ISSN 2616-5775 Vol. 6, Issue 13: 120-126, DOI: 10.25236/AJCIS.2023.061318

Study on the environmental influence factors of stroke itself based on random forest

Weihan Wang[1,*], Wenjing Li[2]

[1]Maynooth International Engineering College, Fuzhou University, Fuzhou, Fujian, 350112, China
[2]Software College, Taiyuan University of Technology, Taiyuan, Shanxi, 030000, China
[*]Corresponding author: 1105263656@qq.com

Abstract: *As a cerebrovascular disease, the number of stroke patients in China has always ranked first in the world, affecting people's life safety. It is very important to predict the population of stroke to prevent stroke and control the number of stroke patients. Different from the conventional method of regression model prediction, this paper uses random forest classification model to predict stroke patients from the perspective of machine learning model. In order to get a model with higher accuracy, this paper adjusts the parameters of cross-validation, classification evaluation and so on to get the optimal model. Experiments have shown that random forest has an improved accuracy rate of 1.9% to 27% compared to other machine learning models. The model can reduce the medical cost, improve the prevention system, and the model ideas can be used to predict similar diseases, and has a certain degree of simulation.*

Keywords: *stroke, population prediction, machine learning, random forest, parameter regulation*

1. Introduction

Stroke, also known as stroke, refers to the interruption of blood supply to the brain caused by cerebrovascular diseases, resulting in brain tissue damage and dysfunction. It can be divided into two main types: ischemic stroke and hemorrhagic stroke. With the aging of population and the acceleration of urbanization, stroke is becoming more and more prevalent, and the burden of stroke disease in China is increasing. In the context of the rapid development of science and technology, the curative effect of stroke has been significantly improved. However, due to the large population base, the number of stroke patients in China still ranks first in the world, and the prevalence rate continues to rise. According to statistics, by 2019, the number of people aged 40 and above suffering from or having suffered from stroke was about 17.04 million[1]. Therefore, while improving the treatment effect of stroke, it is particularly important to study the pathogenesis factors so as to achieve prevention and prediction.

The research on the pathogenesis of stroke at home and abroad has gradually become perfect, which is mainly predicted from the perspective of congenital heredity, living environment and habits. However, the living environment, as a form of external cause, affects the body function and causes the disease, which is different for different individuals[2]. In addition, stroke has the characteristics of chronic non-infection, and the influence of congenital genetic perspective is not a decisive factor. Therefore, in this paper, we only start from the living habits, marriage, smoking, basic diseases, combined with age and gender to predict the disease. Using the form of big data screening, reduce the costs and time costs generated by patient physical examination, as well as the hospital's labor costs and medical costs.

For the prediction of diseases, most of the current models use Cox regression method for prediction, and do not carry out early screening and prevention of stroke [3]. However, the birth of machine learning has derived more methods that can be predicted and classified, providing conditions for the early screening and prevention of stroke. Machine learning can capture complex non-linear relationships to better fit data and make predictions. Machine learning algorithms can also automatically learn from data and extract useful features without human intervention. This is more suitable for the accumulation of subsequent case data and the update iteration of the model when new parameters are introduced.

Stroke disease prediction is a typical binary prediction problem with many influencing factors, diverse data categories and large amount of data. On the other hand, random forest, as a machine learning method using ensemble algorithm, has randomness in the selection of data and features, which overcomes the shortcomings of traditional variable selection methods. At the same time, the

ISSN 2616-5775 Vol. 6, Issue 13: 120-126, DOI: 10.25236/AJCIS.2023.061318

interpretability of the model can be improved by feature importance ranking and partial dependence graph. Compared with other models, the model has higher accuracy in predicting stroke. Moreover, it has good tolerance to noise and is not prone to overfitting [4-5]. In this paper, the random forest method is used to model the stroke problem and obtain a relatively accurate model. This conclusion is helpful for hospitals to make good inference and examination of suspected stroke patients, and it is also helpful for people to adjust their living habits and reduce the possibility of stroke.

2. Data introduction

2.1 Data source and preprocessing

There are many risk factors for stroke, and the data types are varied. The data in this paper were mainly analyzed from gender, age, physical conditions and living conditions. Physical conditions included whether people had high blood pressure or diabetes, blood sugar level and BMI level, and living conditions were subdivided into marital status, type of work, place of residence and smoking status. Nearly 5000 data were drawn from the stroke patients' disease dataset in the whale community.

