UNIVERSITY OF PRINCE EDWARD ISLAND

STAT 4660

Data Visualization and Mining

THE HEART FAILURE PREDICTION

Performed at:

University of Prince Edward Island

By

Quoc Toan Lieu

School of Mathematical and Computational Sciences

October 4, 2021

I. INTRODUCTION

According to World Health Organization (WHO), Cardiovascular diseases (CVDs) are the leading cause of death globally, taking an estimated 17.9 million lives each year. CVDs are a group of disorders of the heart and blood vessels and include coronary heart disease, cerebrovascular disease, rheumatic heart disease and other conditions. More than four out of five CVD deaths are due to heart attacks and strokes, and one third of these deaths occur prematurely in people under 70 years of age. Among 17.9 million in 2019, which accounted for 32% of total deaths worldwide, 85% were due to heart attack and stroke. Moreover, ¾ of CVD deaths take place in low- and middle-income countries. In addition, along 17 million premature deaths, 38% were caused by CVDs.

Therefore, it is essential to obtain a method to predict the CVDs or heart failure specifically. Machine learning, which is one of the most popular and favorite method that can be applied to solve this situation. Machine learning can be used to patients' survival from the patient's data and can individuate the most important features among those included in their medical records.

II. QUESTIONS

There are the questions which are expected to be answered in this project:

1. Which attribute plays an important role in the death event?
2. Whether smokers from any particular gender have more chances of dying earlier?
3. How do smokers/non-smokers' values change the death event?

III. CONCEPTION

According to WHO (2021), Cardiovascular diseases (CVDs) are a group of disorders of the heart and blood vessels. CVDs are a group of disorders of the heart and blood vessels and include (International diabetes federation, 2021):

1. Coronary heart disease: Disease of the blood vessels supplying the heart muscle
2. Cerebrovascular disease: Disease of the blood vessels supplying the brain
3. Peripheral arterial disease: Disease of the blood vessels supplying the arms and legs.

There are 4 risk factors that may increase the chance of developing CVDs:

1. Associated conditions and metabolic risk factors: Hypertension, Diabetes, Chronic kidney disease…
2. Behavioural risk: Smoking, physical inactivity, unhealthy eating habit…
3. Modifiable risk factors: life-course risks, environment risk factors, lack of awareness…
4. Non-modifiable risk factors: age, sex, family history…

Heart attacks and strokes are usually acute events and are mainly caused by a blockage that prevents blood from flowing to the heart or brain. The most common reason for this is a build-up of fatty deposits on the inner walls of the blood vessels that supply the heart or brain. Strokes can be caused by bleeding from a blood vessel in the brain or from blood clots.

WHO defines anaemia as a condition in which the number of red blood cells or the haemoglobin concentration within them is lower than normal. The symptom of this condition can be recognised through fatigue, weakness, dizziness and shortness of breath, among others. The causes of anaemia can be come from nutritional deficiencies, particularly iron deficiency, though deficiencies in folate, vitamins B12 and A are also important causes; haemoglobinopathies; and infectious diseases, such as malaria, tuberculosis, HIV and parasitic infections.

Creatine phosphokinase (a.k.a., creatine kinase, CPK, or CK) is an enzyme (a protein that helps to elicit chemical changes in your body) found in your heart, brain, and skeletal muscles. When muscle tissue is damaged, CPK leaks into your blood. Therefore, high levels of CPK usually indicate some sort of stress or injury to your heart or other muscles. To test CPK, blood is drawn from a vein in your arm (Johns Hopkins Lupus Center)

According to Mitchel & P.GoldmanJean-JérômeGuex (2017), Ejection fraction (EF) is the percentage of blood volume ejected in each cardiac cycle and is a representation of LV systolic performance. It is calculated from the end-diastolic and end-systolic volumes of the left ventricle.

