Title: Assignment 2 - Data Modelling and Presentation
Student ID: S3774430
Student Name and email (contact info): Myeonghoon Sun (S3774430@student.rmit.edu.au)
Affiliations: RMIT University.
Date of Report: 09/05/2022

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

1. Abstract

The report aims to build machine learning models to predict whether heart failure (HF) patients will survive in a bid to effectively allocate human and medical resources to those who most need them. Given the need to label every data point including outliers, classification algorithms are used to fit the training data. After fine-tuning the parameters, a Leave-One-Out-Cross-Validation approach is taken to evaluate the models as the dataset only contains few observations. The k-Nearest-Neighbour model turns out to have the accuracy of 0.73 whereas the decision tree model 0.72. Even though the latter's accuracy is slightly lower than the former, the latter's true positive rate (TPR) is noticeably higher, cementing it as the recommended model with a caveat that neither is industry applicable since both models' true positive rates are too low to be reliable. It is recommended that a model with a higher TPR be built to achieve the project goal of correctly identifying dying HF patients so that they receive maximum medical attention they need to survive.

2. Introduction

Heart failure by nature descends on the unsuspecting. Once heart failure patients come under the wing of medical staff, it is now their responsibility to do all it takes to keep them alive. However, only a fixed amount of time can be allocated for each patient if the medical staff themselves are unsuspecting of who requires the most care, just around the corner of death. With the help of machine learning, this veil could be lifted off, and the knowledge of who stands the lowest chances of survival can allow the medical staff to effectively allocate their resources to the most vulnerable.

The goal of this project is to predict survival of HF patients such that human and medical resources can be best utilised according to the predictions. Therefore, the performance of a model will be measured by the accuracy metric coupled with the true positive rate. With the latter alone, many erroneous predictions to classify the surviving patients as dying will be disregarded, leading to misallocation of the resources.

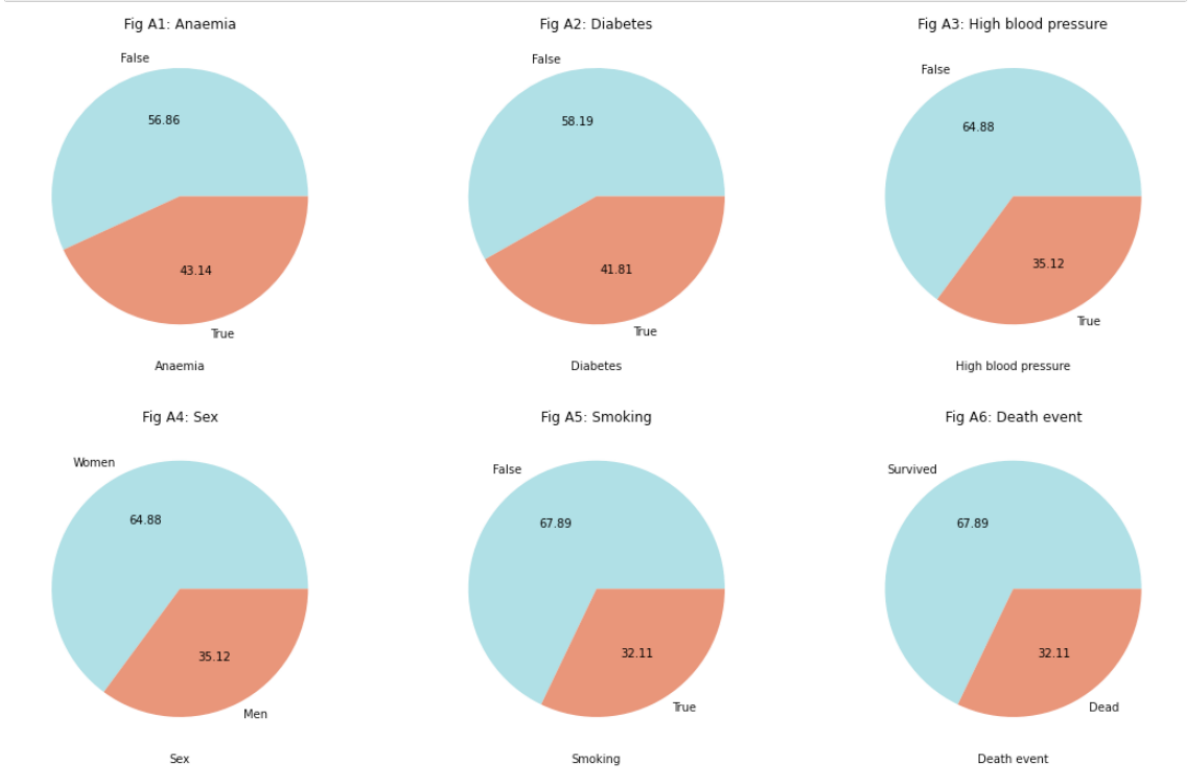
Predictive models will be built on select meaningful features with the help of such classification algorithms as the k-Nearest Neighbour and Decision Tree. Since one of the prerequisites for building an accurate model is to select meaningful features, each of the provided features will be examined below with respect to its association with cardiovascular diseases.

