Problem Statement

Heart disease is the leading cause of death in the United States as well as globally. Annually, it accounts for around 31% of all deaths worldwide and costs billions of dollars, an astounding economic burden. Heart failure occurs when the heart is unable to pump or fill with blood efficiently leading to symptoms such as shortness of breath, fatigue, rapid heartbeat, swollen legs, and it can eventually lead to death.

There are a variety of risk factors for developing heart failure. These include obesity, diabetes, high blood pressure, coronary artery disease, heart attack, among others. There are measurements of heart health and vitality that can be monitored for patients that are at risk for heart failure such as the ejection fraction, or amount of blood that is pumped out by the heart or the creatine phosphokinase level, which can indicate a muscle tissue injury and is used in diagnosing heart attacks.

For this project I analyzed data provided in a publicly available dataset recording heart failure outcomes and clinical measurements for 299 patients in Pakistan that were admitted to the Institute of Cardiology in 2015. By looking at the data provided, I found that the most important measurements to review when predicting whether a patient will experience a negative health outcome are serum creatinine levels, the heart ejection fraction, age, creatinine phosphokinase levels, platelet levels and serum sodium levels.

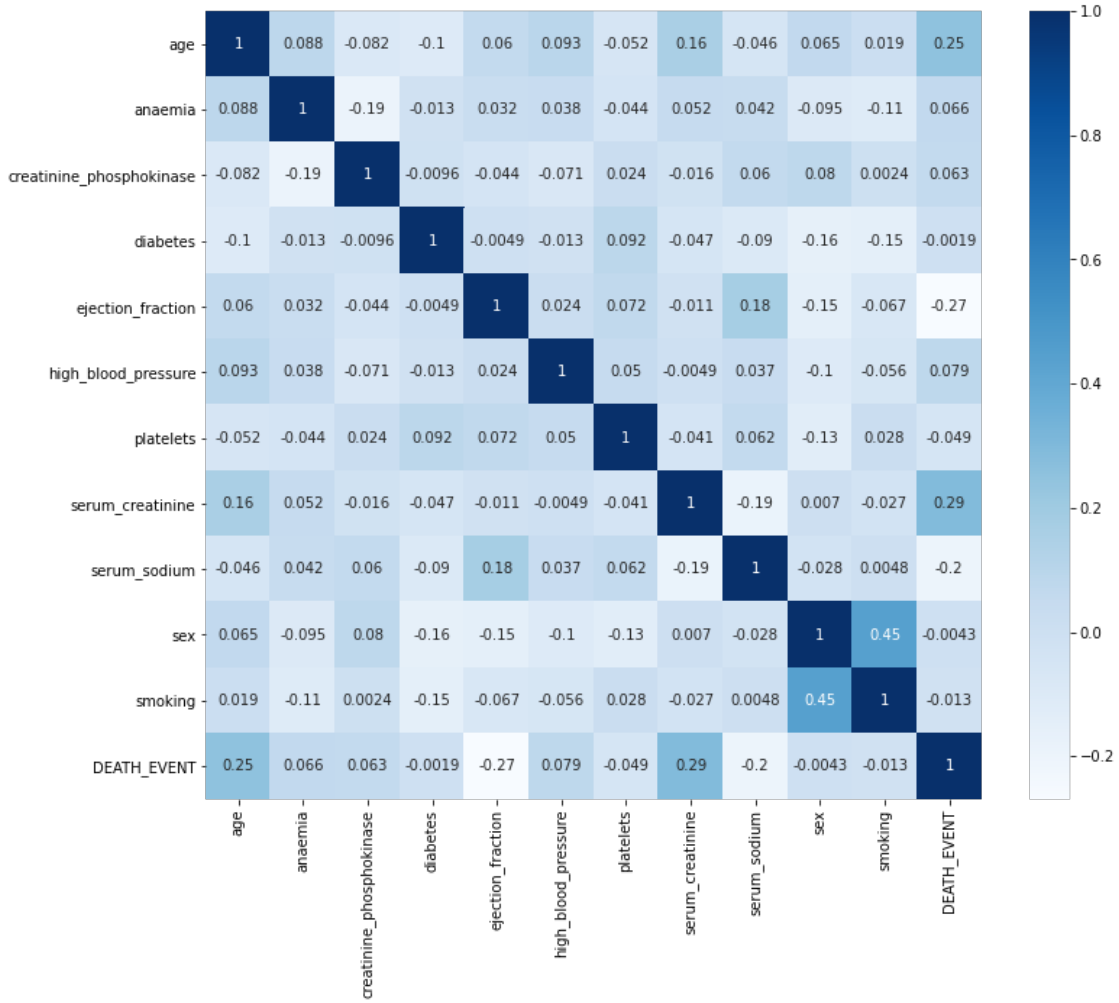
The tuned Decision Tree Classifier I used to model my predictions had an accuracy level of 0.73.

