

**ITEC 621 Predictive Analytics Project**

**Project Name: Monitoring Heart Health:**

**Using machine learning to understand potential risks for death caused by heart failure.**

**Class Section: 001**

**Team Number: 1**

**Team Members:**

Norris Jaber

Rajeev Agrawal

Sha Lu

Yuxi Duan

Jen-Yu Huang

**Last updated:** 04/27/2022  
**Deliverable Number:** 4

1. **Business Question and Case**
   1. ***Business Question***

What factors increase the risk of death due to heart failure?

* 1. ***Business Case***

Cardiovascular diseases (CVDs) are the number one cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide. Heart failure is a common event caused by CVDs and this dataset contains 12 features that can be used to predict mortality by heart failure. Most cardiovascular diseases can be prevented by addressing behavioral risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity and harmful use of alcohol using population-wide strategies. People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidemia or already established disease) need early detection and management wherein a machine learning model can be of great help.

1. **Analytics Question**
   1. ***Outcome Variable of Interest***

Our outcome variable of interest is whether a person died of heart failure. This is represented as a binary variable called “DEATH\_EVENT” in our data set, where 0 = no death and 1 = death. From our outcome variable, we hope to add greater interpretational value to our predictors to allow healthcare providers better understand potential risks each patient may have. Also, the development of a machine learning model for prediction can help healthcare providers in early detection and management.

* 1. ***Main Predictors***

The key predictors of our model include demographic information such as age and gender. This will allow us to see which demographic is more likely to die from heart failure. We will also be including health related factors such as if the patient has anemia, high blood pressure, diabetes, as well as if the patient is a smoker. These predictors are important as it will show us which health related variables are more likely to cause death from heart failure in patients. Knowing these will allow caregivers and doctors to know which of their patients are most at risk. We will also be including more specific predictors such as level of creatinine in the blood and level of sodium for example, which will allow us to see a greater spectrum of patient health. Such predictors are important as they may show how close a patient is to death via heart failure and will allow caregivers to give more precise treatment.

1. **Data Set Description**

For this study, the data set was obtained from Kaggle [1]. This is a data set of 299 patients with heart failure collected at the Faisalabad Institute of Cardiology and at the Allied Hospital in Faisalabad (Punjab, Pakistan), during April – December 2015. The patients consisted of 105 women and 194 men, and their ages range between 40 and 95 years old. All the patients had left ventricular systolic dysfunction and had previous heart failures that put them in classes III or IV of New York Heart Association (NYHA) classification of the stages of heart failure [2].

1. **Descriptive Analytics**
   1. ***Descriptive Statistics of Key Variables***

Average age of the patients within our data is around 60 years of age, with the youngest patient that experienced heart failure at 40 years old and the oldest at 95. Time between visits with health care providers and the event of heart failure was recorded in days, as there was an average between visit time of 130 days, with a minimum of 4 days and a maximum of 285. Biological readings were also performed all on patients recording creatinine phosphokinase level, ejection fraction, serum creatine, serum sodium, and platelets. Of these variables, creatinine phosphokinase had the largest range of readings with creatinine phosphokinase reading at a minimum of 23 micrograms per liter to a maximum of 7861 micrograms/L, and platelets with a minimum of 25100 kiloplatelets per milliliters and a maximum of 850000 kiloplatelets/mL. The averages of these two variables were 581.8 micrograms/L and 263358 kiloplatlets/mL respectively. Serum creatinine and serum sodium had fewer extreme ranges. Serum creatinine exhibited a minimum of 0.5 milligrams per deciliter to 9.4 mg/dL. Serum sodium exhibited a range from 113 milliequivalents per liter to 148 milliequivalent/L. Averages for these two variables were 1.100 mg/dL and 136.6 milliequivalent/L respectively. Finally, the ejection fraction ranged from 14% to 80% with an average of 38%. We cannot determine if this range is considered extreme as percentages of blood in the heart changes per patient. Descriptions of binary variables will be found in our discussion of distributions along with the quantitative variables.

* 1. ***Distribution of Key Variables***

Within our data we have 299 total observations of patients that have experienced heart failure, as well as 13 variables that were measured by care providers for each patient. Of these variables, anemia, high blood pressure, diabetes, sex(male), and smoking are binary variables, where 1 = yes. As seen in Figure 1 of the Appendix of the 299 total patients, 129 had anemia, 125 had diabetes, 105 were considered to have high blood pressure, and 96 were smokers. More males were present in our data with a 194 to 105 ratio.