2.1. Dataset

With the target feature called DEATH_EVENT, there are a total of 12 explanatory features that can be divided into two categories: 5 binary and 7 numerical features.

2.1.1 Binary Features

The binary explanatory features are as follows: anaemia, diabetes, hypertension, sex, and smoking. In addition, the target feature whose values are to be predicted is named DEATH_EVENT. The pie charts below show the percent distribution of the 299 HF patients by each of the binary explanatory features as well as the target feature.



First, anaemia is a frequent comorbidity of heart failure and is associated with poor outcomes from heart failure (Shah, 2022).

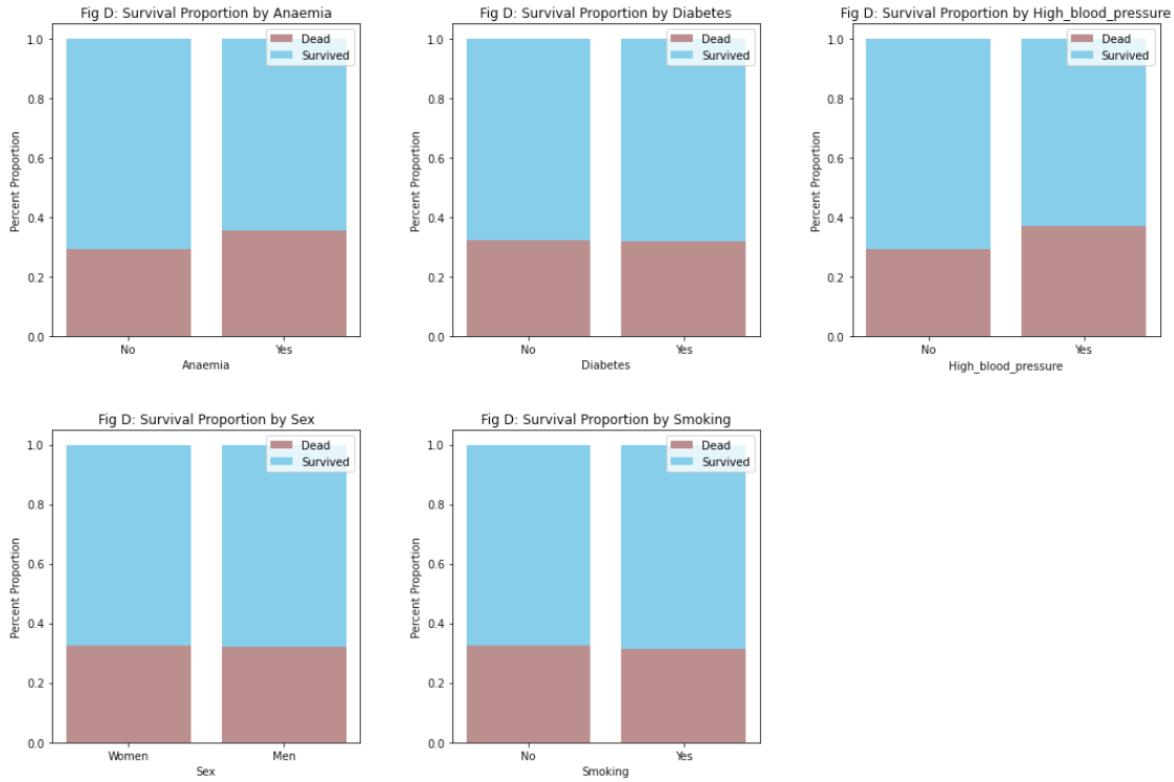
Not only people with diabetes are more likely to have heart failure, but also its related symptoms include the building up of fluid in the lungs, which makes it hard to breathe (Diabetes and your heart, 2022), which could then potentially lower the probability of survival for HF patients under emergency care.

