Predictive Analytics in Healthcare: Heart Failure Patient Insights

Syracuse University, Fall 2022

Joshua Biggs-Bauer, Gianni Conde, Joaquin Rodarte

Introduction

Human beings are one of the most remarkable creatures to inhabit the planet. Their astounding intellectual capabilities and unrivalled determination allow them to achieve most of their deepest desires. However, despite their aptitude for attaining life's wishes, the prospect of death still looms. Although the modernization of medical practices have allowed present day people to live longer, healthier lives than ever before, health complications remain a constant factor.

A health complication that continues to affect many is heart failure. Heart failure manifests from cardiovascular disease and arises when an individual's heart cannot pump an adequate amount of blood to sustain the human body. Vast amounts of medical records detail the symptoms, features, and treatment procedures relating to patient heart failure that could lead to further advancement in cardiovascular research, potentially improving the mortality rates of those afflicted.

There are many stakeholders that would find information regarding heart failure to be useful, including patients and medical professionals. Those currently suffering from some form of cardiovascular disease could view this research as a solution to their heart problems. Those who are deemed at risk of cardiovascular disease could find this information useful by making appropriate lifestyle changes, essentially decreasing their likelihood of future cardiovascular disease diagnoses. Lastly, medical professionals could see it as insight on the current scope of patient heart disease and utilize their findings to cultivate a better future for those afflicted by this devasting illness.

Analysis and Models

About the Data

A data set containing patient medical information regarding heart failure was used in this analysis. The data, titled heart-failure-clinical-data, was collected from April to December of 2015 by the Allied Hospital in Faisalabad and the Faisalabad Institute of Cardiology, both residing in Punjab, Pakistan. The data set was obtained from https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data. It contained information on 299 patients (observations) and 13 clinical measurements (variables). These clinical measurements included age, anemia, creatinine phosphokinase (CPK), diabetes, ejection fraction, high blood pressure, platelets, serum creatinine, serum sodium, sex, smoking, time, and death event. Table 1.1 depicts these clinical measurements' name, accompanied by a brief description. The binary variables in the data, 0 indicates no while 1 indicates yes. In the case of the binary sex variable, 0 indicates female while 1 indicates male.

Table 1.1

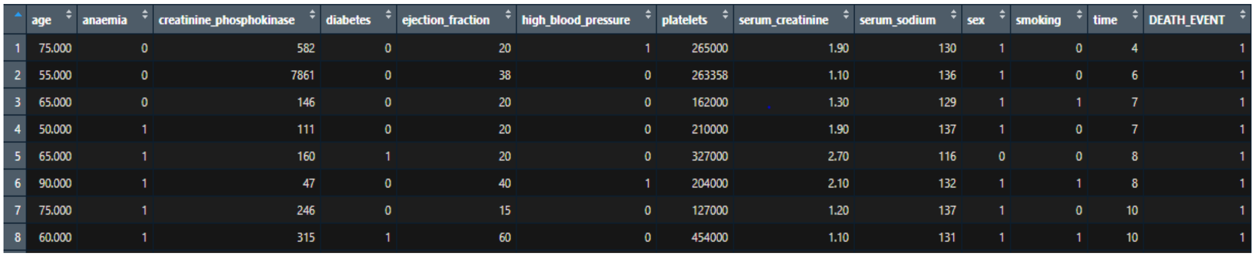
Age	Patient's age (years)	**Sex**	female=0, male=1 (binary)

Anaemia	Decrease of hemoglobin (binary)	**Platelets**	Platelets in bloodstream (kilo platelets/mL)
High blood pressure	Hypertension (binary)	**Serum creatine**	Level of creatinine in bloodstream (mg/dL)
Creatinine phosphokinase (CPK)	Level of CPK enzyme in bloodstream (mcg/L)	**Serum sodium**	Level of sodium in bloodstream (mEq/L)
Diabetes	If a patient has diabetes (binary)	**Smoking**	If the patient smokes (binary)
Ejection fraction	% Of blood leaving the heart at each contraction	**Death Event**	If the patient died during the follow-up period (binary)

Data Preparation and Cleaning

Upon inspection of the data, it appears that there are 299 rows, representing the amount of patients, and 13 columns, indicating the variables representing clinical measurements relating to heart failure. Below is a depiction of the first few instances of the data (Table 1.2).

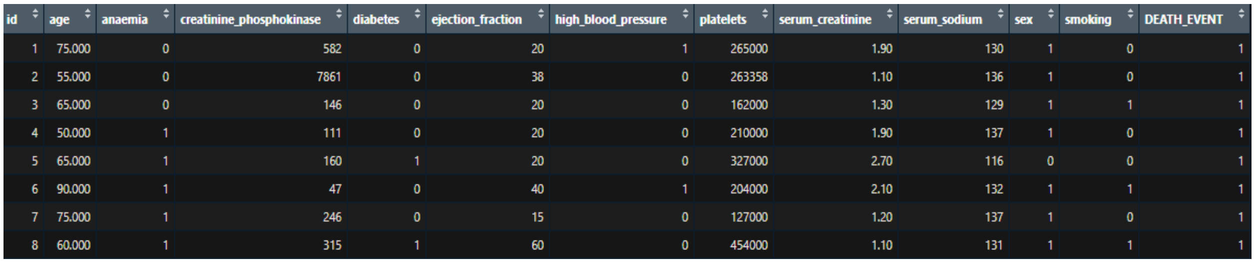
Table 1.2



	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	time	DEATH_EVENT
1	75.000	0	582	0	20	1	265000	1.90	130	1	0	4	1
2	55.000	0	7861	0	38	0	263358	1.10	136	1	0	6	1
3	65.000	0	146	0	20	0	162000	1.30	129	1	1	7	1
4	50.000	1	111	0	20	0	210000	1.90	137	1	0	7	1
5	65.000	1	160	1	20	0	327000	2.70	116	0	0	8	1
6	90.000	1	47	0	40	1	204000	2.10	132	1	1	8	1
7	75.000	1	246	0	15	0	127000	1.20	137	1	0	10	1
8	60.000	1	315	1	60	0	454000	1.10	131	1	1	10	1

To ensure that the data is suitable for analyses, aspects of the data had to either be inspected, manipulated, or added on. Luckily, the data set contained no missing values. However, the time variable proved to be less crucial to the analyses than expected, resulting in the column's removal. Additionally, a patient ID index column was included to solidify the data, which can be viewed below (Table 1.3).

