Asian Basic and Applied Research Journal

Volume 6, Issue 1, Page 70-82, 2024; Article no.ABAARJ.1348

Prediction of Patient's Stroke Vulnerability Status Using Logistic Regression Machine Learning Model

Okpe Anthony Okwori [a*], Moses Adah Agana [b], Ofem Ajah Ofem [b] and Obono I. Ofem [b]

[a] *Department of Computer Science, Federal University Wukari, Nigeria.*
[b] *Department of Computer Science, University of Calabar, Nigeria.*

Authors' contributions

This work was carried out in collaboration among all authors. All authors read and approved the final manuscript.

Article Information

Open Peer Review History:
This journal follows the Advanced Open Peer Review policy. Identity of the Reviewers, Editor(s) and additional Reviewers, peer review comments, different versions of the manuscript, comments of the editors, etc are available here: https://prh.globalpresshub.com/review-history/1348

Original Research Article

Received: 15/06/2023
Accepted: 21/08/2023
Published: 20/06/2024

ABSTRACT

In recent time, machine learning has been widely used in healthcare services due to its efficiency in solving health-related problems through accurate prediction of diseases and medical conditions thereby assisting the physicians to diagnose diseases at an early stage. Machine learning models are equally used to handle complex and high-dimensional ever evolving huge amount of medical data to improve the accuracy and efficiency of disease prediction and diagnosis. This paper aims at applying machine learning model for the prediction of stroke vulnerability among individuals. In particular a Logistic Regression (LR) based stroke prediction model was described and developed using phyton programming language for the prediction of likelihood of stroke occurrence. Stroke usually occur due to blockage of blood flow to the brain cell which causes the brain cells to die as a result of lack of oxygen and nutrients. It is a medical emergency that may result in lasting brain damage, permanent disability and mortality across all ages. To reduce stroke occurrence, there is

Corresponding author: Email: okwori@fuwukari.edu.ng;

***Cite as:** Okwori, Okpe Anthony, Moses Adah Agana, Ofem Ajah Ofem, and Obono I. Ofem. 2024. "Prediction of Patient's Stroke Vulnerability Using Logistic Regression Machine Learning Model". Asian Basic and Applied Research Journal 6 (1):70-82. https://jofresearch.com/index.php/ABAARJ/article/view/141.*

an urgent need for stroke prediction and life style changes. The logistic regression-based stroke prediction model was developed in this paper using the healthcare dataset stroke data obtained from Kaggle machine learning dataset repository. The dataset was preprocessed to improve the prediction performance using various dataset preprocessing techniques such as feature selection, feature encoding, missing values correction, class balancing, outlier detection and correction, feature scaling as well as hyperparameter turning. The preprocessed dataset was used for the training, validation and testing of the logistic regression stroke prediction machine learning model and was evaluated using python Scikit-Learn evaluation metrics such as accuracy score, precision score, recall score, f1-score, specificity score as well as area under receiver operating characteristic curve **(**AUC**-**ROC). After successful evaluation, the model produced a classification accuracy of 81% and AUC-ROC of 90%. This shows that logistic regression model is very efficient in stroke classification using the healthcare dataset and the proposed model has shown improvement over some existing stroke prediction model that uses logistic regression.

Keywords: Stroke; learning; supervised; regression; dataset; preprocessing.

1. INTRODUCTION

Stroke is a neurological disease that emanates from the death of brain cells due to oxygen and nutrient deficiency. It is one of the deadliest diseases in recent times that sometimes results in permanent physical disability and decreased quality of life and hence considered as the most important reason for disability. Stroke is the second foremost reason for dementia and third important reason for death [1]. Early detection of stroke condition improves the possibility of preventing its complications and improves health care and overall management of stroke patients [2], hence a need to use a sophisticated technique like machine learning in stroke prediction to reduce stroke occurrence as the basic goal of every efficient prediction model is to determine the patient risk level to enhance a personalized medical decision making that can improve the patient outcome and overall quality of life [3]. In general, machine learning in healthcare (MLH) is targeted at predicting clinical outcomes based on several predictors. Machine learning is vastly used in system security which is needed by every user [4] and healthcare services as it has demonstrated the ability to achieve human-level or even above in clinical decisions making [5] such as prediction, diagnosis and disease prognosis. In healthcare services, machine learning technology provides algorithms capable of self-learning from analysis of external data about patient medical condition thereby providing an increased quality of diagnosis and treatment. In recent time several advancements have been made in the field of MLH which enable machine learning models to efficiently support expert physicians in providing effective and timely treatments to their patients with increased quality and precision. Properly implemented machine learning application may augment the functionalities of human physicians and redefine patient medical care [6]. This paper aims at demonstrating the use of logistic regression machine learning algorithm in stroke prediction. Logistic Regression is a supervised machine learning model used for classifying events based on probability estimation. It is a very popular algorithm for predicting binary outcomes, like the presence or absence of a disease [7], and hence a very useful tool for identifying individuals at high risk of stroke vulnerability. In general, machine learning algorithms need a training dataset of m input/output pairs ($x^{(i)}$,$y^{(i)}$). The superscript in parenthesis represents individual instances of the training set – for stroke prediction, each instance represents individual patient's records in the dataset to be used for the classification.

