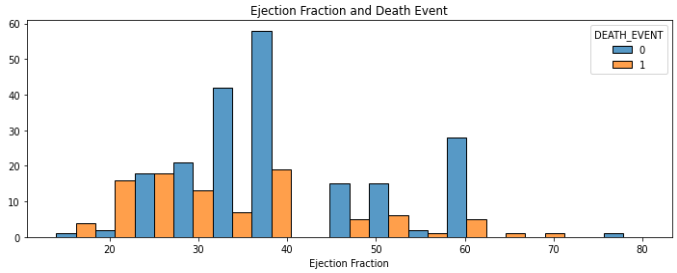


Figure 8. Ejection fraction vs death event

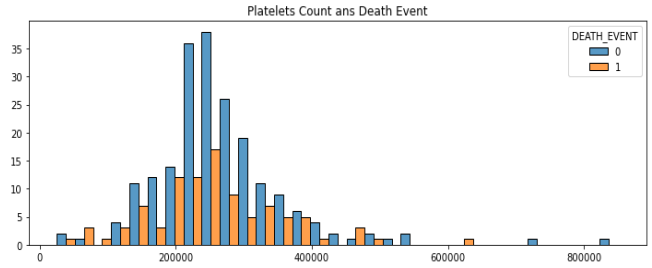


Figure 7. platelets count vs death event

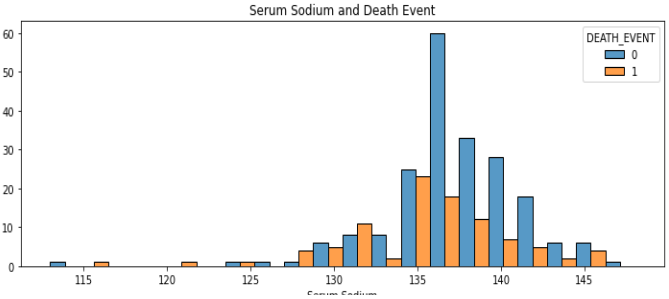


Figure 9. Sodium vs death event

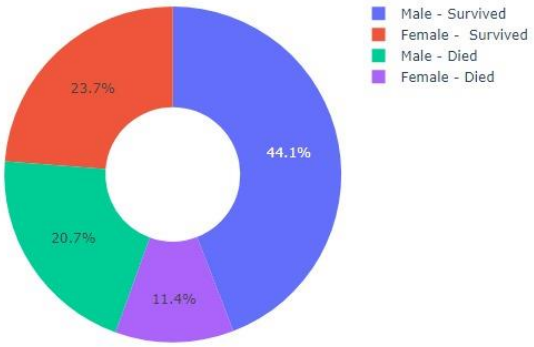
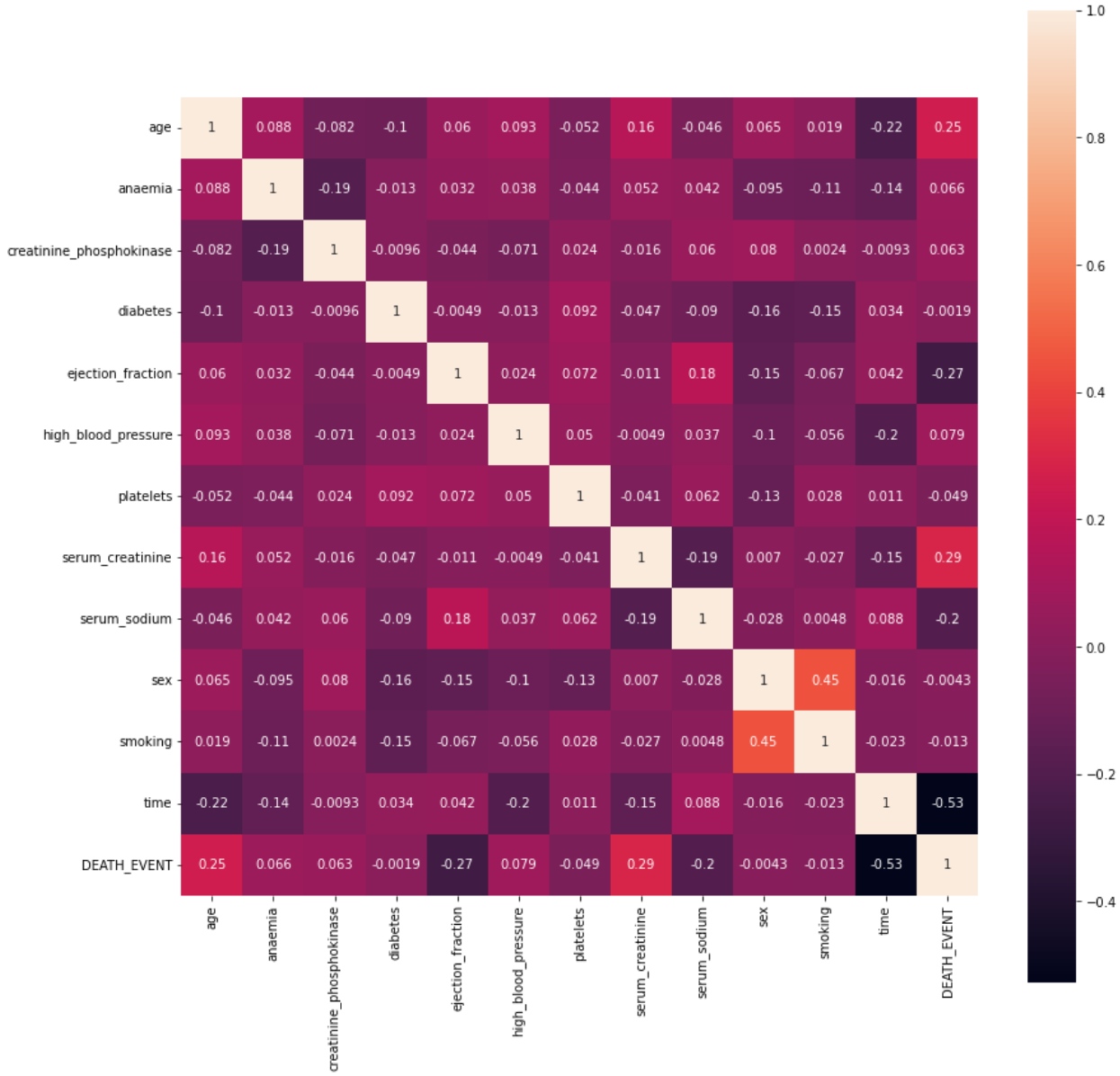