As seen in Figure 2, of the quantitative variables within our data, age and platelet count appears to have a normal distribution. Serum sodium, serum creatinine, and platelets each exhibit tails within their distribution, with both platelets and serum creatinine having right tails and serum sodium having a left tail. However, not taking the tailing into account, the distribution of these variables also appears normal as well. We will need to determine if these tails will be a cause for concern. Creatinine phosphokinase distribution is mostly clustered between a count of 0 and less than 2000, with a very long right tail. Lastly ejection fraction and time (follow-up days), are showing two peaks in the distribution. However, further analysis is needed to determine if this is a cause for concern.

* 1. ***Correlation and Co-Variation Analysis***

According to the correlation matrix (Figure 3), our outcome variable of if a patient died from heart failure is highly correlated with age, serum creatinine, time, ejection fraction, and serum sodium. There does not appear to be correlation issues with our predictors variables asides from sex and smoking.

* 1. ***Data Pre-Processing and Transformations***

Some of the continuous predictors such as CPK and Serum creatinine are right-skewed. One of the ways to make them closer to normal distribution is to take the logarithm. However, we have a sample size of 299 (50+ data points), therefore, the predictors do not have to be normally distributed. Therefore, we leave the continuous predictors without any transformation.

1. **Modeling Methods and Model Specifications**
   1. ***Initial Model Specification***

Since our outcome variable is binary, our initial model will be a logistics model using all 12 variables as predictors. From this initial model, we hope to understand not only how well this model type fits our data, but also which predictors display more significance on our outcome.

* 1. ***Initial OLS or Logit Model Results***

We see in Figure 4, based on the p-values in the following summary output, not all the features in this full model are significant. Variables like age, ejection fraction, serum creatinine, and time (follow-up period) are significant with p-values lower than our alpha level of 0.05. The other predictors such as anaemia, diabetes, high blood pressure, platelets, serum sodium, sex, and smoking are not significant (p-values < 0.05).

* 1. ***Assumption Tests***

To evaluate our initial model, we needed to ensure that the assumptions for the logistic model are met. We already know that our dependent variable is categorical, and that each observation is independent from each other as each observation is a different patient. Our data set is also large enough to allow us to draw valid conclusions from the fitted logistic model. Asides from these first three assumptions, we must ensure that server multicollinearity is not present between the predictors, there are no extreme or influential outliers, and that the independent variables are linearly related to the log odds. From these assumptions, we tested the assumption for multicollinearity as there was no evidence to suggest that the other assumptions were violated.

Multicollinearity could be an issue because it makes it difficult to separate out the impact of individual predictors on the response. We tested for multicollinearity using the conditional index (CI) and variance inflation factor (VIF). Upon testing for the conditional index, we obtained a CI value of 136.49, as seen in Figure 5. This value was very high and led us to believe that we may have a major multicollinearity issue, however when testing the VIF, no predictor returned with a VIF value greater than 2, as seen in Figure 6. This led us to believe that a predictor is highly correlated with the intercept. To remedy this, we centered the data and calculated the CI again. Upon centering, we obtained a much lower CI value of 5.6. Shrinkage methods could also be used to deal with the multicollinearity issue.

* 1. ***Model Candidates and Rationale***

Since our dependent variable is categorical and binary, the model candidates we choose to evaluate are the logistic model, Lasso model, random forest, as well as a boosted tree. We started with a logistic model as it made the most sense for our data, as well as being highly interpretable. This allowed us to see the effects of each variable, regardless of their significance. We also used the shrinkage method Lasso to validate the performance of our logistic model and deal with the multicollinearity. We expected the significant coefficients from the logistic model to be represented in the lasso model.

Both logistic models and lasso have high interpretability which is great to understand how the different biological predictors interact with the chance of death from heart failure, however there are other models that could be better at predictability. We decided to try both a random forest and boosted trees for their additional predictive value. In doing so, we hoped to create a model that would be beneficial for healthcare providers in determining the risk of specific patients.

* 1. ***Model Specification Candidates and Rationale***

We ended up following two model specifications across all the models we tested. A model with all variables, and a smaller nested model with only the most significant predictors. The full logistic model had tolerable multicollinearity after centering, so it is a good model to consider since it contains all the variables, and no information is lost. We also do not have many predictors so felt comfortable keeping in all the predictors. Also keeping all the variables will allow us more interpretability in the result and be less biased.

