Research on Predictive Model Based on Ensemble Learning

Jingyi Zhu[a, #], Anbo Zhang[b, #], Haixin Zheng[*, #]

Zhengzhou University Zhengzhou, China

*Corresponding Author Email: historyhaixin@163.com, [a]jingyizhu95@gmail.com, [b]zzhanganbo325@gmail.com

Abstract. Ensemble learning completes learning tasks by building and combining multiple learners. The use of ensemble learning can make accurate prediction. This paper used the dataset publicly available on kaggle platform. Firstly, this paper preprocessed and performed descriptive statistics on the dataset, based on which this research constructed the prediction model. Three ensemble learning models Random Forest, AdaBoost, and LightGBM were selected to study the data. To prevent overfitting, a 10-fold cross-validation method was used to train the classifiers and the models were tuned using the grid search method. Finally, the three models were compared in terms of Accuracy, Precision, Recall, F1-score, ROC curve and AUC values. The comparison shows that all three models have good performance, and the accuracy of all model predictions are higher than 80%. However, there is a slight difference in classification ability among the models. Random Forest performs best, with an Accuracy of 86.94, Precision of 85.91, Recall of 93.10, F1-score of 0.8936, and AUC of 0.8906. All evaluation indexes are higher, which also verify the feasibility of using ensemble learning algorithms in prediction.

Keywords: Ensemble learning, random forest, 10-fold cross-validation, Grid Search, ROC curves.

1. Introduction

Since 1950, Turing asked "Can machines think?" in Computing Machinery and Intelligence. This question has led to rapid development of artificial intelligence techniques [1]. Machine learning (ML) methods learn to discover patterns in large amounts of data [2]. Various outcomes can be predicted flexibly based on appropriate training, and laws covering all learning systems are computed by induction on learning experience [3,4]. However, since the generalization ability or robustness of individual learners is often poor, some studies have combined multiple learners with certain strategies to form ensemble models to improve the problem-solving ability of the learners [5]. Commonly used integration methods are boosting, bagging, stacking,blending etc.

Ensemble learning is applicable to both supervised learning (classification, regression) and unsupervised learning (clustering), which is a learning process by integrating a set of models (learners) in a specific way to obtain the final prediction results [6]. The prediction effect of ensemble learning algorithm is particularly outstanding [7]. Zheng established a predictive model and developed a corresponding prediction Web system based on ensemble learning algorithms [8]. Gupta et al compared the prediction performance of various machine learning methods including Support Vector Machine (SVM), Naive Bayes (NB), Random Forest (RF) and Gradient Boosting Decision Tree (GBDT) [9]. Yu et al. proposed a joint ensemble learning model that combines basic machine learning model and Boosting algorithm with Stacking to predict [10]. Chen et al. used five algorithms, namely, plain Bayesian, linear regression, decision tree, random forest and gradient-enhanced decision tree, to make predictions using 84 indicator tests, and the AUC of the model can be as high as 0.991 [11].Given the high dimensionality and complex relationships of data, this paper builds a prediction model based on ensemble learning algorithms which can help assess risk and optimize resource utilization.

2. Objects and Methods

2.1. Missing data handling and descriptive statistics

The identification of missing values was performed on the dataset and it was found that there were no missing values in the dataset and no work was needed to fill in the missing values. Subsequently, simple descriptive statistics were performed on the dataset, and the mean, median, and most values of the data were characterized as shown in Table I.

The distribution of the target population can be derived from the table.

Table 1. Descriptive statistics

	count	mean	std	min	50%	max
age	299.0	60.833893	11.894809	40.0	60.0	95.0
anaemia	299.0	0.431438	0.496107	0.0	0.0	1.0
creatinine_phosphokinase	299.0	581.839465	970.287881	23.0	250.0	7861.0
diabetes	299.0	0.418060	0.494067	0.0	0.0	1.0
ejection_fraction	299.0	38.083612	11.834841	14.0	38.0	80.0
high_blood_pressure	299.0	0.351171	0.478136	0.0	0.0	1.0
platelets	299.0	263358.029264	97804.236869	25100.0	262000.0	850000.0
serum_creatinine	299.0	1.393880	1.034510	0.5	1.1	9.4
serum_sodium	299.0	136.625418	4.412477	113.0	137.0	148.0
sex	299.0	0.648829	0.478136	0.0	1.0	1.0
smoking	299.0	0.321070	0.467670	0.0	0.0	1.0
time	299.0	130.260870	77.614208	4.0	115.0	285.0
DEATH_EVENT	299.0	0.321070	0.467670	0.0	0.0	1.0