Figure 10. Gender distribution with death event

In (figure 7) it shows us that patients having a large ejection fraction can increase the chance of survival after a heart failure. (figure 8) shows us that both low and high platelets count seems to be contributing to a patient to decease after a heart failure episode. (figure 9) illustrates that having a low – medium sodium rates will reduce the chance of a patient dying after a heart failure. Finally, we distributed the data according a patient's gender and found out that men were more prone to dying after a heart failure in this study and female had the lowest death percentage (11.4%). Although, men also had the highest surviving rate with (44.1%) while female only had (23.7%). Our analysis suggests that Ejection fraction, Platelets, Serum sodium and sex seems to be good classifiers in our analysis. The last step to conclude our EDA is by plotting all the variables on a heatmap to find the correlation between all our variables. It will also act as a good indicator when the relationship is linear. The correlation matrix can be found below –



The factors that seem to have the highest correlation with death events seems to be serum_creatinine and age. While visualizing all the variables we assumed that platelet count will have a higher survival rate bur in this matrix the correlation score seems to be low. All the other variables like ejection fraction and sodium which we assumed as good classifiers seems to have negative correlation with Death_event. Time seems to be the least correlating factor out of all of them. Therefore, we cannot make any specific assumptions when it comes to all the variables and let our models predict the death event. The correlation table can be found below for reference –

Table 2. Correlation values of all variables

Variables	Correlation
Death event	1.000000
serum_creatinine	0.294278
age	0.253729
high_blood_pressure	0.079351
anaemia	0.066270
creatinine_phosphokinase	0.062728
diabetes	-0.001943
sex	-0.004316
smoking	-0.012623
platelets	-0.049139
serum_sodium	-0.195204
ejection_fraction	-0.268603
time	-0.526964

Analysis and modelling –

As mentioned in the introduction we will be using Classification trees and Stepwise Logistic Regression to build our model. Our primary goal is to see how DEATH_EVENT is influenced by rest of the variables. Since DEATH_EVENT is a binary variable, it becomes easier to build our classification tree. We will build our tree in R and the first step before we start building is to convert the numeric values into factors. Next, we split the training and testing into 80:20, and the reason I chose to split it this way is because it leaves enough data for the model to train on before we test it. After we established our training and testing sets, we constructed a tree which can be found below –

