Optimal Modelling of Stroke Probability Prediction Through Machine Learning

Yuheng Zhu
Hangzhou No.4 High School, Hangzhou, China

Keywords: Stroke Forecasting, Machine Learning, Random Forest, Multilayer Perceptron, Imbalanced Data.

Abstract: Stroke has always been a significant threat to human health. Predicting the occurrence of stroke plays a very important role in reducing the prevalence and lethality of stroke. With the development of machine learning techniques, using machine learning techniques to assist in medical decision-making has become a new area of research. This paper used the random forest algorithm and the multilayer perceptron algorithm to predict the probability of a patient suffering from stroke based on his physiological indicators and compared the performance of the two models. By analyzing the importance of features in the Random Forest model, the health factors more correlated with the probability of stroke were obtained. The random forest algorithm was found to be a more suitable optimization model for predicting the probability of stroke. Among the common health factors, age, average blood glucose level, and body mass index had a greater effect on stroke probability.

1 INTRODUCTION

Stroke has long been a major threat to human health. According to disability-adjusted life years lost (DALYs), stroke is still the second-leading cause of mortality worldwide and the third-leading cause of death and disability combined (Feigin et al 2022). In terms of the total number of cases, the burden increased significantly between 1990 and 2019 (70.0% more incident strokes, 43.0% more stroke deaths, 102.0% more prevalent strokes, and 143.0% more disability-adjusted life years), with lower- and lower-middle-income countries (LMIC) bearing the majority of the burden (86.0% of deaths and 89.0% of DALYs) (Feigin et al 2022).

Machine learning is commonly used to find potential associations from large data sets and make accurate predictions. It has been used in various scenarios, including the medical field. Recent advances in machine learning (ML) have made it possible to classify skin cancer using photos accurately on par with that of a qualified dermatologist and to predict the development of type 2 diabetes from pre-diabetes using data commonly acquired from electronic health records (Gibbons and Gibbons 2019). In conclusion, machine learning and deep learning play an important role in the medical field to assist in medical decision-making with their accurate classification and prediction capabilities (Gibbons and Gibbons 2019).

Accurate and adaptable assessment techniques are crucial for preventing stroke (Wu and Fang 2020). Traditional approaches, like Cox proportional risk modeling, are ineffective for examining intricate non-linear relationships in the data, though (Wu and Fang 2020).

Therefore, the significance of this study is to discover a model that is accurate enough to predict the probability of stroke through machine learning for accurate stroke prediction.

Stroke prediction has profound implications in the field of medicine, such as identifying people at high risk of developing stroke providing preventive treatments to reduce the loss of life and property due to stroke, and revealing hidden factors that lead to stroke to guide people to make healthier lifestyle choices and to reduce the probability of the general public suffering from stroke. This research aims to explore the optimal model for predicting the probability of stroke using machine learning. This study will use the Synthetic Minority Oversampling (SMO) technique for data balancing of the original unbalanced dataset to optimize the model's performance. The Synthetic Minority Oversampling Technique (SMOTE) uses an oversampling method to rebalance the original training set. Data processed by the SMOTE method has the greatest accuracy, precision, recall, and ROC values among the classification prediction models, and the data balanced

Zhu, Y.
Optimal Modelling of Stroke Probability Prediction Through Machine Learning.
DOI: 10.5220/0012810000003885

In *Proceedings of the 1st International Conference on Data Analysis and Machine Learning (DAML 2023)*, pages 159-163
ISBN: 978-989-758-705-4

by this algorithm greatly improves the classification performance of the prediction model (Wang 2023). This study will use two algorithms, the Random Forest Algorithm (RF) and the Multilayer Perceptron Algorithm (MLP), to predict whether a subject is likely to suffer from a stroke based on the subject's eigenvalues such as gender, age, blood pressure, BMI, height, weight, occupation, etc. in the dataset.

The aim of this study is to obtain the optimal model for predicting the probability of stroke by machine learning through a controlled test of the performance of the two algorithms, RF and MLP, as well as to reveal the key factors in common health factors that are related to the probability of developing stroke.

2 METHODS

2.1 Data Source

The data in this study was an open-source dataset obtained from Kaggle.

The dataset is appropriate for training models that forecast the likelihood of a patient having a stroke based on variables such as gender, age, different illnesses, and smoking status. The dataset comprises 5110 rows of patient information. The dataset includes 12 indicators: identification number, gender, age, high blood pressure status, heart disease status, marital status, job type, residence type, blood glucose level, BMI, smoking frequency, and stroke status.

