DOI: 10.15849/IJASCA.240330.11

Int. J. Advance Soft Compu. Appl, Vol. 16, No. 1, March 2024
Print ISSN: 2710-1274, Online ISSN: 2074-8523

LIGHTGBM-Based Machine Learning Model for Stroke Risk Prediction

Van Lam Ho[1], Thi Phuong Thao Le[2], Duy Nguyen V. M[3]

[1] Faculty of Information Technology, Quy Nhon University, Vietnam
e-mail: hovanlam@qnu.edu.vn
[2] TMA Solution QuyNhon, Binh Dinh, Vietnam
e-mail: lephuongthao2597@gmail.com
[3]CTO, Information Technology Unit, Hospital 175, Ho Chi Minh, Vietnam

Abstract

The paper aims to use appropriate machine learning models on real datasets with signs of stroke to make predictions about whether a person is likely to have a stroke and make some necessary recommendations in the patient screening. This study also helps reduce treatment costs and supports remote patient treatment. In this study, we built a machine learning model to assist doctor in predicting a person's risk of Stroke after entering his/her medical examination data at the hospital are factors that cause stroke. People can use this model through an application to track their risk of Stroke. Combining input community data with the expertise of Stroke specialists, we built a dataset with relevant information to predict Stroke. Based on this dataset, we have statistically described the data characteristics as well as the correlated data parameters that may cause Stroke, then we use LightGBM the algorithm to build a machine learning model to predict whether a person to be able to prevent stroke from occurring. The obtained results are going to be applied to assist in predicting whether a person may have Stroke with the input of information about the signs of stroke as previously examined combined with medical practice knowledge about the disease. From this result, it is possible to continue researching and applying artificial intelligence to support diagnosis and treatment of Stroke.

Keywords: *Lightgbm algorithm, Machine Learning, Stroke Prediction, Boosting Algorithm, Model Prediction.*

1 Introduction

Machine learning is the engine which is helping to drive advances in the development of artificial intelligence. It is impressively employed in both academia and industry to drive the development of 'intelligent products' with the ability to make accurate predictions using diverse sources of data [1]. Today, the key beneficiaries of the 21st century explosion in the availability of big data, machine learning and data science have been industries which were able to collect these data and hire the necessary staff to transform their products. The learning methods developed in and for these industries offer tremendous potential to enhance medical research and clinical care, especially as providers increasingly employ electronic health records [2]. Two areas which may benefit from the application of machine learning techniques in the medical field are diagnosis and outcome prediction [3],

Received 30 December 2023; Accepted 28 February 2024

[4]. This includes a possibility for the identification of high risk for medical emergencies such as relapse or transition into another disease state [5], [6].

Stroke is the second leading cause of death, after ischemic heart disease, and accounts for 9% of deaths worldwide. According to the World Health Organization (WHO), fifteen million people suffer stroke worldwide each year. Of these, more than 6 million die and another 5 million are permanently disabled [7].

In the past, strokes often occurred in older people. However, in recent years, there has been a significant growth in the number of stroke cases among younger age groups, including young people and children. This may be related to risk factors such as an unhealthy lifestyle, stress, and underlying medical conditions. Stroke is one of the leading causes of disability and death worldwide. It can cause serious consequences such as loss of mobility, loss of speech, loss of sensation and difficulty recovering [8], [9].

Therefore, the application of machine learning in clinical diagnosis for stroke prediction holds great potential. In this study, we utilized the LightGBM (Light Gradient Boosting Machine) machine learning model to predict the likelihood of an individual experiencing a stroke. Additionally, we employed various evaluation methods to assess the results obtained from the model, determine if it meets the set objectives, analyze the achieved metrics, and make decisions regarding the practical application of the analysis results.

The remainder of this paper is organized as following. Section 2 introduces the data of stroke. Section 3 show algorithm using to build machine learning models and then analysis, evaluate the model and the experimental results of using our method for stroke prediction in our data. Section 4 is conclusion of the paper.

2 Diagnostic Data for Stroke

A stroke usually occurs suddenly when the blood supply to the brain is blocked, interrupted or impaired. Once there, the brain is deprived of oxygen, nutrition, and brain cells begin to die within minutes. People who have had a stroke are at high risk of death if they are not detected and rescued in time. This is one of the most dangerous and common neurological pathologies [9], [10].