Classification tree -

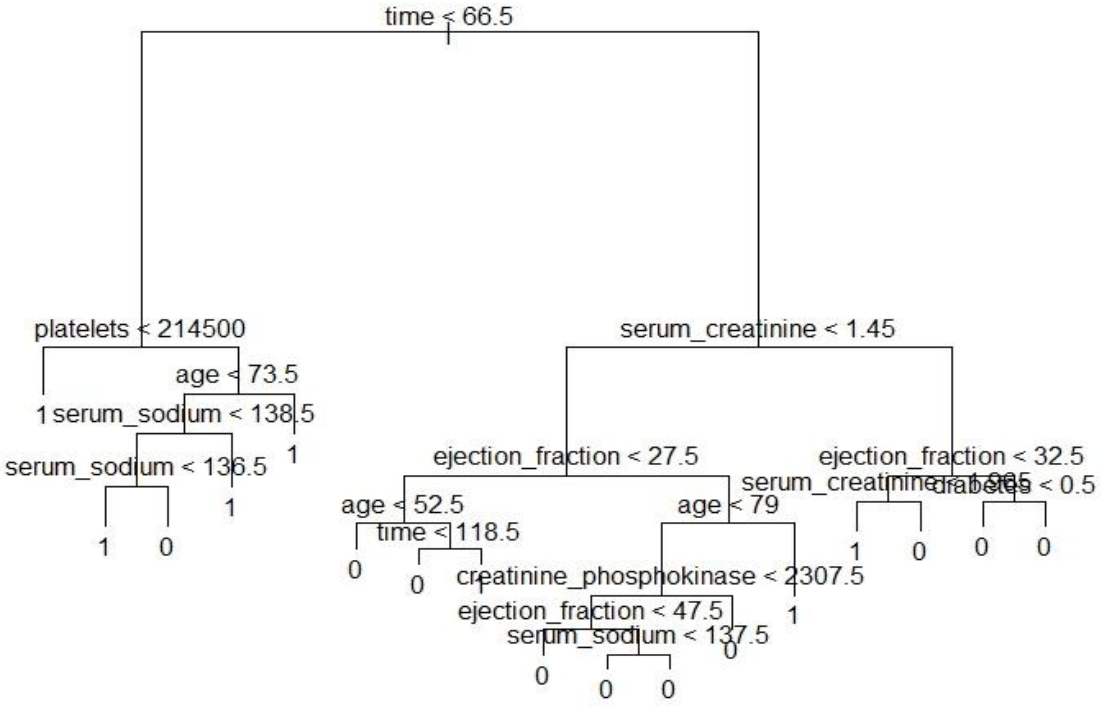


Figure 11. Classification tree heart failure prediction

The variables that were used in the tree are time, platelets, age, serum_sodium, serum_creatine, ejection_fraction, creatinine_phosphokinase, and diabetes. Looks like this tree was grown to full depth and it might be overfitting. Our next choice would be to prune this tree. Also, we want the classification error rate for us to guid the cross validation and the pruning process. The confusion matrix tells us that there is very high false negative rate but also a good false positive rate. We also have to find the length of our tree. By running a K- fold cross validation we can evaluate our model's ability to find the total number of misclassifications. During this process it was clear that the misclassification rate was same when the tree had 2 or 4 terminal nodes. So we can set the length of the tree to 4. The misclassification error seems to be 0.194 before we prune our tree. Below is graph we got from plotting our step function by using the argument FUN=prune.misclass in order for our classification error rate to guide the pruning process.

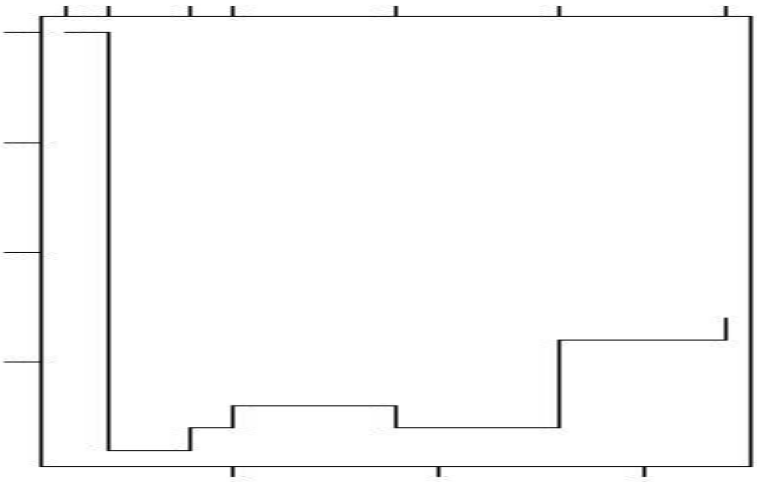


Figure 12. Finding out the depth of the tree.