2.2 Data Prepossessing

This study's data preprocessing primarily encompasses the following aspects: 1. Completing missing values, 2. Encoding non-numerical data, such as text and categories, into numerical labels to help the model use the data correctly, 3. Due to only 250 data samples of stroke patients in the dataset, which represent approximately 4.89% of the total, the data is unbalanced. An imbalanced sample distribution causes small sample sizes in classification with too few features, which makes it difficult to extract patterns, and even if the classification model is obtained, it is prone to overfitting. The limited data samples are the primary culprit (Wang 2023). To address this issue, the SMOTE method is used to oversample in this study to achieve data balance.

2.3 SMOTE

The SMOTE algorithm shows its importance when faced with unbalanced datasets. SMOTE balances the dataset by synthetically increasing the number of new minority class samples by synthesizing new minority class samples so that the model learns the features of the minority class better. The algorithm generates new synthetic samples by selecting minority samples and their nearest neighbors based on feature space interpolation. By employing this technique, classifier performance (in ROC space) can be improved compared to undersampling merely the majority class (Chawla et al 2002).

2.4 Random Forest

Logistic regression models are routinely used for statistical analysis of risk factors. However, in this study, due to the large number of factors affecting the medical process and the intricate relationship, the Random Forest algorithm, as an emerging machine learning algorithm, has a greater advantage in processing such problems (Wen et al 2019). Therefore, Random Forest is chosen as one of the algorithms in this study.

RF is a powerful machine learning algorithm that achieves higher predictive accuracy and generalization by combining multiple decision tree models. At its core is the introduction of randomness, which reduces the risk of overfitting while maintaining model stability through random feature selection and data sampling. RF is suitable for classification and regression tasks and performs well on various data types. It can handle high-dimensional data and complex relationships between features effectively and is well-suited for stroke prediction.

2.5 Multilayer Perceptron

The MLP model is also a powerful machine-learning algorithm. Previous stroke prediction studies demonstrated a superior ability to predict stroke mimics compared to the FABS and TM-Score models, with a higher level of accuracy (Zhang et al 2021).

The MLP model can learn complex patterns and feature representations in data by connecting multiple neuronal layers and nonlinear activation functions (Torres et al 2021). Its forward and backpropagation mechanisms make it adaptable to classification, regression, and feature learning tasks. The flexibility of the MLP allows it to handle a variety of data types and problem domains, including the prediction of stroke in this study. However, as the network structure becomes complex, the training process tends to fall into local optima. Therefore, strategies such as the appropriate number of layers, neurons, and regularisation are crucial for successfully applying MLPs.

2.6 Testing the Performance of the Model

This study used parameters such as accuracy, precision, recall, F1 score, and AUC-ROC score to analyze and compare the performance of the two models (Belete and Huchaiah 2022). These performance metrics are widely used to assess the effectiveness of classification models in different domains.

Accuracy measures the overall correctness of the model's classification, while precision concerns the model's accuracy in positive category prediction. Recall, on the other hand, focuses on the model's ability to recognize positive category samples. The F1 score combines precision and recall and is suitable for scenarios that seek a balance between precision and comprehensiveness. In addition, the AUC-ROC score evaluates the model's ability to classify positive and negative samples and the accuracy of the probability ranking, which is particularly suitable for dealing with unbalanced datasets.

By considering these performance metrics together, this research can gain a deeper insight into the strengths and weaknesses of the two models for the problem of stroke probability prediction, providing strong support for further model selection and optimization.

3 RESULTS

To compare the performance of the two models in stroke prediction, the accuracy, precision, recall, and F1 Score of RF and MLP were calculated separately in this study and plotted in Figure 1.

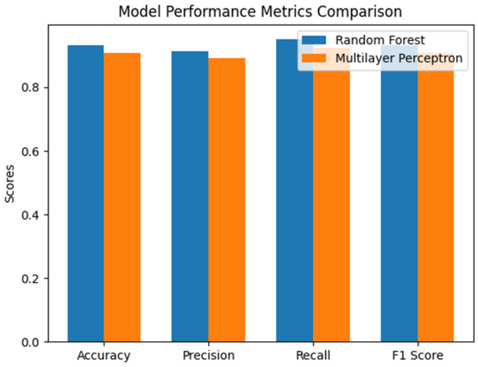

Figure 1: Model performance metrics comparison (Picture credit: Original).

According to our experimental results, the accuracy of the Random Forest model is 0.93, precision is 0.91, recall is 0.95, and F1 Score is 0.93. From these metrics, this paper can see that the Random Forest model performs well in predicting the probability of stroke.