Therefore, understanding the risk of stroke and performing primary prevention are critically important. For training and testing the machine learning model, we utilize a stroke prediction dataset collected from patients coming for medical examination at hospital 175 in Ho Chi Minh city. The raw dataset consists of 242701 observations (rows) and 27 features (columns). Each observation corresponds to one patient, and the important features are variables about the health status of each patient.

2.1 Data preprocessing

The raw data initially contained many unnecessary fields for the machine learning process and several fields were not standardized. We proceeded to eliminate meaningless fields from the existing 27 fields [17]. Given the diverse nature of the job categories, we created a 'work_type' field containing relevant job groups, including: Other, Not working, Officer, Laborer, Selfemployed. After processing, the current dataset includes variables such as Gender, Age, Work_type, Hypertension, Heart_disease, Avg_glucose_level, BMI, Smoking, and Stroke.

Descriptive statistics reveal that the dataset contains missing values specifically in the BMI, Smoking and Avg_glucose_level column. Due to the insignificant proportion of

missing values in the Smoking and Avg_glucose_level fields, we decided to remove records containing null values in these fields. However, for the BMI field, where there were a considerable number of missing values, we addressed this issue by filling the gaps with the median value of the same column. Additionally, we removed outliers with inappropriate BMI values, specifically those greater than 100, as they do not align with the standard BMI index.

These steps were taken to ensure the dataset is cleaned and optimized for classification purposes. By incorporating the information from the variables, we can effectively explore the characteristics of both independent and dependent variables, facilitating the training, evaluation, and fine tuning of models.

2.2 Data Visualization

The target column's stroke has a value of either 1 or 0. There are 222845 patients with stroke value equal to 0, and 17198 patients with stroke value equal to 1.

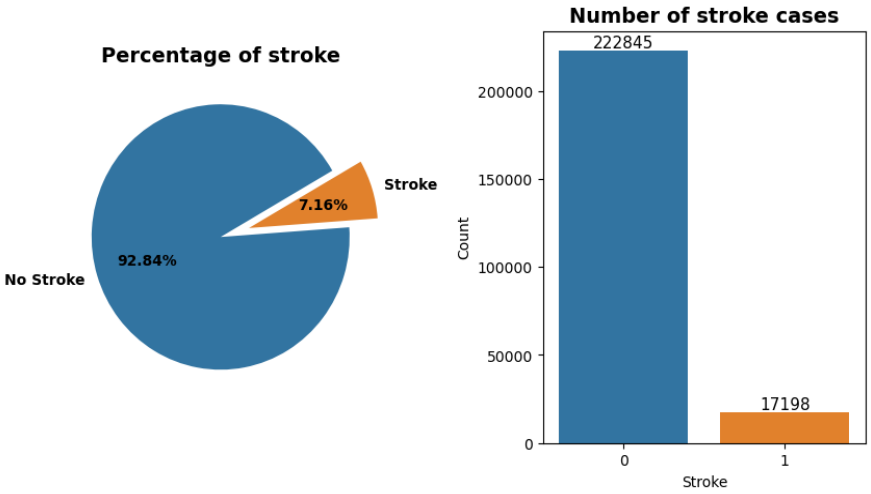

Figure 1. Distribution of Stroke Cases vs Non-Stroke Cases.

From Figure 1, it is clear that every 7 people out of 100 people are having strokes from our sampling data. Moreover, this is a highly unbalanced data distribution, and null accuracy score of this distribution itself is 93%, which employs any dump model should randomly predictions of stroke could reach accuracy of 93%.

The risk factors in high risk population of stroke can be divided into unpreventable factors and preventable factors. The former includes race, age, gender, and family history; the latter includes diabetes, hypertension, heart disease, hyperlipidemia, and smoking [10], [11]. The information about the factors influencing the risk of stroke such is visualized in Figures 2 and 3.