<u>Pruning the tree-</u>

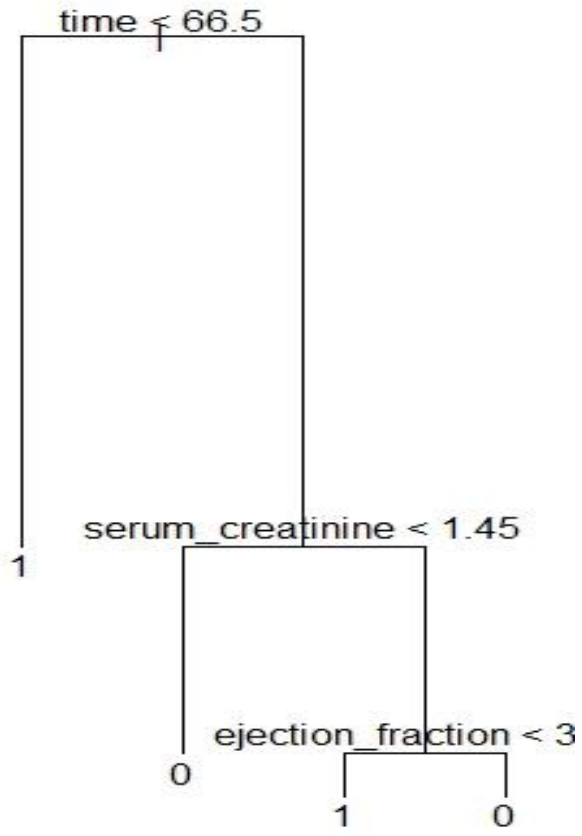


Figure 13. plotting the pruned tree

We can understand that the root node time has a threshold of 66.5 says that the risk of a patient being deceased is high in the first two months and the bottom nodes serum Creatinine is the decision node and ejection fraction is the terminal node. We can understand that serum Creatinine and ejection fractions are two main causes of a patient's death after their heart failure.

The post pruning process has shown us a major improvement in the misclassification error rare since it has come down to 0.088. When looking at our confusion matrix this proves that after pruning has reduced our overfitted tree. The misclassification rate of 0.0909 vs 0.1472 for unpruned vs pruned, is likely because that this function returns the tree training error; training error tends to be lower when a model is having been overfitted - this will be the case more-so with an unpruned tree compared to a pruned tree. Finally, we can apply bagging and random forest to this data. The number of trees are 10, the number of variables that were tried at each split is 12, and the out of bag estimate is 27% (approx.).

Confusion matrix –

	0	1
0	124	32
1	29	41

Accuracy – 0.73 (73 %)

Precision – 0.79 (79%)

Misclassification – 0.21 (21%)

Recall – 0.81(81%)

Stepwise Logistic Regression –

This model would be useful for predictions because it helps to add and remove variables only based on the t- values of the estimated coefficients. By using R studio we can find out the important predictors that will serve us to predict the death event of a patient based on their heart failure records. We will be plotting the Receiver Operating Characteristic curve (ROC) to find out the performance of our model.

As always, we begin our analysis by splitting our data set into training and testing data and we can do a random sampling on the target variable DEATH_EVENT. We will split our data by 80% training and 20% for testing dataset. Our regression has two different models null model and full model where one has no variables and other will possess all variables. By using glm() function we can repeat all possible pairs of predictors into this model. Finally, both models are going to be initiated using the step() function for the stepwise process with a forward direction.

Our model started will all 12 variables first and in the first step it picked up time followed by ejection_fraction, serum_creatinine, age, , sex and serum_sodium. On the other hand, Diaibetes, smoking, platelets, anaemia, creatinine_phosphohinase and high blood pressure were dropped which we can assume that are bad predictors for our model.

