Heart Failure Survival

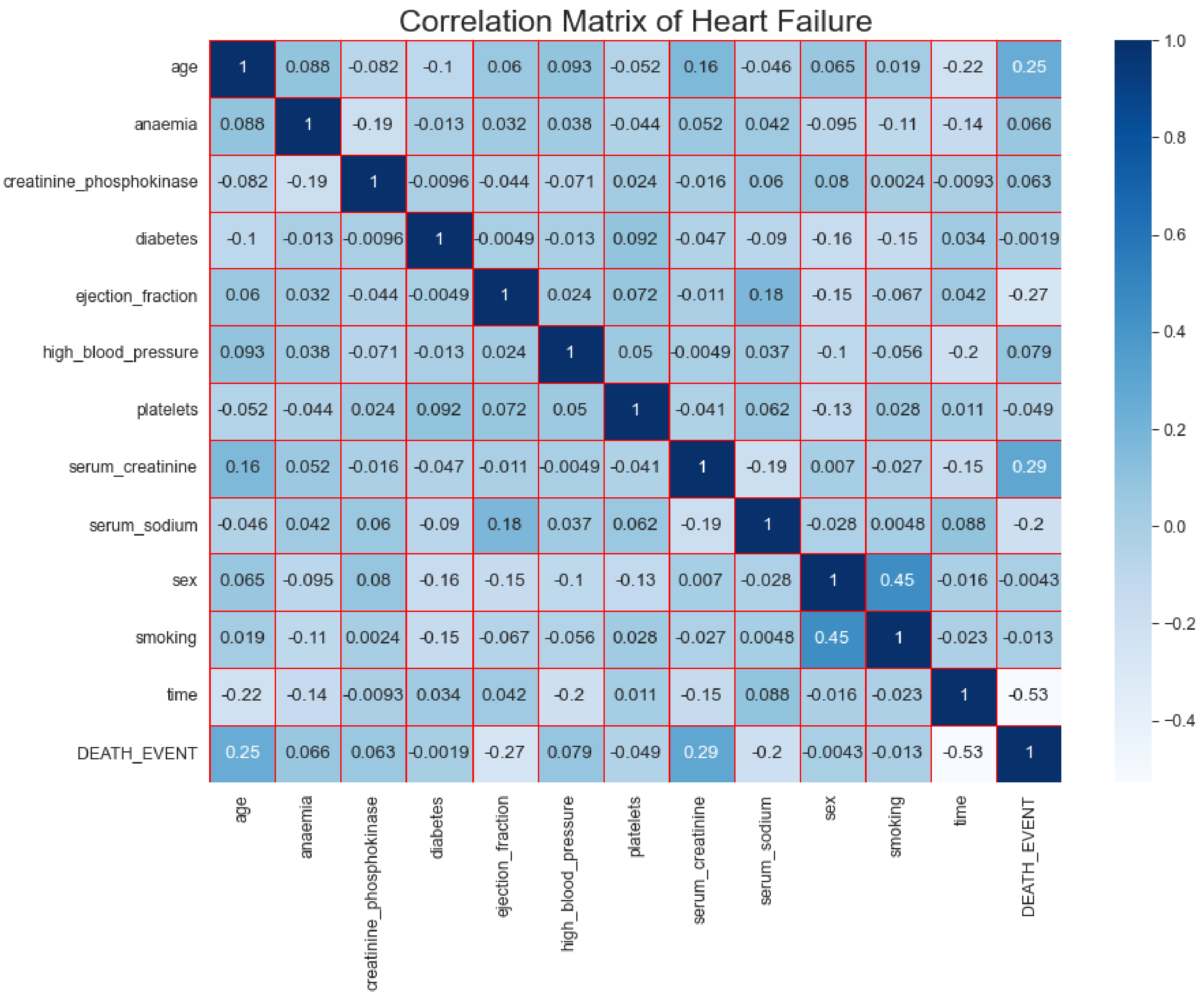
Hear failure (HF) is an insidious cardiovascular disease which kills millions worldwide. There are many factors and components of this disease. The medical world can confidently diagnose this disease based on a number of indicators but cannot yet predict survival with any reasonable precision. This is because each of the factors in this disease is not well correlated with the death event. The conglomeration of factors is no more correlated to the death event than any single factor or grouping of factors. Doctors cannot counsel their patients with confidence about the trajectory of each patient's HF. The medical establishment does not have the tools to determine which test values will predict death with what level of certainty. A good model can potentially be a game-changer in the treatment of HF. Patients have a right to know their chances of survival when facing such a dreadful disease. They need the opportunity to put their affairs in order. If they are expected to survive, they may very well be treated more aggressively, and they must actively manage any other medical issues because they are expected to survive. Knowing which values of which diagnostic test are correlated with death will dictate the directed care necessary to keep the patient alive. The potential savings to insurance companies is significant because there could be an articulate, comprehensive, directed treatment plan. Imagine telling a patient with this dreadful disease that he most likely will survive. The patient can concentrate on a personal care plan with a goal of the best quality of life. Researchers would know which factors to focus on to mitigate the factors most highly correlated factors with death.