Table 1.3

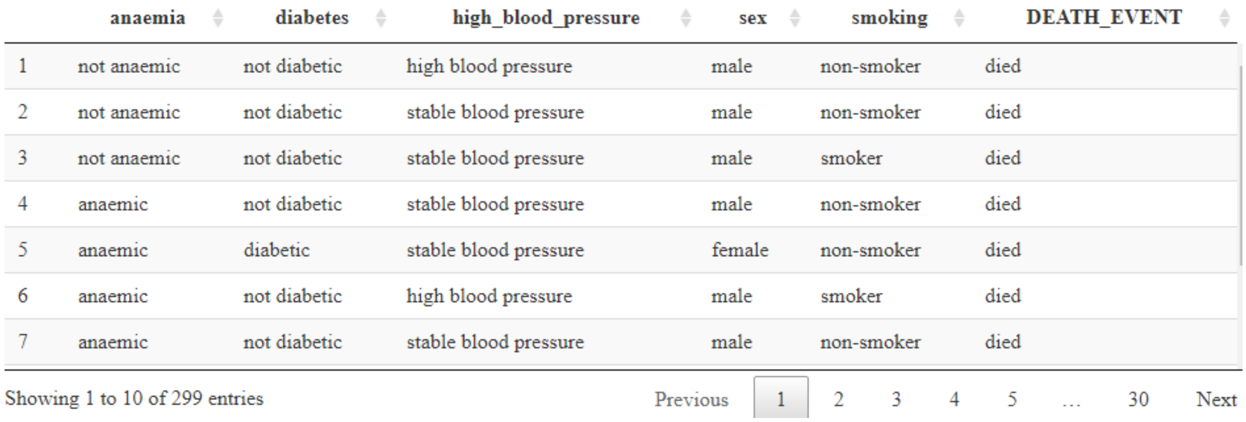


id	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	DEATH_EVENT
1	75.000	0	582	0	20	1	265000	1.90	130	1	0	1
2	55.000	0	7861	0	38	0	263358	1.10	136	1	0	1
3	65.000	0	146	0	20	0	162000	1.30	129	1	1	1
4	50.000	1	111	0	20	0	210000	1.90	137	1	0	1
5	65.000	1	160	1	20	0	327000	2.70	116	0	0	1
6	90.000	1	47	0	40	1	204000	2.10	132	1	1	1
7	75.000	1	246	0	15	0	127000	1.20	137	1	0	1
8	60.000	1	315	1	60	0	454000	1.10	131	1	1	1

Furthermore, since this analysis will contain models and methods pertaining to supervised machine learning, additional data preparations pertaining to each specific method will be necessary. For the supervised machine learning methods, the original data set was split into both training and testing data. The training data consisted of 80% of the overall data, while the testing data consisted of 20%.

The data preparations necessary for the unsupervised machine learning methods include creating a new data frame containing only the binary clinical measurements and replacing the binary values of 0 and 1 with appropriate character descriptions. A portion of this binary data set is pictured below (Table 1.4).

Table 1.4



	anaemia	diabetes	high_blood_pressure	sex	smoking	DEATH_EVENT
1	not anaemic	not diabetic	high blood pressure	male	non-smoker	died
2	not anaemic	not diabetic	stable blood pressure	male	non-smoker	died
3	not anaemic	not diabetic	stable blood pressure	male	smoker	died
4	anaemic	not diabetic	stable blood pressure	male	non-smoker	died
5	anaemic	diabetic	stable blood pressure	female	non-smoker	died
6	anaemic	not diabetic	high blood pressure	male	smoker	died
7	anaemic	not diabetic	stable blood pressure	male	non-smoker	died

Showing 1 to 10 of 299 entries

Data Exploration

Of the 13 clinical measurements, seven are considered to be nominal while six are binary. The nominal variables include id, age, creatinine phosphokinase (CPK), ejection fraction, platelets, serum creatinine, and serum sodium. The binary variables include anaemia, diabetes, high blood pressure, sex, smoking, and death event.

Two correlation plots were run to determine if there were any potential correlations between the data points. The first correlation plot (Figure 1.1) included all aspects of the data, while the second correlation plot (Figure 1.2) contained all aspects except for time.

Figure 1.1

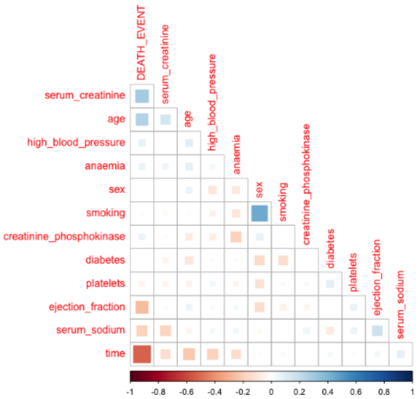


Figure 1.2

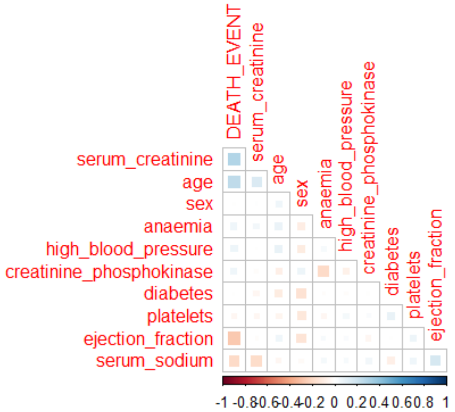
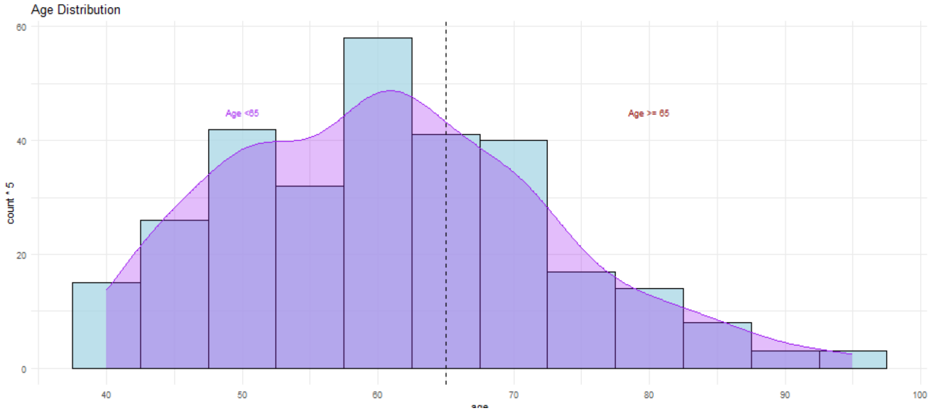