If the blood pressure is high, the heart has to work harder to circulate blood throughout the body. Over time, this extra exertion can make the heart muscle too stiff or too weak to properly pump blood (Mayo Clinic, 2021).

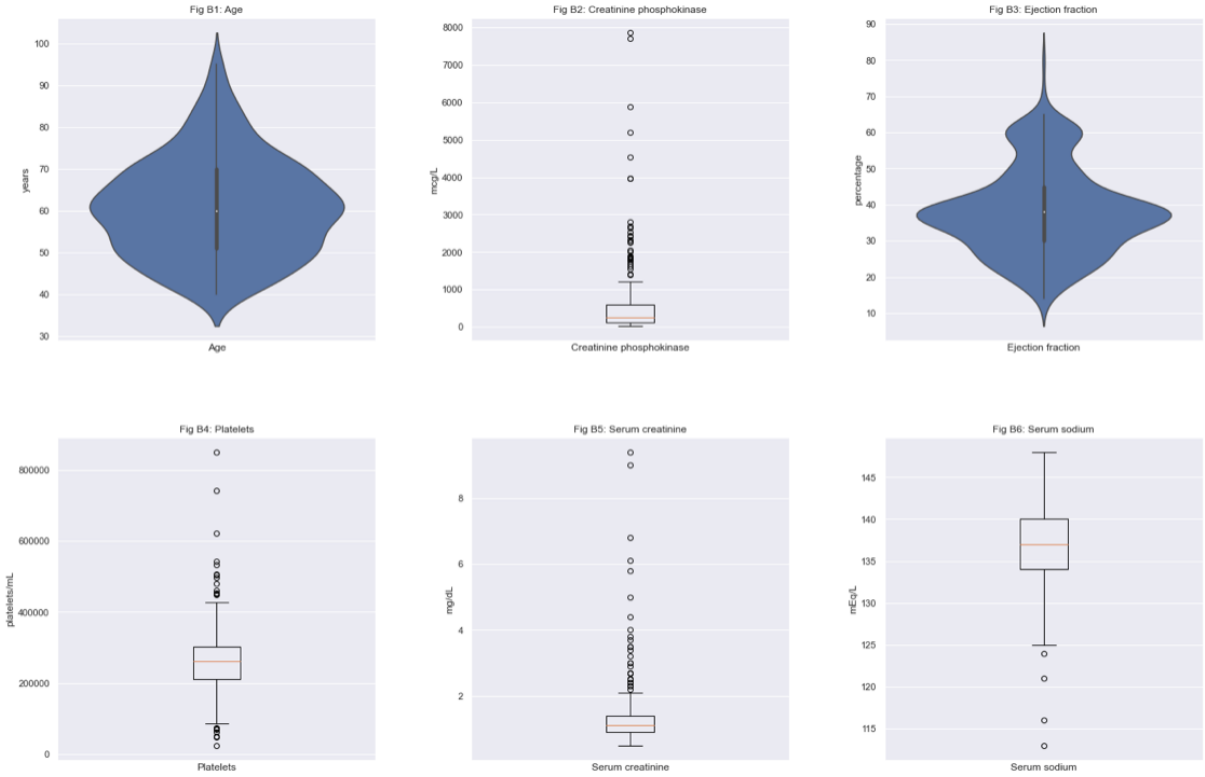
Even though researchers have found women face a 20% increased risk of heart failure or dying within five years after their first severe heart attack compared with men, women are also more likely than men to be older and have a more complicated medical history by the time (American Heart Association, n.d.), which leaves the question of how much of a role gender plays in the likelihood of survival still open-ended.

As with diabetes and high blood pressure, smoking is one of the main causes of heart diseases. Smokers are almost three times more likely to die of heart and blood vessel disease (Vic.gov.au, 2012).

Each of the stacked bar graphs below illustrates the percent proportions of survival by the explanatory features. As far as the dataset is concerned, hypertension and anaemia have slightly higher correlations with death.



