Student Name: Monkolsophearith Prum (s3848409)

Dulguun Bayarmaa (s3761982)

|  |
| --- |
| We certify that this is all our own original work. If we took any parts from elsewhere, then they were non-essential parts of the assignment, and they are clearly attributed in our submission. We will show we agree to this honour code by typing "Yes": *Yes*. |

**Heart Failure Prediction using K-Nearest Neighbors and Decision tree classifiers.**

Monkolsophearith Prum

RMIT, School of Science

[S3848409@student.rmit.edu.au](mailto:S3848409@student.rmit.edu.au)

18 May 2021

Dulguun Bayarmaa

RMIT, School of Science

[S3761982@student.rmit.edu.au](mailto:S3761982@student.rmit.edu.au)

18 May 2021

**Table of Contents**

[**Executive Summary** 3](#_Toc72475919)

[**Introduction** 3](#_Toc72475920)

[**Methodology** 3](#_Toc72475921)

[ **Data Retrieving and Preparing** 4](#_Toc72475922)

[**Results** 4](#_Toc72475923)

[ **Data Exploration** 4](#_Toc72475924)

[o **Summary of features** 4](#_Toc72475925)

[o **Exploring relationship between pairs** 6](#_Toc72475926)

[ **Data Modelling** 8](#_Toc72475927)

[o **K Nearest Neighbors** 8](#_Toc72475928)

[o **Decision Tree** 10](#_Toc72475929)

[**Discussion** 11](#_Toc72475930)

[**Conclusion** 11](#_Toc72475931)

[**References** 11](#_Toc72475932)

# **Executive Summary**

The aim of this project is to predict survival of 299 patients with heart failure with machine learning modelling methods and search for the most accuracy between two classification models, K-nearest neighbour and Decision Tree using Python Programming language and libraries.

The data were summarised and plotted on various ways based on whether they are categorical or numerical data. It is found that 67% of patients (more than half) with heart failure condition did survive. It is also worth noting that, we found out the deceased ones had no high blood pressure, no diabetes, no anaemia and did not smoke. This can be said that the anyone can develop a heart failure without having diabetes, high blood pressure and anaemia. Male patients also showed a higher death rate due to heart failure. However, the data between male and female is unbalanced therefore, this figure may be inaccurate. Additionally, z test hypothesis testing showed that there are significant relationships between age and death, ejection fraction and death, serum creatinine and death and serum sodium and death. This result indicates that further prediction could focus on the blood test result since there is no clear evidence of anaemia, diabetes, high blood pressure and gender having impact on death event for heart failure patients.

To sum up, when using 30% of overall data, it has been found that K-nearest neighbours provide the most accuracy of 87% which outperformed Decision Tree by 1%. Further recommendations include looking into more and better features to extract from this data, and to research using various window lengths.

# **Introduction**

Heart failure is a serious health condition that can happen suddenly or develop over time in both children and adult in which the heart fails to pump enough blood. This may happen when the heart is too weak, or it is not filled with enough blood. There are two types of heart failure: acute and chronic. Heart failure is caused by other medical conditions such as high blood pressure, irregular heartbeat, coronary disease etc. Heart failure may lead to kidney and liver diseases as well as other heart conditions.

In this report, we will discuss research into patients with heart failure, their survival rate and factors that contribute to the heart failure. This dataset is sourced from [UCI Machine Learning Repository](https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records) (Davide Chicco, 2020). The dataset contains the medical records of 299 heart failure patients and collected originally at the Faisalabad Institute of Cardiology and at the Allied Hospital in Faisalabad (Punjab, Pakistan), during April–December 2015. The patients consisted of 105 women and 194 men, and their ages range between 40 and 95 years old. The dataset contains 13 features that indicate their lifestyle, habit, and clinical results.

# **Methodology**

The analysis was done using analytical libraries: pandas, numpy, matplotlib, sklearn and seaborn libraries. We used machine learning (classification) methods to predict the survival of the patients. The classification methods used in this prediction are Decision Tree and KNN (K-Nearest Neighbors). We measured the prediction results through common confusion matrix rates as well.

The analysis was divided into three tasks, data retrieving and preparing, data exploration and data modelling.

# **Results**

## **Data Retrieving and Preparing**

For the first task, pandas and other libraries was first imported to load the data set from CSV file. The dataset was successfully uploaded and ready to be explored. We used “info ()” function from pandas to check the feature types and found no missing values. Out of all variables, 3 features are “float64” type while the rest are “int64” data types. Some features are binary: anaemia, high blood pressure, diabetes, sex, and smoking.

Text

Description automatically generated with low confidence

For categorical variables, they were converted into strings to plot their numerical features for task 2.

