Data Modelling And Presentation

Student ID: S3901999

Affiliations*: RMIT University*

Student name and email: Mohammed Usman E Ghani, *S3901999@student.rmit.edu.au*

Date of report: 25/04/2022

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

Table of Content **page**

Abstract

In recent years, machine learning has been widely implemented in the field of medical research to develop new medical procedures, handle patient records and in developing treatments for chronic diseases like heart failure. The objective of this research project is to predict heart failure patients' survival based on the medical records (various clinical features) on their pre-existing health conditions and their everyday habits.

A dataset containing medical records of 299 heart failure patients from the Institute of Cardiology and Allied hospital was prepared with the help of pandas library in order to ensure good quality data. In order to get a visual context and understanding of the nature of data, various graphs were plotted which provided further insights into meaningful relationships amongst the features. After this, two predictive models, K-Nearest neighbour and decision tree classifier models were selected and built which were further optimised with applying feature selection and hyperparameter tuning appropriately as required. The models were then validated and their performances were evaluated and compared by obtaining accuracy scores, precisions, confusion matrices and various other metrics from the performance reports. These results were thoroughly analysed in order to reach appropriate conclusions.

Keywords: Heart failure, K- nearest neighbour (KNN), decision tree classifier, hyperparameter tuning.

Introduction

Cardiovascular disease (CVDs) is the leading cause of death in humans around the world. It affects around 17.9 million lives each year, representing 32% of all the global deaths (*Cardiovascular*, 2022). Heart failure is a long term condition where the heart muscles are inadequate to pump blood to the rest of the body the way it normally should. High blood pressure, smoking and cholesterol levels are some of the key factors associated with heart failure. Machine learning can be essential in determining complex clinical characteristics of a patient's available medical records in order to achieve better evaluations of the condition. The goal of this project is to predict heart failure patients' survival based on the data in their medical records which will enable potential development of future studies to investigate early treatment of high-risk heart failure patients. Moreover, improved medical decision-making can also be achieved by identifying key risk factors associated with heart failure.

The Heart Failure Clinical Dataset contains medical records (13 clinical features including a target variable) of 299 patients with heart failure from the Institute of Cardiology and Allied hospital, collected during their follow-up period (all patients in the dataset had experienced heart failure after which they were placed into a follow-up period). Out of the 13 features in the dataset, 6 of them were categorical and 7 were continuous features. The categorical features - High blood pressure, smoking, diabetes, sex, anaemia (red blood cells deficiency) and death event (if the patient survived in the follow-up period) consists of binary values. The continuous features in the dataset are age, creatinine phosphokinase, ejection fraction (percentage of blood pumped at each contraction), (CPK enzyme level - tissue damage indicator), serum sodium (sodium content in blood) platelets, serum creatinine (kidney function indicator), and time (follow-up period in days). The dataset consisted of more men (194) than women (105) with their age group being between 40 and 95. The target feature in this scenario is the death event, a binary variable represented with 0 as patient's survival and 1 as patient's death. Amongst the 299 patients, 96 patients (32.11%) died and 203 patients (67.89%) survived.

Since the target variable is a categorical one with two classes (survived or not-survived), the problem here is a binary classification problem. After ensuring the data is in a usable form to perform further data science operations, 2 classification models, K-nearest neighbour model and decision tree classifier model were constructed. This is to represent the overall decision processes in an abstract manner which in this case involves predicting survival status of heart failure patients within a certain period of time (on average 130 days) for patients who have previously suffered heart failures.

Methodology

Overview

Unprocessed raw data can have significant effects on the results that will be drawn from the dataset in the following steps. Therefore is it important to pre-process the dataset to ensure higher- quality data is being used for the subsequent analysis steps and in performing machine learning processes. A range of graphs will be plotted to provide further visual context to heart failure clinical records. After this, two data models will be built and compared against each other.

Retrieving And Preparing Data

Data preparation is a vital step in data science applied in order to yield useful insights, it is crucial to prepare the data before using it for any analytics applications. A copy of the heart failure dataset was created to prevent

any impact on the original dataframe. The data type of each feature was verified to ensure consistency and usability of the dataset.

Checking for missing values

Missing data could be a result of certain data not being recorded or even caused by data entry error. However, when the dataset was checked for missing values using the .isnull() function, it was found that there were no missing values in the entire dataframe.

Checking for outliers

Outliers can heavily impact the model by skewing the overall results. A small amount (two) of outliers were observed in the ejection fraction column. Further research proved that ejection fractions of 70 and 80 are very uncommon in heart failure patients *(Nazario, 2022)* therefore these values were excluded from the dataset.

Checking for inconsistencies

Whitespaces and inconsistencies in a dataset are difficult to notice and can easily cause confusion and issues. Using the '.strip()' function it was ensured that there were no leading or trailing whitespaces in the columns. Moreover, the features were named in an understandable and consistent manner.

Checking for impossible values

Impossible values can contaminate the overall dataset and may lead to biased evaluations which may result in inaccurate conclusions, for example, age of 500 is not possible. All the values in the dataframe were within the appropriate range.