2.1.2 Numerical Features



Each of the numerical features is visualised using either violin plots or boxplots; the former is used in the case where there are not as many outliers but the details of how its values are

distributed provide more insight, whereas the latter is used to intuit the number of outliers at first glance as well as other descriptive statistics.

As seen in Fig B1, the median age is around 60 years old. The plot itself confirms the common sense that anyone regardless of age can have a heart issue and the chances of getting a heart failure increase after the age of 45 for men and 50 for women (Heart Disease & Age | Heart and Vascular, 2022).

Creatine phosphokinase, also known as CPK or CK, is an enzyme found in human body including the heart. When the muscle tissue from such a region as the heart is damaged, CPK leaks into the blood. Therefore, high levels of CPK could indicate injury to the heart (Creatine Phosphokinase (CPK) : Johns Hopkins Lupus Centre, 2022). The CPK normal values range from 10 to 120 micrograms per litre (mcg/L) (Creatine phosphokinase test Information | Mount Sinai - New York, 2022). The case report written by Jad and et al. shows that the CPK level can reach up to 12,000 mcg/L (Jad, 2022), dispelling a suspicion that a great many outliers seen in Fig B2 are impossible values.

Ejection fraction (EF) is a measurement, in percentage, of how much blood the left ventricle pumps out with each contraction. A normal heart's ejection fraction may be between 50 and 70 percent (Ejection Fraction Heart Failure Measurement, 2022) unlike what can be observed in HF patients from Fig B3.

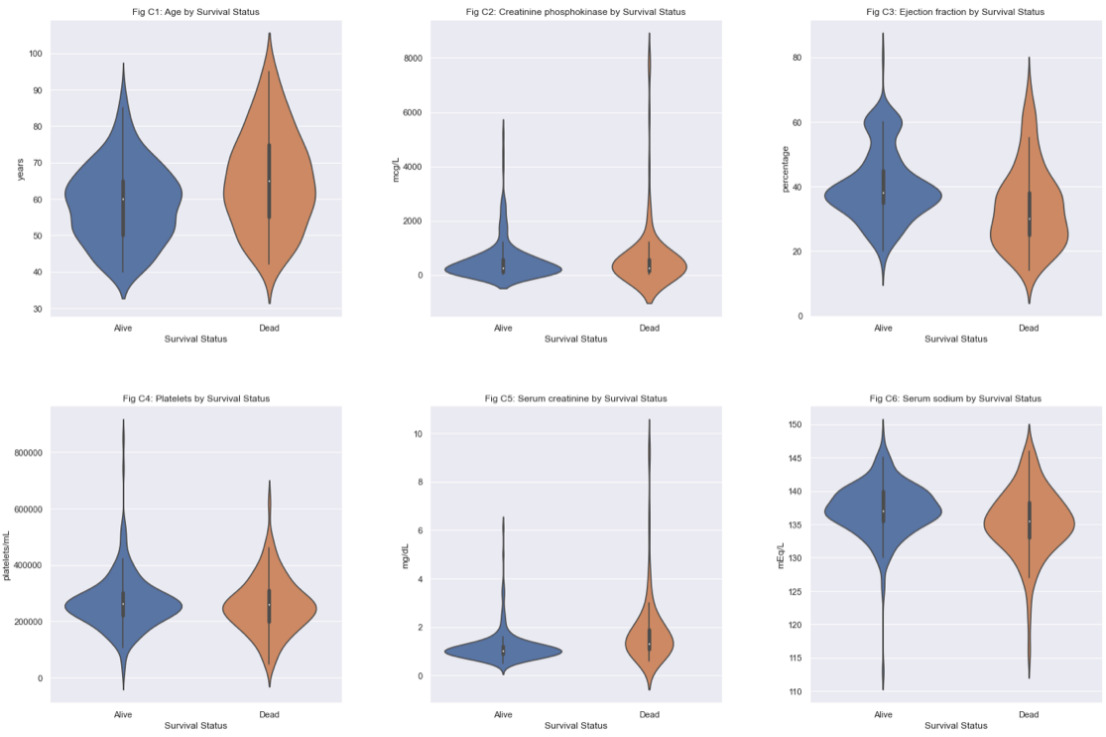
Platelets are specialized cells in the blood that are involved in the formation of blood clots. Blood vessels injured by smoking, high blood pressure, et cetera accumulate plaques, which platelets mistake for an injury and concentrate around them to form a clot; they then cause a detrimental clot that leads to a heart attack (Platelets and Cardiovascular Disease, 2022). The normal platelet count ranges from 150,000 to 450,000 platelets per microliter of blood. Having more than 450,000 platelets is a condition called thrombocytosis. One of its types may be caused by an ongoing condition such as anaemia (What Are Platelets and Why Are They Important?, 2022). A count of less than 50,000 platelets is a seriously low count, but it is still a possible value (Platelet count: definition, low vs normal vs high ranges, 2022).