Figure 1.1 shows that smoking and sex has the highest positive correlation, followed by serum creatinine and death event, and age and death event. It also shows that time and death event are extremely negatively correlated. This implies that time could have miniscule relevance to the analyses, leading to its removal from the data frame. Figure 1.2 also shows that serum creatinine and death event are positively correlated, followed by age and death event, serum age and serum creatinine, and serum sodium and ejection fraction. The most negatively correlated data points in Figure 1.2 are ejection fraction and death event, implying that a patient's ejection fraction has little effect on their mortality rate.

Figure 1.3 was generated to graphically visualize the distribution of age and death event. The histogram shaded in blue displays the layer of patient age and the purple overlapping portion of the graph represents the distribution of death events among patients belonging to their age groups.

Figure 1.3

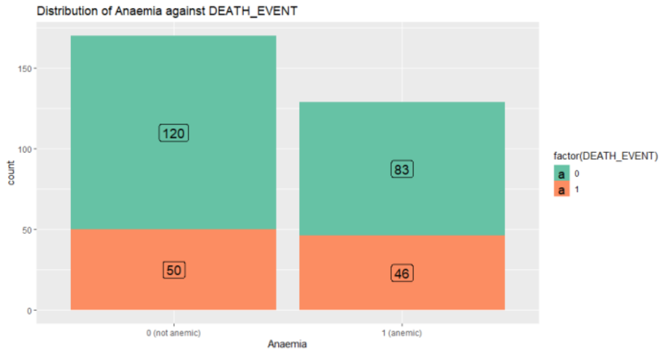


Since there were no participants under the age of 40 and over the age of 95, the overlapping distribution of death events among patient ages begin and end at those points.

Distributions of various clinical measurement against death event were visualized.

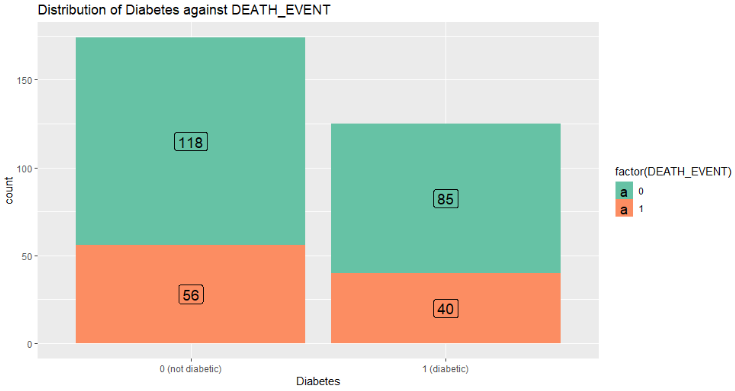
According to the distribution of anemia against death event (Figure 1.4), it appears that of the 299 patients, 170 were not anemic while 129 were anemic. Of the non-anemic patients, 120 survived and 50 died. Of the anemic patients, 83 survived and 46 died.

Figure 1.4



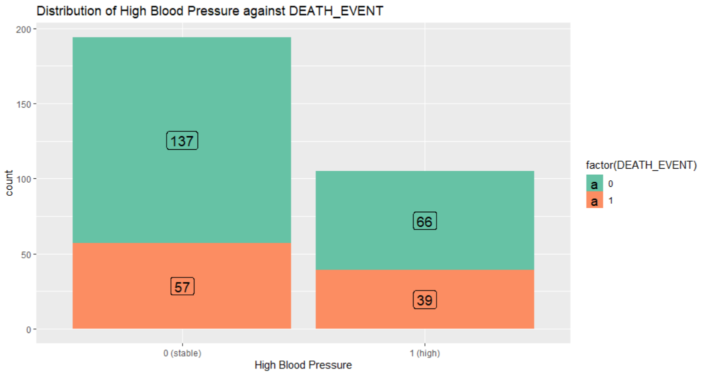
According to the distribution of diabetes against death event (Figure 1.5), 174 were non-diabetic and 125 were diabetic. Of the non-diabetic patients, 118 survived and 56 died. Of the patients who were diabetic, 85 survived while 40 died.

Figure 1.5



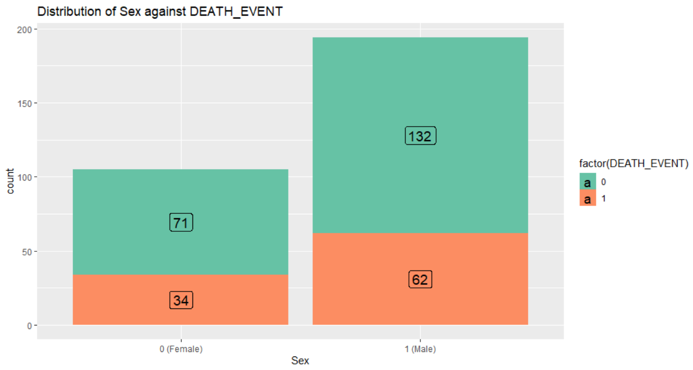
According to the distribution of high blood pressure against death event (Figure 1.6), 194 had stable blood pressure and 105 had high blood pressure. Of the patients with stable blood pressure, 137 survived and 57 died. Of the patients with high blood pressure, 66 survived while 39 died.

