Survival Prediction of Heart Failure Patients

Avneet Kaur
MSc Mathematics

December 15, 2022

Abstract

Heart failure is a serious problem becoming common in adults in addition to elder people. In this project, a data set containing clinical records of patients with heart failure were analyzed. Different machine learning techniques like k-nearest neighbours, logistic regression, classification tree, bagging, random forests, and boosting were implemented to predict the death of patients. Feature ranking was performed to find the most significant predictors. The results showed that boosting with time, creatinine phosphokinase, serum creatinine, and ejection fraction as predictors outperformed.

Contents

1 Introduction

Cardiovascular diseases are a group of problems relating to the heart. The main cause of heart-related diseases is a build-up of plaque in the arteries which hinders the flow of blood. As per WHO, heart-related diseases are the leading cause of death globally, taking an estimated 17.9 million lives each year [3]. In figure 1, the left image shows the top causes of death in the US in 2015 as reported by the Centers for Disease Control and Prevention [8]. It can be seen that heart disease tops in contributing to the total deaths when compared to other diseases with a 23.4% share in the total. In fact, Cancer despite being such a deadly disease was second on the list. Additionally, as per CDC, Heart failure cost the nation an estimated $30.7 billion in 2012 [9]. The right image in Figure

1 shows the proportion of deaths due to various diseases in Canada in 2019 as reported by the Canadian Cancer society. It can be seen that Heart disease was the second leading cause of death in Canada with a contribution of 18.5%.

The main factors leading to heart diseases as reported by various health organisations include unhealthy diet, physical inactivity, use of tobacco, and alcohol consumption. The wrong lifestyle choices manifest as high blood pressure, high blood glucose, high blood lipids, obesity, etc in the human body. The most common Cardiovascular diseases are Heart attack, Stroke, Heart failure, Arrhythmia, and Heart Valve Complications. Particularly talking about heart failure, as per CDC, it is a condition in which the heart is unable to pump enough blood and oxygen to other parts of the body. But it does not imply that the heart has completely stopped working [9]. This means that through proper treatment lives of heart failure patients can be saved. To give proper treatment to patients, doctors need to understand the kind of chemical and biological changes the body goes through during heart failure. To make informed choices, Machine learning models can be developed to analyse the data and make predictions on the survival of heart failure patients.

This problem has been studied by Ahmad et al [1] in 2017. The authors used biostatistical time-dependent models to analyse this data and predict death events. As an extension to this work, Chicco et al [4] applied Machine learning techniques to predict the death event.

The aim of this project was to find the best model to predict the survival of heart failure patients. Additionally, with the help of these models, the features that play a significant role in determining survivals chances of heart failure patients were ranked. Feature ranking would help health practitioners to prioritize the factors they should work on to reduce the chances of death from heart failure.

In this report, I analysed data collected in the follow-up period of patients who suffered from heart failure to determine whether they survived or not. The dataset was taken from UCI Machine Learning Repository. In Section 2, a detailed summary of the dataset is given. Furthermore, the dataset was explored to discover special features of the data set and for a better understanding of the predictor variables that might influence survival chances. In Section 3, the Machine Learning methods used to fit models to the dataset are being discussed. Then, the results of the analysis and inferences are reported in Sections 4 and 5 respectively.

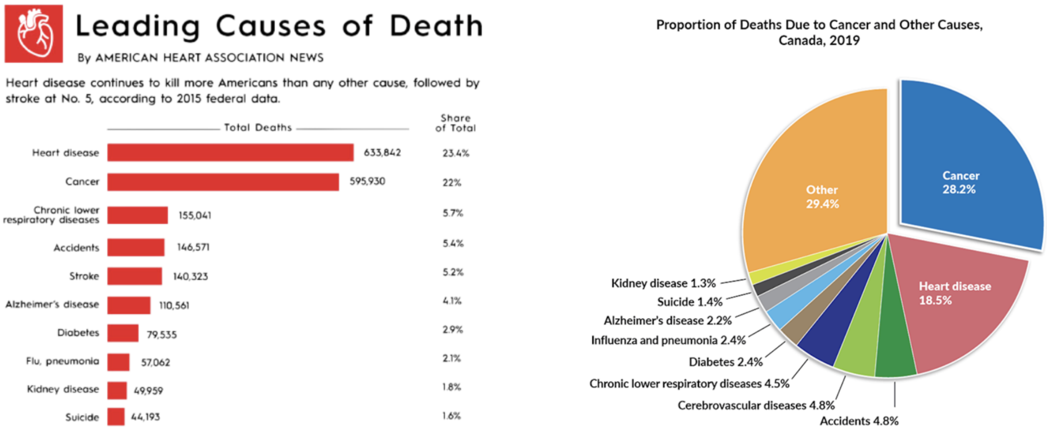

Figure 1: (Left) Death statistics from different diseases in the US in 2015 [8]. (Right) The proportional contribution of some diseases to the total deaths in Canada in 2019 [2].