Creatinine is a chemical compound left over from energy-generating processes in your muscles that healthy kidneys filter out of the blood. An increased level of creatinine may be a sign of poor kidney function (Creatinine tests - Mayo Clinic, 2022). Renal dysfunction is common in HF patients. Cardiac and renal dysfunction may worsen each other through multiple mechanisms (The role of the kidney in heart failure, 2022). The typical range for serum creatinine is 0.74 to 1.35 mg/dL for adult men and 0.59 to 1.04 mg/dL for adult women (Creatinine tests - Mayo Clinic, 2022). Given that creatinine levels of 5.0 or more in adults may indicate severe kidney damage (High, Low, & Normal Creatinine Levels: What This Blood Test Means, 2022), the outliers seen in Fig A5 can be considered reasonable values.

Sodium makes the body hold onto fluid, which means the heart must work harder to pump around the extra fluid in the body. In heart failure, consuming too much salt can worsen associated symptoms (Reducing salt intake with heart failure, 2022). However, hyponatraemia, a condition of having a serum sodium concentration less than 135 mEq/L, is just as detrimental to the health of a HF patient. Therefore, patients with normal sodium levels have a higher chance of survival (The prognosis of heart failure patients: Does sodium level play a significant role?, 2022).

As for the last feature, the follow-up period (in days), it will be dropped because of the great potential for data leakage in determining the survival status of each patient with its presence. The patients in worse conditions subject to death regardless of other attributes are going pass away relatively sooner and a shorter follow-up period will be recorded as a result. This assumption is supported by the high proportion of deaths for patients with a follow-up period of 60 days or fewer: 31 deaths and 4 survivals for 30 days or fewer; 52 deaths and 8 survivals for 60 days or fewer.

Below are the violin plots for the numerical features above grouped by the target feature. As the paper "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone" by Davide Chicco and Giuseppe Jurman indicates, the plots for each of these columns take on distinct shapes. Although CPK 's violin plots in Fig C2 do look more different from each other than the remaining three aside from the two mentioned above, the degree to which they are different is not as pronounced as that seen in Fig C5. These findings will be considered again in the modelling process.



3. Methodology

3.1 Setup

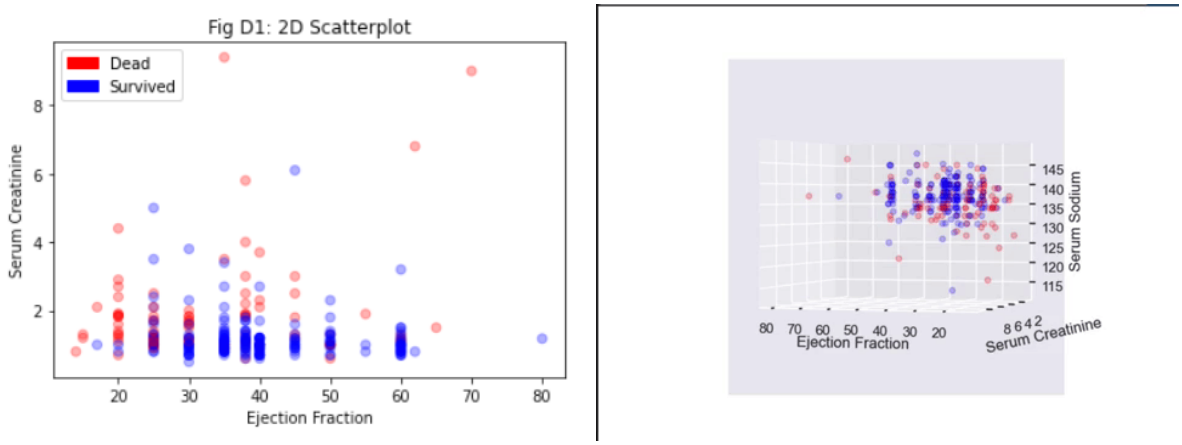
To build models capable of highly accurate predictions, it is important to, first, find out the performance of an untuned model with default settings. In doing so, the baseline for the accuracy of a general model can be determined and it can be then used to discover what parameters need to be adjusted to help improve the accuracy of an ameliorated model. For instance, the first step to the modelling process was to build a makeshift model using the k-Nearest Neighbours algorithm with the default values as parameters. The accuracy of the model turned out to be suboptimal regardless of what values are passed as arguments to the

parameters. It is then concluded that the poor accuracy is most likely due to the inclusion of all the features rather than a select few.

