Predicting Stroke Risk: A Logistic Regression and Data Visualization Approach

Xinyu Wang *

College of Arts and Sciences, New York University, New York, NY, 10003, United States

* Corresponding Author Email: xw2875@nyu.edu

Abstract. Strokes significantly threaten global health, underscoring the urgent need for accurate predictive models to aid in prevention efforts. This study delves into the application of logistic regression and data visualization techniques to analyze a dataset comprising 5110 stroke patients and their characteristics. This paper identifies key attributes associated with stroke risk, including age, hypertension, heart disease, and glucose levels. The logistic regression model demonstrates a high accuracy of approximately 95%, with nuanced insights provided by comparing single-year versus multi-year data models. This paper demonstrates the power of data visualization and machine learning to analyze complex relationships between patient attributes and stroke incidence, concluding that the integration of advanced analytics into clinical practice can significantly bolster stroke prevention, emphasizing a personalized approach to healthcare that prioritizes early detection and intervention. The study contributes to the medical community's arsenal against strokes, offering a robust healthcare predictive modeling framework. The synergy between data science and medicine plays a critical role in mitigating the impact of stroke, enhancing outcomes, and optimizing allocations.

Keywords: Stroke prediction, Losgistic regression, machine learning, data visualization.

1. Introduction

A stroke is a medical emergency usually caused by blood clots blocking or narrowing an artery leading to the brain, reducing blood supply and preventing brain tissue from getting oxygen and nutrients. Stroke has become an increasingly significant health problem [1]. Strokes can result in paralysis, loss of motor skills, and even death. It is the second leading cause of death [2]. Treating strokes consumes significant medical and financial resources. Therefore, preventing strokes is essential for reducing mortality, improving quality of life, and creating economic benefits [3].

Stroke prediction using data visualization is an effective tool for prevention, with numerous online resources and research studies providing insights, methods, and conclusions. Data Visualization is an interdisciplinary field that uses information to provide meaning data with graphs [4]. Integrating these models into healthcare systems allows clinicians to make informed decisions about preventive measures and treatments. Data visualization is being applied successfully to address complex public health challenges [5]. Integrating machine learning models and data visualization tools into healthcare practices is making stroke prevention more precise and effective. For instance, the advantages of applying machine learning in clinical settings have surpass traditional analysis methods, providing patients and healthcare providers with specific recommendations to mitigate stroke risk, such as lifestyle modifications, medical treatments, or monitoring strategies The test input data can be process through a decision tree and then the decision tree will provide a probabilistic decision. Putting several decision trees together creates a decision forest that generates aggregated tree predictions. Regression forests can be applied in anatomy localizations. Systems can be trained to detect abnormalities [6]. Furthermore, the continuous improvement of these predictive models through the incorporation of new data ensures that stroke prevention efforts are based on the most current and comprehensive information available. This iterative process enhances the accuracy of predictions and adapts to the population's emerging trends and risk factors.

This paper addresses three specific questions related to stroke prediction: what attributes are associated with strokes, how can stroke prediction have a higher accuracy, and whether using data from multiple years provides better predictions compared to using single-time-point data. Using data

visualization and machine learning tools, the study will analyze a dataset from Kaggle containing information on 5,110 stroke patients and their attributes, such as gender, age, and smoking status. By leveraging machine learning algorithms and data visualization techniques, the study aims to identify key attributes associated with strokes, build predictive models with high accuracy, and assess the impact of using longitudinal data for predictions.

2. Data and methods

2.1. Data source

The dataset this paper will utilize contains information of 5110 stroke patients. In particular, the 10 attributes that the dataset recorded are the gender of the patient, the age of the patient, whether the patient has hypertension, whether the patient has heart disease, whether the patient is married, the work type of the patient, the residence type of the patient, the average glucose level of the patient, the body mass index of the patient, and the smoking status of the patient.

Kaggle is an online platform primarily used for data science and machine learning. It provides access to a vast collection of datasets that user can explore and use for research and projects. These datasets span various topics, and healthcare is one of the main ones.

Since this paper will use Python as a tool to process and analyze the dataset, the attributes will be categorized first. Age is the only attribute that can be represented with an integer. Gender, marital status, work type, and residence type can be represented by strings. Whether the patient has hypertension or heart disease can be represented by Boolean values (0 for no and 1 for yes). Lastly, average glucose level and body mass index can only be represented using floats since they include numbers after the decimal point.

