# **Assignment 2:Data Modelling and Presentation**

**Prediction of Heart Failure Survival**

**19th May 2021**

Team members details

Student ID: s3841545

Student Name: Sai Ramishetty

Student email: s3841545@student.rmit.edu.au

Student ID: s3815294

Student Name: Manav Makkar

Student email: s3815294@student.rmit.edu.au

Lecturer:

Dr Yongli Ren

|  |
| --- |
| 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 honor code by typing "Yes": *Yes*. |

# Table of contents

1. **Abstract/Executive Summary**………………………………………………………….

2. **Introduction**…………………………………………………………………………………….

3.**Methodology**…………………………………………………………………………………….

4.**Results**………………………………………………………………………………………………

5.**Discussion**………………………………………………………………………………………….

6.**Conclusions**………………………………………………………………………………………..

7.**References**………………………………………………………………………………………….

8.**Appendix**……………………………………………………………………………………………..

**Abstract/Executive Summary**

Heart failure occurs when the blood pumped by the heart is not enough to satisfy the needs of a person. In North America, approximately 379,800 people have died in 2018 due to heart failure. There are several factors that can show impact on how our heart operates. These factors can be helpful in identifying potentially high risked heart failure patients. Machine learning is a tool that can be utilised to predict whether a patient can survive a heart failure and can identify important features/body factors that have a link to heart failure. For this study, we will be using two prediction classification models, kNN and Decision Tree, and determine which model will be used for making accurate predictions.

**Introduction**

Heart failure is a deadly condition that accounts for about one in 50 deaths in Australia. Despite the fact that twice the men experience heart failure than females, it is reported that females are more vulnerable to die from heart failures([Key Statistics: Heart Failure | The Heart Foundation](https://www.heartfoundation.org.au/activities-finding-or-opinion/key-statistics-heart-failure)). On the whole, around 30,000 Australians with heart failures are diagnosed on average every year. There are many symptoms associated with heart failures, such as irregular heartbeat, swelling in legs, dizziness, etc([Heart failure - treatment, causes, living with it and more | healthdirect](https://www.healthdirect.gov.au/heart-failure)). However, there are other health factors that contribute to those symptoms. In fact, those factors are the ones which can determine whether a patient can survive a heart failure or not. Through the reports above, we have gone through the general facts, but for medical treatment, it would be beneficial for the doctors to know whether there is a particular health factor/s that could predict the survival of the patient.

**Methodology**

***Data***

The dataset being used is from [UCI Machine Learning Repository: Heart failure clinical records Data Set](https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records) . Originally, the dataset was collected by Tanvir Ahmad, Assia Munir, Sajjad Haider Bhatti, Muhammad Aftab, and Muhammad Ali Raza (Government College University, Faisalabad, Pakistan). However,  Davide Chicco (Krembil Research Institute, Toronto, Canada) elaborated on the dataset and sent it to University of California Irvine Machine Learning Repository. The data consists of 299 patients in total. The survival is indicated by the variable DEATH\_EVENT. The number 0 means the patient survived, while 1 means the patient died.

***Data Analysis tools***

The Integrated Development Environment selected for the study is Jupyter Notebook. Packages such as pandas, matplotlib, numpy, seaborn, sklearn and math were used to conduct the analysis. The first four packages were mainly used in Data preparation and Data exploration, while sklearn was used for Data Modelling.

***Classification Models***

**KNN(k-Nearest Neighbours Classifier)**

The KNN classifier classifies a data point based on how its neighbour is classified. The letter k represents the number of neighbours near to the new data point. Below is an image displaying how KNN classification process works.Diagram

Description automatically generated

*The star in this image has many neighbours around it, however, when classifying it, the k value determines whether it is class A or class B. When k=3, the star is classified as class B, but, when k=6, it is classified as class A.*

KNN classifier is excellent to use for this study. This dataset has labelled data for the target feature DEATH\_EVENT, which works well for KNN model. Another thing to note is our dataset size is 299, which fairly small.

The perfect value of k is dependent on two things

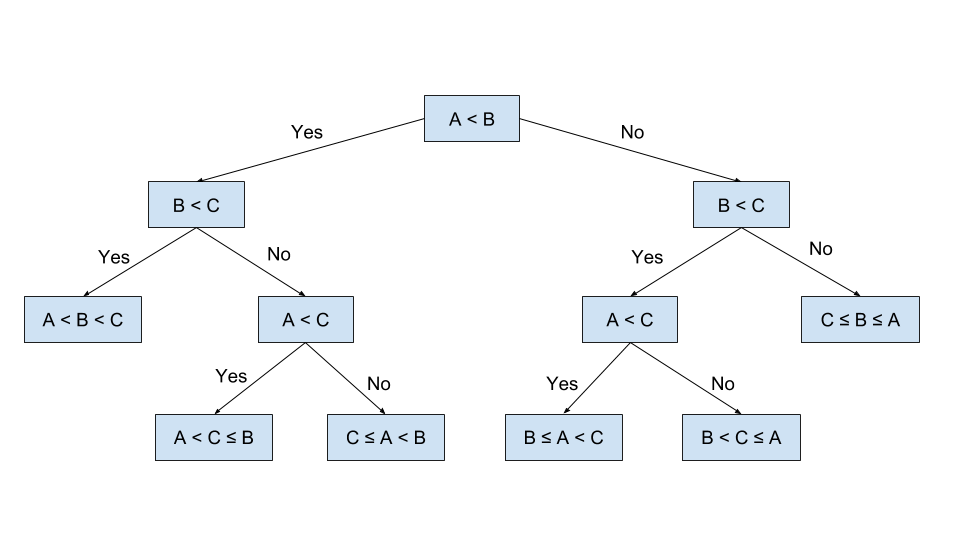
* √(size of y test)[square root of y test size].
* The k value has to be odd as confusion can be avoided

This model will be trained using K-folds validation and Leave-1 out. The parameter n\_splits for K folds will be 5 since our dataset is relatively small and it is a default value. On the other hand, Leave-1 out validation will check the absolute error of predicting actual observations. Both these validations will give an idea on whether the model is able to accurately make predictions. Finally, the model will be broken into two, standardised and non-standardised. This is to make sure we analyse how the presence of outliers affects the accuracy of the KNN model making predictions.

**Decision Tree Classifier**

Decision trees are the one of the commonly used machine learning models. They are non-parametric models which learn by recursively split the predictor space according to the best feature till the tree reaches a reserved depth. Ahead the subsets contain the elements of one class only. The best thing about decision tree model is that if it is used to discriminate samples, it allows us to get the best possible according to the measure (Gini index).

where 𝑝 is the ratio between number of samples of class 𝑗j and total number of samples.



*Here it can be clearly seen for decision tree model works on simple conditions as per dataset.*

The decision trees are simple to understand and can handle both numeric and categorical data. In addition, they require less effort for data preparation and are not affected by any non-linear relationships between the parameters.