Machine Learning based on Survival Prediction Model for Heart Failure Patients

Group member: Qinhan Yu、Jie Lu

February 10, 2022

Contents

Abstract......i

1. Introduction......1

1.1 Background......1

1.2 Research Questions......1

2. Methods......1

2.1 Paticipants and Data collection......1

2.2 Data analysis......2

3. Result......2

3.1 Exploratory Data Analysis......2

3.1.1 Descriptive statistics of dataset......2

3.1.2 Distribution of continuous variable......3

3.1.3 Correlation Analysis......4

3.1.4 Test of significance......4

3.2 Unsupervised learning: Cluster analysis......5

3.3 Supervised learning......5

3.3.1 Decision tree......5

3.3.2 Logistic regression......6

3.3.3 Fully connected neural network......6

4. Discussion......7

5. Conclusion......8

6. Reference......9

7. Contribution......9

Abstract

Heart failure, refers to the failure of the systolic or diastolic functions of the heart, and the venous return blood volume can not be fully discharged from the heart, resulting in venous blood stasis and insufficient blood perfusion in the arterial system. At present, heart failure has gradually become younger, so it is very important to study the influencing factors of heart failure. From the perspective of machine learning, this paper tries to study some factors and survival situation of patients with heart failure by building models, for the purpose of making survival prediction for heart failure patients.

This paper firstly carry out Exploratory Data Analysis, then use classification algorithms: decision tree, logistic regression and fully connected neural network to build the model, and then use 5-fold cross validation and receiver operating characteristic curve to compare the models. Finally, it comes to the conclusion that the three-layer neural network model is the best, with the accuracy rate of 96.6667%, and all three models are above 75%, so a better model can be obtained to predict the survival rate of patients with heart failure.

Key words: Machin learning; Classification algorithm; Neural Network; 5-fold cross validation; Data mining

1. Introduction

1.1 Background

Heart failure, refers to the failure of systolic or diastolic function of the heart, which can not fully discharge the venous return blood volume from the heart, resulting in venous blood stasis and insufficient blood perfusion in the arterial system. At present, heart failure has gradually become younger, so it is very important to study the influencing factors of heart failure. Cardiovascular disease, which will eventually lead to heart failure. has become a disease with a high prevalence, about 17 million people die from cardiovascular every year in world. However, the medical community doesn't have a full grasp of the causes, so machine learning is employed to highlight undetected factors and correlations. Therefore, this paper wants to use machine learning to predict the survival rate of patients with heart failure, so as to establish a suitable model.

1.2 Research Questions

Many researchers have studies on heart failure. The purpose of the study is to predict survival of patients with heart failure by using machine learning. According to this, there are three research questions of this paper:

a) What kind of data should be use?
b) How to choose model and what machine learning methods are used?
c) How to evaluate the model?

2. Methods

2.1 Paticipants and Data collection

In order to solve the question a), a suitable data should be proper for analysis and do machine learning. The data set used was from UCI machining learning repository, named 'Heart failure clinical records Data Set'. The data set contained 299 subjects and 13 clinical features, which were shown in Table 1.

Table 1

Introduction of 13 clinical features

Name	Explain	Type
Age	Age of the patient	Integer
Anaemia	Decrease of red blood cells or hemoglobin	Boolean
High blood pressure	If the patient has hypertension	Boolean
CPK	Level of the CPK enzyme in the blood	Integer

Diabetes	If the patient has diabetes	Boolean
Ejection fraction	Percentage of blood leaving the heart at each contraction	Integer
Platelets	Platelets in the blood	Integer
Sex	Woman or man	Boolean
Serum creatinin	Level of serum creatinine in the bloo	Real
Serum sodium	Level of serum sodium in the blood	Integer
Smoking	If the patient smokes or not	Boolean
Time	Follow-up period	Integer
[target] Death event	If the patient deceased during the follow-up period	Boolean

Table 1 shows the basic information and attributes of data. The target death event is Boolean, so classification algorithm or clustering is suitable for solving the problem in this paper.