Checking for duplicate values

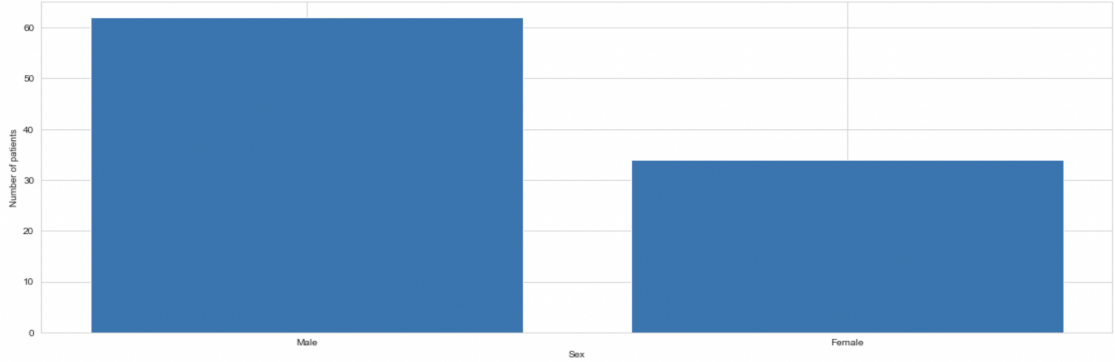
Duplicate observations can be detrimental and can lead to incorrect estimations and other data management issues. Therefore it is important to identify observations that are repeated at an earlier stage. The '.duplicated()' function was used to confirm that there were no values repeated in the dataset.

Data Exploration

Data exploration enables deeper understanding of the data, revealing various characteristics and potential errors in the dataset. It is an important process to improve project efficiency and understand the nature of the data. A range of graphs were plotted to provide further visual context to the heart failure dataset.

Hypothesis: Male patients have a higher risk of dying from a heart failure compared to female patients. Although male and female patients possess almost similar risk factors, the relative significance of these factors differ amongst male and female.

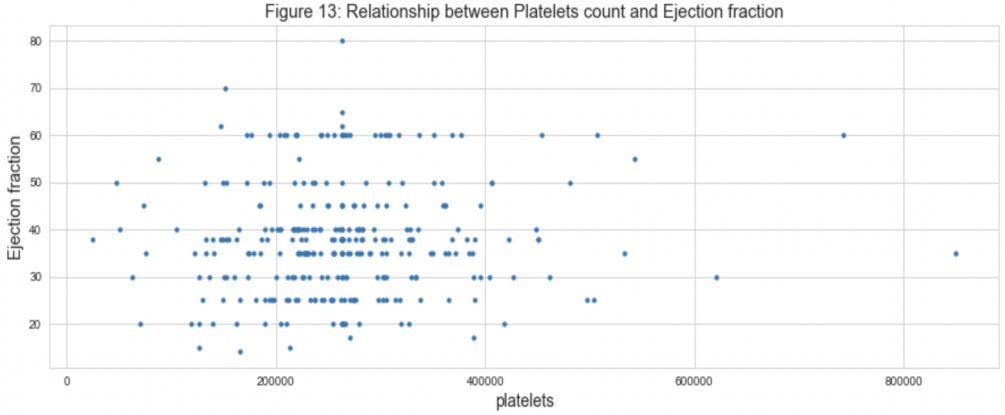
Figure 21: Comparision of death of male and female patients



The result from the graph asserts the stated hypothesis wherein it can be confirmed that a higher number of male patients (62 male patients) have died compared to female patients (34 female patients). It can be observed that the difference in death count in the two sexes is almost twice which suggests that the male patient population is more likely to pass away from a heart failure.

Hypothesise: When there are too many platelets produced in the patient's body it leads to a disorder known as thrombocytosis. It is hypothesised that this may lead to reduced ejection fractions in the patient's body.

Figure 13: Relationship between Platelets count and Ejection fraction



There does not seem to exist any clear correlation between platelet count and ejection fraction in patients' bodies. Without a slope, no trends can be observed, indicating no linear relationship between the features. The data points across the plot are scattered randomly without representing any relationship between the variables. However, it can be observed that the most number of patients (both male and female patients) have a platelet count between 150000 and 400000 and the ejection fraction between 20 and 60 which is consistent with the production levels in heart failure patients (*academic*, *2012*).

Hypothesise: The ejection fraction (amount of blood pumped out by left ventricle with each contraction) percentage tends to fall and the serum creatinine (kidney's performance level) levels tend to rise in heart failure patients.

Figure 12: Serum creatinine and Ejection fraction's relationship with patients' survival status.

