Danielle Saad, Mohammed Ali, Enrique Escobar, Sergey Khegay, Zhehui Zang, Shao Ru Zhang
CIS/STA 3920
Professor Deng
13 December 2020

Final Report

Introduction and Background

The dataset we analyzed is titled "Heart Failure Predictions", and it was originally written by Davide Chicco and Giuseppe Jurman.[1] Our data source is from Kaggle. The dataset was intended to study how machine learning can predict the survival of patients with heart failure from serum creatinine and ejection fraction[2]. Their research is published on BMC Medical Informatics and Decision Making website, which is an open-access journal that publishes original peer-reviewed research articles for health information technologies. Heart failure is the inability of the heart to pump blood, causing malfunctions of other organs due to lack of oxygen and blood. It is a chronic and progressive condition where organs start to fail one by one, eventually leading to death. Some conditions that lead to heart failure cannot be reversed, however treatments can improve signs and symptoms of heart failure, increasing one's life expectancy. Our goal is to utilize the 12 risk factors in the dataset to determine which variable is the most significant in determining whether a person will get heart failure.

In our hypothesis, we predict that diabetes, high blood pressure, and chronic smoking will be significant risk factors in determining heart failure. According to the Mayo Clinic, conditions such as heart attack, high blood pressure (hypertension), faulty heart valves, damage to the heart muscle (cardiomyopathy), birth heart defect (congenital heart defects), abnormal heart rhythms (heart arrhythmias), diabetes, HIV, smoking, obesity, and other risk factors can all damage and weaken the heart, leading to heart failure[3].

Research question

Which risk factors play an important role in determining heart failure?

Research Motivation

Cardiovascular disease is a class of heart conditions that involves the heart and blood vessels, it is synonymous with the term heart disease[4]. It has become one of the most serious and common diseases in recent years. It is the number one cause of death in the United States, with 18.2 million Americans having some form of cardiovascular disease. According to data from the Center for Disease Control and Prevention, "heart disease is the leading cause of death for people; one person dies every 36 seconds in the United States from cardiovascular disease; about 655 thousands Americans die from heart disease each year—that's 1 in every 4 deaths.[5] In 2020, approximately 6.2 million people have heart failure in the United States.[6]" In addition to the number of patients and fatalities, the increase in the scale of medical

---

[1] Larxel. "Heart Failure Prediction." *Kaggle*, 20 June 2020, www.kaggle.com/andrewmvd/heart-failure-clinical-data?fbclid=IwAR119wsqhy87c9M2nOqmHcLATRii6Lp2-Jtc6OlhjsonRjnPU33adhDC9Dg.

[2] 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*. 2020 Feb 3;20(1):16. doi: 10.1186/s12911-020-1023-5. PMID: 32013925; PMCID: PMC6998201.

[3] "Heart Failure." *Mayo Clinic*, Mayo Foundation for Medical Education and Research, 29 May 2020, www.mayoclinic.org/diseases-conditions/heart-failure/symptoms-causes/syc-20373142.

[4] "What Is Cardiovascular Disease?" *www.heart.org*, 31 May 2017, www.heart.org/en/health-topics/consumer-healthcare/what-is-cardiovascular-disease.

[5] "Heart Disease Facts." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 8 Sept. 2020, www.cdc.gov/heartdisease/facts.html.

[6] "Heart Failure." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 8 Sept. 2020, www.cdc.gov/heartdisease/heart_failure.html.

expenses also follows. Therefore, it is important to find a way to suppress the scale of heart disease by conducting a data analytics study to figure out the extent to which different risk factors contribute to heart failure. Our research aims to find the relationship between age, anaemia, diabetes as well as other factors, hoping to give people the guidance of controlling their health and allowing medical staff to conduct effective treatment based on our research.