2.2 Data analysis

To solve question b) and c), this paper analyzed from three aspects: exploratory data analysis(EDA), building the model and evaluating the model. Python and Pytorch were programming language.

Firstly, we used Python for EDA, cluster analysis, and used decision tree, logistic regression to build the model. Meanwhile, we used Pytorch to establish the neural network, and finally the model was evaluated by 5-fold cross validation and receiver operating characteristic curve (ROC). These are the methods used in this paper.

3. Result

3.1 Exploratory Data Analysis

3.1.1 Descriptive statistics of dataset

First, we conducted exploratory data analysis. Using pandas-profiling, which is a package of Python, we found basic information of data.

Table 2

Descriptive statistics of dataset

	N
Number of variables	13
Number of observations	299
Missing cells	0
Missing cells(%)	0%
Duplicate rows	0
Duplicate rows(%)	0%

Table 2 shows there are no missing values in this dataset, so missing values and outliers were not processed in this study.

At the same time, some statistical descriptions of each clinical feature can be obtained from the results. For example, age is one of clinical features in the data set. We can get descriptive statistics of age and obtain the mean value, maximum value, minimum value and so on.

3.1.2 Distribution of continuous variable

Next, we analyzed continuous variables and categorical variables respectively. First of all, we analyzed the distribution of continuous variables. There are seven continuous variables in the data set, which are age, creative phosphokinase, injection fraction, platelets, serum creative, serum sodium and time. We used Python to get the histogram, as shown in figure 1.

Figure 1 shows that there are two features, which are age and creatinine phosphokinase have skewness, so we use Box-Cox transformation to adjust. The adjusted distribution is shown in figure 2.

