Predicting the Risk of having a Stroke using Demographics and Health Factors of a Person(Machine Learning Application)

Uduli T. Iyenshi
Dept. of Electrical and Information Engineering
Faculty of Engineering
University of Ruhuna
Galle, Sri Lanka
iyenshiaut@gmail.com

Kalangi T. Jayakody
Dept. of Electrical and Information Engineering
Faculty of Engineering
University of Ruhuna
Galle, Sri Lanka
thathsarakalangi@gmail.com

***Abstract*— A stroke is a life-threatening emergency occurring when a blood vessel ruptures or a clot disrupts blood flow to a part of the brain. A stroke can impact specific bodily functions if blood flow to the corresponding brain region is affected. In essence, it results in the death of brain cells due to a lack of oxygen, leading to potential long-term consequences. The World Stroke Organization(WSO) claims that stroke remains the second leading cause of death and the third leading cause of death and disability combined in the world. [1] The severity of the impact depends on how quickly the treatment is received. Therefore, early detection of stroke warning symptoms can reduce the stroke severity. The main objective of this study is to forecast the possibility of a stroke occurring at an early stage using some machine learning techniques and algorithms. An unbiased dataset was taken from the Kaggle website to improve the algorithm's and the study's efficacy. For the classification part of this study, Naïve Bayes and Logistic Regression, two supervised learning algorithms, were effectively applied. To evaluate the effectiveness of the models, accuracy, F1 score, specificity, and sensitivity were used, with priority given to F1 score. According to the calculations, for the naïve Bayes model, the F1 score was 0.816, and the accuracy was 0.774. Similarly, for logistic regression, the F1 score was calculated as 0.818, with an accuracy of 0.81. The overall model evaluation indicated that the best-performing model was logistic regression, which is trained with hyperparameters c and solver, resulting in the highest performance in terms of both F1 score and accuracy.**

Keywords—Machine Learning, Logistic Regression, Naïve Bayes, Scaling, Hyperparameter Tuning

I. INTRODUCTION

A stroke or a brain attack is a dangerous medical disorder that occurs when the blood flow to the brain is disrupted or when a blood vessel in the brain bursts, resulting in neurological disability. The National Heart, Lung, and Blood Institute (NHLBI) informs that 87% of strokes can happen due to blood blockage in the brain. As they published, the chance of having a stroke after age 55 will be doubled every 10 years. According to the 2022 version of the Global Stroke Factsheet, the chance of having a stroke has risen by 50% in the past 17 years, and it estimates 1 in 4 people expected to experience one at some point in their lives. WHO claims a 43% increase in deaths due to stroke from 1990 to 2019. [2] Having health conditions including high blood pressure, high cholesterol, heart disease, diabetes, obesity, and sickle cell disease increases the risk of stroke. With the help of machine learning algorithms, significant advancements have been developed to predict medical issues, identify them, and treat numerous of them at an earlier stage. Stroke is one of the several diseases that can be prevented if detected early and can be treated if predicted in the early stages. Machine learning algorithms are the foundation for diagnosing and predicting disease in health, which can be used to overcome lots of issues in earlier stages.

Taking these factors into consideration, this study helps to forecast future stroke risk, which is beneficial to most individuals. Moreover, it guarantees that people can be ready for any situation and can start treatments or medical therapies on time if needed.

In this study, the goal is to predict the risk of having a stroke or not by applying the selected models' algorithms to a reliable dataset sourced from the Kaggle website. This study aims to compare the performance of each model in terms of some different parameters (accuracy, F1 score, sensitivity, specificity) to ultimately determine the best model for predicting the risk of having a stroke among the two models under consideration.

After going through the characteristics of the dataset that was chosen, supervised learning classification algorithms were selected as the most appropriate algorithms to develop the model for this study. Therefore, two supervised learning algorithms, namely logistic regression and naïve Bayes were employed in this study to develop the model.

A. Logistic Regression

Logistic regression analysis is a multidimensional technique to evaluate the relationship between multiple predictor variables and an outcome(target) which is binary(dichotomous) [3] it is a supervised learning algorithm, and for this particular study, logistic regression was successfully applied because of the binary outcome with the probability output. As the output interprets the likelihood of having a stroke, which is more meaningful as a healthcare concern. Due to the relationship between the features and probability of having a stroke may not be strictly linear, logistic regression accommodates non-linear regression more effectively than linear regression. In contrast, the study

assumes there are linear relationships between independent variables and the target, but as it is not strictly linear, it can influence the model.