1.1 Related Works

Logistic regression is actually an effective machine learning classification model as it is used in various binary classifications. Barbosa [8] used logistic regression machine learning classification model to predict heart disease occurrence among individuals. Heart disease dataset obtained from Kaggle website was used. It was preprocessed using various dataset preprocessing techniques. The preprocessed dataset was used for developing the logistic regression model for the heart disease prediction with python programming language and found that the model predicted heart disease with 86% prediction accuracy. In similar vein, a logistic regression model for predicting heart disease was developed by Babatola [9] using cardiovascular disease dataset obtained from Kaggle machine learning dataset repository and recorded a prediction accuracy of 84.9%. In attempt to address the challenges of heart diseases using smart healthcare services, a logistic regression based cardiovascular disease prediction model was developed by Ciu and Oetama [10] using heart disease UCI dataset obtained from Kaggle machine learning dataset repository and found that logistic regression is efficient in binary classification as it achieved 85% prediction accuracy and hence suitable for binary classification. In a sperate research, a logistic regression-based heart disease prediction model was developed by Zhang et al. [11] using a dataset obtained from UCI machine learning dataset repository and found that the model is highly accurate and effective in predicting binary events. Due to the efficiency of logistic regression model in heart disease prediction, researchers extended the application of logistic regression to the prediction of stroke as evident in the work of Eleftherakou [12] where a logistic regression-based stroke prediction model was developed using dataset obtained from Kaggle machine learning repository on Julia programming Language. The dataset was adequately preprocessed and the clean data was used to develop the logistic regression model. It was found that upon evaluation the logistic regression model performs well with 80% prediction accuracy [11] designed experimental research to evaluate the performance of logistic regression and its various variants on the prediction of survival time among lung cancer patients in a retrospective study using a dataset generated from cancer patients and found that the basic logistic regression (LR) and its variant Dual coordinate Descent Logistic Regression technique (DCD-LR) achieved the best results which shows that logistic regression is very suitable for disease classification [12,13] developed a logistic regression based classification model for cardiac disease classification using the University of California Ivain (UCI) dataset and improves the classification efficiency through dataset preprocessing such as missing value identification and correction, feature selection among others. It was found that logistic regression is a very efficient machine learning model for binary disease classification [14] developed a logistic regression and classification tree algorithms for the prediction of type 2 diabetes among Pima Indian women using the Pima Indian dataset obtained from National Institute of Diabetes at the Johns Hopkins University and found that both the logistic regression and the classification tree achieve good prediction performance with the logistic regressing having the highest prediction accuracy [15,16] developed a logistic regression based diagnostic model for Covid-19 using the dataset obtained from 400 patients at Ayatollah Talleghani Hospital, Abadan, Iran. The logistic regression model was implemented using SPSS and the model performance was improved using accurate feature selecting techniques and it was found that the binary logistic regression model was very efficient in diagnosing Covid-19 as evident in its specificity, sensitivity, and accuracy values and hence suitable for binary disease classification.

2. METHODOLOGY

The logistic regression-based stroke prediction model was achieved through the following design steps: General research information, identification of variables, data collection, data preprocessing, splitting of dataset, development of logistic regression model for stroke prediction and logistic regression model evaluation.

2.1 General Research Information

The general research information was obtained from relevant literatures through some published research materials such as journal articles, books, papers, among others that were read to obtain valuable information about stroke prediction using various machine learning models and the strengths of logistic regression in binary classification.

2.2 Data Collection and Variables Identification

The dataset “healthcare-dataset-stroke-data” used in this research was obtained from Kaggle machine learning dataset repository as structured data. The dataset was downloaded and renamed as stroke and converted to comma separated value (CSV) file “stroke.csv” using Microsoft excel for easy data preprocessing using Pandas machine learning library. It originally consisted of 12 columns and 5110 rows. The dataset has 11 independent variables features: id, gender, age, hypertension, heart_disease, ever_married, work_type, residence_type, avg_glucose_level, bmi and smoking_status with one dependent variable class label “stroke”.