2.2. Data Exploration

Based on the above analysis, the following will explore the relationship between each characteristic indicator variable. For the continuous characteristic indicator variables to observe the Pearson correlation coefficient, before doing the Pearson correlation analysis, we need to check whether they obey normal distribution first.

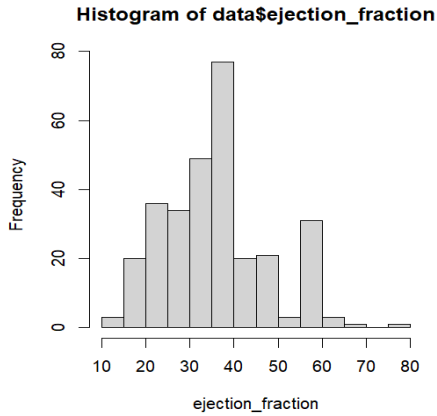

Figure 1. Histogram of ejection fraction

By observing the above graph it can be seen that the approximation obeys the normal distribution. To make the model results more accurate, we use the JB test to check whether the distribution obeys the normal distribution.

2.3. JB test for normal distribution

The steps to perform the hypothesis test are as follows.

H_0:The random variable obeys a normal distribution H_1 :The random variable does not obey a normal distribution

Then calculate the skewness and kurtosis of the variable, get the test value JB*, and calculate its corresponding p-value compare the p-value with 0.05, if it is less than 0.05 then the original hypothesis can be rejected, otherwise we cannot reject the original hypothesis. Using MATLAB, we can get p=0.500>0.05, so the original hypothesis cannot be rejected, so the random variable can be considered to obey the normal distribution. The other continuous variables also conform to the normal distribution.

2.4. Correlation analysis

Correlation analysis can be performed if the normal distribution is met, and the correlation plot of the numerical attributes shows that there is a certain correlation between the characteristic indicator variables, and whether or not death occurs during the follow-up period is related to the length of the follow-up period, whether or not the patient suffers from the underlying disease, the condition of the physical indicators, and the living habits. At the same time, there was no significant covariance among the characteristic indicator variables, so all all characteristic indicator variables were retained for the next step of the study.

