Student ID: Thomas Le, Hue Phuong Le

Student Name: s3491287, s3687477

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*.

Predicting patient survival of cardiovascular failure - A classification model

Assignment 2

Author Information: Thomas Le (s3491287), Hue Phuong Le (s3687477)
Student, Practical Data Science (COSC2670)
RMIT University
College of Science, Engineering and Health
Contact details: s3491287@student.rmit.edu.au,
s3687477@student.rmit.edu.au
Date of report: 23/05/2021

Table of Contents

Executive summary

The aim of this report is to predict the heart failure patients' survival as well as detect the most important risk factors that lead to death. Two different modelling techniques, k-nearest neighbors (K-NN) and decision tree classification are both applied and compared to achieve our goal. We also perform feature selection using hill climbing and correlation to effectively build the prediction models. The results of the models are validated via a 80/20 train/test split as well as k-folds cross validation strategy. Serum creatinine, ejection fraction and age are top three features with a coefficient score of 0.29, -0.27 and 0.25 respectively. The models built with them are also more accurate with an accuracy score of 71.67% for K-NN and 83.33% for Decision Tree. The report concludes that it is possible to predict the survival of heart failure patients by age, ejection fraction and serum creatinine solely, however, selecting appropriate models and metrics are important. It is recommended that for the future dataset with heart failure patients clinical record, either decision tree classifier or an alternative should be used. A metric should be selected carefully, especially for the imbalanced dataset.

Introduction

Cardiovascular diseases (CVDs) are the first cause of death globally, including strokes, heart failure, etc. According to the World Health Organization (WHO), approximately 17.9 million people died from CVDs in 2016, representing 31% of all global deaths. Heart failure (HF) is a condition where the heart does not pump blood properly to fulfill the needs of the body. There are four types of HF: left-sided, right-sided, systolic and diastolic HF. Despite improvement of approaches to treat CVDs, HF remains a serious health problem.

This study focuses on the analysis of a UCI dataset called heart failure clinical record containing the medical records of 299 patients with heart failure, collected during their follow-up period, where each patient profile has 13 clinical features. We aim to build a model using machine learning for the prediction of survival in heart failure patients, and then rank the most important features from the medical record.

In particular, we first detail the cleaning of the dataset, how we explored the data, how features were selected and how classifiers were used in the methodology section. The results including visualization with plausible hypotheses, the ranking of features and the performances of two models are then reported. Finally, we discuss the results in the discussion section and draw some conclusions at the end.

Methodology

Data Cleaning

The data was downloaded from the UCI machine learning repository, discarding patient's follow-up time due to no effect on the survival. A common occurrence throughout the dataset, were outliers. It was especially prevalent in variables: creatinine phosphokinase, platelets and serum creatinine. They were all kept as falling in an appropriate range. Only outliers for ejection fraction features with extremely high values were removed.

Other changes were also made so we could work with the data. Inappropriate values for the patient's age were corrected, and the data type was converted to an integer type rather than a float. Death event column's name was also renamed to lowercase letter version to reduce confusion.

Data Exploration

All columns were investigated when exploring each column. As each feature is an indispensable factor from a patient's medical record, it may have a responsibility regarding heart failure prediction. Hence, discovering all of them would benefit when addressing plausible hypotheses.

As our model is trying to predict the survival of patients, each numerical and categorical variable is compared to death event when exploring pairs of columns:

1. The first pair visualised was the age distribution against death events. This was a logical place to start as it was expected that as one grew older they were more susceptible to heart events and death, hence contributing to the case that age has a strong correlation with the survival of patients.
2. Creatinine phosphokinase (CPK) levels are elevated when there is muscle injury or stress, such as heart attacks. The rationale here was that high levels of CPK would signify death events.
3. The next variable was ejection fraction - the percentage of blood that is pumped through the left ventricle on each beating phase. The plausible hypothesis here is that less ejection fraction is due to heart complications or failure which leads to death.
4. There was not any kind of relationship that could be seen for platelets against death-event. Possible hypothesis was that low platelet levels would mean excess bleeding.
5. Serum creatinine against death where chosen as low levels of serum creatinine suggests possible kidney disease or low blood volume going through the kidneys, which could be a sign of heart failure, hence adding the count to the death toll.
6. Serum sodium and death-event were chosen as possible low levels or high levels of serum sodium could mean adverse health issues related to cardiovascular disease and cause death.
7. Gender was the first categorical variable visualised for its distribution and the proportion of deaths occurred in this population. A plausible hypothesis for this was that males were more likely to die from a heart event.
8. The pie chart for diabetes was generated next, the rationale was that those with diabetes were less likely to survive a heart event due to having another medical condition, which could affect their chances of surviving.
9. Anaemia was also compared to death-event for similar reasons to diabetes.
10. High blood pressure was the next plot pair. The hypothesis here was that high blood pressure could lead to heart damage or attacks, leading to the contribution of death-events.
11. Lastly, smoking was well known that it adversely affected people's health, such as raising blood pressure. It was assumed that smokers were more likely to die than non-smokers.
12. Pair plots of age against diabetes, anaemia, high blood pressure and smoking were graphed to visualise their distribution and to see if there were any hypotheses that could be derived from them.

Feature Selection

Unnecessary and redundant features not only affect the performance but also the accuracy of the algorithm. As a result, selecting the most suitable features for training the model is vital. There are several ways to accomplish this, however, in this study, we will focus on the hill climbing technique as well as correlation.

1. **Hill climbing**: This technique is based on the accuracy score of the model chosen. It involves starting with a subset of one candidate feature and continuously increasing the number of features in the subset if the new subset with an added feature gives a higher score compared to the previous iteration. In case a new subset returns a lower score, a particular feature will be removed. At the end, it has effectively returned an optimal subset of features – a 'peak'. We performed this technique on the dataset with two different models to observe the appropriate set of features for each model.
2. **Correlation**: The corr() method of the pandas dataframe was used to generate a correlation matrix between all the columns of the dataset. Then, we obtained the correlation coefficients between two columns including the target column. A value of exactly 1.0 means there is a perfect positive relationship between two columns, while -1.0 means perfect negative and 0 for no linear relationship. Those scores were compared and checked with the graph visualization to produce a feature ranking.

Data Modelling

Our models were learnt from three sets of features: the whole features, features returned from hill climbing optimisation and top three important features obtained from correlation and visualization.

They were trained and evaluated using two different methods. The first method was using the 80/20 split on

the dataset where a pseudorandom 80% of the data (237 patients) was used for training and another 20% (60 patients) for testing the performance of the model. The second method was the k-folds cross-validation technique. The data was divided into five folds and each part was used once as a test set with the rest as a training set. This method gives us a more stable and comprehensive evaluation of the model as all available data is used for training and testing.

Furthermore, the shuffle was used to get a balanced distribution of data between the training and testing set(s) as well as list of features. Data was randomised and placed in no order to prevent any bias during training the model, otherwise, a deterministic order would give an extremely low accuracy.

The K-Nearest Neighbors Classification Algorithm

K-Nearest Neighbors (K-NN) algorithm is a supervised non-parametric classification technique that was used first to model the cleaned dataset. K-NN will output a class of a category depending on the most common class among its K nearest neighbour. This method is appropriate due to the binary nature of the target classification; whether a patient had survived a heart event (died or not died).

The first step was to find an optimal number used for *k* in the K-NN classification algorithm to fit the training data. In general, large values of *k* reduces noise but makes boundaries between classes less discernible.

K was fine-tuned through a function that returned a minimal error rate so that an optimal value of *k* was chosen. Typically, for binary classifications it is helpful to choose an odd value for *k* as it avoids tied 'votes' for its neighbours, however, *k* was picked on the bases of minimal error in its prediction.

The next parameter used for the algorithm was **'weights'**, this is used to overcome any overfitting that may occur when the data is skewed towards one classification. The 'distance' option was used as it made sense for binary classification, where closer neighbours to the point of object will have a greater influence than neighbours which are further away.

The final parameter chosen was the power parameter **'p'**, which was specified with a value of 2. As we are using a distanced weight for this classification, p=2 will be using the Euclidean metric to measure distance: *sqrt (sum ((x - y) ^2))* for continuous variables.

After the parameters were tuned, the algorithm was used to model training data of three different feature scenarios.

The Decision Tree Classification Algorithm

The decision tree classifier is more challenging to fine-tune than the K-NN algorithm. The reason for this is the number of parameters required is quite large and the range of possible values for them is expansive. There are also various rules to configure the tree, and in this report, we consider the stopping rule with three hyperparameters which are criterion, min_samples_split, min_samples_leaf.