2.2. Methods

2.2.1. Attributes:

To identify attributes associated with strokes, the study needs to find the correlations of each attribute with stroke incidence. This paper uses Python libraries like Pandas NumPy and Matplotlib to find correlations and plot them into graphs. These libraries will be used in the remaining tests as well.

The process starts by loading the dataset from a CSV file using Pandas. After loading the data, the code encodes categorical variables, for example “gender”, into numerical values. This encoding is necessary for correlation analysis, as correlation computations require numerical data.

Once the categorical variables are encoded, it is necessary to calculate the correlation matrix of the dataset. The correlation matrix reveals the strength and direction of the relationships between pairs of variables. The primary focus is to determine how each attribute correlates with stroke.

2.2.2. Algorithms:

Logistic regression is an evaluation method widely used in statistics and machine learning. It is mainly used to analyze the impact of one or more independent variables on a binary dependent variable. Unlike linear regression, the dependent variable of logistic regression is discrete, usually taking the value of 0 or 1, representing two different categories or states. In order to effectively handle this classification problem, logistic regression uses a logistic function, which maps any real value of the input to a probability value between 0 and 1, so that it can be used to predict the probability that a sample belongs to a certain category.

Logistic Regression is a widely used statistical method, mainly used for binary classification problems. Logistic regression is used to predict the probability of an event occurring. Logistic regression is interpretable, easy to implement, and robust to outliers. It provides odds ratios of characteristic effects to help understand the influence of variables. The model structure of logistic regression is relatively simple and easy to understand and interpret. The coefficients output by the model can directly reflect the impact of features on the classification results. Although logistic

regression is relatively simple, overfitting may still occur when the number of features is much larger than the number of samples. At this point, regularization techniques such as L1 or L2 regularization can be used to mitigate overfitting.

This paper also applys several crucial data preprocessing steps to support the logistic regression. Data encoding is converting categorical data into a numerical formet to facilitate computations. Normalization involves giving the dataset a mean of zero and a standard deviation of one to guarantee that all features contribute equally to the model's performance.

2.2.3. Predictions:

Considering the multifactorial nature of stroke, this paper employs logistic regression for its prediction.

First implements logistic regression from scratch to predict stroke occurrences using the dataset. Logistic regression is preferred for stroke prediction due to its interpretability, robustness, and ease of implementation [7]. Relevant features (age, heart disease, average glucose level, hypertension, and marital status) are selected, and the target variable (stroke) is isolated. Normalizing the features to have a mean of zero and a standard deviation of one improves the gradient descent algorithm's convergence.

In order to ensure that the model can be assessed on data that has not yet been seen, the data is then divided into training and testing sets, with 80% utilized for training and 20% for testing.

Logistic regression parameters are initialized to zeros, and the sigmoid function is defined to compute the hypothesis. The cost function, which measures the difference between predicted and actual values, is minimized through gradient descent. Predictions are made using the learned parameters, and the model's accuracy is computed. The prediction function determines the probability of stroke occurrences and makes predictions on the test set.

2.2.4. Comparing:

To find out whether using data from multiple years improves the accuracies of the predictions, this paper splits the dataset into different “years”. Then it compares the performance of models trained on each subset to simulate a scenario where we have data from multiple years.

Evaluations of the accuracy of logistic regression models trained on data from single years versus data from multiple years are used to predict stroke occurrences. It simulates data from multiple years by adding a “year” column with random assignments. Logistic regression is implemented through functions for the sigmoid hypothesis, cost function, and gradient descent, which iteratively minimizes the cost to improve the model. A custom function using NumPy splits the data into training and testing sets.

For each year and for the combined dataset, models are trained and evaluated, printing the accuracies and visualizing them with a bar plot that includes annotations showing accuracy values directly. This comprehensive workflow demonstrates the impact of using a larger dataset spanning multiple years on model accuracy, clearly comparing single-year and multi-year data performance.

2.3. Evaluation Metrices

The efficacy of our logistic regression model for stroke prediction is gauged through a suite of evaluation metrics that offer a comprehensive understanding of its predictive accuracy.

Accuracy serves as the foundation, reflecting the proportion of correctly predicted cases, whether true positive or true negatives, against the total instances [8]. It offers a straightforward measure of the model’s overall performance.

By assessing the model's capacity to accurately detect real stroke cases among all projected positives, precision serves as a counterbalance to accuracy.

