Computer Science & IT, School of Science

COSC2670 Practical Data Science

23rd May 2021

Assignment 2: Data Modelling and Presentation

Report

Student ID	Student Name	Contact details
s3861921	Bharti Sanjeebkumar Sinha	S3861921@student.rmit.edu.au RMIT University, Melbourne Australia
s3853868	Ram Rattan Goyal	S3853868@student.rmit.edu.au RMIT University, Melbourne Australia

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

Table of Contents

Topic	Page Number
Abstract	3
Introduction	3
Methodology	3
- Data Retrieving	4
- Data Preparation	4
- Data Exploration	5
- Data Modelling	10
Results	12
Discussion	12
Conclusion	12

Abstract

The aim of this data science process is to predict the survival of patients with history of heart failures, during their follow up period. Heart failure clinical records Data was used for the analysis of 299 patients. (Dua, D. and Graff, C., 2019). 13 clinical features were noted for each of the patients, including their 'age' (years), whether they had anemia (Boolean), high blood pressure (Boolean) or diabetes. Their level of the CPK enzyme in the blood (mcg/L) and percentage of blood leaving the heart at each contraction (ejection fraction) was recorded as well. The dataset also shows records of patient's sex (binary), platelets level in the blood (kilo platelets/ mL), level of serum creatinine in the blood (mg/dL), level of serum sodium in the blood (mEq/L), whether they smoked (Boolean) and the follow-up period (days). The data science process is supervised as the outcome - 'DEATH_EVENT' which depicts if the patient died during the follow up period is mentioned for each observation in the given dataset. The dataset with all the variables of 299 patient seems adequate to predict a new patient's survival. The results indicate that the level of serum_creatinine, level of ejection fraction and the follow-up period are majorly responsible in determining the death by heart failure. The report concludes that these primary features must be kept under proper limits in order to increase the chances of survival. It is recommended that while dealing with the patients with probable heart failures, these characteristics must be in regular check and be used to restrict the probability of death to some extent.

Introduction

The sedentary lifestyle of people has been increasing and has only exacerbated during these covid situation. This lifestyle combined with food habits lead to problems associated with heart. Heart diseases has been seen as one of the biggest causes of the demise of a large proportion of population. Heart failure is one of the critical heart diseases in which the heart cannot pump or fill adequately. There has been repeated practices in order to reduce the number of deaths caused by heart failure. To deteriorate the numbers of such cases, it may be a significant approach to try and predict the outcome of a patient beforehand. This report will discuss analysis of some medical parameters which could play a crucial role in determining if a person will suffer death while diagnosed with heart failure or not.

Methodology

Source of Dataset: The original dataset version was collected by Tanvir Ahmad, Assia Munir, Sajjad Haider Bhatti, Muhammad Aftab, and Muhammad Ali Raza (Government College University, Faisalabad, Pakistan) (archive.ics.uci.edu, n.d.). The current version of the dataset was elaborated by Davide Chicco (Krembil Research Institute, Toronto, Canada) and donated to the University of California Irvine Machine Learning Repository under the same Attribution 4.0 International (CC BY 4.0) copyright in January 2020. (Dua, D. and Graff, C., 2019).

The data science process involves several steps such as Data Retrieval, Data Preparation, Data Exploration, Data Modeling and Model scoring. Each of the steps involved has been discussed below.

Step 1. Data Retrieving

Data Retrieving process ensures that the data loaded into the workspace using pandas library (*read_csv* function) was identical to the source of the data. *head()* and *tail()* functions of the pandas library reveal the first and the last 5 observations respectively, which give insights into the appearance of the dataset, which has 299 observations and 13 attributes. *info()* function on the dataset, reveals the column names and number of null values in each of the columns. There were no null values found in any of the attributes and the data types of each of the columns were same as those in the source of the dataset.

Step 2. Data Preparation

Data Preparation ensures that the data used for training the model is true representation of reality and is hence free of all errors. None of the columns in the dataset had any impossible or missing values. For instance, there were no age values below 0 or above 100 years. There were no bad values or data entry errors in the categorical variables such as anemia, high blood pressure and diabetes. The percentage of blood pumped by blood at each contraction remained between 14 to 80, i.e., within reasonable limits not exceeding 100 or below 0.