Delete the invalid data. This paper focuses on the influence of acquired factors on the incidence of stroke, so the age of the samples is set to be between 20 and 80 years old, and MAI outliers are identified for blood sugar level and BMI level. It is assumed that the data follow normal distribution, so that the abnormal points fall in 50% of the area on both sides, and the normal values fall in the middle 50%. Finally, 3899 data were obtained by eliminating outliers. In the data, 2688 people did not have stroke, 211 patients had stroke, the data presented imbalance, so use SMOTE to deal with the imbalance of the data and finally get 7376 data.

2.2 Characteristics and analysis of data

Through descriptive analysis of the above data, the statistical table of frequency, as shown in Table 1, shows that the sample frequency has no special circumstances and is suitable for the subsequent model establishment.

Table 1: Frequency Statistics

Name	Option	Frequency	Name	Option	Frequency
Age	[20.0,35.0)	865	Heart Disease	FALSE	7002
	[35.0,50.0)	1232		TRUE	374
	[50.0,65.0)	2278	Marriage Situation	TRUE	5928
	[65.0,80.0]	3001		FALSE	1448
Blood Sugar Level	[55.12,103.43)	4203	Type of Work	Individual Enterprise	4882
	[103.43,151.75)	1395		Government Sector	1482
	[151.75,200.06)	746		Self-employed Rerson	1012
	[200.06,248.37]	1032	Housing Type	City	4862
BMI	[14.1,22.8)	556		Rural Area	2514
	[22.8,31.5)	4315	Smoking Status	Never Smoke	3093
	[31.5,40.2)	2055		Not Quite Clear	1834
	[40.2,48.9]	450		Smoke	1371
Sex	Female	5190		Before Smoking	1078
	Male	2186	Cerebral Apoplexy	FALSE	3688
Hypertension	FALSE	6685		TRUE	3688
	TRUE	691	Amount		7376

3. Regulation of random forest parameters

3.1 Random forest introduction

Random forest model is an integrated algorithm proposed by Breiman, which constructs multiple tree models based on decision trees and makes predictions by combining the results of decision trees. Compared with other machine learning algorithms, random forest model does not need to make

ISSN 2616-5775 Vol. 6, Issue 13: 120-126, DOI: 10.25236/AJCIS.2023.061318

specific assumptions about the data, and has the advantages of strong interpretability, high tolerance and ability to process various types of variables, which is exactly what this paper needs [6].

In addition, the random forest model can effectively deal with data outliers and noise, avoid variable collinearity and overfitting problems, and do not require data to meet the normality hypothesis. The prediction model used for binary prediction problems has good adaptability [7]. In order to get a more suitable random forest model, this paper continues to adjust the parameters of random forest to get a more accurate model.

3.2 Cross check and node split evaluation criteria

Random forest cross-validation divides the dataset into a training set and a validation set, and uses different subsets to train and validate the model multiple times, so as to obtain more accurate model evaluation results. This avoids the need for a single training set. Choosing an appropriate number of folds can improve the accuracy of the model and prevent the occurrence of overfitting.

The node splitting evaluation criterion is used to measure the quality of node splitting in random forest, and the optimal splitting point is selected according to different criteria. Different evaluation criteria will lead to different random forest structures, which will affect the generalization ability of the model. However, the node splitting evaluation criteria may lead to overfitting, so matching the appropriate node splitting evaluation criteria and the number of cross-validation folds is particularly important for the construction of random forest models. In this paper, we compare two evaluation criteria for node partitioning, gini and entropy, and cross-validate them in the range of 30% to 10%. In order to ensure the stability of the data, the model with the same parameters is tested several times to obtain the precision, recall and accuracy of the model.

The accuracy rate represents the proportion of the predicted samples in the total samples, the recall rate represents the proportion of the predicted positive samples in the results of the actual positive samples, and the accuracy rate represents the proportion of the predicted positive samples in the results of the actual positive samples. The larger the above three evaluation criteria, the better the accuracy of the representation model. However, in practice, recall rate and accuracy rate are negatively correlated. In order to better combine recall rate and accuracy rate, the harmonic average F1 of accuracy rate and recall rate is introduced, and Table 2 is obtained.