To further assess the model’s diagnostic capability, the Confusion Matrix for training and testing delineates the true and false positives and negatives, offering a granular view of prediction outcomes [9].

Through the utilization of these metrices, the research not only measures the prediction precision of the model but also pinpoints possible avenues for enhancement. These metrices collectively enable a robust evaluation framework, guiding iterative refinements to enhance the model's predictive prowess in stroke prediction.

3. Results

3.1. Correlation Analysis

As shown in Figure 1, it is a heatmap depicting the correlation matrix of various attributes related to stroke patients. Red denotes a strong positive connection and blue a strong negative correlation. The color intensity represents the degree of the association. From figure 1 it is obvious that age, hypertension, heart disease, and average glucose level have relatively stronger correlations with strokes.

As shown in Figure 2, it is a bar chart illustrating the correlation coefficients of each attribute with stroke. Figure 2 shows that age is the most strongly correlated with stroke, followed by heart disease, average glucose level, hypertension, and marital status. BMI, smoking status, residence type, and gender are relatively weaker. Work type is negatively correlated with stroke. However, the graph indicates that all attributes have a correlation coefficient with stroke below 0.25, suggesting no significant single predictor.

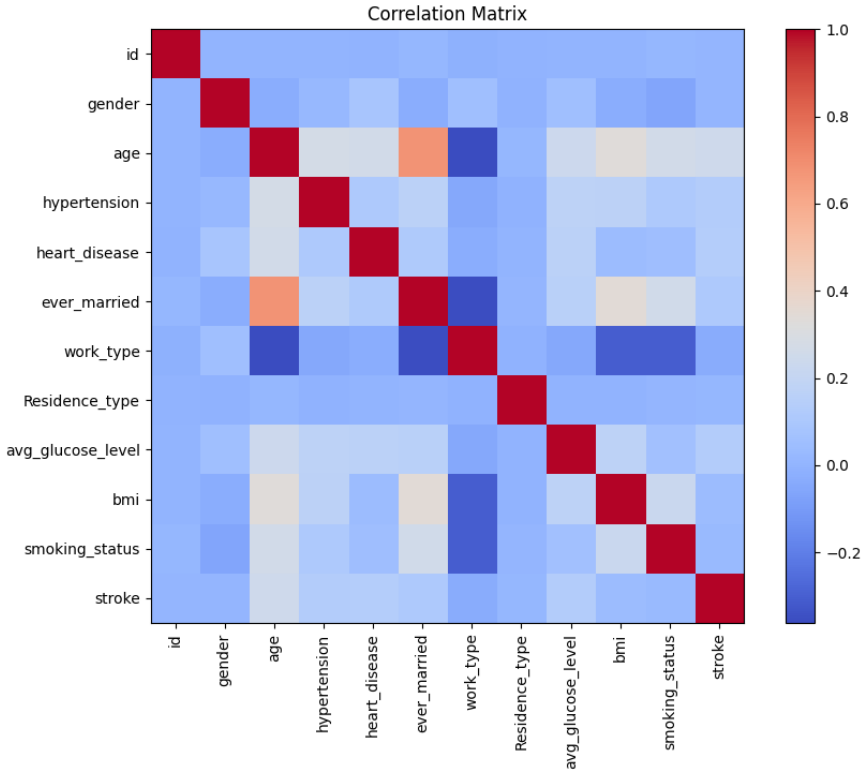

Figure 1. Correlation heatmap

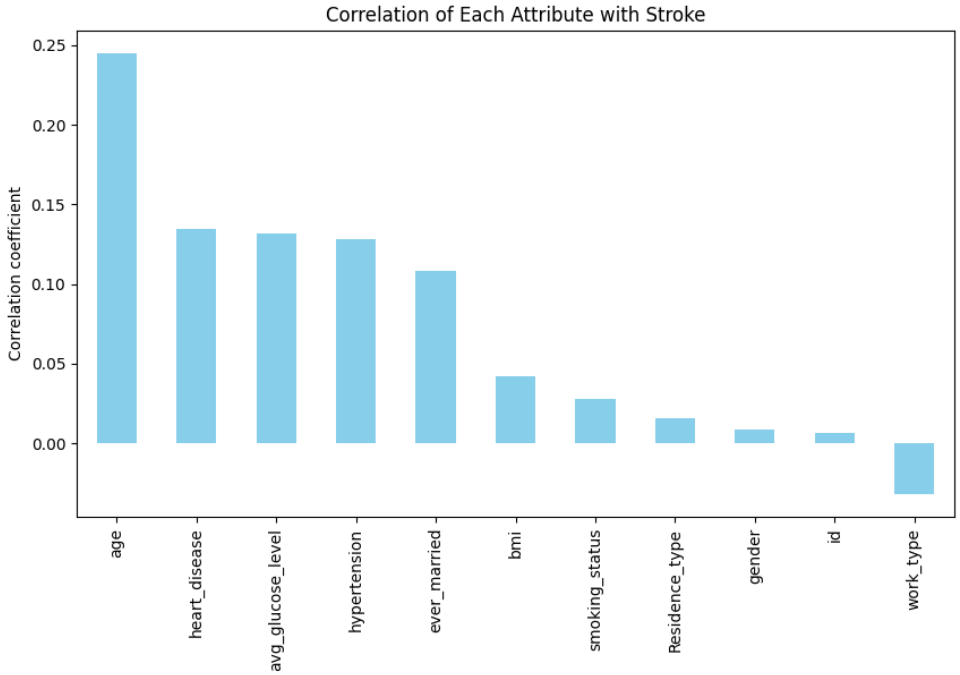

Figure 2. Correlation bar chart

3.2. Accuracy Evaluation

Figure 3 shows the cost function history of a logistic regression model during training. The y-axis represents the cost, while the x-axis represents the number of iterations. The curve illustrates a gradual decrease in the cost over 1000 iterations, indicating that the gradient descent algorithm is successfully minimizing the cost function, leading to better model performance.

Figure 4 is a confusion matrix visualizing the performance of the logistic regression model. True Positives (bottom-right), True Negatives (top-left), False Positives (top-right), and False Negatives (bottom-left) make up the four sections of the matrix. On the y-axis are the actual values, while on the x-axis are the predicted values. The color intensity represents the number of forecasts; red denotes more predictions, while blue denotes fewer. This confusion matrix shows that the model makes over 1000 correct prediction out of 1200 total predictions.