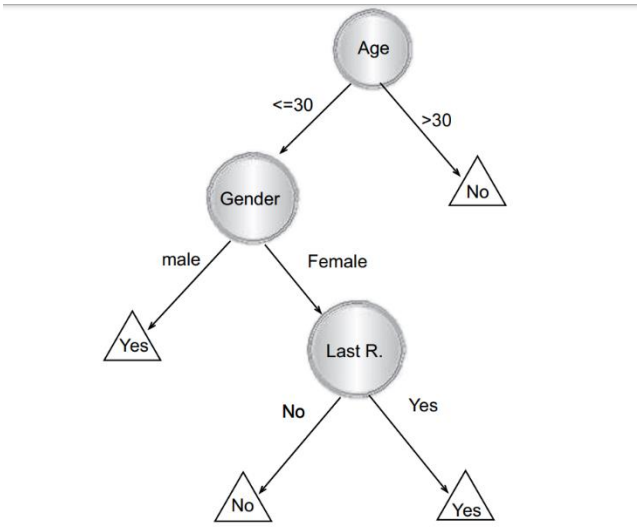
James N George (2000) defines platelets are the smallest of blood cells, being only fragments of megakaryocyte cytoplasm, yet they have a critical role in normal haemostasis and are important contributors to thrombotic disorders. (Kaggle, 2020)

Serum creatinine is an index of a creatinine test which is a way for doctors to measure how well your kidneys are working. Creatinine is a waste product from the normal breakdown of muscle tissue. As your body makes it, it's filtered through your kidneys and expelled in urine. (Hansa D. Bhargava, 2020).

Sodium concentration is maintained in a narrow range of 137 to 142 mEq/L of plasma. The value is 145 to 155 mEq/L of plasma water, a point to be noted because in a few circumstances there are significant changes in the plasma water concentration (George L. Ackerman,1990).

IV. MODEL

One of the most important definition about the Random Forest is decision tree. According to Lior Rokach and Oded Maimon (2005), a decision tree is a classifier expressed as a recursive partition of the instance space. The decision tree consists of nodes that form a rooted tree, meaning it is a directed tree with a node called "root" that has no incoming edges. All other nodes have exactly one incoming edge. A node with outgoing edges is called an internal or test node. All other nodes are called leaves (also known as terminal or decision nodes). In a decision tree, each internal node splits the instance space into two or more sub-spaces according to a certain discrete function of the input attributes values. In the simplest and most frequent case, each test considers a single attribute, such that the instance space is partitioned according to the attribute's value. In the case of numeric attributes, the condition refers to a range.



Source: Lior Rokach and Oded Maimon, 2005

Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error for forests converges to a limit as the number of trees in the forest becomes large. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them (Robert E. Schapire, 2001).

Robert (2001) has made the definition about the Random Forest as "A random forest is a classifier consisting of a collection of tree-structured classifiers $\{h(\mathbf{x}, \Theta_k),\ k = 1, \ldots\}$ where $\{\Theta^k\}$ is the

independent identically distributed random vectors and each tree casts a unit vote for the most popular class at input x”

V. DATASET

The data set was collected on the Kaggle site about the heart failure. The data contains 299 medical records from 299 patients in the Faisalabad Institute of Cardiology and at the Allied Hospital in Faisalabad (Punjab, Pakistan), during April–December 2015. The data set contains 13 different attributes which are age, anaemia, creatinine phosphokinase, diabetes, ejection fraction, high blood pressure, platelets, serum creatinine, serum sodium, sex, smoking, time and death event. The detail of the attributes is demonstrated below.

Attributes	Explanation	Unit
Age	Gender of the patient	Years
Anaemia	Decrease of red blood cells or hemoglobin	Boolean
High blood pressure	If a patient has hypertension	Boolean
Creatinine phosphokinase (CPK)	Level of the CPK enzyme in the blood	mcg/L
Diabetes	If the patient has diabetes	Boolean
Ejection fraction	Percentage of blood leaving	Percentage
Sex	Woman or man	Binary
Platelets	Platelets in the blood	kiloplatelets/mL
Serum creatinine	Level of creatinine in the blood	mg/dL
Serum sodium	Level of sodium in the blood	mEq/L
Smoking	If the patient smokes	Boolean
Time	Follow-up period of heart failure	Days
(target) death event	If the patient died during the follow-up period	Boolean

