Predicting heart failure patient survivability

Matthew Devaney

August 1, 2025

Abstract

Heart failure is an increasingly common and potentially fatal condition that occurs when a person's heart is unable to pump or fill with enough blood to support the body. In this study, I examined a data set of 300 heart failure patients and their associated lifestyle and biological data. The purpose of this study is to 1: determine the biomarkers with the strongest predictive power and 2: create a robust predictive model to determine patient survivability. First, I performed an exploratory data analysis of the patient data. Using these insights, I trained several classifiers to determine survivability. Through this process, I find ejection fraction and serum creatinine are the 2 most effective predictors of a heart failure patient's survival.

1 Introduction

Heart failure, also known as congestive heart failure (CHF), is a condition that occurs when a person's heart is unable to either pump or fill with enough blood to support his or her bodily function. People with heart failure often experience shortness of breath, excess fatigue, and swelling among other symptoms. It is an increasingly common health problem and the leading cause of hospitalizations in older adults. Between 2017-2020, 6.7 million

Americans over the age of 20 had heart failure, a number expected to rise to over 8 million in 2030 [Martin et al., 2024]. In 2022, CHF was mentioned on 457,212 death certificates and responsible for 13.9% of all causes of death in the United States [CDC, 2024]. One study found that patients had a ten-year survival rate of 27% (compared to 75% in the general population) - similar to certain cancers [NICE, 2018]. The economic impact is also significant, with costs related to CHF projected to rise from $30.7 billion in 2012 to $69.8 billion by 2030 [Martin et al., 2024].

The goal of this study is to examine several biomarkers and their associations with CHF patient survival outcomes. In particular, the patient data come from the Faisalabad Institute of Cardiology and the Allied Hospital in Faisalabad in Punjab, Pakistan. They were recorded for a period of time between April-December 2015. All patients had left ventricular systolic dysfunction and previous heart failures.

I will examine the data and visualize relationships found between the target variable (patient survival) and the predictors. After finding the most important predictors, I will create classifiers to predict a patient's survival and evaluate their performance.

2 Data

There are 6 binary variables: death event, anaemia, diabetes, high blood pressure, sex, and smoking. The death event was recorded as whether the patient died during the follow-up time. anaemia occurs when the body does not have enough red blood cells to transport oxygen to tissue and is a common comorbidity of heart failure [Anand and Gupta, 2018]. Diabetes affects how the body uses sugar. One study found patients with diabetes mellitus have two times the risk of developing heart failure [Kenny and Abel, 2019]. High blood pressure occurs when the force of blood pushing against the artery walls is too high. This can lead to growth of the left ventricle, which can

progress to heart failure [Tackling and Borhade, 2023].

The other 7 continuous variables are age, creatinine phosphokinase, ejection fraction, platelets, serum creatinine, serum sodium, and follow up time. Elevated levels of CPK in the blood can indicate muscle tissue damage, such as rhabdomyolosis and acute myocardial infarction [Aujla et al., 2024]. The ejection fraction models the percentage of blood pumped out of the left ventricle per contraction. Fractions less than 40% indicate possible heart failure [Penn, 2022] Abnormal platelet levels can affect the blood's ability to clot and repair blood cells. Too many platelets can lead to heart attack or stroke. Too few platelets can cause frequent bleeding [Williams, 2025]. Abnormal serum creatinine levels can indicate poor kidney function, which can affect heart function [Rochester, 2025]. Lastly, low sodium levels have also been shown to occur alongside heart failure [Rondon and Badireddy, 2023].