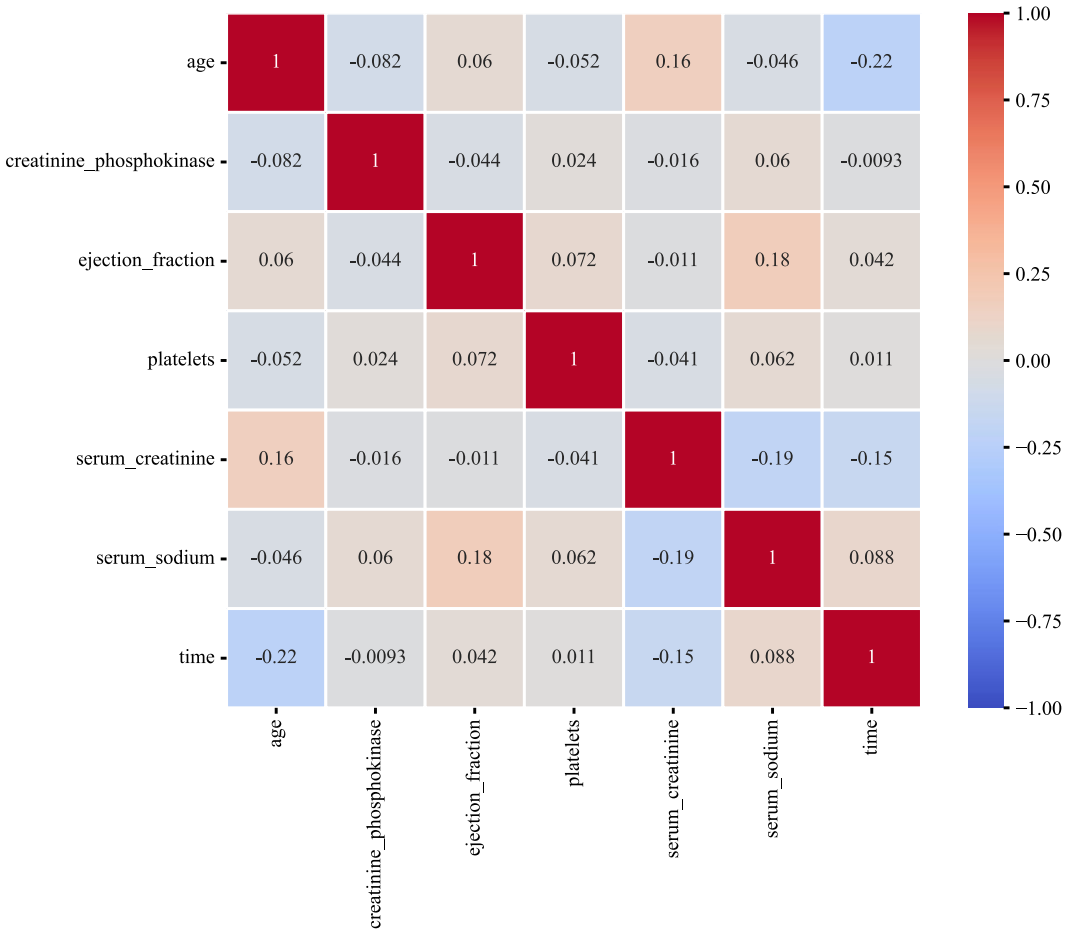

Figure 2. Pearson correlation coefficient between numeric variables

The relationship between the discrete characteristic indicator variables and the target variables can be studied by categorical bar charts, as shown in Figure 3. From the figure, it can be seen that there is no significant difference in survival status between genders, but there are differences in survival status for different physical indicators.

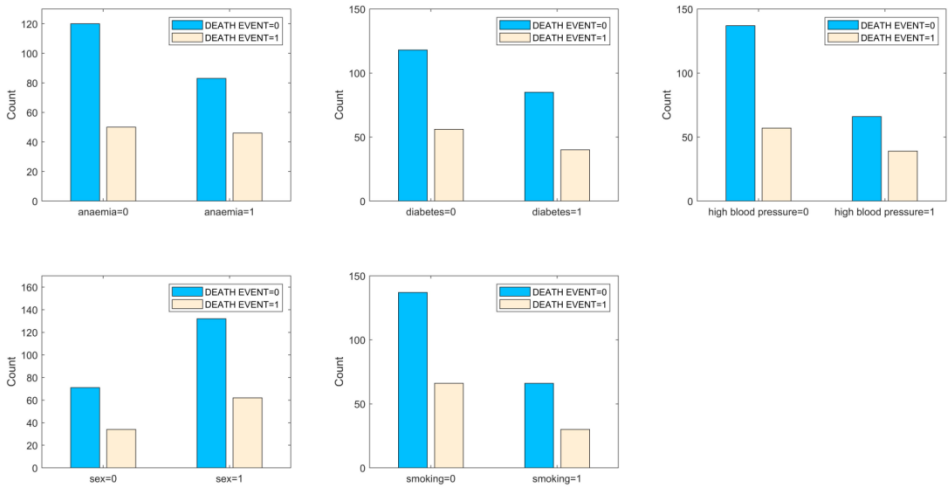

Figure 3. Bar chart of classification

3. Construction of Prediction Model of Death Probability

The relationship between each characteristic indicator variable and the target variable can be roughly derived from the above descriptive analysis, so the prediction model can be established based on the ensemble learning algorithm .