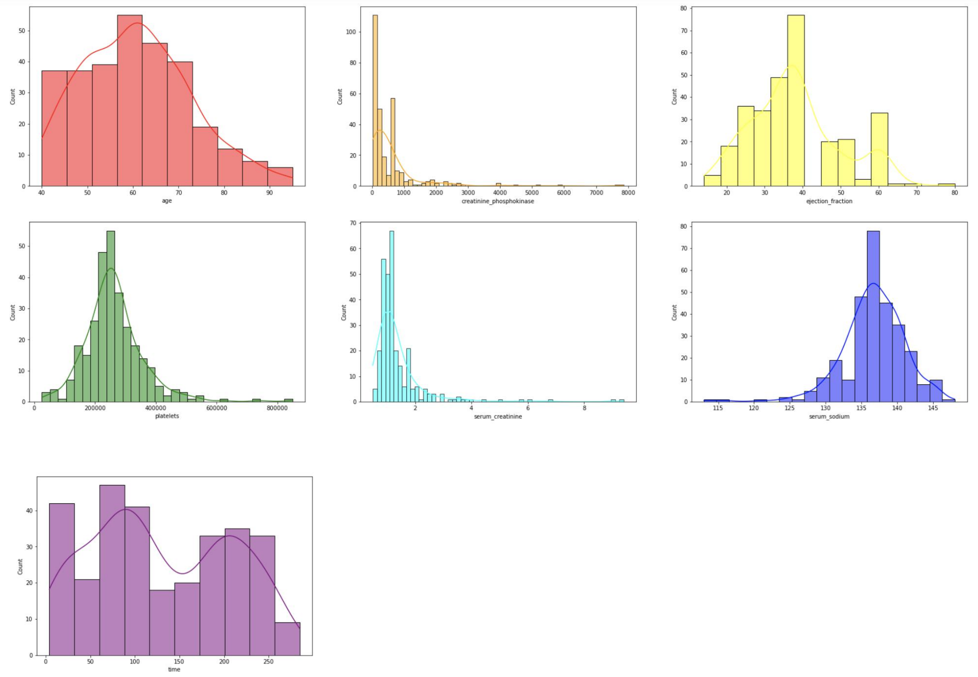

Fig. 1 Distribution of continuous variable

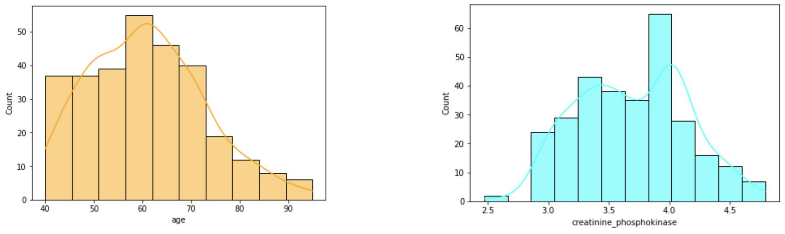

Fig. 2 Adjusted distribution of age and creatinine phosphokinase

3.1.3 Correlation Analysis

For the continuous variables, we made correlation analysis, as shown in figure 3, from which we can get the conclusion that there is no significant relation between any of the continuous variables so we don't need to drop any clinical features.

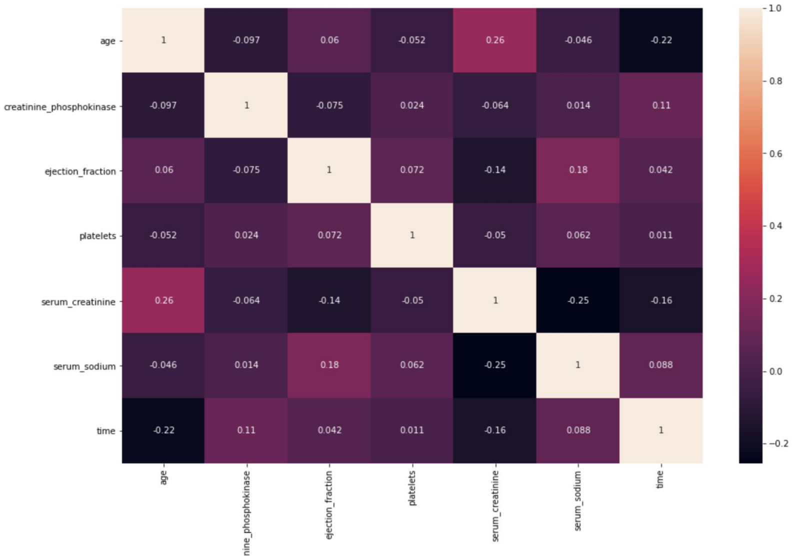

Fig. 3 Heatmap

3.1.4 Test of significance

Then we did significance test, which can help us determine whether the influence between two independent variables was significant.

The chi-square test can be used to determine whether there is significant relation between independent variables, which can simplify the data and make it more universal.

In this paper, we assumed that the original hypothesis H_0 of the chi-square test was that there was no relation between variable 1 and variable 2, and the alternative hypothesis was that there was relation between variable 1 and variable 2. We set the significance level to 0.075, and computer would reject the original hypothesis when the p value is less than 0.075, which meant that there was correlation between variable 1 and variable 2.

We used Python to test the prescription. In the test, we found that there was a case where the p value between smoking and other categorical variables is less than 0.075, as shown in Table 3.

Table 3

p value between smoking and other categorical variables

	Anaemia	Diaetes	High_blood_pressure	Sex
Smoking	0.1618	0.0522	0.4378	4.8999e-06

Table 3 shows that smoking and diaetes, smoking and sex have a significant relationship. When building the model, we can choose to leave only the smoking column to simplify the model.

3.2 Unsupervised learning: Cluster analysis

Because the target variable is categorical variable, we used Python for cluster analysis. Here is the result graph in figure 4.

Intuitively, cluster analysis divides the results into two categories, and we continue to compare the results with the real data, as shown in figure 5.

In figure 5, there should be two of them that are not available, but the distribution in figure 5 is not obvious. So the result is not optimistic. We should consider other machine learning algorithms.