	Min	Max	Mean	Median	Count 0	Count 1
Age	40.0	95.0	60.8	60.0		
CPK (mcg/L)	23.0	7861.0	581.8	250.0		
EF	14.0%	80.0%	38.1%	38.0%		
Platelets	25,100	850,000	263,358	262,000		
SC (mg/dL)	0.5	9.4	1.4	1.1		
SS (mEQ/L)	113.0	148.0	136.6	137.0		
Follow-up (days)	4	285	130	115		
Anaemia					170	129
Diabetes					174	125
HBP					194	105
Sex					105	194
Smoking					203	96
Death					203	96

Figure 1: Summary statistics of the data

3

The percentage of deaths per age group shows a steady increase as patient age increases. ˜36% of patients aged 61 to 90 died and accounted for 50% of all deaths in the data set. The number of deaths is equally distributed between male and female patients, suggesting there is no difference between the two groups. Furthermore, the majority of deaths occur within the first 2 months of follow-up time, with a spike of patients dying between the 5th and 6th month. ˜85% of patients who had a follow-up within the first 2 months died and ˜56% of all deaths were accounted for by these patients. It is possible that the follow-up time is a function of the patient's condition, such that the most sickly patients have shorter follow-up times. Lastly, each of the health conditions provided in the data set have a similar death rate. (Note: patients with multiple conditions were included in each measurement e.g. a patient with anaemia and diabetes was counted once for anaemia and again for diabetes).
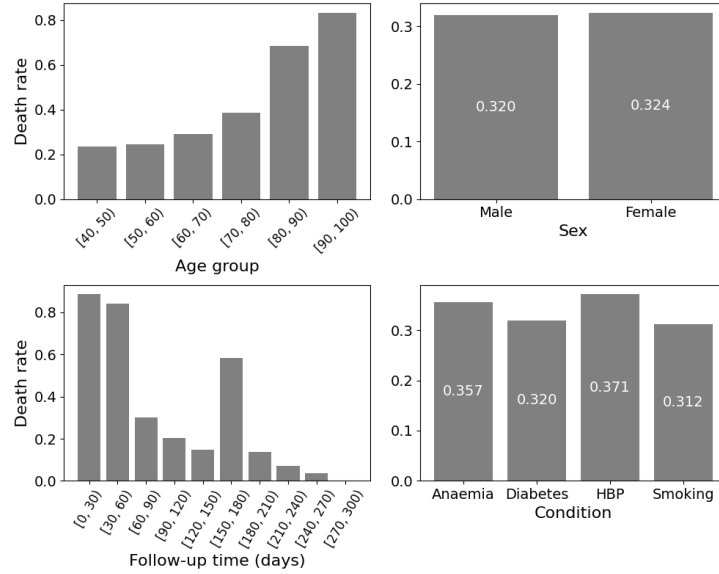

Figure 2: Top left: death rate per 10 year age group. Top right: death rate per sex. Bottom left: death rate per follow-up period. Bottom right: death rate of patients with each condition.

4

The last part of the data I will explore is the correlation between features. From the correlation matrix, we can see that age, anaemia, serum creatinine (SC), and follow-up time have the strongest correlations with death. This indicates that there is at least some relation between the features and patient survival.
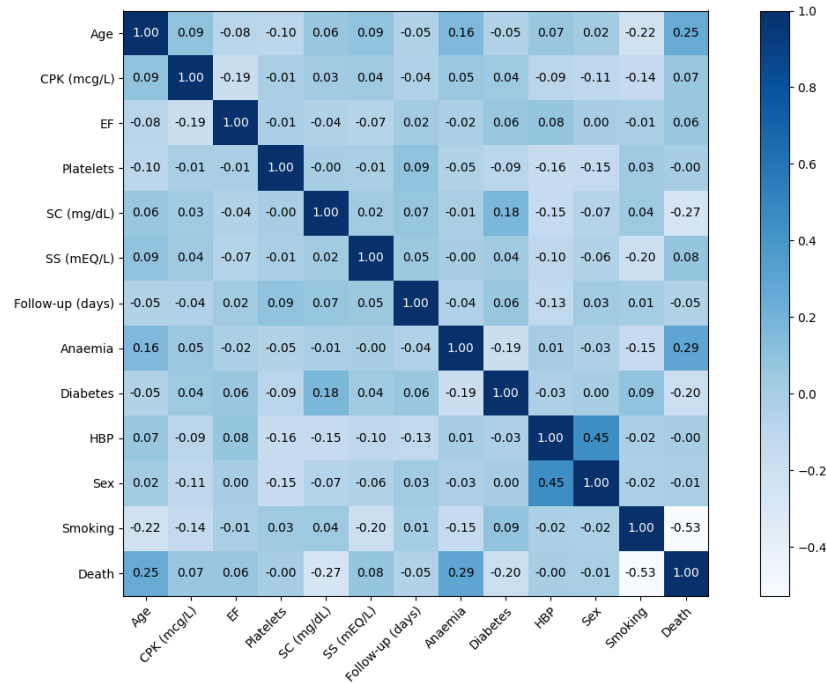

