Predicting survival rate in patients following heart failure using machine learning

15/05/2021

Jingxuan Feng
RMIT University
s3843790@student.rmit.edu.au

Ho Yin Lam
RMIT University
s3889140@student.rmit.edu.au

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

Abstract
With the increased interest and usage of artificial intelligence by tech giants, they offer a varying range of services that improve our quality of life. These can come in the form of digital assistants such Alexa, Google home or embedded in self driving cars. We will also explore the feasibility of utilising machine learning to predict the likeliness patients surviving a heart failure based off a range of medical conditions and readings. We will investigate each attribute, what they represent and whether they play a part in helping to predict survival rate after heart failure.

Introduction
We will be conduction our testing on the "Heart failure clinical records Data Set". This dataset contains information on 299 patients from 2015. The patients included 105 females and 194 males, all ranging between 40 and 95. We will be investigating how the 13 features relating to the patients' medical readings, conditions and lifestyle will affect the survival rate.

Dataset Information [4]

- Anaemia, high blood pressure, diabetes, gender, smoking, and death event are all Boolean values. They will represent whether the feature is present or not.
- Age of the patients is measured in years.
- creatinine phosphokinase (CPK) measures the level of CPK enzyme in the blood. Measured in mcg/L.
- ejection fraction measures the percentage of blood exiting the heart after each contraction.
- Platelets is the number of platelets present in the blood. Measured in kilo platelets/mL.
- serum creatinine is the level of serum creatinine in the blood. Measured in mg/dL.
- Time is the number of days in the follow up period.

Methodology
Data Cleaning

First, we will ensure the dataset is clean before any analysis or modelling is performed. We checked each column for errors as well as outliers. We found that several columns did, in fact include outliers but decided not to remove nor change any of these values as they are likely side effects of a medical symptom and can proceed to possible hypothesis.

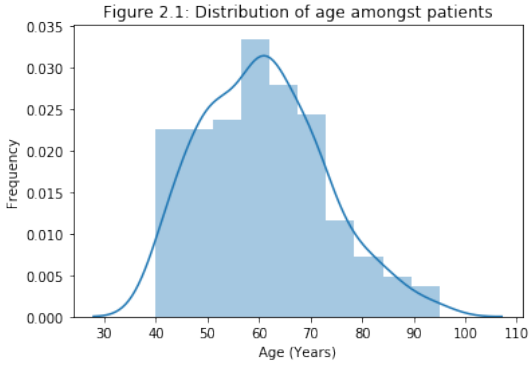
Data Exploration

Afterward, we proceeded to explore the data. We decided to standardise the way we explore the data when looking at the feature by itself. For all Boolean data, we used a pie chart to compare the percentage 'yes' vs 'no' for the feature. This allows us to see how prevalent each case is compared to each other.

For continuous data, we used a frequency histogram. This allows us to visualise how the values are distributed amongst the patients, as well as how they are grouped together. Where a scatter plot was applying to the time variable by index.

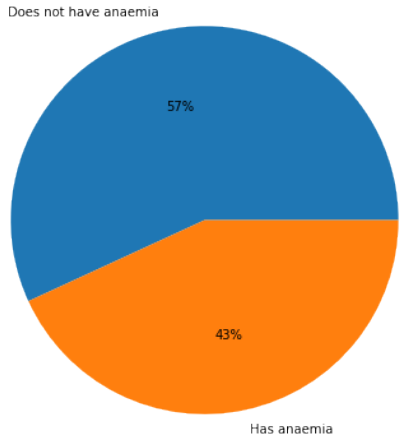
For figure 2.1 that represent the frequency histogram with kernel density estimate of the data for variable age which shows that there is a larger number of younger patients surveyed, evident by right-skewness on the graph.

Figure 2.1: Distribution of age amongst patients



Anaemia which is a Boolean value it gives weather the patient have anaemia or not, from the figure we can see that 43% of the patients have anaemia.