## **Data Exploration**

In task 2 data exploration, data set was summarised, plotted on different graph types and divided into 2 categories, numerical variables and categorical variables for more detailed analysis.

### **Summary of features**

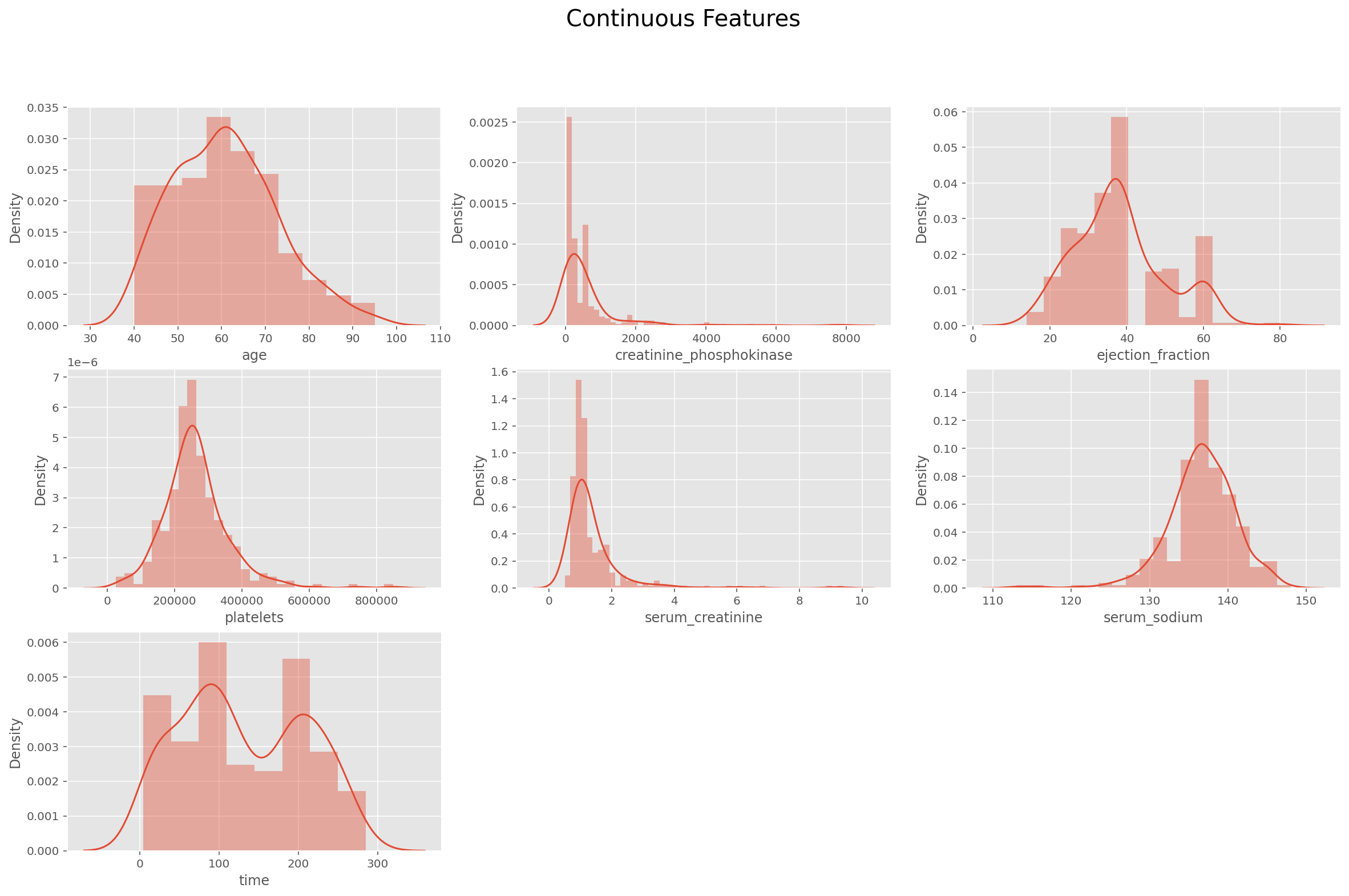
Function ‘df.describe()’ was used to perform summary statistics.

The numerical variables are age, creatinine phosphokinase, ejection fraction, platelets, serum creatinine and serum sodium. The age range of patients appears to be between a minimum of 40, mean of 61.8 (average age) and a maximum of 95 respectively. The creatinine phosphokinase level ranges between a minimum of 23 and a maximum of 7861 with a mean of 581.8. The ejection fraction ranges between a minimum of 14 to a maximum 80 with a mean (average) of 38 in the summary. For platelets, its range is between a minimum of 25100 and a maximum of 850000 with a mean of 263358. The level of serum creatinine ranges between a minimum of 0.5 and a maximum of 9.4 with a mean 1.39. The patients' blood sodium level ranges between 113 and 148 with a mean of 136.6. The normal range is between 135 and 145.