Dataset Description and Variable Introduction

The dataset that was used for the following research contains 299 heart failure patient's medical records. It was collected at the Faisalabad Institute of Cardiology and the Allied Hospital in Faisalabad (Punjab, Pakistan), throughout April to December 2015[7]. The data consists of 105 Women and 194 men ranging between 40 and 95 years old. Furthermore, the dataset consists of 13 features related to clinical and lifestyle information about the patients. The 13 features that were included were binary response: sex (0=Women and 1=Men), anaemia, high blood pressure, diabetes, smoking, and **death event** which all take a boolean measurement (0=No/False and 1=Yes/True). Consequently, the death event will be considered as our dependent variable, which asserts if the patient died or survived before or after the end of the follow-up period, and other 12 features as our independent variables.

For a more scientific explanation, **anaemia** is the decrease of red blood cells of hemoglobin. **Diabetes** is a disease in which the body is not able to produce or respond to the hormone insulin, which results in abnormal metabolism and failure to convert food into energy. **High blood pressure** indicates that a patient has hypertension. **Sex** is the gender of a person. **Smoking** is the act of inhaling and exhaling the smoke of tobacco.

Regarding the numeric features, the **creatinine phosphokinase (CPK)** indicates the level of the CPK enzyme in the blood produced when a muscle tissue gets damaged. The **ejection fraction** indicates the percentage of how much blood the left ventricle pumps out with each contraction. The **platelets** indicate the number of small colourless cell fragments that form clots and prevent bleeding, it is measured in kilo-platelets/mL. The **serum creatinine** is a waste product generated by creatine when a muscle breaks down, it is measured in mg/dL. The **serum sodium** indicates if a patient has normal levels of sodium in the blood, it is measured in mEq/L. The **time** registers the time in days passed since the discharge or treatment. Lastly, we have **age,** which is the age of the patient at the time the record was recorded.

Data Summary

By applying the descriptive statistical analysis as shown in Table 1, we can see the fundamental description and quantitative summary of data. Based on which we can conclude the following tendency and variability. In terms of age, on average patients were in their 60s. Creatine phosphokinase on average measured 581.84 mcg/L, which is beyond the normal value range of 10-120 mcg/L. This can be due to age as well as indication to certain medical issues. An ejection fraction has an average of 38.08% than normal 50-70%, denoting that most of the patients have some sort of vascular diseases. Serum creatinine has an average of 1.39 mg/dL, which is a little higher than the normal range of 0.84 to 1.21 mg/dL. Serum sodium has an average of 136.63 mEq/L, which is within 135 to 145 mEq/L. Moreover, on average, most of the patients were discharged after 130 day. Accordingly, by a categorical summary, we can see that about 35-40% of the patients experienced symptoms of anaemia, diabetes, and high blood pressure. On the other hand, 32% of patients were smokers and around the same number died due to heart failure.

To grasp a greater understanding of data in relation to our motivation, the data has been visualized by boxplots in Figure 1 and histograms in Figure 2 below. Meanwhile, the scatterplot matrix was omitted from the following research since no relationships were shown in the process. In reference to

[7] F. Meng, Z. Zhang, et al. "Machine Learning Can Predict Survival of Patients with Heart Failure from Serum Creatinine and Ejection Fraction Alone." *BMC Medical Informatics and Decision Making*, BioMed Central, 1 Jan. 1970, bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5.

Table 2, the following five features including, age, the amount of serum creatinine, serum sodium, the percentage of ejection fraction, and follow-up time are chosen as the significant variables of this research. By generating the boxplots as shown in Figure 1, the distribution of the significant variables can be seen in dispersion shown by five relevant values. The boxplots show age and serum creatine have a positive relationship while, the ejection fraction, serum sodium, and follow-up time indicate a negative relationship.

By plotting a histogram for each of the 13 variables, as shown in Figure 2, we are able to get a better insight into data distribution. For instance, we were able see that the following study has greater number records of patients between the age range of 60-70. Majority diagnosed with relatively normal laboratory values of serum creatinine is between 0-2 mg/dL and serum sodium between 130-140 mEq/L. However, more than half have ejection fractions that are significantly lower than normal. Additionally, symptoms including anaemia and diabetes having relatively slim impact that cause death; likewise, high blood pressure and an instance of smoking having small impact. Lastly, the sex histogram indicates that most of the patients who died due to heart failure were men.

