Assignment 2

1. Cover Page

Title	Prediction of survival of heart failure patients using 2 different classification models
Author	1. Anson Go Guang Ping (s3767707) 2. Aaron Biju Mathews(s3854257)
Course number	COSC2670
Course name	Practical Data Science with Python
Affiliation	RMIT university
Contact details	1. s3767707@student.rmit.edu.au 2. s3854257@student.rmit.edu.au
Date of report	17/05/2021

Declaration and statement of authorship

1. I/we hold a copy of this work that can be produced if the original is lost/damaged.
2. This work is my/our original work and no part of it has been copied from any other student's work or from any other source except where due acknowledgement is made.
3. No part of this work has been written for me/us by any other person except where such collaboration has been authorized by the lecturer/teacher concerned.
4. I/we have correctly acknowledged the re-use of any of my/our own previously submitted work within this submission.
5. I/we give permission for this work to be reproduced, communicated, compared and archived for the purpose of detecting plagiarism.
6. I/we give permission for a copy of my/our marked work to be retained by the school for review and comparison, including review by external examiners. I/we understand that:
7. plagiarism is the presentation of the work, idea or creation of another person as though it is my/our own. It is a form of cheating and is a very serious academic offence that may lead to exclusion from the University. Plagiarized material can be drawn from, and presented in, written, graphic and visual form, including electronic data and oral presentations. Plagiarism occurs when the origin of the material used is not appropriately cited.
8. plagiarism includes the act of assisting or allowing another person to plagiarize or to copy my/our work.

Student signature/s

I/we declare that I/we have read and understood the declaration and statement of authorship.

1. *AnsonGo*

2. *AaronMathews*

2. Table of Content

Contents

3. Executive summary

The aim of the report is to build an ideal classification model that can predict the survival of heart failure patients using datasets which have similar features to Heart failure clinical records Data Set. Medical records of heart failure patients were retrieved, preprocessed and explored, and two classification models (k-nearest neighbors and decision tree) were trained using them. These models were then used to predict the survival of patients. Overall, the results indicate that the decision tree model has a better performance as compared to the other. The report concludes that decision tree model can predict the survival of heart failure patients more accurately. It is recommended that prediction model should be widely used in clinical fields to determine the conditions of heart failure patients to increase the accuracy of diagnosis.

4. Introduction

Cardiovascular diseases kill approximately 17 million people globally every year, and they mainly exhibit as myocardial infarctions and heart failures (The Guardian 2019). Heart failure occurs when the heart is unable to pump enough blood to the body, and it is usually caused by diabetes, high blood pressure, or other heart conditions or diseases (NHLBI n.d.). Available electronic medical records of patients quantify symptoms, body features, and clinical laboratory test values, which can be used to perform biostatistics analysis aimed at highlighting patterns and correlations otherwise undetectable by medical doctors. Machine learning can predict patients' survival from their data and can individuate the most important features among those included in their medical records (Chicco & Jurman 2020, p. 1). This report will discuss the steps in building our classification model while analyzing the relationship between the clinical features and survival of heart failure patients along the way.

5. Methodology

5.1 Data Retrieval and Preparation

For this report, we made use of the publicly available dataset which contained medical records of 299 patients who had heart failure, collected during their follow-up period, where each patient profile has 13 clinical features. This data was made to go through different data-preprocessing steps to make it ready for exploration and model building. After reading the data into a Jupyter Notebook, the shape of the data will be investigated. The next step is to investigate the data types of the features present in the dataset. It should match the description mentioned on the website. If there are any differences, it should be rectified. The data entry errors for all the features will be handled. This also handled any data entry errors for those features are as value other than 0 will be coerced as True, otherwise False. Then we look at the descriptive statistics of the data, and to check for unusual values. We then check whether there are any duplicate records present in the dataset. We also check for the presence of NA values. If any column contains more than 50% of the data as NA values there are dropped, otherwise they are replaced by the mean of the corresponding column. For the numeric features, we will use boxplots to look at the range of the features. The box plots will also be used to identify any outliers or data entry errors. A research paper (Chicco & Jurman 2020, p. 1) which uses this dataset, and has a similar goal has provided the range of all the features. This range values will be used to identify and eliminate any observations that can be considered as either outliers or data entry errors. To maintain the representation of the data, the outliers, and data entry errors, i.e.,

the values outside the range limits will be clipped instead of replacing it with mean. The feature was converted to match the range provided in the paper.