The dataset is described in Table 1 as shown.

The top view of healthcare-dataset-stroke-data used in the system is shown in Table 2.

2.3 Data Pre-processing

This is a process of removing data inconsistencies such as out of range values, missing values, noise, unformatted data among others from the dataset. The raw healthcare-dataset-stroke-data used in this work needs to be properly cleaned before it can be used for training, testing and validation of the stroke perdition models for optimal performance. The various techniques used in preprocessing the dataset is discussed below.

i) Feature Selection: Feature selection is a process of choosing the optimal subset of dataset features in order to reduce model complexity, enhance the computational efficiency of the models, and reduce generalization error introduced due to noise by irrelevant features, reduce the entire processing cost and enhance the overall model performance. The healthcare-dataset-stroke-data used in this work contains some less relevant features such as id that was eliminated using correlation technique as the features are simply made up of metric variables. In the healthcare-dataset-stroke-data, the feature Id has correlation coefficient of 0 which shows that it does not correlate with the target variable and hence eliminated from the dataset to reduce the dataset dimension from 12 to 11.

ii) Feature Encoding: Conversion of categorical (string or nominal) data to numeric data was done using LabelEncoder technique. In this technique, every category in the feature is assigned a value from 1 to N where N represents the number of categories in that feature. The healthcare-dataset-stroke-data consists of categorical features such as gender, ever_married, work_type, Residence_type and smoking_status that were encoded using python scikit-learn label Encoder as shown in Fig. 1.

iii) Missing Values: The healthcare-dataset-stroke-data contains missing values in the smoking_status and bmi variables. The missing values in the bmi variable were filled using the computed mean of the available data within the bmi variable field while all data associated with missing values in the smoking_status were simply removed from the dataset.

iv) Class balancing: Most classification applications are highly inefficient due to problem of class imbalance in dataset used for training the classification algorithm. The healthcare-dataset-stroke-data used in the system was an imbalance dataset with two main target classes: vulnerable to stroke and not vulnerable to stroke target classes. The stroke vulnerable target class has 4.87% of the total dataset population while the not stroke vulnerable target class has 95.13% of the total population of the healthcare-dataset-stroke-data as shown in Fig. 2.

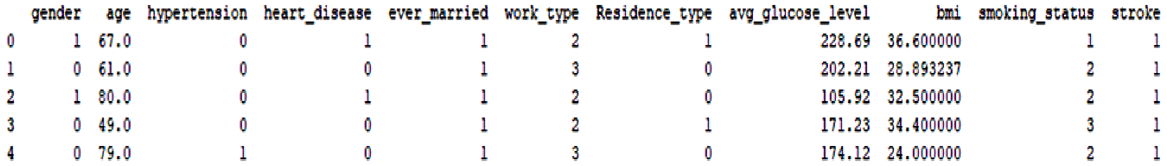
In order to balance the ratio of occurrence of the target classes, Synthetic Minority Oversampling Technique (SMOTE) was adequately employed in the system. The SMOTE method generates synthetic data samples for the minority class by picking a point randomly from the minority class and calculating its k-nearest neighbors and fixing the synthetic point in between them. The SMOTE prevents overfitting that is normally associated with most other random oversampling techniques thereby obtaining a good resampled dataset capable of giving reliable predictions. The balanced dataset generated through the synthetic minority oversampling technique consists of 50% stroke vulnerable instances and 50% no stroke instances as shown in Fig. 3.

Table 1. Dataset description

Feature No.	Feature Name	Feature Description
1	Id	Unique identification number for each data point in the dataset
2	Gender	Male or Female
3	Age	Number of years of a patient
4	Hypertension	Presence or absence of hypertension
5	Heart_disease	Presence or absence of heart disease
6	ever_married	Married or not marredS
7	work_type	Children, Private, Never worked, Govt. job or Self employed
8	Residence_type	Urban or rural residence
9	avg_glucose_level	Average quantity of glucose in the patient body
10	Bmi	The body mass index of the patient
11	smoking_status	Never smoked, formally smoked or smokes
12	Stroke	Presence or absence of stroke

