Author: Linda Nguyen

PREDICT SURVIVAL OF PATIENTS WITH HEART FAILURE

Table of Content

Abstract

This project aims to investigate which clinical factors could be used to predict survival of heart failure patients. Machine learning models was conducted based on the heart failure dataset collected in 2015 (Machine Learning Respository, 2020). It was medical records of 299 patients. The project applied two classification methods: K-Nearest Neighbours (KNN) and Decision Tree Classifier (DTC). The model achieved 91% accuracy for KNN and 87% accuracy for DTC. We applied these machine learning classifiers to rank the importance level of each feature. Overall, the results indicated that the number of check-up periods, ejection fraction and serum creatinine are top three predictive features for survival prediction on heart disease patients. The report concluded that the models predict on these three clinical features obtained results more accurate comparing to using all 11 features. Moreover, it is recommended that it is critical for patients with cardiovascular issues making regular follow-up visits, improve their ejection fraction and serum creatinine.

1. Introduction

Heart failure is one of cardiovascular disease that affects pumping action of the heart muscles (What is heart failure, 2020). It has killed approximately 17 million people globally every year (WHO, 2021). Heart failure happens when our heart is not able to delivery enough blood to our body. This fatal disease often causes shortness of breath or fatigue. People have serious high blood pressure, diabetes, other heart condition have high chance to get heart failure. In this context, medical records are mainly considered as the most useful resource. This helps to discover new efficient tool for medical sectors predicting survival for heart disease patients. Recently, several studies have been conducted using different demographics and conditions working with different data sources. They aim to investigate which factors can be used to predict heat failure (Chicco & Jurman , 2020). In this report, we'll aim to answer the research question: **"Which features can be used effectively to forecast survival of heart failure patients?"**

2. Methodology

In this project, we used the heart failure dataset. It included medical records of 299 patients who had heart failure, collected during their follow-up period, where each patient profile has 13 clinical features: age, anaemia, high blood pressure, creatinine phosphokinase (CPK), diabetes, ejection fraction, platelets, sex, serum creatinine, serum sodium, smoking, time and death event as target feature.

We used machine learning models to predict survival. Particularly, K-Nearest Neighbours (KNN) and Decision Tree Classifier (DTC) were mainly utilised. Before modelling, we explored the relationship between the target feature 'death' with other features through data exploration step. We then used Extra Trees Classifier technique to rank the importance level of the features. We then began to model by training and evaluating two models. Since both exploration data results and feature ranking approaches clearly identified time, ejection fraction and serum creatinine as top three important features, we built models based on these three factors. To improve the predicted results, we also ran parameter tuning. Particularly, different

parameters for ‘weights and ‘p’ were selected for KNN model and RandomizedSearchCV method used for the DTC model. We then use k-fold cross-validation to evaluate performance and choose classifier parameters. This will test the model’s ability to predict new data. This step ensures the model scores do not depend on how we selected train and test dataset. Lastly, we compared the accuracy score between two models.

2.1. Descriptive statistician summary

	age	CPK	ejection_fraction	platelets	serum_creatinine	serum_sodium	time
count	299.000000	299.000000	299.000000	299.000000	299.00000	299.000000	299.000000
mean	60.829431	581.839465	38.083612	263358.029264	1.39388	136.625418	130.260870
std	11.894997	970.287881	11.834841	97804.236869	1.03451	4.412477	77.614208
min	40.000000	23.000000	14.000000	25100.000000	0.50000	113.000000	4.000000
25%	51.000000	116.500000	30.000000	212500.000000	0.90000	134.000000	73.000000
50%	60.000000	250.000000	38.000000	262000.000000	1.10000	137.000000	115.000000
75%	70.000000	582.000000	45.000000	303500.000000	1.40000	140.000000	203.000000
max	95.000000	7861.000000	80.000000	850000.000000	9.40000	148.000000	285.000000

Figure 1: Descriptive statistics for numeric variables

Figure 1 shows that there are 299 observations. Let’s interpret the results of ‘time’, the mean follow-up period is 130 days greater than the median which is 115 days. This means the data appear to be skewed to the right. Also, the results show the standard deviation of time is 77. In normal distribution, most of the observations are within 1 standard deviations of the mean. (reference empirical rule). We can interpret other variables in the same way to understand figure. In another way, we could visualise these features as below.