Data Analysis Method Description

The research first employs logistic regression through validation set approach and backward selection to find out the best-fit regression model, then examines the model by double checking on collinearity, and finally verifying our results through the usage of random forest.

To begin our data mining techniques, we used the validation set approach to split the data into test (validation set) and training data. The validation model is fit on the training set, and the fitted model is used to predict the responses for the observations in the validation set. This approach allows us to estimate test error rate, which we then utilized for the improvement of the model accuracy of logistic regression and random forest. We used a logistic regression model with multiple inputs due to a binary output (boolean 0/1). Specifically, for the binary outputs, the research measures whether the DEATH_EVENT variable measures 0=False/No and 1=True/Yes, for whether the patient died or survived. There will be two logistic regressions in our final output, one with all variables, and the other with significant variables obtained through backward selection.

We used logistic regression with a backward selection method to help make a good inference. Logistic regression with backward selection allows us to eliminate insignificant variables based on the threshold $p>0.10$ one at a time. We then compared the model accuracy of the two regressions to get an idea of the best-fit regression model. Since we are conducting inference research, logistic regression with backwards selection allows us to produce significant variables that have an impact on our dependent variable, rather than focusing on increasing the accuracy rate of a good prediction.

Additionally, collinearity is checked in the logistic regression model by employing the VIF function in R, which will indicate the existence of colinearity if the VIF value is greater than 5 or 10. Checking collinearity will enhance the examination on the accuracy and comprehensiveness of the logistic regression model.

Finally, random forest is a classification algorithm that consists of numerous decision trees that merge the trees together to obtain a more accurate and stable prediction. Random forest is not only an accurate algorithm, but it is also the go to method when dealing with classification. Thus, using a random forest for this project will provide a higher accuracy of our model. It also identifies the importance of each variable and to what degree of importance it has on the model, as shown in Figure 3.

Method Evaluation

In the logistic regression method, we used the validation set approach by splitting our data set into training and test sets. We first ran the logistic regression in the training dataset by using the glm function, it included all of our 12 independent variables. Then, we used the predict function on our test set to predict the probabilities. We found that the accuracy rate of our logistic regression is 88%, with an error rate of 12%, these results are very assuring for our research. It shows that 132 out of 150 observations from our test set were correctly predicted, further ensuring the model we are using is fairly accurate.

Furthermore, to continue ensuring that our model is accurate, we decided to make a smaller logistic regression model that contained only the variables that were significant. We did so by implementing the backward selection method. This method allowed us to remove the insignificant variables with largest p-values one by one until we reached a stopping threshold of less than or equal to 0.10. By doing this, it removes variables that have high standard error. Standard error is the averaged distance that the observed value falls from the logistic regression. As shown in Table 2, variables such as platelets and creatinine phosphokinase have low standard error values, which is great, but have a really high p-value and are eliminated by the backward selection. In the end, we concluded that age, ejection fraction, serum creatinine, serum sodium and time are the most significant variables in our model.

After we discovered our significant variables, we used logistic regression and found the accuracy rate to be 86.7% and error rate to be 13.3%. Although the accuracy decreased slightly from the full logistic regression model, it is acceptable because our goal is to make a good inference and interpretation with the significant variables of the data, rather than focusing on increasing the accuracy rate of a good prediction. The decrease in accuracy rate can be due to the limitations posed by our data. According to Table 2, the residual deviance increased from 116.8 with all 12 of our variables to 119.25 with only 5 of our independent variables. An increase in residual deviance means there is a lack of fit within our model, but since it only increased 2.5, the effect is not that severe. It also explains why our accuracy rate for the small model is lower than the larger model with all 12 independent variables. We then used VIF to check for any collinearity for our significant variables. As seen in Table 3, our five significant variables have a VIF value between 1.05 and 1.27 which means there is no evidence of problematic amounts of collinearity because VIF values are less than 5 or 10.