Data Wrangling

The dataset I analyzed was very clean so I conducted minimal data cleaning. I examined the data for any missing information and did not find any. I dropped the time column as it correlated with the death outcomes data exactly and thus did not provide any insight.

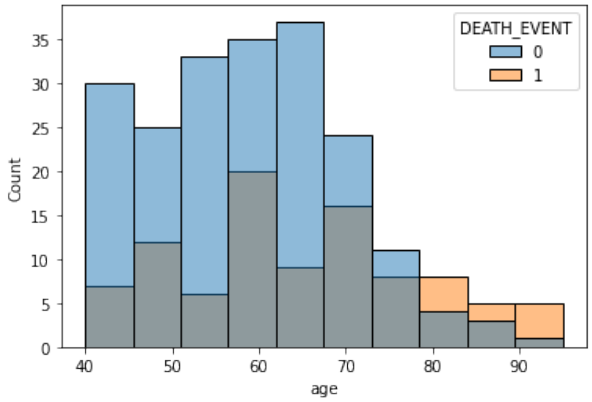
Exploratory Data Analysis

I initially plotted my data to see which variables seemed to have a strong relation to the death outcome variable. The scatter plots and histograms I initially created were not particularly illuminating. I plotted the variables as a heat map and found there were several variables that were correlated with the death outcome variable – notably, age and serum creatinine. Serum sodium and ejection fraction had a strongly negative correlation as well, which I thought was worth examining in my analysis.

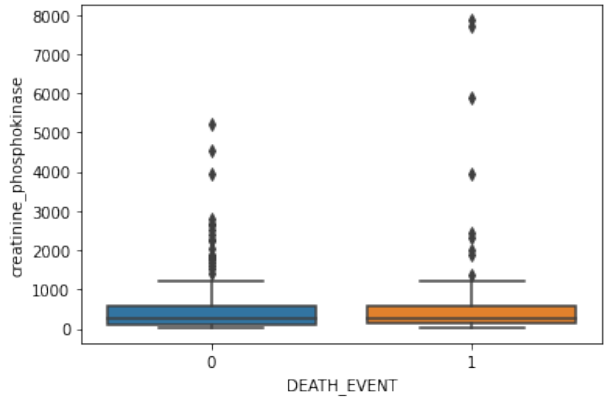


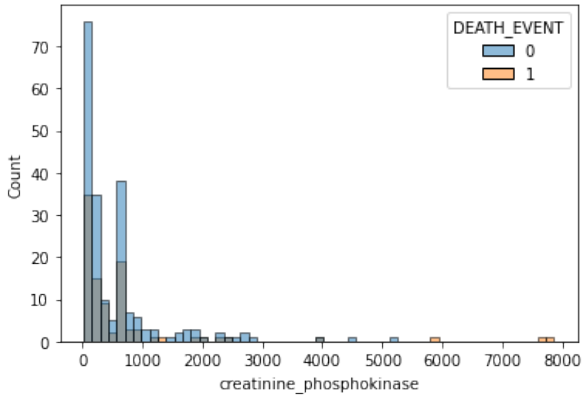
I created histograms and boxplots for variables I was most interested in examining.