2.2. Explore relationships between target feature and other features

Before we start exploring the relationship between features, we should see how imbalanced the classes are. That is, how many patients have and don’t have heart disease.

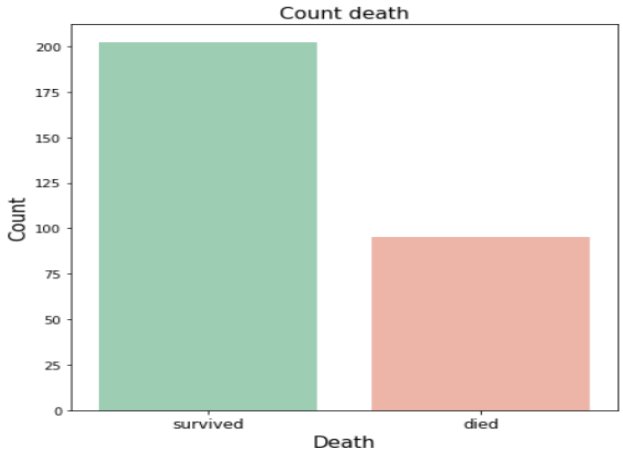

Figure 2: Count number of death and survival of dataset

Figure 2 represents an imbalance of the two classes. That is 202 survival and 95 deceased patients

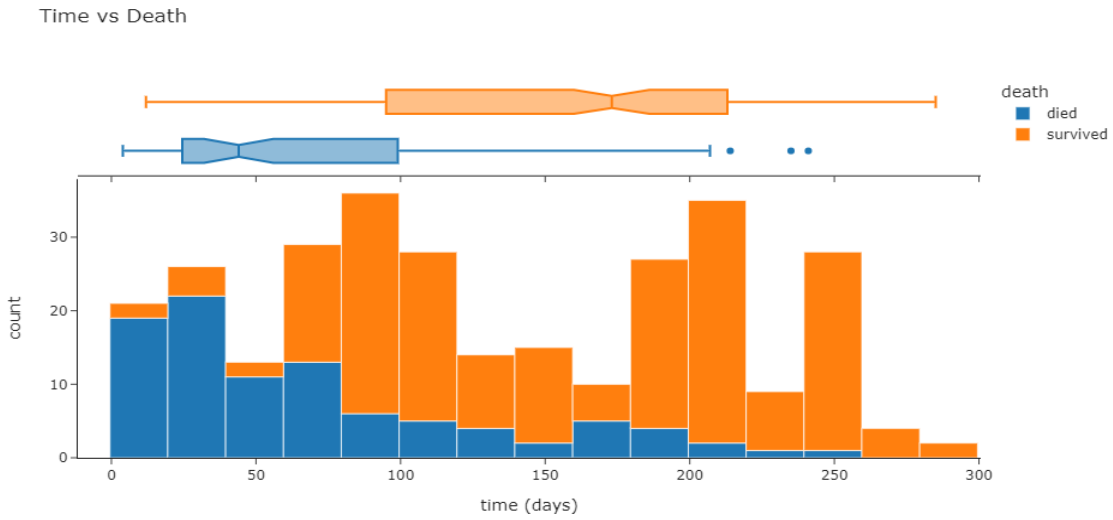

Figure 3: Time vs. Death

Figure 3 shows died people are clearly has less period of days doing their follow-up check up while survived patients are people doing it more regularly.

that number of period patients do follow up check up on their disease are extremely important.

Time of the patient's follow-up visit for the disease is crucial in as initial diagnosis with cardiovascular issue and treatment reduces the chances of any fatality. It holds and inverse relation.

