

# Feature Impact Analysis and Machine Learning-based Stroke Probability Prediction

Zhengshu Zhang<sup>1</sup>

<sup>1</sup>College of Food Science, Sichuan Agricultural University, Yaan, Sichuan, 625014, China  
Email: Forestzhang@seu.edu.mk

**Abstract.** Stroke is an acute cerebrovascular disease with a high mortality rate after onset. More than ten millions of people worldwide are suffering from stroke and with millions of them are losing their lives each year. Therefore, studying the factors and probability of stroke is of great value for people's health and life today. With the development of artificial intelligence, the machine learning has become one of the best tools to help people predict the more accurate future values in many areas. As the result, this experiment aims to use different machine learning models to predict stroke probability and do the model comparison to provide model reference for future researchers. Due to the different optimal machine learning models corresponding to different types of disease factors, this experiment first selected quantifiable continuous variables as a combination of features for research. This experiment found that when the features are a combination of age, glucose, and BMI, the logistic regression model is an ideal choice because its accuracy and recall rates are 95% and 100% respectively. According to this study about the feature impact analysis, age has the greatest impact on the predicted stroke probability, followed by glucose, and BMI has the smallest impact.

**Keywords:** Probability of Stroke, Machine Learning, Model Comparison, Feature Impact Analysis.

## 1 Introduction

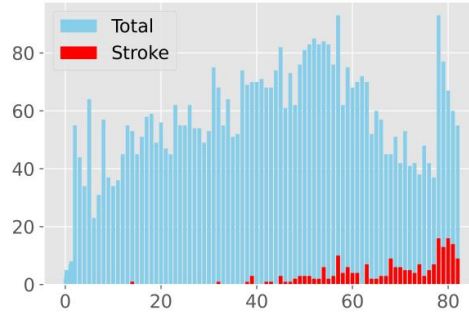
Stroke is the second leading cause of death worldwide and stroke patients also have a high risk of disability [1]. Among them, age; hypertension; diabetes; smoking and others are all factors that will affect stroke [2]. Peter Appelros and Ingegerd Nydevik conducted a risk assessment of stroke patients using the Stroke Scale [3]. Matthew Chun and Robert Clarke used machine learning models such as random forest and support vector machine to identify individuals at high risk of stroke [4]. T Kansadub and S Thammaboosadee use decision trees, naive bayes and neural network algorithms for predicting stroke [5]. It can be seen that the research on stroke probability has great scientific value, and machine learning models are an important research method among them. Machine learning can use varies algorithms to train past data to predict values, it is an important branch of artificial intelligence [6]. Common machine learning models are mainly divided into regression models and

classification models. Regression models are often used to study continuous variables, including linear regression, polynomial regression, etc. Their main task is to fit the input data in the form of linear or polynomial functions and then use functions to predict future values. [7]. Polynomial regression is supposed to perform better when relationship between two variables is not linear [8]. Classification models are commonly used to study discrete variables or texts, including logistic regression and decision trees. Logistic regression often uses medical data as input variables, and in addition to predicted values, it can also predict probabilities [9]. Decision trees can establish a classification system for multiple covariates, categorizing populations into dendritic fragments for prediction [10]. The aim of this experiment is to investigate the suitable model when features are continuous variables. In this way, the study decided to use four machine learning models - the linear regression, polynomial regression, logical regression, and decision tree to predict the stroke probability of three continuous variables - age, average glucose level and BMI. Then compare their performance and concentrate on the impact of these features on the prediction results.

## 2 Method

### 2.1 Dataset

The dataset is from a file called healthcare dataset from Fedsoriano [11].



**Fig. 1.** Stroke and total number in different age.

Figure 1 shows that almost no patients getting stroke in the sample between the ages from 0 to 40 years old. In the population aged from 40 to 80 years old, especially over 60 years old, there is a significant soar in the number of stroke patients.

Figure 2 shows that when the value of BMI is below 40, the proportion of stroke patients stabilizes among 0 to 0.2. When the value is above 40, the proportion of stroke patients soars to 0.5. It may be due to the lack of data to cause there are few bars in this area.

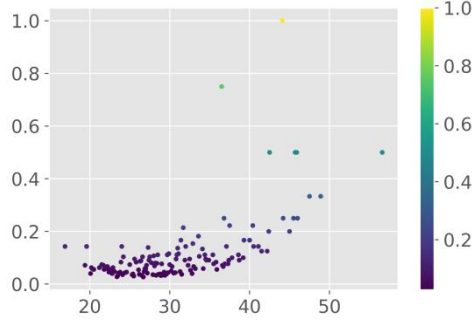


Fig. 2. Stroke proportion in different BMI.

