PREDICTING DEATH BY HEART FAILURE

Matthew James Kingsbury

Abstract

Heart failure and other related cardiovascular diseases (CVDs) are leading causes of death for individuals across the globe. According to the World Health Organisation, an estimated 17.9 million people died from CVDs in 2019, representing 32% of all deaths across the globe.(1) In the UK, detailed in a 2024 article by the British Heart Foundation, the number of individuals dying of heart failure before the age of 75 has reached an all-time high.(2) It is imperative that we take immediate societal action towards tackling the ever-growing heart disease epidemic, and thus this paper looks to leverage and analyse data related to signs of impending heart failure.

This analysis reads in data from a public heart failure predictor dataset, and trains a Logistic Regression model and Support Vector Machines (SVMs) to analyse the relationships in the data and highlight key predictors that may lead to an elevated risk of heart failure. This research performs feature engineering, principle component analysis, and trains three different SVM model kernels - linear, polynomial and radial basis function - to capture the exact relationships that underlie the feature and target variables. Other methodologies such as confusion matrices, classification reports and accuracy scores are used to evaluate each models' effectiveness at classifying the dataset.

Key words. Heart Failure, Logistic Regression, Support Vector Machines, Classification

1. INTRODUCTION

The aim of the project is to effectively predict whether a heart failure death event has occurred given a variety of features describing the condition of an individual. A heart failure death event is simply a binary value which denotes whether the individual in the dataset died due to heart failure.

Heart failure can be caused by a variety of different internal and external factors, however for the scope of this analysis the dataset focuses on measurable medical signs describing aspects of a person's physical health. Some features such as anaemia and high blood pressure are given as binary boolean values which simply denote if the individual has been diagnosed with the condition. Other features are continuous, such as the individual's ejection fraction reading which denotes the percentage of the blood volume in your heart which is pumped out per heartbeat. Finally, some binary variables such as sex and smoking look to capture additional variables that may contribute to the probability of heart failure.

I hypothesise that with a varied set of features, the model will be able to capture a more holistic idea of what variables do or do not correlate to an elevated risk of heart failure. It is important to note that while these variables may be correlated to heart failure, they may not necessarily be a cause of heart failure. Either way, this model could help support further analyses in the medical field with applications in coronary condition diagnosis.

Not all features in the dataset directly link to heart failure on the surface. This analysis looks to see if other bodily conditions such as diabetes may also correlate with an elevated risk of heart failure. This understanding may be useful to ensure that individuals with certain conditions are able to receive necessary treatment if they are found to have an elevated risk.

The base dataset contains 299 entries over 13 features including the target variable. There are 194 men and 105 women analysed, with ages ranging from 40 to 95 years of age, including a mixture of smokers, diabetics, anaemics and individuals with high blood pressure.

2. OBJECTIVES

Successful completion of this analysis involves training two separate machine learning models on the dataset, both of which will perform classification against the death event target variable. From the logistic regression model, capturing the co-efficients to see how each feature either positively or negatively correlates to the predictions, and by what magnitude they influence the model. Evaluation of the model is key, including analysing a confusion matrix and its accuracy score to understand the strengths and shortcomings of the model. I am going to perform Principle Component Analysis (PCA) on the data for dimensionality reduction, before training three separate SVM models with different kernel types to try and analyse the degree of the relationship between the feature and target variables. These models will then also be analysed for their accuracy, before I will evaluate this project as a whole to understand which processes provided valuable insight, which areas could be improved, and how this research can be extended upon and utilised in real-world scenarios.

3. LITERATURE REVIEW

When assessing the validity of the project scope, it is important to look at existing research in the field. The project definition was highly influenced by the research paper titled "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone" written by *Davide Chicco and Giuseppe Jurman.*(3) They analysed a 2015 dataset to find the most prominent feature risk factors and found that the serum creatinine level and ejection fraction level were the most important. They developed multiple models to predict purely on these two features. I decided to utilise the same dataset(4) to perform my analysis, instead maintaining a holistic predictive approach to see how each feature can provide relevant information to the predictions, including engineering new features from the existing data to capture hidden relationships. This dataset falls under the Attribution 4.0 International license.(5)

One interesting formula that can be captured through the existing features in the dataset is the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) equation, which looks to estimate glomerular filtration rate (GFR), a measure very closely linked to correct kidney function as it essentially measures the rate of blood being filtered by the kidneys. A significant 2009 paper "A New Equation to Estimate Glomerular Filtration Rate" by *Andrew S. Levey et al.*(6) outlined the effectiveness of the equation at estimating the GFR of any given individual based on their age, sex and serum creatinine level.