Ejection fraction is the second most important feature. It is quite expected as it is basically the efficiency of the heart.

Age of the patient is the third most correlated feature. Clearly as heart's functioning declines with ageing

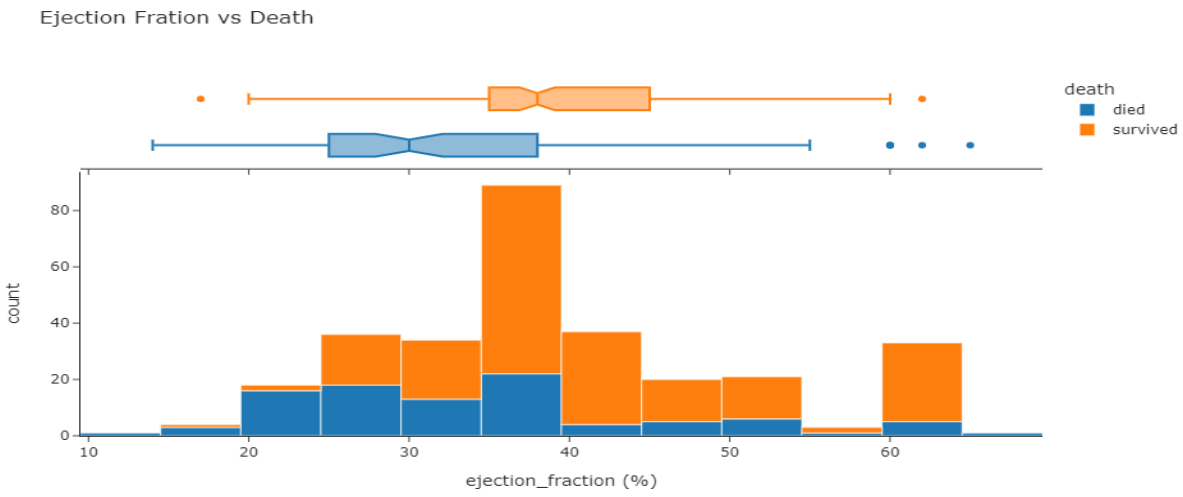

Figure 4: Time vs. Ejection Fraction

Figure 4 clearly indicates that the higher percentage of ejection fraction the higher survived chance. This means that percentage of blood leaving the heart at each contraction can predict the fatality chance of heart disease.

