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.

ASSIGNMENT 2

Machine Learning in prediction of survival of patients with heart failure

Ngoc Bao Vy Le
s3828276@student.rmit.edu.au
RMIT Melbourne Australia

Date of report: 23/05/2021

Table of Content

Abstract ….....3

1. Introduction ….....3
2. Methodology ….....3
3. Result

3.1. Data Preparation 3

3.2. Data Exploration 3

3.2.1 Continuous features 3

3.2.2 Categorical Features.....4

3.2.3 Feature's association 5

3.3. Model Building

3.3.1 Feature engineering and Model selection.....7

3.3.2 Data Modelling.....8

4. Discussion.....9
5. Conclusion.....10

References.....11

Abstract

Heart failure is a problem that affects millions of people all over the world and is the leading cause of death globally. This study investigates how effective machine learning can be in predicting the future survival of patients who have suffered with heart failure using clinical records collected from the Allied Hospital in Faisalabad (Pakistan). The dataset was cleaned before being analyzed and supervised learning techniques were used to train the classification models, with the Hill Climbing technique being used for feature selection. The K-Nearest Neighbors and decision tree classifiers were used to make the predictions. A diverse range of results were found after analyzing the data, some contrary to previously recorded findings and others not. We believe this to be the result of dataset not being varied or large enough to accurately represent the true relationship between these pairs of attributes. This is once again reflected in the bias found in the decision tree. We conclude that machine learning can be effective in predicting the survival of these patients and that it can be used to predict the survival of patients with other illnesses and diseases, however, to make the most accurate predictions, a large and dynamic dataset must be used.

1. Introduction

According to the World Health Organization (WHO) in 2017, heart failure was the number one cause of death in the world. Even though the WHO states that most heart disease deaths in the world are in low and middle-income countries, it is also a significant problem in the developed world with heart failure being the leading cause of death in Australia in 2019 (Australian Bureau of Statistics 2020, Health section) and America in 2018 (Centers for Disease Control and Prevention 2018). The purpose of this report is investigating the effectiveness of machine learning in predicting survival of patients with heart failure by using heart failure clinical records.

The dataset that this report will be centered around is collected at Faisalabad Institute of Cardiology and at the Allied Hospital in Faisalabad (Punjab, Pakistan), during April–December 2015. It contains 12 independent predictive variables and a target variable for 299 observations.

The report will explore all the variables, observe the associations between them and use them for machine learning to predict the death event after the following period of patients with heart failure. There are different data modeling techniques will be used to train the data set, where its results as well as the recommendation of the data set will be discussed.

2. Methodology

The report used heart failure clinical data set from UCI website for machine learning to predict the death event during follow-up period of patients with heart failure. Before being analyzed, the data was cleaned to make sure that it was in the good condition for further stage. Since the target variable was in categorical form, supervised learning technique was used to train the model which was classification. In addition, the Hill Climbing technique was used for feature selection. To predict patient's death event, the report employed two different machine learning classification areas which were K-Nearest Neighbors and decision tree. To validate the chosen classification models, the data set was split in to two parts which were 25% for testing and 75% for train.

3. Results

3.1 Data preparation

According to code from data preparation part, the data set was quite clean. There was no missing value, no impossible value and typos. However, the data type for age and platelets were in float. therefore, they got typed cast to integer.

3.2 Data exploration

3.2.1 Continuous features

	age	creatinine_phosphokinase	ejection_fraction	platelets	serum_creatinine	serum_sodium	time
count	299.000000	299.000000	299.000000	299.000000	299.00000	299.000000	299.000000
mean	60.829431	581.839465	38.083612	263358.026756	1.39388	136.625418	130.260870
std	11.894997	970.287881	11.834841	97804.236869	1.03451	4.412477	77.614208
min	40.000000	23.000000	14.000000	25100.000000	0.50000	113.000000	4.000000
25%	51.000000	116.500000	30.000000	212500.000000	0.90000	134.000000	73.000000
50%	60.000000	250.000000	38.000000	262000.000000	1.10000	137.000000	115.000000
75%	70.000000	582.000000	45.000000	303500.000000	1.40000	140.000000	203.000000
max	95.000000	7861.000000	80.000000	850000.000000	9.40000	148.000000	285.000000

Table 1: Statistical quantitative description of the numberical features

From the statistical quantitative description of the numerical features table and the code from data exploration part, the report analyzed the distribution of age, ejection fraction, creatine phosphokinase, serum creatinine, serum sodium and platelets.