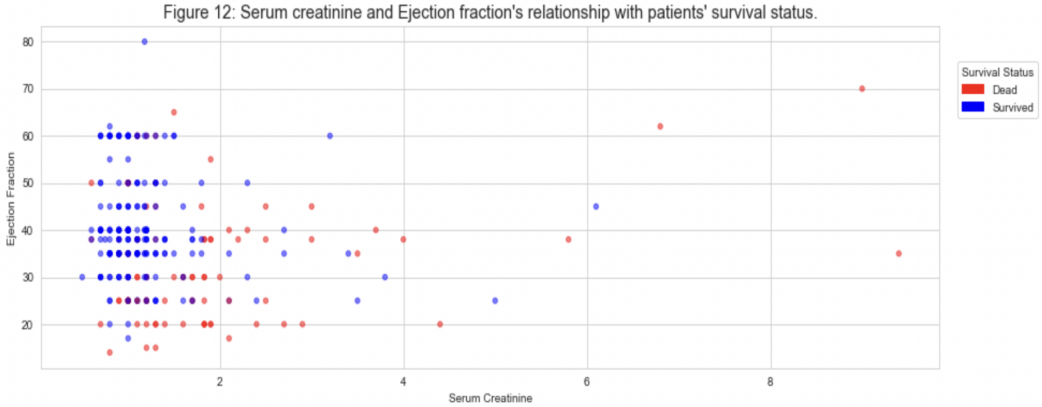


Figure 12 represents a scatter plot of patients' ejection fraction and serum creatinine, coded by patients' survival and death data points. It can be inferred from the graph that the surviving patients possessed lower serum creatinine and higher ejection fractions compared to the patients who did not survive (until the end of their time period) with relatively higher serum creatinine levels and lower ejection fraction. This could greatly aid in measuring the severity of a patient's heart failure condition.

Hypothesis: High creatinine phosphokinase is a result of stress on muscle tissues and heart. Heart failure patients with higher levels of creatinine phosphokinase are more likely to suffer low ejection fractions in their bodies.

Figure 20: Relationship between Ejection fraction and Creatinine phosphokinase

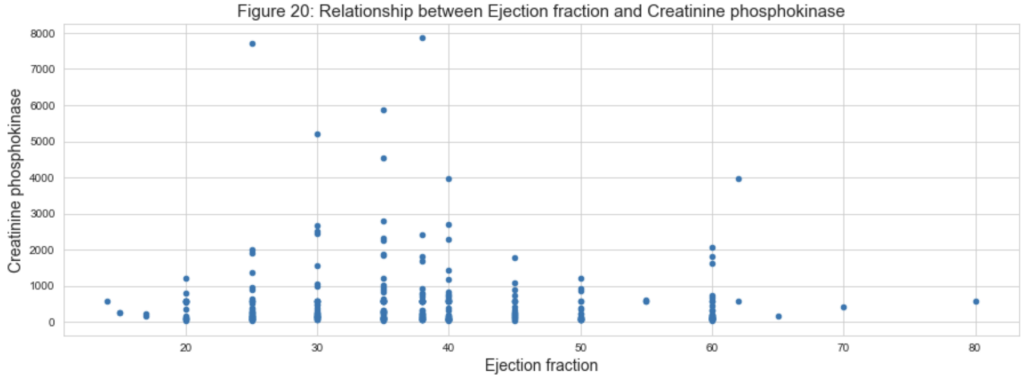
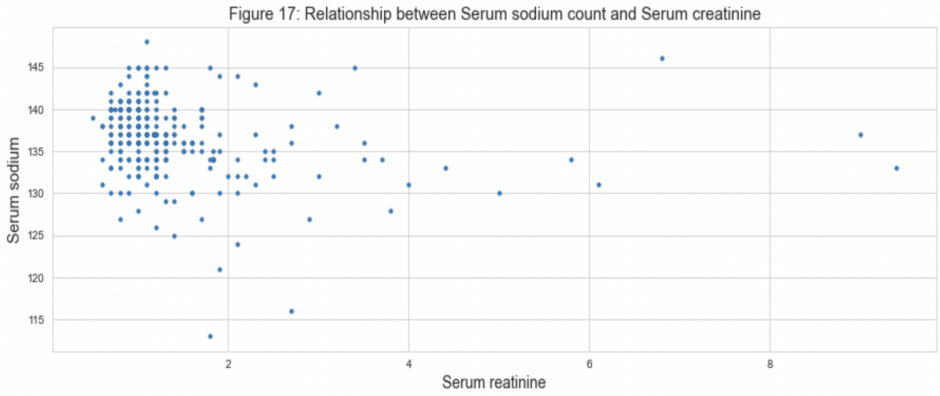


Figure 20 represents the relationship between the ejection fraction of percentage of blood that leaves the heart ventricle at each heartbeat and the creatinine phosphokinase level in the patient's body. It can be deduced from the scatterplot that there is no obvious correlation among the two variables, however, it can be observed that the data points are clustered around certain values due to the fact that the ejection fraction levels are expressed as percentage. Moreover, a strong clustering can be found in the range of 20 and 60 indicating nearly normal ejection levels among the patients.

Hypothesis: Higher sodium intake may result in a rise in serum sodium levels leading to increased serum creatinine in the body. Therefore, as the serum sodium level increases, the serum creatinine may also increase.