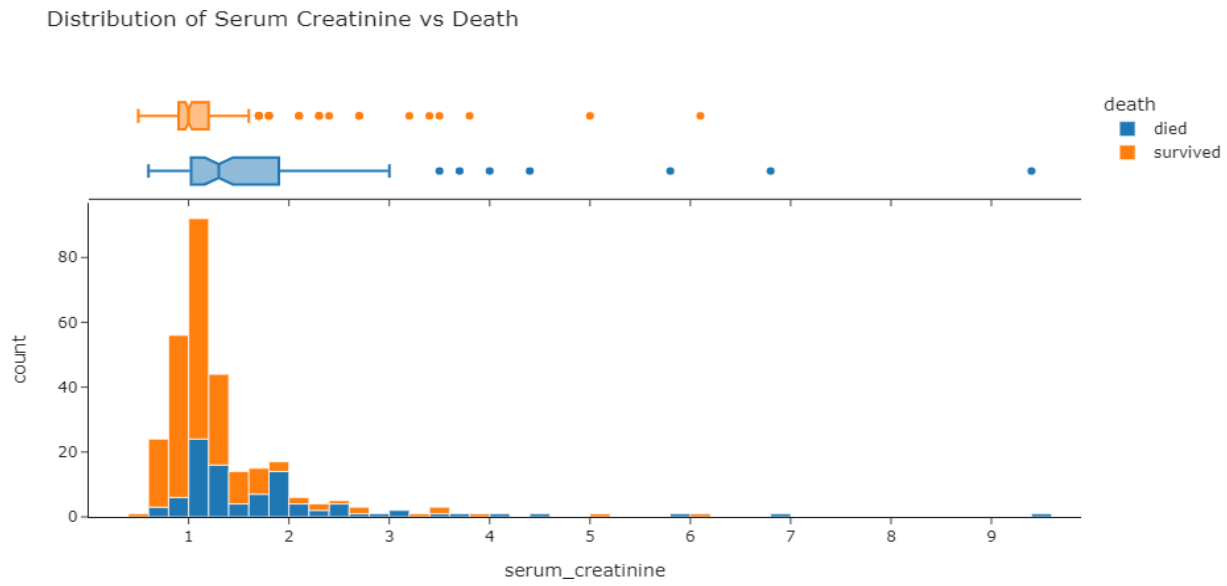

Figure 5: Death vs. Serum Creatinine

Figure 5 clearly shows that the higher amount of mg/dL serum creatinine the patient has the higher fatality chance they have.