Neither missing values nor negative values were observed in the other columns such as platelets, serum_creatinine, serum_sodium, sex, smoking, time and death event. However, there were some outliers observed in the 'platelets' and 'serum_creatinine' column. The mean of 'platelets' was 263358.029264 kiloplatelets/mL and 'serum_creatinine' was 1.39388 mg/dL. The maximum value of both columns was 840000 kiloplatelets/mL and 9.4 mg/dL respectively. While their upper quartiles were 303500 kiloplatelets/mL and 1.4 mg/dL respectively, such a hike in the maximum value could not be explained and therefore the presence of outliers was confirmed using a boxplot. These were replaced using the median of the columns to ensure as least as information as possible to be removed in this process. The follow-up time of the patient remained within 0 to 300 days while serum_sodium levels remained under 150 mEq/L for all patients. The binary columns such as sex, smoking and DEATH_EVENT had two possible values – 0 and 1.

Removal of errors play significant role in data science process as it can lead to faulty decision making which could have adverse effect on the system in which the model is deployed. (Source: Lecture material week 1 slides, Y.Ren)

Step 3. Data Exploration

From Figure 1, it is observed that the minimum and maximum ages of the patients who had heart failures are 40 and 95 respectively. The mean age of patients in the dataset is about 61 years. Three quarters of the patients under study are below 70 years of age and about a quarter are below 51 years of age. Histogram distribution reveals that the number of people who had heart failures and were recorded in the dataset, between ages of 40 to about 75 was higher than those between approximately 75 and 95.

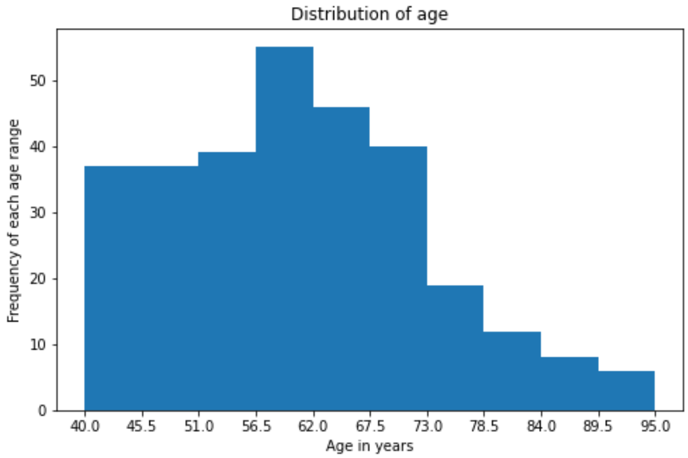


Figure 1

129 of 299 patients had anemia. The mean age of patients with anemia is 62.03 years and those who don't have anemia is about 60 years. It can be observed that composition of anemic (35% for death, 65% for non-death) and non-anemic patients (29% for death, 71% for non-death) within death cases (death, non-death) during their follow-up period in the scope of the concerned dataset are comparable and hence will not bias the model significantly. It was concluded from calculating death rate of anemic and non-anemic patients. From Figure 3, it was found that 56.86% patients in the study were not anemic. Figure 2 reveals about 42% of the patients were diabetic and Figure 4 highlights about 65% of the patients did not have high blood pressure, amongst those considered for the study.