Table 2. Top view of healthcare-dataset-stroke-data

Id	Gender	Age	Hypertension	Heart_Disease	Ever_Married	Work_Type	Residence_Type	Avg_Glucose_Level	BMI	Smoking_ Status	Stroke
9046	Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
51676	Female	61	0	0	Yes	Self-employed	Rural	202.21	N/A	never smoked	1
31112	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
60182	Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
1665	Female	79	1	0	Yes	Self-employed	Rural	174.12	24	never smoked	1
56669	Male	81	0	0	Yes	Private	Urban	186.21	29	formerly smoked	1
53882	Male	74	1	1	Yes	Private	Rural	70.09	27.4	never smoked	1
10434	Female	69	0	0	No	Private	Urban	94.39	22.8	never smoked	1
27419	Female	59	0	0	Yes	Private	Rural	76.15	N/A	Unknown	1
60491	Female	78	0	0	Yes	Private	Urban	58.57	24.2	Unknown	1
12109	Female	81	1	0	Yes	Private	Rural	80.43	29.7	never smoked	1
12095	Female	61	0	1	Yes	Govt_job	Rural	120.46	36.8	smokes	1
12175	Female	54	0	0	Yes	Private	Urban	104.51	27.3	smokes	1
8213	Male	78	0	1	Yes	Private	Urban	219.84	N/A	Unknown	1



```
   gender   age  hypertension  heart_disease  ever_married  work_type  Residence_type  avg_glucose_level        bmi  smoking_status  stroke
0       1  67.0             0              1             1          2               1             228.69  36.600000               1       1
1       0  61.0             0              0             1          3               0             202.21  28.893237               2       1
2       1  80.0             0              1             1          2               0             105.92  32.500000               2       1
3       0  49.0             0              0             1          2               1             171.23  34.400000               3       1
4       0  79.0             1              0             1          3               0             174.12  24.000000               2       1
```