Figure 3: Correlation matrix showing the pairwise correlations between all 13 features. Darker colors indicate a strong positive correlation, while lighter colors indicate a strong negative correlation.

3 Analysis

Now that I have have explored the data set, I will create models to predict the survival outcome of a heart failure patient. After, I will compare them to choose a final model to fine tune. First, I will address the scale of the data. Features like 'platelets' and 'serum creatinine' are on vastly different scales.

To alleviate this, I will standardize all of the continuous features. Also, I have chosen to omit the 'follow-up time' feature to focus on features that could be measured immediately.

I will start by fitting the following machine learning models: logistic regression, support vector machine (with RBF kernel), random forest (500 trees), and XGBoost (1000 trees) using all remaining features. I will assess the model performance with 10-fold cross validation using recall as a scoring metric. Recall is used to emphasize the importance of detecting all positive examples of heart failure. I will also use the following scoring metrics on the test data set: accuracy, precision, F1, ROC AUC, and Matthews' correlation coefficient (MCC). Lastly, to deal with the imbalanced classes (96 deaths vs 203 survivals), I employ the 'balanced' feature of scikit-learn's estimators to penalize wrong predictions on the minority class.

Full model	Mean CV recall score	Standard deviation
Logistic regression	0.73	0.13
SVM	0.69	0.18
Random forest	0.48	0.20
XGBoost	0.59	0.47

Figure 4: 10-fold CV results of models fitted with all features.

As you can see, logistic regression with a balanced class penalty performed the best, while the random forest with 500 decision trees was the weakest. Before I proceed with changing the models, let's examine the learning curves for each of the 4 models. Unfortunately, using the full feature set results in overfitting. In particular, the random forest and XGBoost models have a large gap between training and test performance, which could be due to poor hyperparameter choice.
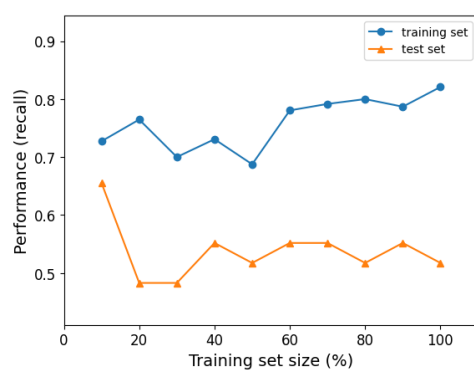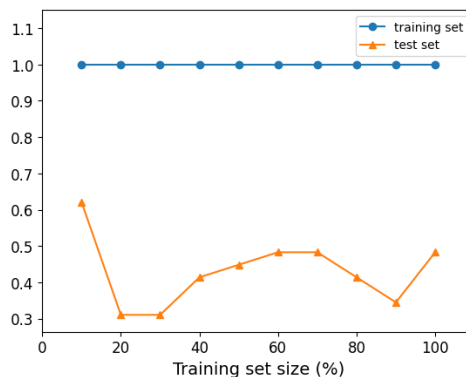
(a) Logistic Regression
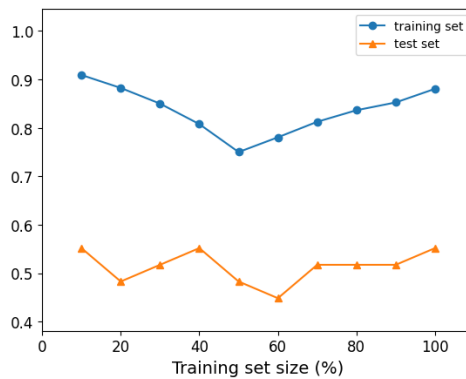
(b) SVM
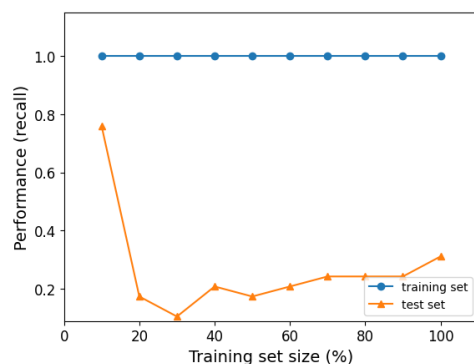
(c) Random forest