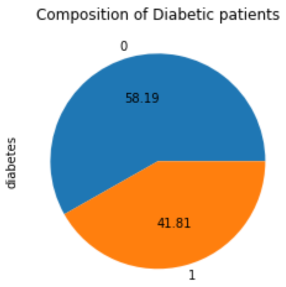


Figure 2

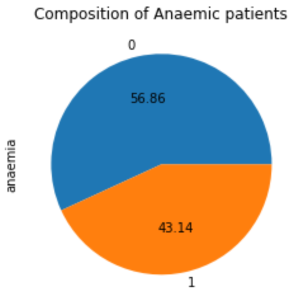


Figure 3

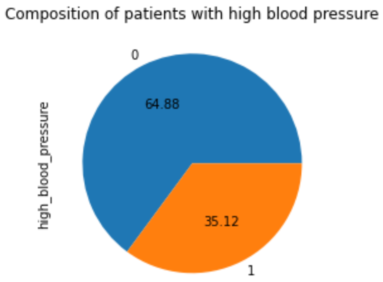
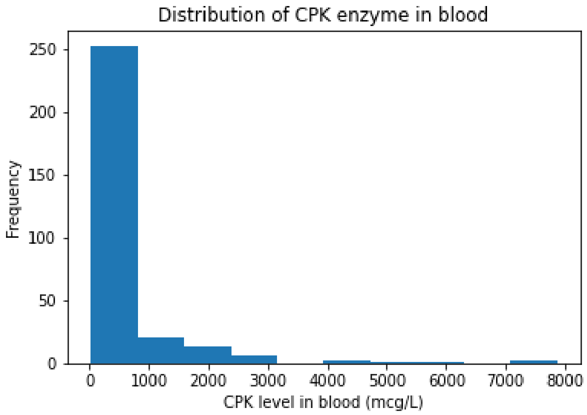
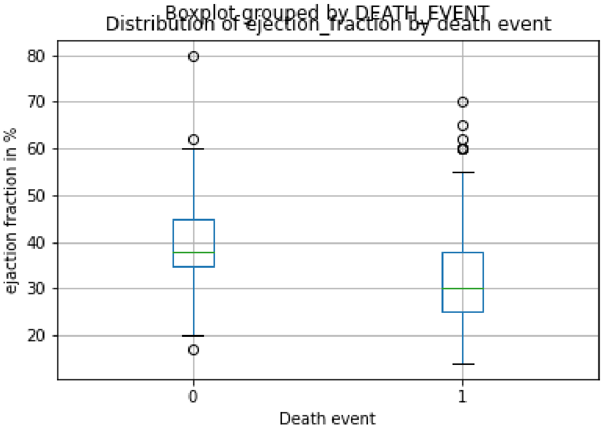


Figure 4

Figure 5 highlights the minimum and the maximum value of the amount of creatinine phosphokinase (CPK) enzyme found in the blood to be 23 mcg/L and 7861 mcg/L approximately. Histogram distribution of the level of CPK in blood depicts that most of the patients had the CPK level below 1000 mcg/L, and 7 observations had values above 3500 mcg/L. .

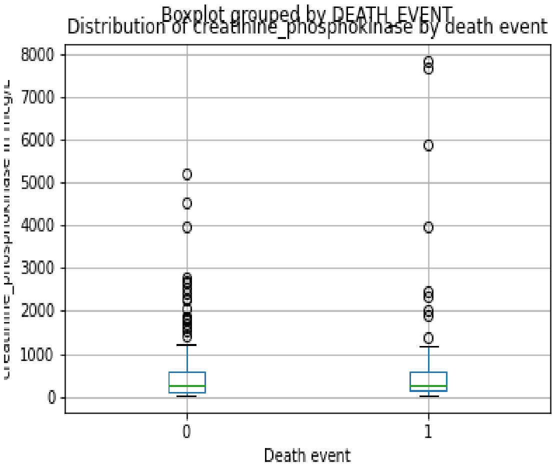
Figure 5





Ejection fraction which depicts percentage of blood leaving the heart at each contraction did not go below or above 100, with maximum being 80% and minimum being 14%. 105 patients with high blood pressure were noted in the dataset. A significant insight from the figure 6 is that the average ejection fraction in patients who died were lower (30%) than those who survived (about 39%).

Figure 6.



From Figure 7 it can be observed that there are several patients with CPK levels above 1291.5 mcg/L (Q3+ 1.5(IQR)). The average CPK levels of patients who died was about 540 mcg/L and for those who didn't die it was about 670 mcg/L as concluded from Figure 7 box plot. Combining this with insights from figure 5, it can be understood that the distribution of CPK level is left skewed with 250 patients (out of 299) with their CPK level between 0 – 1000. Therefore, as part of improving the models, more patients with higher CPK level should be included in the data for better prediction.

Figure 7

Using the medium of a scatter plots it was observed that patients with platelet levels between 150000 kiloplatelets/dL and 350000 kiloplatelets/dL generally had their serum creatinine levels between 0.8 to 1.2 while the levels of serum_sodium within the same range of platelets was less distributed and fell among the domain of 130 mEq/L to 140 mEq/L. Females had higher concentration of platelets than men but non-smokers and smokers generally had the same levels of platelets as observed from two respective boxplots. No information could be derived for the follow-up period from the platelet levels.