According to the experimental results, the multilayer perceptron model has an accuracy of 0.91, a precision of 0.89, a recall of 0.92, and an F1 score of 0.91.

Overall, the findings obtained from this experiment show that RF slightly outperforms MLP in the four aspects of Accuracy, Precision, Recollection, and F1 Score.

This indicates that RF has a more comprehensive, stable, and superior performance in predicting the probability of stroke as compared to MLP.

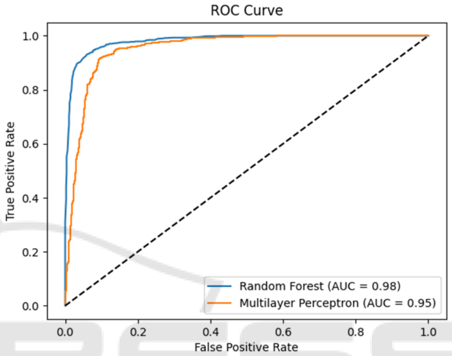

Figure 2: ROC curve (Picture credit: Original).

The ROC curves of the two models were then plotted in Figure 2 for comparison in this study.

The ROC curves were created by plotting the relationship between True Positive Rate and False Positive Rate at different thresholds. The ROC-AUC is the area under the ROC curve, and its value ranges from 0 to 1. The closer the ROC-AUC is to 1, it indicates that the model has a better ability to discriminate between positive and negative samples, and the closer it is to 0.5 it indicates that the model performance is poor.

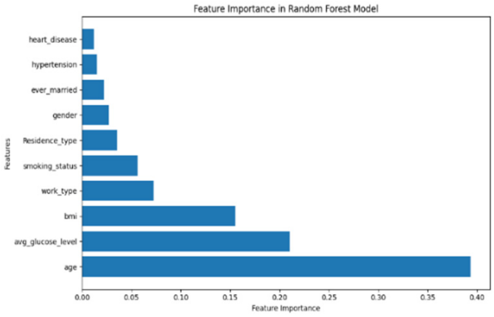

Figure 3: Feature importance in RF model (Original).

Figure 3 shows that the ROC-AUC value of RF is about 0.98, while the ROC-AUC value of MLP is about 0.95.

These findings underscore the superiority of the RM model in this particular context, showcasing its potential for stroke forecasting. Nonetheless, the MLP model has a slightly lower but commendable, discriminative capability.

Figure 3 shows the relative importance of each feature for predicting the probability of stroke in the random forest model. The importance score for each feature represents the degree to which the feature contributes to the model decision. Based on the chart, the following conclusions could be draw:

3.1 Age

Age has the highest importance in the model with an importance score of 0.3936. this suggests that age is a key factor in predicting the probability of stroke, probably because of the significant correlation between stroke risk and age.

3.2 Avg_glucose_level

Average blood glucose level is high in the model with an importance score of 0.2103. high blood glucose may be associated with an increased risk of stroke.

3.3 BMI (Body Mass Index)

Body mass index has a relatively high significance in the model with an importance score of 0.1552, suggesting that the relationship between body weight and stroke is also an important predictor.

3.4 Work_type

Work type has a relatively small but still significant contribution to predicting the probability of stroke, with an importance score of 0.0721, which may indicate that there is an association between different types of work and the probability of having a stroke.

3.5 Smoking_status

Smoking_status has a relatively small effect but still contributes with an importance score of 0.0564. smoking may increase the risk of stroke.

3.6 Residence_type

Gende, ever_married, hypertension and heart_disease have a smaller effect on the probability of stroke, with a Feature Importance of 0.0353, 0.0275, 0.0153, and 0.0123, respectively.

Overall, age, mean blood glucose level and body mass index were the most important predictors of stroke probability, with other characteristics having lesser but still contributing effects.

4 DISCUSSION

This study compares the performance of two machine learning algorithms, RF and MLP, in predicting stroke probability. By comprehensively comparing the accuracy, precision, recall, and F1 score metrics in Fig.1 and the ROC-AUC values in Fig.2, it is concluded that Random Forest has a more comprehensive, stable, and superior performance in predicting stroke probability.

Based on the findings of this study, the following points can be suggested to optimize further and improve the performance of the stroke probability prediction model (Assmann 2002). Due to the excellent performance of the Random Forest model, future research can focus on how to perform parameter tuning of the RF model to improve the prediction accuracy. However, considering that the multilayer perceptron may still have advantages in some aspects, future research can also consider combining the models of random forest and multilayer perceptron for joint decision-making to take advantage of the strengths of each model.

