Machine Learning Assignment 2

February 2023

1 Introduction

1.1 Background

Cardiovascular disease is one of the most deadly diseases worldwide, and its disability and mortality rates are increasing every year, making it one of the most important diseases for the global medical community. Heart failure is one of the most serious consequences of cardiovascular disease. If patients with cardiovascular disease or those at high risk of cardiovascular disease can be screened early, machine learning can help screen those at risk based on the results.

This paper is based on a dataset of electronic medical records of heart failure patients from Pakistan to predict survival in patients with cardiovascular disease and to screen features in the dataset as biomarkers to predict survival in patients with cardiovascular disease. Several methods in machine learning can be used to accomplish this task, and this paper will use generalised additive model and tree-based model.

1.2 Exploratory Data Analysis

This data was collected from a group of heart failure patients who attended the Institute of Cardiology and Allied hospital Faisalabad-Pakistan from April to December 2015.

The data recorded the basic profile of 299 heart failure patients, comprising 194 male and 105 female patients, whose ages ranged from 40-95. In terms of patient information, the data contains 13 variables which comprise clinical data, physical information and lifestyle information. Specifically, the 13 variables were age, anaemia, creatinine phosphokinase, diabetes, ejection fraction, high blood pressure, platelets, serum creatinine, serum sodium, sex, smoking, time and death event.

An initial collation of the dataset reveals that the variable DEATH_EVENT in the dataset is a target/response variable with a binary variable type; the other 12 variables are independent/predictor variables, which contain both binary and continuous variables of the data type. The specific types of variables and the statistical results (including distribution, mean, standard deviation, quartiles, minimum and maximum values, etc.) are detailed in **Table 1**.

Feature	**Explanation**	**Data type**
age	Age of the patient	Numerical
anaemia	Decrease of red blood cells or hemoglobin	Boolean,0: No
creatinine phosphokinase	Level of the CPK enzyme in the blood	Numerical
diabetes	the patient has diabetes or not	Boolean,0: No
ejection fraction	Percentage of blood leaving the heart at each contraction	Numerical
high blood pressure	the patient has hypertension or not	Boolean,0: No
platelets	Platelets in the blood	Numerical
serum creatinine	creatinine in the blood	Numerical
serum sodium	sodium in the blood	Numerical
sex	Male or female	Boolean, 0: Female
smoking	the patient smokes or not	Boolean, 0: No
time	follow up time	Numerical
death event	The target event, if the patient died during the follow-up period	Boolean,0: No

Table 1: Features of Dataset

In particular, it is noted that there is an imbalance in the data set, with the data set showing 96 death events (32.11%) and 203 survived patients (67.89%).

Feature	Full size	Dead patients(32.11%)	Survived patients(67.89%)
anaemia(No)	56.86%	52.08%	59.11%
anaemia(Yes)	43.14%	47.92%	40.89%
diabetes(No)	58.19%	58.33%	58.13%
diabetes(Yes)	41.81%	41.67%	41.87%
high blood pressure(No)	64.88%	59.38%	67.49%
high blood pressure(Yes)	35.12%	40.62%	32.51%
sex(Female)	35.12%	35.42%	34.98%
sex(Male)	64.88%	64.58%	65.02%
smoking(No)	67.89%	68.75%	67.49%
smoking(Yes)	32.11%	31.25%	32.51%

Table 2: Statistics of Category features

Feature	count	mean	std	min	25%	50%	75%	max
age	299	60.83	11.89	40	51	60	70	95
creatinine phosphokinase	299	581.84	970.29	23	116.5	250	582	7861
ejection fraction	299	38.08	11.83	14	30	38	45	80
platelets	299	263358.03	97804.24	25100	212500	262000	303500	850000
serum creatinine	299	1.39	1.03	0.5	0.9	1.1	1.4	9.4
serum sodium	299	136.63	4.41	113	134	137	140	148
time	299	130.26	77.61	4	73	115	203	285