Table 1: Summary of features

Age	age of the patient (in years)
Anaemia	if a patient has low red blood cells or haemoglobin (0: false, 1: true)
Creatinine phosphokinase (CPK)	level of the CPK enzyme in the blood (mcg/L)
Diabetes	if the patient has diabetes (0: false, 1: true)
Ejection fraction	percentage of blood pumped by the heart at each contraction (percentage)
High blood pressure	if the patient has hypertension (0: false, 1: true)
Platelets	platelets in the blood (kiloplatelets/mL)
Serum creatinine	level of serum creatinine in the blood (mg/dL)
Serum sodium	level of serum sodium in the blood (mEq/L)
Sex	woman or man (0: woman, 1: man)
Smoking	if the patient smokes or not (0: false, 1: true)
Time	follow-up period after heart failure (in days)
Death event	if the patient died during the follow-up period (0: false, 1: true)
where mcg/L: micrograms per litre, mg/dL: milligram per decilitre, mEq/L: milliequivalents per litre, and mL: microlitre.	

2 Data

In this project, I analysed a dataset taken from UCI Machine Learning Repository [11] originally collected by Ahmad et al [1] in 2017. This dataset contains clinical records of 299 patients in Pakistan during their follow-up period. Each clinical record consists of 13 features consisting of both binary (Anaemia, Diabetes, High blood pressure, Sex, Smoking, DEATH_EVENT) and numeric variables (Age, Creatinine phosphokinase, Ejection fraction, Platelets, Serum creatinine, Serum sodium, Time) to predict the death of patients after heart failure. A summary of the 13 features is given in table 2.

By looking at the summary of this dataset, the following observations were made:

- The range of age of patients considered here was 40 to 95 years. This may reflect that heart failure does not affect people younger than 40 years. However, cases of heart failure for less than 40 years have been rising. The average age of the patients was 60 years with the most affected age group being 50 to 70 years.

- Ejection fraction is the percentage of blood pumped out by the heart at each contraction. The ejection fraction in a healthy adult is between 50-70% [7]. The data showed a mean of 38% which was much less than the healthy range. Additionally, more than 75% of the patients had ¡45% ejection fraction. So, a low ejection fraction was a common feature among patients with heart failure.

- Serum sodium is a measure of the sodium in the blood. The healthy range of serum sodium is 135-145 milli equivalents/L [10]. The patients who died had slightly lower levels of serum sodium.

- Serum creatinine is an indicator of the performance of the kidney- the lower the better. The healthy range of serum creatinine for adult women is 0.59 to 1.04 mg/dL [6]. The mean value of serum creatinine for women was calculated as 1.38 mg/dL from the data, which was much higher than the normal range. From clinical reports, it is found that women are more likely to die after heart failure than men. This is one of the features that can be used to explain the higher likelihood of death.

- The healthy range of CPK levels is 10 to 120 micrograms per liter [5]. In this dataset, the average value (581 mcg/L) and median (250 mcg/L) were much higher than normal.

- A normal person has a platelet count of 150,000–450,000 kiloplatelets/mL of blood [12]. In this data set, 75% of the patients had a platelet count within this range. So, the patients were healthy on this ground. Moreover, there is no evidence that platelets play a role in determining survival chance from Heart failure.

Other highlights of the dataset are as follows:

- The correlation map (see Figure 2) showed a positive correlation between smoking and sex. The rest of the variables were mostly uncorrelated.

- Class imbalance (see Figure 3): there were uneven proportions of the population in the following classes:
  - The patients comprised 194 men and 105 women.
  - Around 32% of the patients died after heart failure and the remaining survived.

- And lastly, k-means clustering was implemented using just the numerical predictor variables to figure out if there was any underlying group in the data set. The Scree plot did not show the existence of any clusters (see Figure 4).

This data set had some limitations. The data set was very small, a larger data set would have ensured diversity in training data and would have helped to fit better models. In addition, the data set didn't contain a record of other relevant features of the patients like weight, height, etc. These features may also play a role in predicting the death of the patient. Moreover, the criteria to determine the value of binary variables were not stated. Lastly, additional data to generalise the model to a larger population, and different geographies would have been useful.