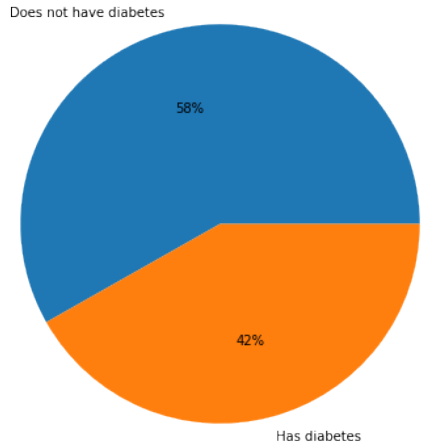
Figure 2.2 Percentage of patients with and without anaemia



For Creatinine Phosphokinase as observed in the figure in the ipynb file , the majority of patients have a CPK level within the IQR of 465.5mcg/L. Values higher than the upper outlier threshold of 1280.25 could indicate the patient has underlying health issues or symptoms

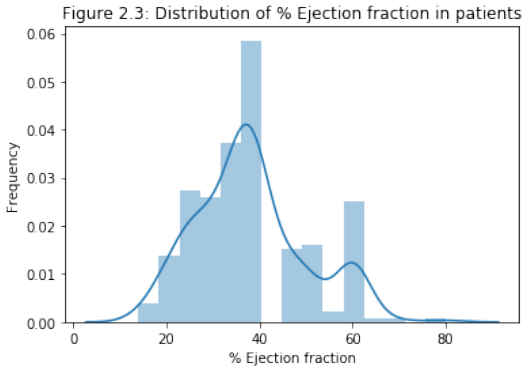
Diabetes, as shown in the figure, we can see here that a large number (42%) of patients in the dataset has diabetes.

Figure 2.4 Percentage of patients with and without anaemia



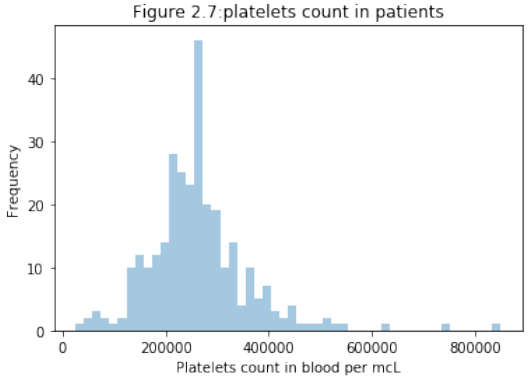
Ejection fraction, which is how much blood the left ventricle pumps out with each contraction. (American Heart Association, 2017) Figure 2.2 shows the ejection fraction of the patient and as the

graph shows us, a large portion of the patients are distributed within the IQR. There are only a few patients located beyond the upper outlier threshold of 67.5, which could be an indicator of health complications.

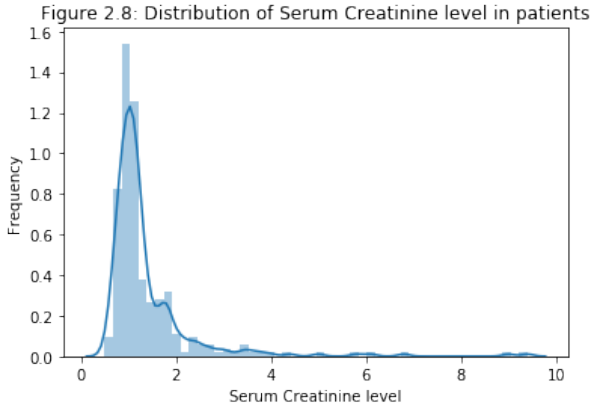


From the figure in the python file we can see that more than half (65%) of patients in the dataset has does not have high blood pressure

Figure 2.7 represent the histogram of the patient platelets, with the y axis being the number of patients, and the x axis is the platelets counts in blood per mcL, whereas the ideal range of platelets in human are 150,000 to 400,000 per mcL(Balduini & Noris, 2014) from the figure we can see that most of the patients data lies in between the ideal range level.