3.2 Feature Selection

The hill climbing algorithm for feature selection has the major downside of indiscriminately producing local and global maximum solutions. To accommodate to this shortcoming, the algorithm was executed 50,000 times to extract the three most frequently selected features by a landslide: ejection_franction and serum_creatinine, 21,730 and 19,999 times respectively with a count of 14,935 for serum_sodium immediately after them; the three features were initially chosen to build a comprehensible model.

As seen below, given the distribution of data points in the two and three dimensions (blue indicating survival and red death), fitting a model using classifiers is deemed more appropriate than using clusters; even if DBSCAN is used, outliers are going to be left without labels and the well-performing model, if ever built after countless mindless parameter tuning without the domain expertise, runs the risk of overfitting the available data, unable to be effectively used when the model's performance is rendered abysmal in the face of unseen data.



3.3 Modelling with K-Nearest-Neighbours

3.3.1 Parameter Tuning

As depicted in Fig D1, data points are clustered on the x-axis above the 17 discrete values from 15 to 80. Given this characteristic, the value for the number of neighbours for KNN is set to 17 (299 [the total number of observations] / 17 [the number of unique values found in ejection_fraction]). It can be also seen in the figure that neighbours of the same category are mostly close to each other, specifically the points between 30 and 40 on the x-axis between 0 and 2 on the y-axis. If any of these data points were a query point, it would make sense the data points on the same vertical line should have a greater influence in determining what category it belongs to. Therefore, it is for this reason that weights parameter is set to "distance". Because the features are engineered from the hill-climbing algorithm, each of them is qualified to be given an equal weight in the distance calculation and, hence the power parameter for the Minkowski metric is set to 2.

3.3.2 Leave-One-Out Cross-Validation

Since the dataset only contains 299 observations, the validation process warrants the Leave-One-Out-Cross-Validation (LOOCV) approach, making complete use of the given data.

3.4 Modelling with Decision Tree

3.4.1 Parameter Tuning

According to an empirical study on hyperparameter tuning of decision trees (DT), the ideal min_samples_split value is within the range of 2 to 40 while the ideal min_samples_leaf_value is between 1 to 20 (Mithrakumar, 2019). Given this information, all the combinations of min_samples_split and min_samples_leaf values were tried to find the most ideal combination; any value passed to the min_samples_split parameter and a value around 20 passed to the min_samples_split reliably produced high accuracy scores. When the model is evaluated, the max_depth is set to 4 to allow the tree to grow deep enough to capture major relations between features so as not to overfit the data and simple enough to understand its structure.

3.4.2 Leave-One-Out Cross-Validation

As with the previous model, the LOOCV approach was taken to evaluate the performance of the model in terms of accuracy [(True Positive + True Negative) / the number of target values].

4. Result

4.1 KNN

The Table A1 presents the counts of correct predictions in the diagonal cells from left to right and its accuracy 0.73. The 3-dimensional figure below identifies wrong and accurate predictions.