Technically, if no stopping criteria are set, the nodes will be splitted until all leaves are pure and hence, the tree will be complex. The deeper we allow the tree to grow, the more chances overfitting will happen. Being noticed from the result acquired from a vanilla tree, overfitting occurs due to one or two samples in each leaf.

- **criterion**: "gini" index was chosen because it is faster to compute compared to entropy and both of them do not make much difference in the tree performance.

- **min_samples_split** and **min_samples_leaf**: When min_samples_split prevents a node from expanding if it contains less than the quantity of samples given, min_samples_leaf guarantees a minimum number of samples required in a leaf node of the tree. A function using k-folds cross validation was created to tune them. It tries all possible combinations of those two parameters and returns the combination with the best F1 score.

To avoid misleading, F1 score was chosen instead of the traditional classification accuracy. As this is an imbalanced class problem, the model can decide the best thing to do is to always predict the patient to be survived (death event = 0) and achieve a higher accuracy score. F1 score is useful when seeking a balance in an uneven class distribution. The second choice was made for the range of parameters' value. Because the regions for the minority class will be very small, lower values should be chosen. Therefore, the range was predefined to be between one and two to 20 for min_smples_leaf and min_samples_split respectively.

The tuning function was used before modelling data with each set of features. As there are three sets of features, the hyperparameters were tuned three times. The appropriate results were obtained and passed onto a decision tree classifier.

Results

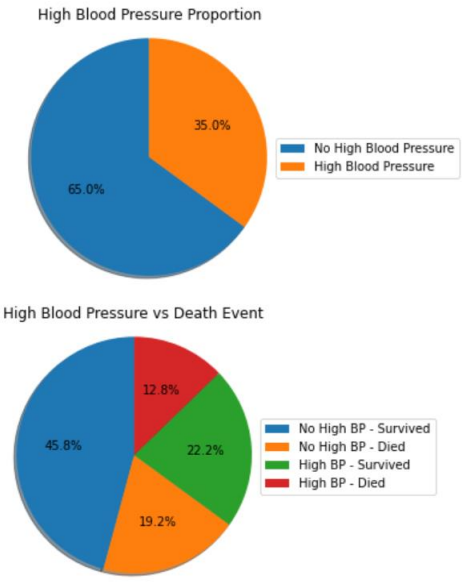
Visualization

More than 10 different pairs of columns were explored and in this report, those with the most interesting results are listed. There are five hypotheses:

Hypothesis #1: High blood pressure (HBP), smoking and diabetes has no significant distribution to predict the survival of patients having heart failure.

This hypothesis was explored by a visualization to observe the relationship between the death event and HBP, smoking and diabetes respectively. Pie chart was chosen because they are binary variables. The figure below describes the relationship between the HBP and death event:

Figure 23: High Blood Pressure vs Death Event



We assumed that patients having cardiovascular disease with HBP are more likely to die. However, the figure shows that there are 35% of patients with HBP but only 12.8% died.

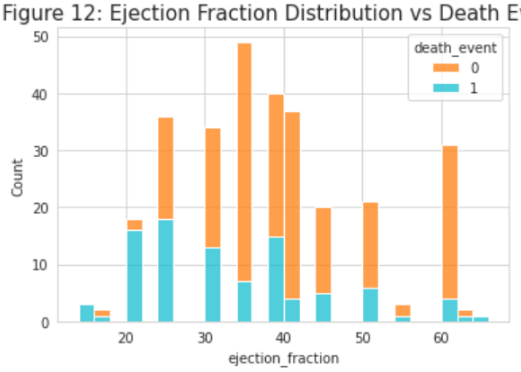
Similarly for patients having diabetes and smoking, there are 42.1% of patients being diabetic but only a third of them were dead (13.5%). For smoking, only 10.1% of smokers died out of 32.3%.

Hypothesis #2: Heart failure patients with lower ejection fraction percentages face a higher risk of death.