Fig. 1. Top view of the encoded healthcare-dataset-stroke_data

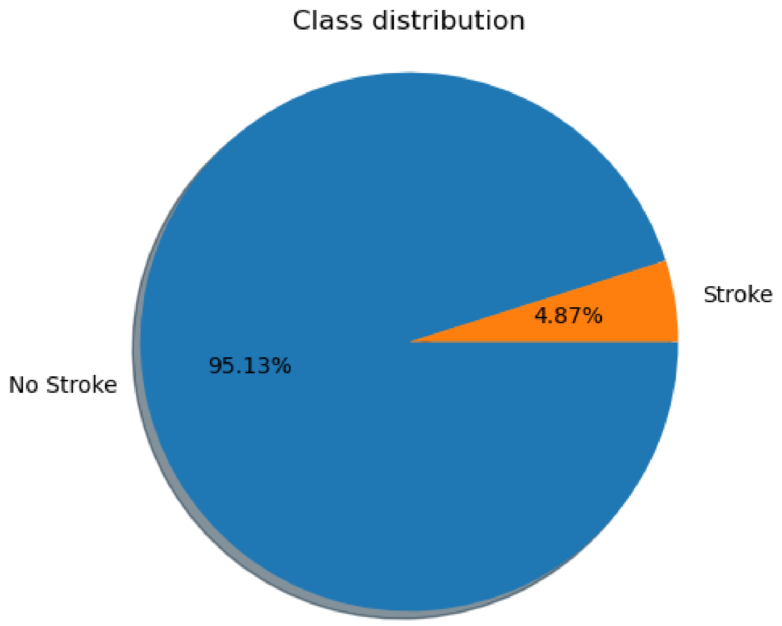


Fig. 2. Imbalance target class distribution of healthcare-dataset-stroke-data

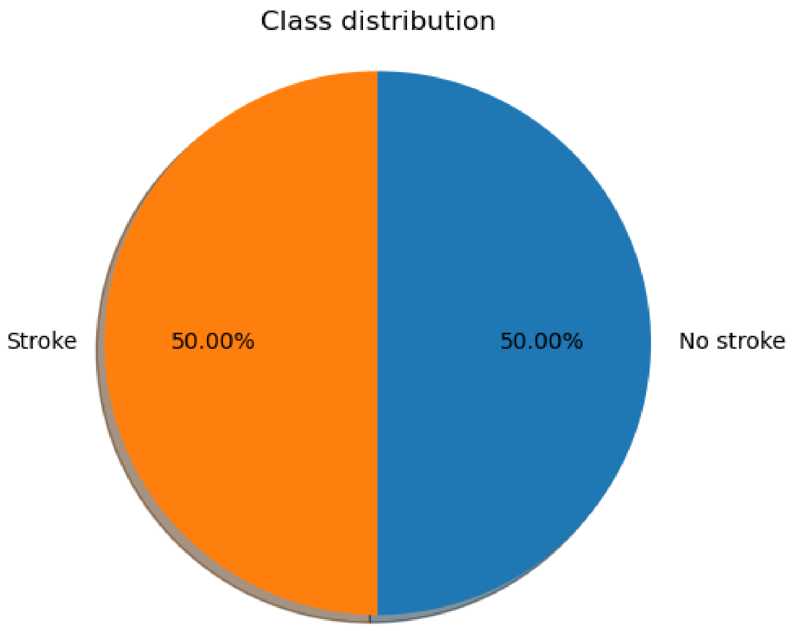


Fig. 3. Balanced target class distribution of healthcare-dataset-stroke-data

v) Outliers: An outlier in dataset is simply a data point that is apparently different from the other data points in that dataset feature. The healthcare-dataset-stroke-data used in this work contains some outliers in both the avg_glucose_level and bmi features that must be remove. All data points in these features that were greater than four (4) times the feature's standard deviation (STD) were considered as outliers and were replaced by the mean value of the features to prevent it from skewing the training process of the algorithms and to ensure shorter training time that leads to very accurate models that produces optimal result.

vi) Feature Scaling: This is a technique of controlling the range of values of the independent variable in the dataset. Gradient Descent Based Algorithms like logistic regression require feature scaling before training in order to speed up the training process, increase model numerical stability and to ensure that all features contribute equally to the prediction model. In this work, the z-normalization (standardization) method was used to normalize the independent features of the healthcare-dataset-stroke-data. In the standardization technique the feature values are centered around the mean with a unit standard deviation.

vii) Hyperparameters turning: Hyperparameter of a machine learning model simply refers to those external configurations of a model whose values are not obtainable from the data used in building the model. In this work the various hyperparameters were optimally tuned to ensure high efficiency of the predictive capability of the resulting models. The logistic regression model was configd to use (LBFGS) solver that is very suitable for optimization problem due to its limited memory requirement.

2.4 Split Dataset into Training, Validation and Testing Sets

After the entire data preprocessing exercise, the preprocessed stroke.csv had 9722 rows with 11 features that were split into training dataset, validation dataset and testing dataset respectively using sklearn python library. Out of the 9722 data records, 80% (7776) data items were used for training, 10% (973) data items were used for validation and 10% (973) data items were used for testing of the models respectively.

2.5 Design of the Logistic Regression Machine Learning Model for Stroke Prediction

Logistic regression is a supervised machine learning model used for classifying events based on probability estimation. In general, machine learning algorithms need a training dataset of m input/output pairs ($x^{(i)}$,$y^{(i)}$), where the superscript in parenthesis represents individual instances of the training set – for stroke prediction, each instance represents individual patient record in the dataset used for the classification. The inputs from the dataset were represented in the form of feature vectors. For each input (a patient record) in the dataset $x^{(i)}$, there exists a vector set of features [x_1; x_2; …; x_n] corresponding to the values of the individual stroke risk factors of the patient. The feature i for input $x^{(j)}$ is referred to as $x_i^{(j)}$ and is simplified in this research as xi. The basic target of this binary logistic regression is to train a stroke predictor capable of making a binary decision about a set of new input patient records. To facilitate efficient decision making, a sigmoid classifier was used. To achieve the classification using sigmoid classification function (logistic function), we consider a single patient input record of stroke vulnerability x, which is represented by the feature vector [x_1; x_2; … ; x_n]. The output of the classifier y can either be 1 (which means the patient is vulnerable to stroke) or 0 (which means the patient is not vulnerable to stroke). To show that a patient is vulnerable to stroke (in which case we say the decision is "positive to stroke"), we computed the probability $P(y = 1|x)$ and to show that a patient is not vulnerable to stroke (in which case we say the decision is "negative to stroke"), we computed the $P(y = 0|x)$. To achieve this computation, the logistic regression learns a vector of weights and a bias term from the training set. Every weight w_i is a real number, which is assigned to one of the input features x_i. The weight w_i shows the importance of that particular input feature to the entire classification decision. These weights could either be positive or negative depending on whether the feature shows stroke vulnerability or not. In a similar vein, the bias term (intercept) just like the weight is also a real number that is added to the weighted inputs. Whenever a new data (test data) is fed in for prediction decision after learning the weights during training, the prediction model first multiplies the features in each patient record x_i by the weight of the feature w_i, computes the sums of the weighted features, and finally adds the bias term b to the computed sum to produce a single value result z that represents the weighted sum of the evidence for the class (i.e. vulnerable or not vulnerable to stroke) as shown in equation (1).

$$z = \left(\sum_{i=1}^{n} w_i x_i\right) + b \quad (1)$$

Where n = number of features in the patient record that is used for the prediction decision.

