Title: Practical Data Science with Python – Assignment 2

Student ID: s3899679

Student Name and email (contact info): Liam Whitaker, s3899679@student.rmit.edu.au

Affiliations: RMIT University.

Date of Report: 25/05/2022

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*.

1. Introduction

Heart failure and similar heart diseases kill nearly 17 million a year [1] and affects millions more. These conditions are often difficult to diagnose early, and can be more-so challenging to treat, as the cardiovascular system (primarily the heart) is a delicate structure. Cardiovascular failure can develop from the result of old age, diabetes, heart attack, damage to the heart muscle, arrhythmia, hypertension, and more [2].
The goal of this investigation is to be able to predict the likelihood of surviving heart failure using statistical and AI modelling of a dataset gathered from hundreds of heart failure patients.

2. Methodology

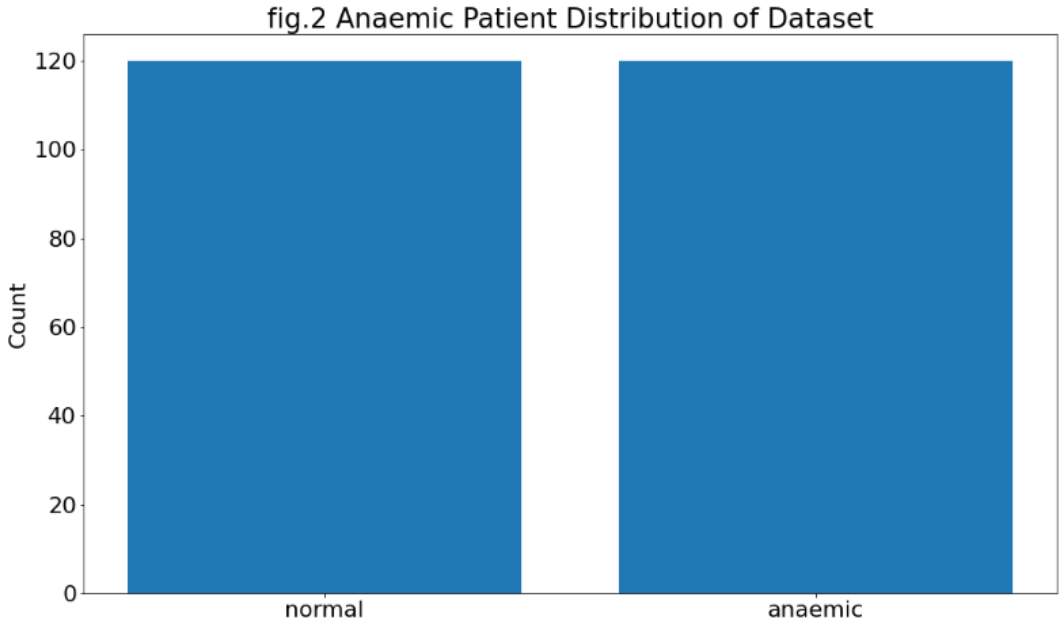
The first steps with data modelling is to ensure the data that is being input is clean and correct. Likewise for each feature of the dataset, the patient's data will be checked for impossible or missing values and repeating or duplicate responses.
From initial assessment of the dataset, it was discovered that the 'age' feature was described as a floating-point value, meaning some age values were input as a decimal. These values were found to be two instances with the age of '60.667' years, which were subsequently rounded down and returned to the dataset.
After this, each integer feature was converted to an appropriate integer size. While this doesn't make a significant difference for a dataset as small as the one given (299 patients) it is good practice to do this for larger datasets in order to improve the memory efficiency of the dataset. This integer optimisation was carried out for the 'age', 'ejection_fraction', 'serum_sodium', and 'time' features.
A discrepancy was found between the feature 'platelets' description (Table 1 of source document [1]) and its actual values. In this table, it was described as 'kiloplatelets/mL', however was given in 'platelets/mL'. The values for this feature were converted and rounded to the nearest 2 digits.
Through further investigation of the data, a major discrepancy was discovered. For the 'platelets' and 'creatinine_phosphokinase' features, there were 34 and 25 instances of a repeating value respectively. For these highly variable features, it is highly improbable that there would be instances of exactly repeated. Unfortunately, these patient observations had to be removed from the dataset in order to preserve the integrity of the outcome in the modelling stage.
After this process was complete, the size of the dataset decreased from 299 to 240.