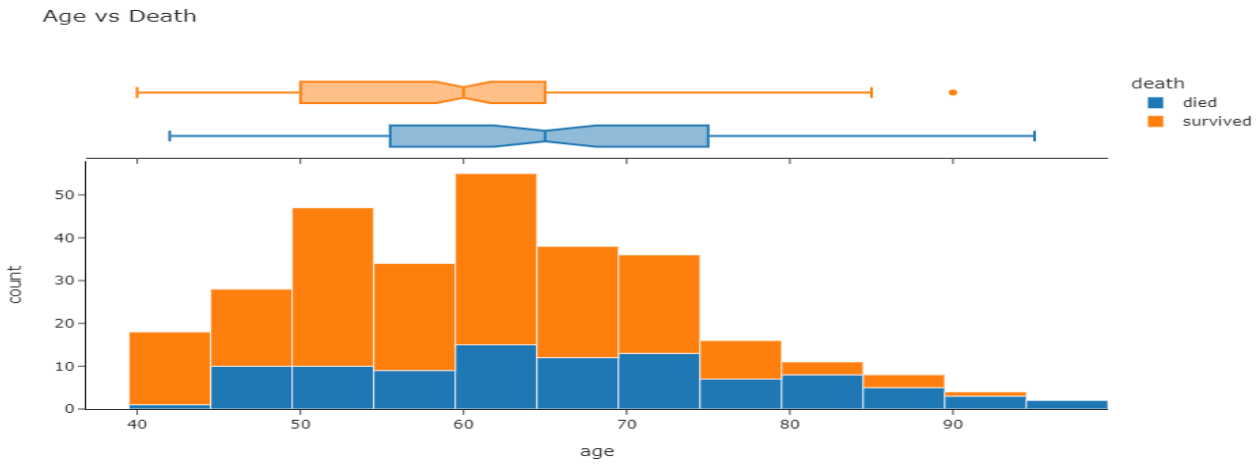

Figure 6: Death vs Age

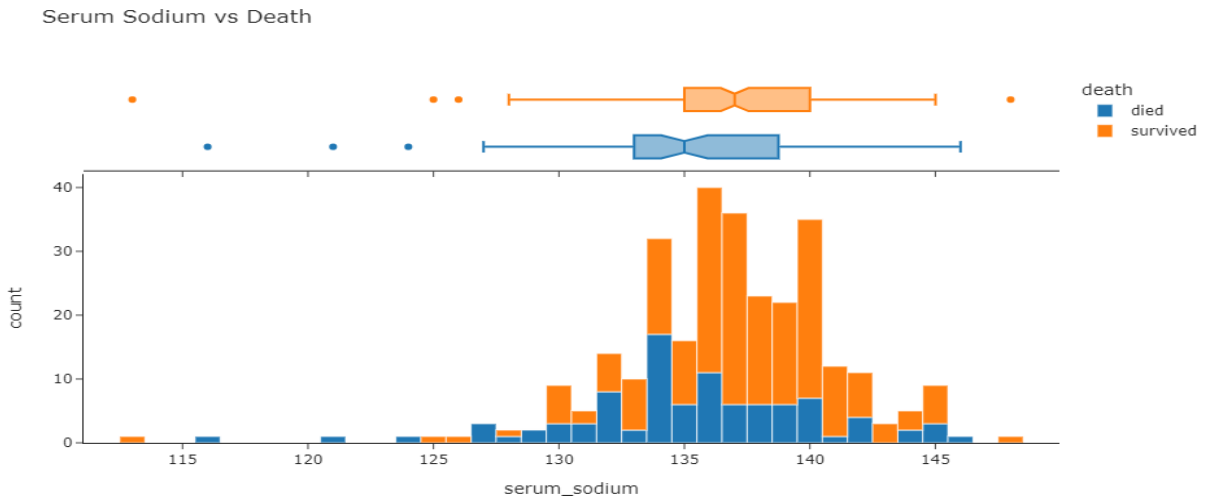

Figure 7: Death vs Serum Sodium

Figure 6 and figure 7 show the distribution of old people with higher serum sodium has more chance to get heart issues. However, it does not show very clear. Therefore we should consider whether if these two features are useful to predict heart disease.

Next, let's explore the correlation between features in this dataset.

