# Heart Failure Death Prediction

Badger Analysts
Shaonan Wang & Yumian Cui

Heart failure has been a vital cause of death in recent years. With the dataset containing 12 health characteristics and collected from nearly 300 patients, we aim to predict whether a person would die of heart failure based on his/her health attributes and which attributes would influence the death event more to offer health advice.

Through data exploration, data modeling, and model evaluation, we found Random Forest to be the model with optimal performance applied to this dataset, and logistic regression comes after closely. With the results of both models combined, we can see that variables of serum_creatinine, age, ejection_fraction, and time are most correlated to the death event variable.

Since predicting whether the patient would die is a classification problem, we choose Logistic Regression and Random Forest as our models. Logistic Regression is a predictive analysis to estimate the dependent binary variables based on one or more nominal, numerical, or ordinal variables. It can provide not only the magnitude of correlations between variables but also the direction of the association (positive or negative). However, it has limitations of assuming linearity between the dependent and independent variables, which is rare in the real-world case.

Different from logistic regression as a probabilistic model, Random Forest is also a commonly used machine learning algorithm for classification problems. It is improved upon the decision tree so as to provide higher accuracy and lowered variance. It is especially known for its nonlinearity which is why we choose the algorithm.

We first import the classifiers from sklearn, get a preliminary 5-fold cross-validation score, and then do the hyperparameter tuning to improve upon the model even more. Then we borrowed a set of metrics to evaluate which model is a better fit. Both models do well, but Random Forest is slightly better. As to the performance measurement for the Random Forest, the final results we get are 88% accuracy, 79% precision, 79% recall, and 93% AUC score. It turns out that the data preprocessing step is equivalently significant as the model building because before the final adoption of the model, we ended up running the code twice--- first with outliers and then without outliers to compare their performance. The version without outliers is seen with accuracy increased by 2-3%.

After data modeling is the feature selection. Based on feature importance generated by the random forest model, we can see preliminarily that serum_creatinine, age, ejection_fraction, and time are more heavily weighted factors. However, we cannot see in what direction it is related to the death_event variable, so we proceeded with conducting analysis by logistic regression. As the proportion of DEATH_EVENT=1 is larger than 5% (32.59%) in the dataset after removing the outliers, we can use the Logistic Regression model to predict the value of DEATH_EVENT. By identifying the correlations between DEATH_EVENT and other features, we choose features highly correlated with the death as different independent variable (x) combinations in the logistic regression modeling. Then we evaluate whether these features are

likely to predict the death based on the accuracy scores of the confusion matrix between actual dependent variable values and predicted values in both testing and training sets.

By choosing groups of features with high positive correlations, we find that the accuracy is around 75% for training sets and the difference of the accuracy between training and testing sets is 5%~7%. Then we divide the testing and training sets into 50/50 instead of 30/70. The training set accuracy is raised to 76%, and the accuracy difference is only 1% between training and testing. This means that we can increase the sample size to train the model for more accurate predictions. We also choose a group of features highly negatively related to the death event, and the training set accuracy increases to 82%. Then features highly positively or negatively related to the death event are combined together, and the model accuracy remains around 82%. This shows that our model can more accurately predict factors preventing heart failure death, and identify features highly correlated to the heart failure death.

In conclusion, serum_creatinine, age, ejection_fraction, and time are the features most affect heart-failure death. Older people with higher levels of serum creatinine in the blood are more likely to die of heart failure, and people with a higher percentage of blood leaving the heart at each contraction and follow-up more frequently are less likely to induce heart failure death. Other features like high blood pressure and smoking are likely to cause heart-failure death as well. Therefore, in order to prevent the heart-failure death, we would suggest older people monitor the levels of serum creatinine and blood pressure for lower values and exercise more to strengthen the heart muscles to push more blood out of the heart.