For random forest, the out of bag error estimate is seen as an unbiased test, and our OOB error rate was 19.46% which means it has an accuracy rate of 80.54%. When the forest is built, each tree is tested with samples that are not used in building that tree, this allows an internal error estimate that is averaged for all trees produced in a random forest, creating the OOB error estimate.

Conclusions

In conclusion, we have determined that the most important variables for predicting death by heart failure are age, ejection fraction, serum creatinine, serum sodium and time, according to the statistical methods discussed above. Originally, we expected risk factors such as diabetes, chronic smoking and high blood pressure to have a greater impact of experiencing heart failure based on our research on heart failure. Based on the model and summary statistics, we interpret that higher age, lower ejection fraction, higher serum creatinine, lower serum sodium, and lower follow-up time are good indicators of an event of death due to heart failure. In addition, our analysis found that time holds the most significance and importance over the remaining risk factors. Therefore, the time variable has the biggest effect on death by heart failure, followed by serum creatinine, ejection fraction, serum_sodium, and then age.

Our results make sense because having a lower time, which is the number of days passed before following up on treatment or discharge, indicates that original treatment was not successful and patients have a higher chance of dying from heart failure. Having a higher serum creatinine indicates that there are more waste products generated by creatine when a muscle breaks down, so this is a sign of heart failure since not enough blood and oxygen are being supplied to organs to function properly. Lower ejection fraction indicates that a lower percentage of blood is being pumped out of the left ventricle with each contraction. This leads to heart failure symptoms of shortness of breath, fatigue, weakness, swelling, and irregular heartbeat. These symptoms signal that the body is about to experience heart failure. Lower serum sodium indicates that a patient has a lower than normal amount of sodium in the bloodstream, increasing the chance of death due to heart failure. Lastly, higher age makes sense because the older you get, the more health complications a person experiences, and the harder the body and heart has to work to keep a person alive.

As a result, we reject our initial hypothesis. However, these findings are limited to the research methods, data set we used, and the data/variables we have accessibility to. For example, the validation estimate of the test error rate can be highly variable, depending on which observations are included in the

training set and which observations are included in the test (validation) set. In addition in the validation set approach, only a subset of the observations are used to fit the model. Since statistical methods tend to perform worse when trained on fewer observations, validation set error rate may tend to overestimate the test error rate for our model fit on the entire data set. Additionally, there may be other variables not included in the dataset which may better predict death by heart failure. More research in this field is definitely needed in order to help our medical professionals provide us with their best care. With that in mind, we believe our model is still a good fit, and medical professionals may monitor these risk factors with more caution when dealing with patients with heart failure.

Practical Implications

The practical implications from examining this research question to predict the patient's survival and classify the most important factors that lead to heart failure is the ability to help save lives, prevent further complications, and save money. By creating a model that enables early detection of potential heart failure, patients are able to seek out early treatment to reduce the risk of heart failure. There are many types of medications and steps that individuals can take like lifestyle changes and needed medication. By identifying the most important risk factors associated with heart failure, people can make the necessary adjustments to either control or reduce those risk factors for heart failure. Additionally, early detection of heart failure can avoid the severe complications that transpire if one does experience heart failure. Complications like kidney failure, heart valve problems, liver damage, and heart rhythm problems are a few life-threatening complications that arise and can be prevented.[8] Lastly, the cost of health services and medication to treat heart failure is expensive and cost the United States approximately 30.7 billion dollars in 2012. In 2015, the total cost of treatment of heart diseases in the United States reached 215 billion dollars.[9] The value in which this machine learning application can bring to the world is tremendous and can provide medical professionals crucial insights.