The age of patients was maximum at 95, minimum at 40, mean at 60.83 and median at 60 years. The distribution of age is right skewed which 50 percent of patients are in ranges 51 –70 years of age.

Creatine phosphokinase level was at the maximum at 7861, minimum at 23, mean at 581.84 and median at 250 mcg/L. Creatinine phosphokinase level skewed to the right and there were 50 percent of patients had creatinine phosphokinase in range 116 -582 mcg/L.

The ejection fraction level was at maximum at 80, minimum at 14, mean at 38.1 and median at 38 percent. The level of ejection fraction among patients were distributed evenly as it had a symmetrical shape which 50 percent of patients have ejection fraction level in range 30-45 percent.

Serum creatinine of patients had the maximum at 9.4, minimum at 0.5, mean at 1.4 and median 1.1 percent. Level of serum creatinine had a positive skew to the right which 50 percent of patients had serum creatinine in range 0.9-1.4 percent.

The serum sodium level of patients had the maximum at 148, minimum at 113, mean at 136.6 and median at 137 mEq/L. Level of serum sodium were distributed normally as it had a bell shape which 50 percent of patients had serum sodium level in range 130-140 mEq/L.

Level of platelets were maximum at 850000 and minimum at 25100, mean at 263358 and median at 262000 Kil o platelets/mL. The distribution of platelets level is positively skewed to the right which 50 percent of patients had platelets level in range 212500- 303500 Kilo platelets/mL.

3.2.2 Categorical Features

	anaemia	diabetes	high blood pressure	smoking
without	170	174	194	203
with	129	125	105	96

Table 2: Statistical quantitative description of the categorical features

Based on the above table, there were 129, 125, 105 and 96 patients with anaemia, diabetes, high blood pressure and smoking, respectively. When there were 170, 174, 194 and 203 patients without anaemia, diabetes, high blood pressure and smoking, respectively.

3.2.3 Feature's association

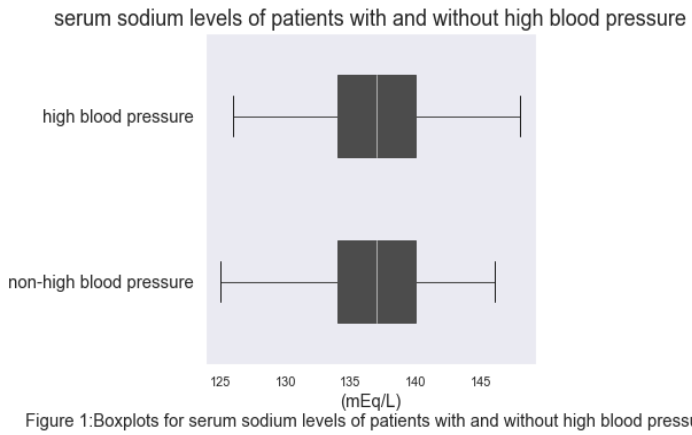

Figure 1:Boxplots for serum sodium levels of patients with and without high blood pressure