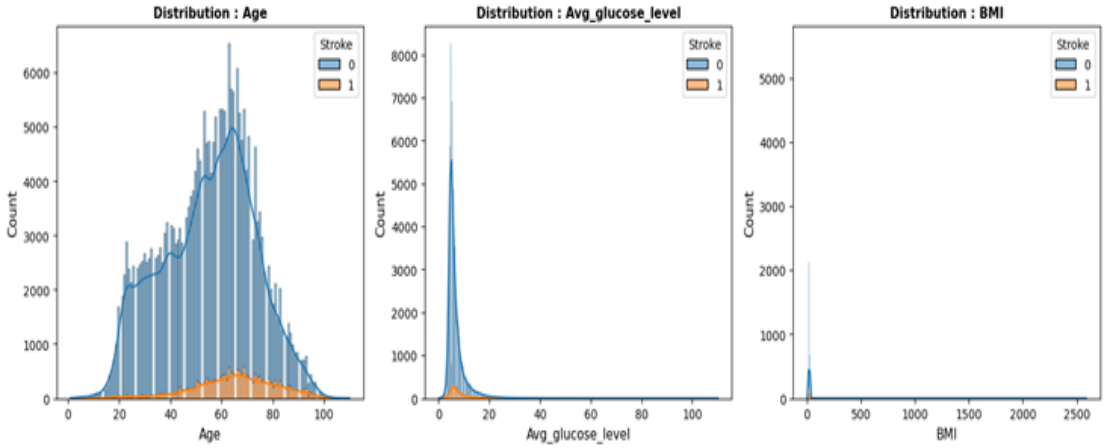

Figure 2. Visualizing distribution of continuous feature (age, avg_glucose_level, bmi).

There is a strong correlation between age and the risk of stroke. From Figure 2, we can observe that older individuals are more prone to strokes compared to younger individuals, with the highest number of cases occurring between the ages of 55 and 75.

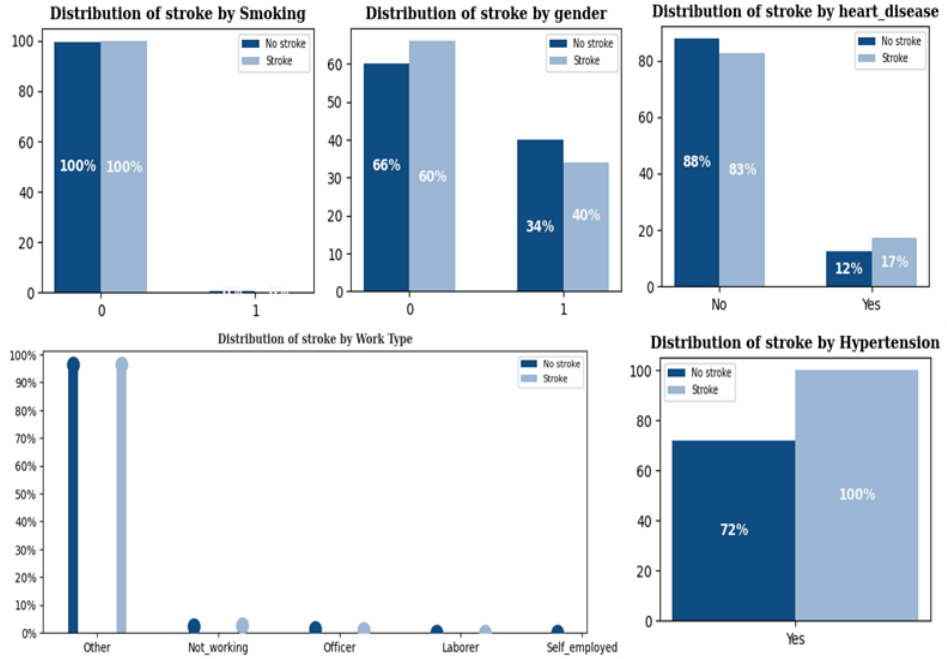

Figure 3. Visualizing distribution of stroke risk by some features.

The distribution of BMI follows a nearly normal distribution, but there are some values skewed to the right with fewer occurrences. The graph indicates that the majority of individuals attending medical appointments do not smoke. Due to the large number of unique data points in discrete features, obtaining detailed information becomes challenging. Therefore, we will convert features such as age, BMI, and avg_glucose_level into categorical features for visualization purposes. Individuals without heart disease have a higher risk of stroke compared to those with the condition. All stroke patients have high blood pressure, indicating a significant impact of hypertension on this condition.