5.2 Data Exploration

5.2.1 Univariate analysis

After data preprocessing, the Heart failure clinical records Data Set should contain the following 13 features:

Clinical Features	Description	Data types
Age	Age of the patient (years)	float64
Anaemia	Decrease of red blood cells or hemoglobin (boolean)	bool
High blood pressure	If the patient has hypertension (boolean)	bool
Creatinine phosphokinase (CPK)	Level of the CPK enzyme in the blood (mcg/L)	int64
Diabetes	If the patient has diabetes (boolean)	bool
Ejection fraction	Percentage of blood leaving the heart at each contraction (percentage)	int64
Platelets	Platelets in the blood (platelets/mL)	float64
Sex	Woman or man (binary)	bool
Serum creatinine	Level of serum creatinine in the blood (mg/dL)	float64
Serum sodium	Level of serum sodium in the blood (mEq/L)	int64
Smoking	If the patient smokes or not (boolean)	bool
Time	Follow-up period (days)	int64
[target] Death event	If the patient deceased during the follow-up period (boolean) As it can be suspected based on the above features	bool

Table 1: Description of all clinical features and their data types (Dua, D. & Graff, C. 2019)

Before we start the data exploration step, we decided to drop the time feature from the dataset as from our point of view, the time post no effect on the physical condition of patients and is considered as a noise which interferes the model building process. First, we will look at the descriptive statistics of all numerical features after they have been preprocessed. By comparing the mean values with the real-world statistics of normal human, we can have an insight into the relationship between the features and the death event of heart failure patients. For example, a normal heart's ejection fraction may be between 50 and 70 percent (American Heart Association 2017) while the ejection fraction mean values of the dataset has a mean of 38 percent which indicates heart failure patients tend to have lower ejection fraction level and that might have minimal or significance effect on the death rate. For all the categorical features, we are plotting pie graphs to show the proportion of the relative category of that feature. It enables us to spot extreme distribution in categories which will affect our model prediction accuracy. For numerical features, we are plotting histograms for each feature to monitor their sampling distribution and to identify their behaviors.

5.2.2 Bivariate analysis

Scatter matrix is plotted to visualize two numerical features and boxplot is plotted to visualize numerical feature against categorical features. First, all features are plotted against death event to visualize their relationship with death event as it is our target feature, and we want to explore the correlation. To visualize the trend of categorical features against death event, we first calculate the rate of death event by dividing the count of true death event of that category by the total sample count of that category in that feature. Finally, a bar graph of the respective

feature against rate of death event is plotted to observe the difference in rate of death event between the two categories of that feature. This is because when a bar graph of death event grouped by a categorical feature is plotted, we interpret the rate of death event by looking at the size of gap of the plot between two categories. Difference in sampling distribution between two categories might cause the size of gap to bring misconception to their respective rate of death event. To visualize the trend of numerical features against death event, we first split the data into bins based on their quantile values to ensure each bin has the equal number of samples thus ensure the validity of the graph. Then the count of true death event is divided by the total sample count of that bin to calculate the rate of death event for each bin. Finally, a bar graph is plotted to observe the pattern of the rate of death event when the numerical values increase in a more direct way. Besides that, we also plot boxplot of numerical features against death event to compare the mean values of that features when grouped by death event. Moving to the next part, we choose ten pair of attributes which can generate plausible hypothesis and generate a visualization graph for each pair of attributes. Before moving to task 3, we find the pairwise correlation of all columns in the dataset which we may not discover.