Table 2: Results of cross-checking and node splitting evaluation criteria

Cross Validation	gini				entrogy			
	Precision	Recall Rate	Accurate Rate	F1	Precision	Recall Rate	Accurate Rate	F1
Three	0.859	0.859	0.86	0.859	0.846	0.846	0.846	0.846
Four	0.857	0.857	0.858	0.857	0.864	0.864	0.866	0.864
Five	0.855	0.855	0.857	0.855	0.855	0.855	0.856	0.855
Six	0.857	0.857	0.857	0.857	0.838	0.838	0.839	0.838
Seven	0.862	0.862	0.862	0.862	0.865	0.865	0.866	0.865
Eight	0.864	0.864	0.865	0.864	0.869	0.869	0.871	0.868
Nine	0.866	0.866	0.868	0.866	0.853	0.853	0.854	0.853
Ten	0.874	0.874	0.876	0.874	0.858	0.858	0.858	0.858

In order to properly combine the data of the three and obtain the most ideal model, the average value of F1 and accuracy rate is taken as the evaluation value for comparison under the same circumstances, as shown in Figure 1.

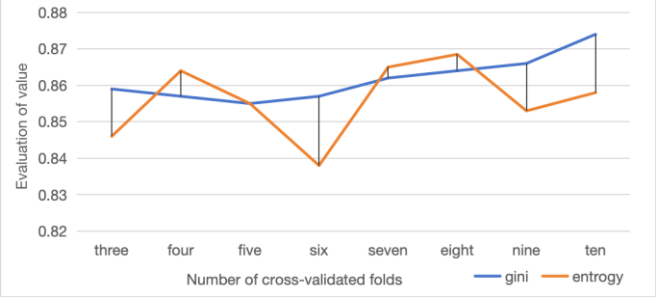

Figure 1: Comparison between gini and entrogy

ISSN 2616-5775 Vol. 6, Issue 13: 120-126, DOI: 10.25236/AJCIS.2023.061318

As can be seen from the evaluation value comparison graph of gini and entrogy, the classification prediction effect of random forest is the best when the cross check is 10 fold and the evaluation criterion of node splitting is gini.

3.3 The minimum number of samples of internal split points

In a random forest, the primary role of the minimum sample number parameter (hereinafter referred to as the "parameter") for internal node splitting is to control tree growth. This parameter defines the minimum number of samples an internal node needs before splitting. It controls the depth of the random forest model.

This data set is small, which is suitable for using smaller parameters. Therefore, this paper controls the parameters between 2 and 7. Combined with the above conclusions, the cross-check is set at 10 percent, and the node split evaluation criterion is gini as the prerequisite, and the experiment is conducted to obtain the relationship between parameters and evaluation value, as shown in Table 3.

Table 3: Parameter results of minimum sample number for internal node splitting

Parameter	Precision	Recall Rate	Accurate Rate	F1
2	0.86	0.86	0.862	0.86
3	0.868	0.868	0.869	0.868
4	0.865	0.865	0.867	0.865
5	0.873	0.873	0.874	0.872
6	0.853	0.853	0.853	0.853
7	0.857	0.857	0.86	0.857

And the relationship between parameters and evaluation value, as shown in Figure 2.

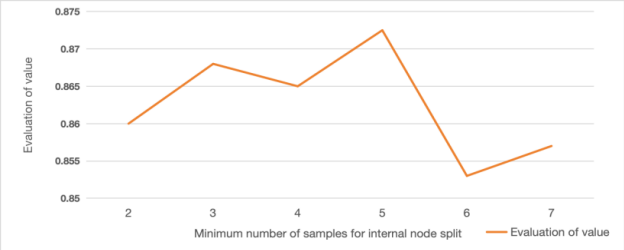

Figure 2: The relationship between the minimum sample number parameter of internal node splitting and the evaluation value

As can be seen from Figure 2, when the parameter is 4, the evaluation value is the highest, and the model has a good prediction function. With the increase of the parameter, the evaluation value can be inferred as a trend of fluctuation and decline.