References:

[8] Pruthi, Sandhya, et al. "Heart Failure." *Mayo Clinic*, Mayo Foundation for Medical Education and Research, 29 May 2020, www.mayoclinic.org/diseases-conditions/heart-failure/symptoms-causes/syc-20373142.

[9] "Heart Failure." Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 8 Sept. 2020, www.cdc.gov/heartdisease/heart_failure.htm.

Figure 1: Boxplots of Significant Variables

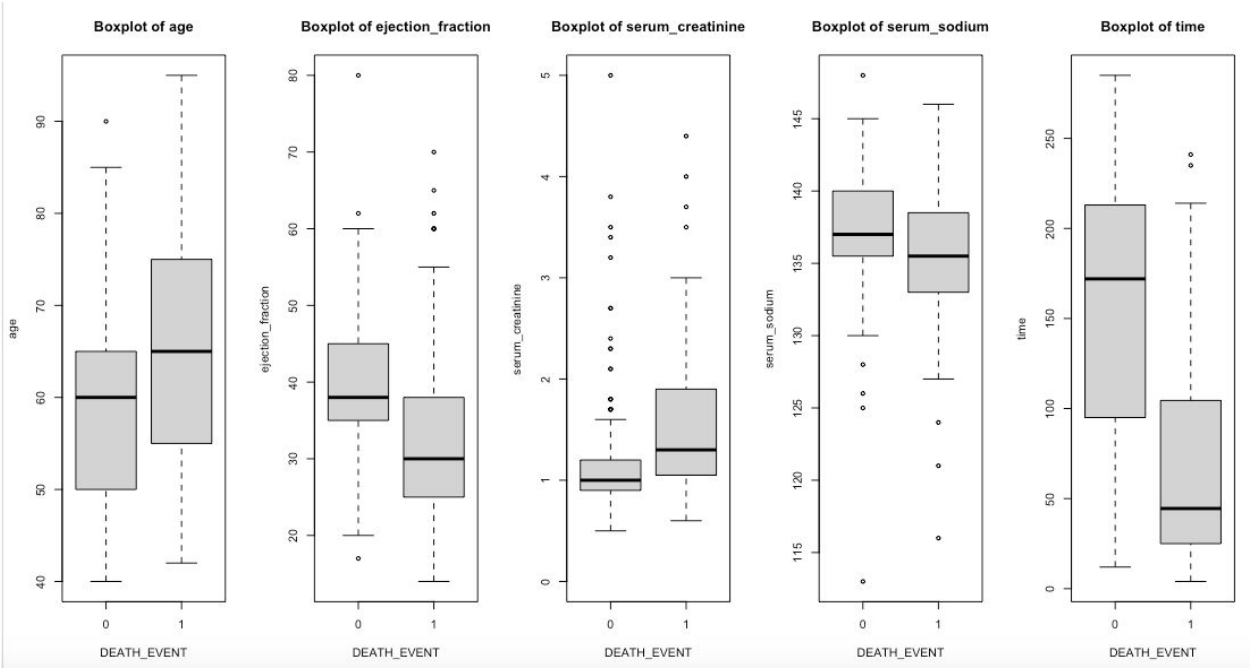
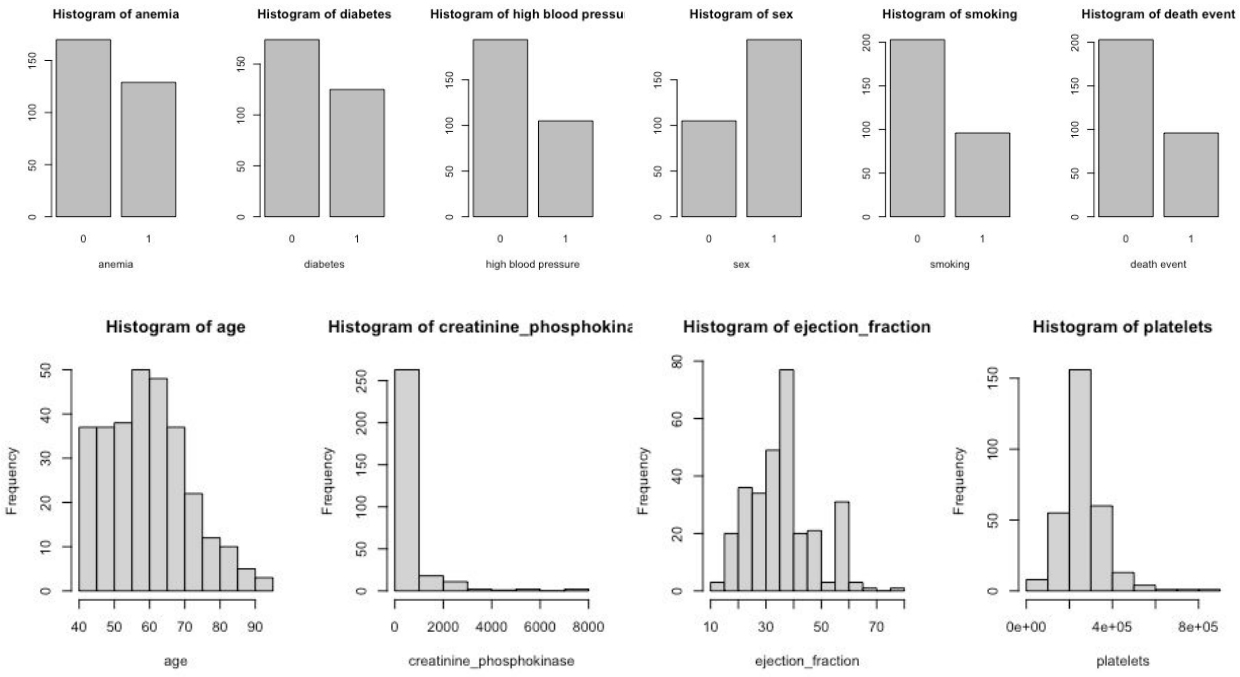