The advent of electronic medical records has opened many of these intractable problems to the field of machine learning. These algorithms can potentially find real patterns when none are apparent to the trained eyes of medical professionals or statisticians. I work with a dataset containing records of 299 HF patients in Pakistan. This dataset is housed in the UCI Machine Learning Repository. There are 194 males and 105 females, of whom 72 and 34 respectively died. The imbalance of data and the small sample size are both issues which must be contended with. Ejection fraction is a major indicator of HF. It measures the amount of blood pumped out of the left ventricle during each beat. A patient presenting with an ejection fraction less than 40% most assuredly suffers from left ventricular systolic dysfunction, or left-sided heart failure. But this is not necessarily a death sentence in the short term.

The binary features of the dataset include anemia, hypertension, diabetes, and smoking. These features inject great unclarity into the data. The dataset uses 36% as the cutoff value for hematocrit. This is reasonable because the normal hematocrit for males is a minimum of 41% and for females the normal minimum is 36%. There are no numeric values given for the anemic subjects. Maybe severe anemia is positively correlated with the death event? The same holds true for hypertension. There is a great range of both systolic and diastolic values even within the hypertensive range. The metadata does not indicate what cutoff line was used to establish hypertension. How severely diabetic and which type of diabetes are the positive hypertensives? We are not even privy to what makes a patient positively a smoker. How many cigarettes a day and for how many years has the subject smoked? Are pipes, cigars, or e-cigs included? There is much lacking in these features and it is very understandable why we cannot confidently extrapolate from them to a death event.

The numeric features include CPK which is an indicator in the blood of muscle damage. HF is manifest in heart muscle damage. Low levels of sodium in the blood may be caused by HF. When muscle breaks down creatine metabolizes to a waste product called serum creatinine. The dataset does not provide any information about other concomitant or other comorbidities. These abnormal blood values are associated with a myriad of diseases. Age is given as a whole year value. Subjects have a median age of 60.

There are no null values in the dataset. The continuous variable features are not normally distributed. These patients are all ill and their values are perforce abnormal. There exist outliers in both the ejection-fraction column and in the platelets column. Some HF subjects are able to presently pump blood efficiently. Most HF patients exhibit poor ejection fractions. Aside from these outliers, males and females exhibit similar ejection fractions. The correlation coefficients show no good correlations. The best correlations to the death-event are 0.29 for serum-creatinine, 0.25 for age, and –0.27 for ejection-fraction. Although time is correlated to the death-event, time is not useful because it refers to the follow-up period. It is not informative for how to better treat patients, and therefore I drop the time variable. The dataset does not include any more information related to follow-up care, another great omission which presumably affects the survival outcome.


Correlation Matrix of Heart Failure

It is imperative to get a most accurate picture of the data because the correlations are so dubious. After performing a Pearson coefficient on the dataset I performed a partial correlation to better appreciate how each feature correlates with the death-event controlling for the other features. I then calculated the Kendall Rank Correlation Coefficient because it is a non-parametric hypothesis test. Here serum creatinine showed the greatest correlation at 0.31, outscoring even ejection fraction at –0.25.