Table 1: the description of the attributes

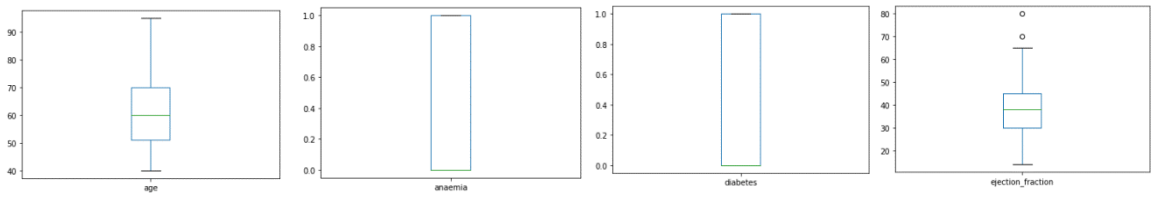
VI. RESULT

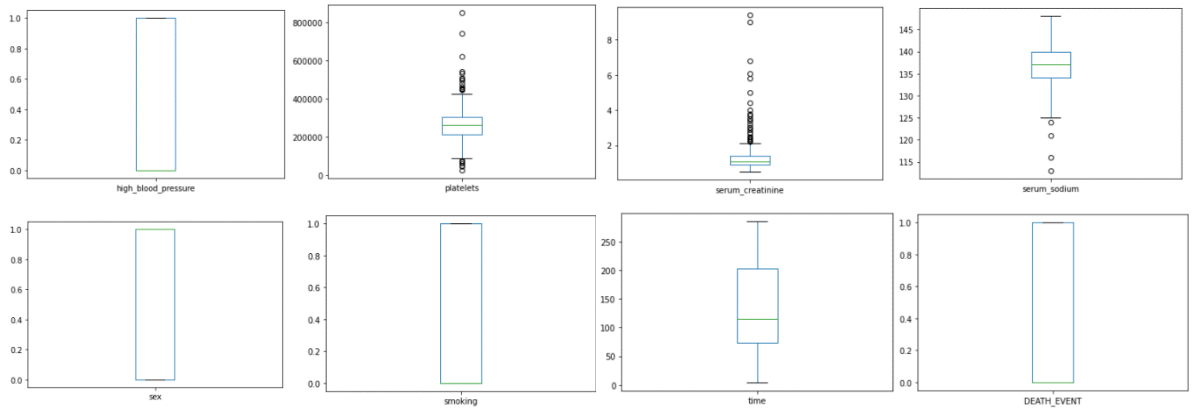
First, it is necessary to performance some descriptive analysis to look out for some information about data set. The picture below shows that there is no null value in the data set.

```
data = pd.read_csv("C:/upei/Fall2021/visulization/project/heart_failure_clinical_records_dataset.csv", delimiter=',')
df = pd.DataFrame(data)
print(df.isnull().sum())
```

```
age                         0
anaemia                     0
creatinine_phosphokinase    0
diabetes                    0
ejection_fraction           0
high_blood_pressure         0
platelets                   0
serum_creatinine            0
serum_sodium                0
sex                         0
smoking                     0
time                        0
DEATH_EVENT                 0
dtype: int64
```

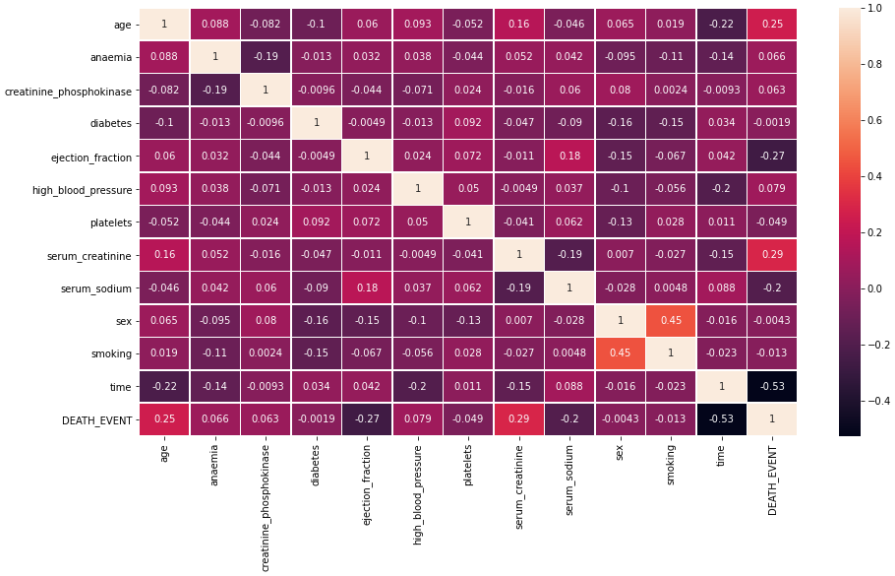
Then, each attributes of the data set will be checked. It can be seen through the box plots below.