B. Naïve Bayes

It is a supervised learning algorithm as well as a probabilistic classifier since it is based on Bayes' theorem. The fundamental concept of the Naïve Bayes algorithm is that the existence of one feature or parameter is not necessary for the presence of other parameters, hence the existence of one feature is independent of the existence of other features. Since the dataset contains categorical features, naive Bayes assumes that the features are independent of class labels, so that the calculations will be simplified. Moreover, naïve Bayes provides probability estimation, allowing for the measurement of the uncertainty in stroke risk predictions. In contrast, the assumption of feature independence might be limiting in this study as certain features, hypertension, and heart disease could be correlated and the dependency between features may not be accurately captured.

II. METHODOLOGY

The study utilized the dataset of stroke prediction that can be accessible on the Kaggle website, consisting of 12 columns and 5110 rows. (mention the table where it can be found) The output column, 'stroke' is a binary value of either 1 or 0, where 1 indicates the presence of a stroke risk, and 0 represents the absence of a stroke risk. notably, the dataset is substantially imbalanced, with the higher frequency of 0s in the stroke column. In the data set, there are 249 instances of a stroke risk (1) and 4861 instances without a stroke risk (0).

Table II.1 Visualization of feature selection

	Attribute	Attribute Description
1	id	Unique Identifier
2	gender	"Male", "Female" or "Other"
3	age	age of the patient
4	hypertension	0 if the patient doesn't have hypertension, 1 if the patient has hypertension
5	heart_disease	0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
6	ever_married	"No" or "Yes"
7	work_type	"children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
8	Residence_type	"Rural" or "Urban"
9	avg_glucose_level	average glucose level in blood
10	bmi	body mass index
11	smoking_status	"formerly smoked", "never smoked", "smokes" or "Unknown"* ("Unknown" in smoking_status means that the information is unavailable for this patient)
12	stroke	1 if the patient had a stroke or 0 if not

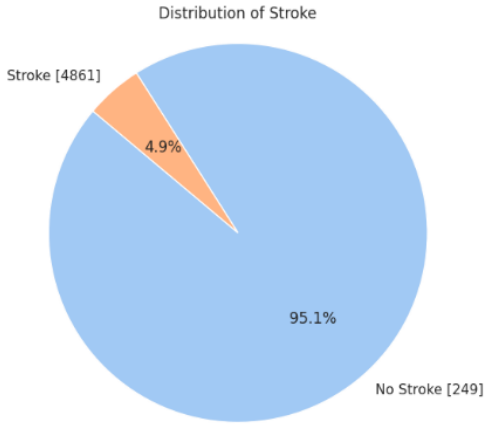

Figure II.1 Distribution of Stroke Before Undersampling

As the data set mentioned, the 12 attributes are described above. According to the dataset, the output variable will be the stroke, so it is the predictor, and other features will be used to predict the stroke.

A. Data Preprocessing

The "gender" feature is included in a data row labeled as "other", which can be removed since the decision was based on its limited impact on modeling, considering the other categories have a significantly higher frequency in the dataset.

The features "Gender, ever_married, work_type, residence_type, smoking_status" are of object type, indicating those are categorical features. To make them compatible with modeling algorithms, those features were converted to numerical values by using one-hot encoding. The encoded columns were then concatenated to the original data while removing the corresponding categorical columns for further analysis.

The correlation between each feature was analyzed using the correlation matrix.

To address the imbalance, preprocessing techniques, and under sampling we employed. This method aims to balance the dataset and enhance the predictive accuracy. Figure 1 illustrates the visual representation of class counts before and after the under-sampling with RandomUnderSampler(). After doing the under sampling the value of stroke with 0 has decreased to the level of 249, while the value of stroke with 1 has remained the same.

The dataset was divided into training and testing sets, allocating 20% for testing. This percentage aims to have a balance, ensuring a significant set for modeling while having a majority for the training set to robust the models. To standardize "age","avg_glucose_level", and "bmi" features, Standard Scaler was used because of the effectiveness in normalizing features, thus make the comparison. This method was chosen over others as it centers data around zero and scales to unit variance, hence preventing features with larger scales. The training data was fit_transfomed, while the test dataset was transformed to maintain consistency in scaling.