In a modern implementation with complete information, the CKD-EPI equation would be applied differently depending on the racial demographic of the individual, however we do not have racial descriptions in this dataset. The naive implementation equation I decided on utilised the age, sex and creatinine features:

$$GFR = \frac{140 - a}{c * s}$$

In this equation, a represents the age of the individual in years, c represents the serum creatinine level in mg/dL, and s is the 'sex factor' which is 0.9 for men and 1.2 for women. In clinical practice, this generalisation without a race variable may cause issues due a lack of exact accuracy however for the scope of this analysis adding this feature to the dataset to capture a scientifically-supported relationship across the data may provide insight on whether a correlation exists between GFR and heart failure.

There have been various investigations into the application of machine learning technology, analysing various features to understand their predictive ability. In "Improving Heart Disease Prediction Using Feature Selection Approaches" by *Saba Bashir et al.*, the researchers develop a variety of different models over the features given in the UCI Heart Disease dataset to try and predict heart disease diagnoses.(7) There is some crossover in the feature set, including age and sex. The paper states that both Logistic Regression and Logistic Regression SVM both achieve very high accuracy on the feature-set (82.56% and 84.85% respectively) demonstrating their capability at predicting heart disease diagnoses. I decided to utilise these models to try and predict the heart failure death events on the dataset used in this analysis.

Another investigation "Heart failure disease prediction and stratification with temporal electronic health records data using patient representation" by *Ye Liang and Chonghui Guo* analyses a very large set of temporal features to see if networks can be applied to enable immediate prevention and treatment for individuals deamed at risk, as well as sway policy at scale regarding head disease treatment processes.(8) Their comprehensive analysis found that age was a very strong predictor of heart disease characteristics, sodium was a strong predictor of heart disease characteristics, and creatinine was a moderate predictor of heart disease across three different phenotypes. These phenotypes were discovered by implementing a K-Means algorithm on all individuals who were positively diagnosed with heart disease. This may well imply that there exists inherent clustering within the dataset used for this analysis, given some feature crossover. However, clustering is generally outside the scope of this research.

4. DATA PROCESSING

The dataset(4) was freely sourced from Kaggle as a CSV file with 14 features, however I decided to drop the 'time' column as it bares no relevance to the scope of this research, as I am purely looking to find correlations between bodily variables and the risk of heart failure. The included features are:

- Age (Continuous, Integer)
- Anaemia (Binary)
- Creatinine Phosphokinase (Continuous, Integer)
- Diabetes (Binary)
- Ejection Fraction (Continuous, Integer)
- High Blood Pressure (Binary)
- Platelets (Continuous, Integer)
- Serum Creatinine (Continuous, Float)
- Serum Sodium (Continuous, Integer)
- Sex (Binary)
- Smoking (Binary)
- Death Event (Binary, Target Variable)

A sample of the dataset displaying the first 8 features is shown in the image below

	age	anaemia	cph	diabetes	ejf	hbp	platelets	creatinine
0	75.0	0	582	0	20	1	265000	1.9
1	55.0	0	7861	0	38	0	263358	1.1
2	65.0	0	146	0	20	0	162000	1.3
3	50.0	1	111	0	20	0	210000	1.9
4	65.0	1	160	1	20	0	327000	2.7

Every entry has the complete set of values, and there were no duplicate entries, thus no extensive data cleanup was required. As previously mentioned, I made the decision to engineer some additional features to capture relationships between variables across the dataset, being mindful of the potential effects of multicollinearity:

- Sodium-Creatinine Ratio (Continuous, Float)
- GFR (Continuous, Float)
- Ejection Fraction * Age (Continuous, Float)
- GFR * SCR (Continuous, Float)