3.4 The maximum proportion of features considered when dividing

In determining the parameters of the random forest method, we need to choose a feature to partition, which determines the proportion of features randomly selected to participate in the partition at each partition. A smaller proportion of features can increase randomness and make the decision tree more diverse. Thus, the influence of specific features on model prediction is reduced, and the model has more generalization ability. If the proportion of features considered in the division is large, each decision tree may choose similar features for division, resulting in high correlation between decision trees, and the effect of the model may not be as good as that in the case of high diversity.

In order to get the most suitable partition when considering the maximum feature ratio method. In this paper, based on the above conditions, the fold number of cross-validation is determined to be ten folds, and the best feature proportion is selected by comparing auto, log2 and sqrt strategies through cross-validation.

Table 4: The maximum proportion of features considered in the division

	Precision	Recall Rate	Accurate Rate	F1
Sqrt	0.85	0.85	0.85	0.85
log2	0.852	0.852	0.854	0.852
Auto	0.863	0.863	0.863	0.863

ISSN 2616-5775 Vol. 6, Issue 13: 120-126, DOI: 10.25236/AJCIS.2023.061318

It can be seen from Table 4 that the maximum feature ratio of auto is most suitable for this model.

3.5 Comparison with other machine classifications

The development of machine learning is accompanied by a number of classification models similar to the random forest model. They function similarly to random forests, but use a different approach to processing data to predict outcomes. In order to verify that random forest has a good ability to solve this problem in many mature models. By comparing the test results of other models, Table 5 is obtained.

Table 5: Comparison of different machine learning types

Types of machine learning	Precision	Recall Rate	Accurate Rate	F1	Evaluation Value
Random Forest	0.858	0.858	0.858	0.858	0.858
Decision-making Tree	0.839	0.839	0.844	0.839	0.839
ExtraTrees Classify	0.835	0.835	0.835	0.834	0.8345
Adaboost Classify	0.819	0.819	0.82	0.819	0.819
Naive Bayes Classify	0.806	0.806	0.806	0.806	0.806
BP Neural Network	0.789	0.789	0.789	0.789	0.789
Support Vector Machine (SVM)	0.584	0.584	0.62	0.559	0.5715

As can be seen from Table 5, random forest has the highest evaluation value and is most suitable for the prediction of stroke data to obtain more accurate conclusions.

4. Analysis of Experimental Conclusion

4.1 Analysis of results

Random Forest not only predicts the data, but also summarizes and generalizes the training set, and obtains the eigenimportance and confusion matrix heat map of the influencing factors in the training set.

Feature importance reflects the primary and secondary of each influencing factor. Among the influencing factors in this experiment, age is the most influential factor to the model, followed by blood sugar level. The confusion matrix heat map shows the relationship between the prediction results of the classification model and the actual labels. It contains four important indicators: true positive, true negative, false positive and false negative. In general, the higher the true positives and true negatives, and the lower the false positives and false negatives, the better the model's performance and the more accurate the prediction.

