Muzahidul Islam
DAT 315

Decision Tree Model Shows High Accuracy in Determining Heart Failure Patients Survival

Abstract

Along with cancer and other respiratory diseases, cardiovascular diseases are some of the most frequent causes of death worldwide. Our research conducted data analysis on heart failure patients to determine what factors are most associated with survival of a patient with heart failure and to predict their survival using various modeling techniques. 9 different modeling techniques were utilized with various tuning combinations for a total of 20 models. From these techniques, it was determined that serum creatinine and ejection fraction were the two most useful predictors in determining the survival of the patient. The most accurate model was able to predict survival or death of the patient with 80% accuracy.

Introduction

The data in our study contained the medical records of 299 heart failure patients collected at the Faisalabad Institute of Cardiology and at the Allied Hospital in Faisalabad (Punjab, Pakistan), during April–December 2015. It is from a public database and can be utilized under the license Creative Commons Attribution 4.0 International (CC BY 4.0) : https://plos.figshare.com/articles/dataset/Survival_analysis_of_heart_failure_patients_A_case_study/5227684/1

Medical records are usually protected, but having access to them and applying data analysis can help us draw useful conclusions for medical care of the patient. Our data specifically looked at heart failure patients as it can be difficult for doctors to predict when heart failure may occur and the best medical care to apply to these patients. Our study was concerned with the survival of the patient, and applying statistical techniques on the medical dataset can allow us insight into what

factors doctors should possibly consider the most during treatment. The dataset contained 11 different variables and their corresponding values for each of the 299 patients. For reference, 203 or 68% of the patients survived and 96 or 32% of the patients died. The first task is to apply a scatterplot matrix between each of the quantitative variables to determine if any of them are strongly correlated with each other. There is no strong correlation between any of our quantitative predictor variables as evidenced by **Figure 1**, so we do not need to consider removing any of them before applying our modeling techniques.

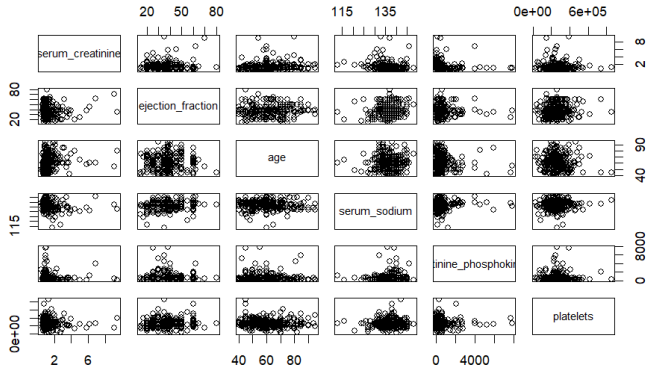

Figure 1.

Our analysis determined that the two most important factors in predicting the survival of the patient are ejection fraction percentage and serum creatinine levels. Ejection fraction is the percentage of blood pumped out from the heart during each contraction. **Figure 2** shows that the median ejection fraction percentage for the patients that survived was higher than the median ejection fraction percentage for the patients that did not survive.

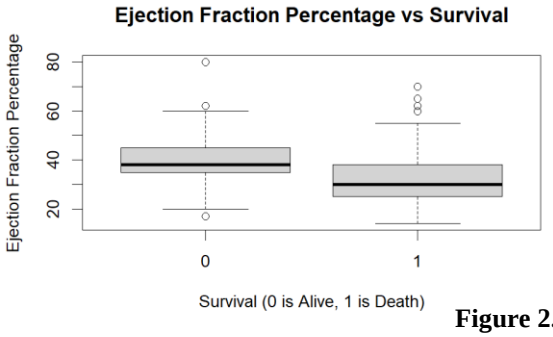

Figure 2.

For serum creatinine, the waste product generated by creatine when the muscles break down, we can see that the median waste was higher for patients that died than patients who survived (**Figure 3**).