Table 3: Statistics of Numerical features

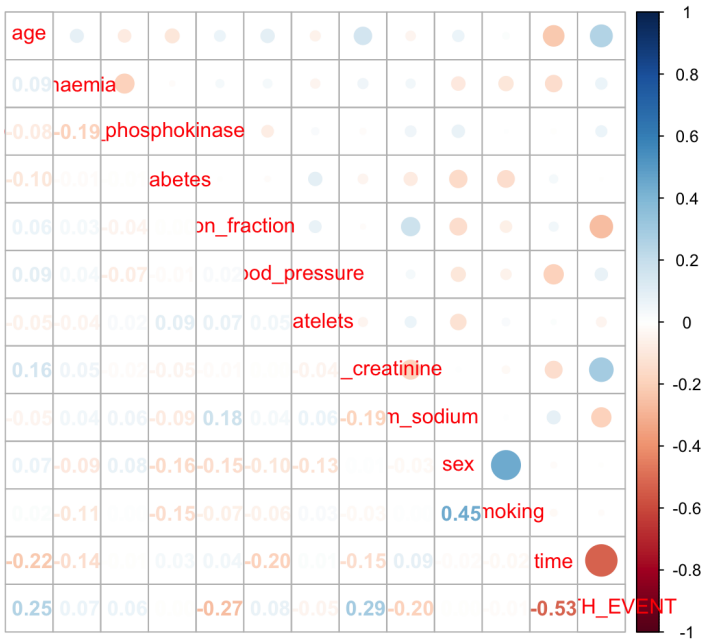

Figure 1: Correlation of each variable

Based on picture 1, we can assume that age, ejection fraction, serum creatinine, serum sodium and time are more correlated with the event of death.

2 Analysis and Modeling

2.1 Methods

2.1.1 Generalized Additive Model

The Generalized Additive Model (GAM) is an extended linear model that models complex relationships between variables by introducing non-linear smoothing functions (e.g. natural splines) and can be applied to model non-linear, non-monotonic, high-dimensional data with greater flexibility. GAM allows for more flexible adaptation to data and discover potential non-linear relationships, resulting in improved predictive accuracy and data exploration.

For this heart disease dataset, which contains a relatively large number of features and the relationship between the predictor and response variables is unclear, and the response variables are binary, this dataset involves a typical binary classification problem. In this paper, the GAM model is chosen to solve both the classification and prediction problems, as the GAM model can simultaneously test whether there is a linear relationship between the variables and also select the best predictor variables. Therefore, the GAM model was first applied in processing this dataset.

2.1.2 Random Forest

Random Forest is a classification tree based machine learning algorithm that can be used for classification and regression problems. For classification problems, Random Forest can learn from the data in the training set to obtain a model that can classify unknown data. In prediction, the unknown data is fed into the model and the random forest will classify it according to the patterns in the training data and output the prediction results.

For feature selection, the random forest can assess the importance of features by calculating the importance of each feature in the tree. This method can be used for feature selection, where the features that contribute most to the prediction results are selected to improve the prediction accuracy of the model.

When constructing a classification tree, the random forest randomly reselects n observations from the original data, some of which are selected multiple times and some of which are not selected, as a Bootstrap resampling method. At the same time, the random forest randomly selects some of the variables from the k independent variables for the determination of the classification tree nodes. In this way, the classification tree may be constructed differently each time. In general, the random forest randomly generates several hundred to several thousand classification trees and then selects the tree with the highest degree of repetition as the final result.

In addition, random forests have the following advantages over single decision tree models. Random forests can reduce the risk of overfitting of single trees and improve generalisation performance. Random forests can handle high-dimensional data and are able to select important features and reduce dimensionality. Random forests have better robustness and can handle missing values, outliers and other situations. Random forests are able to parallelise processing and speed up the training process.

As random forests can be used for both classification and regression problems, the role of random forests is to construct classifiers when the response variable is a categorical variable, which is consistent with the response variable (DEATH_EVENT) for this dataset. Therefore, we choose Random Forest as a computational method here. Since the process of constructing a tree with random forests is a process of filtering the importance of features, using random forests allows us to select the most important features.

2.2 Results

2.2.1 Results of GAM

When building the GAM model, 80% of the dataset is randomly selected as training data and the remaining 20% as test data. The GAM model contains two parts, one for the parametric estimation of the cofactors that enter the model in parametric form, and the other for the non-parametric estimation of the smoothing cofactors.

In the results of the parametric coefficients, it can be noted that all variables did not reach significance and only gender reached borderline significance with $p = 0.102$, suggesting that male patients may be more likely to death, but this needs to be verified by collecting more data.