Figure 3 illustrates that as the average glucose level increases, the proportion of stroke patients shows a phased increase, with a significant jump in the proportion at 150mg/dl and 250mg/dl, from 6% to 20% and 24% to 43%.

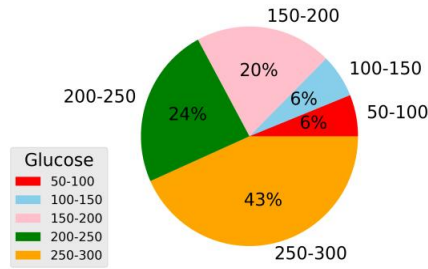


Fig. 3. Stroke proportion in different glucose.

## 2.2 Prediction Model

**Linear Regression Model.** Using a linear regression model to fit the data and calculate the probability. The model will get the relationship function between label and feature, and the relationship function is a simple linear function.

**Polynomial Regression Model.** The polynomial regression model reflects the nonlinear relationship between feature and label. In this experiment, only the quadratic polynomial model and the linear regression model were used for comparison.

**Logistic Regression Model.** Liu Lu and Gao Ying conducted research on stroke related issues using logistic regression models [12]. So, this study decides to add the logistic regression model for predicting stroke probability. The logistic regression model uses the sigmoid function to obtain output results between 0 and 1, which meets the probability prediction range required by this experiment and is one of the most common binary classification models.

**Decision Tree.** Using the decision tree to train the data and predicting the stroke probability. It focuses on nonlinear classification problems, dividing the training set according to the hierarchical structure of the tree. In this study, the maximum height of the tree was set to 3. The use of this model aims to achieve better performance while above three models perform unexpected.

### 2.3 Evaluation Matrices

MAE indicates the mean absolute error. It shows the difference between the prediction and the test. MSE presents the mean squared error. It takes more attention to the outlier and like the MAE, the smaller its value demonstrates the model is better. Accuracy calculates the proportion of correct prediction in all of the test data and recall calculates the proportion of positive predictions which are correctly identified among all of the data. The higher the accuracy and recall values, the better the performance of the model.

## 3 Result

The study uses the above four machine learning models to get the value of prediction in the certain conditions. By the method of control variates, this study observes the result through controlling one of the features to change while others are stable. Comparing the result in certain conditions, it will be directly to find the degree of impact of features on the value of prediction. According to the result of 2.2.3, the best performance models will be paid more attention to and take on greater weight in this part to find the connection between features and possibility.

### 3.1 Model Comparison

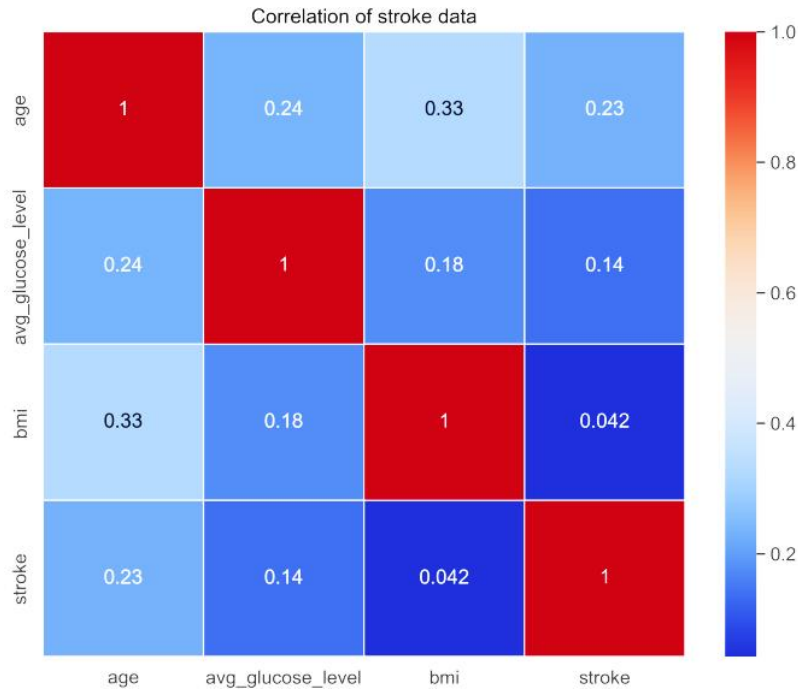
Table 1 shows the value of evaluation metrics about four machine learning models in the possibility of stroke. According to the score of MAE, MSE, Accuracy and Recall of logistic model are 4%, 4%, 95% and 100% which directly means that this machine learning model has the best performance among four models. What is more, the linear has the worst performance for its Mae and Accuracy. In this study, the performance of polynomial regression model and the decision tree is close to each other and better than the linear regression model but worse than the logistic regression model because of their lower Accuracy and Recall score.