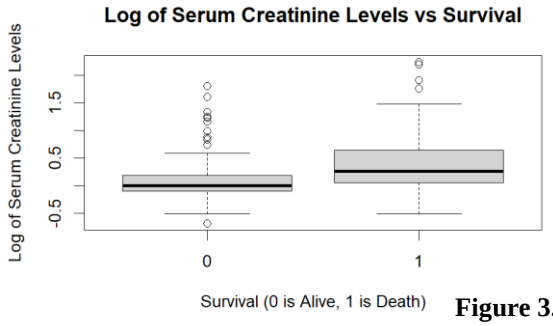

Figure 3.

(The log was taken to get a better scale of the y-axis so we could see the box and whisker plot better.)

Methods

The study utilized 9 different modeling techniques: logistic regression, KNN, linear discriminant analysis, ridge regression, lasso regression, bagging, random forest, boosting, and support vector machines. The 299 observations from the data were split into test and training with an 80/20 split. The seed was 5 and all techniques were applied to this data split to compare their accuracies. For logistic regression, all 3 variable technique selection methods (Forward, Backward, and Stepwise) were used with SBC as the selection criterion. The logistic regression analysis considered all possible two-way interactions between variables in the model. The KNN method was ran by scaling the continuous variables only. We utilized cross validation to identify an optimal k of 7. The test proceeded with linear discriminant analysis and then ridge and lasso regression. Ridge and lasso considered all possible two-way interaction like the logistic regression model from earlier. Lasso provided variable selection but didn't improve on the accuracy compared to ridge. Then we moved on to using decision trees for predictions. Bagging was the first method, followed by random forest. They all used Gini Index as the loss function. 5000 trees or bootstrap samples were used to train every model. Random forest tried 3 different values for m, which is the number of predictors considered for each split. The considered values for m were $\sqrt{p}$, p/2, and p = 1, where p represents the number of predictors in our dataset. Our data had 11 predictors. Random forest with m=p/2 resulted in the most accurate model in the study and is more thoroughly discussed later. For boosting, 5000 trees were used during training. 3 different values were considered for λ or the shrinkage, and depth for each tree considered values of 2 and 4. This equated to 6 different combinations of boosting. A few of them resulted

in similar models. The final technique considered was support vector machines. The study compared linear, radial, and polynomial with a degree of 3. For SVM linear it found the best cost was 0.1. For radial, the ideal cost was 1 and $\gamma = 0.5$, and for polynomial of degree 3 the ideal cost was 1.

Results

Model	Accuracy	TPR	TNR
Logistic Regression Backward SBC	0.7500	0.6154	0.7872
Logistic Regression Forward SBC	0.6833	0.4615	0.7447
Logistic Regression Stepwise SBC	0.7500	0.6154	0.7872
KNN (K Nearest Neighbors)	0.7000	0.3333	0.8571
Linear Discriminant Analysis	0.7333	0.3889	0.8809
Ridge Regression	0.7000	0.2222	0.9048
Lasso Regression	0.7000	0.2778	0.8809
Bagging	0.7667	0.4444	0.9048
Random Forest m=sqrt(p)	0.7833	0.4444	0.9286
Random Forest m=p/2	**0.8000**	**0.5000**	**0.9286**
Random Forest m=1	0.7500	0.2222	0.9762
Boosting $\lambda = 0.1$ and $d = 2$	0.7500	0.4444	0.8809
Boosting $\lambda = 0.01$ and $d = 2$	0.7667	0.4444	0.9048
Boosting $\lambda = 0.001$ and $d = 2$	0.7333	0.3889	0.8809
Boosting $\lambda = 0.1$ and $d = 4$	0.7500	0.4444	0.8809
Boosting $\lambda = 0.01$ and $d = 4$	0.7667	0.4444	0.9048
Boosting $\lambda = 0.001$ and $d = 4$	0.7500	0.4444	0.8809
SVM linear cost = 0.1	0.7167	0.3333	0.8809
SVM radial cost = 1 and $\gamma = 0.5$	0.7167	0.3333	0.8809
SVM polynomial cost = 1 and degree = 3	0.7167	0.3333	0.8809

Figure 4.