Our formula looks like this - glm(formula = DEATH_EVENT ~ time + ejection_fraction + serum_creatinine + age + serum_sodium, family = "binomial", data = heart)

The stepwise regression has generated the best predictors based on Akaike Information Criterion(AIC). The AIC values of each step can be found below –

	AIC
Step 1	377.35
Step 2	283.07
Step 3	262.08
Step 4	243.41
Step 5	236.3
step 6	235.49

After printing their summary, we can find the estimate and standard error our coefficients. serum_creatinine and age are the only values with positive estimate regression while time, ejection_fraction and serum_sodium has negative estimates and the final AIC values seems to be 235.49 and the fishing scoring iterations is 6. Time, age and serum sodium seems to have the lowest standard error of 0.002, 0.015, and 0.038.

Now that our model is built we can proceed to the testing phase with the same data set to predict the death event of a patient. We can classify the binary events based on the probabilities that are generated and we will be using 0.5(50%) as the rate of the binary events which means anything above 50% will be evaluated as a death event. After we have classified, we can produce the confusion matrix to check our accuracy and predictions.

Confusion matrix –

	0	1
0	37	1
1	3	18

From the confusion matrix we can tell 55 out of 59 observations are predicted accurately and this model seems to be good at predicting the number of people who will be alive which is 37 and the number people who will be dead which is 18.

Accuracy – 0.93 (93 percent)

Precision – 0.97(97%)

Misclassification - 0.06

Recall – 0.92 (92%)

Roc curve –

To produce the specificity and sensitivity of a model we can use the ROC curve and later plot the regression of DEATH_EVENT. This curve will also help us to distinguishing between classes

The area under the curve (AUC) - **0. 0.9362**

In (figure.14), Since our curve seems to be high this tells us that our model is good at predicting between patients and death event. Since our value is near to 1 it has a good measure of separability. The ROC curve can be found below -

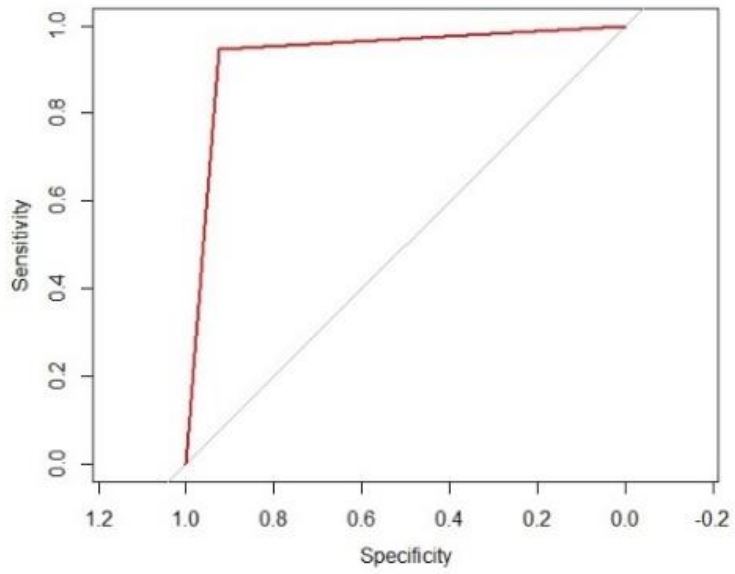


Figure 14. ROC - AUC curve

Regression Plot –

In our plot below (figure 15) we can visualize the residuals against the line and there are a lot of patients who are alive than deceased patients. As you can see patients with probability above 50 are deceased. The plot clearly shows us the relationship between the rank and probability of death event., Moreover, this seems to be a pretty good model to predict the DEATH_EVENT of the patients in our data set.

