Machine Learning and Statistical Methods for Predicting Survival of Patients with Heart Failure

Lingjie Jia *

Department of Mathematics and Statistics, University of Glasgow, Glasgow, UK

* Corresponding Author Email: 2534535J@student.gla.ac.uk

Abstract. Heart failure is a branch of heart disease, which causes millions of people to die worldwide. This research is an investigation of survival prediction of patients with heart failure. The type of this prediction is a binary classification problem. The machine learning models used in this paper are logistic regression, decision trees, random forests, support vector machines, and artificial neural networks. The methods for evaluating the performance of prediction models are k-fold and stratified k-fold cross-validation. The results of 2 cross-validation indicate that logistic regression has the best performance. In addition, according to the feature ranking method in the literature, it can be observed that the prediction of heart failure mainly relies on serum creatinine, ejection fraction, and follow-up time. The conclusion is that the logistic regression, which only involves features: serum creatinine, ejection fraction, and follow-up time, is well suited for predicting the survival of patients with heart failure.

Keywords: Machine learning, Biostatistics, Survival.

1. Introduction

Approximately 17 million people worldwide die from cardiovascular diseases every year, with the main symptoms being myocardial infarction and heart failure. Heart failure occurs when the heart cannot pump enough blood to meet the body's needs. By investigating the factors that cause heart failure and making effective prevention, it is possible to reduce the mortality rate of mental failure [1]. With the development of information technology, researchers can more easily collect data on heart disease patients and use statistical and machine learning methods to analyze the causes of heart disease. This paper mainly develops machine learning methods to investigate the survival prediction of heart failure through the given data set and analyses the main influencing features of heart failure. In section 2, the basic situations of the chosen dataset will be introduced. Section 3 is about methods used in this research, in which there are five machine learning methods: logistic regression, decision tree, random forest, support vector machines, artificial neural networks, and two cross-validation methods: k-fold, and stratified k-fold. In addition, there are some features ranking methods, which will be discussed through literature [1]. The section 4 is about the results of this research, which will show the performance of survival prediction models. The final section 5 is a short conclusion of research.

2. Dataset

The dataset used in this research is collected from the UC Irvine platform, the dataset is also known as Heart failure clinical records [2]. This dataset is also used in literature “Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone” [1]. Figure 1 describes the details of features in the dataset. The target feature is a boolean type, which follows that this heart failure prediction is a binary classification problem. The total number of patients in the dataset is 299. The next section will provide some classification methods.

Feature	Explanation	Measurement	Range
Age	Age of the patient	Years	[40, ..., 95]
Anaemia	Decrease of red blood cells or hemoglobin	Boolean	0, 1
High blood pressure	If a patient has hypertension	Boolean	0, 1
Creatinine phosphokinase (CPK)	Level of the CPK enzyme in the blood	mcg/L	[23, ..., 7861]
Diabetes	If the patient has diabetes	Boolean	0, 1
Ejection fraction	Percentage of blood leaving the heart at each contraction	Percentage	[14, ..., 80]
Sex	Woman or man	Binary	0, 1
Platelets	Platelets in the blood	kiloplatelets/mL	[25.01, ..., 850.00]
Serum creatinine	Level of creatinine in the blood	mg/dL	[0.50, ..., 9.40]
Serum sodium	Level of sodium in the blood	mEq/L	[114, ..., 148]
Smoking	If the patient smokes	Boolean	0, 1
Time	Follow-up period	Days	[4,...,285]
(target) death event	If the patient died during the follow-up period	Boolean	0, 1

Figure 1. Detail of features [1].

3. Methods

3.1. Logistic Regression (LR)

LR is a supervised machine learning algorithm mainly used for binary classification problems [3]. The learning and prediction procedures revolve around assessing the probability of binary classification. For the LR model to operate effectively, the class variable must be binary. In most datasets of problems of heart failure prediction, the columns of target feature comprise binary values, "0" denoting patients with no chances of heart failure and "1" signifying those predicted as heart failure patients. On the other side, independent variables can take on binary classified, nominal, or polynomial types [4]. The LR equation is expressed as follows:

$$\log\left(\frac{p}{1-p}\right) = a + b \times x, \tag{1}$$

Where p is the probability, a is the intercept, b is the regression coefficient of x, and x is the predictor variable [5].