5.3 Data modelling

We create 2 classification models (k-nearest neighbors and Decision Tree) using the preprocessed data. We will be choosing the best model of the two, to serve as an automation prediction model for real-world datasets with similar features as Heart failure clinical records dataset. Our model building process consists of the following stage:

1. Split the pre-processed data into train and test sets. The splitting would be stratified to retain the same proportion of target variable's classes in the train and test datasets. The test dataset will be used as the final test (unseen data).
2. Repeat the following steps for both the models.
3. Split the train data from step 1 into train and validate datasets. Even here, the splitting would be stratified. The performance of the trained model is validated on the validate dataset.
4. Select the best features by performing feature selection using hill climbing algorithm. The first parameter is selected only if it has an accuracy score of more than 50%.
5. Parameter tuning is performed to find the best parameters for the features selected in the previous step.
6. The train dataset from Step 1 is fitted on the model.
7. A new model is trained using the previously selected best features and optimal parameters. k-fold validation strategy is used to validate the previously trained model. We use the average score of the new model trained using k-fold strategy and compare it with the earlier trained model's accuracy. If there is a significant difference in the scores, we discard that model and go back to step 3 and start fresh with a new set of features. If there is not much difference between both the models, we continue forward.
8. The model is made to predict the test data and the performance of the model is evaluated.

6. Results

6.1 Data Preparation

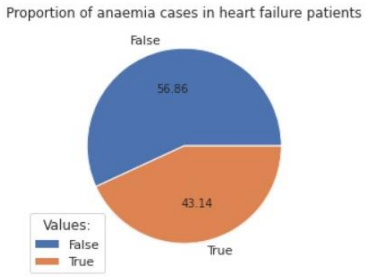
After loading the dataset into the Jupyter Notebook for analysis, the data was investigated upon to find and eliminate any issues if present. The shape of the data was observed and compared to the details present on the website (Dua, D. & Graff, C. 2019) and no issues were found. The datatype of the features was observed, and it was found that there is a difference in the datatypes present in the data from the ones mentioned on the website. The Boolean features were converted into numeric datatypes. Appropriate actions were taken to rectify it. When

Boolean features were updated as Boolean datatypes, that automatically eliminated any data entry errors from them. There were no duplicate values or NA values present. One feature(platelets) had the values in platelets/mL, but the website had the units in (kiloplatelets/mL). The boxplots identified many outliers in few numeric features, but there were all within the range mentioned in the paper (Chicco & Jurman 2020, p. 1) and hence are acceptable.

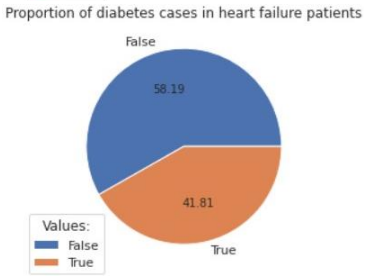
6.2 Data Exploration

6.2.1 Univariate Analysis

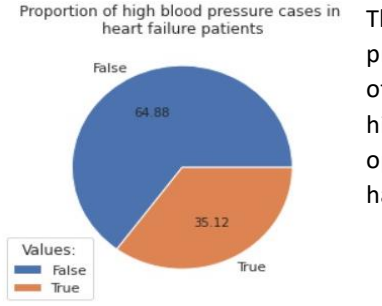
After preprocessing the data, we investigated each feature present, to find patterns present within it. Appropriate visualization graphs and descriptive statistics were produced to find the patterns and are mentioned below:



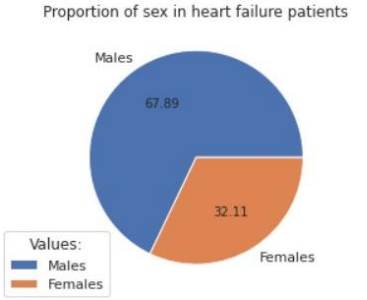
We could see that the data set is almost evenly split between the anemic patients. This visualization shows that decrease of red blood cells or hemoglobin does not directly contribute to heart failure, as both anemic and non-anemic patients can suffer heart failures.



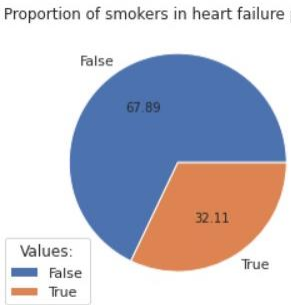
We could see that there is a slight increase in the number of patients who had heart failure but do not have diabetes as compared to patients with anaemia. Even this pie chart signifies that diabetes is not a major contributor towards heart failure.