Pandas allows the application of a custom function across each entry in a dataset to create a new feature, which I utilised for implementing the GFR equation outlined previously

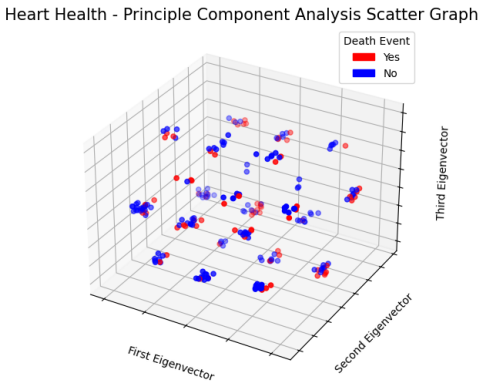
```
# Serum Sodium to Serum Creatinine Ratio
data['scr'] = data['sodium'] / data['creatinine']

# Estimate for CKD-EPI Glomerular Filtration Rate
def pseudoGFR(d):
    sex_factor = 0.9 if d['sex'] == 1 else 1.2
    return (140 - d['age']) / (d['creatinine'] * sex_factor)

data['gfr'] = data.apply(pseudoGFR, axis = 1)

# Interaction Terms trying to catch Relationships in Data
data['ejf_age'] = data['age'] * data['ejf']
data['gfr_scr'] = data['gfr'] * data['scr']
```

These additional features cover some important ratios and medical calculations, as well as acting as interaction terms to check if the relationship between the target variable and any given feature is affected by a second feature. As creatinine, sodium and age were important features in the studies analysed previously, I decided to focus on them particularly when engineering these additional features. As none of the features are categorical, one-hot encoding or other similar methods were not required for this analysis.

To ensure cross-validation, I split that dataset into a training set and a test set. The test set encompassed 20% of the total entries, and the entries were selected at random with the random seed 16. The model would only 'see' the test set at testing time, as the test data plays no role in the training process. This is to ensure that the outcomes of the model are generalisable to never before seen data.

It is important to apply normalisation when handling these features so that each features has the opportunity to contribute equally to the outcome of the model. Each feature was Min-Max scaled for the range 0 to 1 in preparation for model training.

However, it is important to note that the dataset size compared to the number of features is quite small. There could be concerns for severe underfitting is the model is unable to utilise enough data to effectively find the relationships in the data. I decided to apply PCA dimensionality reduction to the dataset and plot the outcome to see if there are any particular clusters that may visibly stand-out frm the plot.



The PCA Scatter Graph does not clearly demonstrate any particular groupings amongst the first three eigenvectors in regards to the target variable. This cemented my beliefs that the scope of this project would be better suited to a classification task over any form a cluster-based, unsupervised learning.

I took a different approach to dimensionality reduction by performing feature selection instead of applying PCA, as it would help develop an understanding of which features contribute most to the model.

5. METHODOLOGY

Firstly, I am going to run the standardised, non-PCA data through a logistic regression model as I want to capture the importance of the features without dimensional reduction. This data will be very useful for medical professionals as they can understand exactly which features are more likely to be a pre-cursor to heart failure. Logistic regression is a great starting point for analysis as we can test to see if the data is linearly separable and the model co-efficients directly correspond to the importance of features in the classification process. Logistic regression is used for binary classification - it takes a linear combination of input features and applies the sigmoid function:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

The sigmoid function produces a value in the range 0 to 1, so to perform binary classification the sigmoid output must be mapped to the value of either 0 or 1 as these represent the two possible states of the target variable:

$$\text{y} = \begin{cases} 1 & \text{if } \sigma(x) \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$$

As the death event target variable is represented by a binary value denoting whether the individual died from heart failure, I believe logistic regression is the best starting point for starting my analysis of the dataset, especially due to the fact that the logistic regression models performed well across a more complicated feature-set in the research analysed in the literature review.

I created a logistic regression class instance, including a light l2 regulariser. This is to try and prevent the possibility of overfitting to the dataset, allowing the model to perform better during cross-validation.