3 Methods

This project was being implemented in R. The Machine learning methods used for the binary classification of survival were: k-nearest neighbours, logistic regression, classification tree, bagging, random forests, and boosting.

- k-Nearest Neighbours algorithm was implemented using the knn() function. The validation approach was applied, i.e., the data was split into 80% training and 20% testing set. Then, the kNN model was fit for a range of values of k between 1 to 20. The test and train errors were determined and plotted for each value of k to find the optimal k.

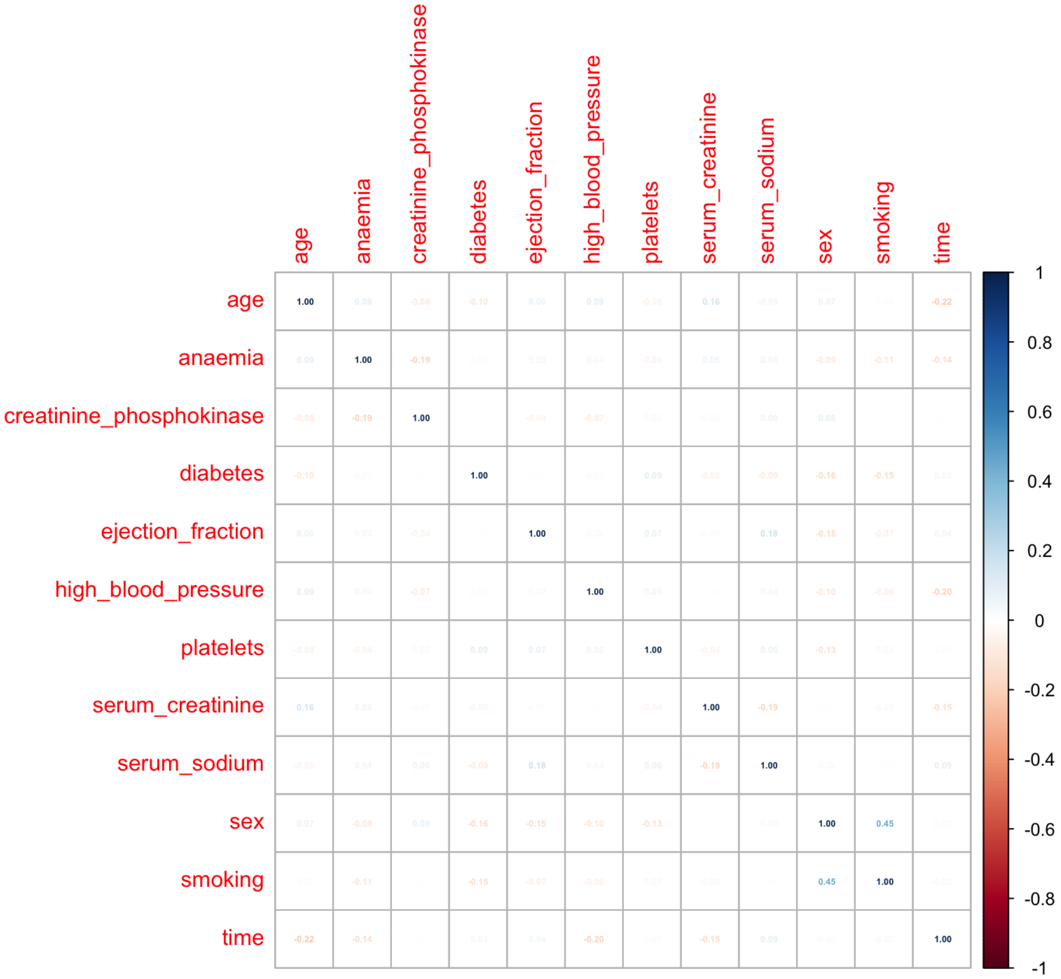

Figure 2: Correlation plot

- Logistic regression was implemented using the glm() function with DEATH_EVENT as the response variable and all other variables as predictors. Then, the backward selection was applied to remove the insignificant variables. Then, the model was fit on the training set (80% of the data) with just the significant variables as predictors. A confusion matrix was constructed after making predictions of the DEATH_EVENT on the test data (20% of the data). Furthermore, The logistic regression technique was also implemented using 5-fold and 10-fold Cross-validation approaches. Lastly, the misclassification rate was determined from the resulting confusion matrices of the validation and cross-validation approach.