(d) XGBoost

Figure 5: Learning curves of the 4 classifiers

To alleviate the overfitting, I will use a random forest to assess feature importance and perform feature selection to shrink the number of model parameters. From the figure, it is clear that serum creatinine and ejection fraction are the two most important features in the data set. Now, I will train the models again, this time using just serum creatinine and ejection fraction as the feature input.
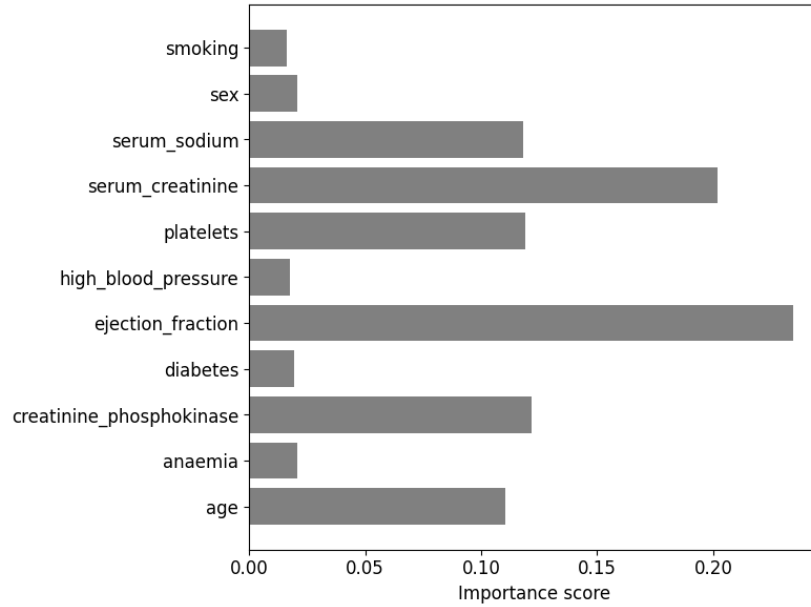
7

Figure 6: Feature importance ranking from a random forest with 500 trees.

Based on the cross validation performance and the learning curves, all 4 classifiers improved when using just 2 features as opposed to the entire feature set. Using less features likely reduced the variance of the models, which explains the smaller gap between the training and testing curves.

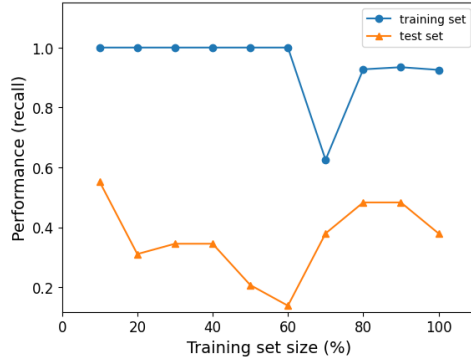| Reduced model	Mean CV recall score	Standard deviation
Logistic regression	0.78	0.14
SVM	0.70	0.19
Random forest	0.64	0.24
XGBoost	0.71	0.22