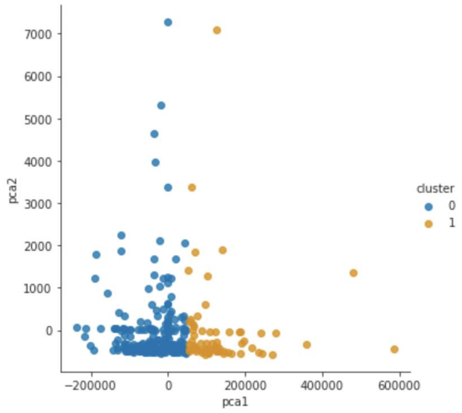

Fig. 4 Result of cluster analysis

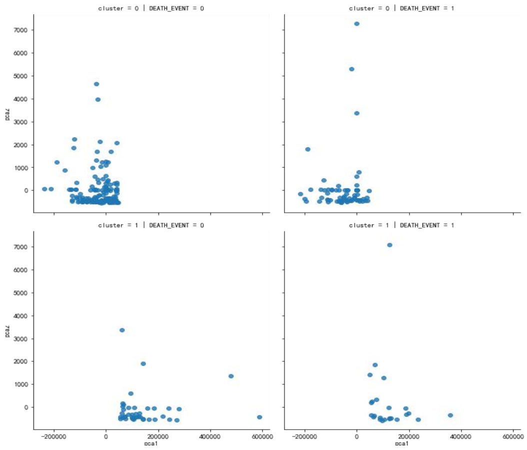

Fig. 5 Comparison graph with real value

3.3 Supervised learning

3.3.1 Decision tree

Decision tree is one of the most commonly used categorical algorithms in supervised learning. It is mainly used when the target variable is a classification variable.

In this paper, we used Python to build decision tree model. First, we divided the processed data into two data sets, and the training set and test set are divided into 7:3. Then, we built decision tree with a depth of 5, and the programming results show that the accuracy rate of train set is 93.78%, and the accuracy rate of test set is 82.22%. In order to get the classification of decision tree more concretely, we visualized the tree, and the visualization result is shown in figure 7.

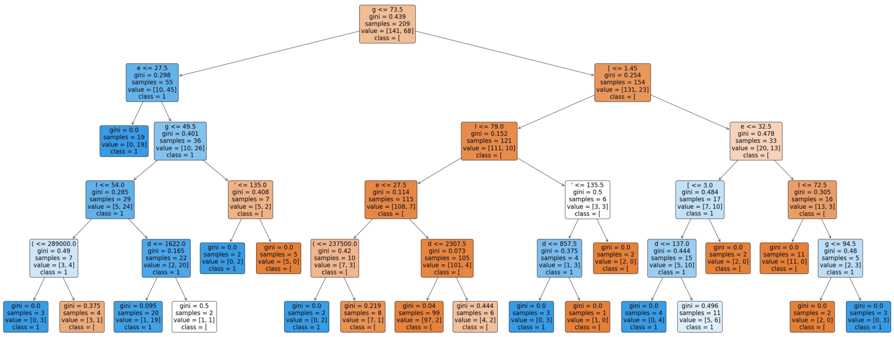

Fig. 7 Visual tree diagram

3.3.2 Logistic regression

Logistic regression is also a good classification algorithm in supervised learning. In this paper, we used Python to build logistic regression model. The training set and test set are divided into 7:3. Then, we built logistic regression, and the programming results show that the accuracy rate of train set is 84.21%, and the accuracy rate of test set is 83.33%. At the same time, we analyzed the importance of variables, and found that age and time are more important in the model, which is shown in figure 8.