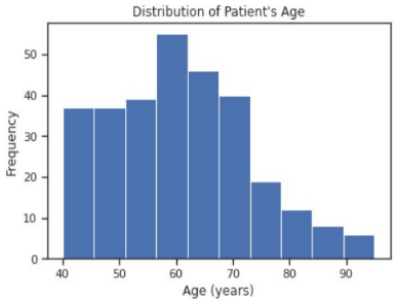
There is a significant increase in the number of patients without high blood pressure as compared to the earlier to features where patients were without other aliments but still had a heart failure. Contrary to the general notion that high blood pressure leads to heart failure, the data provides a completely opposite story. Almost 65% of the patients who suffered heart failure did not have high blood pressure.



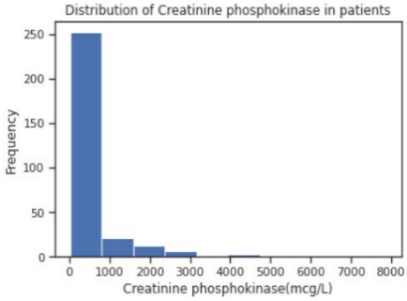
Males contribute to a high percentage of the number of patients with heart failure as compared to females. The percentage is almost double. Males and females can have different ranges of clinical features, as males tend to have a higher range of clinical features as compared to women.



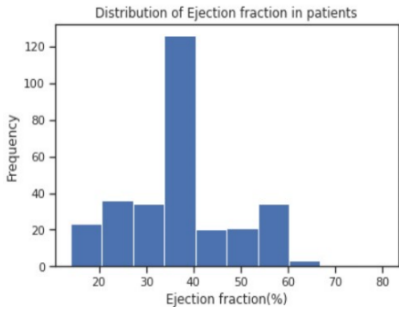
The data reported that only one-thirds of the patients with heart failures were smokers and the rest two-thirds were not.



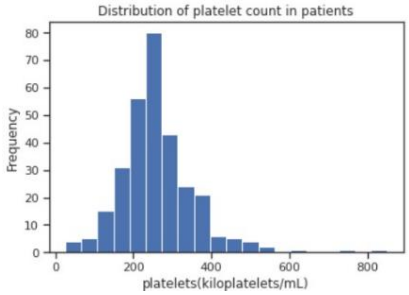
The age of the patients with heart failure lies between 40 to 95. All the patients are older adults and senior citizens. The distribution of the data does not look like a normal distribution, but the mean and median are almost the same at 60 years.



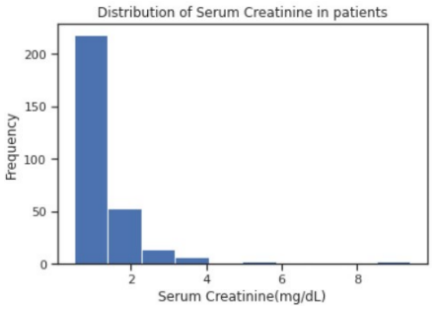
The distribution of creatinine phosphokinase in the dataset is highly positively skewed. It ranges from as low as 23 to as high as 7861. There are a few observations above 3000 which is stretching the graph and making it look more skewed. The mean value is 581.8 mcg/L while the median is 250 mcg/L. The normal range is from 10-20mcg/L (UCSF HEALTH 2019), but as it is seen, the smallest value in the dataset (i.e., 23) is above the normal range.



The ejection fraction is percentage of the blood pumped by the heart. A normal range for ejection fraction is 50-70% (American Heart Association 2017), but the data set has the range from 14-80%. The mean and the median are same with the value 38%. 75th percentile of the data lies below 45%, which is lower than the normal range for a healthy heart.



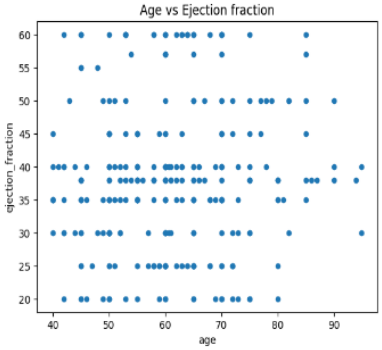
The distribution of the platelets count among the patients looks like a normal distribution. The mean and median are almost the same with the values 263 kiloplatelets/mL and 262 kiloplatelets/mL, respectively. The normal range of platelet count for a human body is between 150 to 450 kiloplatelets/mL (Marlene W., n.d). There are quite a few observations that are outside this normal range.