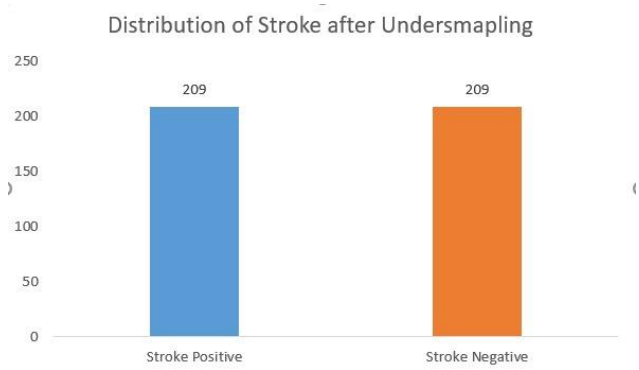

Figure II.2 Distribution of Stroke after Under sampling

B. Algorithm Explanation

1) Logistic Regression

Out of all the different types of logistic regression algorithms, binary logistic regression was used as the dataset runs with a binary outcome. The logistic function (sigmoid function) is as follows.

$$P = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

This function has an S-shaped curve. P is the probability of a binary outcome. And it ranges between 0 and 1. β_0 is the intercept term in the logistic regression. β_1 is the coefficient associated with the predictor variable x, here, x is the predictor variable. That is the input to the function.

This function predicts the probability of an instant belonging to a specific class. For a model with logistic regression, in the context of sci-kit-learn's Logistic Regression() model fit the data to a logistic regression model assuming a linear relationship between predictor variables and the binary outcome. After fitting the data, an internal process is applied to adjust coefficients(β_0 and β_1) to align with the best-fit line. Hence the predictions are made using the logistic function.

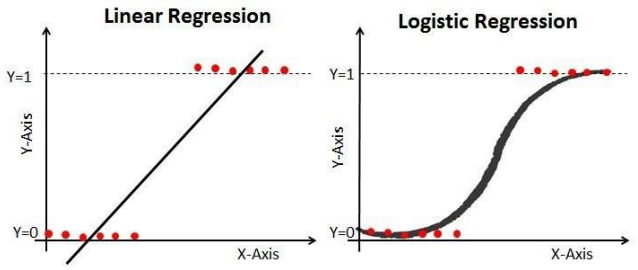

Figure II.3 Graph Representation of Linear Regression and Logistic Regression [4]

2) Naïve Bayes

Naïve Bayes classifiers are considered simple probabilistic classifiers that apply Bayes' theorem. Bayes theorem provides a way to calculate posterior probability using other different probabilities.

$$P(A|X) = \frac{P(X|A) \cdot P(A)}{P(X)}$$

In this function, $P(A|X)$ is the posterior probability, $P(X|A)$is the likelihood, $P(A)$ is the prior probability and $P(X)$ is the evidence. In this study, the Gaussian Naïve Bayes was used assuming that feature values are continuous and have a Gaussian distribution. For each feature, Gaussian Naïve Bayes estimates the mean and the standard deviation for each class. GaussianNB () is a class in the sci-kit learn library that implements Gaussian Naïve Bayes. When the training data is fit with GaussianNB (), it internally estimated the parameters of the Gaussian Distribution for each class.

C. Modeling and Post-Processing

Each algorithm was fit to the dataset and then by predicting the values for the test dataset, the model effectiveness was evaluated. For that, the model scores, accuracy, precision, recall, F1 score, confusion matrix, and classification report were used.

To have optimal results in the performance of the model, hyperparameter tuning was done. In logistic regression, the model is fine-tuned through a grid search cross-validation. After initializing the logistic regression model, a grid of hyperparameters is defined with different regularization strengths and optimization solvers. The model's performance is evaluated using K-fold cross-validation with 20 folds to enhance the robustness. After following the grid search, the best hyperparameter values were identified, and the model was refitted with those parameter values. This process ensures that the model is fine-tuned to maximize its predictive capabilities.

Gaussian naïve Bayes model is fine-tuned with the process through the cross-validation with 5 folds to enhance the robustness in assessing the model's generalization. It was fine-tuned with the hyperparameter including different prior probabilities and var_smoothing values. After identifying the best hyperparameter values, the naïve Bayes model was refitted with these parameters.

III. Results

A. Comparison Before Hyperparameter Tuning

1) Logistic Regression

Also before hyperparameter tuning, the logistic regression model maintains superior performance with higher accuracy, precision, recall, and f1 score. The confusion matrix depicts a balanced distribution between true positives, true negatives, false positives, and false negatives, indicating that well-performed classifier.

2) Naïve Bayes

Before fine-tuning the naïve Bayes model, it shows comparatively lower accuracy and precision than the logistic regression model. While it achieves perfect recall, the precision is affected, hence resulting in a lower f1 score. Further, the confusion matrix depicts the challenges in correctly predicting class 0, leading to a disparity in precision between the two classes.