A histogram was generated as part of exploration showing the distribution of ejection fraction by dead and

survived patients. The relationship is shown in the figure belows:

Figure 12: Ejection Fraction Distribution vs Death Event

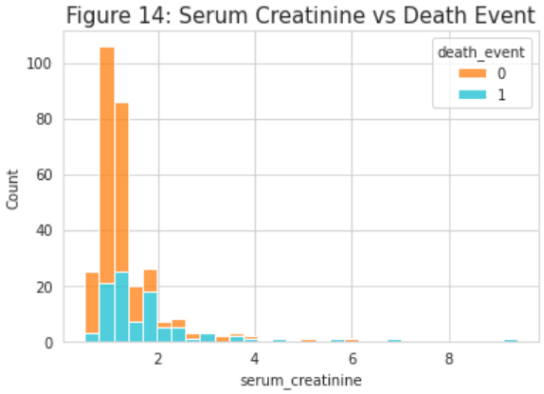


As we expected, most of the patients in the dataset have a reduced ejection fraction of smaller than 40% and those with less than 30% tend to die. In contrast, patients with a normal ejection fraction of more than 50% are more likely to survive.

Hypothesis #3: An increase in serum creatinine level causes death in cardiovascular heart disease patients.

The histogram chart belows shows the distribution of serum creatinine, categorized by the survival status of patients (0 for survived and 1 for dead patients):

Figure 14: Serum Creatinine vs Death Event

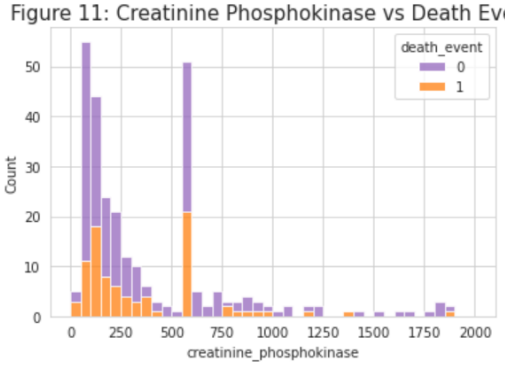


It is clearly seen that there is a strong correlation between the serum creatinine and death event. At an elevated serum creatinine level of greater than 1.4 mg/dL, the proportion of deceased patients increases. As serum creatinine level decreases, patients tend to survive.

Hypothesis #4: Whether or not creatinine phosphokinase (CPK) is an important factor associated with the survival prediction.

The graph belows display the distribution of CPK, labelled by survival status:

Figure 11: Creatinine Phosphokinase vs Death Event

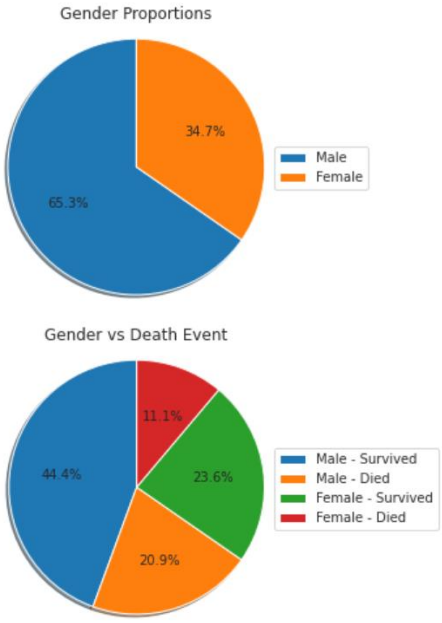


It was assumed that high levels of CPK in the blood and heart attack at the same time would closely be associated with deaths. Nonetheless, deaths occurred within normal CPK levels (24-200 mcg/L), this is probably due to the considerable quantity of patients with normal CPK from the dataset.

Hypothesis #5: Gender distribution is not equal, indicating that sex as a predictor of the survival may cause misleading.

The pie chart belows shows the gender distribution by the survival status:

Figure 20: Gender vs Death Event



There is an imbalanced gender ratio of male and female in the dataset (65.3% to 34.7%). As a result, the percentage of male patients who died is double compared to female ones (20.9% to 11.1%). Generally, it remains still unclear whether men and women have similar survival or not.