As it is demonstrated through the plots above, most of the attribute's do not have outliers. There are only 3 attributes that has the outliers, especially platelets and serum creatinine. In the regression models, the outliers will a major effect on the accuracy of them. With random forest, the model will handle it by binning them. Therefore, it is not necessary to handle the outliers manually.

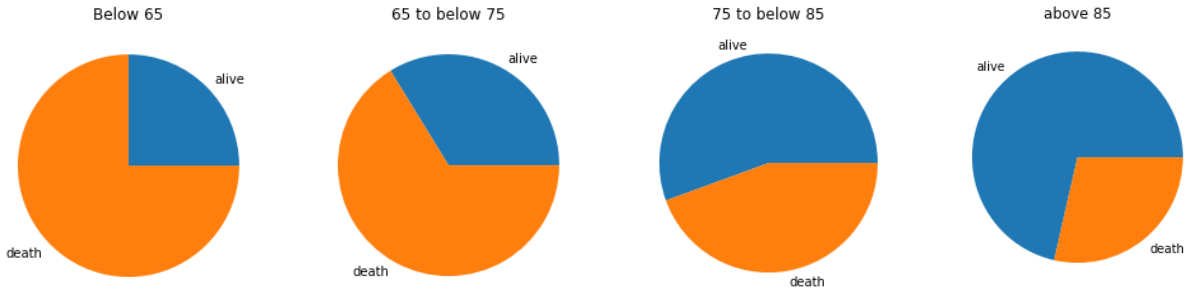
The objective of the project is looking for the attributes that can predict the death event of patient who have heart failure. One of the popular ways is to look for the correlation coefficient between each attribute with death event. The value of correlation coefficient will be showed in the matrix below.



As the matrix result, looking at the death event column, the attribute time, serum sodium, serum creatinine, and ejection fraction have the vital effect on the death event due to the correlation coefficient stays close to the value 1 and -1. Mean while, the other attributes have less impact on the death event because the correlation coefficients are closed to the value 0. To be more specific, some pie charts of the attributes will be shown to demonstrate the relation between them with death events.

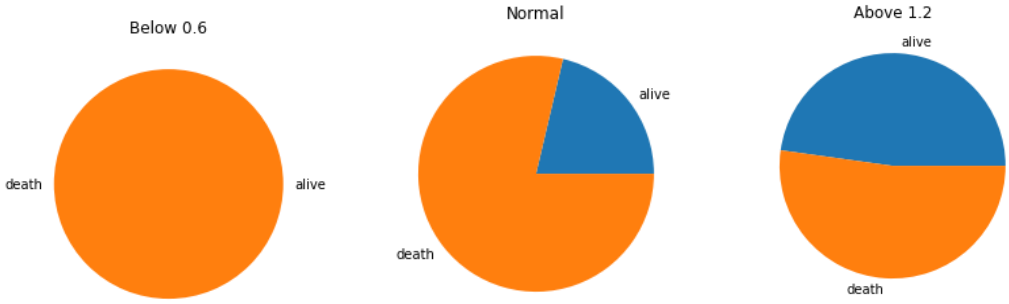
The first 3 pie charts demonstrate the proportion of death and alive event during the patients in the different period of ages. The graphs are demonstrated below. As it can be seen through the chart, it

seems like the older patient would be get less heart failure than the younger one. This is not appropriate as conventional concept. This can be explained that the number of data is not large enough in each group.