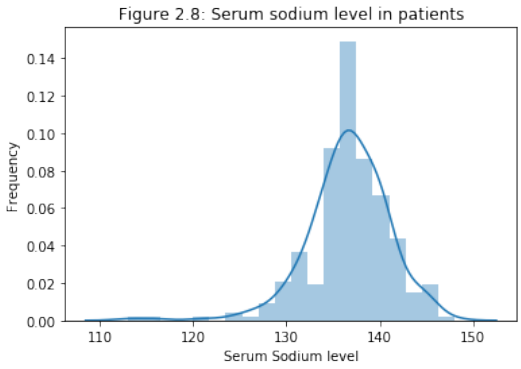


Serum Creatinine which can see from the figure that a very strong right skewed shape for the data.



The ideal range of serum sodium in human are 135-145 mmol/L (McKee et al., 2015) and from the figure 2.9 and from the code we can see that 214 out of 300 patients have the ideal range of serum sodium level.

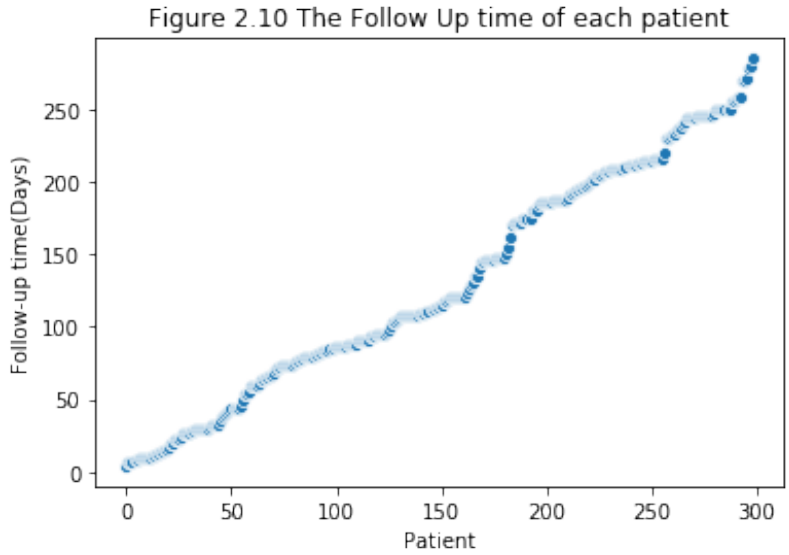
Figure 2.8: Serum sodium level in patients



65% of the patients are male, as well as from the data we can also analysis that about 68% of the patients does not smoke.

For the time factor, we can see that the follow up time increase with patients from 4 days to 285days.

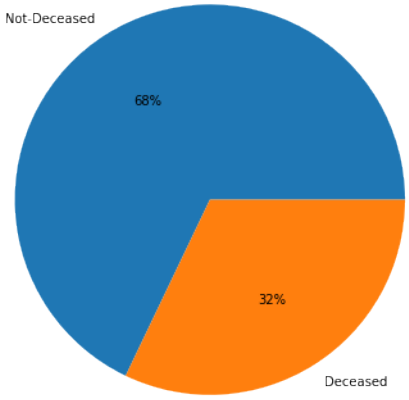
Figure 2.10 The Follow Up time of each patient



I

Along with the follow up days data, weather the patients were deceased during the follow up period were shown in figure 2.11, and 32% of the patients were deceased during the time.