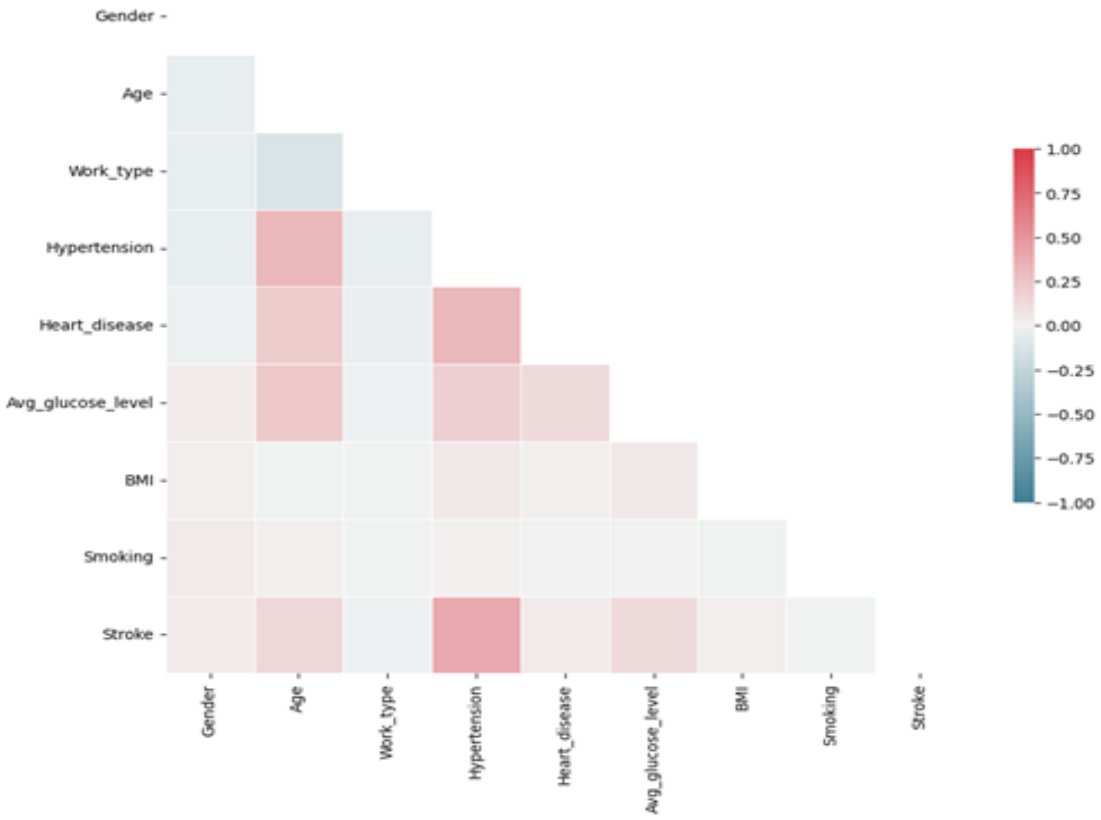

Figure 4. Correlation analysis in the utilized stroke prediction dataset.

Since the characteristics in the dataset are individual, it is necessary to analyze the correlations between features to examine the relationships among them, as well as their relationships with the target variable.

From Figure 4, we can find that hypertension and age and stroke are strongly positively correlated. Since this dataset is for classification, it is important to check the correlation between stroke and other variables. Using the information, we can decide to use which variables to predict the stroke.

3 Machine Learning Model

3.1 LightGBM Algorithm

Gradient boosting decision tree (GBDT) is a long-standing model in machine learning. Its main idea is to use a weak classifier, decision tree, and iterative training to get the optimal model, which has the advantages of good training effect and is not easy to over fit. LightGBM (Light Gradient Boosting Machine) is a framework that implements the GBDT algorithm and can be used for sorting, classification, regression, and various other machine learning tasks. Under the condition of not reducing the accuracy, the speed is increased by approximately ten times, and the memory occupied is reduced by approximately three times, which has the advantages of high training efficiency, low memory occupation, high precision, and support for parallelism, and GPU can be used to process large-scale data. [12], [13], [14].