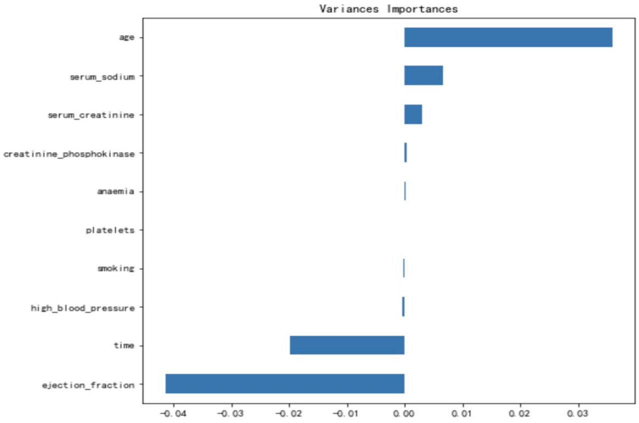

Fig. 8 Variances importances

3.3.3 Fully connected neural network

In order to judge whether death will occur due to heart failure according to the patient’s condition, we also used a fully connected neural network model to train and test it. In this part, we mainly built a fully connected neural network with one, two and three hidden layers, and implement the code through Pytorch. Since the data types of each feature in the dataset we used were different, and the span of the data were also quite different. So in order to train the model better, we normalized the data.

First, because when we set the number of hidden layer neurons to 84 in the previous training of the fully connected neural network with only one hidden layer, the training effect was better, in the hyperparameter setting, the number of layer neurons was set to 84. Then built the network and used relu() function as its activation function. According to the result we trained, we can find that the distribution range of loss values were basically at the level around 0.4, 0.5 and 0.6. After that, we tested the network which we trained. We can find that the accuracy is around 81%.

Second, we built a fully connected neural network with two hidden layers. We also set the number of neurons in each layer to 84 and used the relu() function as the activation function. Also, according to the result we trained, loss values of model were rarely greater than 0.6 and the accuracy we calculate is about 83%.

When building a fully neural network with three hidden layers, the number of neurons in each layer was still 84, and relu() function was used as the activation function. The loss values of this model were concentrated at the level of 0.1 and 0.2, and the accuracy rate can be as high as 91%.

4. Discussion

In the previous content, we used the heart failure dataset to train and test the cluster analysis, decision tree, logistic regression and neural network models. It can be seen from the model test results that the effect of learning the heart failure dataset by cluster analysis is not good. Using decision tree, logistic regression and neural network models, the final learning effect has reached a good level. In addition, by comparing the final training effects of the three models, we can also see that the learning effect of decision tree is slightly low, the learning effect of logistic regression is better than decision tree, and the learning effect of the neural network model is better than the previous two.

However, it is inaccurate to judge the learning effect of the model only by the accuracy rate we calculated. In order to make the results more accurate, we mode a more detailed study of the accuracy of these three models.

First, we used a 5-fold cross validation. We use Weka to apply decision tree and logistic regression, and average the five scores obtained by fully connected neural network. According to our test, for the decision tree, it ended up with an average score of 78%, the average scores of logistic regression is around 82%, and the average scores of fully connected neural network is 96%, which is concluded in Table 4.

Table 4

Accuracy Rate of 5-fold cross validation

	Decision tree	Logistic regression	FC neural network
Accuracy Rate	78.929%	82.2742 %	96.6667%

This result clearly shows that the accuracy of the fully connected neural network is significantly higher than the previous two models.

Second, we also plot the ROC curves of the three models. As shown in figure 9, the area under the ROC curve of the decision tree is 0.78, the area under the ROC curve of the logistic regression is 0.83 and the area under the fully connected neural network is 0.88. According to this result, we can also get that the training effect of fully connected neural network is the best, which is same as the results we obtained before.