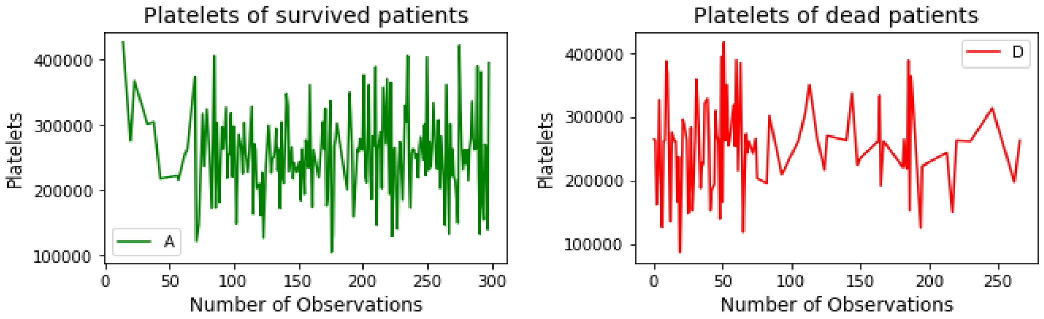
The plots between serum creatinine and serum sodium revealed some interesting findings in which it was found that patients with serum creatinine levels in the range 0.7 mg/dL to 1.3 mg/dL mostly had their serum sodium between 138 mEq/L to 143 mEq/L. Females had lower amounts of serum creatinine than males unlike in the case of platelets as per the information derived from the respective boxplot. The concentration of serum creatinine in the case of non-smokers and smokers was found to be near equal which was quite an unprecedented finding. Patients with serum sodium levels less than1.3 mg/dL were observed to have a wide domain of follow-up days from 4 days to 283 days while the higher ones usually live up to either a small number of days (less than 124) or days more than 164.

In case of serum_sodium, both males and females follow the same tradition as in the case of platelets i.e. which females had a higher concentration of serum_sodium than males as concluded from a boxplot. Using the same type of graph, it was seen that most non-smokers seem to have their levels between 134 mEq/L and 140 mEq/L but most smokers had their serum sodium levels starting from 136 mEq/L with the same upper limit as non-smokers. However, the most interesting observation was informed by the scatter plot in this case in which it was seen that people who lived the maximum number of days had their serum sodium levels greater than 134 mEq/L.

As high as 95.8% of the whole population of smokers was shared by males and only a small percentage of 4.2% of females were found to be smokers as depicted by the pie graph. Both females and males lived up to the same number of days as portrayed by the boxplot.

From the boxplot between smokers and non-smokers, the average expectancy of the number of days for smokers and non-smokers was seen to be near equal while the greatest number of non-smokers seemed to have more days.

The process of data exploration with these columns in which graphs of these with the target variables did reveal some useful inference which could be decoded into some useful information. The patients who survived generally had their serum creatinine levels within 0.6 mg/dL o 1.5 mg/dL while those who could not had their levels above 1.5 mg/dL. This makes the serum creatinine information very crucial. But in case of serum sodium , no such relation could be observed. And so is the case with platelets in



which the deceased patients and those who made it alive nearly share the same domain or the concentration of platelets.