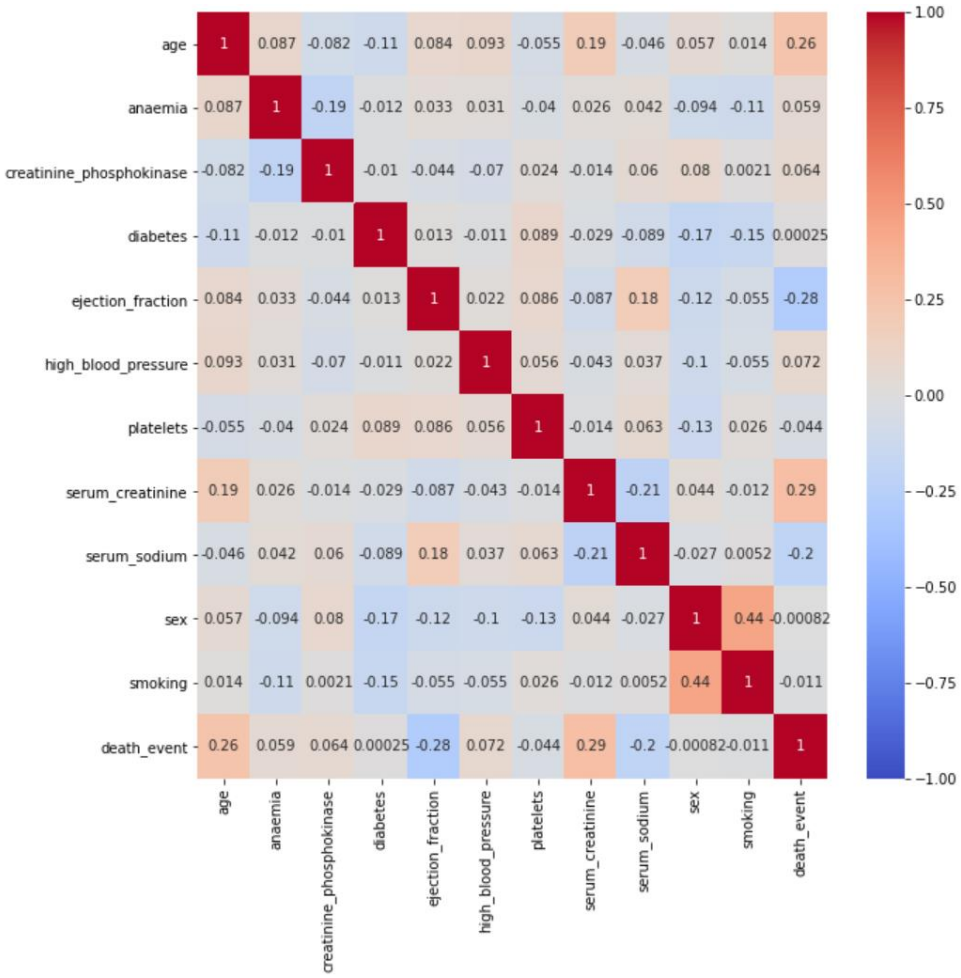
Feature Selection

The hill climbing using K-NN returned two features: age and ejection fraction.

The hill climbing using a vanilla decision tree returned five features: creatinine phophoskinase, diabetes, serum creatinine, sex and smoking.

As obtained from visualization, three important features are selected which are age, ejection fraction and

serum creatinine. Comparing with a correlation matrix, it identified that serum creatinine, ejection fraction and age with a coefficient score of 0.29, -0.27 and 0.25 respectively are top three clinical features. A correlation matrix of all columns is observed belows:



K-NN Classification

K-NN Classification on all features

The accuracy for K-NN model with all features in the dataset on an 80/20 train/test split were as follows:

	precision	recall	f1-score	support
0.0	0.73	1.00	0.85	44
1.0	0.00	0.00	0.00	16
accuracy			0.73	60
macro avg	0.37	0.50	0.42	60
weighted avg	0.54	0.73	0.62	60

After k-folds cross-validation, these were the resulted accuracy scores:

```
[fold 0] score: 0.75000
[fold 1] score: 0.81667
[fold 2] score: 0.72881
[fold 3] score: 0.71186
[fold 4] score: 0.35593
```

Feature selection with Hill Climbing

The following accuracy scores were based on feature selection with simple hill climbing:

	precision	recall	f1-score	support
0.0	0.68	0.94	0.79	36
1.0	0.80	0.33	0.47	24
accuracy			0.70	60
macro avg	0.74	0.64	0.63	60
weighted avg	0.73	0.70	0.66	60