In the results of Non-parametric estimation for approximate significance of smooth term, it can be found that 7 features used as smoothing factor which used Chi-square test statistic for testing the hypothesis.

Parametric coefficients				
Covariates	**Estimate**	**Standard Error**	**Z value**	**p-value**
Intercept	0.1103	0.5164	0.214	0.831
sex	-0.8154	0.4980	-1.637	0.102
smoking	0.0425	0.5109	0.083	0.934
diabetes	-0.2156	0.4431	-0.487	0.627
high blood pressure	-0.1990	0.4344	-0.458	0.647
anaemia	-0.1983	0.4374	-0.453	0.650
Approximate significance of smooth terms(Non-parametric)				
Smooth Covariates	**Edf**	**Ref. df**	**Chi.sq**	**p-value**
age	1.001	1.003	5.682	0.01729 *
ejection fraction	1.964	2.478	17.077	0.00063 ***
serum sodium	3.202	3.946	4.797	0.29453
serum creatinine	3.471	4.232	11.716	0.02361 *
platelets	2.263	2.869	3.076	0.45457
creatinine phosphokinase	1.000	1.000	1.588	0.20760
time	5.111	6.205	48.930	<2e-16 ***

Edf: Estimated degrees of freedom; **Ref.df**: Degrees of freedom before smoothing; **Chi.Sq**: Chi square value. **Significance Level**:'***' 0.001; '**' 0.01; '*' 0.05; '.' 0.1. **R-sq.(adj)** 0.563;**Deviance explained** = 53.9%; **UBRE(Un biased risk estimator)** = -0.18344

Table 4: Results for GAM of Heart disease data analysis

As can be seen in Table 4 the four features of age, projection fraction, serum creatinine and time all reached significant levels. The p-value for age was 0.01729, for projection fraction 0.0006 and for serum creatinine 0.02361. The p-value for time was ¡2e-16, but time as a follow-up time is not a physical or pathological feature of the patient that we do not focus on here.

The results suggesting that the older the patient with cardiovascular disease, the higher the risk of death. The relationship between ejection fraction and patient death may be that the lower the ejection fraction score, the more likely the patient is to die. The higher the index of serum creatinine, the higher the probability that the patient will die.

From Table 2, it can also be seen that the GAM model has an adjusted R-squared value of 0.563, of which 53.9% of the deviance is explained. the UBRE (Un biased risk estimator) score is 0.18344. these scores indicate that the model is not particularly effective.

As seen in Figure2, age and creatinine phosphokinase may have a linear relationship with the response variable, while all other non-parametric variables are non-linear with the response variable and their non-linearity is related to their smoothness.

Applying the test data to this GAM model, we can see a prediction accuracy of 95%, precision of 95%, recall and specificity both at 100% and an F1 score of 97.44%. This prediction result may seem good, but based on the model results it shows that the model does not perform very well. Considering what we found in our exploratory analysis, the data is highly imbalanced and the imbalance is particularly high for this test data. Therefore, it is likely that the high prediction accuracy is due to the imbalance in the dataset.