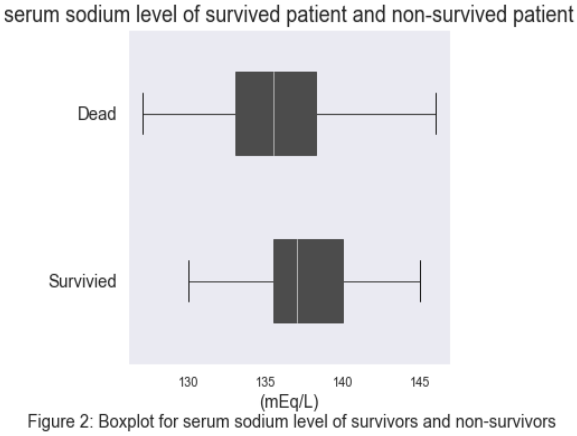

Figure 2: Boxplot for serum sodium level of survivors and non-survivors

Based on the association between serum sodium of patients with high blood pressure and without high blood pressure graph, the median, interquartile ranges, and overall ranges serum sodium level of patients with and without high blood pressure are same. In the other hand, the maximum and minimum values of serum sodium level for patients with high blood pressure are slightly greater than that of patients without high blood pressure. Serum sodium level of patients without high blood pressure appears to be slightly skewed to the left which means many patients have elevated level of serum sodium in this group. While the distribution of serum sodium level of patients with high blood pressure is symmetric. Overall, two batches of data look as if they were genetically distributed in a comparable way. The direction of this data indicates that high blood pressure condition does not vary with serum sodium level. This is contrary with much of the scientific literature, some of which claimed that there is a strong direct association between serum sodium and blood pressure. People with high blood pressure have elevated level of serum sodium (Wilfried et al, 2020).

The association between serum sodium levels of survived patients and non-survived patients graph shows that the median serum sodium level of patients who survived are greater than those who did not. The whole interquartile range of serum sodium level of survivors lies on the right side of non-survivor median. The interquartile and overall ranges of the data set are greater for non-survivors. The distribution of Serum sodium levels for the survivors and non-survivors appears to be slightly skewed to right. Overall, the direction of this data indicates that increased serum sodium level with better outcomes for survival of heart failure patients. This is contrary to (Macro et al, 2012), since they have found an association of increased serum sodium level with worse outcomes for survivors of heart failure patients.

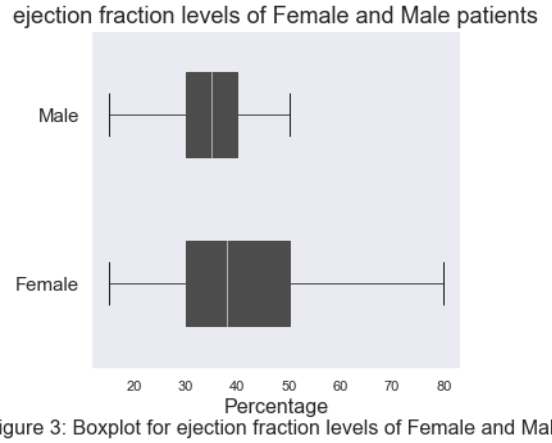

Figure 3: Boxplot for ejection fraction levels of Female and Male

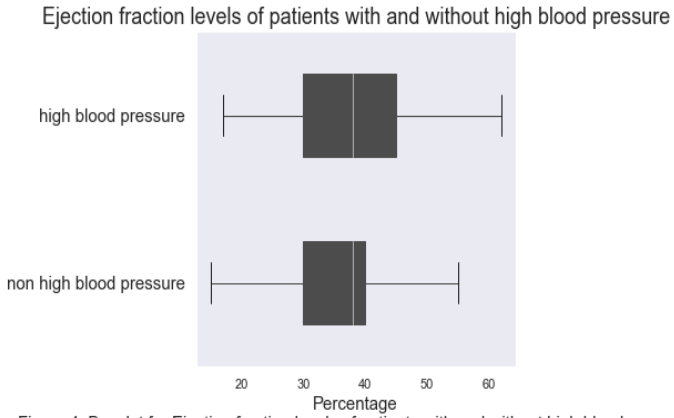

Figure 4: Boxplot for Ejection fraction levels of patients with and without high blood pressure

From the distribution of ejection fraction for Male and Female graph, there is a large discrepancy between the ejection fraction levels of males and females. Whilst the median value for females is only slightly larger than males, the range of ejection fraction levels is significantly larger than males and the interquartile range is approximately double. These findings suggest that there is a larger variance of ejection fraction levels within women compared to within men and that on average, women have higher ejection fraction levels than men. This find is contrary to US National Library of Medicine 2008, since there is no association between sex and ejection fraction.

The above boxplots display the difference in ejection fraction levels of patients with high blood pressure and without it. It can be viewed that the minimum and maximum values of patients with non-high blood pressure are lower than that of patients with high blood pressure. The median value for both sets of patients is quite similar, however, the upper quartile is much greater for those with high blood pressure. Both facts suggest that people with high blood pressure are more likely to have higher ejection fraction levels than those who do not have high blood pressure. This is contrary to (Ming et al, 2014), since it has been reported that those with a reduced ejection fraction levels tend to be hypotensive.