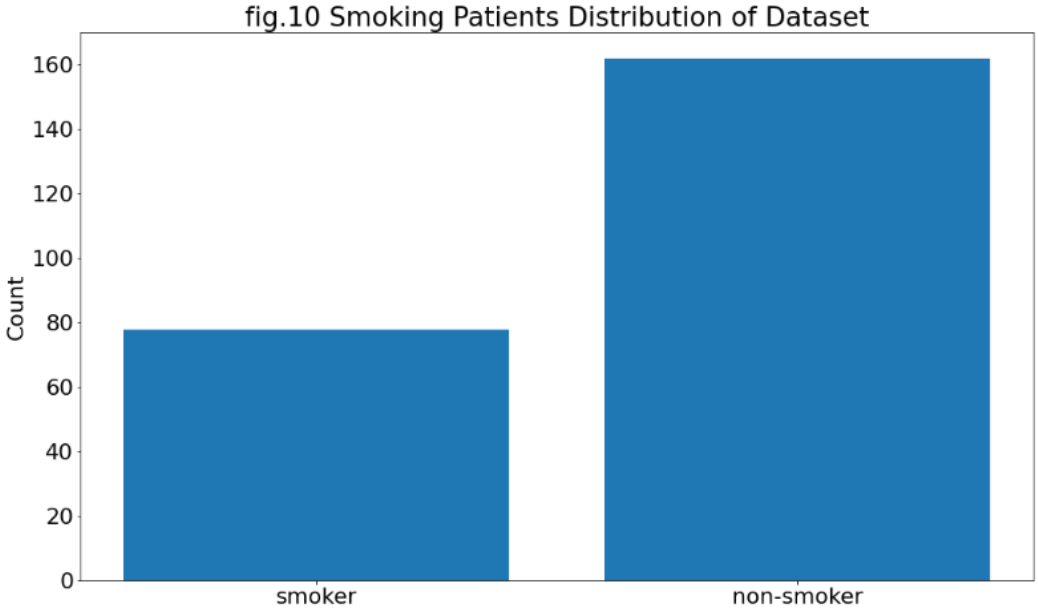
Now that the dataset was clean, some categories were pre-processed to make working with the dataset easier for the exploration process. The major changes made in this section primarily made the responses for the binary features ('smoking', 'diabetes', etc.) more descriptive. For example, the values for 'high_blood_pressure': the values '1' were changed to 'hypertensive' whereas the values '0' were changed to 'normal'. This was carried out for all binary features of the dataset except for 'sex', where the original source of the data never described which value was for 'male' and 'female'. A copy of the DataFrame was made under the label 'df_before_pp' as the numerical values are required for the data modelling process.

Now that the pre-processing was complete, that data could be explored and analysed, then modelled using machine learning. The results to these

3. Results

The first step was to explore the dataset and its many variables. This process begun by outlining the distribution of results for ten primary features of the dataset, where a few interesting facts were obtained. To start, exactly 50% of the patients in the dataset were classified as anaemic (fig.2), where the international average lies around 25-30% [3][4]. This is a significant difference and could be a cause for concern with its relationship with heart failure. Another fact that was discovered is that 33% of the patients in this dataset are smokers (fig.10).





Exploring deeper into the dataset, the relationship between CPK (creatinine phosphokinase) and serum creatinine was investigated. Serum creatinine is found in the blood when the liver breaks down the CPK enzyme, which is released into the blood due to muscle or tissue damage, typically with the heart [1]. So in theory, liver failure as an external factor may contribute to, or be somewhat related to, heart failure in patients.