Age

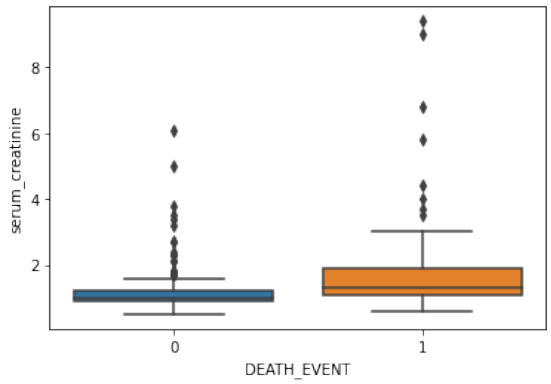


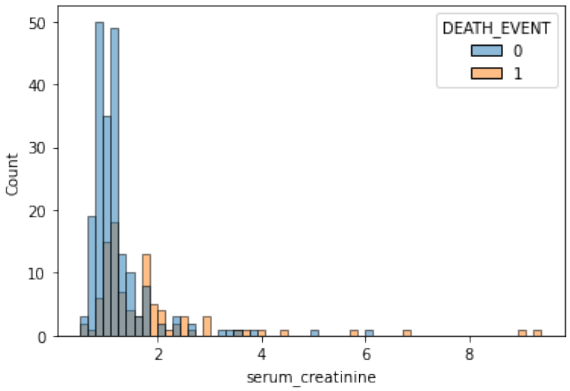
Creatinine Phosphokinase Levels



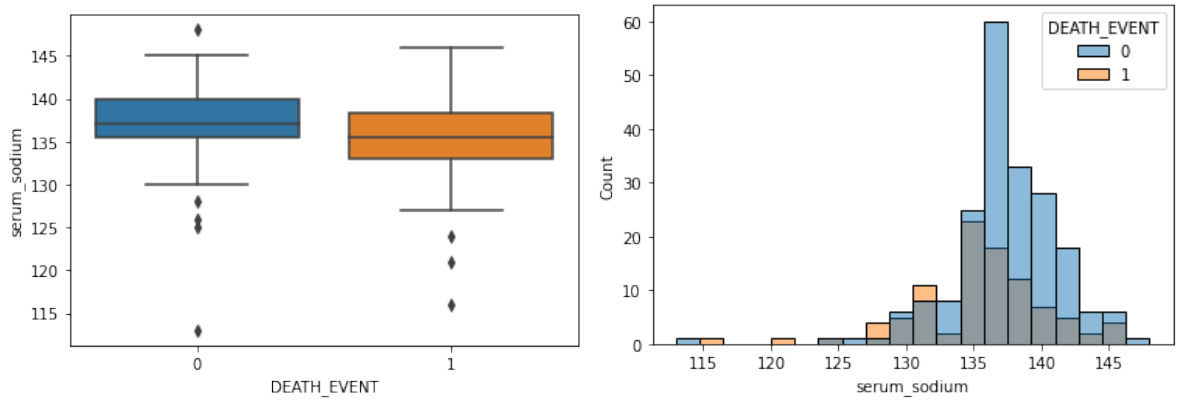


Serum Creatinine Levels





Serum Sodium Levels



Based on the exploratory data analysis, I was most interested in seeing how the age, serum creatinine levels, and ejection fraction affected the death event outcome. It appears that older patients are more likely to die, and older men are more likely to die than older women. Increased age seems to be associated with higher rates of death, lower levels of serum creatinine seem to be associated with lower rates of death, and a higher ejection fraction seems to be associated with lower rates of death. According to the box plot and histograms I created, higher levels of serum sodium may also be associated with lower rates of death.