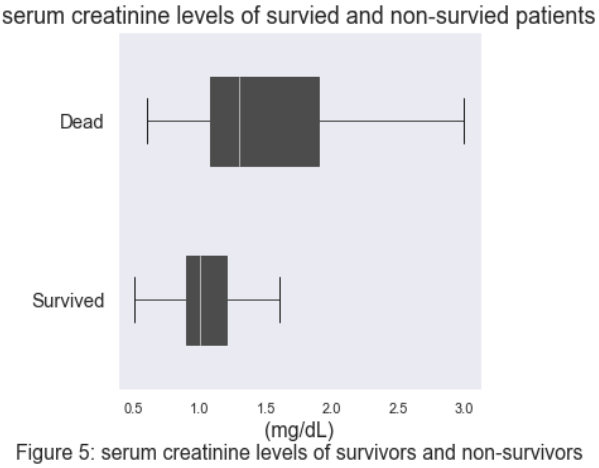

Figure 5: serum creatinine levels of survivors and non-survivors

The above graph shows that a significant difference in serum creatinine levels between patients who survived the follow-up period and those who did not. Firstly, the lowest amount of serum creatinine that was found in those who survived was lower than the lowest amount found out of the people who died. Secondly, the highest amount of serum creatinine found in patients who survived is less than half of what was found to be the highest amount in those who died. This resulted in a major difference in ranges between the two data sets. This ultimately means that the range of serum creatinine levels within those who survived is less than half of the levels found in those who died. It was also found that the values at Q1, Q2 and Q3 were all significantly higher for the patients who died compared to the patients who did not. From this data, we can see an association between higher serum creatinine levels and a higher chance of death. This is contrary to

(Davide &Giuseppe, 2020) as they have claimed indicated that long-term mortality increases with higher creatinine levels.

Based on the association between serum sodium of patients with and without diabetes graph in the code for data exploration, the serum sodium interquartile and overall ranges of patients with and without diabetes are quite similar. In the other hand, the median, maximum and minimum values of serum sodium level are greater for patients without diabetes. Distribution of Serum sodium level of patients with diabetes slightly skewed left. While the distribution of serum sodium level of patients without diabetes is symmetric. Overall, two batches of data look as if they were genetically distributed in a comparable way. The direction of this data indicates that diabetes condition associated with lower serum sodium level. The same find has been documented by (Giuseppe et al, 2017) since they have found low serum sodium level has been associated with increased insulin resistance, a condition that causes higher blood sugar and insulin levels.

The association between ejection fraction levels of smoking and non-smoking patients graph in the code for data exploration shows that the median, interquartile and overall ranges of ejection level is greater for non-smoking patients. The distribution of non-smoking patients is slightly to the right where smoking-patients is slightly to the left. Overall, the direction of the data indicates that the ejection fraction level of non-smoking patients is higher than smoking patients. This find is logical, since (Ming et al, 2014) has claimed that the tobacco smoking is associated with reduced in ejection fraction level.

As seen in the distribution of platelets for patients with and without high blood pressure graph in the code for data exploration, the platelet level of patients with high blood pressure is positively skewed, whilst the boxplot presenting the platelet levels of patients with non-high blood pressure is negatively skewed. While the median values are approximately the same, the minimum value for those with non-high blood pressure is significantly lower than that of those with high blood pressure. On the other hand, the maximum platelet level of those with high blood pressure was higher than those who did not have high blood pressure. This means that the range of both distributions was similar, thus it can be concluded that the patients with high blood pressure had a higher number of platelets on average than the patients who did not have high blood pressure. The same find has been documented by US National Library of Medicine 2016, since they have found increased platelet activation is involved in the pathogenesis of elevated blood pressure.

Based on the data which compares the distributions of ages for patients with and without anemia graph in the code for data exploration, it can be seen that within the group of patients with anemia, their maximum age was lower than the group of patients without anemia. Additionally, on average patients with anemia were younger than those without. Both facts suggests that the patients without anemia were able to live longer than those patients with it. The same find has been reported by (Bruce et al, 2006), since anemia are associated with increasing in risk for all-cause mortality.