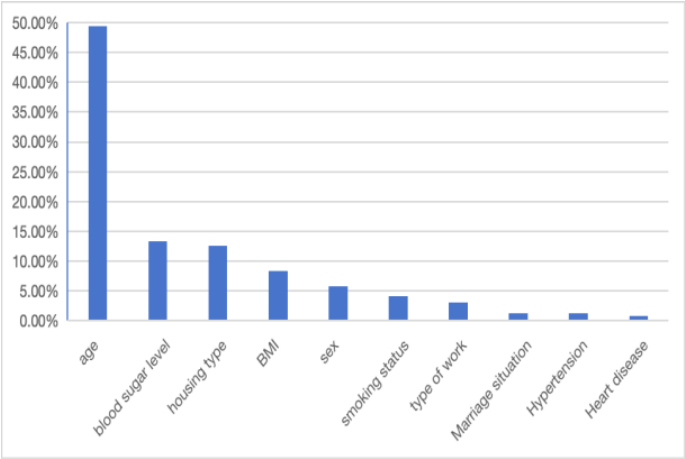

Figure 3: Importance of features

ISSN 2616-5775 Vol. 6, Issue 13: 120-126, DOI: 10.25236/AJCIS.2023.061318

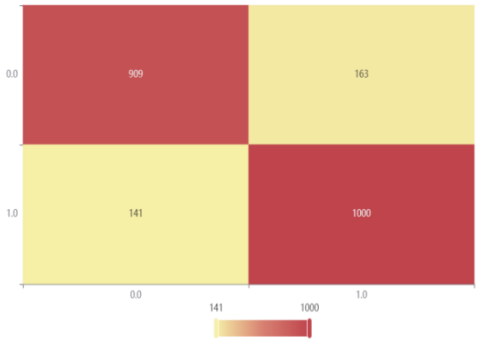

Figure 4: Thermal map of confusion matrix

As can be seen from Figure 3 and Figure 4, this model has a good predictive effect, and age is an important factor that causes stroke. From the controllable influencing factors, blood sugar regulation and the choice of housing type are the decisive factors to avoid stroke, keep blood sugar in the normal range, and choose a living environment with good natural environment and less pollution. It can better avoid the occurrence of stroke.

4.2 Experimental Conclusion

It can be seen from the final conclusion that random forest has a high accuracy in predicting stroke, and it can adapt to the addition of more influential factors to form a more perfect model with better scalability, and the model can be modified and adapted through relevant parameters. Table 6 shows the model parameters of this experiment.

Table 6: Relevant data of the final model of this adaptation

Parameter name	Parameter value	Parameter name	Parameter value
Data segmentation	0.7	The maximum proportion of features considered when dividing	auto
Data shuffling	TRUE	Minimum number of samples of internal node splits	5
Cross verification	10	Minimum sample number of leaf nodes	1
Evaluation criteria for node splitting	gini	The minimum weight of samples in leaf nodes	0
Decision tree quantity	100	The maximum depth of the tree	10
There are put back samples	TRUE	Maximum number of leaf nodes	50
Out-of-pocket data testing	FALSE	The threshold value of node partition impurity	0

However, as a prediction method of machine learning, random forest cannot obtain an accurate and concise formula. And a certain amount of data is required for training in order to get accurate predictions.

5. Conclusion

This model is helpful for hospitals to have a better judgment of stroke patients, improve people's living habits, and reduce the incidence of stroke. The idea of machine learning classification is also widely used to predict similar types of diseases, to determine the key factors causing such diseases and vulnerable populations.

However, in terms of data scale, the data of this experiment is small, and the influence factors collected are limited due to the influence of data sources. As far as the model is concerned, on the basis of this model, heuristic algorithms can be further considered for optimization to obtain a more suitable model, or the parameters of the random forest can be adjusted more carefully to seek the optimal solution.

ISSN 2616-5775 Vol. 6, Issue 13: 120-126, DOI: 10.25236/AJCIS.2023.061318

References

[1] Lin X , Zeng D , Cheng L ,et al.Study on the influence factors of music piracy in china based on SEM model[C]//International Conference on Service Systems & Service Management. IEEE, 2015. DOI:10.1109/ICSSSM.2015.7170311.

[2] Wang X. Correlation between cerebral apoplexy incidence and meteorological environment factors in Shenyang area [J]. Chinese Clinical Rehabilitation, 2006(36):12-13.

[3] Hou Yumei, Zhang Chenyang, Su Yanlin. Prediction of ischemic stroke risk based on support vector machine [J]. Modern Preventive Medicine, 2019, 46(15):2692-2695+2700. (in Chinese)

[4] PEI Zehua, Ge Miao, Li Hao et al. Study on environmental factors affecting HDL-C in middle-aged and elderly people in China based on random forest model [J]. Journal of Geoinformation Science, 2002, 24(07):1286-1300.

[5] Yan Guanghua, Chen Xi, Zhang Yun. Study on the Distribution Pattern and Influencing Factors of Shrinking Cities in Northeast China Based on Random Forest Model. Geographical Science, 2021, 41(05):880-889. DOI:10.13249/j.cnki.sgs.

[6] Guan Jun, Zhang Shaopeng, Ren Yue et al. Spatial and temporal differentiation and influencing factors evolution of agricultural net carbon sinks in China based on random forest model [J/OL]. China environmental science: 1-13 [2023-10-19]. https://doi.org/10.19674/j. cnki.issn1000-6923. 20230928.002.

[7] Ling Xiaodan, Wang Luoqi, Zhao Keli et al. Study on spatial distribution characteristics of soil available nutrients in Pecan forest based on random forest method [J/OL]. Acta ecologica sinica, 2024 (02): 1-14 [2023-10-19]. HTTP: / / https://doi.org/10.20103/j.stxb.202301130090.