This relationship was described in figure 11 and its parts, where CPK is graphed against serum creatinine. From this, it was found that CPK and serum creatinine are loosely inversely related (fig.11a), as predicted. This would hypothetically be describing how the body reacts to different levels of CPK in the blood, where a high amount of CPK and a low amount of serum creatinine indicates that muscle tissue in the body has been damaged, however the liver is failing to break down the CPK enzyme, whereas a low amount of CPK and a high amount of serum creatinine indicate similarly that muscle tissue in the body has been damaged, however the liver has broken the CPK enzyme down successfully. Low amounts of both of these values imply that there is little to no tissue damage in the body, and subsequently there is likely to be little cause for concern for the patient's mortality (fig.11b).

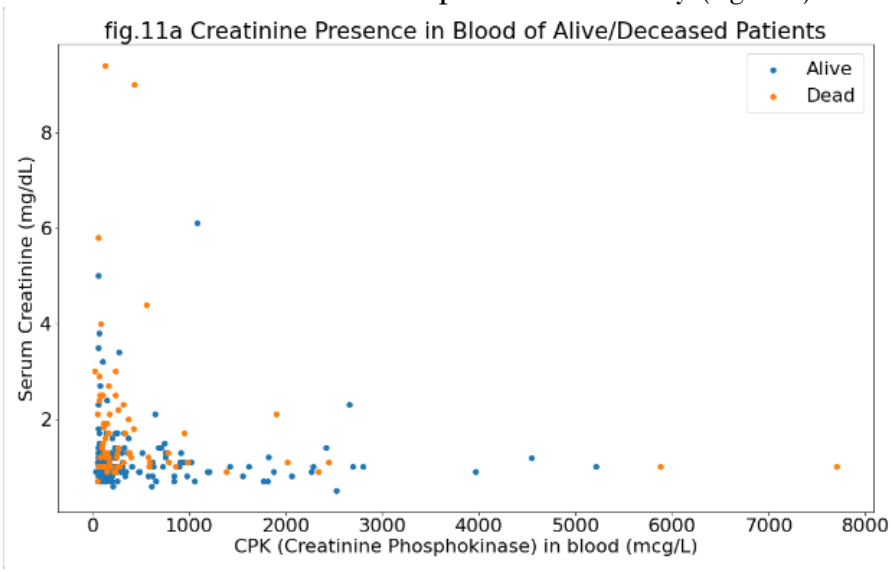


fig.11a Creatinine Presence in Blood of Alive/Deceased Patients

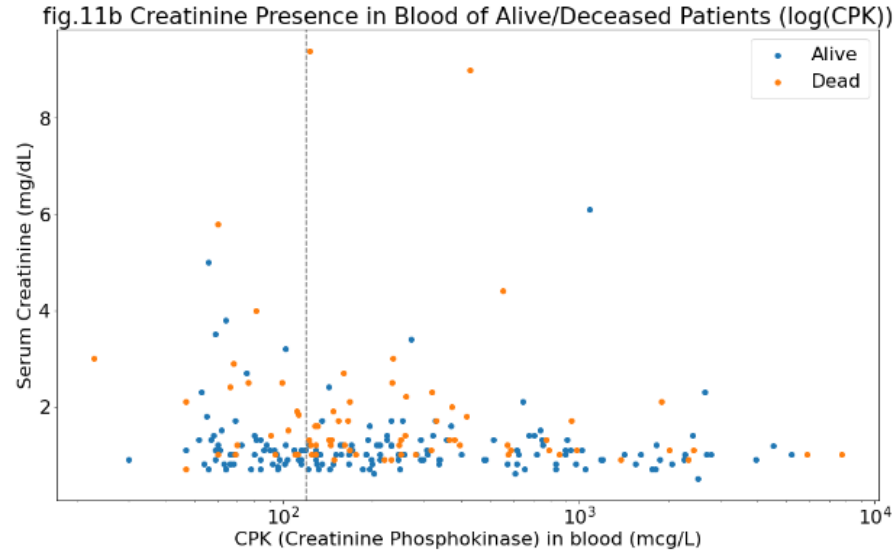
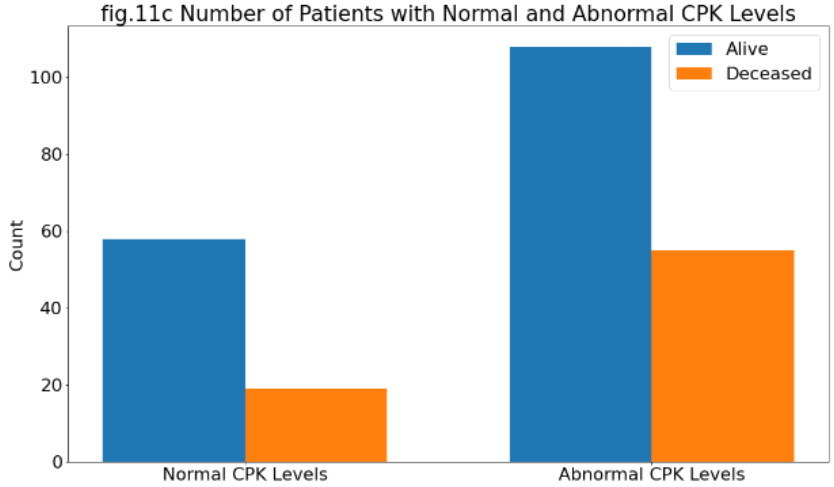