Figure 8

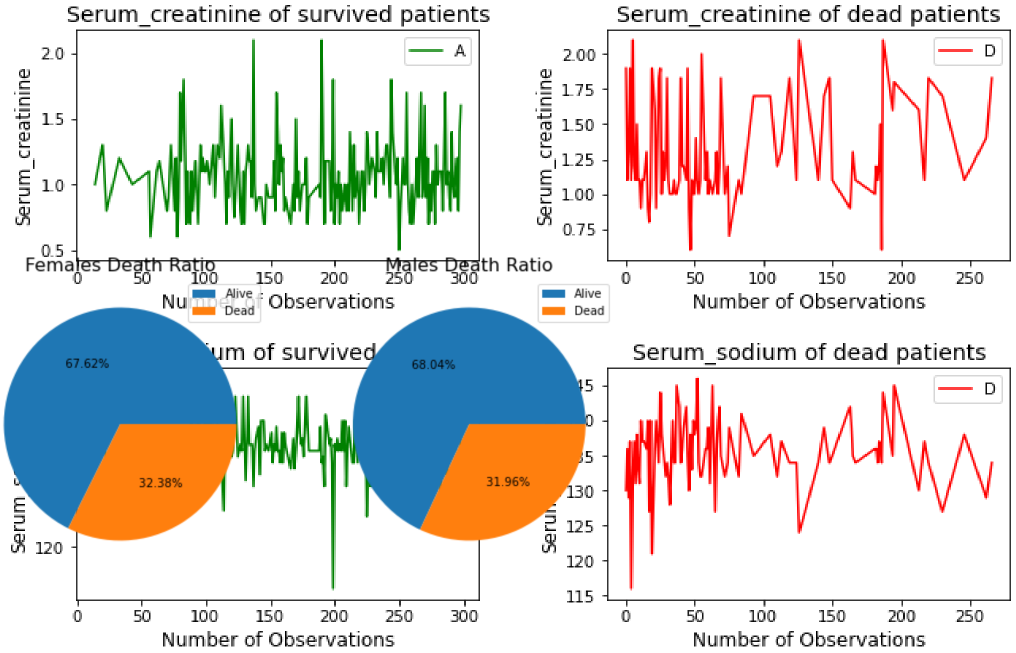
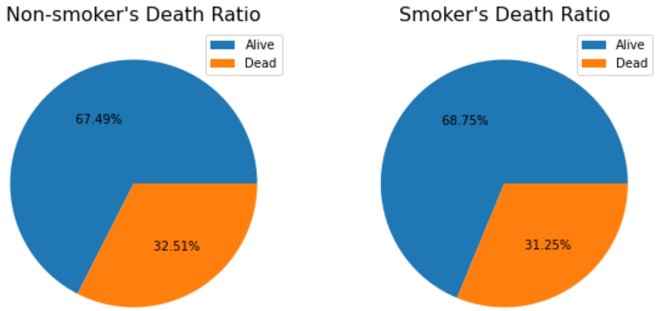


Figure 9

When we consider gender to have a relation with the death ratio of the patients it is seen that nearly 67.42% females survived the mishap of heart failure while 32.38% of them met their demise, a similar trend was seen in the case of males where 68.04% of the males were alive and 31.96% of them died. Hence, no relation could be derived from here.



A similar trend is observed in accordance with the people who smoke and don't smoke. The ones who made it alive and don't smoke shared a composition of 67.49% and those who smoke had a proportion of 68.75% respectively while the rest were deceased.

Figure 10

One of the most interesting facts that this report is to find in the upcoming sections are based on the observations of figure 10. As it can be clearly seen that the lines representing the patients who met their demise lived different number of days from 10 days to around 200 days, and those who survived the heart failure had their follow-up period in a similar domain which is from around 0 days to more than 250 days .Both have a very similar trend which could mean that this data could not yield some good results at the end but the contradiction to this was observed in the latter part of the report.