Figure 4 above shows all the different model selection techniques and their corresponding accuracy, true positive rate, and true negative rate. The accuracy represents the proportion that the model correctly identified survival or DEATH_EVENT from our dataset, when fitting the trained model to the test data. TPR or true positive rate represents how often the model correctly predicted the positive class, the patient dying, when fitted to the test data. TNR, or true negative rate, represents how often the model correctly predicted the patient surviving when fitted to the

test data. All the models seemed better at predicting patients surviving. This is likely due to there being more observations for that outcome in our data. Around 203 patients survived, while 96 died. The final chosen model utilized random forest with m, the number of predictors considered for each split, as 11/2, rounded down to 5. This model proved to be the most accurate in fitting to the test data, and it was not as bad at misclassifying the patients that died, compared to some of the other models. The confusion matrix when fitting the model to the test data is displayed in **Figure 5.** To calculate TPR$= \frac{\text{TP}}{\text{TP+FN}} = \frac{9}{9+9} = 0.50$ and TNR$= \frac{\text{TN}}{\text{TN+FP}} = \frac{39}{39+3} = 0.9286.$ There is an argument for also considering the Backwards SBC model because it had a better TPR, but it had a much lower TNR. For doctors, it can be argued that they may care more to identify patient's survival to not waste unnecessary resources in the hospital.

	Actual	
Predicted	0	1
0	39	9
1	3	9

Figure 5.

Prediction Factors

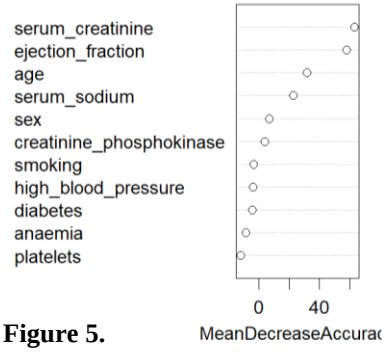

Figure 5.

The random forest model also allows us to identify what the most important factors are in determining the survival of a patient. **Figure 5** shows that serum creatinine and ejection fraction are the most important factors. The x axis shows the average decrease in accuracy when leaving these variables out of our model. For these two variables, it is above 40%, which is considerably high. Age also shows relevance, and this is logical since increases in age may lead to more likely death outcome, holding all else fixed. Platelets, anaemia, diabetes, high blood pressure, and smoking did not show too much relevance as a predictor in helping affect the accuracy of the model.

Muzahidul Islam
DAT 315

Discussion

This study displays the power of applying decision trees and other techniques in helping to possibly determine the survival of heart failure patients. The modeling techniques used were able to generate around 75% accuracy in determining a heart failure patient's survival. Specifically, random forest modeling can also be used to help a doctor determine the most important factors that they should possibly look at when determining treatment for a heart failure patient. It can be argued that doctors in the medical field may use their own form of decision tree reasoning to diagnose illnesses and apply treatment. A patient goes in for an appointment, fills out forms listing their symptoms, and through experience and training the doctor determines the best method to move forward. This system doesn't lead to perfect diagnoses and many patients aren't given the proper treatment. In fact, there have been studies which show that medical errors may be the third leading cause of death in the United States (Anderson JG). With the recent breakthroughs in deep learning and the rise of Large Language Models, we may already be seeing doctors and other professionals use them to aid in their work, since they allow much easier interpretation then analyzing the studies of a research paper like this. To reiterate, this study can only be applied to heart failure patients in a specific hospital in Pakistan. To help generalize this study on many other types of medical patients requires collecting and analyzing data from many races and regions. However, the legal implications complicate that in many areas of the world. The study shows promise in applying decision tree modeling and other techniques to health data, but that needs to be used in combination with experience from doctors in the real world to help aid in their work. Leaving the decisions in a sensitive field like the medical industry to be left purely up to doctors seems outdated with the advancements in AI and other

statistical fields. Statistics is a powerful tool, and it can possibly be useful in helping the medical industry move forward.

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. Ahmad, Tanvir; Munir, Assia; Bhatti, Sajjad Haider; Aftab, Muhammad; Ali Raza, Muhammad (2017): DATA_MINIMAL.. PLOS ONE. Dataset. https://doi.org/10.1371/journal.pone.0181001.s001
3. Anderson JG, Abrahamson K. Your Health Care May Kill You: Medical Errors. Stud Health Technol Inform. 2017;234:13-17. PMID: 28186008.