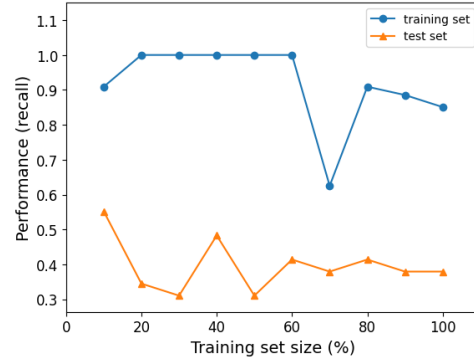Figure 7: 10-fold CV results models fitted with reduced features
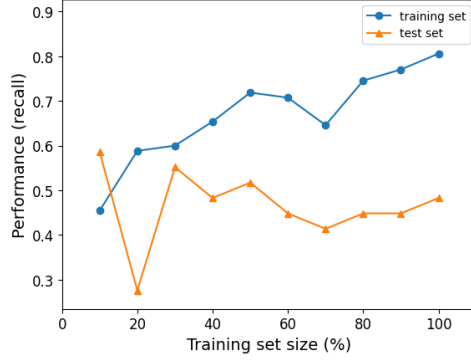
(a) Logistic Regression
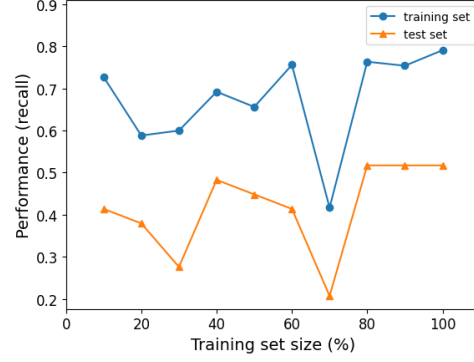
(b) SVM

(c) Random forest

(d) XGBoost

Figure 8: Learning curves of the 4 reduced classifiers

The tree-based methods are still suffering from high variance, so I will implement a grid search to find better hyperparameter values. The best random forest model contained the following parameters: 'max_depth': None, 'min_samples_leaf': 4, 'min_samples_split': 2, and 'n_estimators': 200. Furthermore, the best XGBoost model contained the following parameters: 'colsample_bytree': 0.7, 'gamma': 0, 'learning_rate': 0.01, 'max_depth': None, 'min_child_weight': 2, 'n_estimators': 200, and 'subsample': 0.7. This improved both classifiers greatly, with each achieving a cross validation score of 0.78 and 0.81, respectively.

The final performance test was performed with the remaining 30% of the data set not used for training. Overall, the models performed similarly well, with the SVM and XGBoost models performing slightly better than the logistic regression and random forest model. It is important to keep in mind that these results come from just two predictors.

Model	Accuracy	Recall	Precision	AUC	F1	MCC
Logistic regression	0.67	0.48	0.48	0.62	0.48	0.24
SVM	0.70	0.52	0.54	0.65	0.53	0.31
Random forest	0.68	0.48	0.50	0.63	0.49	0.26
XGBoost	0.70	0.48	0.54	0.64	0.51	0.29

Figure 9: Various scoring metrics based on the test data set

4   Conclusion

The goal of this study was to predict heart failure patient survival. Through exploratory data analysis, I found that the risk of death increased as the patient's age increased. Conversely, the risk of death decreased as follow-up time increased. I found that male and female patients had the same death rate. The health conditions included in the data (high blood pressure, anaemia, diabetes, and smoking) followed a similar trend, with only small differences in death rate between patients with each condition. Then, I created a correlation matrix and found that age, anaemia, serum creatinine (SC), and follow-up time were most strongly correlated with patient survival outcome.

After finding some influential predictors, I created 4 classifiers (logistic regression, SVM, random forest, XGBoost), using the full feature set. After

evaluating the cross validation results, I used a random forest model to assess variable importance in hopes to reduce overfitting. I refit the models using just the two most important features, which reduced the degree of overfitting. Since the tree-based methods were still overfitting, I implemented a grid search to find the best hyperparameters for these two methods. Finally, I used the test dataset to evaluate the performance of the tuned classifiers.