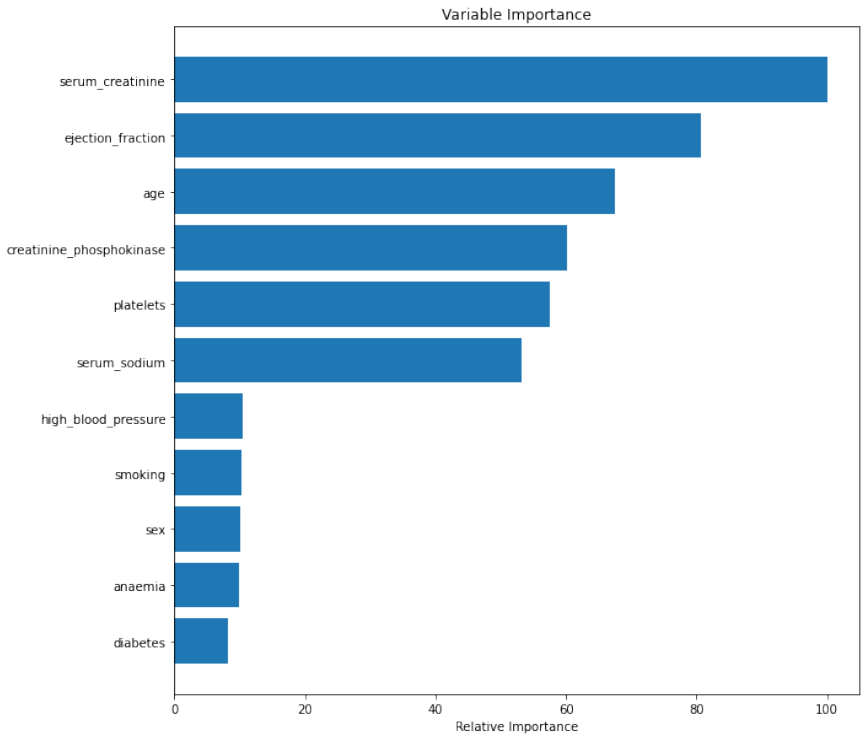
Model Selection

For my model selection, I tested 3 machine learning classification models: Logistic Regression, Random Forest Classifier and Decision Tree Classifier. I wanted the model to accurately predict instances of a death event outcome. Before analyzing my models I applied a standard scaler in order to normalize my data as several of the variables such as platelets, creatinine phosphokinase levels, and serum creatinine were a different order from the others.

Based on this table, the most important features for the Logistic Regression are age, serum creatinine levels, ejection fraction, and serum sodium levels. Serum sodium and ejection fraction are the most negatively correlated with a death event outcome as their coefficient value is less than one.

	coef
age	1.064225
serum_creatinine	1.006299
high_blood_pressure	1.000871
anaemia	1.000793
creatinine_phosphokinase	1.000202
diabetes	1.000028
platelets	1.000000
smoking	0.999869
sex	0.999805
serum_sodium	0.982212
ejection_fraction	0.940943

For the Random Forest Classifier, the most important features are serum creatinine levels, ejection fraction, age, creatinine phosphokinase levels, platelet levels, and serum sodium levels, shown in the graph below.



For the Decision Tree Classifier, the most important features are serum creatinine levels, ejection fraction, and age.

I evaluated the performance of the three models using the cross-validation score, specifically the average cross validation score when using 8 folds. I tested two separate Decision Tree Classifiers, one using entropy as the criterion and the other using gini. These both performed poorly however. I also used a Decision Tree Classifier with entropy as the criterion and a max depth of 3, which did not perform quite as well as the Random Forest Classifier, but is computationally less expensive. The difference in accuracy for the two models was 0.7359 for the Random Forest Classifier and 0.7325 for the Decision Tree Classifier. The Logistic Regression had an accuracy of 0.7255, so the most accurate and simultaneously least computationally expensive model is the Decision Tree Classifier.

Based on this information, it might be proposed that healthcare systems can accurately predict the outcome for heart failure patients based on a few measurements, such as age, ejection fraction and serum creatinine levels when using the Decision Tree model and thus make decisions about how closely patients need to be monitored and how aggressively they need to be treated, etc.

Future Research

I think applying this model to a larger dataset could provide more insight into the importance of these variables and allow health providers to make more informed decisions about patient care, treatment and monitoring.

Health outcomes in patients with cardiovascular disease vary greatly from state to state in the US and vary between countries as well. Future research into regional and global differences would also increase the accuracy and robustness of the model.