- Classification tree was built again with DEATH_EVENT as the response variable and all other variables as predictors using the tree() function. Then, cost complexity pruning was used to get a sequence of best subtrees obtained by cutting the bushy tree. The cross-validation approach was used to find the optimal number of terminal nodes such that the cross-validated test error minimizes.

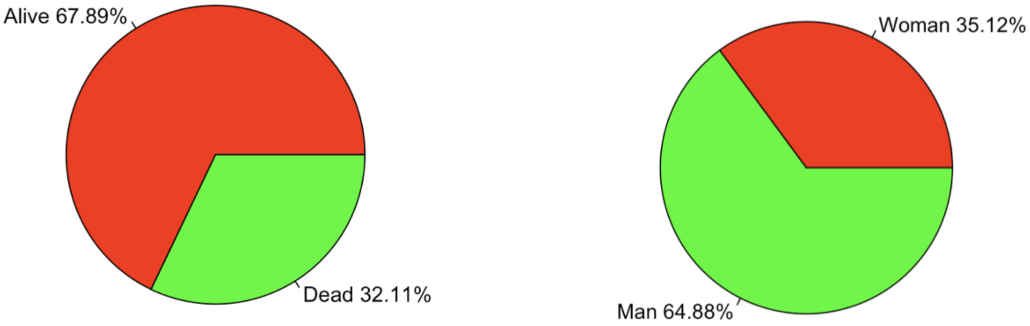

Figure 3: Pie charts showing class imbalance

- Bagging was implemented with DEATH_EVENT as the response variable and all the variables as predictors using the randomForest() function. The Variable importance plots were plotted to find the top significant predictors. And then, Bagging was re-implemented with the top 3 predictors.

- Random forests model was implemented on the data in a similar way as bagging. And then, it was re-implemented using the top 3 predictors extracted from the Variable importance plots.

- Finally, Boosting was implemented using the gbm() function with 3000 trees. The significant predictors were determined from the summary output of the implemented boosting model. And then the validation approach (70:30 train-test split) was applied with the top 3 significant predictors (leaving platelets), d=1 (interaction.depth), and B=3000 (n.trees). In addition, 10-fold cross-validation was applied with the same 3 predictors, d=1, and B=520. The misclassification rates were determined from the confusion matrices.

In the next section, the results obtained from these methods are listed.

4 Results

- kNN: In Figure 5, the test and train errors were plotted with $\frac{1}{k}$ on x-axis. k=12 gives the minimum test MSE of 0.2667.

- Logistic Regression: The summary output of regressing DEATH_EVENT on all other variables reported a subset of the predictors (age, ejection fraction, serum creatinine, and time) as significant. On implementing backward selection, the final model had the same predictors as listed above. The validated misclassification rate of the model with the significant predictors was calculated as 0.2167. The 5-fold and 10-fold cross-validation gave a misclassification rate of 0.2105 and 0.1739 respectively.

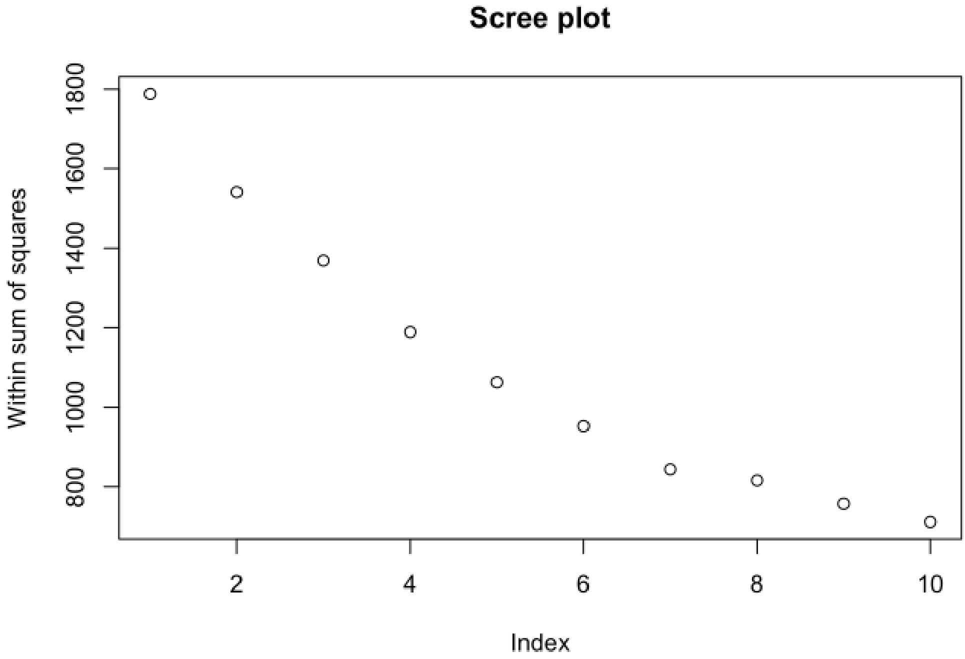

Figure 4: Scree plot of k-means clustering

- Classification tree: The cost complexity pruning (see Figure 6) suggested pruning the bushy tree with 17 terminal nodes to have just 8 of them. The cross-validated misclassification rate of the small tree (see Figure 7) came out to be 0.1839.