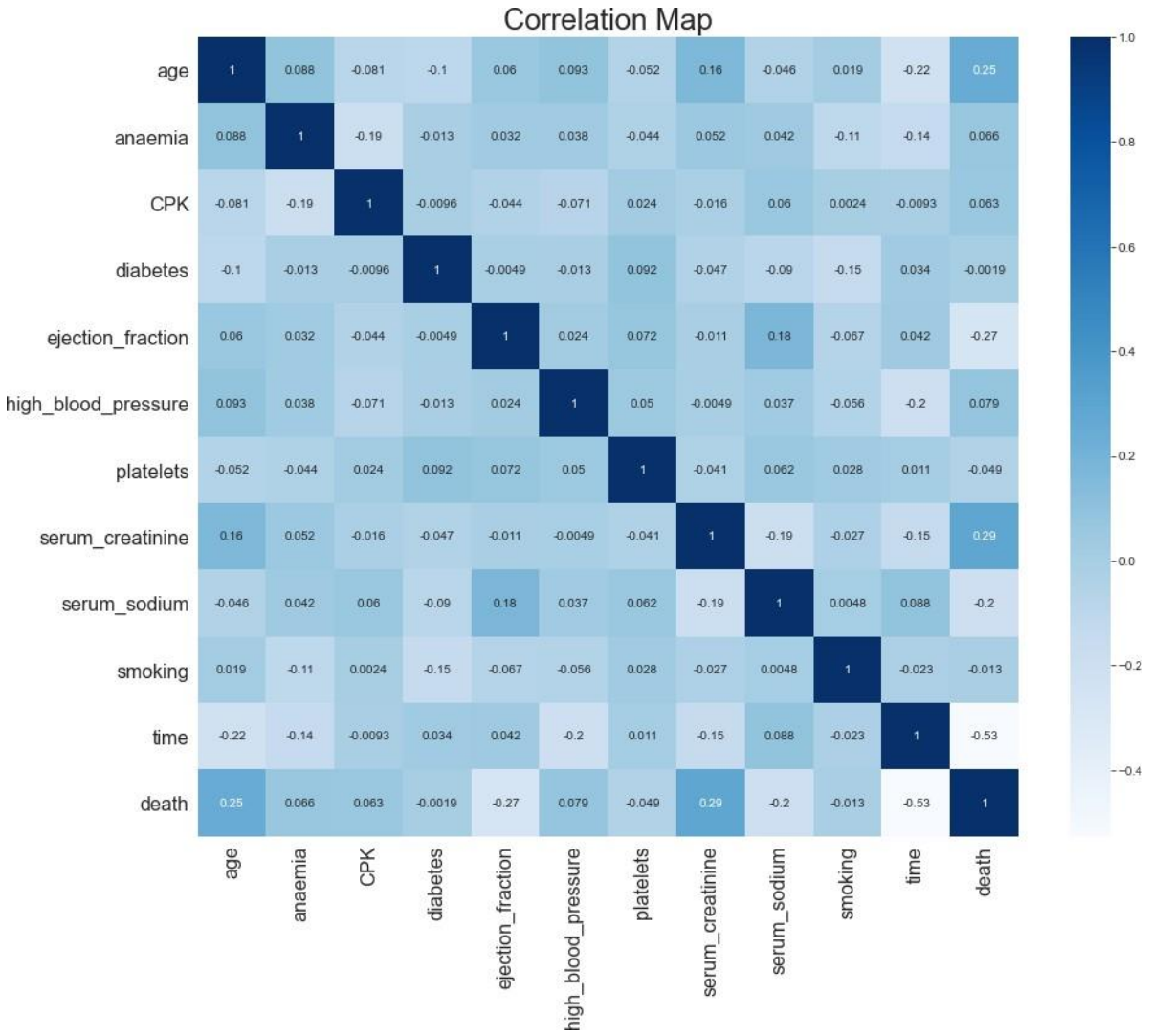

Figure 8: Correlation map of all features

Look at the last row of figure 8., we can see that time, serum creatinine, ejection fraction, age and serum sodium are the most correlated with the target variable ‘death’. Other variables have insignificant correlation with death.

2.3. Feature Importance

We’ll use Extra Decision Tree Classifier to rank the important level of the features.

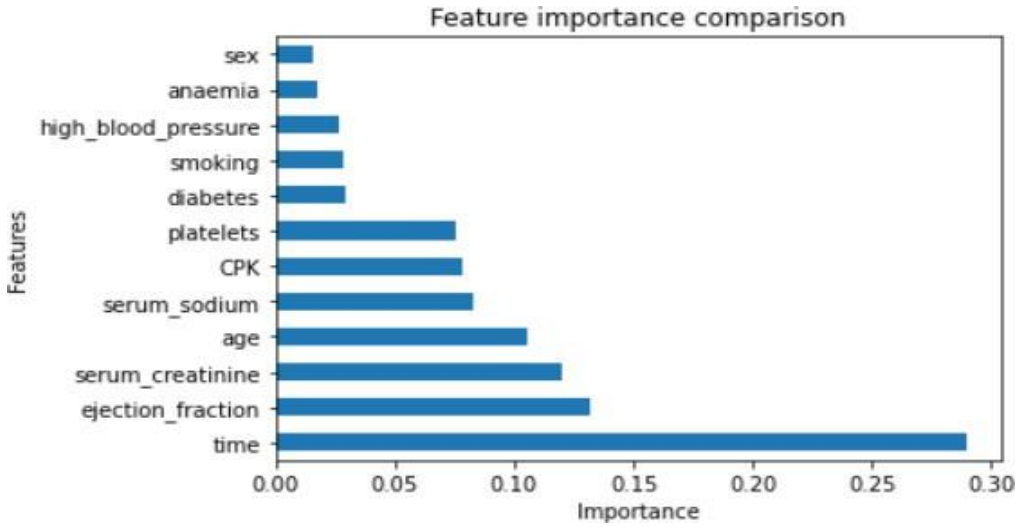

Figure 9: Feature importance ranking

Figure 9 clearly show that time, ejection fraction and serum creatinine are the three most important features. Hence, they will be used to predict death event in this project. Let's do further exploration the relationship between the target variable 'death' and these five features. We then can determine which feature are useful to predict heart disease.