3.1. Model introduction

3.1.1. Random forest

Random forest is an evolutionary version of the Bagging algorithm, the idea is still Bagging, that is, based on the use of decision trees, randomly sampled from the data set, build multiple decision trees, training each decision tree in the model, for the classification problem, using a simple voting method as shown in Equation (1), the category with the most votes or one of the categories is the final model output . The main features of this model are faster processing of high-dimensional data and better stochasticity and noise immunity of the model, which is less likely to be overfitted [12].

$$H(x) = \frac{1}{T}\sum_{t=1}^{T} h_t(x) \tag{1}$$

3.1.2. AdaBoost

AdaBoost is the most famous representative of boosting algorithm, which is an integrated algorithm for classification problems. It constructs a strong classifier with a linear combination of weak classifiers as shown in Equation (2), and the performance of the weak classifiers need not be too good, so that a very accurate strong classifier can be constructed. The main feature is that the division of labor is clear, the training of the latter model is done on the basis of the former model. And the strong classifier is constructed when a specified number of iterations or an expected error rate is reached [13]. The ceiling of the model is high, and it is good at solving difficult problems.

$$H(x) = \sum_{t=1}^{T} \alpha_t h_t(x) \tag{2}$$

3.1.3. LightGBM

LightGBM is a distributed gradient boosting algorithm in machine learning, GBDT (gradi- ent boosting decision tree) based on gradient boosting tree [14]. Unlike the traditional decision tree model, the LightGBM algorithm uses a depth-first splitting strategy, as shown in Figure 4, which means that the global samples are considered each time the leaf nodes are split, reducing the possibility of the number of post-pruning operations. It uses an additive model as well as continuous iteration to reduce the error residuals of the previous round to achieve the effect of regressing or classifying the data, which has the advantages of good training effect, not easy to overfit, and flexible handling of various types of data.

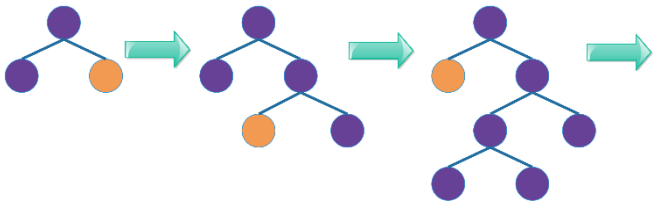

Figure 4. Depth-first split strategy

The three ensemble learning algorithms have obvious differences in model design, each with its own advantages and shortcomings, and their core differences are shown in Table II.

Table 2. Comparison of Three Integrated Algorithms

Algorithm	Model thought	Segmentation algorithm	Gain calculation
Random forest	Bagging	Random segmentation	Gini coefficient
AdaBoost	Boosting	pre-sorted	Optimization derivation formula
LightGBM	Boosting	histogram	Optimization derivation formula

3.2. Selection of performance Evaluation Indexes of Model

The various indexes of each model are compared and the optimal model is selected. The optimal model will be tuned to obtain the optimal results.

The confusion matrix is used to classify the prediction results, which are divided into true positive (TP), true negative (TN), false positive (FP), and false negative (FN). As shown in Table III.

Table 3. Classification Result Confusion Matrix

Classification	Forecast result	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

The Accuracy rate is the ratio of correctly classified samples to the total samples in the classifier. The calculation formula is

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

The prediction results are evaluated by several indexes: Precision, Recall, F1-score, ROC curve and AUC. The specific formulas are as (4) to (8).

The precision P and recall R are defined as

$$P = \frac{TP}{TP + FP} \tag{4}$$

$$R = \frac{TP}{TP + FN} \tag{5}$$

Among them, precision and recall are contradictory metrics, in general, recall tends to be low when precision is high, and precision tends to be low when recall is high. This leads to the introduction of a performance metric that combines both, namely F1-score, which is defined as

$$F1 - score = \frac{2 \times P \times R}{P + R} \tag{6}$$

For the binary classification problem, the Sensitivity (Threshold) of the classifier can be adjusted to obtain different classification results, and the accuracies under various sensitivities are connected into a curve which is the ROC curve. The ROC curve takes the True Positive Rate (TPR) as the y-axis and the False Positive Rate (FPR) as the x-axis. The two formulas are

$$TPR = \frac{TP}{TP + FN} \tag{7}$$

$$FPR = \frac{FP}{TN + FP} \tag{8}$$

The advantage of the ROC curve is that when the distribution of the two types of samples changes, the curve does not change much, so that the model itself can be evaluated more objectively. At the same time, the area enclosed by the curve and the x-axis can be observed to directly observe the goodness of the model. The larger the enclosed area, the better the performance of the model. As shown in Figure 5, A, B, and C respectively represent different models. The area enclosed by curves A, B, and C decreases in order, and the model performance becomes worse in order.