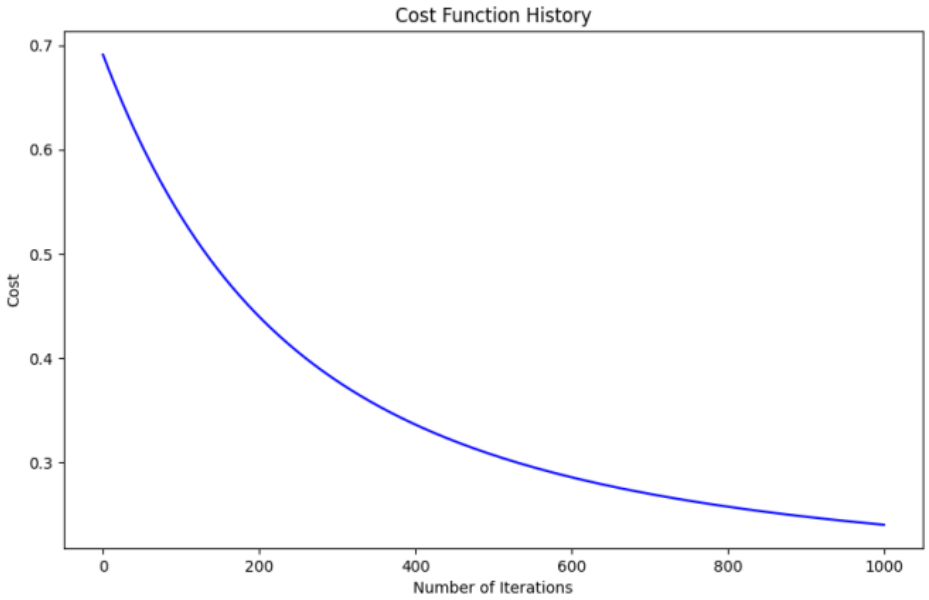

Figure 3. Cost Function

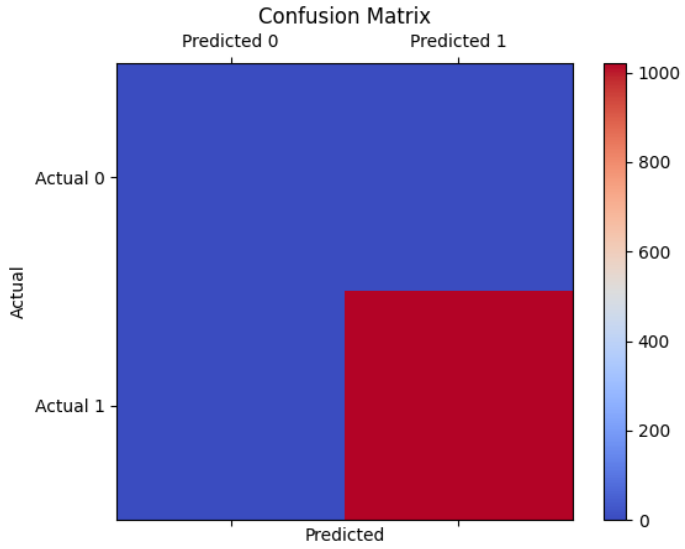

Figure 4. Confusion Matrix

Figure 5 is a bar chart illustrating the accuracy of logistic regression models trained on data from individual years versus a combined dataset spanning multiple years. The y-axis represents the accuracy percentage, while the x-axis lists the different years and the multi-year dataset. Each bar shows the model's accuracy for the respective year or dataset, with the exact accuracy value annotated above each bar. The models for the single years show accuracies of 95.45%, 96.43%, 95.56%, and 94.14%. The combined multi-year model shows an accuracy of 93.93%. This visualization suggests that while individual-year models perform slightly better, the multi-year model still maintains high accuracy, indicating that using data from multiple years does not significantly degrade performance and can still be effective for prediction.

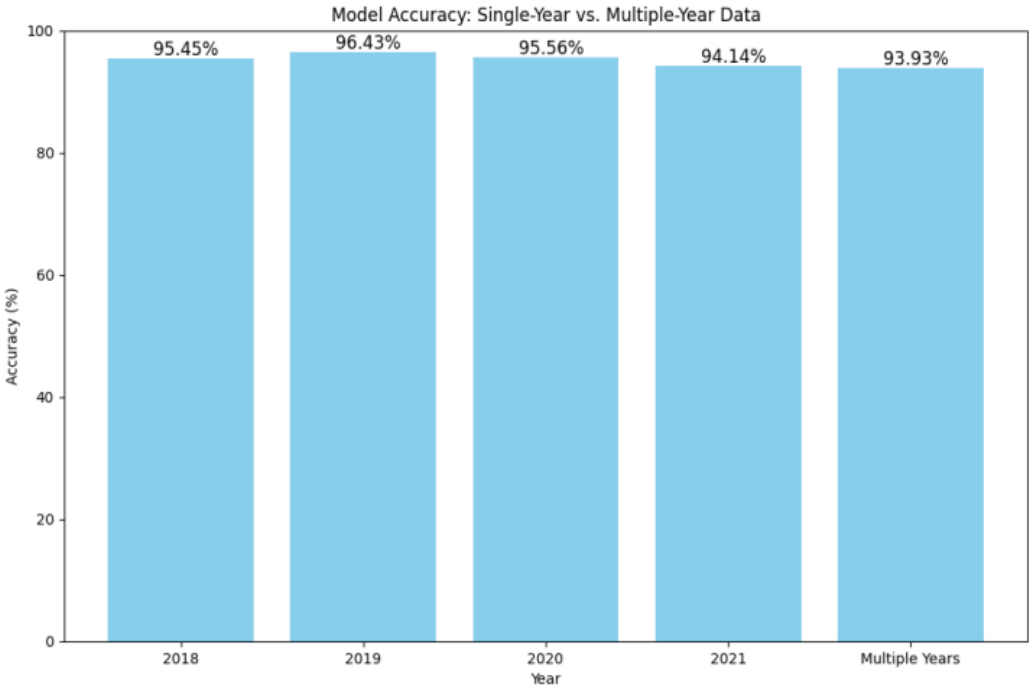

Figure 5. Model accuracy bar graph

3.3. Comparison and Discussion

Logistic regression is advantageous for its interpretability, ease of implementation, and robustness to outliers. It provides odds ratios for feature effects, aiding in understanding variable impacts. However, it assumes a linear relationship between the log-odds and predictors, which may not always hold [10]. Unlike decision trees or neural networks, logistic regression is less flexible in capturing

complex patterns but offers greater simplicity and transparency. It may underperform in cases where interactions between features are significant or the data is highly non-linear, which is where ensemble methods or deep learning might excel. Despite the simplicity and transparency of logistic regression being its advantage, it may perform poorly when dealing with highly nonlinear and complex feature interactions. In this case, integration methods (e. g., random forest, gradient lifting trees) or deep learning models (such as convolutional neural networks, recurrent neural networks) may be more suitable. Logical regression is an ideal choice when the number of features is small and the relationship between the features is relatively simple.