Figure 1.6



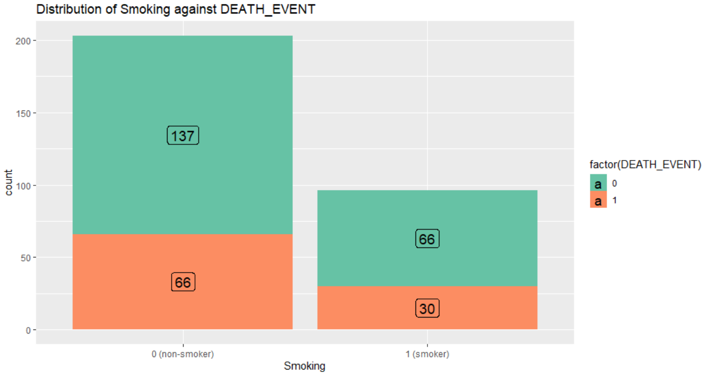
According to the distribution of sex against death event (Figure 1.7), 105 were female while 198 were male. In terms of the female patients, 71 survived while 34 died. In terms of the male patients, 132 survived while 62 died.

Figure 1.7



According to the distribution of smoking against death event (Figure 1.8), 203 were non-smokers while 96 were smokers. Regarding non-smokers, 137 survived while 66 died. Regarding those who were smokers, 66 survived while 30 died.

Figure 1.8



Next, the distribution of the data for patient age (Figure 1.9) revealed a range of 51 between the lowest and highest values present. The lowest age recorded is 40 years old and 91, the oldest age recorded and an outlier in the dataset.

Figure 1.9

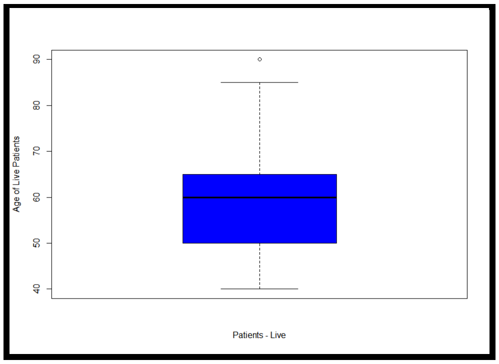
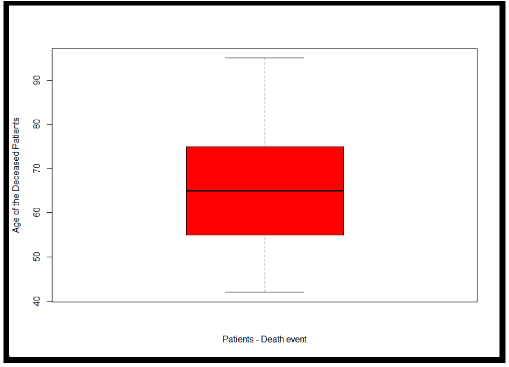


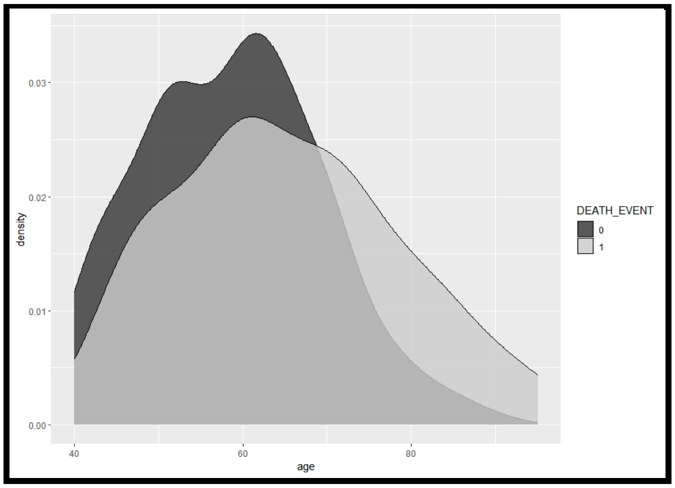
Figure 1.10



The patient death event data subset actually reveals that one of the outliers that reported a death event above the 90-year age. This is a clear sign that the distribution of the data ranges below 71 and 50 years of age with 40 being the youngest and 91 the oldest. The center of the data hovers around 60 years of age for the live patients (Figure 1.10). The patient live data subset revealed an outliers that reported no data event was of the age of 91. This is a clear sign that the distribution of the data ranges below 71 and 50 years of age with 40 being the youngest and 91 the oldest. The center of this boxplot represents a higher mean for the death event subset around 5 years difference.

Given that age has tendency to report younger patients as less commonly to die from heart failure than older patients (Figure 1.11), the decision tree model analysis may be hampered by the data category as to skew results away from specific medical data. This may include substance levels in the body, lifestyle, or underlying conditions in the dataset which ultimately are best suited to medical treatment.

Figure 1.11



Models and Methods

To analyze the heart failure data, various machine learning techniques will assist in determining which clinical measurements most attribute to death events due to heart failure within the scope of the data set. Methods such random forest, support vector machines (SVMs), decision tree, and association rule mining will be implemented in this analyses. The random forest model will develop a group of decision trees that take the results of the various samples and make a consensus. The SVM model will analyze the heart failure patient data for classification and regression analysis. The decision tree model will display patient variable decisions based on the possible consequences of the clinical measurements in regard to heart failure. Lastly, the association rule mining model will discover rules that will predict the occurrences of clinical measurements based on the occurrences of other variables in the heart failure data set.

Analyses Goals and Parameters

The goal of this analysis is to determine which of the provided clinical measurements will best assist in determining patient death event by heart failure.

Results

Random Forest

The initial random forest model (Figure 2.1), containing a seed set at 9, displayed a consensus of the results of various samples from a group of many decision trees. The confusion matrix (Figure 2.2) results stated that the random forest model had a sensitivity of 97.7%, specificity of 52.6%, precision of 82.4%, Kappa of 0.57, and an accuracy of 83.9%. Thus, the model could predict a death event of about 84% of the time.