Figure 2: Histogram of Full Model Variables



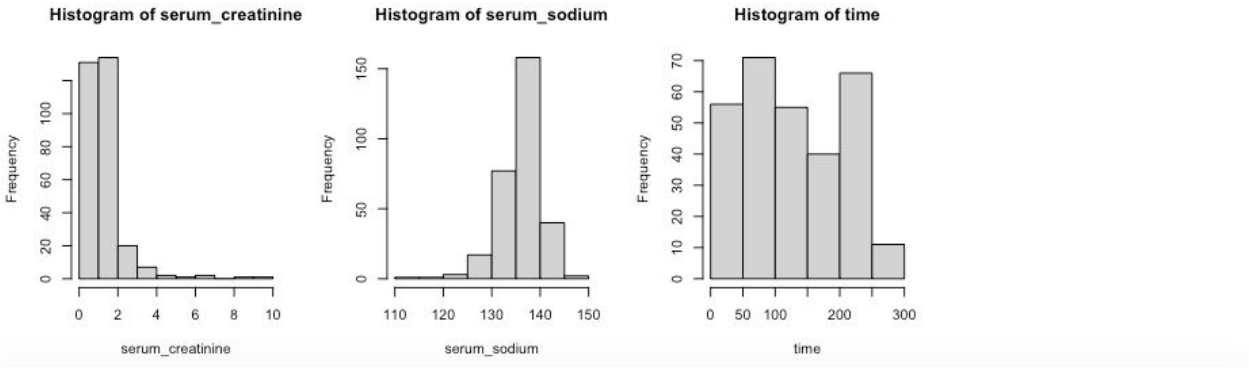


Figure 3: Variable Importance Plot

rf.Heart

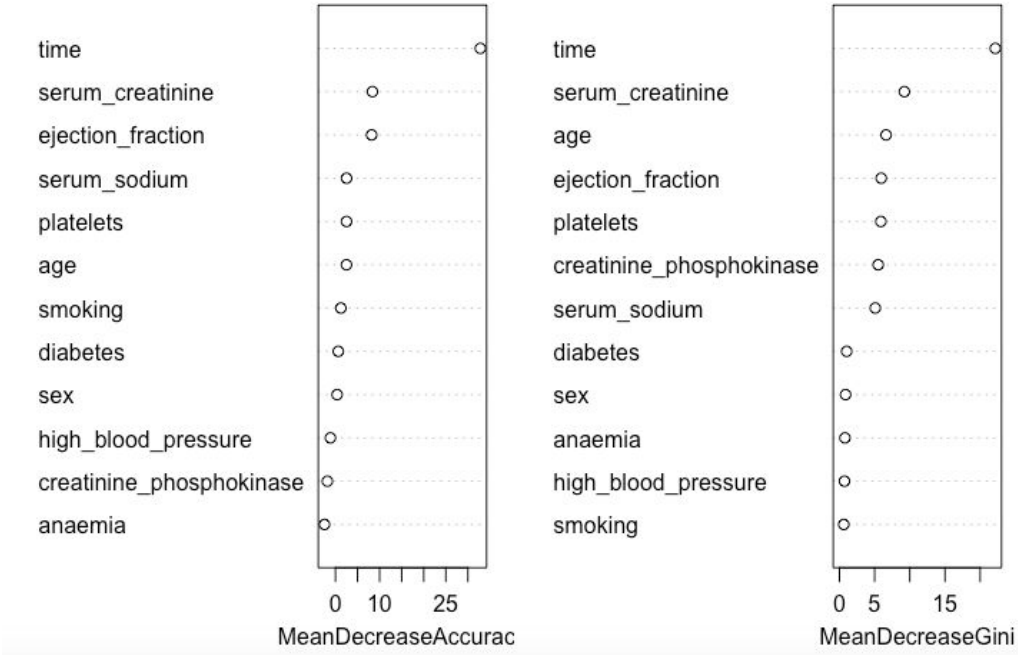


Table 1: Summary Results

Numeric Variables	Mean	Median	St. Dev.	Min	Max
Age	60.83	60	11.89	40	95
Creatinine_phosphokinase	581.84	250	970.29	23	7861
Ejection_fraction	38.08	38	11.83	14	80
Platelets	263358.03	262000	97804.2	25100	850000
Serum_creatinine	1.39	1.1	1.03	0.5	9.4
Serum_sodium	136.63	137	4.41	113	148
Time	130.26	115	77.61	4	285
Categorical Variables	**Yes (1)**	**No (0)**			
Anaemia	129	170			
Diabetes	125	174			
High_blood_pressure	105	194			
Smoking	96	203			
Death_event (DV)	96	203			
Categorical Variables	**Male (1)**	**Female (0)**			
Sex	194	105			