After training, I retrieved the feature names and the co-efficients of the model and created a table that orders the features by magnitude of impact on the model predictions.

I set a feature importance threshold at 1.0 and only train the consecutive models on features that scored above this baseline. I decided NOT to include creatinine as I believe its correlation to the target variable may be fully captured in the GFR and sodium-creatinine (SCR) features. These features alongside ejection fraction and age will be used to train the SVM models.

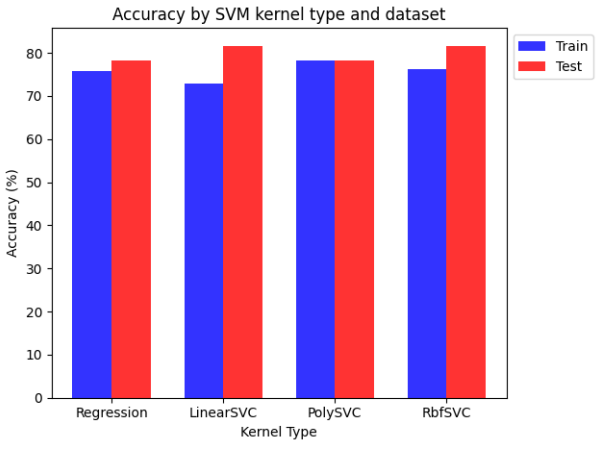
Negative weight values imply that the feature has a negative correlation with the target value, and a positive weight value implies a positive correlation with the target value.

Table 1: Feature Weights and Importances

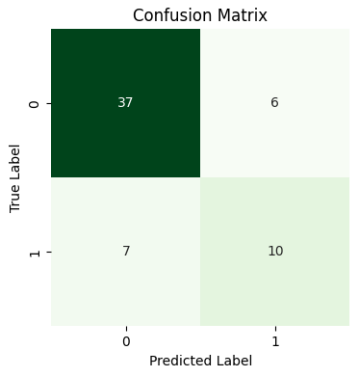
Feature	Weight	Importance
ejf	−1.53	1.53
age	1.21	1.21
gfr	−1.19	1.19
scr	−1.11	1.11
creatinine	0.74	0.74
cph	0.69	0.69
gfr_scr	−0.67	0.67
sodium	−0.59	0.59
ejf_age	−0.43	0.43
hbp	0.39	0.39
platelets	−0.33	0.33
anaemia	0.20	0.20
smoking	0.10	0.10
diabetes	−0.07	0.07
sex	0.00	0.00

I created a new training-test split of the original data only including these important features, min-max normalised the new data and then trained three separate support vector machine models with different kernels, all other parameters the same.

The first kernel type is the linear kernel, which is effective for linearly separable data. The second kernel type is the polynomial kernel which can find polynomial relations between data which is not linearly separable. The third kernel type is the radial-basis function (RBF) kernel which can map input to an infinite dimensional space, and is especially effective when no prior knowledge is known about the data. I tested each trained model on the dataset and compared the training and testing accuracy results alongside the result for the standard logistic regression model.



I was confident in selecting the features for the SVC models due to the fact that the logistic regression model appears to generalise well to the test data, with an accuracy score of 78.33%. The exact model predictions on the test set are outline in the following confusion matrix:



All models trained on this dataset performed strongly, all attaining training and test accuracies about 70%, with both LinearSVC and RbfSVC scoring 81.67% on the test data. All models generalised well to the test set. It is highly likely that with the strong performance of both the logistic regression and LinearSVC models that the data is predominantly linearly separable, however I still believe that choosing the Support Vector Machines model to test across different types of relationship including polynomial and radial basis functions was the most practical choice in terms of accuracy but also speed of implementation. It also allowed a method to purely isolate just the kernel to see how that variable affected the predictions for each model.

```
svc = svm.SVC(kernel = 'poly', C = 0.5).fit(X_train_norm, y_train)
y_pred_poly = svc.predict(X_test_norm)
y_poly = svc.predict(X_train_norm)

acc_poly_train = accuracy_score(y_train, y_poly) * 100
acc_poly_test = accuracy_score(y_test, y_pred_poly) * 100

print(f"Train Set Accuracy Score: {acc_poly_train:.2f}%")
print(f"Test Set Accuracy Score: {acc_poly_test:.2f}%")
```