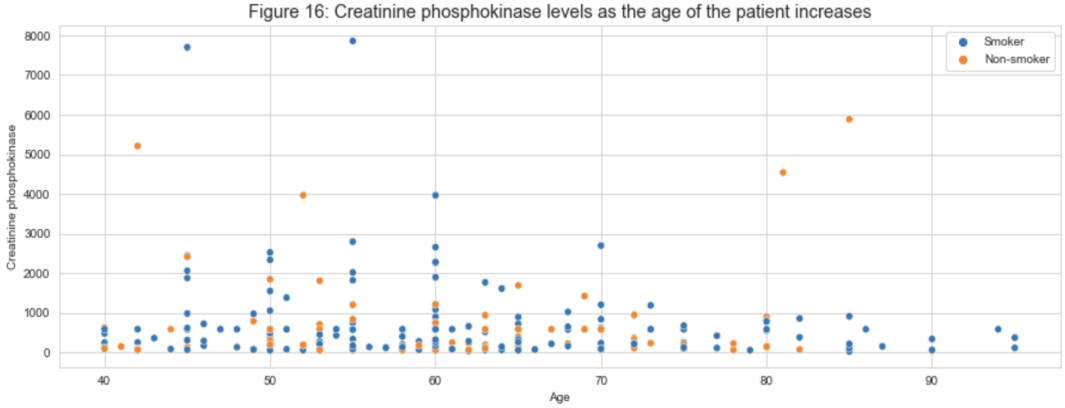
Figure 17: Relationship between Serum sodium count and Serum creatinine



In figure 17, a strong clustering can be observed in the scatterplot where the sodium creatinine levels are in the range of 130 and 145 and the serum creatinine levels between 0 and 2. A group of data points clearly follow the same general pattern indicating that the inputs from this range are likely to produce similar specific results. In addition, several other data points are scattered around the graph indicating variables are not correlated.

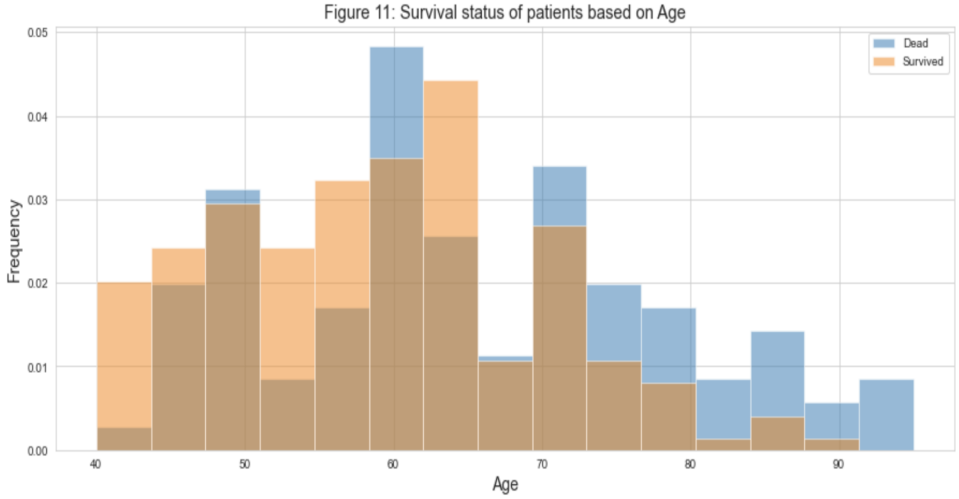
Hypothesis: Cigarette smoking is associated with high risk factors in heart failure diseases development of humans. The creatine phosphokinase levels in heart failure may rise with their age along with depending upon their smoking status.

Figure 16: Creatinine phosphokinase levels as the age of the patient increases



High creatinine phosphokinase is linked with people who suffer from heart diseases. However, it can be inferred from the graph that the creatinine phosphokinase levels in heart failure patients is independent of their cigarette smoking status. Moreover, creatinine phosphokinase does not increase or decrease in level other than a slight insignificant hike in the 60's age range meaning it is not controlled by the age factor of the patients, suggesting absence of any correlation between the features.

Figure 11: Survival status of patients based on Age



Hypothesis (figure 11): Older patients are more likely to die from a heart failure compared to the younger patients as the risk of death from heart failure increases with age (can be labelled as a personal risk factor).

The histogram in figure 11 represents the distribution of heart failure patient's age differentiated by their death status. Up until the age of 65 mostly the dominance of patient's surviving status can be observed wherein more number of patients survived during their follow-up period. There is a high rate of survival at this range, however distribution after this range (after the 65 age range), suggests that more people passed away than people who survived. This indicated that the older patients, after the age of around 70 are highly likely to pass away from a heart failure than patients who are at a younger age.