Figure 2.1

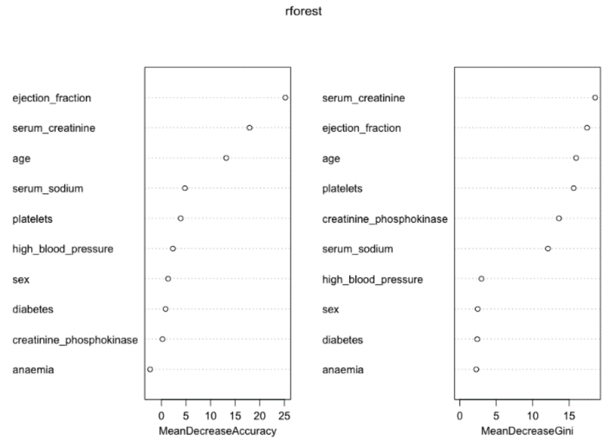
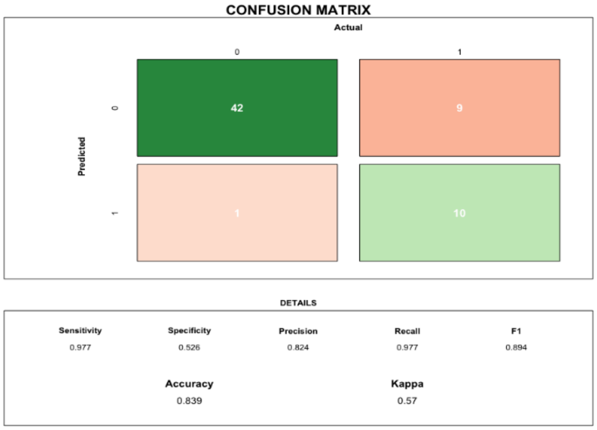


Figure 2.2



The second random forest model (Figure 2.3) contained a seed set at 20 and a data split where the training data held 70% of the data, while the testing data held 30%. The confusion matrix (Figure 2.4) results stated that the random forest model had a sensitivity of 98.2%, specificity of 35.7%, precision of 75%, Kappa of 0.398, and an accuracy of 77.1%. Thus, the model could predict a death event of about 77% of the time.

Figure 2.3

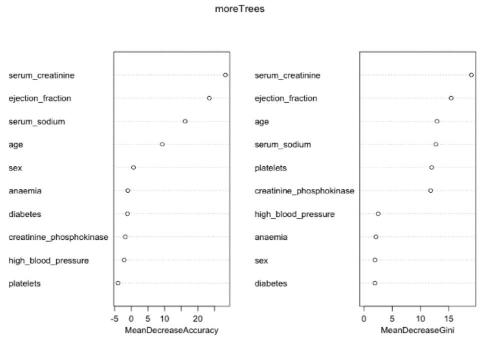
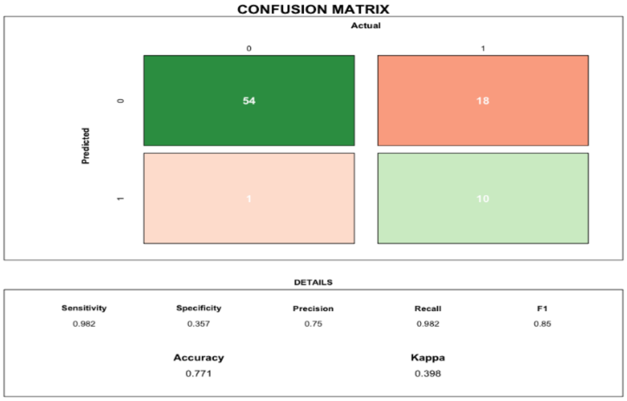


Figure 2.4

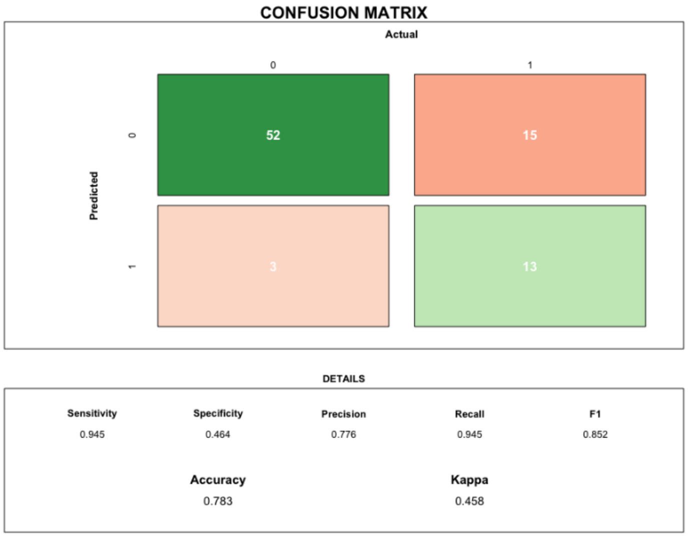


Despite the second model's adequacy, it appears that the initial random forest yielded the best model.

SVM

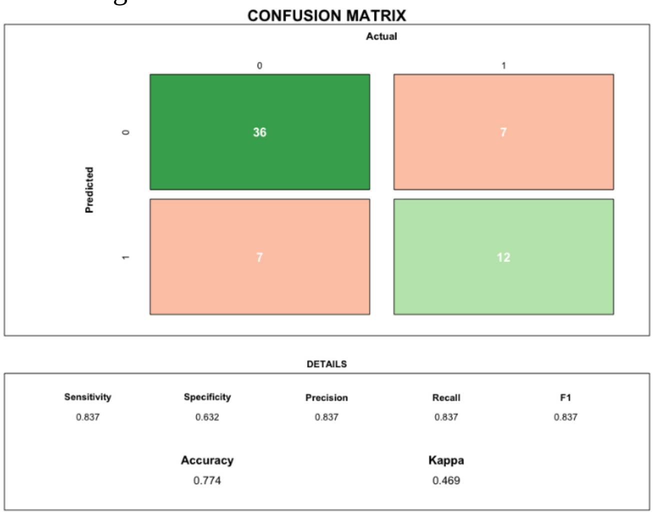
The first SVM model (Figure 2.5), the seed was set at nine with a data split where the training data consisted of 80% of the data and 20% assigned to the testing data. The model could predict a death event 78 percent of the time.

Figure 2.5



The second SVM model (Figure 2.6), had a seed set at twenty with a data split where the training data consisted of 70% of the data and 30% assigned to the testing data. The model could predict a death event 77% of the time.