Figure 2.11 Percentage of patients are deceased during the follow up period

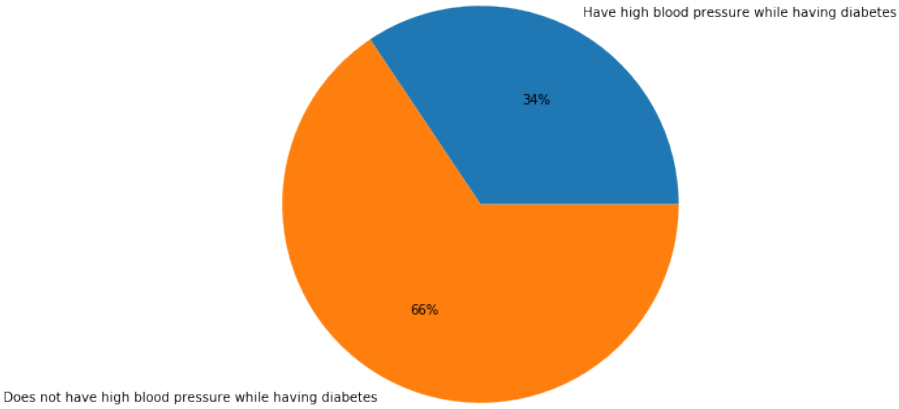


When exploring relationships between pairs of attributes, we used a scatter diagram. The scatter diagram allows us to check for correlation between two features. This helps us locate figure out which features to pay closer attention to when investigating the likeliness of survival. Some pie charts were also used to show the ration of different variables.

6 variables were chosen in order to compound 15 pairs of attributes: Age, Diabetes, High Blood Pressure, smoking, Serum sodium and death event.

Comparing diabetes and high blood pressure, we make a hypothesis that the patients with diabetes will be likely to have high blood pressure. With the total number of people with diabetes were 125 and though analysis we can see that 43(34%) of the patient have high blood pressure while having diabetes and which does not show a very strong positive correlation between these two variables however still related and could be one of the factors which need further investigation.

Figure 2.11 Percentage of patients have high blood pressure while having diabetes

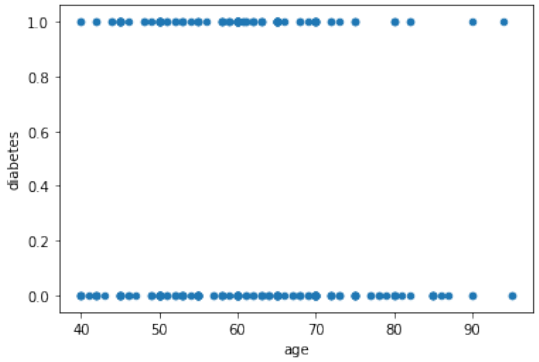


Similar hypothesis was decided for diabetes and death event, we assume that the patients will have higher chance that deceased while having diabetes though analysis Which can see that 32% of the patient have deceased while having diabetes, that also does not show a very strong correlation between these two variables but still have some connection and could be one of the factors which need further investigation.

For the pair Smoking while having diabetes, deceased while having high blood pressure, smoking while having high blood pressure, deceased while smoking. Similar hypothesis was made which each pair of

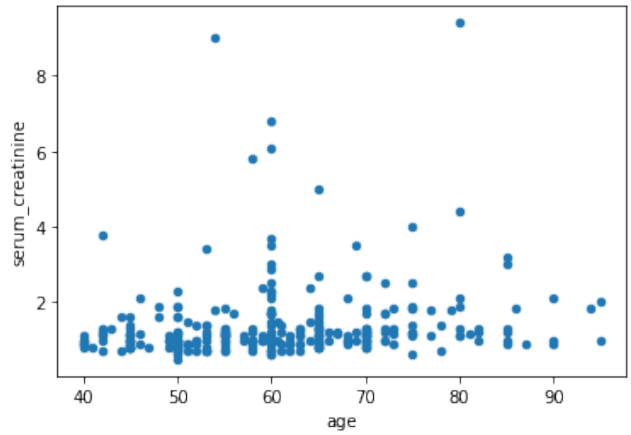
attribute shows a similar behaver and does not show a very strong positive correlation but at the same time can be one of the factors to the according variables. The percentage are listed below:
24% of the patient are smoking while having diabetes.
37% of the patient have deceased while having had high blood pressure.
29% of the patient are smoking while having high blood pressure.
31% of the patient were deceased while smoking.