Data Modelling

In data science, data modelling is training of a machine learning algorithm to predict the labels based on the features. It involves training the model to find patterns and make decisions from a dataset that is previously unseen. After gaining better insights into the heart failure clinical dataset it was discovered that the time feature which indicates the follow up period of a patient after experiencing a heart failure is irrelevant in the modelling process since this information is only applicable to patients who survived until the end of follow-up period, not

for the patients who passed away. Excluding this will enable more accurate survival predictions for all the patients.

Two classification models, k-nearest neighbour and decision tree classifier were constructed, acknowledging the fact that the problem being solved is a binary classification problem. The models developed were then compared against each other in terms of their performances.

Data split proportion and justification

Since the dataset is small in size with 299 observations, the train-test split procedure was utilised to divide the dataset into 2 subsets, 80% to fit the model and remaining 20% for testing. The two-dimension size of the heart failure dataset was reduced to one dimension, converting it to a numpy array. A random state was also selected for the data to ensure consistent random samples are used. Each observation in the dataset is uniquely represented where the data is independently distributed without any records of the patient's duplicate visits.

K-nearest neighbour classification

The K- nearest neighbour algorithm seeks to classify data points according to the similarity measures of the nearest neighbours. All the required modules related to the K-Neighbours classifier were imported for the model construction.

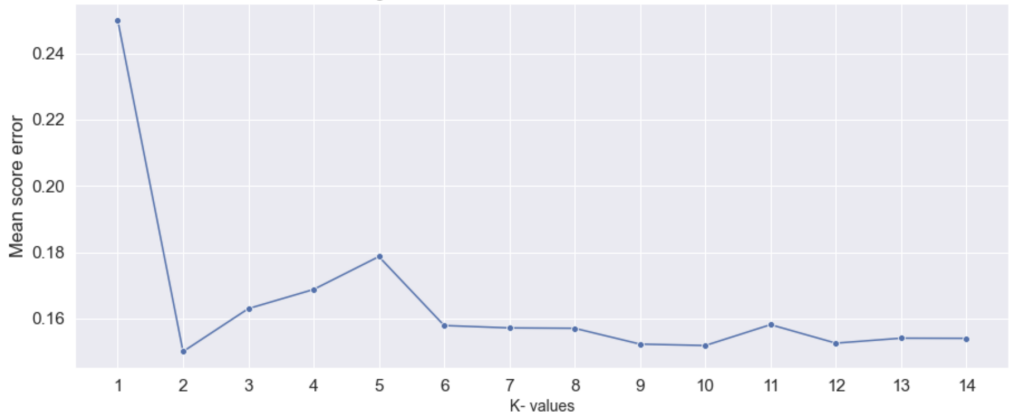
Feature selection

K- nearest neighbour can greatly benefit from feature selection wherein the algorithm accuracy can be improved by selecting the most significant and influential features. Given the small size of the dataset, the feature selection method selected for this model was a hill climbing technique process which can be used to generate the most optimal results (employs a heuristic approach for optimising the problem). The hill climbing technique was applied to reduce the number of less significant variables to improve the overall performance of the predictive model. This was achieved by selecting a reasonable K- value (k-value of 5) to start the computation. Various scores with different features were obtained which were useful in further analysis.

Hyperparameter tuning

The k-nearest neighbours is a non-parametric algorithm. In implementation of the k- nearest neighbour classifier, there are 2 parameters required, the value of k and the distance function. After filtering the dataset features, an appropriate k-value was evaluated by calculating mean squared error for different values of k ranging from 1 to 15 to further build the model. The k- value of 2 with the lowest mean square error 0.15 was selected for parameter tuning. Moreover, the default value of leaf size 30 and p value parameter of 2 (Euclidean distance, computed with the formula $(\sum_{i=1}^{n} |x_i - y_i|^p)^{1/p}$) were maintained since the results generated from these values indicated a good model performance. In addition, a classification report along with a confusion matrix was generated for performance assessment of the model.

Figure 22: Mean score error for different k values



Model validation

In order to verify that the model is performing as expected to achieve its intended purpose, K- fold cross validation was applied using a value of k as 5 to divide the dataset to 5 subsets and its performance was evaluated. The scores obtained from these folds were averaged for clearer representations. These have been further discussed in the results section of the report.

Decision trees classification

Decision tree classifier seeks to recursively distinguish classes by partitioning the training set. All the required modules related to the decision tree classifier were imported. A decision tree was then constructed and fitted with the '.fit()' function. With the training data as input to perform estimates corresponding to properties on the unseen data. A classification report along with a confusion matrix was again generated to further assess the model.

Hyperparameter tuning

An experimental approach was again utilised to come up with the most suitable parameters to optimise the model architecture. A range of max depth values from 1 to 20 were looped and the accuracy percentage of the model with each of the max depth values was evaluated. The max depth value of 5 with the highest accuracy score of 82% was selected. Furthermore, by tuning the criterion parameter, changing it from the default 'gini' index to 'entropy' a model with slightly improved accuracy of 83% was achieved. It was also perceived that with the help of hyperparameter tuning a decision tree classifier model can be constructed based on solely two features, ejection fraction and serum creatinine (however, this was not further explored as it does not help in addressing the research goal of this report). Further, rather than allowing the decision tree to expand until the last leaf nodes, setting a limit of 5 maximum leaf nodes led to a condensed tree structure.