For the nested models however, we elected to use only the most significant predictors; being age, ejection fraction, serum creatinine, and follow-up time. We attempted a quadratic model, but the coefficients were found insignificant. Other larger models with some more predictors were also an option. Across the different models, we found that the 4 significant variables in the logit model had relatively higher variable importance. Also, multicollinearity was not an issue in the logit model with these 4 variables. We confirmed that the 4 variables were the most significant by running a stepwise variable selection method. When utilizing the model at an alpha level of 0.15, we obtained a model with 5 variables with a high CI value. Upon switching to an alpha of 0.05, these multi-collinearity issues were resolved and age, ejection fraction, serum creatinine, and follow-up time were the only variables found significant. We elected to keep these 4 variables throughout all the modeling methods to allow for better comparison.

* 1. ***Cross-Validation Testing and Final Model Selection***

From our total 8 models, we cross-validation tested them to determine which models performed the best in terms of interpretation and prediction power. To do so, we divided the data into testing and training sets. Using the training set, we refit the models before evaluating their performance using the test set.

Class imbalance is a common problem when working with real-world data. The performance of the ML model is degraded because of it as it biases the model toward the majority class at the expense of the minority class. The class distribution of the test data set should mirror that of the original dataset because a model's performance against the test data is a proxy for its generalizability against unseen data. However, the training data should be balanced prior to the modeling process. Once balanced we fit the models with this training set.

* + 1. The Logistic Model

We fitted our logistic model with the two specifications, being the full set of predictors and the nested specification using only the 4 predictors deemed significant. According to the confusion matrices in Figure 7, the full model obtained an accuracy of 73.3 with an error rate of 26.7%, while the nested specification model obtained an accuracy of 75%. The full model had a higher sensitivity and lower specificity as well than the nested model. Finally, the false positive rate obtained by the full model was 35.9% compared to the nested model’s 30.8%.

* + 1. The Lasso Model

We followed the same procedure when creating the Lasso models. According to Figure 8, the full model performed better than the smaller nested model with an accuracy of 78.3% compared to the nested specification model’s 75%. The full model also had a lower error rate and slightly her sensitivity and specificity compared to the specified model. It is worth noting that the smaller specified model performed the same as the specified Logistic model. We anticipated this as the lasso model did not shrink any of these 4 variables as they were all deemed significant.

* + 1. Random Forest Model

Both the Logistic and Lasso models are great for interpretation, however we wanted to include models that could be stronger at prediction. We first used the Random Forest Model to attempt to generate greater accuracy. We kept the specifications the same as the logistic and lasso models to allow for better comparison. According to Figure 9, both the full and nested model for random forest performed better than both the logistic and lasso models. The full model obtained an accuracy of 83.3% while the nested model obtained 81.7%. Despite having higher accuracy, both random forest models had the same sensitivity rate of 76.2%, lower than both the logistic and lasso models.

* + 1. Boosted Tree Model

The final model method tested was the Boosted Tree Model. According to Figure 10, both boosted tree models obtained the highest accuracy when compared to all models. The nested model had an accuracy rate of 86.7%, slightly higher than the full model with an accuracy of 85%. The nested boosted tree model had sensitivity and specificity rates that were very close to each other of 85.7% and 87.2%. This is the closest these rates have been across all modeling endeavors.

* + 1. Final Model Selection

To compare all models together, we created ROC plots for both the full and nested specification models. (Figures 11 and 12) When comparing these ROC plots, we see that each method performs similarly to each other, with differences in the area under the curve. Among all models, the full boosted tree performed the best with an AUC of 0.921, however all models had similar AUC values with the lowest AUC of 0.902 obtained by the full random forest model. Based on these results, all the classifiers performed very well so we decided that is the best model because of its high interpretability. For predictive purposes, we decided that the full boosted tree would be the best as it has the largest AUC value.

1. **Analysis of Results**

Comparing all models, together we see that all of them performed similarly. The logistic regression model which contains only the significant variables we found was the best for interpretation. The variables age and serum creatinine were found to have a positive correlation with death from heart failure. For each year increase in age, the log odds of dying from heart failure increases by 0.04, that is when all variables are held in constant. As with serum creatinine, on average and when holding all variables constant, when serum creatinine increased by 1 milligram per deciLiter, the log odds of death by heart failure increased by 0.72. Ejection fraction and time had a negative correlation with death. When holding all variables constant, on average per 1% increase in ejection fraction, the log odds of death by heart failure decreased by 0.07. As for time, when holding all variables in constant, when time between doctor visits increased by 1 day, the log odds of death by heart failure decreased by 0.02. As for prediction, the full boosted model performed the best, and should be used by healthcare providers to determine risk of death for patients. This model does not have much interpretability however, this can be used in conjunction with the logistic model to better explain how each of these significant variables plays a role in their heart health.