Logistic regression is the gold standard for modeling. I created a logit baseline model both on the entire df and death event and on ejection fraction and death event. I standardized the df with StandardScaler. I had to use the non-standardized target column because the logistic regression model interpreted the standardized column as continuous, even though the target included only 2 values... -0.69 and 1.45 for 0 and 1 respectively. Whereas the logit model on only ejection fraction had an accuracy of 67%, the logit model on all features scored only 64%. The MCC on ejection fraction alone is 0.37 which would be one of the higher scores of any model. By contrast the logit model on all features had an MCC of only 0.25. Using the most highly independent, correlated features of ejection fraction and serum creatinine in a logit model did not improve the model. It produced a 64% accuracy and an MCC of 0.25. Adding age to these two features produced a logit model with 67% accuracy and an MCC of 0.30, not better than the logit model of ejection fraction alone to death event.

A major part of model building on this intractable dataset is choosing the best features to model on. The tools of ML must be used in feature selection. PCA reduced the features by one but produced a model faring no better than 67% accuracy and an MCC of 0.30, by no means a success. RFE with cross-validation reduced the optimal features to 3 and performed well. The f-classif between each feature and the target produced results of a group of best features led by age and followed by serum creatinine, ejection fraction, and serum sodium. All other features produced minimal values.

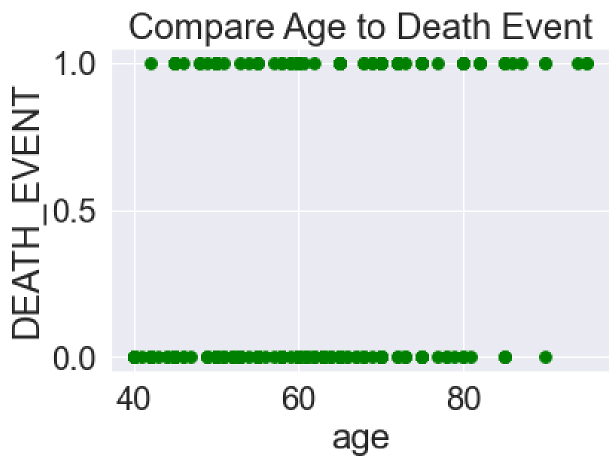
I next created a RandomForest model. This produced a 64% accuracy and an OOB error of 0.24. The all-important recall on death-event is only 47%. This was with tweaking the train/test split from 75%/25% to 85%/15% to increase the train size to give the model more samples to learn on. This comes with the tradeoff of having such a small test set. It behooves us to run the model on the test set multiple times to assess the range of predictions.

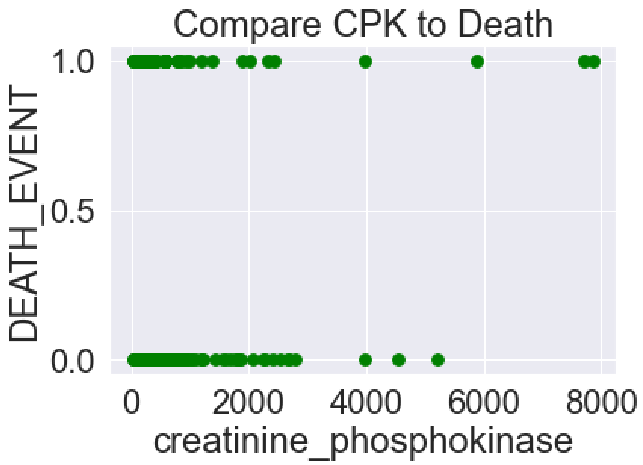
The Gradient Boosting Classifier using StandardScaler and GridSearchCV produced an accuracy of 69% and a recall on 1 of 53%. Using the 3 best features in Gradient Boosting did not improve the model significantly. There is potential for hyperparameter tuning to create a better Gradient Boosting Classifier.