Figure 2.6

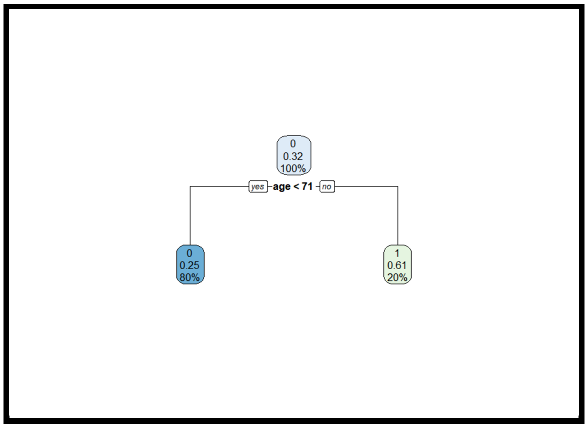


It appears that the SVM models did not perform at the same levels as the random forest models.

Decision Tree

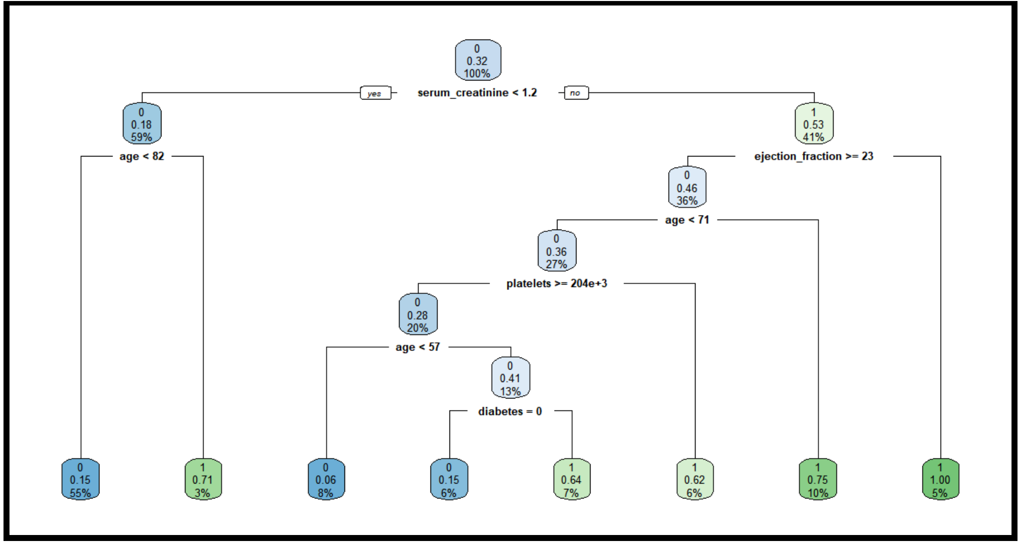
Several factors needed to be considered when creating a decision tree analysis model. First, there needed to be a way to remove age as a factor considering it does not provide any insight into a patient's overall health alone. The first decision tree (Figure 2.7) examines this occurring in the data, for any patients over the age of 71, there exists a higher likelihood of a death event in the dataset.

Figure 2.7



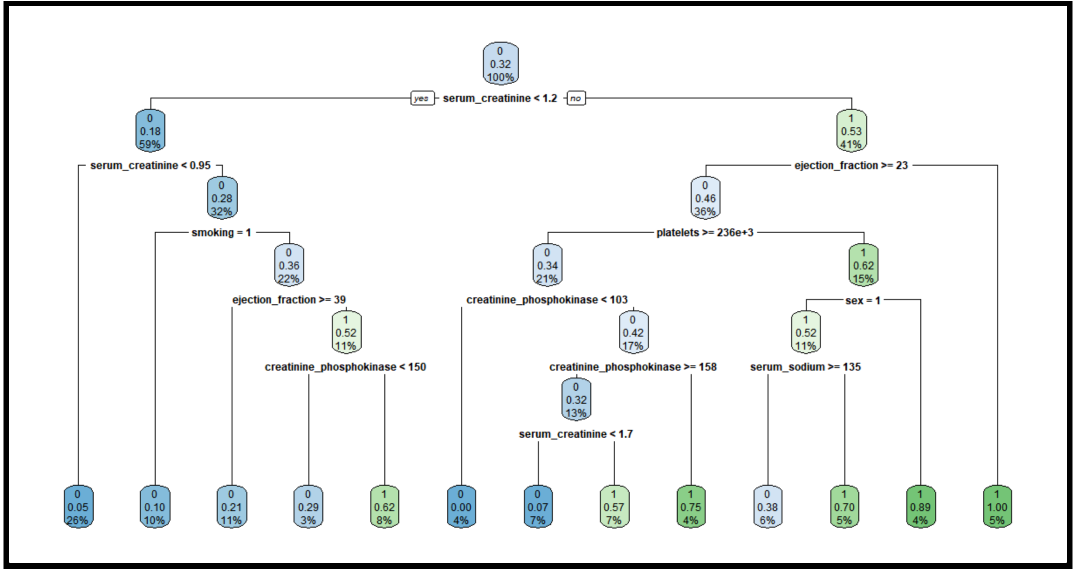
For the decision tree analysis, it was crucial to remove certain data columns as they had no relevance to the data or outcome. This included the time and patient ID values. In this model (Figure 2.8), it is revealed age does contribute as a significant node to determine death event, and it is not preferable to have a non-medical related data category dominating the data analysis model. Information that is actionable, informative and can be treated or examined is preferred. This model reported a 61% accuracy.

Figure 2.8



For the decision tree analysis, it was critical to remove non-informative data columns as they had no relevance to the focus of this analysis. The second data subset excluded the time, patient ID, and age column values. This model (Figure 2.9) reported a 74% accuracy.