- Bagging: The bagging model with all the variables as predictors resulted in an OOB error rate of 18.06%. The variable importance plots (see Figure 9) showed time, ejection fraction, and serum creatinine as the top 3 predictors. Implementing bagging with the 3 predictors mentioned resulted in an OOB error rate of 17.39%.

- Random Forests: Implementing random forests algorithm with all the variables as predictors resulted in an OOB error rate of 14.05%. The variable importance plots (see Figure 8) showed time, ejection fraction, and serum creatinine as the significant predictors. Random forests with these 3 predictors gave an OOB error rate of 15.05%.

- Boosting: The variable importance plot obtained by implementing boosting with all the variables is given in Figure 10 The top 4 predictors (excluding platelets) came out to be time, creatinine phosphokinase, serum creatinine, and ejection fraction. The boosting algorithm implemented with the mentioned significant predictors, validation approach, B=3000 gave a misclassification rate of 0.1889. And, 10-fold cross-validation with the same predictors, B=520 trees gave a misclassification rate of 0.1575.

The error rates of these methods are summarized in table 4.

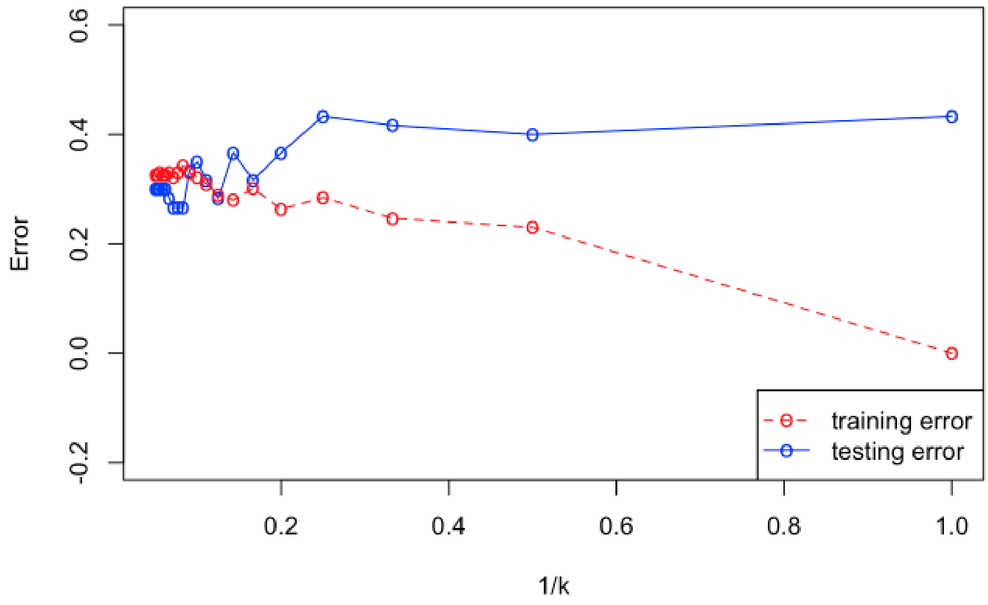

Figure 5: Error vs $\frac{1}{k}$ plot for different number of nearest neighbours

5 Discussion

- Exploratory data analysis: the exploration of data in Section 2 revealed some key features of the patients of heart failure. Patients who died of heart failure had lower levels of ejection fraction and serum sodium, and higher levels of serum creatinine and CPK. Platelets did not play a role in predicting the death event. Sex and smoke were positively correlated but both variables were insignificant for the implemented ML models. And there was a class imbalance both in the Death event and Gender of the patients. Lastly, the scree plot obtained by implementing k-means clustering with just the numerical predictors for different values of k is given in Figure 4. A gradual decrease in within sum of squares suggested that there were no underlying clusters in the data.

- k-Nearest neighbours: the Error vs. $\frac{1}{k}$ plot showed the typical trend in the train error (monotonic decrease) and test error (u-shape) with the increase in flexibility.