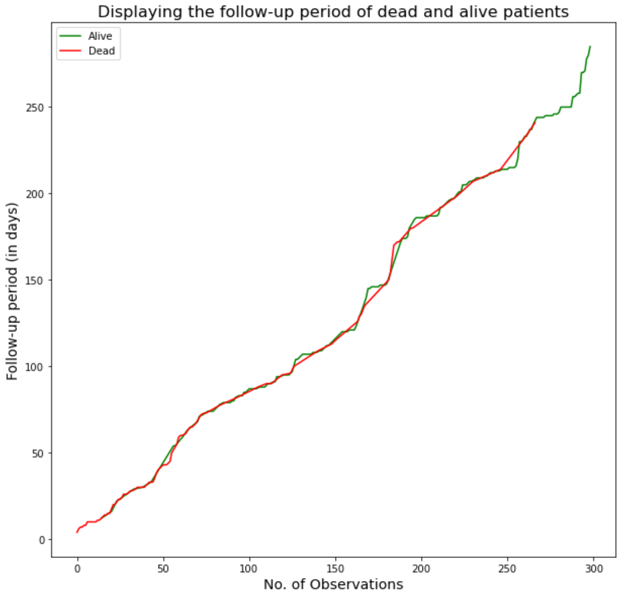


Figure 11

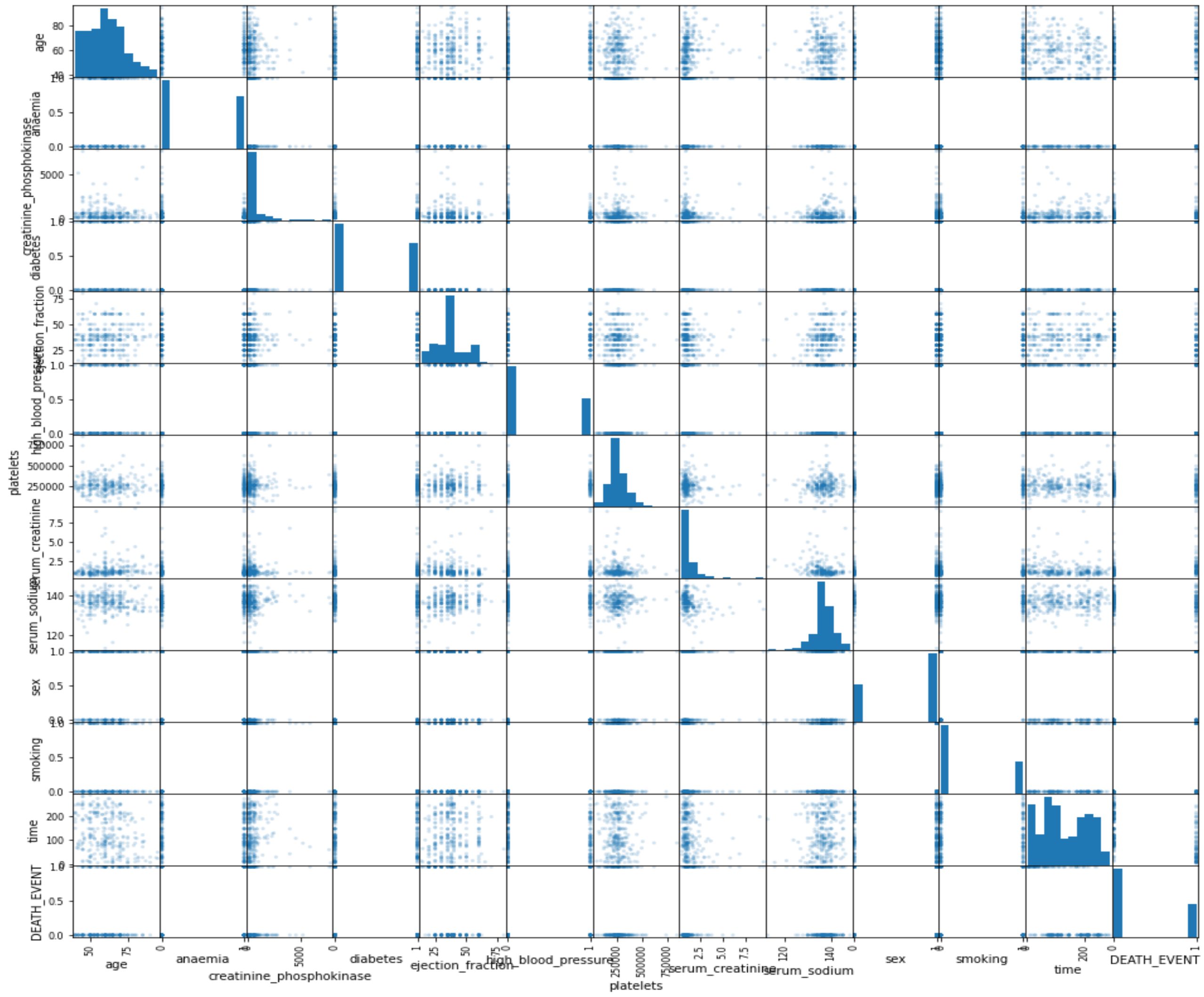


Figure 12 Scatter Matrix: Scatter plot between columns

Scatter plot helps us understand the relationship between different variables in the dataset. If the relation between any two variables is linear then it indicates high dependency between them and can significantly affect the model prediction. The models considered for the given research problem are k-NN and Decision tree. K-NN assumes that there is no multi-collinearity between variables. Figure 12 indicates there is no multi-collinearity between variables.

Step 3 Data Modelling

For prediction of patient's survival (0 or 1) from the attributes during the follow up period, two classification models were developed, kNN and Decision Tree.

kNN

KNN algorithm works by identifying k nearest neighbors of a new observation from the training data. Validation strategies deployed were holdout validation and k-folds validation.