		Actual	
		0	1
Predict	0	57	3
	1	0	0

Table 5: Prediction Result of GAM

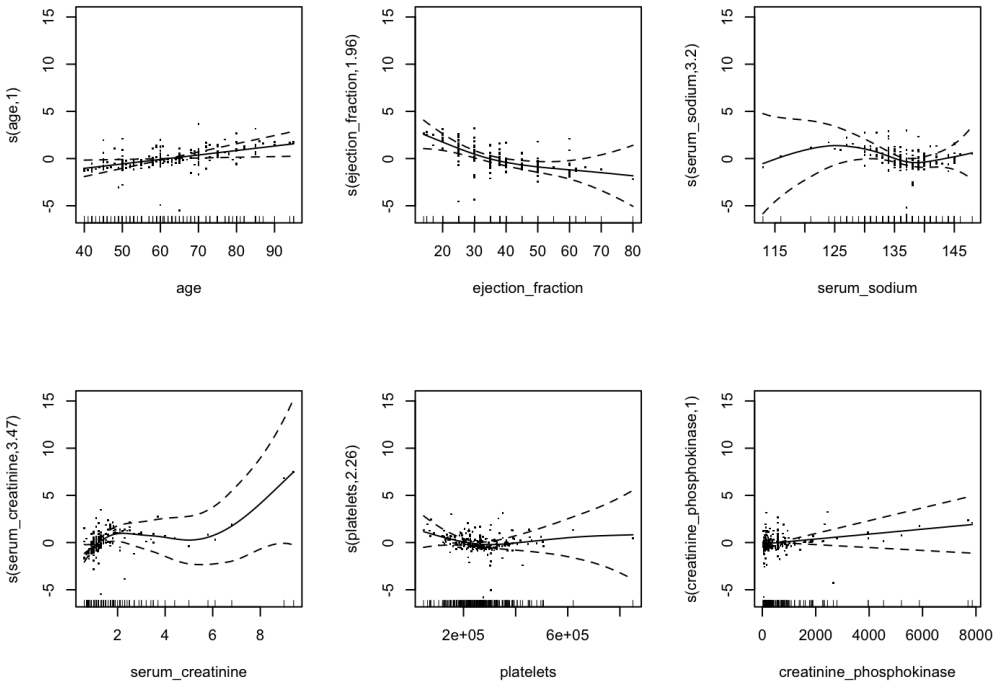

Figure 2: the Partial Dependence Plot (PDP)

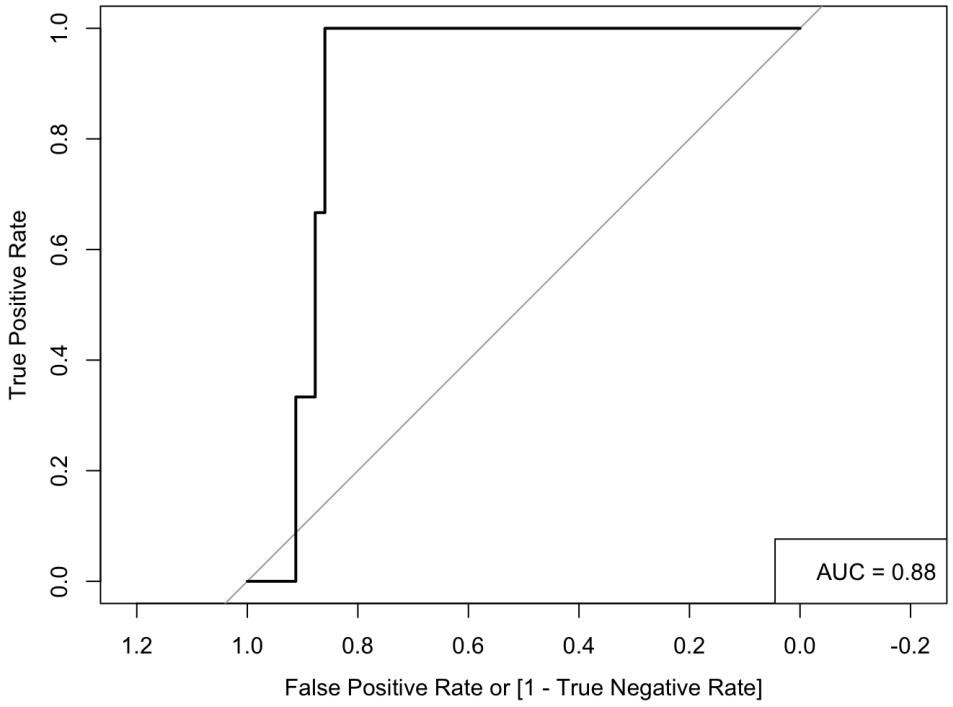

Figure 3: ROC of GAM

Based on the AUC data in the ROC curve, which shows AUC = 0.88, the model predicts average per-

formance when looking purely at the AUC value. However, the quality of the model can be obtained by comparing it with the AUC of other models.