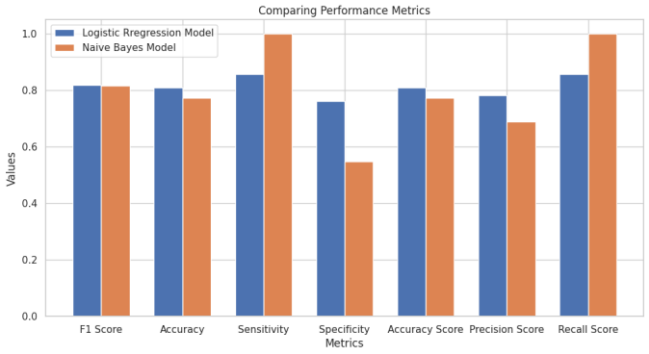

Figure III.1 Compare Performance Metrics between Logistic Regression and Naïve Bayes Models

B. Comparison After Hyperparameter Tuning

After doing the hyperparameter tuning for both models, the performance has increased for both models. Before tuning, the logistic regression model showed an accuracy of 0.595, with a precision of 0.553 and a recall of 1.0. however, after tuning, the model's accuracy increased to 0.810, along with a precision of 0.783 and recall of 0.857. This improvement is affected to enhance the f1 score to 0.818, showing a more balanced trade-off between precision and recall. Further, the confusion matrix reveals a notable reduction in misclassifications, especially predicting class 0, as indicated by the decrease in false positives.

Similarly, in the naïve Bayes model, before tuning the accuracy was 0.60 with a precision of 0.553 and recall of 1.0. After doing the tuning, the accuracy increased to 0.810, the precision to 0.783, and a recall of 0.857. The enhancement affected the f1 score which increased from 0.711 to 0.818, emphasizing the model's improved ability to balance recall and precision. The confusion matrix also reveals that a reduction in misclassifications, especially false positives, contributes to the enhancement of the model's accuracy.

IV. Discussion

By examining the performance metrics, including the accuracy, precision, recall, and f1 score, the tuned logistic regression model stands out as the best model out of the two models. The logistic regression models performed higher metric values rather than the naïve Bayes model, showing that it can correctly classify instances and maintain a more balanced precision-recall trade-off.

The interpretation of the confusion matrix emphasized the logistic regression model's improved predictive capabilities, notable in minimizing misclassifications, such as false positives and false negatives. As this study is related to the healthcare sector, this aspect is critical. Moreover, the logistic regression model's simplicity makes it transparent in the decision-making process making it easier for healthcare professionals and patients to trust the model's predictions. As the challenges in achieving a comparable level of performance in the naïve Bayes model, limitations may occur as this study goes through in the healthcare sector. In conclusion, due to the model's interpretability and the balance between accuracy and fairness, the logistic regression model is a more reliable and ethically sound choice for the specific task of stroke prediction in this dataset.

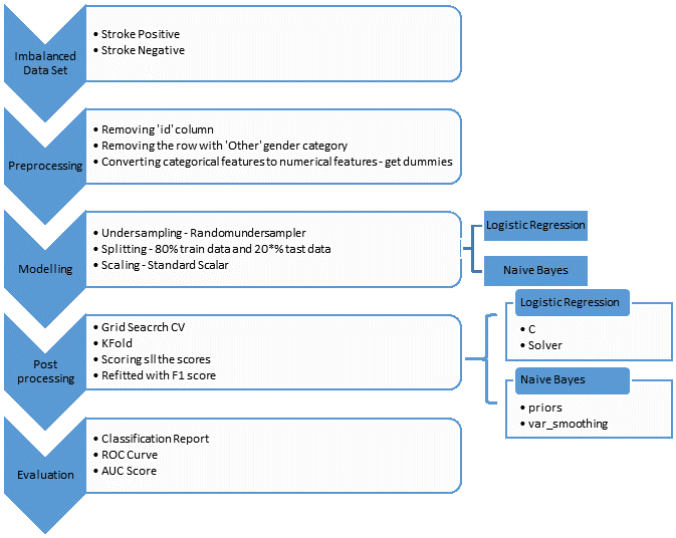

Figure IV.1 The Workflow of the Proposed Methodology

References

[1] "PubMed," 17 January 2022. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/34986727/.

[2] "World Health Organization," 29 October 2022. [Online]. Available: https://www.who.int/srilanka/news/detail/29-10-2022-world-stroke-day-2022.

[3] C. S. P. R. A. Priya Ranganathan, "Common pitfalls in statistical analysis: Logistic regression," pp. 148-151, 2017.

[4] "ResearchGate," [Online]. Available: https://www.researchgate.net/figure/Linear-Logistic-Regression-Sigmoid-function-Applied_fig3_348064698. [Accessed 23 01 2024].