During hold out validation, the data split was 75%-25% for training and testing respectively. All the columns were considered for training and predicting the survival of a patient. The kNN model can adapt to different p values which stands for methods of distance calculations such as Manhattan distance (p = 1) and Euclidean distance (p=2), and 'uniform' and 'distance' weights which indicate how much should k neighbors weigh-in. In uniform weights, all the k neighbors contribute equally to the prediction whereas in distance weight, closer ones contribute more. The values of nearest neighbors were changed iteratively from 2 to 7 for uniform weights and Euclidean distance, 'distance' weights and Euclidean distance, 'distance' weights and Manhattan distance. The best accuracy was obtained from k = 5, 'distance' weights and Manhattan distance with accuracy score of 0.786. This can be understood as prediction improves when close neighbors contribute more and Manhattan distance with a lower p value. High value of p results in ignoring several columns. The accuracy can be observed in table 1.

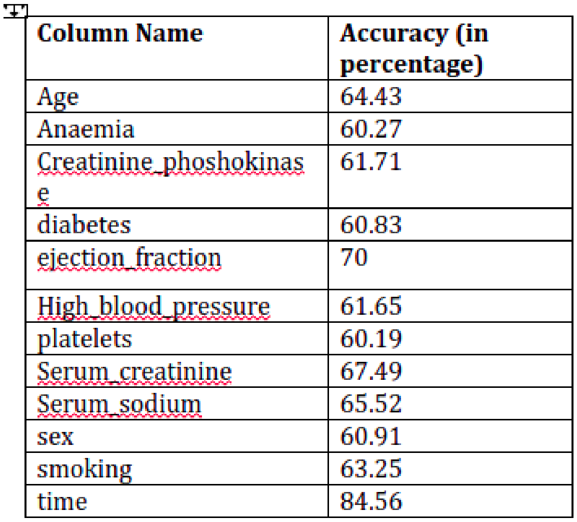
Cross validation strategies such as k-fold are more rigorous. Different parameters for k-NN models were used for different splits of the dataset. Value of k for k-NN model was used from the set {3,5,7} and value of k for k-fold validation was used from the set {5,7}. For each value of k-fold, all three kNN models were trained with model parameters such as uniform weights and Euclidean distance, 'distance' weights and Euclidean distance, 'distance' weights and Manhattan distance. Overall 3NN model performed better than 5NN and 7NN models. The accuracy score of k=3 model improved significantly from 0.77333 when holdout validation was used to 0.94915 when 5-fold validation was used.

The overall performance of the model may look satisfactory to some extent, however there is still some scope for improvement as the marvelous concept of feature engineering still waits to be applied. Some of the following approaches were taken in order to identify the best columns for data modelling.

The accuracy of all the columns were found to be as demonstrated in the table (the accuracy could slightly differ in case the code is executed again).

Table 1

The columns 'time', 'serum_creatinine' and 'ejection_fraction' seemed to hold a high significance in determining the target variable. These columns when fitted to the k-Nearest Neighbours model yield a max accuracy of around 93.33% after the calculation of the perfect value of k in different iterations. This when validated using KFold validation across different splits yield upto an accuracy of around 97%.



Column Name	Accuracy (in percentage)
Age	64.43
Anaemia	60.27
Creatinine_phoshokinase	61.71
diabetes	60.83
ejection_fraction	70
High_blood_pressure	61.65
platelets	60.19
Serum_creatinine	67.49
Serum_sodium	65.52
sex	60.91
smoking	63.25
time	84.56

Decision Tree

Decision tree is a more explainable model with various parameters such as criterion to measure the quality of split, max_features to consider for split, min_samples_leaf indicates minimum number of samples required to consider a node a as leaf. Decision trees are prone to overfitting. Thus, choice of parameters plays a crucial role in decision tree modelling.

First decision tree was modeled with default parameters. With precision of 84% and recall of 85%. To understand the impact of the number of samples to split a node for decision, all the parameters except 'min_samples_split' values were kept unchanged. It was noticed that upon increasing number of samples to consider for node splitting there was no change in model scoring. However, an increase in max_depth led to reduction in the accuracy of the model. In the last iteration, as accuracy improved with increasing min_samples_split and increasing max_depth separately, so these were changes together to see the net effect. Changing max_depth and min_samples_split simultaneously leads to 84% precision and 85% accuracy.