LightGBM is fundamentally a learning algorithm based on a tree. While other such similar algorithms grow horizontal trees horizontally, Light GBM grows along leaf wise, or simply

put, vertically. The leaf with max delta loss will be chosen by the algorithm for growth. This helps reduce said loss in subsequent iterations. This also results in much better accuracy than existing gradient boosting algorithms [15]. However, the leaf-wise method can also lead to overfitting if not carefully controlled. Thanks to its remarkable advantages, LightGBM has gained widespread usage in various fields, including credit prediction, natural language processing, and data exploration [16].

The LightGBM algorithm [14]

Input:

Training data: $\mathrm{D} = \{(\chi_1, \mathrm{y}_1), (\chi_2, \mathrm{y}_2), \dots, (\chi_\mathrm{N}, \mathrm{y}_\mathrm{N})\}, \quad \chi_\mathrm{i} \in \chi, x \subseteq \mathbb{R}, \mathrm{y}_\mathrm{i} \in \{-1, +1\}$

Loss function: $L(y, \theta(\chi))$

Iterations:

M; Big gradient data sampling ratio:a; slight gradient data sampling ratio: b;

1: Combine features that are mutually exclusive (i.e., features never simultaneously accept nonzero values) of $\chi_i, i = \{1, \dots, N\}$ by the exclusive feature bundling (EFB) technique

2: Set $\theta_0(\chi) = \arg\min_c \sum_i^N L(y_i, c)$

3: For m =1 to M do

4: Calculate gradient absolute values:

$$r_i = \left| \frac{\partial L(y_i, \theta(x_i)}{\partial \theta(x_i)} \right|_{\theta(x)=\theta_{m-1(x)}}, i = \{1, \dots, N\}$$

5: Resample data set using gradient-based one-side sampling (GOSS) process:

$topN = a \times len(D);\ randN = b \times len(D)$

$sorted = GetSortedIndices(abs(r))$

$A = sorted[1{:}\,topN];\ B = RandomPick(sorted[topN{:}\,len(D)], randN);$

$D = A + B$

6: Calculate information gains:

$$V_j(d) = \frac{1}{n}\left(\frac{(\sum_{x_i \in A_i} r_i + \frac{1-a}{b}\sum_{x_i \in B_i} r_i)^2}{n_i^j(d)} + \frac{(\sum_{x_i \in A_r} r_i + \frac{1-a}{b}\sum_{x_i \in B_r} r_i)^2}{n_r^j(d)}\right)$$

7: Develop a new decision tree $\theta_m(x)'$ on set D'

8: Update $\theta_m(\chi) = \theta_{m-1}(\chi) + \theta_m(\chi)$

9: End for

10: Return $\tilde{\theta}(x) = \theta_M(x)$

3.2 Model for stroke prediction

The stroke dataset from Hospital 175, Ho Chi Minh, Vietnam, after being processed, cleaned, and encoding the categorical features into numerical values, can be utilized as input for machine learning models.

In practice, a problem can be approached by multiple methods, resulting in different models. Faced with such choices, we have conducted research and selected LightGBM, an algorithm that utilizes boosting techniques to create a powerful predictive model by combining multiple weak models into an ensemble. The prediction model is built using 70% of the data for training and 30% for testing, which are randomly split from the dataset.

Gradient boosting algorithms have been widely used to effectively address many machine learning problems, and one such algorithm is XGBoost, known for its speed and accuracy. However, LightGBM has gained more attention and is being used more frequently than XGBoost. LightGBM offers faster training speed and comparable model accuracy. Moreover, it provides users with more hyper parameters to fine-tune the model. Therefore, using the same gradient boosting technique, we trained and compared both the XGBoost and LightGBM models. Additionally, to gain a broader perspective, we also compared the LightGBM model with other machine learning methods such as Logistic Regression, Naive Bayes, and K Nearest Neighbors. The training and comparison results are presented in Table 1.