3. Results

3.1. Overview

In this section, we employed several methods to predict the survival chance of the patients. We firstly demonstrated the results for survival prediction obtained on the whole dataset using KNN method. This showed how bad it is to predict survival on all 11 features. Next, we'll analyse the obtained results predicted on 3 selected features: time, ejection fraction and serum creatinine. In this step, we'll build two machine learning models using KNN and DTC. For both methods, we split the dataset into 207 selected patients for train set and 90 selected patients for test set. Our prediction results showed that KNN outperformed DTC by obtaining 91% accuracy prediction. The DTC obtained 87% accuracy. We'll test if the results can be improved by tunning parameters. The next step is to use analysis results of k-fold cross validation. We obtained the highest score of k-fold cross validation of 89% for KNN and 83% for DTC which is not much lower than our original predictions. Lastly, we'll compare accuracy rate between the models.

3.2. Model outcomes

3.2.1. KNN- Survival machine learning prediction on all 11 clinical features

The results show that survival prediction cannot be predicted on all given features.

	precision	recall	f1-score	support
0	0.73	0.80	0.76	65
1	0.32	0.24	0.27	25
accuracy			0.64	90
macro avg	0.52	0.52	0.52	90
weighted avg	0.62	0.64	0.63	90

Figure 10: Classification report of KNN using all 11 features

Figure 10 shows that the test set included 90 patients. The model can accurately predict 64% of the time. In other words, this model is predictable only 64 % correctly in the test set if they can be survival or not. The precision indicated that it is 32% of the time when the model predicted that a patients had heart failure. Similarly, the model was able to correctly predict only 24% of the time in patients who actually had heart disease.

3.2.2. KNN - Survival machine learning prediction on 3 clinical features: time, ejection fraction and serum creatinine

	precision	recall	f1-score	support
0	0.91	0.97	0.94	65
1	0.90	0.76	0.83	25
accuracy			0.91	90
macro avg	0.91	0.86	0.88	90
weighted avg	0.91	0.91	0.91	90

Figure 11: KNN classification report predicted on 3 features

In these results, the test set included 90 patients. The model can accurately predict 91% of the time. In other words, this model can predict 91% of the patients in the test set whether they have heart disease or not. The precision indicated that it is 90% of the time when the model predicted that a patients had heart failure. Similarly, the model was able to correctly predict only 76% of the time in patients who actually had heart disease.

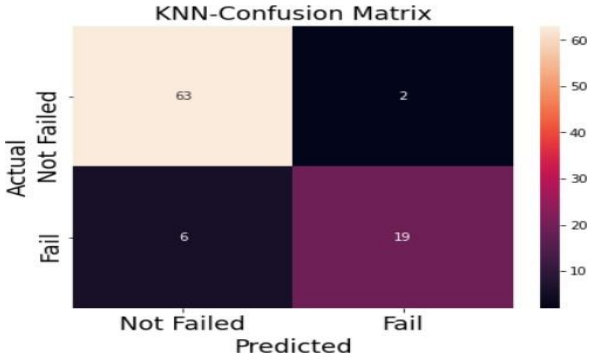

Figure 12: Confusion matrix

Figure 12 shows that KNN obtained the top results on the true positives and true negative. This means our models obtained good results.

3.2.3. Decision Tree Classifier - Survival machine learning prediction on 3 clinical features: time, ejection fraction and serum creatinine

	precision	recall	f1-score	support
0	0.91	0.91	0.91	65
1	0.76	0.76	0.76	25
accuracy			0.87	90
macro avg	0.83	0.83	0.83	90
weighted avg	0.87	0.87	0.87	90

Figure 12: Decision Tree classification report predicted on 3 features

In these results, the model can correctly predict 87% accuracy. In other words, this model can predict 87% of the patients in the test set whether they have heart disease or not. The precision indicated that it is 76% of the time when the model predicted that a patients had heart failure. Similarly, the model was able to correctly predict 76% of the time in patients who actually had heart disease.

Figure 14: Confusion matrix

Figure 14 shows that DTC obtained the top results on the true positives and true negative. This means our models obtained good results.