- Logistic Regression: The summary output of regression with all the variables as predictors gave age, ejection fraction, serum creatinine, and time as the significant variables. The backward selection technique did not change the insignificance of predictors at each step and hence resulted in the same 4 predictors as significant at the end.

- Classification tree: the CV plot (see Figure 6) obtained from cost complexity pruning showed the minimum misclassification rate for a tree with 8 terminal nodes. The pruned tree (see Figure 7) had time, serum creatinine, ejection fraction, age, and creatinine phosphokinase at the decision nodes which suggested these variables to be significant. 4 out of these were consistent with the significance results obtained from Logistic regression.

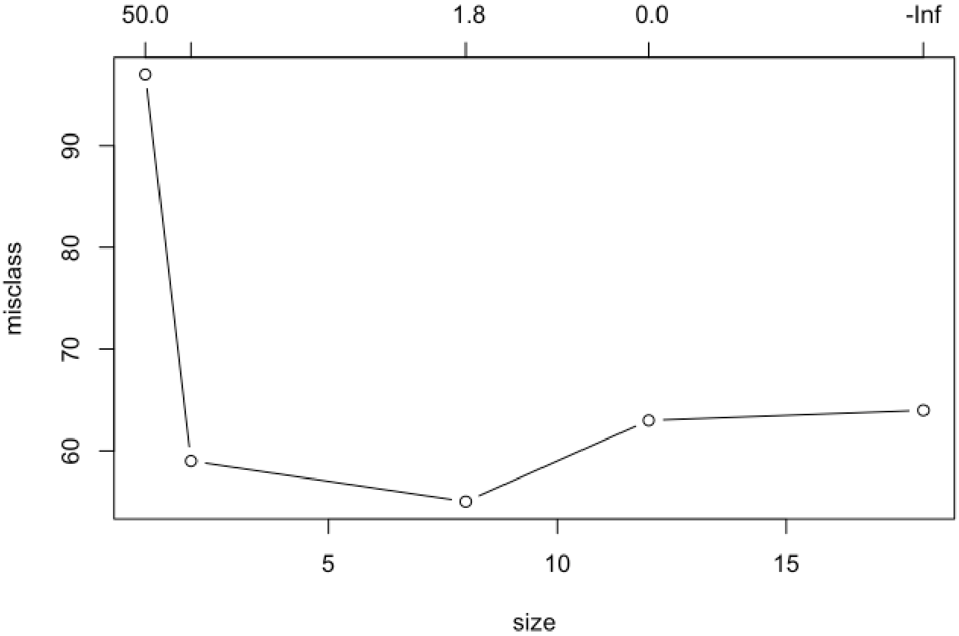

Figure 6: CV Plot to determine number of terminal nodes the bushy tree need to be pruned to

- Bagging: the variable importance plots (see Figure 9) showed that the mean decrease in accuracy and gini index was maximum for time, ejection fraction, and serum creatinine. Age seemed to be significant in the left plot but not in the right one. So, it was not considered as significant.
- Random forests: Similar to bagging, the top 3 significant predictor variables (see Figure 8) were time, ejection fraction, and serum creatinine. Age showed as the fourth significant predictor in both the importance plots but it was not considered because its significance level was half that of the third predictor.
- Boosting: the variable importance table (see Figure 10) gave time, creatinine phosphokinase, serum creatinine, platelets, and ejection fraction as the top 5 predictors. Platelets were not considered because there is no evidence that platelets play a role in predicting the survival of heart patients. Moreover, including platelets as a predictor increased the misclassification rate. Note that a small number of stumps were required with 10-fold cross-validation compared to the validation approach to attain better accuracy. This was because in the cross-validation approach results for 10 different train and test sets were averaged out.