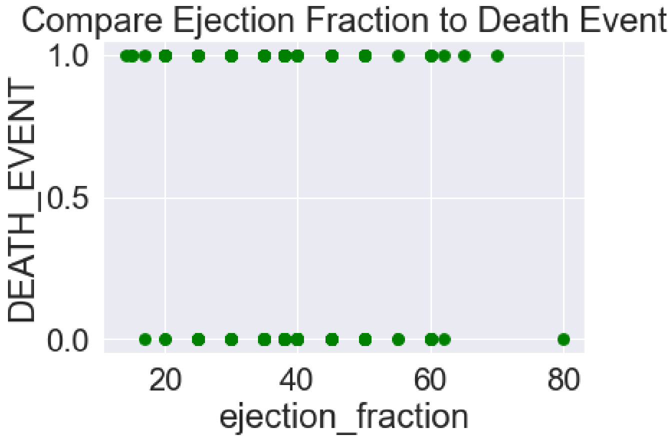
The Look at One Rule Classifier produced a not unexpected poor accuracy of 62% and an abysmal recall on death-event of 26%. Linear Discriminant Analysis produced an accuracy of 69% and an MCC of 0.35. Fitting this model to the three best features did not improve it. A Support Vector Machine model with a GridSearchCV produced only 62% accuracy. A Gaussian Naïve Bayes model with GridSearchCV also produced 69% accuracy. A KNORA-Eliminate model produced only 64% accuracy on the test set. The MCC is only 0.27. A KNORA-Union model did not improve accuracy. A KNORA model with Random Forest Ensemble did not improve on the 69% accuracies of the other models.

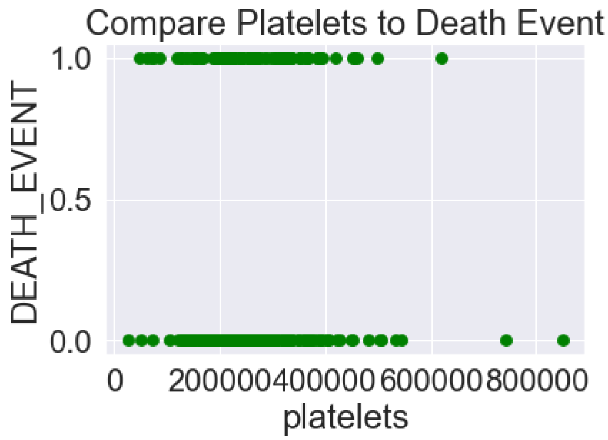
Ensemble models have become increasingly popular for complex modeling. The k-Nearest Neighbor Oracle KNORA-Eliminate model proved to be the best performing. It produced an accuracy of 73%. It produced a recall on death event (1) of 68%. The MCC of 0.45 trounced all other models.

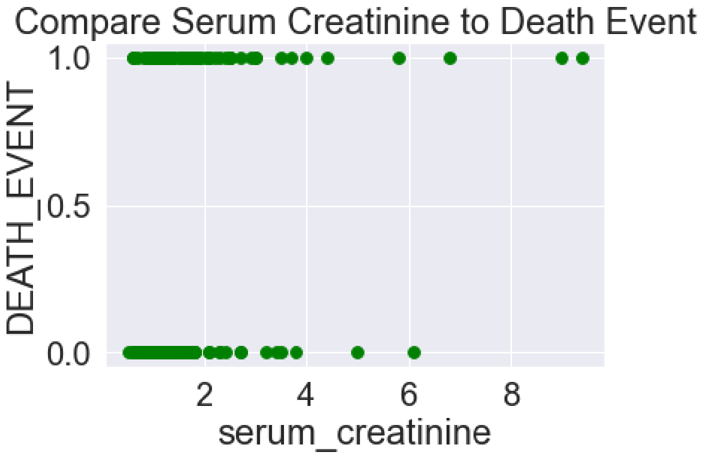
These varied models show what kind of problem there is in modeling this dataset for the death-event. While there is potential to squeeze out a better performing model with robust hyperparameter tuning, there does not appear to be a high ceiling for model improvement. The plots below show little correlation between the feature and the death event.