Table 2

Decision tree (considering all features)	Precision	Accuracy
max_depth=None, min_samples_split=2, min_samples_leaf=1 --->default	85%	77%
max_depth=3, min_samples_split=3, min_samples_leaf=1	88%	88%
max_depth=3, min_samples_split=4, min_samples_leaf=1	88%	88%

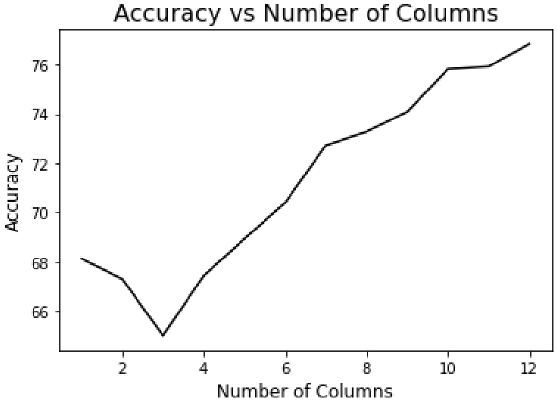
max_depth=3, min_samples_split=8, min_samples_leaf=1
max_depth=3, min_samples_split=20, min_samples_leaf=1
	88%	88%

criterion='gini', max_depth=6, min_samples_split=3, min_samples_leaf=1,	88%	82%
max_depth=4, min_samples_split=4, min_samples_leaf=1,	84%	85%

Decision tree was used as another algorithm and the graph to the left depicts the accuracy calculated for over 50 iterations with different number of columns. From Figure 13, it can be concluded that in case of a decision tree the average accuracy did seem to change a bit but still the results were lower than that of k-

Nearest Neighbors. However, with feature engineering applied to this model, the metrics yielded a better result than k-Nearest Neighbours.

Figure 13



Results

Knn	accuracy score
2	0.62667
3	0.68000
4	0.73333
5	0.78667
6	0.73333
7	0.74667

Table 1

Include tables of 5NN and 7nn

The performance of 5NN improved from accuracy score: 0.73333 (default weights = uniform, Euclidean distance) to 0.78667 (weights = distance, Euclidean distance) during hold out validation. The accuracy further improved to 0.91525 (weights = uniform, Euclidean distance) and 0.95238 (weights="distance") during 5fold validation. The accuracy improved to 0.95238 when weights = distance and Euclidean metric (p=2) was used for 3NN model was the best performing model.

Folds	Test Score
1	74.41
2	62.79
3	65.11
4	79.07
5	76.74
6	83.33
7	97.62

Better results were noticed in all the folds when the model was presented with featured columns. The test score calculated as the accuracy of the prediction saw non-linear trend from descending from 74.41% to 62.79%, to ascending from 62.79% to 65.11%, the highest being 97.62%. (Note- the percentages could be slightly different but nearly same in case of different iterations).

Discussion

With the presence of some known medical parameters of a patient, it is possible to predict the outcome of the heart failure situation to some extent. Certain features have more significance and may dominate the spectrum of prediction. However, the data presented for training in this case isn't valuable enough to produce excellent results and there are chances of unprecedented data leakage. In the future, with the help of some bigger size and more relevant data, this research could yield higher rewards.

Conclusion

The results of this model can be used to predict survival of patients given parameters such serum_creatinine, level of ejection fraction and the follow-up period. The data is not balanced and does not represent the population of people with incidences of heart failures.

The knn model was found to be 95% accurate with 7fold cross validation and 3 nearest Neighbour in place. Decision model performed well with default values considering all the features, criterion='gini', max_depth=None, min_samples_split=2, min_samples_leaf=1.

This research has proved that with the availability of relevant parameters, this model could be reproduced to help in predicting other types of diseases.

The study of this data has revealed that there are instances where unidentifiable data leakage could be present and there are still less implementations where we could solve these types of problems. There needs to be work done on this field.

References

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository *UCI Machine Learning Repository: Heart failure clinical records Data Set,* Irvine, CA: University of California, School of Information and Computer Science. Available at: < UCI Machine Learning Repository: Heart failure clinical records Data Set >