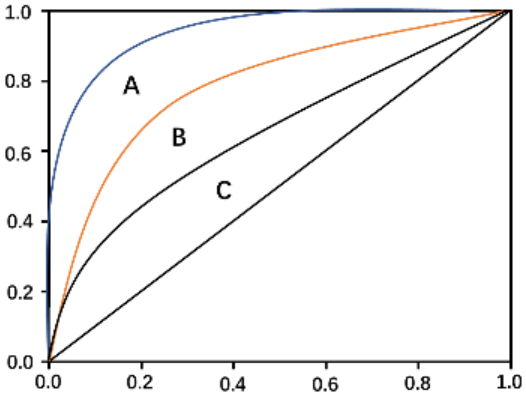

Figure 5. The comparison of ROC curve

3.3. Model training and prediction

After determining the model parameters, this paper is based on Python language and combined with sklearn machine learning library for training. Meanwhile, in order to prevent overfitting, a 10-fold cross-validation method is used for training, and obtain the corresponding prediction results and evaluation indexes, as shown in Tables IV~VI.

Table 4. Random Forest Evaluation Indicators

Accruracy (%)	Precision (%)	Recall (%)	F1-score	AUC
84.30	87.14	90.14	0.8867	0.8967

Table 5. Adaboost Evaluation Indicators

Accruracy (%)	Precision (%)	Recall (%)	F1-score	AUC
79.90	84.29	87.19	0.8571	0.8533

Table 6. Lightgbm Evaluation Indicators

Accruracy (%)	Precision (%)	Recall (%)	F1-score	AUC
80.62	80.59	75.28	0.7701	0.8859

4. Adjustment of Model Parameters

In order to ensure that the model has a better effect, it is necessary to tune the parameters of the model, and this paper uses the grid search method to adjust several parameters that have a greater impact on the model. The parameters after tuning are shown in Table VII. The comparison of the models after tuning is shown in Table VIII.

Table 7. Parameter Value

Parameter Name	Parameter Description	Random Forest	AdaBoost	LightGBM
n_estimators	number of trees	10	29	100
max_depth	maximum depth of tree	300	20	2
min_samples_split	Minimum number of samples to separate	2	2	2
min_samples_leaf	Minimum number of samples on leaf nodes after separation	1	1	1
learning_rate	learning rate	0.1	0.1	0.1

Table 8. Comparison of Models After Parameter Adjustment

	Accruracy (%)	Precision (%)	Recall (%)	F1-score	AUC
Random forest	86.94	85.91	93.10	0.8936	0.8906
AdaBoost	81.70	84.98	88.76	0.8926	0.8760
LightGBM	81.66	82.69	79.48	0.7984	0.8863

The ROC curves of the three models after tuning the parameters are shown in Figure 6.

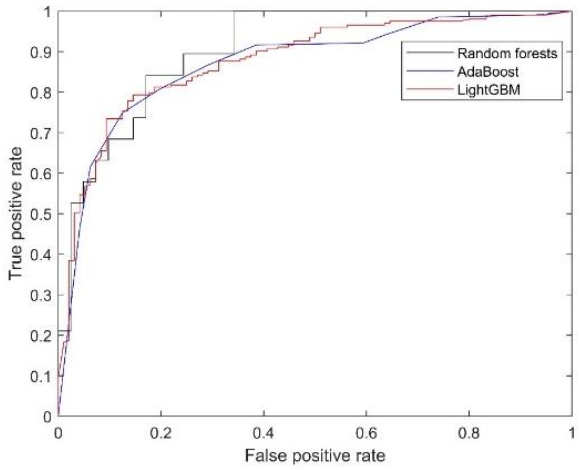

Figure 6. The roc curves of the three models

By comparing the evaluation indexes of each model and observing the ROC curve above, it is found that all three models have good performance. The accuracy of Random Forest, AdaBoost, and LightGBM goes up respectively and all exceed 80%. Random Forest has the best performance with an Accuracy of 86.94, Precision of 85.91, Recall of 93.10, F1-score of 0.8936, and AUC of 0.8906, which are all slightly higher than the other two models, so Random Forest is the better performing model.