✓ 0.0s

```
Train Set Accuracy Score: 78.24%
Test Set Accuracy Score: 78.33%
```

```
svc = svm.SVC(kernel = 'rbf', C = 0.5).fit(X_train_norm, y_train)
y_pred_rbf = svc.predict(X_test_norm)
y_rbf = svc.predict(X_train_norm)

acc_rbf_train = accuracy_score(y_train, y_rbf) * 100
acc_rbf_test = accuracy_score(y_test, y_pred_rbf) * 100

print(f"Train Set Accuracy Score: {acc_rbf_train:.2f}%")
print(f"Test Set Accuracy Score: {acc_rbf_test:.2f}%")
```

✓ 0.0s

```
Train Set Accuracy Score: 76.15%
Test Set Accuracy Score: 81.67%
```

6. EVALUATION

While these models do appear to perform well on both the training and tests sets, the process of course has its fair share of drawbacks. Firstly, the dataset is comprised of a moderate number of features but only has 300 rows of data. It is hard to train a model effectively on such a small set of data, and the performance on the data set could be explained by its smaller size. Without question, more entries of data would prove beneficial for training models to test for these features. Perhaps some form of oversampling could be implemented to alleviate this. That being said, narrowing down just the top features for the SVM models likely helped them to capture more obscure relationships in the data.

The model shows neither signs of overfitting nor underfitting, following the training and test results discussed previously.

One success in particular was the feature engineering and feature selection stage. Generally, the engineered features positively affected the overall performance of the model and were selected as features for the SVM models. These relationships could be insightful for medical professionals looking to make connections between bodily conditions.

What may be of interest is performing the same classification task through a decision tree model and capturing the importance of features from that model and comparing the importance of features to that of the logistic regression model. It may provide further insight or support of certain features being more likely to contribute to heart failure. Also, an area of interest as an extension to this paper would be unsupervised clustering methods, potentially with the application of identifying particular groups within this dataset who may be at greatest risk of heart failure.

Also, developing a greater understanding of the co-linearity and correlation between feature variables may provide a greater understanding of how to prepare the data set for learning. Perhaps looking to plot a correlation heatmap or correlation table might have provided deeper insight during the feature selection stage of the analysis.

Overall, I am very pleased with the outcome of the project, as I believe it extends upon existing knowledge of the subject area with some interesting, novel additions that look to tread new ground - and does so successfully.

REFERENCES

[1] World Health Organisation https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds), accessed Friday 8th March 2024

[2] British Heart Foundation https://www.bhf.org.uk/what-we-do/news-from-the-bhf/news-archive/2024/january/early-heart-disease-deaths-rise-to-14-year-high *"Early heart disease deaths rise to 14-year high"*, accessed Friday 8th March 2024

[3] BMC Medical Informatics and Decision Making https://bmcmedinformdecismak.biomedcentral .com/articles/10.1186/s12911-020-1023-5 "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone" by Davide Chicco and Giuseppe Jurman, 2020

[4] Kaggle https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data "Heart Failure Prediction" Dataset used for this analysis, accessed February 14th 2024

[5] CC Attribution 4.0 International License https://creativecommons.org/licenses/by/4.0/

[6] ACP Journals https://www.acpjournals.org/doi/epdf/10.7326/0003-4819-150-9-200905050-00006 "A New Equation to Estimate Glomerular Filtration Rate" by Andrew S. Leve et. al, May 5 2009, American College of Physicians

[7] IEEE Explore https://ieeexplore.ieee.org/abstract/document/8667106 "Improving Heart Disease Prediction Using Feature Selection Approaches" by Saba Bashir et al., 18th March 2019

[8] Science Direct https://www.sciencedirect.com/science/article/pii/S020852162200119X "Heart failure disease prediction and stratification with temporal electronic health records data using patient representation" by Ye Liang and Chonghui Guo, accepted 27 December 2022, Institute of Systems Engineering, Dalian University of Technology, Dalian, Liaoning, China