The categorical features are anaemia, diabetes, high blood pressure, sex and smoke. There are 129 patients with anaemia, 125 patients with diabetes, 105 patients with high blood pressure, 105 female and 194 male patients and 96 patients who smoke.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Patient with / who are/ who do | Anaemia | Diabetes | High blood pressure | Sex: Male / Female | Smoke |
| Yes - 1 | 129 | 125 | 105 | 105 /Female | 96 |
| No - 0 | 170 | 174 | 194 | 194 /Male | 203 |

* + **Exploring each column**



As seen in the density plots above:

The continuous (numerical) features are not normally distributed and contain outliers. Creatine\_phosphokinase, serum\_creatinine, ejection\_fraction and serum\_sodium features are skewed. We’ll explore them in detail to see if they affect the DEATH\_EVENT.

Chart, bar chart

Description automatically generatedChart, pie chart

Description automatically generated

The number of male and female patients in this dataset is imbalanced therefore it might give misleading information.

In other graph of categorical features, the number of patients who have anaemia, diabetes, high blood pressure and who smoke are less than those who has no diabetes, anaemia, high blood pressure and who does not smoke. We will analyse whether these factors contribute to the survival of the patients. (Please refer to the ipynb file).

### **Exploring relationship between pairs**

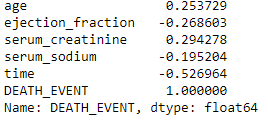
Chart, bar chart

Description automatically generated

We do not see that the anaemia, diabetes, high blood pressure and smoking habit affect the death event in the graph above. Because the patients who have anaemia, diabetes, high blood pressure and who smoke have lower death event.

Chart, bar chart, treemap chart

Description automatically generated



We can see columns:

'serum\_creatinine' and 'age' show a positive significant correlation with the DEATH\_EVENT.

Columns:

'time' 'ejection\_fraction' 'serum\_sodium' show a negative significant correlation with the DEATH\_EVENT.

Text Box

Diagram

Description automatically generated

To confirm our investigation, the hypothesis testing was done for pair of features using ztest with significance level 0.05. Our findings are:

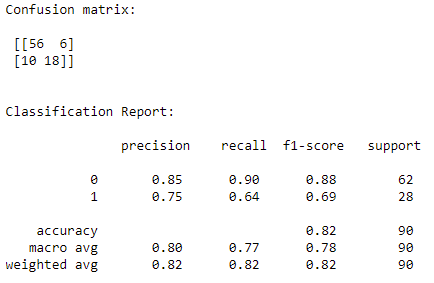
* Hypothesis\_Status for age = Reject Null Hypothesis : Significant
* Hypothesis\_Status for anaemia = Fail to reject Null Hypothesis : Not Significant
* Hypothesis\_Status for creatinine\_phosphokinase = Fail to reject Null Hypothesis : Not Significant
* Hypothesis\_Status for diabetes = Fail to reject Null Hypothesis : Not Significant
* Hypothesis\_Status for ejection\_fraction = Reject Null Hypothesis : Significant
* Hypothesis\_Status for high\_blood\_pressure = Fail to reject Null Hypothesis : Not Significant
* Hypothesis\_Status for platelets = Fail to reject Null Hypothesis : Not Significant
* Hypothesis\_Status for serum\_creatinine = Reject Null Hypothesis : Significant
* Hypothesis\_Status for serum\_sodium = Reject Null Hypothesis : Significant
* Hypothesis\_Status for sex = Fail to reject Null Hypothesis : Not Significant
* Hypothesis\_Status for smoking = Fail to reject Null Hypothesis : Not Significant
* Hypothesis\_Status for time = Reject Null Hypothesis : Significant

## **Data Modelling**

Finally, for the third task, data modelling, the data was treated as classification task. KNN classifier and Decision Tree Visualization were the two classification models used. To train the data, death event was set as y (the target variable) and ejection fraction, serum creatinine, time, serum sodium and age were set as the test/train subjects. The data then were split into test set and training set with the test size of 30%.