The distribution is a highly positively skewed distribution. The mean is 1.39 mcg/dL, and the median is 1.1 mcg/dL. The average range of the presence of serum creatinine in the human body is between 0.59 to 1.35 mcg/dL (Mayo Clinic Staff 2021). The minimum value present in the dataset is less than the normal range with the value 0.5 mcg/dL. The mean is also the outside the normal range. The maximum value recorded is a massive 9.4 mcg/dL.

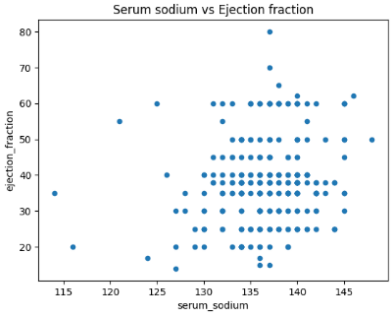
6.2.2 Exploring relationships between pair of attributes

By comparing all the clinical features with the death event, it is found out that sex, age, ejection fraction, serum creatinine has significance effect on the death event of heart failure patients. By getting a hold of this information, we can choose what attribute pairs to explore to find possible secondary relationship that may affect the death event. Moving on, we will be showing the graph of attribute pairs we choose to explore and their respective hypothesis.



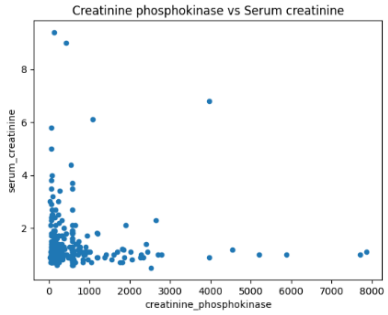
Hypothesis: If age increases, ejection fraction decrease. When age increases from 40, the body metabolic and organ function start to decline, thus causing ejection fraction to decrease.

Observation: No specific relationship is observed.



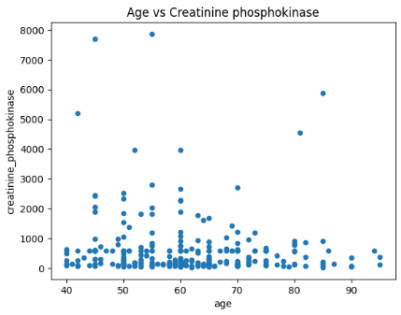
Hypothesis: If serum sodium level increases, ejection fraction decreases. High sodium intake reduces cardiovascular event which leads to lower ejection fraction

Observation: No specific relationship is observed



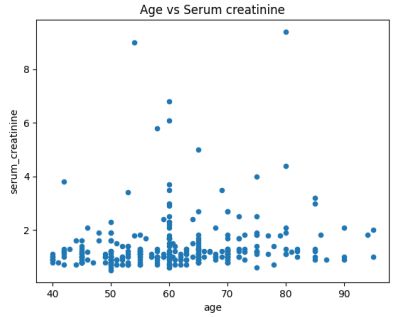
Hypothesis: If creatinine phosphokinase level increase, serum creatinine level decrease. Increase in creatinine phosphokinase speed up the process to break down serum creatinine, thus lower the level of serum creatinine in blood.

Observation: When creatinine phosphokinase level increases, serum creatinine level increases



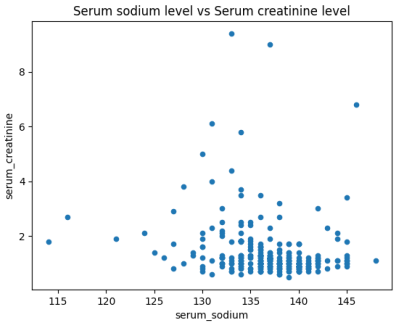
Hypothesis: If age increases, the creatinine phosphokinase level decreases. When age increases from 40, the body metabolic and organ function start to decline, thus causing number of creatinine phosphokinase enzyme to decrease.