K-NN Classification on age, ejection fraction and serum creatinine

The classification report for the K-NN model for serum creatinine, ejection fraction and age on an 80/20 split:

	precision	recall	f1-score	support
0.0	0.69	0.94	0.80	36
1.0	0.82	0.38	0.51	24
accuracy			0.72	60
macro avg	0.76	0.66	0.66	60
weighted avg	0.74	0.72	0.69	60

The accuracy scores after k-folds cross-validation was used for the top three features:

```
[fold 0] score:  0.71667
[fold 1] score:  0.68333
[fold 2] score:  0.71186
[fold 3] score:  0.71186
[fold 4] score:  0.74576

Average score is: 0.7138983050847457
```

It appears that in all three models, K-NN had a higher accuracy in predicting patients that had survived a heart event. This is most likely due to the distribution of patients being mostly alive (68%), therefore, for any value of *k*, the object point will be more likely to be surrounded by neighbours of surviving patients.

Consequently, this led to a poorer prediction in true positives; patients that died. This is highlighted in the first model, where all features were used. It had a recall of 1.0, predicting all the true negatives correctly, but was unable to predict any true positives with a recall score of 0.0. This model also had the highest accuracy score of 0.73 due to its higher precision in predicting all true negatives correctly.

The K-NN model for the top three features were able to produce a more balanced prediction between surviving patients and dead patients, with an accuracy score slightly lower than the ones with all features; 0.72. It was able to predict 38% of true positives correctly, compared to 0% with all features.

A more well-balanced classification may be the better model with the top three features producing a superior macro and weighted average (0.76, 0.74 vs 0.37, 0.54) even with a slightly lower accuracy score than the

model with all features (0.72 vs 0.73).

Decision Tree Classification

Decision Tree Classification on all features

A classification report obtained from training the data with all features in a 80/20 train/test split is as follows:

	precision	recall	f1-score	support
0	0.67	0.78	0.72	36
1	0.56	0.42	0.48	24
accuracy			0.63	60
macro avg	0.61	0.60	0.60	60
weighted avg	0.62	0.63	0.62	60

After k-folds cross-validation, the following F1 scores were reached for each of the 5 folds:

```
[fold 0] score:  0.62500
[fold 1] score:  0.72727
[fold 2] score:  0.50000
[fold 3] score:  0.56410
[fold 4] score:  0.61111

Average F1 score for 5 folds: 0.6054972804972805
```

Decision Tree Classification on features returned from Hill Climbing feature selection

A classification report obtained from training the data with features from Hill Climbing feature selection in a 80/20 train/test split is as follows:

	precision	recall	f1-score	support
0	0.76	0.66	0.71	44
1	0.32	0.44	0.37	16
accuracy			0.60	60
macro avg	0.54	0.55	0.54	60
weighted avg	0.64	0.60	0.62	60

After k-folds cross-validation, the following F1 scores were reached for each of the 5 folds:

```
[fold 0] score:  0.48276
[fold 1] score:  0.42105
[fold 2] score:  0.45161
[fold 3] score:  0.47619
[fold 4] score:  0.60000

Average F1 score for 5 folds: 0.4863229263369771
```

Decision Tree Classification on age, ejection fraction and serum creatinine

A classification report obtained from training the data with age, ejection fraction and serum creatinine solely in a 80/20 train/test split is as follows:

	precision	recall	f1-score	support
0	0.89	0.89	0.89	44
1	0.69	0.69	0.69	16
accuracy			0.83	60
macro avg	0.79	0.79	0.79	60
weighted avg	0.83	0.83	0.83	60

After k-folds cross-validation, the following F1 scores were reached for each of the 5 folds:

```
[fold 0] score:  0.53846
[fold 1] score:  0.74286
[fold 2] score:  0.63415
[fold 3] score:  0.57895
[fold 4] score:  0.62857

Average F1 score for 5 folds: 0.6245967639549155
```

Due to the imbalance of the dataset (32% of patients died and 68% survived), this model on all sets of features obtained better results for the true negative, rather than the true positive. As explained above, this is the reason why F1 score was used instead of the accuracy metric.