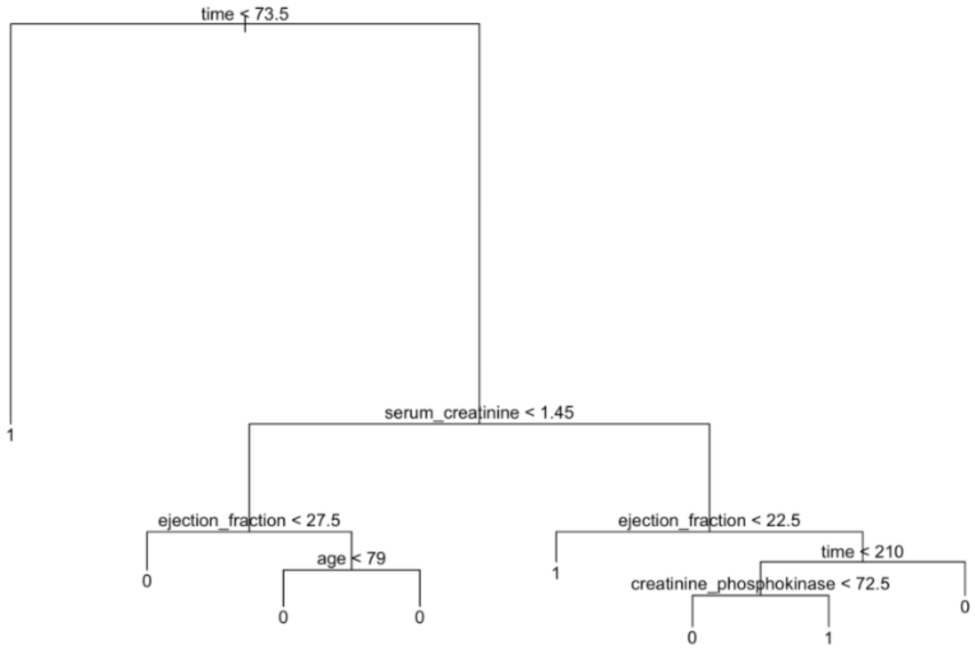

Figure 7: Pruned Classification tree

```
##                                               var    rel.inf
## time                                         time 32.9603315
## creatinine_phosphokinase creatinine_phosphokinase 14.4260462
## platelets                               platelets 12.8422576
## serum_creatinine                 serum_creatinine 12.2368054
## ejection_fraction               ejection_fraction 10.5601155
## age                                           age  9.0616487
## serum_sodium                         serum_sodium  4.1036708
## anaemia                                   anaemia  1.2165921
## diabetes                                 diabetes  0.7620049
## high_blood_pressure           high_blood_pressure  0.7520982
## sex                                           sex  0.7504656
## smoking                                   smoking  0.3279635
```

Figure 10: Variable importance table from Boosting algorithm

From table 4, it can be seen that Random forests with DEATH_EVENT as the response variable and all the variables as predictors performed the best with a misclassification rate of around 14%. Generally, boosting is expected to perform better. But here we got an error of 15.75% with 10 fold cross-validation approach and time, creatinine phosphokinase, serum creatinine, and ejection fraction as the predictors. Since variable significance analysis from other models revealed that all the variables were not significant in predicting the death event, Random forests may not be the