2.2.2 Results of Random Forests

The random forest can solve both feature selection and classification prediction problems, so from Table 6, features can be selected based on feature importance, where the features with high importance values include time, serum creatinine, ejection fraction, serum sodium and age. The important features selected by the random forest are the same as those that reach significance in the GAM model.

time	serum_creatinine	ejection_fraction	serum_sodium	age	platelets
0.448400	0.178062	0.157850	0.071292	0.070110	0.042276
creatinine_phosphokinase	sex	smoking	anaemia	diabetes	
0.021889	0.003538	0.002908	0.002356	0.001319	

Table 6: Feature Importance

		Actual	
		0	1
Predict	0	155	9
	1	19	56

Train Accuracy : 0.8828

Table 7: Train Confusion Matrix

		Actual	
		0	1
Predict	0	35	4
	1	9	12

Test Accuracy : 0.7833

Table 8: Test Confusion Matrix

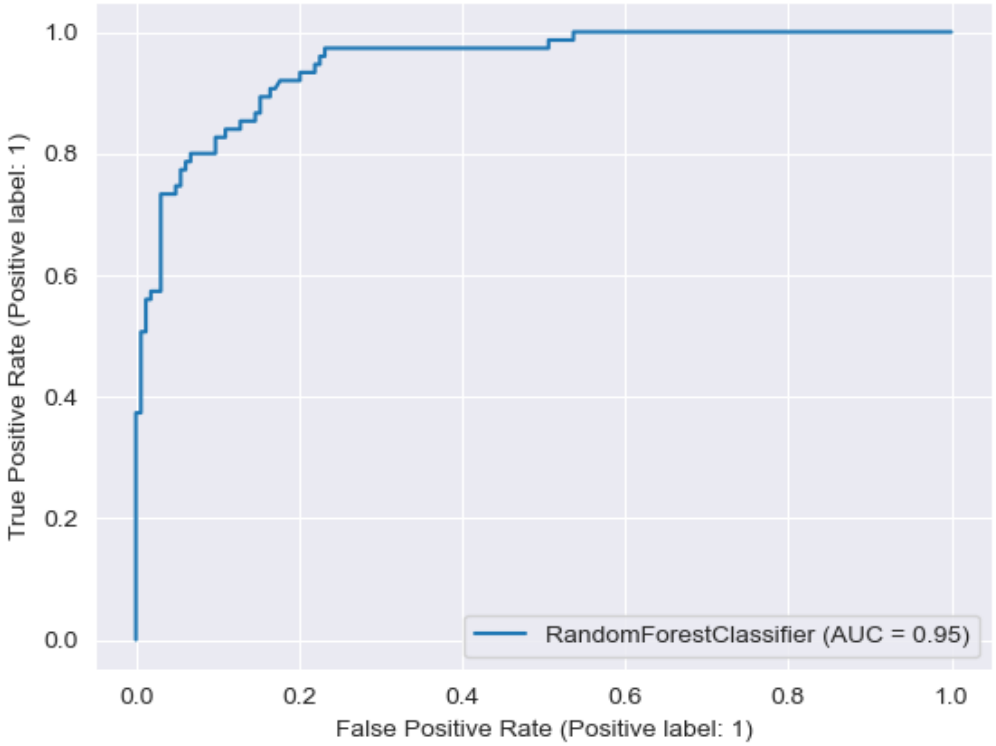

Figure 4: Caption

Since the definition of time is unclear and time is not a physical or pathological index, we assume that the

effect of time on death events will not be particularly significant, and therefore we do not focus too much on time here as we did with the GAM model.

According to the Train Confusion Matrix and the Test Confusion Matrix, the Train Accuracy reaches 0.8828, while the test accuracy is 0.7833. The AUC value of the ROC curve reached 0.95, indicating that the model performed well in prediction.

Method	Accuracy	Precision	Recall	Specificity	F1 Score	AUC
GAM	95%	95%	100%	100%	97.4%	0.88
Random Forests	76.6%	76.9%	47.6%	92.3%	58.8%	0.95

Table 9: Caption

2.2.3 Comparison of two methods

A comparison of the data from Table 9 shows that all indices of GAM are better than the data from the random forest, except for the AUC index. But here it cannot be simply assumed that GAM is better than Random Forest. The possible factor here lies in the fact that the GAM and Random Forest programs are written in two files, and the bias is caused by the inconsistency between the selected training dataset and the test dataset. We found with the test dataset of GAM that the randomly selected dataset of GAM has a greater imbalance.

Looking at the ROC curves and AUC values, random forests may work a little better. This needs to be explored further by re-tuning the dataset and other parameters so that a better performing model can be selected.

3 Discussion and Further work

The results from both GAM and random forest models showed that serum creatinine, ejection fraction, age and time were all important features in predicting the likelihood of death in patients with cardiovascular disease, but in terms of physiological indicators/pathological indicators, serum creatinine, ejection fraction can be used as a biomarker to screen for patients at high risk of cardiovascular disease.