fig.11b Creatinine Presence in Blood of Alive/Deceased Patients (log(CPK))

Looking at this relationship further, there is a significant increase in mortality rate for patients where CPK is above 120mcg/L, which is the normal level [5] (fig.11c). This mortality rate increases from 25% to 34%, which implies a significant relationship between the presence of CPK in the blood and a patient's chance of survival.

fig.11c Number of Patients with Normal and Abnormal CPK Levels



Finally, a nearest neighbour classification model was created to predict the mortality of patients with respect to their records. The model was trained using 90% of the data, and this is due to the small size of the dataset, being only 240 patients long. For a dataset of this size, increasing the training size significantly improves the resulting model. The K-value used for the model was decided to be 13, as a test was conducted on all reasonable values from 1-20, where the primary goal was to more accurately predict true negatives, or correct death guesses.

Resulting adjacency matrix:

```
array([[17,  1],
       [ 2,  4]], dtype=int64)
```

Classification report:

```
: # Classification report
  print(classification_report(y_test, pred))
```

	precision	recall	f1-score	support
alive	0.89	0.94	0.92	18
dead	0.80	0.67	0.73	6
accuracy			0.88	24
macro avg	0.85	0.81	0.82	24
weighted avg	0.87	0.88	0.87	24

4. Discussion

With an accuracy of 88%, this model has been built to an acceptable level, at least for the size of the dataset given. Out of 24 predictions, there were only 2 false negatives

and 1 false positive. This is considered an acceptable distribution, though the results can be improved drastically. While the dataset provided was limited, it was further crippled by the erroneous values as detected in the data cleating step of the process, in which it lost 59 total observations. If this were to be conducted in the future, a larger and more trustworthy dataset is highly recommended, as the process will not only become easier, but will additionally yield more accurate results.

5. Conclusion

Through this data modelling task, the relationship between a given heart failure patient's features and their mortality was somewhat successfully modelled, and a k-nearest-neighbour model was able to predict the likelihood of a patient's death to the degree of 88% accuracy.

6. References

[1] - Chicco, D., Jurman, G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Med Inform Decis Mak 20, 16 (2020). https://doi.org/10.1186/s12911-020-1023-5

[2] - Heart Foundation (2021). Home | The Heart Foundation. [online] www.heartfoundation.org.au. Available at: https://www.heartfoundation.org.au/Conditions/Heart-Failure

[3] - Diagnostics, E. (n.d.). Who has the highest risk of developing anemia? [online] www.ekfdiagnostics.com. Available at: https://www.ekfdiagnostics.com/who-has-the-highest-risk-of-developing-anemia.html

[4] - Cleveland Clinic (2020). Anemia: Symptoms, Types, Causes, Risks, Treatment & Management. [online] Cleveland Clinic. Available at: https://my.clevelandclinic.org/health/diseases/3929-anemia

[5] - Mount Sinai Health System. (n.d.). Creatine phosphokinase test Information | Mount Sinai - New York. [online] Available at: https://www.mountsinai.org/health-library/tests/creatine-phosphokinase-test