Table A1: k-Nearest Neighbour Model

Accuracy: 0.73

	Predicted: NO	Predicted: YES
Actual: NO	176.0	27.0
Actual: YES	55.0	41.0

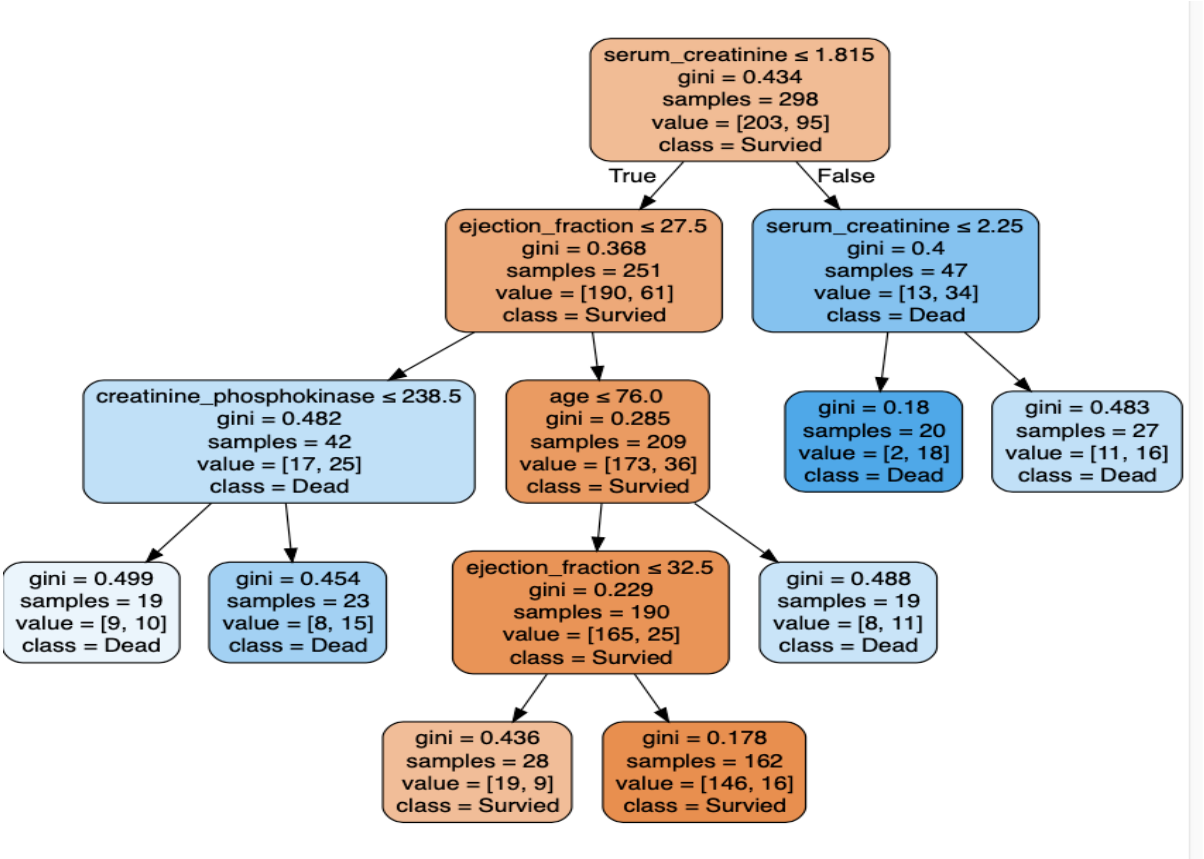
4.2 DT

The Table A2 presents the counts of correct predictions in the diagonal cells from left to right and its accuracy 0.72. The diagram of a binary tree below visualises the last classifier used in the LOOCV process.

Table A2: Decision Tree Model

Accuracy: 0.72

	Predicted: NO	Predicted: YES
Actual: NO	168.0	35.0
Actual: YES	48.0	48.0



5. Discussion

Both models have correctly predicted whether HF patients will survive a little better than 70 percent of the time. The DT algorithm was able to identify the most significant features, serum_creatinine and ejection_fraction (Chicco and Jurman, 2020), without the use of an additional algorithm for feature engineering as executed to build the well-performing KNN model. The 3-dimensional figure for the KNN model above clearly shows the enigmatically mislabelled data points, bringing light to the imperfect nature of a model in the absence of the domain knowledge. Even though both models have almost identical accuracy rates, the decision tree model is more effective in addressing the goal of the project; its true positive rate is greater, meaning that more dying patients will be correctly identified and attended to at the cost of misidentifying slightly more surviving patients and thus misallocating the resources to them in comparison to the alternative outcome of the KNN model.

6. Conclusion

Considering the goal of the project, the resultant models are far from being reliable, despite the accuracy of 0.70, as half the dying HF patients will be lumped into the same category as the correctly identified surviving patients. To make the models applicable on the ground, their true positive rates need to be improved.