Whereas time needs to be further explored in terms of its meaning in relation to patient death. It is common knowledge that there is a high correlation between age and death in older people at high risk of cardiovascular disease.

There is an area for improvement in comparing the merits of the two models. We wrote the two models in two files when training the models, resulting in inconsistent training and testing data for a random sample of the two models, which makes it difficult to compare the merits of the two models effectively.

There are areas for improvement in comparing the two models. We wrote the two models in two files when training the models, resulting in inconsistencies between the randomly sampled training data set and the test data used for the two models, which makes it difficult to compare the two models effectively.

In the case of both models, there are areas for improvement. For the GAM model, we were unable to determine the relationship between the predictor and response variables without processing the data in greater depth, and therefore we constructed the model using the same smoothing function for all variables except the categorical variables, without further consideration of choosing a smoothing function such as natural smoothing or B splines. In addition, we chose default values for the choice of degrees of freedom with the functions without calculation. Therefore, having obtained the GAM model, the corresponding functions can be reselected and the degrees of freedom adjusted according to the features of each variable. Also, with the model, we found that only a very small number of these key features have the greatest impact on prediction, so the best subset can be selected on this basis to retrain the model, which will train a better performing model.

For GAM models, there is also the method of K- folds cross-validation that can be used to evaluate GAM models, which reduces the random fluctuations of the model by iterating over multiple datasets, thus improving the reliability of the model. Alternatively, boosting as a general method, applying this method to GAM as well, may also result in a better model.

Random forest as an integrated learning method based on decision trees, in order to decision trees can have better performance, the accuracy of the model can be improved by continuing to increase the number of

decision trees, but as the number of decision trees increases, the problem of over-fitting needs to be considered; and adjusting the parameters of the decision trees is also a way to improve the accuracy of the model, such as adjusting the depth of the tree, the minimum number of samples of leaf nodes, the minimum division of samples and other parameters, may also improve the performance of the model.

It is also a more effective approach if the best features can be selected based on the feature selection algorithm and then the best feature dataset can be used for training; the model accuracy can also be improved by increasing the number of samples in the training dataset, for instance, by increasing the number of samples through data augmentation.

4 Executive Summary

The data in this paper is a dataset containing electronic medical record information for 299 patients with cardiovascular disease, recording thirteen key pieces of information for each patient, both personal information such as gender and age, as well as daily behavioural habits such as whether or not they smoke, along with pathology-related information such as the presence or absence of hypertension, and finally a variable for whether or not the patient died. Of these, whether the patient died is the target event that we are most interested in and the other variables are the ones that we need to investigate in depth. The aim of this paper is to use a machine learning approach to explore the prediction of whether a patient will die from heart failure based on this closely related information about cardiovascular disease. Also, to explore which of these 12 variables can be used as biomarkers to predict death from heart failure, in other words, to screen for variables that, when seen as abnormal, can be used by doctors to determine that a patient is likely to die from heart failure, so that they can take action in advance.

The methods chosen for the study were the Generalised Additive Model (GAM) and Random Forests, which are more commonly used in machine learning. The generalised additive model is an extension of the linear logic model, so understanding the generalised additive model can be understood by first understanding the linear relationships. In the simplest sense, a linear relationship is one in which the response variable increases by one unit after a certain increase in one variable, for example, a child's weight may increase by 10kg for every 10cm increase in height. A generalised linear model is one in which there are multiple independent variables in some complex events, and the linear relationship between the multiple independent variables can be simply understood as a generalised linear model, for example, a child's weight is not only linearly related to height, but may also be linearly related to waist circumference, for example, for every 1cm increase in waist circumference, weight increases by 2kg, so there is a linear relationship between height, waist circumference and weight. To predict weight based on height and waist circumference, the model should be constructed by considering both height and waist circumference, and the functions of height and waist circumference and weight should be added together to form a complete model. The generalised additive model (GAM) is an extension of the linear model in that the GAM does not require a linear relationship between the variables. For example, the relationship between age and weight may be increasing with age at a young age, but at an older age, weight may be decreasing with increasing age. At this point, a linear model would not be able to predict this, whereas GAM can add non-linear variables to the model by constructing a non-linear model function. Just as we mentioned the relationship between height, waist circumference and age and weight, the model of height, requirement and weight have their own coefficients, while the relationship between age and weight requires the construction of a special predictive function and then adding the three together. This is a simple illustration of the GAM, which is not very accurate but can be quickly understood. once the GAM is constructed, it is possible to find out which variables are more sensitive to the target event and these are the best variables to use for prediction.