The LightGBM model relies on the features extracted from the dataset. The selection of these features is often based on experience and cannot guarantee the best performance.

The LightGBM algorithm helps in selecting important features based on their importance scores. One advantage of using gradient boosting is that after constructing the boosting trees, retrieving the importance scores for each feature is relatively straightforward. In general, feature importance provides scores that indicate the usefulness or value of each feature in building the boosted decision tree model.

Table 1. Comparison among the the machine learning model

Models	Accuracy	Precision	Recall	F1-score	ROC-AUC
LightGBM	0.9287	0.99	0.9287	0.0027	0.5
XGBoost	0.9286	1.00	0.9286	0	0.5
Logistic Regression	0.9286	1.00	0.9287	0	0.5
Naive Bayes	0.7411	0.72	0.9997	0.3548	0.86
K Nearest Neighbors	0.8997	0.94	0.9444	0.2804	0.61

It also helps in identifying a set of significant and meaningful features to include in the model, thereby improving prediction performance and reducing model complexity [10], [15]. We have ranked the importance of all variables in the LightGBM model to gain a better understanding of the role of each variable, as shown in the Figure 5.

From Figure 5, it can be observed that the five most important variables in the prediction model are age, average glucose level, BMI, hypertension. These variables significantly contribute to predicting whether a patient is at risk of stroke or not

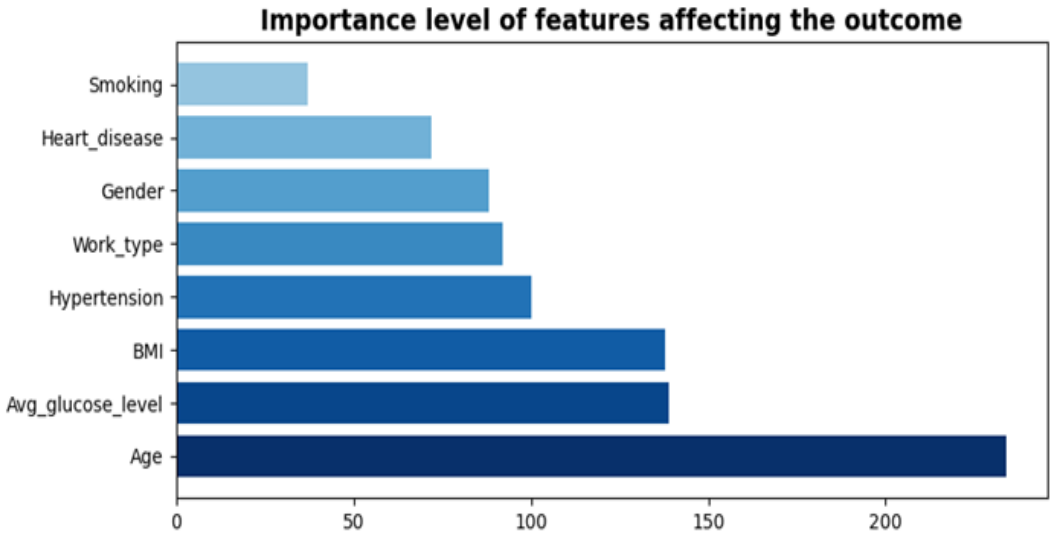

Figure 5. Ranking of feature importance in the LightGBM model.

3.3 Evaluation of the prediction performance of model

3.3.1 Labels of figures and tables

A learning curve shows how the generalization error, training error, and the complexity of a stochastic machine are interconnected. It reveals how the machine's performance improves with more training examples. By analyzing the learning curve, we can make informed decisions about the optimal training set size, model complexity, and potential regularization techniques to enhance the machine's performance.