The initial step for constructing a predictive model through LR is to use the univariable analysis to investigate the unadjusted associations between each variable and the predicted class. Subsequently, both categorical and continuous variables in the dataset are involved in the p-value test, and variables with results lower than 0.25 are selected and integrated into multivariate analysis [6]. Following this, the second step involves incorporating all selected variables with a p-value less than 0.25 into the multivariable logistic regression model. In the ultimate iteration of this model, variables with a p-value surpassing 0.05 are filtered out. Moving to the third step, smoothed scatter plots are generated to evaluate the relationship between continuous variables and the logit scale outcome. The fourth step involves scrutinizing for potential interactions among the chosen predictor variables [5].

3.2. Decision Tree (DT)

DT is a tree-based approach in which each path from the root is delineated by a series of data separation criteria until a Boolean result is reached at a leaf node. It serves as a hierarchical representation of knowledge relationships, characterized by nodes and connections. In the context of classification using relationships, nodes act as decision points [7]. This tree-based algorithm is a typical classification method in machine learning. Figure 2 is an example of a DT.

DT algorithm can classify patients based on known mortality features of heart failure patients. Predictors are divided into Class A (high risk) and Class B (low risk). DT requires minimal data preparation before building a predictive model, and its performance is not highly affected by non-linear data distribution. The main problem is overfitting, which occurs when a model exhibits high levels of accuracy during training, resulting in inaccuracies or the testing accuracy being significantly lower than the training accuracy [5].

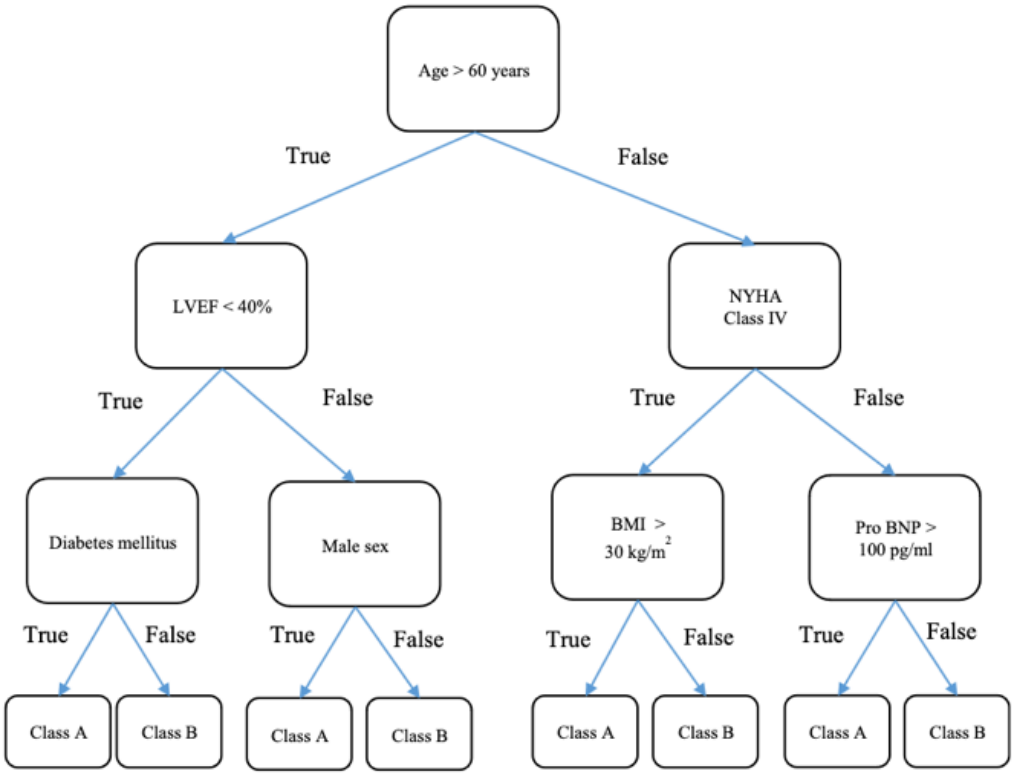

Figure 2. An example of DT [5].