For comparing age and diabetes, we assume that elder people with age 65 and above will have higher chance to have diabetes. However, 35.7% of the elder patients and 45.7% of the younger patients are having diabetes, which mean the assumption does not have sufficient evidence to support it.



Similar situation was performed for other pair of attributes with serum sodium and pair of attributes with age variables.

For the pair of serum sodium to age variables, there 128 patients not having an ideal range of serum sodium where 44 were age 65 and above, 84 were below age 65. There are total 115 patient age 65 and above and 184 patient below age 65, which though analysis 38.3% of the patient age 65 and above were not having an ideal level of serum sodium on the other hand 45.7% of the patient with age below 65 were not having an ideal level of serum sodium, which might shows that the increase in age could be a factor that less likely to have the problem of having a non-ideal level of serum sodium.



Data Modelling

We choose to use two classification models: Decision Tree and K-Nearest Neighbours from the sklearn library. We will use 70% of the dataset to train the model while the remaining 30% will be used to evaluate the accuracy of our model. As for the K value, I ran it through a for loop to check for the best k value to use, as well as graphing out the corresponding accuracy level for every other k value.

When selecting our features, I have used a correlation matrix and selected the features manually based off the results from the chi-square.

Results

The Accuracy for Decision tree model: 75.556 %

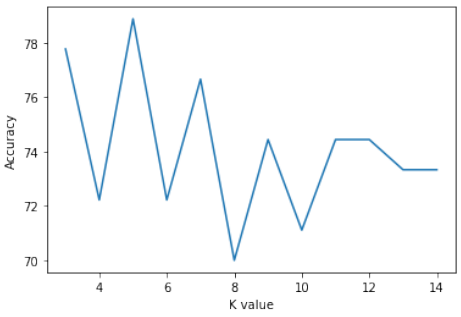
```
# this code calculates for the best k value to use
accuracy = []
best_knn =None
prev_acc = 0
best_k = 0
k_range = [i for i in range(3,15)]
for i in k_range:
  KNN = KNeighborsClassifier(n_neighbors=i)

  # Train the model using the training sets
  KNN.fit(X_train, y_train)

  #Predict Output
  y_pred_knn= KNN.predict(X_test)
  curr_acc = metrics.accuracy_score(y_test, y_pred_knn)*100
  accuracy.append(curr_acc)
  if prev_acc<curr_acc:
    best_knn = KNN
    best_k = i
  prev_acc = curr_acc
print('Best K value to use: ',best_k)
```

the best K value to use which is 11 though above coding

```
In [129]: plt.figure()
          plt.xlabel('K value')
          plt.ylabel('Accuracy')
          plt.plot(k_range, accuracy)
          plt.show()
```