**Table 1.** Evaluation metrics of models

Models	MAE (%)	MSE (%)	Accuracy (%)	Recall (%)
Linear regression	9	4	61	93
Polynomial regression	8	4	73	89
Logistic regression	4	4	95	100
Decision tree	5	4	81	73

### 3.2 Feature Impact Analysis

**Feature Correlation.** According to Figure 4, the heat map illustrates the correlations between three features and stroke in the dataset leveraged in this study. The value of the heat map is supposed to be between -1 and 1, and the larger the absolute value, the stronger the correlation. Among them, -1 represents complete negative correlation, +1 represents complete positive correlation and 0 represents no relationship. From the study, the correlation between age and stroke is strongest, followed by glucose level, while the correlation between BMI and stroke is very weak.



**Fig. 4.** Feature correlation of dataset

**Feature Correlation in the Prediction.** According to Table 2, when the values of glucose and BMI are set to 120mg/dl and 20 control constants, the prediction results obtained by the four machine learning models significantly increase with age. In other three models, this upward trend is more obvious. When the sample age is 80 years old, the probability of stroke is 16.4%, 16.8%, and 17.2%, which are higher than the predicted values of the linear regression model.

It can also be observed that when the values of age and BMI are set to 60 and 20, the prediction results of the four machine learning models show an upward trend with the increase of glucose, but this upward trend is not significant compared to the upward trend caused by age. Among them, the results of the decision tree are quite unique. During the process of increasing glucose from 80mg/dl to 140mg/dl, the

probability of stroke did not increase and remained stable at 4.4%. When glucose reached 160mg/dl, the predicted result suddenly increased to 13.1%.

**Table 2.** Possibility of stroke in certain conditions prediction

(age, glucose, BMI)	Linear (%)	Polynomial (%)	Logistic (%)	Decision tree (%)
(40, 120, 20)	5.2	0.3	1.3	2.3
(50, 120, 20)	7.1	2.9	2.6	2.3
(60, 120, 20)	9.0	6.4	5.0	4.4
(70, 120, 20)	10.9	10.9	9.3	10.2
(80, 120, 20)	12.8	16.4	16.8	17.2
(60, 80, 20)	7.3	6.1	4.0	4.4
(60, 100, 20)	8.2	6.2	4.5	4.4
(60, 140, 20)	9.9	6.8	5.6	4.4
(60, 160, 20)	10.8	7.3	6.2	13.1
(60, 120, 15)	9.6	6.4	4.8	4.4
(60, 120, 25)	8.5	6.4	5.2	4.4
(60, 120, 30)	8.0	6.2	5.3	4.4
(60, 120, 35)	7.4	6.0	5.5	4.4

When the values of age and glucose are set to 60 and 120mg/dl, there is a small fluctuation in the stroke probability values as BMI increases. In a linear regression model, when BMI increases from 15 to 35, the probability of stroke slightly decreases from 9.6% to 7.4%. In the logistic regression model, the opposite situation occurred, with the probability of stroke slightly increasing from 4.8% to 5.5%.

## 4 Discussion

### 4.1 Performance Differences Between Models

It can be seen that different machine learning models exhibit different performance in predicting stroke probability. Due to the same dataset and the fixed division, the preprocessing stage of the data will not affect the results. Therefore, the performance differences are considered in the following aspects: Firstly, the impact of algorithms: differences caused by different algorithm logics used by different models. For example, there are significant differences in algorithmic logic between linear regression models and decision tree models. Linear regression involves linear fitting of data while decision tree models construct trees to classify data. Secondly, choose of features: because three features in the form of continuous numbers are selected in this experiment, linear regression model and polynomial regression model can show good performance in dealing with problems. Theoretically, while other discrete or descriptive features that may affect stroke probability are incited, the performance of linear regression model and polynomial regression model will decrease while the performance of logical regression model and decision tree model will increase. Thirdly, choose of evaluation indicators: Although Mae and Mse are more effective

for the evaluation of linear regression and polynomial regression models. But for logistic regression and decision tree, in addition to the four evaluation indicators selected in this experiment, Precision, F1-score, etc. are also reference evaluation indicators. Therefore, different choices of evaluation indicators can also lead to differences between models.