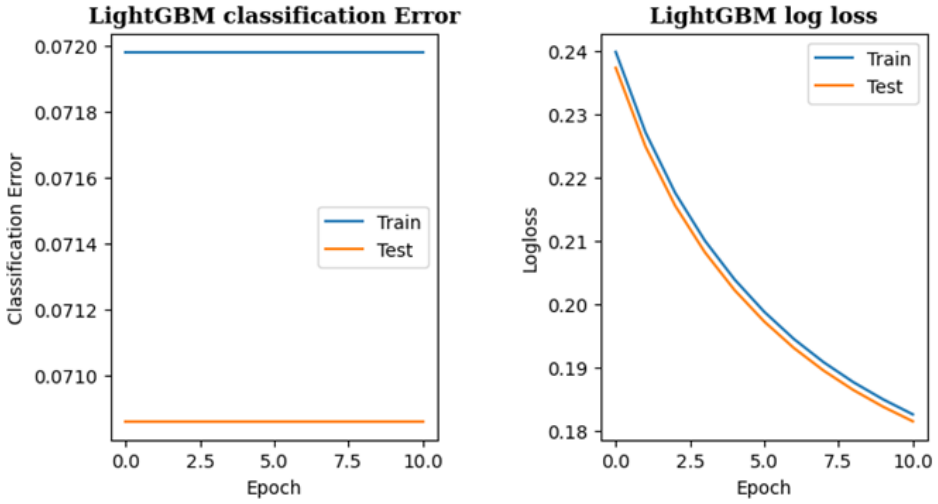

Figure 6. LightGBM Learning Curve.

From the Figure 6, it looks like there is an opportunity to stop the learning early, perhaps somewhere around epoch 5 to epoch 7.

3.3.2 Confusion Matrix

The confusion matrix in machine learning is used to evaluate the performance of a classification model. It provides an overview of the model's predictive ability by

comparing the model's predictions against the actual labels in the test dataset. The confusion matrix helps in assessing the model's accuracy, precision, recall, and other performance metrics by quantifying the number of true positives, true negatives, false positives, and false negatives.

The predictive model using LightGBM for the stroke dataset consisting of 72013 records has a Confusion Matrix displayed in Figure 7.

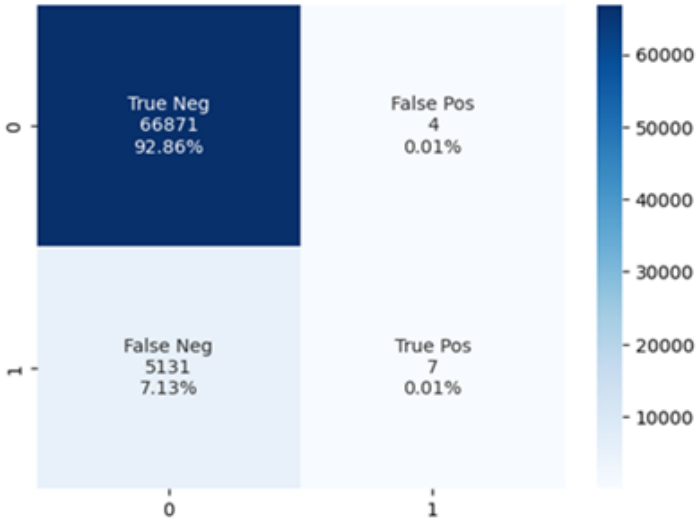

Figure 7. Confusion matrix.

The accuracy is the proportion of the number of correctly classified samples to the total number of samples in the given data. It is determined by the following formula: Accuracy = (TP + TN)/(TP+TN+FP+FN). The accuracy of the LightGBM machine learning model is 92.87%.

4 Conclusions

In this study, we have presented the steps of data analysis and the construction of a machine learning model using the LightGBM algorithm applied to the prediction of strokes from the stroke prediction dataset collected from patients coming for medical examination at hospital 175 in Ho Chi Minh city. With the machine learning model achieving a prediction accuracy of over 92%, it demonstrates significant promise for the future.

With this approach, the proposed method leverages community data through surveys to assist the machine learning model in achieving higher accuracy, thereby supporting prevention, diagnosis, and treatment of diseases and reducing treatment costs. The machine learning model for predicting stroke risk is packaged and embedded into a web application to assist users in understanding their own disease susceptibility. Moreover, it aids doctors in using the application to communicate with patients and evaluate support for model enhancements through their expertise and practical results.

References

[1] Jordan MI, Mitchell TM. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), 255–260.

[2] Jenni A. M. Sidey-Gibbons & Chris J. Sidey-Gibbons. (2019). Machine learning in medicine: a practical introduction. *BMC Medical Research Methodology*, 19, 1-18.