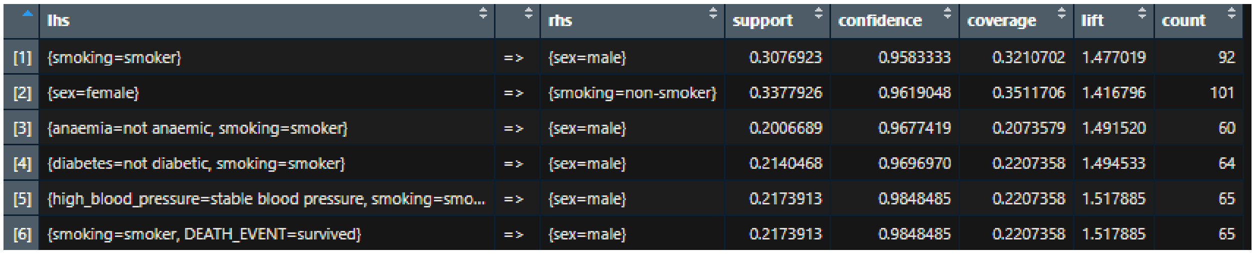
Figure 2.9



Association Rule Mining

Association rule mining was performed on the binary clinical measurements of the heart failure data set, including anemia, diabetes, high blood pressure, sex, smoking, and death event. After various adjustments of the Apriori algorithm’s parameters, a successful set of 20 rules were generated after setting the support to 0.13 and confidence to 0.9. A portion of the top 20 rules can be viewed in Table 2.1.

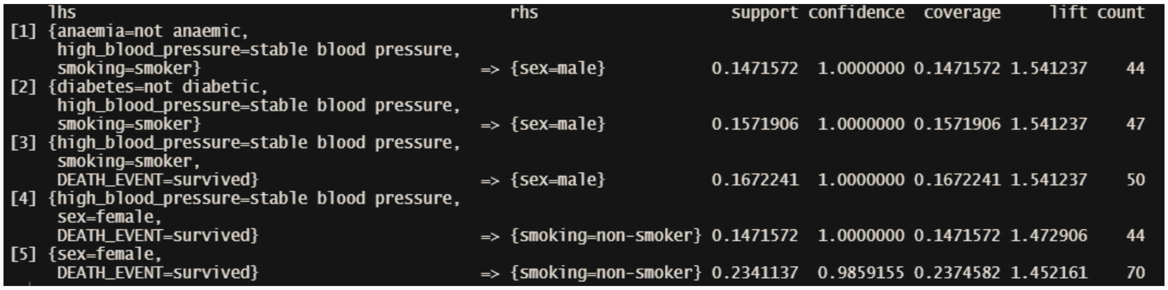
Table 2.1



	lhs		rhs	support	confidence	coverage	lift	count
[1]	{smoking=smoker}	=>	{sex=male}	0.3076923	0.9583333	0.3210702	1.477019	92
[2]	{sex=female}	=>	{smoking=non-smoker}	0.3377926	0.9619048	0.3511706	1.416796	101
[3]	{anaemia=not anaemic, smoking=smoker}	=>	{sex=male}	0.2006689	0.9677419	0.2073579	1.491520	60
[4]	{diabetes=not diabetic, smoking=smoker}	=>	{sex=male}	0.2140468	0.9696970	0.2207358	1.494533	64
[5]	{high_blood_pressure=stable blood pressure, smoking=smo...	=>	{sex=male}	0.2173913	0.9848485	0.2207358	1.517885	65
[6]	{smoking=smoker, DEATH_EVENT=survived}	=>	{sex=male}	0.2173913	0.9848485	0.2207358	1.517885	65

The top five rules in terms of confidence (Table 2.2) displayed various rules. For example, the 4th rule states that if a patient has stable blood pressure, is female, and survived, they are 100% likely to be a non-smoker.

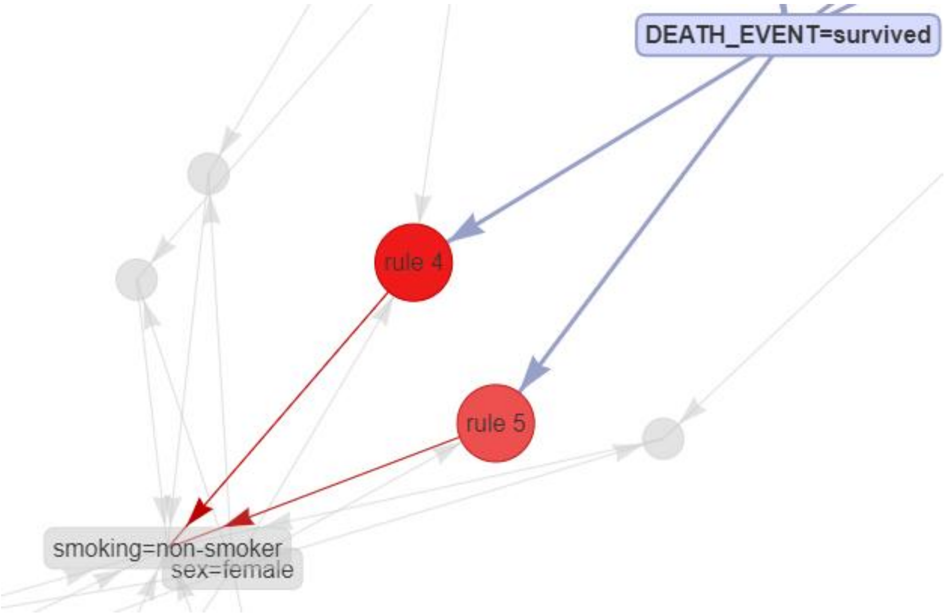
Table 2.2



```
    lhs                                              rhs                     support confidence  coverage     lift count
[1] {anaemia=not anaemic,
     high_blood_pressure=stable blood pressure,
     smoking=smoker}                              => {sex=male}            0.1471572  1.0000000 0.1471572 1.541237    44
[2] {diabetes=not diabetic,
     high_blood_pressure=stable blood pressure,
     smoking=smoker}                              => {sex=male}            0.1571906  1.0000000 0.1571906 1.541237    47
[3] {high_blood_pressure=stable blood pressure,
     smoking=smoker,
     DEATH_EVENT=survived}                        => {sex=male}            0.1672241  1.0000000 0.1672241 1.541237    50
[4] {high_blood_pressure=stable blood pressure,
     sex=female,
     DEATH_EVENT=survived}                        => {smoking=non-smoker} 0.1471572  1.0000000 0.1471572 1.472906    44
[5] {sex=female,
     DEATH_EVENT=survived}                        => {smoking=non-smoker} 0.2341137  0.9859155 0.2374582 1.452161    70
```