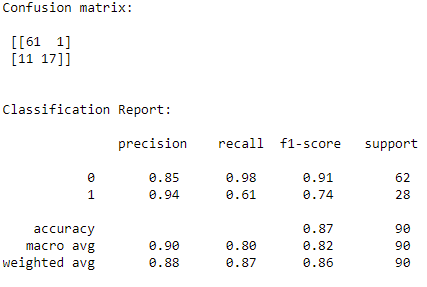
### **K Nearest Neighbors**

The K neighbours classification conducted as part of the project demonstrated how various input parameters or data subsets can have a significant effect on the output. An initial analysis was conducted on a subset of the data to investigate the influence of various input parameters. We started the first analysis using default settings for weight, metrics, and p (3). The outcomes were quite successful with an accuracy of 82%.



To see the differences of the accuracy, we basically create a for-loop to model the data using KNN where “n\_neighbors” ranging from 1 to 30 and appended it to an array. We found the lowest error rate of 0.12 at K value equals 8.

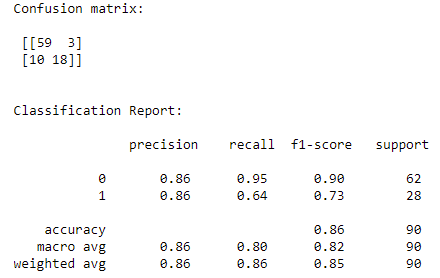


By setting 8 for “n\_neighbors”, we performed the analysis again and have achieved the best accuracy of 87% which confirmed the best results of using KNN.

### **Decision Tree**

For Decision Tree model, we used the same train/test data with the same test size. The steps are identical to K Nearest Neighbors, except that this time we experimented with different values for the hyperparameters 'depth' and 'min samples split' to see which produced the best results. A depth of 5 seemed like a good compromise, any further depth would risk overfitting the platform. Setting the hyperparameter 'min samples split' to 40 reduces the chance of overfitting much further without impacting the test result. We will ensure that all inner nodes have at least as many class instances by setting the minimum samples split.

After the analysis, we got the result from using the model, the accuracy achieved was 86%. It is found that for this dataset, Decision Tree is almost as good as KNN.

The graph for this model can be open using “Graphviz.exe”. To open it, make sure you have Graphviz installed and open the “heart\_failure.dot” using the program.

# **Discussion**

In Task 2, we expected the high blood pressure, smoking habit and diabetes would have conjunctions to DEATH\_EVENT or impact the cause of death in this dataset of 299 patients. After analysing, it turned out that these did not affect the cause of death but level of the creatinine phosphokinase, ejection fraction, level of creatinine and level of sodium level have significant roles in cause of death among deceased ones.

Heart failure data is best classified with K-nearest neighbour as it gives the highest level of accuracy although the accuracy of decision free is not far behind. If the depth has been increased, decision might have performed even better than K-nearest neighbour. However, doing so would risk overfitting.

K-nearest neighbour scores for f1 and accuracy level are only 0.01 higher than that of decision tree. Therefore, it can be said that in classifying heart failure data, there exists a slim preference, if any, of choosing one model over the other in term of accuracy.

With the occurrences of survival (denoted with 0) rate being nearly twice as much death (denoted with 1), the data is uneven. This unevenness, thus, has impact on the precision of both models. For example, K-nearest neighbour and decision tree models managed to classify survival rate or ‘0’ with great precision as f1-scores were at 91% and 90%, respectively. However, their precision for classifying death or ‘1’ is only at 74% for K-nearest neighbour and 73% for decision tree.

# **Conclusion**

In conclusion, in classifying data for heart failure, K-nearest neighbour model shows a slightly more accurate level than decision tree. Therefore, it would be unfair to conclude that it is the ‘best’ model rather than a ‘better’ model as its accuracy level is only 0.01 more superior. Additionally, the test could be improved by using balanced instances of each of the data when making the model. Also, we found that high blood pressure, diabetes, anaemia, smoke factors were not the causes of death for the heart failure patients in this project. Main factor for death event was level of the creatinine phosphokinase, ejection fraction, level of creatinine and level of sodium level in blood. However, more research studies with more patients needs to be done for accurate research conclusion.

**References**

# David Chicco & Gieseppe Jurman 2020, *Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone, viewed 15 May 2021,* [*https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5*](https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5)

National Heart, Lung, and Blood Institute 2021, *Heart Failure Also Known as Congestive Heart Failure,* NIH, viewed 19 May 2021, https://www.nhlbi.nih.gov/health-topics/heart-failure

Subhash Meena 2020, Statistics for Analytics and Data Science: Hypothesis Tesing and Z-Test vs T-test, viewed 22 May 2021, <https://www.analyticsvidhya.com/blog/2020/06/statistics-analytics-hypothesis-testing-z-test-t-test/>

Yogesh Agwaral 2019, Hypothesis Testing in Machine Learning using Phyton, viewed 21 May 2021, <https://towardsdatascience.com/hypothesis-testing-in-machine-learning-using-python-a0dc89e169ce>