When the platelet levels of patients who survived was compared to patients who did not, the data showed that the range of platelet levels found in those who died during the follow up period is larger than those who survived, having both a lower minimum value and higher maximum value. The interquartile range was also larger for those who died, however, the median values for both sets of patients was similar, as was the distribution over their respective ranges. This suggests that having either lower or higher platelet levels than average increases the risk of death. This trend is supported by evidence from (Villines et al, 2020), which states that a high count of platelets can result in blood clots and a low count can result in spontaneous bleeding, both of which increases the risk of death.

3.3 Model Building

3.3.1 Feature engineering and Model selection

There were 12 predictive features related to the death event during follow-up period of patients with heart failure in the data set. Since death event variable was in categorical form, Classification is the best model to use for the report. Hill climbing technique was used to pick the best features for prediction where features were selected based on their accuracy score. Only the features which the high scores were chosen.

Feature selection for KNN

Random state to shuffle the data was 4, then split the data into 25% for testing and 70% for training with random state at 2. After applying Hill climbing technique to the heart failure data set, there were 6 features selected which were time, anemia, sex, diabetes, high blood pressure and ejection fraction. These features were extracted for the data modelling.

Feature selection for Decision Tree

Random state to shuffle the data was 6, then split the data into 25% for testing and 70% for training with random state at 0. After applying Hill climbing technique to the heart failure data set, there were 4 features selected which were serum sodium, time, platelets, and anaemia. These features were extracted for the data modelling.

3.3.2 Data Modelling

The report used K-Neighbors and Decision Tree for classification.

	precision	recall	f1-score	support
0	0.93	0.98	0.95	54
1	0.94	0.81	0.87	21
accuracy			0.93	75
macro avg	0.94	0.90	0.91	75
weighted avg	0.93	0.93	0.93	75

Table 3: Classification report for KNeighbors classifier

	precision	recall	f1-score	support
0	0.87	0.89	0.88	54
1	0.70	0.67	0.68	21
accuracy			0.83	75
macro avg	0.79	0.78	0.78	75
weighted avg	0.82	0.83	0.83	75

Table 4: Classification report for Decision Tree Classifier

K-Neighbors Classifier

From the set of selected features, we split it into 25% for testing and 75% for training with the random state at 2 as this random state gave the best accuracy score. We found the best K value at 9 after testing K in range 3-15 and compared their scores. K-Neighbors classifier gave a score of accuracy at 0.93. The accuracy of positive prediction was 0.93 for survivors and 0.94 for non-survivors. The fraction of positives that was correctly identified for survivors was 0.98 and non-survivors was 0.81.

Decision Tree Classifier

From the set of selected features, we split it into 25% for testing and 75% for training with the random state at 7 as this random state gave the highest accuracy score. We found the best parameters for Decision Tree classifier which entropy, splitter at best, min samples split at 2, min sample leaf at 7 and default for other parameters. Decision Tree classifier gave a score of accuracy at 0.83. The accuracy of positive prediction was 0.87 for survivors and 0.70 for non-survivors. The fraction of positives that was correctly identified for survivors was 0.89 and non-survivors was 0.67.