Model validation

The K- fold cross validation was applied to verify that the model is performing to achieve its intended purpose using a k value of 5. The scores obtained from these folds were averaged. This further helps in generalising the model, which results in better predictions on unknown data.

Comparing classification models

After constructing the two classification models, they were compared against each other in regards to various performance metrics to determine which of the models is suitable in predicting the heart failure death event. The results are further discussed in the results section.

Results

Applying the methodologies provided a wide range of insights in the development process which helped in determining the best classification model. A variety of tests with different changing variables in each scenario led to reaching more informed conclusions.

Training the K-nearest neighbour classification model with training data and testing it produced the following results shown in table 1

.

Table 1

Classification report for K-Nearest Neighbour

	precision	recall	f1-score	support
0	0.77	0.95	0.85	42
1	0.75	0.33	0.46	18
accuracy			0.77	60
macro avg	0.76	0.64	0.66	60
weighted avg	0.76	0.77	0.73	60

Table 2

Classification report for decision tree

	precision	recall	f1-score	support
0	0.85	0.93	0.89	43
1	0.77	0.59	0.67	17
accuracy			0.83	60
macro avg	0.81	0.76	0.78	60
weighted avg	0.83	0.83	0.83	60

After performing all the required processes in developing the model, it can be inferred from table1 that the K-nearest model computed the overall classification accuracy of 77% with 95% accuracy for surviving patients and 33% for the non-survival. The f1-score, harmonic mean of recall and precision produced 0.85 and 0.46 for surviving and non-surviving patients respectively. In addition, there were 42 true responses that supported the model prediction for surviving and 18 responses for the non-survival patients.

Training the Decision tree classification model with training data and testing it produced the results shown in the above table 2. In the decision tree classifier after performing all the required processes in developing the model, it can be inferred from the above table 2 that this model computed an overall classification accuracy of 83% with 93% accuracy for surviving patients and 59% for the non-survival. The f1 score, harmonic mean of recall and precision, produced a score of 0.89 and 0.67 for surviving and non-surviving patients respectively. In addition, there were 43 true responses that supported the model prediction for surviving and 17 responses for the non-survival patients.

The confusion matrices provide further insights into measuring recall, precision and predicted combinations. Figure 23 represents the confusion matrix of the K-nearest neighbour model wherein there are 40 values classified as true negative and 6 as true positives out of the 60 observations. In figure 24, the decision tree classifier model consisted of 40 true negatives and 10 true positives for surviving and non-surviving patients respectively.