3.3. Random Forest (RF)

RF is an ensemble learning method that involves aggregating a large number of DTs, resulting in reduced variance compared to a single DT [8]. This method is a supervised machine learning method, and it is widely used in investigating both classification and regression problems [3].

When developing a classification model by the RF algorithm, the initial step involves the random selection of samples from a provided dataset. Subsequently, a DT is constructed for each sample, and predicted outcomes are generated from each tree. The final predicted class is then determined by selecting the class with the highest number of votes from the predictions made by all DTs. In the case of a regression model, where the output is numerical, the final prediction is derived from the mean or average value of the predicted outputs. In addition, the RF algorithm exhibits proficiency in handling datasets with missing values, and it demonstrates robust performance on large datasets, showcasing an ability to rank features based on their importance. However, it is important to acknowledge that a significant drawback of the RF algorithm is its computational expense, demanding more training time compared to many other algorithms [5]. Figure 3 shows a process of constructing an RF algorithm.

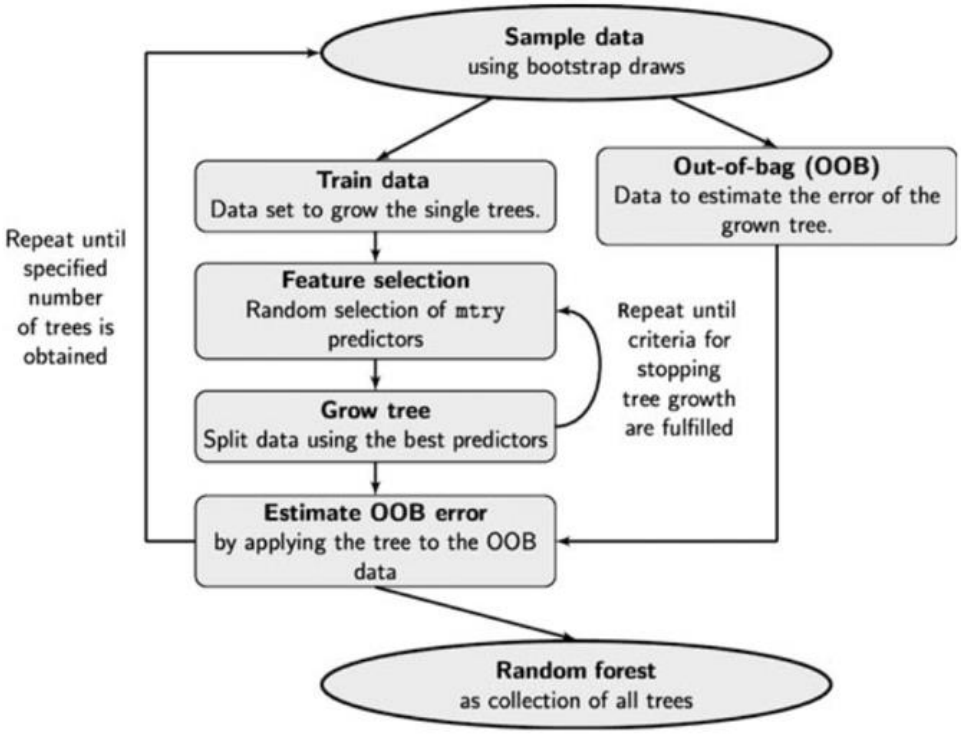

Figure 3. Process of RF algorithm [9].

3.4. Support Vector Machines (SVM)

SVM is effective for both linear and non-linear data classification and regression analysis. One of the key advantages of SVM is its ability to handle multi-class classification within a single dataset [5].

SVM achieves classification by focusing on the data points located at the edges of each class. It establishes a line or hyperplane that separates the classes. The optimal hyperplane is determined by selecting the maximum distance between the support vectors (observations on the class edges) and the hyperplane. In situations where the data cannot be linearly separated, a kernel function is employed to transform the data from a lower-dimensional space to a higher-dimensional space, making it separable [5].