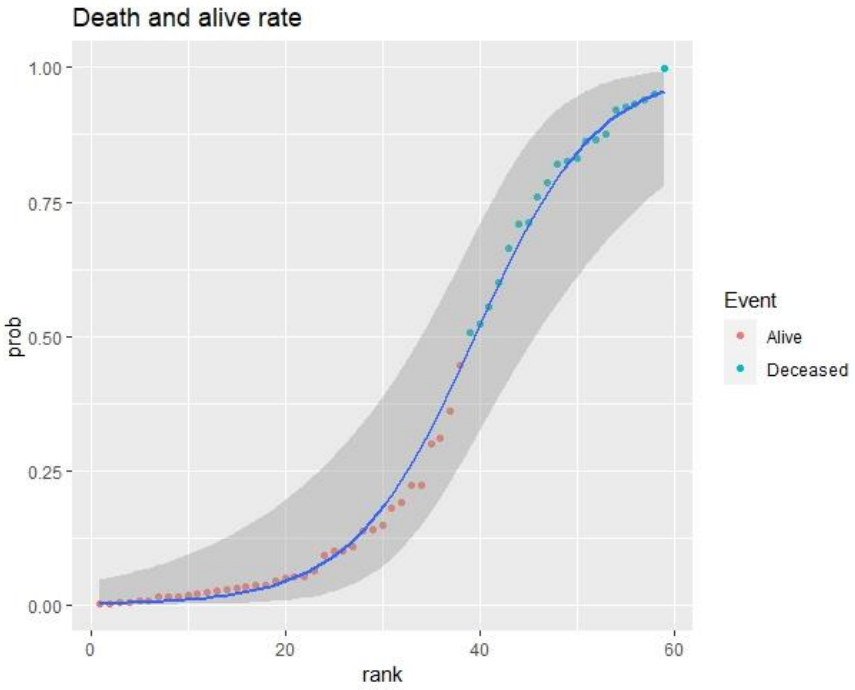


Figure 15. Regression plot of alive and dead patients.

Model Comparison –

Table 3. Model comparison Table

Model name	Accuracy	Precision	Misclassification rate	Recall
Classification tree	**0.73 (73 %)**	**0.79 (79%)**	**0.21 (21%)**	**0.81(81%)**
Stepwise logistic regression	**0.93 (93 percent)**	**0.97(97%)**	**0.06**	**0.92 (92%)**

Both our models were split the same way 80%training and 20% testing. Having Compared both from reduced to full models, it is evident that our stepwise logistic regression model outperformed and has the better accuracy rate (93%). Even the precision score of our regression model (97%) was higher compared to our classification tree (79%). All our models were mostly measured using the confusion matrix and Roc curve for the regression model. It is fair to say that stepwise with forward direction can be used even when extra data is added and the accuracy score will remain higher than the classification tree. Hence, we chose simple models to and compared the results of both of them. In conclusion, stepwise regression model will be the preferred model for the heart failure prediction data set and is much preferred when we have to find the death event of a patient.

Discussions-

Two models were developed to predict death after a heart failure, and both models demonstrated acceptable sensitivity and specificity. Although stepwise regression had the better accuracy rate, our classification tree was also significantly decent. Applying machine learning strategies to the health sector seems to be very useful as even simple models that were used in this report did a fairly good job in identifying the good predictors and death rate with just the details of a patient during the follow up period. Although we weren't able to implement deep learning techniques such as Neural networks like CNN and DNN, it would have been very useful for our analysis. We were not able to implement neural networks because our dataset was not big enough and it was not a very high dimensional dataset to train it. One thing we were particular about whilst building our models is improve the efficiency to predict death because after a heart failure it has many after effects that would carry minor or major risk factors.

Having more predictors like symptoms patients had after a heart failure or other clinical parameters would have helped to improve the accuracy of both our models to predict the death event of a patient. The data we used in our models were small and having more patients and their history would have helped us during the training process of our models. Most influential predictors in both the models were Serum_creatinine, age and ejection fraction. Stepwise logistic regression also happened to have the lowest misclassification rate compared to the classification tree. During our modelling process, we wanted a model which fits data well and not to be too complex. In our EDA we found out that some predictors like Days, Age, Ejection_fraction, platelets and sodium were very useful but both our models had serum _creatinine, time and ejection fraction in common and were very influential predictors. In general, we can say that the predicted outcomes are almost the same as the actual outcomes. Only our classification tree seemed to be overfitting before the pruning

process which is one of the reasons why it had a higher misclassification rate and lacked in accuracy. Our training datasets were trained using numerous machine learning algorithms on this data set to make it an efficient and effective model in terms of recall.

Summary –

Heart failure has become a major issue because of all the easy access people have for fast food and other harmful habits. Looking at our dataset of people who had a heart failure, we have noticed fluctuating levels of high blood pressure, anaemia and diabetes in several people in the beginning of our analysis. But we observed that people are more prone to dying in the first two months of a heart failure and people should take an effective approach in maintaining the level serum creatinine in their blood, and doctors have to be more observant of the percentage of blood leaving the heart at each contraction (Ejection fraction) in the first two months after a heart failure. All our models built with complex machine learning algorithms concluded that the above factors are extremely important for someone who suffered a heart failure. We need to integrate more risk/death prediction models into the health care system which could save millions of lives.