3.3. Comparison of two models

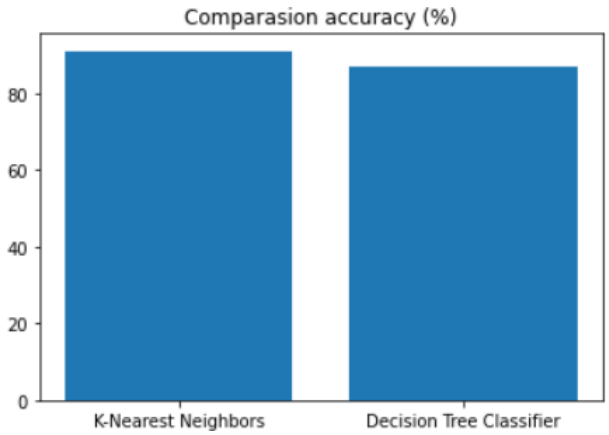

Figure 13: Comparation accuracy between KNN and DTC

Figure 13 visualises the percentage of prediction accuracy of two models built in this project. We can clearly see the model built using KNN obtained a higher accuracy rate compared to that's of DTC. Through these results, we can be recommended to use the KNN model to predict survival prediction of heart failure patients.

4. Discussion

The results showed that it's possible to predict whether a patient can survive with heart disease based on their follow-up period, ejection fraction and serum creatinine. Moreover, the accuracy rate is higher if the prediction made on these three features instead of all given features. These results are useful for medical settings in many cases, particularly when doctors could not obtain all clinical information, they are encouraged to at least analyse number of time patients made check-up visits, their ejection fraction and serum creatinine. These features can you them a correct prediction whether their patients can survive. Additionally, it is worth to notice there are additional studies need to be carried on ensuring this machine learning results can be utilised into medical settings.

Comparing to the original study which concluded that ejection fraction and serum creatinine are top two features achieved the highest prediction, our analysis achieved highest accuracy when prediction made on these features: time, ejection fraction and serum creatinine. The importance feature raking also generated different results which age is on 4th position and serum sodium is on 5th position. The last positions lie on sex and anaemia. While the original study ranked anaemia on the 3rd ranking position.

5. Conclusion

In this project, the feature importance model selected: time, ejection fraction and serum creatinine as top three relevant features. Moreover, the approaches in this project showed that we can use machine learning effectively with classification of health records of heart failure patients.

The results of this analysis can become a new efficient method for medical professionals who want to predict survival chance of patients. Therefore, this discovery is potential to effect in medical practice. Indeed, cardiovascular doctors can predict whether a patients could survive by concentrate on analysing three factors: their follow up period, ejection fraction and serum creatinine level.

However, this study exits several limitations. Firstly, the size of this dataset is too small which is observed on only 299 patients. Also, the dataset is very inbalanced between the number of survival and deceased observations. Secondly, the dataset missed out some potential features of patients including occupation history and exercise routine, physical features such as height, body mass index, weight. We can say a larger dataset with additional information would help us to generate more reliable results.

Regarding future development, we'll approach results using other classification method of machine learning such as Support Vector, Cat Boost, Random Forest. Clustering is also considered an effective method to obtain the good results for this project. Also, alternative datasets related to fatal disease including breast cancer, diabetes, cervical cancer will be applied using this machine learning approach.

References

Chicco, D., & Jurman , G. (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Medical Informatics and Decision Making*, 2.

Machine Learning Respository. (2020). Retrieved from Heart failure clinical records Data Set: https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records?msclkid=735f8bd7cf6e11ecb07f348bc3df59f6

What is heart failure. (2020). Retrieved from Heart Foundation: https://www.heartfoundation.org.au/conditions/heart-failure

WHO. (2021, June). *Cardiovascular diseases (CVDs)*. Retrieved from World Health Organization: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)?msclkid=dbcf2adecf6d11ecb2c4407ac41d1a70