1. **Conclusions and Lessons Learned**
   1. ***Conclusions from the Analysis***

With the goal of providing health care providers better interpretational power, the logistic model performed the best. We saw that the variables, age, serum creatinine, ejection fraction, and time were all very significant predictors in whether a patient died from heart failure. Knowing how these predictors influence the outcome will allow healthcare providers to provide better care and allow them to understand the potential causes of concern for their patients. In doing so, they can mitigate the risks for their patients and hopefully result in less death from heart failure. Interesting outcomes from our analysis show a negative correlation between follow-up time and death from heart failure. We would have expected time between visits to be positively correlated as it is more time away from receiving health care, however this could also imply that doctors were already monitoring their patients with higher risk and were set on a more regular schedule.

In terms of prediction, we found that the full boosted tree had the best overall performance. Using this model will allow healthcare providers to predict the possibility of a specific patient using the patient’s own biological data. This will allow healthcare providers to know how severe a patient’s health is.

* 1. ***Project Issues, Challenges and Lessons Learned***

One limitation of the current study is that it may only be valid on a similar data set as was used for this study, which was sourced from a very specific location. Further research is needed to check if similar results are seen for data collected elsewhere. If an additional external dataset with the same features from a different geographical region had been available, we would have used it as a validation cohort to verify our findings.

Another limitation of the present study that we report is the small size of the dataset (299 patients): a larger dataset would have permitted us to obtain more reliable results. Additional information about the physical features of the patients (height, weight, body mass index, etc.) and their occupational history would have been useful to detect additional risk factors for cardiovascular health diseases.

**Appendices**

1. **Data Information**

Our data set was obtained from Kaggle. This is a data set of 299 patients with heart failure collected at the Faisalabad Institute of Cardiology and at the Allied Hospital in Faisalabad (Punjab, Pakistan), during April – December 2015. The patients consisted of 105 women and 194 men, and their ages range between 40 and 95 years old. All the patients had left ventricular systolic dysfunction and had previous heart failures that put them in classes III or IV of New York Heart Association (NYHA) classification of the stages of heart failure.

1. **Visuals, Graphs and Plots**

Chart, bar chart

Description automatically generated

*Figure 1*

Chart, diagram

Description automatically generated

*Figure 2*

Chart, scatter chart

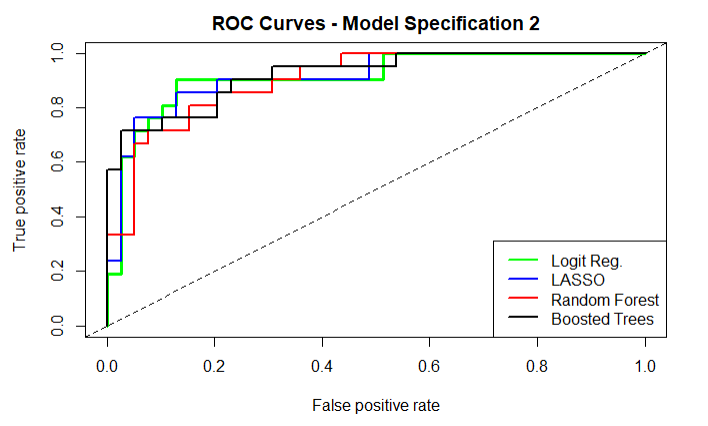
Description automatically generated

*Figure 3*

Chart

Description automatically generated

*Figure 11*

**

*Figure 12*

1. **Quantitative R Output**

Table

Description automatically generated

*Figure 4*



*Figure 5*



*Figure 6*

1. **Cross-Validation Tables**

A picture containing text, receipt

Description automatically generated

*Figure 7 The Logistic Model CV Tables*

A picture containing text, receipt

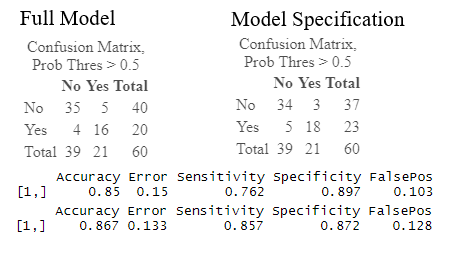
Description automatically generated

*Figure 8 The Lasso Model CV Tables*

Text

Description automatically generated

*Figure 9 Random Forest Model CV Tables*



*Figure 10 Boosted Trees CV Tables*

1. **References**
2. <https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data>
3. Bredy C, Ministeri M, Kempny A, Alonso-Gonzalez R, Swan L, Uebing A, Diller G-P, Gatzoulis MA, Dimopoulos K. New York Heart Association (NYHA) classification in adults with congenital heart disease: relation to objective measures of exercise and outcome. Eur Heart J – Qual Care Clin Outcomes. 2017; 4(1):51–8.