4. Conclusion

This paper shows the power of data visualization and machine learning to analyze the complex relationships between various patient attributes and the incidence of stroke. The combination of data visualization and machine learning can help researchers deeply understand the complex relationship between patient attributes and stroke incidence, identify potential risk factors, and provide data support for clinical decision making.

Through the analysis of a dataset of 5110 stroke patients, this paper has identified key attributes that exhibit correlations with stroke occurrences. The logistic regression models employed have demonstrated commendable accuracy of around 95%, with the multi-year dataset maintaining a high level of performance. The findings suggest that while individual-year models marginally outperform the multi-year model, the latter still provides robust predictions.

The evaluation metrics have provided a clear and quantifiable assessment of the model's performance. These metrics have not only validated the model's efficacy but have also highlighted areas for potential improvement. The comparison with other methodologies, such as decision trees and neural networks, reveals that while logistic regression offers interpretability and robustness, it may fall short in. scenarios characterized by complex patterns or non-linear relationships.

This paper demonstates that machine learning combined with healthcare practice can significantly improve the precision of stroke prevention efforts. The continuous refinement of predictive models through the incorporation of new data ensures that these tools remain at the forefront of medical innovation. With the wide application of medical imaging and electronic health records, logistic regression can achieve real-time prediction and decision support in a big data environment. For example, with cloud computing and distributed computing technologies, logistic regression can handle massive amounts of stroke related data to achieve rapid prediction. Update the model through real-time data to adapt to the dynamic changes of the data and maintain the effectiveness of the model. Furthermore, logistic regression can be used in personalized medicine for risk assessment of stroke, to help develop personalized prevention and treatment options. Identify high-risk groups through risk assessment, conduct early screening and intervention to reduce the incidence of stroke. Finally, logistic regression is used to classify the patients to provide personalized monitoring and treatment options.

Ultimately, integrating these predictive models into clinical decision-making processes can revolutionize stroke prevention, offering a more personalized and proactive approach to healthcare. The synergy between data science and medicine will undoubtedly play a pivotal role in mitigating the devastating impact of strokes, improving patient outcomes, and optimizing the allocation of healthcare resources.

References

[1] E. D. Simon, O. Olatunji, U. D. Itanyi, & J. O. Aiyekomogbon. Computerized tomographic brain findings in hypertensives with stroke in Abuja, Nigeria. Journal of Radiation Medicine in the Tropics, 5 (1), 2024, 7 - 13.

[2] I. Kortazar-Zubizarreta, A. Pinedo-Brochado, I. Azkune-Calle, U. Aguirre-Larracoechea, M. Gomez-Beldarrain, & J. C. Garcia-Monco. Predictors of in-hospital mortality after ischemic stroke: a prospective, single-center study. Health Science Reports, 2 (4) 2019.

[3] J. V. E. Gerstl, S. E. Blitz, Q. R. Qu, A. G. Yearley, Philipp Lassarén, & R. Lindberg, et al. Global, regional, and national economic consequences of stroke. Stroke, 54 (9), 2023, 2380 - 2389.

[4] A. Bachmeier, & M. Heroux. Data Visualization through Graphic Representation, 2023.

[5] J. Shoultz. Geographic information systems in public health data visualization, analysis and sharing, 2015.

[6] A. Callahan, & N. H. Shah. Machine learning in healthcare. Key Advances in Clinical Informatics, 2017, 279 - 291.

[7] C. G. Soh, Y. Zhu, & T. L. Toh. A regularized logistic regression model with structured features for classification of geographical origin in olive oils. Chemometrics and Intelligent Laboratory Systems 2023.

[8] Lijuan Zhang. Data quality: the accuracy dimension (book review).

[9] B. Gaye, & A. Wulamu. Sentiment Analysis of Text Classification Algorithms Using Confusion Matrix, 2019.

[10] S. Simsin, & J. Khiewyoo. A Report of the Management of Continuous Independent Variables in Logistic Regression Analysis in Medical and Public Health Journals in Thailand 2017.