Figure 23: Confusion matrix heat map for KNN

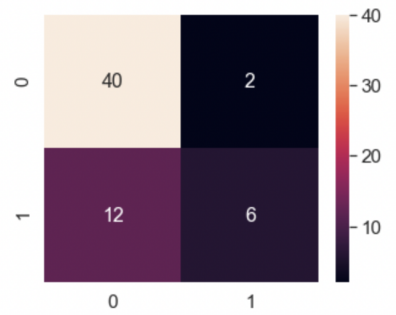


Figure 24: Confusion matrix of Decision tree classifier

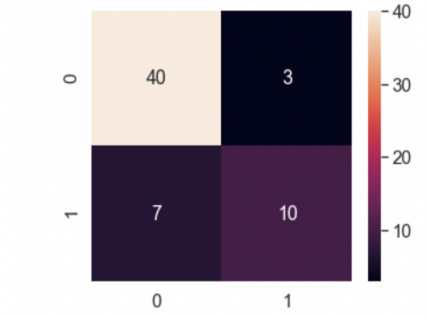
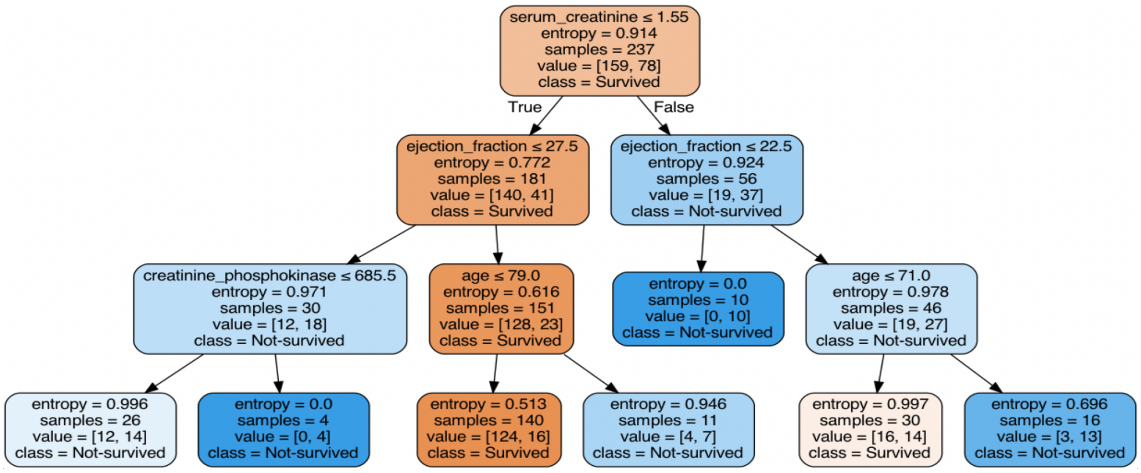
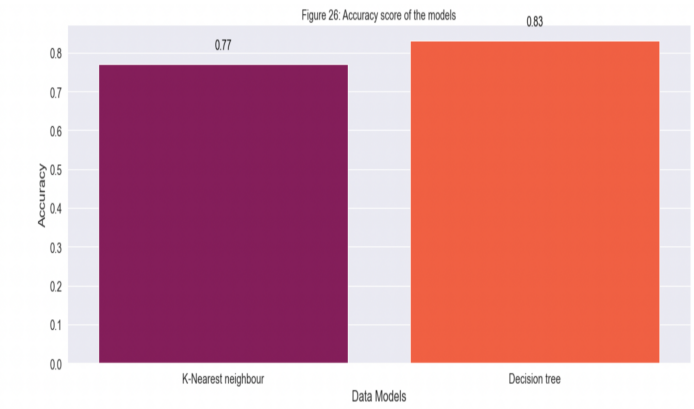
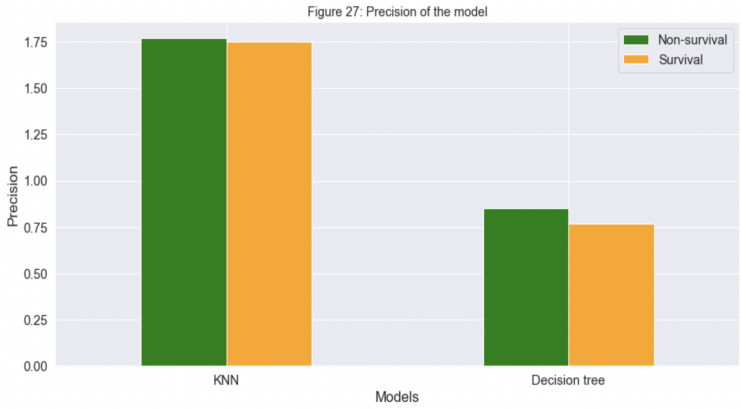


Figure 25. Decision tree classification



The above decision tree structure was plotted using the Graphviz GUI to derive clearer visual observations. The flow chart diagram demonstrates various outcomes from a binary decision survived and not-survived. The impurity of each node is measured by the entropy, a high entropy can be observed starting from the root node and as the tree follows down as nodes are splitted into sub-nodes. The 3 levels of diagram further reduces the complexity of the tree structure.





The average scores of 0.67 for K- nearest neighbour and 0.74 for decision tree were obtained from the K-fold cross validation method after which graphs representing the accuracy and precision of the models constructed were plotted to aid in model comparison in order to draw appropriate conclusions. Figure 26 and 27 represent the comparison of the models in terms of accuracy and precision.

Discussion

The predictive models decision tree classifier and K-Nearest neighbour classifier selected for the data set aided in identifying the survivability of the heart failure patients based on previous health conditions, their lifestyle and blood sample readings. Feature selection and tuning the hyperparameter values of the k- nearest neighbour and decision tree accordingly led to discerning various new facts and details. Performing evaluations of the predictive models by comparing them in regards to the ground truth with the help of accuracy scores, confusion matrices and other performance reports enabled drawing more appropriate conclusions.

K- nearest neighbour classification model overview

Since feature selection in K- nearest neighbour model facilitates the modelling, the hill climb feature selection method was utilised to individuate the most important features among all the included features in the patient's

medical records, given the small size of the dataset. The resulting features were Serum Creatinine, Ejection fraction, Age, Anaemia, Sex, Diabetes, Blood pressure. It was also astonishing to discover that the smoking status of the patient was not included in the feature selection process but again, the dataset contains data of patients who have already had a heart failure experience, therefore the smoking status would tend to be a smaller factor amongst other factors. In addition, to obtain the best results from the algorithm, the most appropriate distance metric 'euclidean' distance was selected accordingly wherein the neighbours are evaluated with this metric function and for each new data point, only a fixed number of neighbours in this case k = 2 (as this had the least mean square error and in this case due to the small size of the dataset a K value of 2 can be considered appropriate) are taken into account.

Decision tree classification model overview

Another widely known model, decision tree classifier was employed and its survival predictability was evaluated. Since the model learns by recursively dividing the sample space according to the most significant features (greedy approach) until the tree reaches a constrained depth, the features that a feature selection model like hill climb would have eliminated are automatically not utilised by the decision tree in the decision making process. The measure in this case that provided the most appropriate results is the entropy in which the value intervals are between 0 and 1. However, a high entropy was observed when the tree was visualised (in figure 26) suggesting lack of homogeneity. Hyperparameter tuning enabled further control of the model behaviour and the complexity of the tree was also reduced by setting the appropriate parameter values.