7. References

- Shah, R., 2022. *Anemia associated with chronic heart failure: current concepts*. [online] National Library of Medicine. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3569049/> [Accessed 10 May 2022].
- Centres for Disease Control and Prevention. 2022. *Diabetes and your heart*. [online] Available at: <https://www.cdc.gov/diabetes/library/features/diabetes-and-heart.html> [Accessed 10 May 2022].

- Mayo Clinic (2021). *Heart Failure - Symptoms and Causes*. [online] Mayo Clinic. Available at: https://www.mayoclinic.org/diseases-conditions/heart-failure/symptoms-causes/syc-20373142.
- American Heart Association. (n.d.). *Women found to be at higher risk for heart failure and heart attack death than men*. [online] Available at: https://newsroom.heart.org/news/women-found-to-be-at-higher-risk-for-heart-failure-and-heart-attack-death-than-men.
- Vic.gov.au. (2012). *Smoking and heart disease*. [online] Available at: https://www.betterhealth.vic.gov.au/health/HealthyLiving/smoking-and-heart-disease.
- memorialhermann. 2022. *Heart Disease & Age | Heart and Vascular*. [online] Available at: <https://memorialhermann.org/services/specialties/heart-and-vascular/healthy-living/education/heart-disease-and-age> [Accessed 10 May 2022].
- Johns Hopkins Lupus Center. 2022. *Creatine Phosphokinase (CPK) : Johns Hopkins Lupus Center*. [online] Available at: <https://www.hopkinslupus.org/lupus-tests/clinical-tests/creatine-phosphokinase-cpk/> [Accessed 9 May 2022].
- Jad, D., 2022. *Neuroleptic Malignant Syndrome: A Case Aimed at Raising Clinical Awareness. - PDF Download Free*. [online] docksci.com. Available at: <https://docksci.com/download/neuroleptic-malignant-syndrome-a-case-aimed-at-raising-clinical-awareness_5a34af6ad64ab2d5be907526.html> [Accessed 10 May 2022].
- Mount Sinai Health System. 2022. *Creatine phosphokinase test Information | Mount Sinai - New York*. [online] Available at: <https://www.mountsinai.org/health-library/tests/creatine-phosphokinase-test> [Accessed 9 May 2022].
- www.heart.org. 2022. *Ejection Fraction Heart Failure Measurement*. [online] Available at: <https://www.heart.org/en/health-topics/heart-failure/diagnosing-heart-failure/ejection-fraction-heart-failure-measurement> [Accessed 9 May 2022].
- Circulation. 2022. *Platelets and Cardiovascular Disease*. [online] Available at: <https://www.ahajournals.org/doi/10.1161/01.cir.0000086897.15588.4b> [Accessed 10 May 2022].

- Hopkinsmedicine.org. 2022. *What Are Platelets and Why Are They Important?*. [online] Available at: <https://www.hopkinsmedicine.org/health/conditions-and-diseases/what-are-platelets-and-why-are-they-important> [Accessed 10 May 2022].
- Sightdx.com. 2022. *Platelet count: definition, low vs normal vs high ranges*. [online] Available at: <https://www.sightdx.com/knowledge-center/platelet-count> [Accessed 10 May 2022].
- Mayoclinic.org. 2022. *Creatinine tests - Mayo Clinic*. [online] Available at: <https://www.mayoclinic.org/tests-procedures/creatinine-test/about/pac-20384646> [Accessed 10 May 2022].
- Oxford University Press. 2022. *The role of the kidney in heart failure*. [online] Available at: <https://academic.oup.com/eurheartj/article/33/17/2135/483602> [Accessed 10 May 2022].
- Health.qld.gov.au. 2022. *Reducing salt intake with heart failure*. [online] Available at: <https://www.health.qld.gov.au/__data/assets/pdf_file/0034/147877/cardiac_salt.pdf> [Accessed 10 May 2022].
- National Library of Medicine. 2022. *The prognosis of heart failure patients: Does sodium level play a significant role?*.
- eMedicineHealth. 2022. *High, Low, & Normal Creatinine Levels: What This Blood Test Means*. [online] Available at: <https://www.emedicinehealth.com/creatinine_blood_tests/article_em.htm> [Accessed 10 May 2022].
- Mithrakumar, M. (2019). *How to tune a Decision Tree?* [online] Medium. Available at: https://towardsdatascience.com/how-to-tune-a-decision-tree-f03721801680.
- Chicco, D. and 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*, 20(1). doi:10.1186/s12911-020-1023-5.