Meanwhile, based on the analysis of the Feature Importance of each parameter in the RF model in the study, this study obtained that age, average blood glucose level, and body mass index have a greater impact on the probability of stroke. Therefore, healthcare professionals can pay special attention to these factors to screen people with high risk of stroke and reduce the risk of death from stroke. At the same time, healthcare professionals can monitor and adjust patients' average blood glucose levels and body mass index to reduce the probability of stroke. Future studies could further delve into the specific associations between these factors and stroke onset to provide more targeted guidance for prevention and treatment.

Despite this study's series of meaningful results, some limitations exist. First, the research data in this study only covered specific regions and populations, which may have geographical bias. Future studies may consider introducing more regions and diverse population samples to improve the model's generalization ability.

Second, although several common health factors were considered in this study, other factors, such as genetic factors, may still affect stroke. Further studies could try to incorporate more potential influencing factors into the model to improve the comprehensiveness of the prediction.

In summary, our study provides some insights into optimizing stroke probability prediction models. Still, further in-depth studies are needed to overcome the limitations to achieve more accurate and reliable predictions.

5 CONCLUSION

This study delves into the two main algorithms, RF and MLP, in machine learning for predicting stroke probability. The main finding of our study is that RF slightly outperforms MLP in several aspects, such as accuracy, precision, recall, and F1 score, showing a more comprehensive, stable, and superior performance. By comparing the ROC-AUC scores, it was demonstrated that RF has a higher ability than MLP in recognizing stroke, which further solidifies its superiority in stroke probability prediction.

Therefore, the random forest algorithm is a more suitable optimization model for predicting stroke probability. When constructing a stroke prediction model, choosing an appropriate algorithm is crucial to obtaining accurate and reliable prediction results.

This study bridges the gap of machine learning in predicting stroke probability and provides a basis for constructing more accurate models for assisted medical decision-making. Second, this study provides new ideas for actively preventing stroke by explaining the key factors more strongly associated with the probability of stroke. This study helps medical professionals and data analysts study stroke through machine learning algorithms, enabling them to make accurate stroke probability predictions and explore stroke risk factors.

This study still has shortcomings, such as the limitation of the dataset. In the future, researchers can explore deeply in the following aspects. First, researchers can consider introducing more advanced machine learning algorithms into the study to improve the prediction performance further. Meanwhile, researchers can conduct interdisciplinary collaborative research, which can better understand the mechanism and prediction methods of stroke by combining the knowledge of clinical medicine and data science.

In conclusion, despite the limitations of this study, the stroke probability prediction model can be improved through continuous efforts and in-depth research to reduce the threat of stroke to human health.

REFERENCES

VL. Feigin, M. Brainin, B. Norrving, et al, World Stroke Organization (WSO): Global Stroke Fact Sheet 2022. International Journal of Stroke. 2022, vol.17, no.1, pp.18-29.

J. Sidey-Gibbons, C. Sidey-Gibbons, Machine learning in medicine: a practical introduction. BMC Med Res Methodol, 2019, vol.19, pp.64.

Y. Wu, Y. Fang, Stroke Prediction with Machine Learning Methods among Older Chinese. International Journal of Environmental Research and Public Health, 2020, vol.17, no.6, pp.1828.

Z. R. Wang, Application of Machine Learning Methods in Stroke Risk Prediction (Master's Thesis, Guangzhou University), 2023.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 2002, vol.16, pp.321-357.

T. C. Wen, B. Y. Liu, Y. N. Zhang, A study of risk factors for unplanned readmission within 31 days in ischaemic stroke patients: a random forest model. Chinese Journal of Evidence-Based Medicine, 2019, vol.19, no.5.

Z. Zhang, D. Zhou, J. Zhang, et al, Multilayer perceptron-based prediction of stroke mimics in prehospital triage. Scientific Reports, 2022, vol.12, no.1, pp.17994.

J. F. Torres, D. Hadjout, A. Sebaa, et al, Deep learning for time series forecasting: a survey. Big Data, 2021, vol.9, no.1, pp.3-21.

D. M. Belete, M. D. Huchaiah, Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results. International Journal of Computers and Applications, 2022, vol.44, no.9, pp.875-886.

G. Assmann, P. Cullen, H. Schulte, Simple scoring scheme for calculating the risk of acute coronary events based on the 10-year follow-up of the prospective cardiovascular Munster (PROCAM) study. Circulation, 2002, vol.105, no.3, pp.310-315.