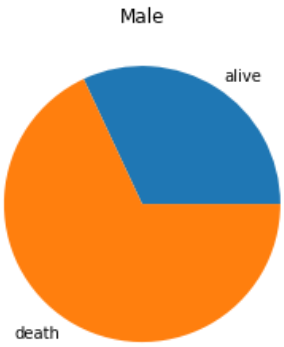


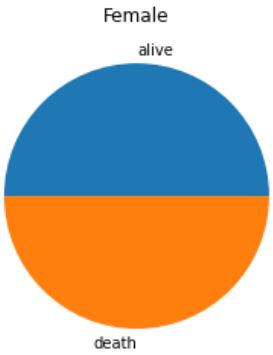
Next, the time attributes and death event should be considered. The time will be rescaled by separate into 3 different group which are 1 month, 1 to 6 months and above 6 months. As it is demonstrated, it can be guessed that the patients with the has the larger time for next heart failure has less chance to be survived than comparing to the one with the shorter following time.

Another attribute will be demonstrated is serum creatinine. The pie charts below will demonstrate the distribution of death and alive event. There are 3 level of serum creatinine which are low (below 0.6), normal (0.6 to below 1.2) and high (above 1.2). the graphs show that the patient high level of serum creatine seem to be lightly death when heart failure happening.

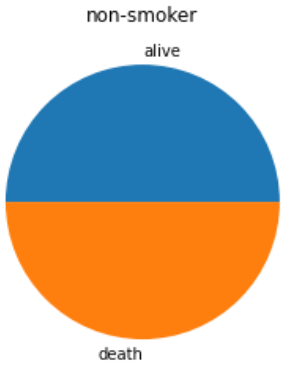


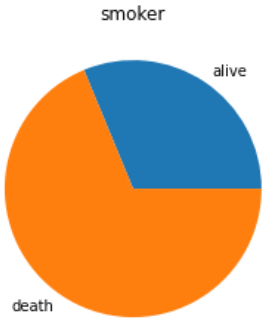
Another problem that may be discuss is that is there a differences in the death event between male and female. The bar charts below show that it maybe the male patients have a larger chance of death when heart failure event happens. In this model, death event attribute will be used as the predicted variable with the label of 1 (death) and 0 (alive) while the rest of attributes will be predicting variable.





Beside that, as the question was assigned in the beginning, the proportion of death and alive patients is demonstrate through the graph below. The graphs show that it is likely the non-smoker patients have a large chance of survive from heart failure than the smoker.





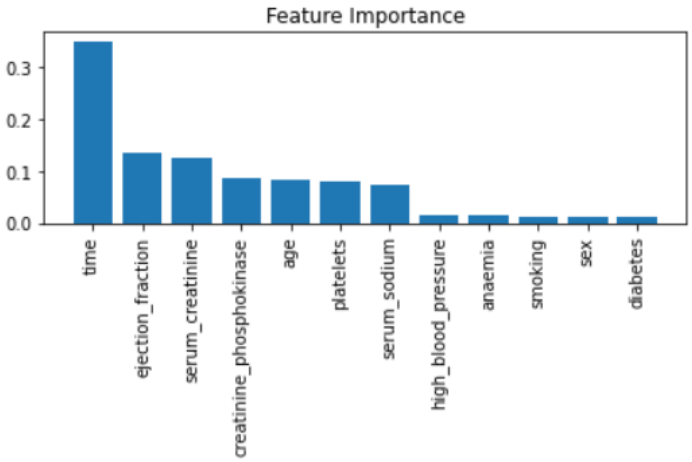
All the graphs above just give some subjective consideration about the data set, it can not conclude that these attributes really dominate whether the death event may happen or not. In order to do that, the Random Forest model will be applied to predicting the death event in the future of patient. Beside that, 10- fold-cross-validation will also be applied to increase the accuracy of the Random Forest. Basically, 299 samples will be split into 10 different folds where one them will be use as the testing set, and the rest of 9 will be training set. In addition, each fold will be used in turn of the testing set. The result of the model is demonstrate by the picture below.

```
def randomForrest(X,Y):
    clf=RandomForestClassifier(n_estimators=100)
    scores = cross_val_score(clf, X, Y, cv=10)
    print(np.average(scores))
```