The best performing models were the SVM and XGBoost model, according to Figure 9. The random forest and logistic regression were not far behind in terms of performance, though. For using just 2 predictors, the performance of the models is quite impressive.

Through exploring and modeling the data, I came to the conclusion that the 2 most important predictors of a heart failure patient's survival are ejection fraction and serum creatinine. Ejection fraction models the heart's blood flow, which is obstructed in a heart failure patient. Muscle damage can increase serum creatinine levels, so higher levels may also indicate heart failure. The classifier I built demonstrate their predictive power. These variables should be examined thoroughly in a patient who has heart failure. Careful monitoring of these variables may be able to predict who needs care most urgently.

4.1 Limitations and further analysis

This data set featured imbalanced classes, with about one-third of patients dying and the other two-thirds of patients surviving. It would be beneficial to have a larger sample size and more balanced classes. If balanced classes could not be obtained, then resampling techniques or synthetic methods such as SMOTE could be applied. Moreover, additional predictors would be beneficial such as a person's height and weight. For further analysis, one could produce modeling involving the follow-up time as a variable, as it was shown to be highly correlated with the death event.

References

[hea, 2020] (2020). Heart Failure Clinical Records. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5Z89R.

[Anand and Gupta, 2018] Anand, I. S. and Gupta, P. (2018). Anemia and iron deficiency in heart failure. *Circulation*, 138(1):80–98.

[Aujla et al., 2024] Aujla, R. S., Zubair, M., and Patel, R. (2024). Creatinine phosphokinase. *StatPearls*.

[Bruin, 2018] Bruin, E. (2018). House prices: Lasso, xgboost, and a detailed eda.

[CDC, 2024] CDC (2024). About Heart Failure — cdc.gov.

[Chicco and Jurman, 2020] Chicco, D. and Jurman, G. (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Medical Informatics and Decision Making*, 20.

[Kenny and Abel, 2019] Kenny, H. C. and Abel, E. D. (2019). Heart failure in type 2 diabetes mellitus. *Circulation Research*, 124(1):121–141.

[Martin et al., 2024] Martin, S. S., Aday, A. W., Almarzooq, Z. I., Anderson, C. A., Arora, P., Avery, C. L., Baker-Smith, C. M., Gibbs, B. B., Beaton, A. Z., Boehme, A. K., Commodore-Mensah, Y., Currie, M. E., Elkind, M. S., Evenson, K. R., Generoso, G., Heard, D. G., Hiremath, S., Johansen, M. C., Kalani, R., Kazi, D. S., Ko, D., Liu, J., Magnani, J. W., Michos, E. D., Mussolino, M. E., Navaneethan, S. D., Parikh, N. I., Perman, S. M., Poudel, R., Rezk-Hanna, M., Roth, G. A., Shah, N. S., St-Onge, M.-P., Thacker, E. L., Tsao, C. W., Urbut, S. M., Spall, H. G. V., Voeks, J. H., Wang, N.-Y., Wong, N. D., Wong, S. S., Yaffe, K., Palaniappan, L. P., on behalf of the American Heart Association Council on

Epidemiology, Committee, P. S., and Subcommittee, S. S. (2024). 2024 heart disease and stroke statistics: A report of us and global data from the american heart association. *Circulation*, 149(8):e347–e913.

[NICE, 2018] NICE (2018). Chronic heart failure in adults. pages e26–e28.

[Penn, 2022] Penn (2022). Ejection fraction: What the numbers mean.

[Rochester, 2025] Rochester (2025). Creatinine (blood).

[Rondon and Badireddy, 2023] Rondon, H. and Badireddy, M. (2023). Hyponatremia. *StatPearls*.

[Tackling and Borhade, 2023] Tackling, G. and Borhade, M. B. (2023). Hypertensive heart disease. *StatPearl*.

[Williams, 2025] Williams, M. S. (2025). What are platelets and why are they important?