Observation: When age increases, creatinine phosphokinase level decreases.



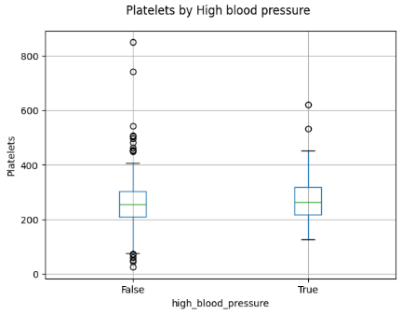
Hypothesis: If age increases, the serum creatinine level decreases. When age increases from 40, the body metabolic and organ function start to decline, causing number of creatinine phosphokinase enzyme to decrease, thus reduce the decomposing of serum creatinine in blood.

Observation: When age increases, a concave downward graph of serum creatinine level is observed



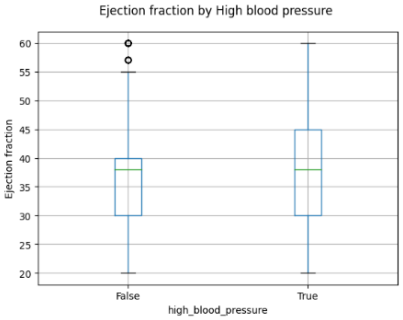
Hypothesis: If serum sodium level decreases, serum creatinine level decreases. Lower serum sodium level indicates better kidney function, thus more creatinine is removed from the body.

Observation: When serum sodium level decreases, serum creatinine level decreases



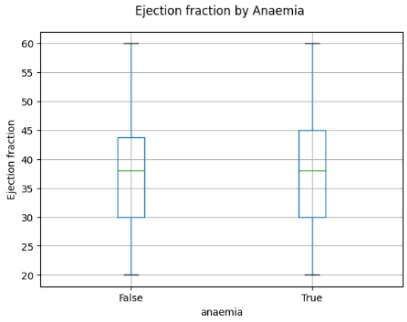
Hypothesis: If number of platelets is high, then the chance of getting high blood pressure is higher. Increase in platelets causes thrombocytosis which form blood clotting in blood vessels and narrow its radius, thus increasing blood pressure.

Observation: Patients with high blood pressure have slightly higher platelets mean count in their blood.



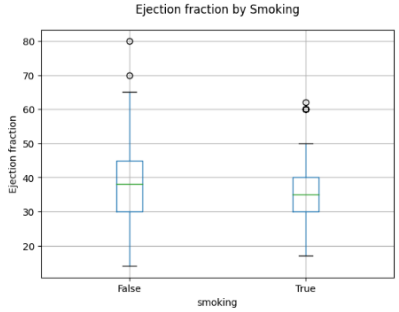
Hypothesis: If ejection fraction is low, then the chance of getting high blood pressure is lower. Lower ejection fractions mean lower pumping force from left ventricle of heart; thus, blood pressure will be lower.

Observation: Similar mean ejection value is observed but high blood pressure with true value has higher 75% quantile value of ejection fraction



Hypothesis: If patient has anaemia, then he will have higher ejection fraction. Anemic patients have lower oxygen carrying capacity in blood, so blood needs to be pumped faster throughout the body, thus increasing the ejection fraction.

Observation: No specific relationship is observed



Hypothesis: If patient has smoking habits, then the ejection fraction is lower. Harmful chemicals in cigarette cause damage to heart and blood vessels which leads to lower ejection fraction.

Observation: Smokers has lower mean ejection fraction

6.3 Data Modelling

6.3.1 Feature Selection and Parameter Tuning

Both the models went through feature selection process and out the 2, the decision tree had selected 5 features. KNN gave an accuracy score of 0.75 after using the 5 features. Even after selecting the most optimal parameter the score did not improve. While Decision Tree had the score of .75 by selecting just 2 parameters. Parameter tuning further improved Decision Tree's accuracy to 0.79.

6.3.2 Model Validation and Model Performance

Validation models of both KNN and Decision Tree had the same score of 0.74. This is quite close to the trained model's score and hence both the trained models are not discarded and used to predict the test data.

After selecting the best features and optimal parameters, both the models were tested on the unseen test data.