The equation (1) above can be represented using the dot product notion of the two vectors: weights (w_i) of the feature and the value of the features (x_i). The dot product of two vectors a and b can be defined as the sum of the products of the corresponding elements of each vector and can be expressed using dot product notation as a.b. Therefore, equation one can be rewritten as:

$$z = w.x + b \quad (2)$$

since the vector weights (w) are real numbers, the value of z in (2) lies between -∞ to +∞ and hence does not correspond to the expected probability function whose values should lie between 0 and 1 only. To enable the value of z to lie between 0 and 1 so as to conform with probability function, z is passed through a sigmoid (logistic) function as shown in equation (3).

$$y = \sigma(z) = \frac{1}{1+e^{-z}} = \frac{1}{1+\exp(-z)} \quad (3)$$

The sigmoid (logistic) function in equation (3) maps a real valued number into a range of [0,1] which is highly suitable for expressing probability. Consequently, the result of appling the sigmoid function to the sum of the weighted features is simply a value between 0 and 1. This resulted values are considered as a probability output if the sum of p(y = 1) and p(y = 0) is equal to 1. Therefore:

$$\begin{aligned} P(y=1) &= \sigma(w \cdot x + b) \\ &= \frac{1}{1+\exp(-(w \cdot x + b))} \\ P(y=0) &= 1 - \sigma(w \cdot x + b) \\ &= 1 - \frac{1}{1+\exp(-(w \cdot x + b))} \\ &= \frac{\exp(-(w \cdot x + b))}{1+\exp(-(w \cdot x + b))} \end{aligned} \quad (4)$$

One of the basic properties of sigmiod function is:

$$1 - \sigma(x) = \sigma(-x) \quad (5)$$

Therefore, from equation (4) and (5) above p(y = 0) could be expressed as:

$$\sigma(-(w \cdot x + b)).$$

Therefore, given an instance x, the probability p(y=1|x) can be computed from the above algorithm. For us to make a decision about a test instance x, we simply say yes if the p(y = 1|x) is greater than 0.5 and say no if it is less than or equal to 0.5. The value 0.5 is called a decision boundary.
i.e.

$$\hat{y} = \begin{cases} 1 & \text{if } P(y=1|x) > 0.5 \\ 0 & \text{otherwise} \end{cases}$$

To achieve the prediction using the healthcare_dataset_stroke_data in accordance to the above functions the following algorithmic steps were taking.

Algorithm 1: logistic regression prediction algorithm

Step 1: Import the relevant libraries
Step 2: Import the dataset
Step 3: Read the dataset
Step 4: Pre-process the dataset
Step 5: split the dataset
Step 6: Create the logistic regression model object
Step 7: Fit the model
Step 8: Predict result
Step 9: Evaluate the model

The efficient representation of the above algorithmic steps is shown diagrammatically in the proposed system architecture of Fig. 4.

2.6 Logistic Regression Training

To generate the logistic regression model, a logistic regression algorithm was imported from the python sklearn library and the various hyperparameters were tuned. To train the model, the training dataset was fit into the Logistic Regression model and evaluated using both stratified 10-fold cross-validation method and confusion matrix. Confusion Matrix is a tabular conception of the predicted output of the model and the true output of the tested sample instances. The confusion matrix for a given machine learning model can only be generated if the true values for the test data are known beforehand. It takes the form of n x n matrix where n represents the number of target output classes. Every row of a confusion matrix depicts a particular instance of prediction (i.e. positive prediction or negative prediction) while every column depicts a particular instance in the actual class of the tested sample data. In this logistic regression based stroke prediction model which is a binary classification problem, the confusion matrix generated is a 2x2 matrix of two types of correctly predicted values (TP and TN) as well as two types of wrongly predicted values (FP and FN).

Where:

True positives (TP): Predicted positives and are actually positive.
False positives (FP): Predicted positives and are actually negative.
True negatives (TN): Predicted negatives and are actually negative.
False negatives (FN): Predicted negatives and are actually positive.

Fig. 5 shows the general representation of a confusion matrix.

The true positive (TP), false positive (FP), true negative (TN) and false negative (FN) values obtained from the confusion matrix are used in this paper to compute the values of the following logistic regression model evaluation metrics.

i) Accuracy score: The classification accuracy of a machine learning model defines its ratio of the total number of correct predictions to the total number of input instances predicted, i.e. the total number of correct predictions divided by all the predictions. It usually demonstrates how frequent the classification model produces the correct result. It is given by the formula:

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{6}$$

ii) Precision Score: The precision score is the ratio of all correctly classified positive instances to the total number of positive predicted instances. i.e., it determines the actual number of positive instances from the total instances classified as positive. It computes the classifier's ability in classifying positive instances appropriately. It is the true positives value divided by the total number of predicted positive values. It is given by the formula:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{7}$$

iii) Recall score: This matrix determines the number of actual positive instances that were correctly classified as positive with the machine learning model. It shows how often the logistic regression model was able to predict actual positive values correctly. A good recall value is very vital in disease prediction because it is better to raise a false alarm than to allow a stroke vulnerable patient to go undetected. The recall score is sometimes referred to as Sensitivity score or True Positive Rate which is expressed as the ratio of all positive data points that are correctly classified as positive (TP) to all positive data points in the entire input instances (FN and TP). It is given by the formula:

$$\text{Sensitivity} = \text{Recall} = \frac{\text{TP}}{\text{FN} + \text{TP}} \tag{8}$$

iv) F1- score: This matrix determines the harmonic mean (weighted average) of the Precision and Recall values generated by the logistic regression machine learning model.

$$\text{F} - \text{Measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}$$

v) Specificity score: Specificity score shows the ratio of the number of actual negative target classes that were correctly classified as negative by the classifier to the total number of actual negative target classes in the dataset. It therefore determines how well the classifier is able to detect actual negative cases from all the available negative instances. A low value of specificity shows that more actual negative target classes were classified positive, i.e. there is an increase in the value of false positive. Specificity is given by the formula:

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}} \tag{10}$$

vi) Area Under Receiver Operating Characteristic Curve (AUC-ROC): This is a graphical method of summarizing the performance of the logistic regression machine learning classification model over all possible thresholds. In this graph, the True Positive Rate is plotted on the (y-axis) against the False Positive Rate plotted on the (x-axis) and the threshold for assigning observations to a given class is varied accordingly. This curve determines the number of correct positive classifications that can be gained with increase in the number of false positive classifications. The receiver operating characteristic (ROC) curve is not a single value but entire graph hence to be able to compare the ROC curve of a given classifier with other classifiers, the area under the curves (AUC) are computed.

3. RESULTS AND DISCUSSION

The logistic regression-based stroke prediction model was evaluated using confusion matrix and its related matrices such as Accuracy score, Precision score, Recall score, F1 score, Sensitivity score, Specificity score and the Area Under Cure (AUC) scores respectively. Table 4 below shows the confusion matrix obtained after the successful testing of the logistic regression stroke prediction model with the test dataset.

From Table 3, a total of 973 data instances were used as test sample dataset and out of the 973 data instances used for the prediction, 389 are positive instances that were actually classified as positive (true positive), 403 are negative instances that were actually classified as negative (true negative), 98 are actually negative instances that were classified as positives (false positive) while 83 are actually positive instances that were classified as negatives (false negative) respectively. The various predictions obtained from the confusion matrix were used to compute the values of the associated metrics as shown in Table 4 and Fig. 6.

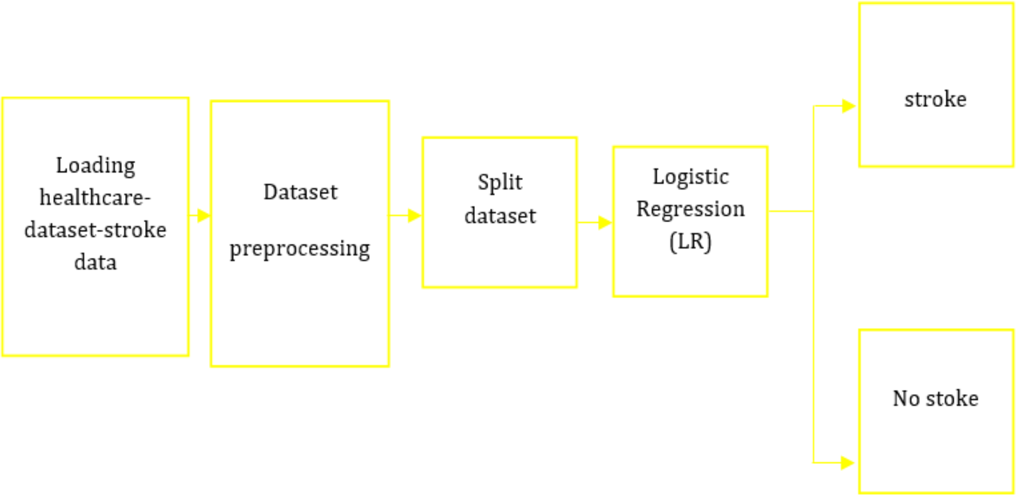
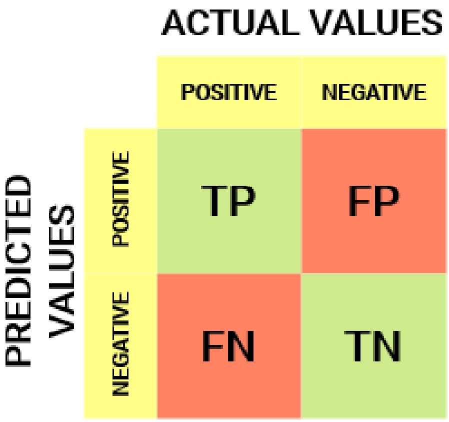