Table 2: Logistic Regression with Backward Selection Method Results

	Dependent variable: Death_Event							
	rating							
	OLS							
Variables:	1	2	3	4	5	6	7	8
Intercept	1.49E+01 .	1.47E+01 .	1.50E+01 .	1.50E+01 .	1.57E+01 .	1.52E+01 .	1.49E+01 .	1.46E+01 .
	(8.81E+00)	(8.75E+00)	(8.70E+00)	(8.74E+00)	(8.62E+00)	(8.53E+00)	(8.41E+00)	(8.34E+00)
Age	4.31E-02 *	4.35E-02 *	4.36E-02 *	4.35E-02 *	4.15E-02 *	3.95E-02 *	3.90E-02 *	3.79E-02 .
	(2.047E-02)	(2.04E-02)	(2.03E-02)	(2.04E-02)	(1.99E-02)	(1.96E-02)	(1.95E-02)	(1.93E-02)
Ejection_fraction	-6.09E-02 **	-6.08E-02 **	-6.13E-02 **	-6.06E-02 **	-6.18E-02 **	-5.99E-02 **	-5.92E-02 **	-6.11E-02 **
	(2.186E-02)	(2.19E-02)	(2.19E-02)	(2.17E-02)	(2.16E-02)	(2.13E-02)	(2.13E-02)	(2.12E-02)
Serum_creatinine	6.69E-01 **	6.66E-01 **	6.74E-01 **	6.66E-01 **	6.64E-01 **	6.86E-01 **	6.79E-01 **	6.79E-01 ***
	(2.186E-01)	(2.19E-01)	(2.18E-01)	(2.15E-01)	(2.15E-01)	(2.11E-01)	(2.09E-01)	(2.05E-01)
Serum_sodium	-1.03E-01 .	-1.02E-01 .	-1.06E-01 .	-1.07E-01 .	-1.10E-01 .	-1.08E-01 .	-1.07E-01 .	-1.03E-01 .
	(6.139E-02)	(6.09E-02)	(6.01E-02)	(6.03E-02)	(5.99E-02)	(5.95E-02)	(5.87E-02)	(5.80E-02)
Time	-2.10E-02 ***	-2.09E-02 ***	-2.10E-02 ***	-2.06E-02 ***	-2.12E-02 ***	-2.12E-02 ***	-2.07E-02 ***	-2.08E-02 ***
	(4.398E-03)	(4.39E-03)	(4.38E-03)	(4.26E-03)	(4.18E-03)	(4.20E-03)	(4.09E-03)	(4.10E-03)
Creatinine_phosphokinase	1.51E-04	1.53E-04	1.51E-04	1.67E-04	1.64E-04	1.40E-04	1.47E-04	
	(1.994E-04)	(2.00E-04)	(2.00E-04)	(1.96E-04)	(1.94E-04)	(1.90E-04)	(1.91E-04)	
High_blood_pressure	-3.94E-01	-3.72E-01	-3.56E-01	-3.64E-01	-3.69E-01	-3.47E-01		
	(4.954E-01)	(4.89E-01)	(4.87E-01)	(4.87E-01)	(4.85E-01)	(4.84E-01)		
Sex	-4.61E-01	-4.12E-01	-4.08E-01	-3.68E-01	-3.88E-01			
	(5.524E-01)	(5.27E-01)	(5.27E-01)	(5.16E-01)	(5.14E-01)			
Diabetes	2.95E-01	2.80E-01	2.39E-01	2.51E-01				
	(5.054E-01)	(5.02E-01)	(4.92E-01)	(4.91E-01)				
Anaemia	-1.87E-01	-2.24E-01	-1.95E-01					
	(5.287E-01)	(5.15E-01)	(5.09E-01)					
Platelets	-1.05E-06	-1.00E-06						
	(2.441E-06)	(2.46E-06)						
Smoking	1.82E-01							
	(5.833E-01)							
Residual Deviance	1.169E+02	1.170E+02	1.171E+02	1.173E+02	1.175E+02	1.181E+02	1.186E+02	1.193E+02

Table 3: VIF- Checking for Collinearity

Significant Variables	VIF
Age	1.05
Ejection_fraction	1.21
Serum_creatinine	1.19
Serum_sodium	1.09
time	1.27

Table 4: Importance IV Variable in Random Forest

Independent variables:	0	1	Mean Decrease Accuracy	Mean Decrease Gini
Age	-0.92	5.4	2.5	6.63
Anaemia	-2.52	-1.03	-2.48	0.78
Creatinine_phosphokinase	-2.17	-0.47	-1.83	5.51
Diabetes	0.84	-0.29	0.64	1.03
Ejection_fraction	6.23	5.08	8.19	5.93
High_blood_pressure	-2.04	0.55	-1.14	0.71
Platelets	1.7	2.15	2.51	5.89
Serum_creatinine	3.83	8.63	8.37	9.23
Serum_sodium	-0.43	4.66	2.52	5.09
Sex	-1.02	2.04	0.38	0.87
Smoking	0.32	1.79	1.23	0.62
Time	27.41	27.1	32.82	22.15