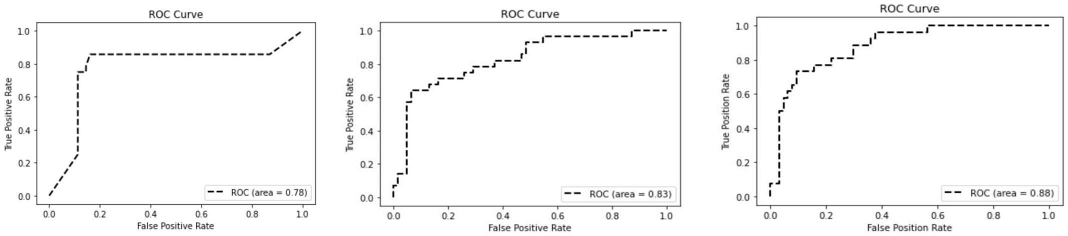

Fig. 9 ROC of three models

In addition to the comparison among the different models, the models themselves also have certain contrasts. In the process of constructing fully connected neural network, we created neural networks with one, two and three hidden layers. In the previous tests, we got an accuracy of about 0.81 for a fully connected neural network with one hidden layer, about 0.83 for a fully connected neural network with two hidden layers, and about 0.91 for a fully connected neural network. From this result, we can also see that without overfitting, when there are more hidden layers of the neural network, that is to say, the more complex the structure of the neural network is, the better the learning effect will be, and the ability to predict is also stronger.

5. Conclusion

All in all, among cluster analysis, decision tree, logistic regression and neural network which we chose for this project, the cluster analysis is not ideal for the learning effect of the heart failure dataset, but the learning effect of the other three types of models is good, and the accuracies of these three models are acceptable, which can achieve above 80%. But whether it is by predicting the ratio of the correct number to the total number during the test, or by the method of 5-fold cross validation, or by the way of drawing the ROC curve, the fully connected neural network has shown highest performance.

For the fully connected neural network itself, if we can appropriately increasing the number of hidden layers to make the network more complex without causing over-fitting, then we can also get better training results.

6. Reference

[1]Tanvir Ahmad, Assia Munir, Sajjad Haider Bhatti, Muhammad Aftab, and Muhammad Ali Raza: "Survival analysis of heart failure patients: a case study". PLoS ONE 12(7), 0181001 (2017).

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

[3]孙铁铮,于泽灏.基于机器学习的心脏病例分类预测研究[J].电脑知识与技术,2021,17(26):96-97+104.DOI:10.14004/j.cnki.ckt.2021.2607.

[4]黄嵩.基于机器学习的心脏病预诊系统的研究[D].浙江理工大学,2021.DOI:10.27786/d.cnki.gzjlg.2021.000522.

[5]叶苏婷,潘媛媛,毕迎春.基于决策树算法的心脏病发病预警模型研究[J].电脑知识与技术,2020,16(19):187-189.DOI:10.14004/j.cnki.ckt.2020.2325.

[6] Qin, Zhang, L., Liu, M., Xu, Z., & Liu, G. (2021). ASFold-DNN: Protein Fold Recognition based on Evolutionary Features with Variable Parameters using Full Connected Neural Network. IEEE/ACM Transactions on Computational Biology and Bioinformatics, PP, 1–1. https://doi.org/10.1109/TCBB.2021.3089168

[7]Anderson, Mason, J. C., & Ellacott, S. (1997). Mathematics of neural networks : models, algorithms, and applications / edited by Stephen W. Ellacott, John C. Mason, Iain J. Anderson. (1st ed. 1997.). Springer Science Business Media B. V. https://doi.org/10.1007/978-1-4615-6099-9

7. Contribution

The topic setting, process, presentation and report of this project are all completed by two team members, and the specific division of labor is as follows.

Topic: Qinhan Yu, Jie Lu

Introduction: Jie Lu

Methodology: Qinhan Yu

Result:

EDA & Decision tree & Logistic &Cluster : Qinhan Yu

Neural network(one hidden layer & two hidden layer & three hidden layer): Jie Lu

Discussion: Qinhan Yu, Jie Lu

Conclusion: Jie Lu

Slides: Qinhan Yu, Jie Lu

Report: Qinhan Yu, Jie Lu

Additional(Code link): *https://github.com/fymobo/ML_Project/blob/main/project.ipynb*