Models Comparison

The K- Nearest neighbour and decision tree classifier models are both non-parametric supervised learning models having their own advantages and disadvantages. With the primary objective of analysing better overall performance, several other metrics would also influence the model selection phase. Examining the statistics from the reports generated during the model analysis assisted in this comparison. In terms of the prediction accuracy both models provided a considerably close score with the K- nearest neighbour predicting 77% of the data accurately and and the decision tree classifier with 83% accuracy. The precision metric used to identify the percentage of classification correctness was found to be high for both positive and negative classes in both models, however, the classifications in the decision tree model slightly outperformed the K-nearest neighbour classifier results for predicting the patients death status as survived. The K-nearest neighbour achieved a precision of 85% and 77% for predicting the survivability as survived and not survived respectively, whereas the K-nearest neighbour had a precision of 77% for survived patients and 75% for the non-surviving patients suggesting that the decision tree has a slightly better ability to classify the surviving cases. In this particular instance where a patient's whole survivability, a vital prognosis to consider is being predicted, one would need to be considerate in revealing the outcome to the patient's survival. The recall performance metric in which the correct surviving positive predictions are quantified out of all positive predictions that could have occured, will greatly assist in addressing this prognonsi problem. In this regard, the K- nearest neighbour classifier had a higher recall score of 95% for predicting survivability as survived and a lower score of 33% for non-survival. Moreover, the f1-score of the decision tree model (89% for survival and 67% for non-survival) was slightly higher compared to the K-nearest neighbour model (85% and 46%).

As the decision tree grows in size, it loses its generalisation capability and gets more susceptible to model overfitting (the higher value of maximum depth causes overfitting, and a lower value causes underfitting), negatively impacting the performance on the new data. The training data would then be fitted in tightly such a way that it results in inaccurate predictions of survival outcomes on the untrained data. Switching to K- Nearest neighbour reduces this risk of overfitting, however, it adds in an obstacle of selecting the best K value for modelling. Moreover, the K- Nearest neighbour is very sensitive to outliers. As it is an instance-based algorithm

that uses distance as a criteria, any outliers in the dataset will lead to biassed outcomes and affect the class boundaries.

Along with the predictive power of the models, the explanabilty is a major factor in assessing the models. It is possible to visualise the process of the decision tree algorithm where a decision is taken at each node. This will enable easier understanding and provide a more visualised context of the decision making process where the outputs can be easily interpreted without any complex knowledge of statistics. The K- nearest neighbour on the other hand has no such attributes other than visualising the decision boundary changes as the parameters are tuned.

After considering all the benefits and drawbacks of both the models, It can be concluded that the decision tree is a more suitable machine learning model for predicting the survivability of the heart failure patients. In addition to having a higher accuracy and precision in classifying the data, decision tree trends to perform better on dataset with smaller number of features such as in this heart failure prediction case. Although the decision tree model is prone to overfitting, the reduced number of branches will significantly decrease these chances. Another major advantage of employing the decision tree is that it does not necessarily require feature selection to be applied unlike the K-nearest neighbour model where hill climb approach was utilised. All these characteristics of the decision tree model will provide a powerful approach to address the goal of predicting the survivability of heart failure patients. Therefore, the decision tree classifier can be considered as a more appropriate model to help in addressing the research goal.

Conclusion

In the world of fast paced medical advancements, modern technology and healthcare with hands clasped have progressed throughout the years, driving various positive changes in the healthcare industry like never before. Therefore, it is rational to desire utilisation of machine learning methods to seek technological progression for developing heedful medical innovations and get more prepared for the upcoming occurrences. In this research, the heart failure patients' condition was better understood through various means to answer the research question of predicting patients' survival status based on their medical history records; in order to make more informed decisions in the future, considering it is the leading cause of death throughout the world. This led to inferring conclusions that recognising the key factors and relationships to reliably predict the patient's survival and effective use of machine learning models (in this case the decision tree with better overall performance) for classification of heart failure patients' medical data will ultimately extend the applications of data science in the cardiology branch and facilitate health professionals in ameliorating the condition of patients with heart failure diseases.

References

Who.int. 2022. Cardiovascular diseases. [online] Available at: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1 [Accessed 27 April 2022].

academic.oup.com. (n.d.). Validate User. [online] Available at: https://academic.oup.com/eurheartj/article/40/Supplement_1/ehz747.0353/5594427 [Accessed 3 May 2022].

Nazario, B. (2022). What Does Ejection Fraction Have to Do With Heart Failure? [online] WebMD. Available at: https://www.webmd.com/heart-disease/heart-failure/features/ejection-fraction#:~:text=Normal%20EF%20is%20in%20the [Accessed 6 May 2022].