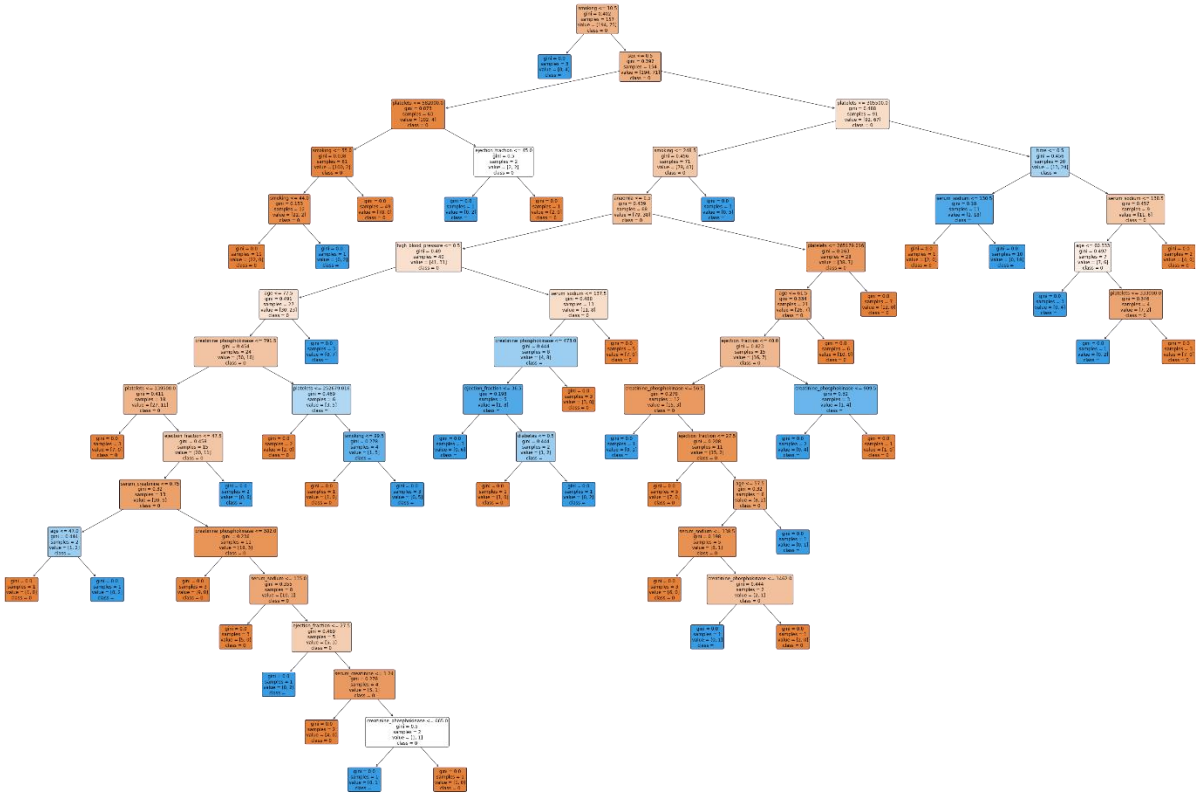
```
randomForrest(X,Y)
```

```
0.7689655172413794
```

The accuracy of the model is around 77%. This is an acceptable index due to the lack of samples in the data set. Therefore, with the currency data set, the Random Forest model can predict where the patient may be death in the future.



The graph above shows the role of the attributes to the death event. Time, ejection fraction, serum creatinine are the attributes that play the most important role in the deaths event. While the diabetes, sex and smoking do not contribute much to the predicting death event. This is similar to the correlation efficiency above. One of the decision trees on the forest is demonstrated below.