3.5. Artificial Neural Networks (ANNs)

ANNs exhibit similarities to neurons, the fundamental units of the central nervous system. They consist of three main components: an input layer, a hidden layer, and an output layer. The input layer represents the extracted features used by the model, which can be continuous or categorical variables. These features serve as inputs to predict one or multiple categories represented by the output layer. The nodes in the input layer connect with each node in the hidden layer, and their connections are initially assigned arbitrary values ranging from 0 to 1. The weighted sum of the input signals is then fed through an activation function, typically the sigmoid function. This function transforms the weighted sum, converting negative values to values close to 0 and positive values to values close to 1. Nowadays, most neural networks are designed with multiple hidden layers, enhancing their ability to classify nonlinear data [5].

3.6. Cross-validation (CV)

CV is a data resampling method used to evaluate the generalization ability of a predictive model and prevent overfitting [10]. It is an approach for selecting suitable models. In this paper, two CV methods, k-fold and stratified k-fold CV, are used in prediction models. Furthermore, the Monte Carlo stratification method is introduced in section 3.7.

In the k-fold, the using dataset split into k disjoint subsets and each one of the subsets will be used as a test set for one time. In the other words, the model is trained by different unions of k-1 disjoint subsets [11]. K-fold provides more test results, allowing a high fault tolerance rate during model analysis. However, since the data set is randomly divided, there may be an imbalance in the distribution of values of target variable in the test set or training set. The extreme case that may occur is that all values of the target variable in the training set or test set are "0". The stratified k-fold is a booster of k-fold. In addition to dividing the dataset like k-fold, the ratio of values of target variable for each fold is fixed, which ensure that no extreme case happen [10].

3.7. Features Ranking

Feature selection has long been an important research topic in the fields of pattern recognition, machine intelligence, and data mining. It can help simplify the model and exclude some features that are not relevant to predictions [12].

D. Chicco et al. use the Monte Carlo stratification method to divide the dataset [2] in section 2 into 100 random subsets, each subset containing different training and testing sets with ratio 70% and 30%. They then use Mann–Whitney U and Chi square test to analysis the features ranking of the given dataset. The following figure 4 is the results of tests which indicate that the key features in this dataset are serum creatinine and ejection fraction [1].

Chi squared test

Rank	Feature	*p*-value
1	Ejection fraction	0.000500
2	Serum creatinine	0.000500
3	Serum sodium	0.003998
4	Age	0.005997
5	High blood pressure	0.181909
6	Anaemia	0.260370
7	Creatinine phosphokinase	0.377811
8	Platelets	0.637681
9	Smoking	0.889555
10	Sex	1
11	Diabetes	1

Results of the application of the chi squared test between each feature and the target feature death event

Mann–Whitney *U*

Rank	Feature	Test *p*-value
1	Serum creatinine	0
2	Ejection fraction	0.000001
3	Age	0.000167
4	Serum sodium	0.000293
5	High blood pressure	0.171016
6	Anaemia	0.252970
7	Platelets	0.425559
8	Creatinine phosphokinase	0.684040
9	Smoking	0.828190
10	Sex	0.941292
11	Diabetes	0.973913

Results of the univariate application of the Mann–Whitney *U* test between each feature and the target feature death event

Figure 4. Result of Chi squared and Mann-Whitney U test [1].

4. Results

This research investigates the prediction of survival of patients with heart failure using five machine learning models and two CV methods described in section 3. Table 1 developed by Python is the result of the mean performance of CV of the LR, DT, RF, SVM and ANNs model with all clinical features in dataset [2]. Following the result of those 3 indicators, LR by stratified k-fold has the best performance, followed by RF. On the other side, the overall performance of ANNs is the worst.