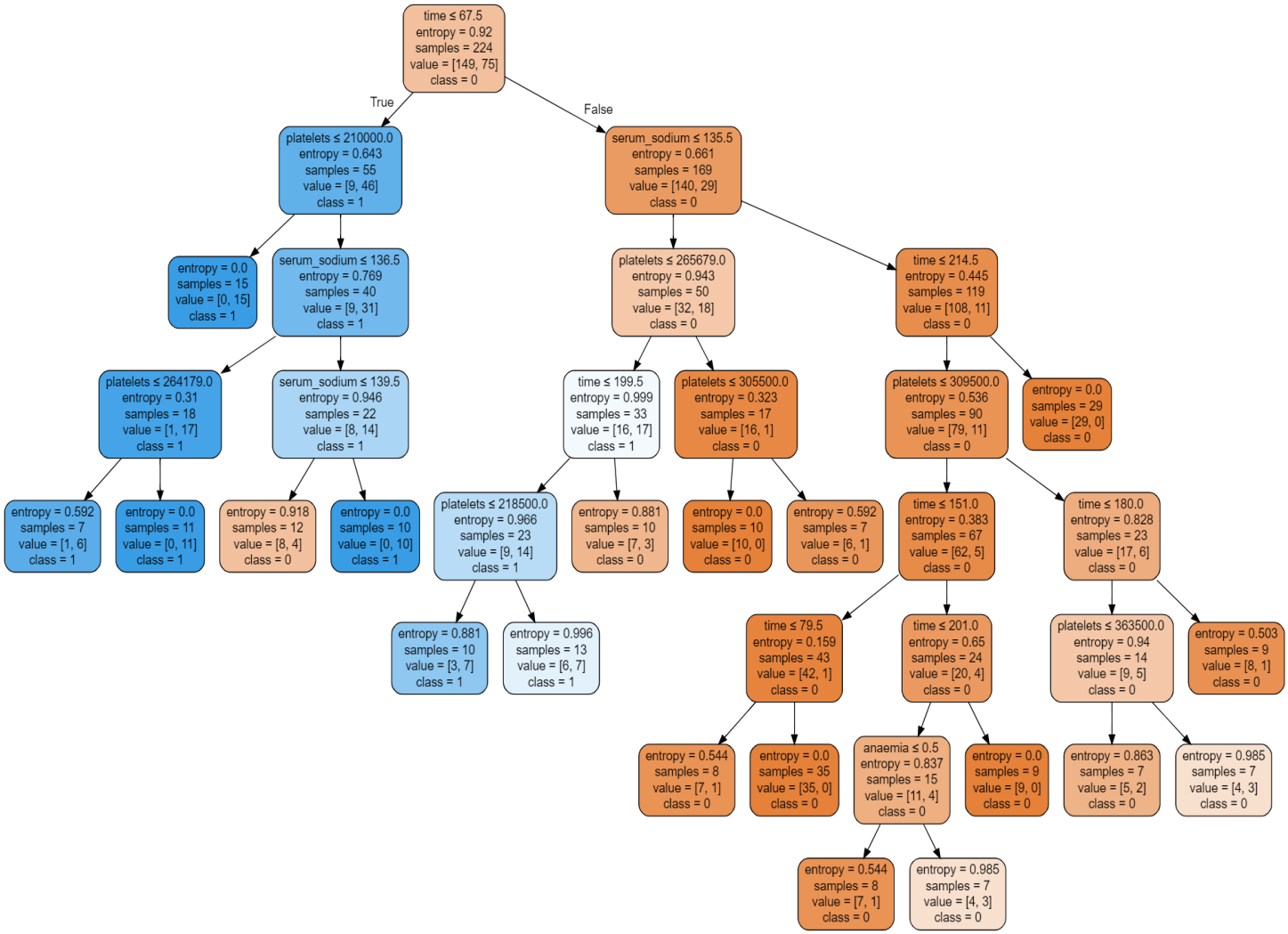

Figure 6: Decision Tree Classifier Visualization

4. Discussion

There are many ways to investigate the effectiveness of machine learning in predicting the survival of patients with heart failure using heart failure clinical records, however, some ways work better than others. We chose to model the data using two techniques: the K-Neighbors classifier and the Decision Tree classifier. Both classifier methods have their advantages and disadvantages, with the k-neighbors method being simpler and to the point compared to the more detailed and visual decision tree method, which presents an easy-to-read diagram of the prediction. When using k-neighbors, it gave an accuracy score of 0.93, which was similar to the accuracy score found with the decision tree, which was 0.83. If we were just to look at the accuracy scores alone, it would seem to be obvious that KNN can more accurately predict the data than a decision tree can, however, it can be seen above in the decision tree diagram that there is a much larger number of branches on the right side of the tree than the left side. This is due to the bias stemming from the initial split, allowing for survivors to dominate the branches compared to non-survivors and influences the dataset. The bias is the result of the data only being obtained from a single clinic in Pakistan and only having a limited and small number of patients to base the calculations off. This bias cannot be seen using the k-neighbors classifier. KNN is also considered a ‘lazy learner’ as it simply stores training data and waits until it is given test data to classify, whereas a decision tree learns from the training data and is typically faster than KNN as it has pre-calculated algorithms. Based on these facts, we have come to find that it is better to use a decision tree than a k-neighbor classifier. That being said, the histograms provided in the coding that compare two sets of data overlap with each other majority of the time because of the bias within the dataset. This would indicate that using classification methods with this dataset is not the most appropriate way to calculate predictions.