[3] A. Papadopoulou, D. Harding, G. Slabaugh, Marouli, P. Deloukas. (2022). Prediction of atrial fibrillation and stroke using machine learning models in UK Biobank. *medRxiv*.

[4] Senjuti Rahman, Mehedi Hasan, and Ajay Krishno Sarkar. (2023, January). Prediction of Brain Stroke using Machine Learning Algorithms and Deep Neural Network Techniques. In *European Journal of Electrical Engineering and Computer Science* (Vol. 7 (1), pp. 23-30). EJECE.

[5] You, Jia. (2023). Development of machine learningbased models to predict 10-year risk of cardiovascular disease: a prospective cohort study. *Stroke and Vascular Neurology 2023*.

[6] Van Lam Ho et. al. (2023). Sentiment Analysis by Lexical Analysis Combined with Machine Learning. *International Journal of Advances in Soft Computing & Its Applications*, 15(2).

[7] Rodrigo R, Fernández-Gajardo R. (2013). Oxidative stress and pathophysiology of ischemic stroke: novel therapeutic opportunities. *CNS Neurol Disord Drug Targets*, 12(5), 698-714.

[8] Smita Patil, Rosanna Rossi, Duaa Jabrah and Karen Doyle. (2022). Detection, Diagnosis and treatment of Acute Ischemic Stroke: Current and Future Perspective. *Frontiers in Medical Technology*, 4, 748949.

[9] Christian Grefkes and Gereon R. Fink. (2020). Recovery from stroke: current concepts and future perspectives. *Neurological research and practice*, 2(1), 1-10.

[10] Ho Van Lam, Vu Tuan Anh, Pham Thi Hoang Bich Diu, Tran Xuan Viet. (2021). APPLING MACHINE LEARNING TO PREDICT MELASMA. *International Journal of Computer Science and Information Security*, 19(11), 1-10.

[11] Xue Y, Chen S. (2022). The Prediction Models for High-Risk Population of Stroke Based on Logistic Regressive Analysis and Lightgbm Algorithm Separately. *Iranian Journal of Public Health*, 51(5), 999.

[12] Xu Y, Cai W, Wang L, Xie T. (2021). Intelligent Diagnosis of Rolling Bearing Fault Based on Improved Convolutional Neural Network and LightGBM. *Shock and Vibration*, 1-8.

[13] Cui B, Ye Z, Zhao H, Renqing Z. (2022). Used Car Price Prediction Based on the Iterative Framework of XGBoost+ LightGBM. *Electronics*, 11(18), 2932.

[14] Ke G, Meng Q, Finley T, Wang T. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in neural information processing systems*, 30.

[15] Ponsam JG, Gracia SJ, Geetha G. (2021, December). Credit Risk Analysis using LightGBM and a comparative study of popular algorithms. In *2021 4th International Conference on Computing and Communications Technologies (ICCCT)*, (pp. 634-641). IEEE.

[16] Zhang H, Ge L. (2019, October). Optimal Feature Selection for EMG-Based Finger Force Estimation Using LightGBM Model. *In 2019 28th IEEE international conference on robot and human interactive communication (RO-MAN)*, (pp. 1-7). IEEE.

[17] https://www.cdc.gov/stroke/about.htm. Access 18/02/2024.

Notes on contributors

Van Lam Ho received his Ph.D. degree in computer science and engineering at Yuan Ze University, Taiwan in 2016. He is lecturer at the Department of Information Technology, Quy Nhon University of Vietnam. His main teaching and research interests include Algorithm, Machine Learning and Data Science. He has published several research articles in international journals of Artificial Intelligence, Data Science, Computer Science.

Thi Phuong Thao Le is currently pursuing her Master's degree in Data Science at Quy Nhon University while working as a Developer Engineer at TMA Solutions.. Her primary work and research interests lie in application software, web development, and she possesses expertise in Machine Learning and Data Science.

Duy Nguyen V. M Master of Computer Science, Ho Chi Minh City University of Technology - HUTECH. He is Chief information technology officer at Military Hospital 175. His main teaching and research interests include Machine Learning, Artificial Intelligence, Computer Vision and Data Science.