5. Conclusions

This paper uses data from the kaggle platform for the example analysis. In the construction of the prediction models, three ensemble learning models, Random Forest, AdaBoost and LightGBM were selected. The models were optimized using the grid search method, and finally the three models were compared in terms of Accuracy, Precision, Recall, F1-score, ROC curve, and AUC values. The comparision shows that all three models have good performance, and the accuracy of all model predictions are higher than 80%. However, there is a slight difference in classification ability among the models. Random Forest performing the best, because all evaluation indexes are higher. Random Forest performs the best with an Accuracy of 86.94, Precision of 85.91, Recall of 93.10, F1-score of 0.8936, and AUC of 0.8906

However, there are some limitations in this study. Firstly, the data used are from open platforms, which have some limitations in terms of data quantity, quality and applicability. This research will consider using real big data for model construction and prediction in the future. Secondly, the algorithmic models used in this paper are all in the category of ensemble learning. This research can consider selecting different types of machine learning algorithms for improvement and comparison in the future to build better risk prediction models.

References

[1] Turing AM, Haugeland J. Computing machinery and intelligence. Cambridge, MA: MIT Press, 1950.

[2] Efimov, I.R.; Fu, S.N.; Laughner, J.I. (Eds.) Cardiac Bioelectric Therapy: Mechanisms and Practical Implications; Springer: Berlin/Heidelberg, Germany, 2021; pp. 335–352.

[3] Au-Yeung, W.-T.M.; Sahani, A.K.; Isselbacher, E.M.; Armoundas, A.A. Reduction of false alarms in the intensive care unit using an optimized machine learning based approach. NPJ Digit. Med. 2019, 2, 1–5.

[4] Au-Yeung, W.-T.M.; Sevakula, R.K.; Sahani, A.K.; Kassab, M.; Boyer, R.; Isselbacher, E.M.; Armoundas, A. Real-time machine learning-based intensive care unit alarm classification without prior knowledge of the underlying rhythm. Eur. Hear. J. Digit. Health 2021, 2, 437–445.

[5] NAG K, PAL N R. A multiobjective genetic programming-based ensemble for simultaneous feature selection and classification[J]. IEEE Transactions on Cybernetics, 2015, 46(2): 499-510.

[6] MENDES-MOREIRA J, SOARES C, JORGE A M, et al. Ensemble approaches for regression: a survey[J]. ACM Computing Surveys, 2012, 45(1): 1-40.

[7] Liu, AY, Yang SW, Li LZ. Research advances in the application of machine learning in disease prediction[J]. Journal of Nursing,2021,28(7):30-34.

[8] Xiaoyan Zheng. Research on cardiovascular disease prediction system based on machine learning[D]. Beijing Jiaotong University,2018.

[9] S. Gupta, D.T. Ko, P. Azizi, et al. Evaluation of machine learning algorithms for predicting readmission after acute myocardial infarction using routinely collected clinical data[J/OL]. Canadian Journal of Cardiology, 2019.

[10] K. Yu, X. Xie. Predicting hospital readmission: a joint ensemble-learning model[J]. IEEE Journal of Biomedical and Health Informatics, 2020, 24(2): 447–456.

[11] Chen Q, Zhang B, Yang J, et al. Predicting intensive care unit length of stay after acute type A aortic dissection surgery using machine learning. Front Cardiovasc Med, 2021, 8: 675431.

[12] Cheng Wang, Rui Gao. Research on random forest improvement algorithm based on feature simplification [J]. Computer Technology and Development, 2020, 30(3): 40-45.

[13] Miaoyan Zhang. Research on railroad fastener ensemble detection method based on Adaboost [D]. Lanzhou Jiaotong University,2018.

[14] Peng C, Zhan WL, Zhou XH. Research and im- plementation of abnormal mail detection method based on random forest algorithm [J]. Journal of Hunan uni- versity of technology, 2020, 34(1):70-76.