The role of some attributes can be figured out by dropping them out of the model. If the accuracy of model is less than more than 15% after dropping the attribute, that one will have a big important role in the predicting the death event. The table below will show the accuracies of model after dropping some attributes

Dropping attribute	Accuracy
Time	0.7326
Ejection fraction	0.7722
Serum creatinine	0.7489
Creatinine phosphokinase	0.7589
smoking	0.7656
diabetes	0.7622
sex	0.7522
Original	**0.7689**

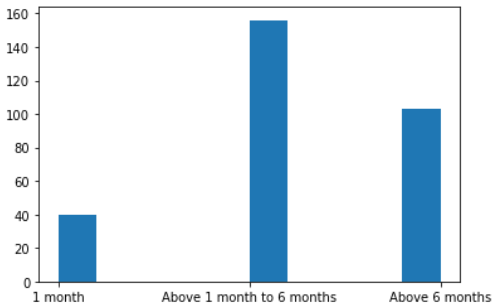
The table above shows that there is a big decrease in the accuracy when the attributes which are considered as playing the big role in predicting death event. Mean while, there is the minor decrease in the accuracies of the dropping unimportant attributes. However, there is no major decrease of the accuracy. It can conclude that there is no attribute that dominates the model. Though it can be seen that when attribute Ejection fraction was eliminated from the model, the accuracy is increasing a little bit. This is because too many attributes may lead to overfitting in the model. However, the difference is big enough to conclude that whether the model is overfitting or not.

Next, another question that need to be answer is that whether the attributes, but death event can predict the time that for the next heart failure likely to happen. To do that, the time attribute will be rescale into 3 different group of below 1 month, 1 to 6 months and above 6 months. The accuracy of the Random Forest is only 49.17%.

```
randomForrest(X_time,Y_time)
```

```
0.4917241379310345
```

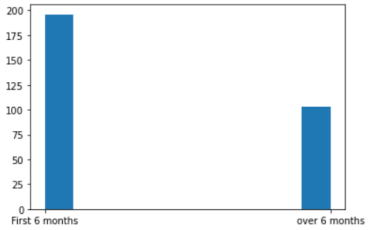
The reason for the low accuracy is maybe there is time is divided equally. It can be seen through the histogram below.



Then, it is necessary to rescale the time attributes to check whether the accuracy could be increased. Time attribute will be divided into 2 group with the time of 6 months. Then, the accuracy can be increase to around 62%.

```
plt.hist(Y_time)
```

```
(array([196.,   0.,   0.,   0.,   0.,   0.,   0.,   0.,   0., 103.]),
 array([0. , 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1. ]),
 <BarContainer object of 10 artists>)
```




```
randomForrest(X_time,Y_time)
```

```
0.6254022988505746
```

VII. Conclusion

The Random Forest perform very well on predicting the death event of the patients with the certain health conditions. 77% is an acceptable accuracy for such a problem. However, if more samples are provided, the random Forest model can be more efficiency. This outcome can help to monitor the patient with history of heart failure and poor condition to avoid the death vent of heart failure. However, the model does not have enough information to predict the time for the next heart failure to happen.

REFERENCES

I. Ackerman., G. L. (1990). *Clinical Methods: The History, Physical, and Laboratory Examinations.* Boston: Butterworths.

II. Bhargava, H. D. (2020). *Creatinine Test.* https://chiasehieubiet.com/a-to-z-guides-creatinine-and-creatinine-clearance-blood-tests/.

III. federation, I. d. (2021). Diabetes and cardiovascular disease. *https://idf.org/our-activities/care-prevention/cardiovascular-disease.html*.

IV. federation, W. h. (2017). The world most common cause of death CARDIOVASCULAR DISEASES global facts and figured. *, https://world-heart-federation.org/wp content/uploads/2017/05/WCC2016_CVDs_infographic.pdf*.

V. George, J. N. (2000). Platelets. *THE LANCET*, Vol 355 .

VI. Kaggle. (2020). Heart prediction. *https://www.kaggle.com/andrewmvd/heart-failure-clinical-data*.

VII. Lior Rokach, Oded Maimon. (2005). *The Data Mining and Knowledge Discovery Handbook .* Springer.

VIII. Mitchel P.GoldmanJean-JérômeGuex. (2017). Noninvasive Examination of the Patient Before Sclerotherapy. *Treatment of Varicose and Telangiectatic Leg Veins*, 100-136.

IX. Schapire, R. E. (2001). Random Forests. *Machine Learning*, 5–32.

X. WHO. (2021). Cardiovascular diseases (CVDs). *https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)*.