Table 1. The mean performance of CV of models with all clinical features

Models	CV method (k=5)	Accuracy	Roc_Auc	Average_Precision
LR	K-fold	0.773164	0.794831	0.598557
LR	Stratified k-fold	0.778927	0.949407	0.927901
DT	K-fold	0.725989	0.563369	0.365184
DT	Stratified k-fold	0.652147	0.609884	0.429141
RF	K-fold	0.779774	0.803650	0.615759
RF	Stratified k-fold	0.692260	0.844030	0.662399
SVM	K-fold	0.679831	0.507070	0.388174
SVM	Stratified k-fold	0.678927	0.493903	0.364091
ANNs	K-fold	0.666271	0.485200	0.340655
ANNS	Stratified k-fold	0.602260	0.546357	0.357167

D. Chicco et al. [1] concluded that survival of patients with heart failure can be predicted from serum creatinine and ejection fraction alone. In addition, they also mentioned that "follow-up time" needs to be involved in prediction. Table 2 developed by Python, provides the result of the mean performance of CV of models with those 3 features alone. It indicates that the models in this research involving features: serum creatinine, ejection fraction, and "follow-up time" alone, are all suitable models to predict the survival of patients with heart failure since the results in Tables 1 and 2 are similar. Following the result, LR by stratified k-fold still has the best performance. However, the overall performance of ANNs in Table 2 is much better than the result in Table 1.

Table 2. The mean performance of CV of models with features: serum creatinine, ejection fraction and time alone

Models	CV method (k=5)	Accuracy	Roc_Auc	Average_Precision
LR	K-fold	0.783164	0.811287	0.645992
LR	Stratified k-fold	0.795593	0.951449	0.900775
DT	K-fold	0.732825	0.604571	0.383645
DT	Stratified k-fold	0.485706	0.448254	0.336998
RF	K-fold	0.782994	0.756041	0.656206
RF	Stratified k-fold	0.652260	0.771537	0.606195
SVM	K-fold	0.709831	0.614475	0.475468
SVM	Stratified k-fold	0.778927	0.685983	0.739230
ANNs	K-fold	0.673164	0.730059	0.575042
ANNS	Stratified k-fold	0.745593	0.903116	0.826336

5. Conclusion

This research, according to all the performance results of the five models, observes that LR is the best-performing model. Furthermore, by feature ranking method, there are three key features: serum creatinine, ejection fraction, and follow-up time. The five models can predict the survival of patients with heart failure from those key features alone since the differences in mean values of indicators between Table 2 and Table 3 are low. Moreover, the overall performance of models using 3 key features alone is better than using all features. Hence, it is concluded that the LR only using key features alone is better than LR using all features. In summary, using LR to predict the survival of patients with heart failure from features only serum creatinine, ejection fraction, and follow-up time is the best strategy for this research problem.

References

[1] D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart failure from. serum creatinine and ejection fraction alone," BMC Medical Informatics and Decision Making, 2020, 20 (1), 16.

[2] "Heart failure clinical records," UCI Machine Learning Repository, 2020.

[3] V. Grgi´c, D. Mu˘si´c, and E. Babovi´c, "Model for predicting heart failure using Random Forest and Logistic Regression algorithms," in IOP Conference Series: Materials Science and Engineering, 2021, 1208, 012039.

[4] F. S. Alotaibi, "Implementation of Machine Learning Model to Predict Heart Failure Disease," International Journal of Advanced Computer Science and Applications, 2019, 10 (6).

[5] D. Mpanya, T. Celik, E. Klug, and H. Ntsinjana, "Machine learning and statistical methods for predicting mortality in heart failure," Heart Failure Reviews, 2021, 26 (3), 545 – 552.

[6] Z. Zhang, "Model building strategy for logistic regression: purposeful selection," Annals of translational medicine, 2016, 4 (6).

[7] B. Charbuty and A. Abdulazeez, "Classification based on decision tree algorithm for machine learning," Journal of Applied Science and Technology Trends, 2021, 2 (01), 20 – 28.

[8] P. Probst and A.-L. Boulesteix, "To tune or not to tune the number of trees in random forest," The Journal of Machine Learning Research, 2017, 18 (1), 6673 – 6690.

[9] A.-L. Boulesteix, S. Janitza, J. Kruppa, and I. R. K¨onig, "Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics," Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2012, 2 (6), 493 – 507.

[10] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, The elements of statistical learning: data mining, inference, and prediction. Springer, 2009, 2.

[11] D. Berrar et al., "Cross-validation." 2019.

[12] R. C. Prati, "Combining feature ranking algorithms through rank aggregation," in the 2012 International Joint Conference on Neural Networks (IJCNN). Brisbane, Australia: IEEE, 2012, 1 – 8.