The prediction score for dead patients from three key features is 0.69, which is much higher compared to all features (0.42) and features from hill climbing (0.44). In terms of surviving patients, the score obtained is 0.89, also better than other two sets of features. Unsurprisingly, although the scores vary across 5 different folds, the average F1 score for top three features is still higher compared to others.

Overall, the performance of Decision Tree classifier on age, ejection fraction and serum creatinine solely outperformed for both 80/20 train/test split and k-folds cross-validation.

Discussion

Pair plots during data exploration

1. **Age vs Death:** It appeared that the proportion of deaths increased as age increased, however, there were more frequent deaths between 60 and 70 rather than 70 and onwards.
2. **CPK vs Death:** The majority of patients that had died had relatively normal levels of CPK (24-200 mcg/L).
3. **Ejection Fraction vs Death**: The graph visualised does indeed show a trend of higher deaths with ejection fractions lower than 40%.
4. **Platelets vs Death:** The graph appeared to be normally distributed with deaths occurring at normal levels of platelets.
5. **Serum Creatinine vs Death:** Serum creatinine against death event showed a strong right skewed relationship. With levels lower than 2 mg/dL showing an increased number of deaths.
6. **Serum Sodium vs Death:** Distribution for serum sodium was left skewed with deaths occurring at normal levels of 135 to 145 mEq/L. Values below 135 could possibly suggest death due to hyponatremia due to renal or heart failure. No special relationship between the two variables could be seen from this visualisation.
7. **Gender vs Death:** The population were about two-thirds male and one third female. The pie chart showed that male died more than females because of how disproportional the ratio between male and female were.
8. **Diabetes vs Death:** The chart showed that those with diabetes died less than those without diabetes, suggesting no correlation between diabetes and death-event.
9. **Anaemia vs Death:** Patients with anaemia also showed that they died less than those without anaemia, implying little association between the two variables.
10. **High Blood Pressure vs Death:** Only 12.8% of patients with high BP died. This may be due to only

one-third of the population having high BP.

11. **Smoking vs Death:** The generated pie chart showed that two-thirds of patients did not smoke and that there was a higher percentage of deaths in non-smokers, implying small associations between smoking and death-event.
12. **Age vs Diabetes, Anaemia, High Blood Pressure and Smoking**: There were no significant relationships that could be seen for these feature pairs that could improve predictions of patient survival.

Feature Selection

From the hypotheses and correlation, it stated age, serum creatinine and ejection fraction as top three key features. Hill climbing for K-NN also identified ejection fraction and age as top features with highest accuracy score. Surprisingly, results obtained from hill climbing for the decision tree differ. Creatinine phosphokinase (CPK), diabetes, sex, serum creatinine and smoking were selected, however, from the correlation, they were lower-ranking.

Data Modelling

As we concluded that models running on top three features returned more accurate results, in this section, we will compare two models based on their results from that features set only. We see that decision tree outperformed K-NN and produced more balanced results on the test set as it performs better with no overfitting. The prediction for the true positives is more accurate with a recall score of 68.75% (K-NN = 37.5%). Although K-NN retuned a higher recall score on the true negatives (94.44%) than decision tree (88.86%), decision tree still obtained a higher accuracy score of 83.33% (K-NN = 71.67%) and F1 score of 83% (K-NN = 69%).

Recommendation

It is noted that the dataset contained a small number of observations (297 samples) with uneven distribution. Therefore, more data should be added and different models (Logistic Regression, Random Forests, etc.) with appropriate metrics should be carefully employed as well to improve the learning process of the model.

Conclusion

Through this project, we had the opportunity to overcome difficult challenges to create valuable insights for the patients. As a result of our exploration and modelling, age, ejection fraction and serum creatinine are the most important risk factors regarding heart failure patients' survival. And the decision tree classifier with a recall or F1 score is much more effective to predict survival. As a limitation of this study, we were able to report the results from only two classification models. Employing more methods from different machine learning areas will help to identify the best learner and obtain more reliable results. Moreover, if the larger dataset with more dead patients' medical record is available, it will be useful to verify our findings. Regarding future developments, we plan to test the dataset with more machine learning models and apply our approach to alternative datasets of other illnesses.

References

Archive.ics.uci.edu. 2020. UCI Machine Learning Repository: Heart failure clinical records Data Set. [online] Available at: <https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records>.

WHO. Cardiovascular diseases: Key Facts; 2017. Available from: <http://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)>.