These methods of machine learning proved to be effective in predicting the survival of patients with heart failure as they had accuracy scores, however, further research should be conducted on data that has been collected from many different locations to make the dataset more dynamic, as well as using more methods of classification, to further validate these findings.

5. Conclusion

The K-Neighbor and Decision Tree classifier methods that we used demonstrated that machine learning can be used to effectively predict the survival of patients with heart failure, however, it can be seen that there is still room for improvement and different methods could provide greater effectiveness in the ability to make accurate predictions. Most importantly, the data used to make these predictions needs to large enough and have enough variation to produce the most accurate results. These same techniques could be used to predict the survivability of other illnesses and diseases with appropriately sized and dynamic datasets. More research can be done to see how different classification methods respond to different data and how the accuracy of their predictions may vary.

Reference

Australian Bureau of Statistics 2020, Causes of Death, Australia, viewed 22 March 2021, <https://www.abs.gov.au/statistics/health/causes-death/causes-death-australia/2019>.

Centers for Disease Control and Prevention 2018, Heart Disease Facts, viewed 22 March 2021, <https://www.cdc.gov/heartdisease/facts.htm> .

Centers for Disease Control and Prevention. 2020. High Blood Pressure Symptoms, Causes, and Problems | cdc.gov. [online] Available at: <https://www.cdc.gov/bloodpressure/about.htm> [Viewed 18 May2021].

Heart Research Australia 2021, Risk Factors, viewed 22 May 2021, <https://www.heartresearch.com.au/heart-disease/risk-factors/>

US National Library of Medicine 2008, USA, Patient Sex Does Not Modify Ejection Fraction as a Predictor of Death in Heart Failure: Insights from the approach Cohort, viewed 21 May 2021,

< https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2596502/ >

Davide, C &Giuseppe, J 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.

Wilfried, M, Kevin, D, Jeffrey, T, Pieter, M, Christian, M, Johan, L, Wilson, T, Hadi, S, Frederik, V, Francesco, O, Loreena, H, Dilek, U, Mitch, L, Patrick, R, Marco, M, Alexandre, M, Petar, S, Frank, R & Andrew, C, 2020, Evaluation of kidney function throughout the heart failure trajectory – a position statement from the Heart Failure Association of the European Society of Cardiology, *European Heart Journal* ,Vol. 22, Iss.4, pp.584-603.

Giuseppe, R, Cristiana, V & Petar, S, 2017, Heart Failure in Patients with Diabetes Mellitus, *European Heart Journal*, vol.3, no.1, pp.52-55.

Macro, M, Gad, C, Mihai, G, Livio, C & Adriaan, V, 2012, The role of the kidney in heart failure, *European Heart Journa*l, Vol.33, iss.17, pp.2135-2142.

Ming, L, Chin, C, Bryan, Y, Qing, Z, Yat, L, Rui, L, Jon, Sanderson, Andrew, C, Jing, S, Grabriel, Y and Cheuk, Y, 2014, Albumin levels predict survival in patients with heart failure and preserved ejection fraction, *European Journal Heart Failure*, vol.14, Iss.1, pp.39-44.

US National Library of Medicine 2016, USA, An association of platelet indices with blood pressure in Beijing adults: Applying quadratic inference function for a longitudinal study, viewed 21 May 2021,<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5265936/#:~:text=Several%20studies%20have%20reported%20that,with%20cardiovascular%20morbidity%20and%20mortality.&text=Increased%20platelet%20activation%20and%20aggregation,associated%20with%20hypertensive%20risk%20factors. >

Bruce, C, Braden, M, Jiangui, Z, Marcello, T, Scott, K and Brenda, H, 2006, Impact of anemia on hospitalization and mortality in older adults, *Blood Journals*, vol.107, iss.10, pp.3841-3846.

Villines, Z, Keith Fisher, J, Medical News Today 2020, USA, What do high or low platelet count levels mean?, viewed 22 May 2021, < https://www.medicalnewstoday.com/articles/medical-team#Meet-Our-Medical-Network>