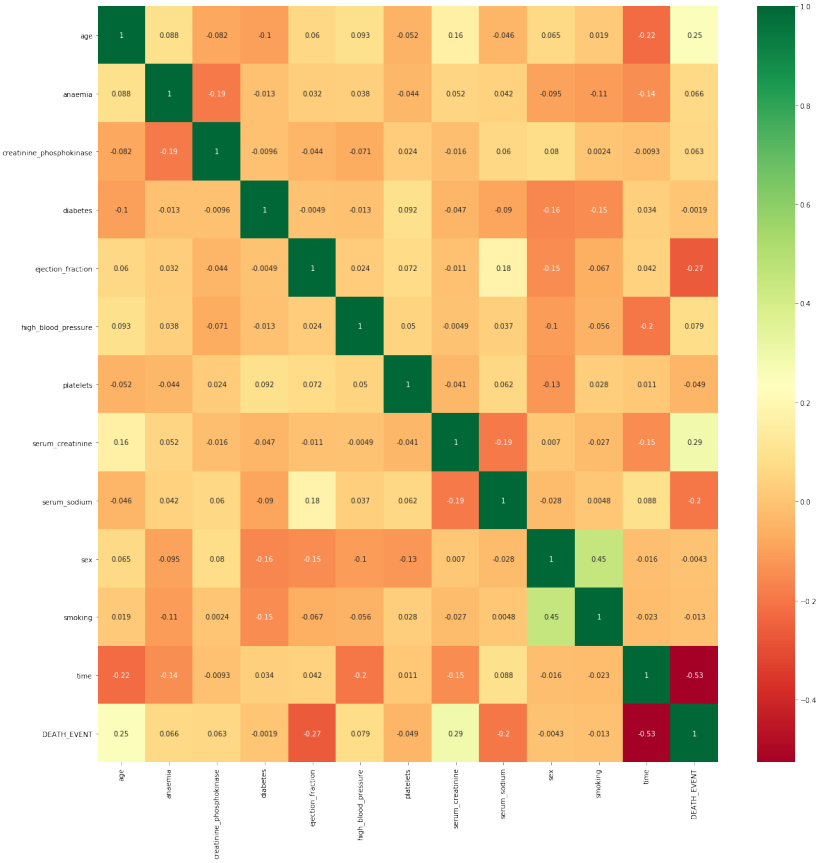
As we can see from the figure, the code above had selected the best k value to use when comparing accuracy between different K values and the corresponding prediction accuracy.

Then the accuracy of KNN model: 74.444 %

Discussion

With the Data Exploration we can see that most of the attribute compared shows a relationship with different variables however doesn't have a very strong positive correlation in order to support the hypothesis that need further information or more investigation.

The Correlation matrix for further feature identification



When observing the last row, we can see that DEATH_EVENT has a considerable correlation (negative or positive) with age, ejection_fraction, serum_creatinine, serum_sodium and time. These are the same features identified from chi-square test however high blood pressure has a relative low correlation value. But which will be kept, other columns will not be used.

The classification models perform with the 70% data from the data set and randomly choose 30% of the database for testing, and the rest will be used for training.

Conclusion

From two classification models Decision Tree and K-Nearest Neighbours from the sklearn library and though the above coding we can see that Decision Tree model have a slightly better accuracy with 75.556% to 74.444%

References

- [1] 'Correlation Matrix in Python - Practical Implementation - AskPython'. https://www.askpython.com/python/examples/correlation-matrix-in-python (accessed May 14, 2021).
- [2] 'Plotting Correlation Matrix using Python', *GeeksforGeeks*, Nov. 23, 2020. https://www.geeksforgeeks.org/plotting-correlation-matrix-using-python/ (accessed May 14, 2021).

- [3] M. Ebrahim, 'Python correlation matrix tutorial', *Like Geeks*, Jun. 17, 2020. https://likegeeks.com/python-correlation-matrix/ (accessed May 14, 2021).
- [4] Assia Munir, Sajjad Haider Bhatti, Muhammad Aftab, Muhammad Ali Raza, and Tanvir Ahmad, 'Heart failure clinical records Data Set', *UCI Machine Learning Repository*. https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records (accessed May 03, 2021).
- [5] D. Chicco and G. Jurman, '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. 1, p. 16, Feb. 2020, doi: 10.1186/s12911-020-1023-5.
- [6] T. Ahmad, A. Munir, S. H. Bhatti, M. Aftab, and M. A. Raza, 'Survival analysis of heart failure patients: A case study', *PLOS ONE*, vol. 12, no. 7, p. e0181001, Jul. 2017, doi: 10.1371/journal.pone.0181001.
- [7]C. L. Balduini and P. Noris, "Platelet count and aging," Haematologica, vol. 99, no. 6, pp. 953–955, 2014.
- [8] M. McKee, S. Exall, D. Stuckler, and A. Wolff, "'Normal' serum sodium concentration among inpatients over 65 admitted to hospital: an observational study," Postgraduate Medical Journal, vol. 92, no. 1083, pp. 21–26, 2015.