By the name of Random Forest, we can assume that the method is related to trees. And a random forest is an extension of a decision tree. In line with the logic of understanding GAM, to understand random forests, one first understands decision trees as well. A decision tree is a classification and regression method based on a tree structure. A decision tree model is like a tree with roots and bifurcations, as can be seen in Figure 5, which is the result of using a decision tree to model this dataset. From the Figure it can be seen that the variable time contains all the samples like the root of the tree. Based on the value of time it is possible to classify the samples into less than 67.5 days and more than 67.5 days, less than 67.5 days is more likely to have a death, here it is like a tree bifurcating into the next level and two other variables appear, the two variables will bifurcate again until all the variables are found. We can find that the closer the variable is to the root, the more important it is. And a random forest is the construction of very many very pairs of such decision

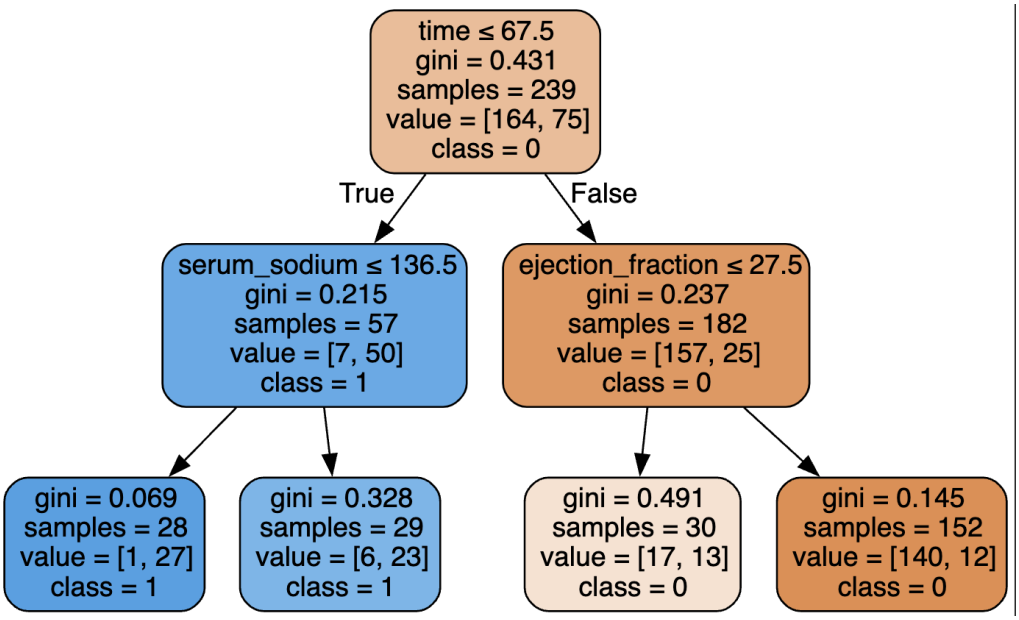

Figure 5: Decision Tree

trees, which are brought together into a forest, from which the model obtained after many rounds of fitting is selected. A decision tree is based on the entire data set to build a tree, while a random forest is a random sampling of a portion of the data, and the variables contained in each portion of the data are also randomly selected, so that by randomly building a completely different tree, the purpose of sampling all the data and all the variables is achieved, and eventually a random forest is formed to produce the best random forests model.

After building the model, we can make predictions on similar data, and the predictions include accurately predicting what should happen, accurately predicting what should not happen, incorrectly predicting what should happen as not happening and incorrectly predicting what should not happen as happening. With these events it is possible to predict accuracy etc.

These machine learning methods have a wide range of applications in real life, such as the medical applications we have studied in this paper, and can also be applied to financial risk control to prevent fraud, and have been used to make product recommendation systems, etc. The GAM can also be used in a wide range of applications, such as psychology, economics and ecology, where many variables are involved and the relationship between the variables is not clear, and can be used to predict financial markets, weather changes, etc.; biology and chemistry, where there are strong interactions between variables, can also use the GAM.