	KNN	Decision Tree
Accuracy	0.68	0.75
Misclassification Rate	0.32	0.25
Recall (True Positive Rate)	0.42	0.58
False Positive Rate	0.2	0.17
True Negative Rate	0.8	0.83
Precision	0.5	0.57
Prevalence	0.32	0.32
F-Measure	0.46	0.57

Table 2: Classification report of KNN and Decision Tree models

From the Table 2, we can see that the accuracy of KNN has dropped by 0.07, while the decision tree has a drop of a mere 0.04. The Decision Tree has a better performance as compared to the KNN model. Out of the 60 observations in the test data, the KNN predicted 33 True Negatives and 8 True Positives. The Decision Tree performed better and provided 33 True Negatives and 11 True Positives. The Decision Tree model has a better score in all categories including the prominent ones like Recall rate, Precision score, F-measure, etc. However, the difference in the rates is not much significant.

7. Discussion

The Decision Tree model has the upper hand as compared to the KNN model. It outperformed KNN on almost every aspect. It also takes uses less features and gives a better score. Decision Tree was able to provide a better classification using just serum creatinine and ejection fraction as compared to the KNN, where it used 5 features, but still had a less accuracy. The results are along similar lines of the research paper (Chicco & Jurman 2020, p. 1) that demonstrated that the 2 features (serum creatinine and ejection fraction) can predict better as compared to using other features. One thing to point out that distance-based models (e.g., KNN) are highly sensitive when features that have large differences between their ranges and can affect their accuracy, while it is not an issue for tree-based model (e.g., Decision Tree). Maybe performing standardization during pre-processing can improve the performance of the KNN model.

8. Conclusion

This report showed the potential in identifying key clinical features and focusing on them, rather than relying on all the data available. By having to focus on just serum creatinine and ejection fraction this model will help the doctors and the medical staff to initiate lifesaving treatments early and thus have a higher chance of helping a patient survive after a heart failure.

9. References

- The Guardian 2019. UK heart disease fatalities on the rise for first time in 50 years, viewed 17 May 2021, <https://www.theguardian.com/society/2019/may/13/heart-circulatory-disease-fatalities-on-rise-in-uk>
- National Heart Lung and Blood Institute (NHLBI). Heart failure, viewed 17 May 2021, <https://www.nhlbi.nih.gov/health-topics/heart-failure>
- Chicco & Jurman 2020, ' Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone ', BMC Medical Informatics and Decision Making, vol. 20, no. 16, pp. 1, https://bmcmedinformdecismak.biomedcentral.com/track/pdf/10.1186/s12911-020-1023-5.pdf
- Dua, D. and Graff, C. 2019, UCI Machine Learning Repository, "Heart failure clinical records Data Set", Irvine, CA: University of California, School of Information and Computer Science, viewed 22 May 2021, <https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records>
- American Heart Association 2017, Ejection Fraction Heart Failure Measurement, American Heart Association, viewed 17 May 2021, < https://www.heart.org/en/health-topics/heart-failure/diagnosing-heart-failure/ejection-fraction-heart-failure-measurement#:~:text=A%20normal%20heart's%20ejection%20fraction,failure%20with%20preserved%20ejection%20fraction>
- University of California San Francisco 2019, MEDICAL TESTS- Creatine phosphokinase test, viewed on 17 May 2021, <https://www.ucsfhealth.org/medical-tests/003503#:~:text=Normal%20Results,per%20liter%20(mcg%2FL)>
- Marlene W., M.D., JOHNS HOPKINS MEDICINE, What are Platelets and Why are They Important?, viewed on 17 May 2021, < https://www.hopkinsmedicine.org/health/conditions-and-diseases/what-are-platelets-and-why-are-they-important#:~:text=A%20normal%20platelet%20count%20ranges,150%2C000%20is%20known%20as%20thrombocytopenia>
- Mayo Clinic Staff 2021, MAYO CLINIC, Creatinine tests, viewed on 17 May 2021, <https://www.mayoclinic.org/tests-procedures/creatinine-test/about/pac-20384646#:~:text=The%20typical%20range%20for%20serum,52.2%20to%2091.9%20micromoles%2FL>