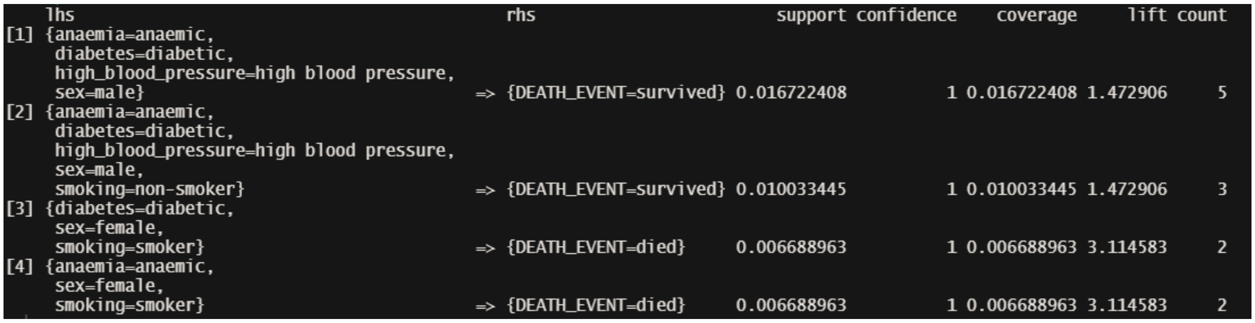
Below is a portion of an HTML widget (Figure 2.10) of the top 5 rules. It appears that death event, indicating survival in this scenario, is pointing towards a dark red circle, which represents a strong rule 4 stated previously. Rule 4 points towards patients being non-smokers, implying that being a non-smoker can lead to patient survival.

Figure 2.10



Next, the RHS of the Apriori algorithm was set to death event to discover rules that associated clinical variables to patient mortality, which can be viewed in Table 2.3. For example, the third rule states that if a patient is diabetic, female, and is a smoker, she is 100% likely to have died.

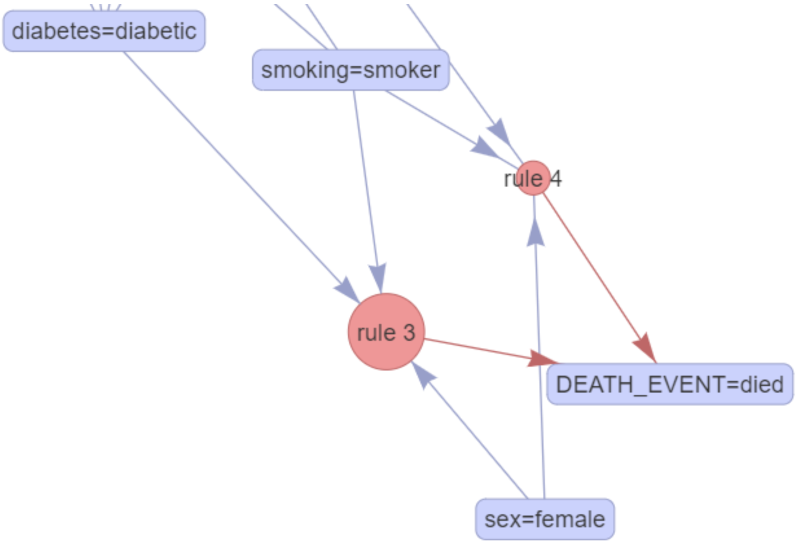
Table 2.3



```
    lhs                                                rhs                         support confidence    coverage     lift count
[1] {anaemia=anaemic,
     diabetes=diabetic,
     high_blood_pressure=high blood pressure,
     sex=male}                                     => {DEATH_EVENT=survived} 0.016722408          1 0.016722408 1.472906     5
[2] {anaemia=anaemic,
     diabetes=diabetic,
     high_blood_pressure=high blood pressure,
     sex=male,
     smoking=non-smoker}                           => {DEATH_EVENT=survived} 0.010033445          1 0.010033445 1.472906     3
[3] {diabetes=diabetic,
     sex=female,
     smoking=smoker}                               => {DEATH_EVENT=died}     0.006688963          1 0.006688963 3.114583     2
[4] {anaemia=anaemic,
     sex=female,
     smoking=smoker}                               => {DEATH_EVENT=died}     0.006688963          1 0.006688963 3.114583     2
```

Another HTML widget was created for when the RHS (Figure 2.11) has been set to death event. It appears that diabetic, smoker, and female are pointing towards Rule 3, which then points to death event = died.

Figure 2.11



This is another instance where smoking has led to patient death, further implying that there is a strong association between smoking and death by heart failure.

Conclusions

There appears to be a connection between serum creatinine and ejection fraction abnormalities and a higher chance of heart failure and death. While this is the case, other intriguing factors did not impact a possible death event as highly. Diabetes and high blood pressure did not seem to contribute to heart failure and death as the variables mentioned previously. While these are by no means something that can be ignored, it is intriguing that they did not play as big of a role.

Initial analysis using the decision tree model revealed that age is the most dominant factor considering heart failure death events for patients suffering from cardiovascular ailments. Within this category of medical conditions, it is understandable that older patients are more likely to die from heart failure. In the heart failure dataset, ages ranged from 40 to 91 years old. A distribution of the ages between dead and surviving patients varied with an average difference of 5 years, surviving patients leaning toward the younger side. Ruling out age as a factor helped better accommodate the analysis to more relevant medical test and prognosis with variables such as ejection fraction and serum creatinine being a better variable to monitor and target medically. Given serum creatinine Level of creatinine in the bloodstream above of 1.2 mg/dL and Ejection fraction value above 23, a patient would follow a branch that leads to a death event. This is followed by the nodes of a smoker and a significant level of platelets (2.3M) in the patient's blood.

In terms of the binary clinical measures, there appears to be a prevalent connections between smoking and death events resulting in patient death. There were various instances of strong associations between patient survival when the heart failure patient in question is a non-smoker.