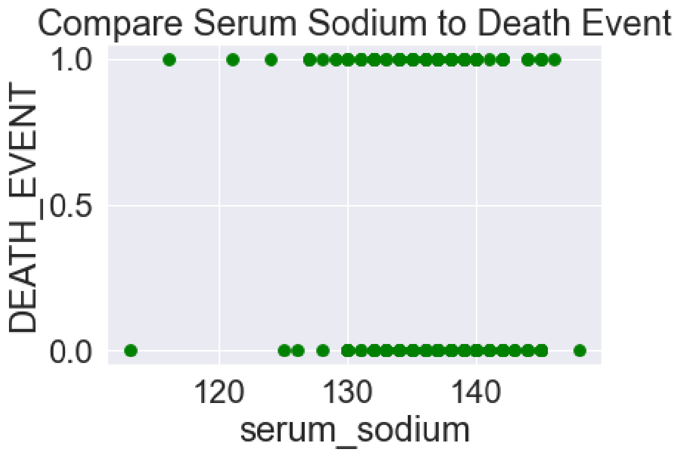












While objectively the models are lacking in performance the scientific medical community has embraced this idea of leveraging ML to both choose the best features to model with, and to choose models to give doctors a better idea of the probability of survival. The medical establishment has been at a loss to predict HF survival. At the very least the modeling shows how doctors need not have every lab result in the dataset to predict as well as or better as with every result. The top two to four features are enough to predict survival as well as using all features.

Depending on the use case will be what type of model to use. What balance does the doctor or patient need? Some models achieved a higher precision statistic. The higher the score, the better chances that a patient will not be advised that he will not survive if he will in truth survive. Recall was rarely more than 50%. This means that false negatives abounded. Too many patients would be advised that they would most likely survive, when in truth they will die. One would assume that wrongly telling a patient that he is expected to survive is nigh unacceptable. It may behoove the data scientist to trade in some accuracy and precision in order to limit the false negatives as much as possible. The MCC must nevertheless always be valued because it gives an objective statistic telling one the quality of this binary classification by stripping away random correct guesses and counting only the correct guesses due solely to the model.

The model may be deployed and used to immediately improve the care of HF patients. The best models have been scoring 74% accuracy. These models have been struggling to achieve MCCs of 0.4. However, there is great room for improvement. With the proliferation of electronic medical records, there is a veritable treasure trove of potential data to be leveraged. One can access these records to produce datasets of 2000, 20,000, and 200,000 subjects and try ML to 'learn' the data to create good models. Many refinements are in the offing. There are so many HF patients that one can model for a chosen sex in a given locale, for a given age to get a better model. One can control for other comorbidities. Unfortunately, many HF patients also suffer from pulmonary and renal issues which may positively correlate to the death outcome. We are not privy to what other diseases the subjects of our study were suffering from. I am confident that there are better models in the offing, if only because there exists so much more data.

Best model per type	Accuracy%	Precision	Recall	F1-score	MCC	AUC %
Logit model on top 3 features	67	0.68 0.64	0.81 0.47	0.74 0.55	0.30	64
RandomForest	64	0.67 0.60	0.77 0.47	0.71 0.53	0.25	62
GradientBoosting with GridSearch	69	0.70 0.67	0.81 0.53	0.75 0.59	0.40	67
XGBClassifier on top 3 features	71	0.76 0.65	0.73 0.68	0.75 0.67	0.41	71
One Rule	62	0.62 0.62	0.88 0.26	0.73 0.37	0.19	57
LinearDiscriminantAnalysis	69	0.70 0.67	0.81 0.53	0.75 0.59	0.35	67
SVM with GridSearch	62	0.64 0.58	0.81 0.37	0.71 0.45	0.20	59
Naïve Bayes(Gaussian) with GridSearch	69	0.69 0.69	0.85 0.47	0.76 0.56	0.35	66
k-Nearest Neighbor Oracle KNORA-Eliminate	73	0.77 0.68	0.77 0.68	0.77 0.68	0.45	73
KNORA-Union	62	0.68 0.55	0.65 0.58	0.67 0.56	0.23	62
KNORAU Choosing Classifier	69	0.68 0.55	0.65 0.58	0.67 0.56	0.23	62
KNORA with RandomForest Ensemble	69	0.73 0.63	0.73 0.63	0.73 0.63	0.36	68