## 4.2 Differences in Prediction Results Among Models

When predicting under specific conditions, there are some abnormal probability values that should be discarded. It can be seen that when age, glucose, and BMI are set to 40, 120 and 20, the probability values of linear regression model are significantly higher than others. The reason for this difference is that the algorithm logic of the linear regression model is to fit the data as linearly as possible. Therefore, in some intervals with significant differences from linear fitting, the obtained values will have significant differences. Another obvious anomaly area is that when the age in the decision tree model is set to 60, regardless of how the glucose and BMI change, the prediction probability is mostly 4.4%. This indicates that the decision tree model is not sensitive to changes in glucose and BMI, and only affects the results when there are significant changes in glucose and BMI. When the glucose increases to 160, the probability will increase to 13.1%.

## 5 Conclusion

This study finds that when features are set to continuous variables such as age, average glucose level and BMI, the logistic regression model performs best with 95% Accuracy and 100% Recall. Therefore, it is recommended to use the logistic regression model to predict the probability of stroke in the future to obtain the most accurate results. The features selected in this experiment have the characteristics of continuous variables. Therefore, when studying the impact of other continuous variables such as blood pressure, blood lipids, etc. on the probability of stroke, the results of this experiment can be referred to and the logistic regression model can be selected to compare its performance with more models that don't mentioned in this study. In addition, this experiment also found that three features have varying degrees of impact on the probability of stroke. Age has the greatest impact, followed by glucose and BMI has the smallest impact. Therefore, for middle-aged and elderly people with a high probability of stroke, controlling glucose and BMI to prevent further increase in stroke risk is very important. In the future, experiments will also further discuss the impact of other discontinuous variables such as gender, smoking status, and location on stroke probability, as well as appropriate machine learning models to study their relationship with stroke, in order to help people choose accurate machine learning models for stroke probability prediction based on different features and understand the severity of their impact on stroke probability. With the result of severity of impact, people are supposed to pay more attention to some physical and chemical indicators in daily life to keep health.

## References

1. Li, Q., Wu, H., Yue, W., Dai, Q., Liang, H., Bian, H., et al.: Prevalence of stroke and vascular risk factors in China: a nationwide community-based study. *Scientific reports*, 7(1), 6402 (2017).
2. Cubrilo-Turek, M.: Stroke risk factors: recent evidence and new aspects. In *International Congress Series 1262*, 466-469 (2004).
3. Appelros, P., Nydevik, I., Seiger, A., & Terént, A.: Predictors of severe stroke: influence of preexisting dementia and cardiac disorders. *Stroke*, 33(10), 2357-2362 (2002).
4. Chun, M., Clarke, R., Cairns, B. J., Clifton, D., Bennett, D., et al.: Stroke risk prediction using machine learning: a prospective cohort study of 0.5 million Chinese adults. *Journal of the American Medical Informatics Association*, 28(8), 1719-1727 (2021).
5. Kansadub, T., Thammaboosadee, S., Kiattisin, S., & Jalayondeja, C.: Stroke risk prediction model based on demographic data. In *2015 8th Biomedical Engineering International Conference*. 1-3 (2015).
6. Baştanlar, Y., & Özuysal, M.: Introduction to machine learning. *miRNomics: MicroRNA biology and computational analysis*, 105-128 (2014).
7. Maulud, D., & Abdulazeez, A. M.: A review on linear regression comprehensive in machine learning. *Journal of Applied Science and Technology Trends*, 1(4), 140-147 (2020).
8. Ostertagová, E.: Modelling using polynomial regression. *Procedia Engineering*, 48, 500-506 (2012).
9. Lever, J., Krzywinski, M., & Altman, N.: Logistic regression: Regression can be used on categorical responses to estimate probabilities and to classify. *Nature Methods*, 13(7), 541-543 (2016).
10. Song, Y. Y., & Ying, L. U.: Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, 27(2), 130 (2015).
11. Stroke Prediction Dataset, URL: <https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset?datasetId=1120859&searchQuery=logistic+regression> 2023/5/5
12. Liu, L., & Gao, Y.: Study on the correlation between traditional Chinese medicine syndrome and short-term prognosis of ischemic stroke using logistic regression model and repeated-measures analysis of variance. *Journal of Chinese Integrative Medicine*, 10(9), 983-990 (2012).