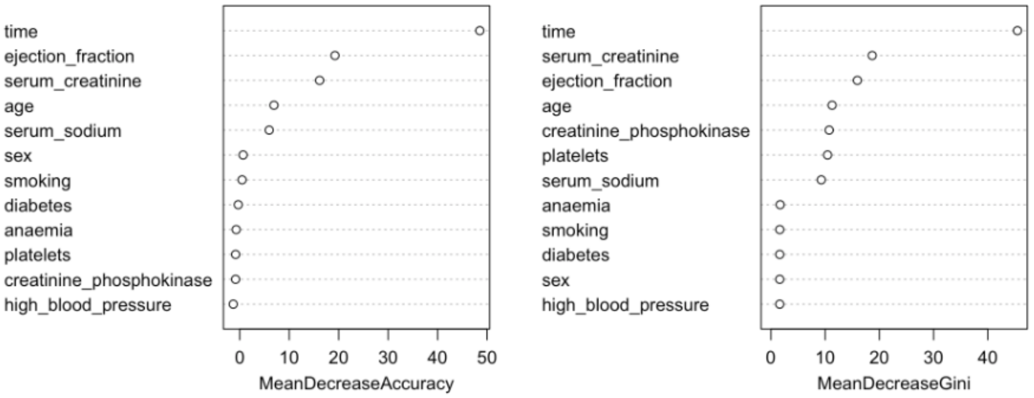

Figure 8: Variable importance plot from Random forests algorithm

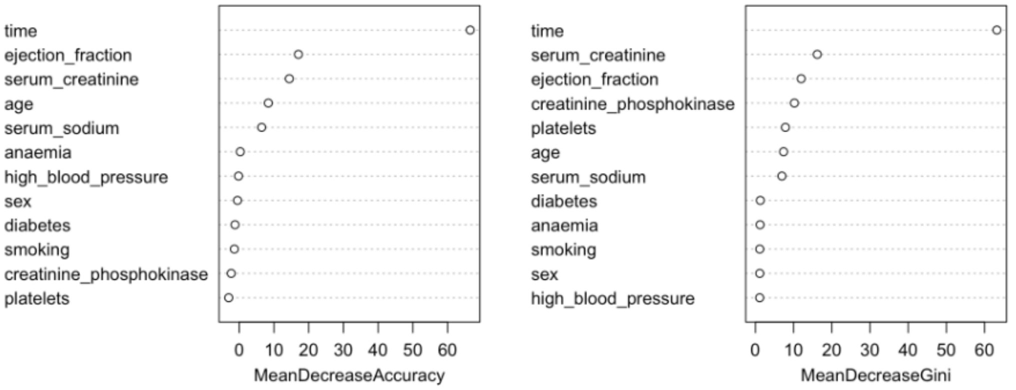

Figure 9: Variable importance plot from Bagging algorithm

optimal model. With a slight decrease in accuracy but a significant increase in interpretability, Boosting was found as the optimal model.

Chicco et al [4] found serum creatinine and ejection fraction to be significant predictors. In this project, time and creatinine phosphokinase were additionally found to be significant. The authors fitted various ML models and found the stratified logistic regression model with ejection fraction, serum creatinine, and time as predictors to be an optimal model. However, the authors also mentioned Boosting and Random forests to perform well on different metrics which is consistent with my result. The authors also explored the application of ML models without considering the follow-up time as a predictor. However, time should not be removed because it was the most significant predictor in all the models.

Regarding future work, since there was a class imbalance in the data set, different performance metrics can be evaluated and compared. F1 score is popular in the case of imbalanced datasets. Moreover, false negatives, i.e., predicting a death case as alive, is more costly here, so Recall might also be a reliable metric to compare. And lastly, this model can be used to predict on data from different geographies and on people of different age groups to find if it is generalisable.

It was surprising that diabetes, high blood pressure, and smoking were not significant in pre-

Table 2: Misclassification rates of all the methods with Validation or Cross-validation approaches and with different predictors

Method		**Misclassification rate**
kNN	k=12	0.2667
Logistic Regression	Validation	0.2167
	5-fold CV	0.2105
	10-fold CV	0.1739
Classification tree	10-fold CV	0.1839
Bagging	All predictors	0.1806
	With time, ejection_fraction, and serum_creatinine	0.1739
Random forests	All predictors	0.1405
	With time, ejection_fraction, and serum_creatinine	0.1505
Boosting	Validation with time, creatinine_phosphokinase, serum_creatinine, and ejection_fraction	0.1889
	10-fold CV with time, creatinine_phosphokinase, serum_creatinine, and ejection_fraction	0.1575

dicting the death of patients. These factors are generally highlighted by the health authorities to keep a check on. The results of this analysis showed that death of heart patients can be predicted based on time, creatinine phosphokinase, serum creatinine, and ejection fraction levels. However, more confirmatory research should be done before implementing these ideas in clinical practices.

References

[1] Tanvir Ahmad et al. “Survival analysis of heart failure patients: A case study”. In: *PloS one* 12.7 (2017), e0181001.

[2] *Cancer statistics at a glance*. URL: https://cancer.ca/en/research/cancer-statistics/cancer-statistics-at-a-glance.

[3] *Cardiovascular diseases (CVDs)*. URL: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)#.

[4] Davide Chicco and Giuseppe Jurman. “Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone”. In: *BMC medical informatics and decision making* 20.1 (2020), pp. 1–16.

[5] *Creatine phosphokinase test*. URL: https://www.mountsinai.org/health-library/tests/creatine-phosphokinase-test#.

[6] *Creatinine tests*. URL: https://www.mayoclinic.org/tests-procedures/creatinine-test/about/pac-20384646#.

[7] *Ejection Fraction Heart Failure Measurement*. URL: https://www.heart.org/en/health-topics/heart-failure/diagnosing-heart-failure/ejection-fraction-heart-failure-measurement.

[8] *Heart disease, stroke death rates increase following decades of progress*. URL: https://www.heart.org/en/news/2018/05/01/heart-disease-stroke-death-rates-increase-following-decades-of-progress.

[9] *Heart Failure*. URL: https://www.cdc.gov/heartdisease/heart_failure.htm#.

[10] *Hyponatremia*. URL: https://www.mayoclinic.org/diseases-conditions/hyponatremia/symptoms-causes/syc-20373711#.

[11] *UCI Machine Learning Repository*. URL: https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records#.

[12] *What Are Platelets and Why Are They Important?* URL: https://www.hopkinsmedicine.org/health/conditions-and-diseases/what-are-platelets-and-why-are-they-important#.