Fig. 4. Architectural model of the proposed system



		ACTUAL VALUES	
		POSITIVE	NEGATIVE
PREDICTED VALUES	POSITIVE	TP	FP
	NEGATIVE	FN	TN

Fig. 5. Confusion matrix

Table 3. Confusion matrix for logistic regression

N = 973		**Actual values**	
		Positive (YES)	**Negative (NO)**
Predicted values	Positive (YES)	TP = 389	FP = 98
	Negative (NO)	FN = 83	TN = 403

Table 4. Evaluated results of logistic regression stroke prediction model

Matrix	**Accuracy**	**Precision**	**Recall**	**F1 score**	**Sensitivity**	**Specificity**	**AUC**
Value	0.81	0.80	0.83	0.82	0.83	0.80	0.90

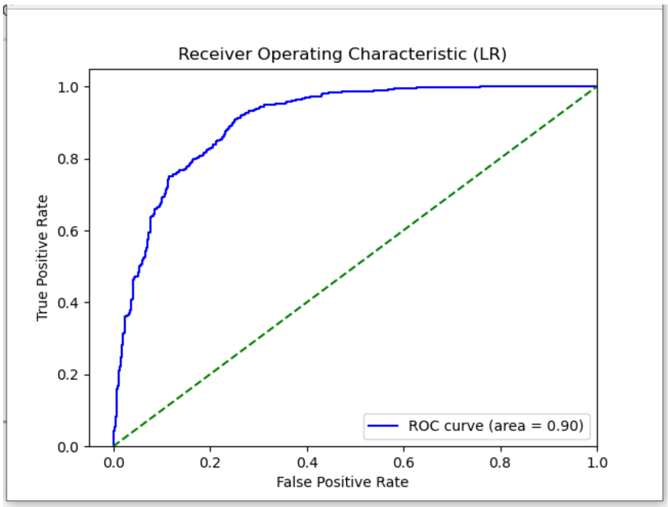


Fig. 6. Area under receiver operating characteristic (ROC) cure (AUC) for logistic regression

Table 5. Summary of comparative analysis of this study with related literatures

S/NO	**Author(S)**	**Title**	**Accuracy of result**
1	Barbosa (2020)	Prediction Model of Heart Disease with Logistic Regression	86%
2	Babatola (2020)	Heart Disease Prediction: A Logistic regression implementation from python scikit-learn	84.9%.
3	Ciu and Oetama (2020)	Logistic Regression Prediction Model for Cardiovascular Disease	85%
4	Eleftherakou (2022)	Stroke prediction: Logistic Regression with Julia	80%
5	Proposed work	Stroke prediction using logistic regression machine learning model	81%

From Table 4 above it can be seen that logistic regression has predicted accuracy of 81%, precision of 80%, recall of 83%, F1 score of 82%, sensitivity of 83%, specificity of 80% and AUC of 90%.

The AUC of 90% in Fig. 6 shows that the algorithm is ninety percent efficient in predicting stroke vulnerability using the healthcare dataset stroke data.

3.1 Comparative Analysis of the Research with Existing Works

The results obtained from this empirical research was compared with some results of related works in available literatures as shown in Table 5.

Table 5 above shows that logistic regression is a good choice of algorithm for binary classification as it achieves classification accuracy within the range of 80% to 86% in all its applications reviewed in this work. The difference in the model performance may be largely dependent on the dataset, preprocessing technique as well as the choice of programming language for the implementation. In particular, proposed system outperforms it closest related work in [9] where the logistic regression model was used to predict stroke with 80% prediction accuracy while the proposed system has 81% stroke prediction accuracy. This slight improvement may be as a result of high efficiency of python programming language in building machine learning model and the preprocessing technique used.

4. CONCLUSION

This research demonstrated the relevance of machine learning for efficient decision making in the field of health care services. Stroke prediction is one of the major techniques of reducing stroke occurrence and severity as it leads to early preventive measures. In this paper, the efficacy of logistic regression classification model was investigated in predicting stroke vulnerability. Experiment was conducted using the dataset obtained from Kaggle machine learning dataset repository on python programming language and the prediction results were evaluated using confusion matrix and its associated metrics. The logistic regression algorithms performed very well in predicting stroke vulnerability among individuals using the healthcare dataset stroke data as shown in its AUC values of 90%. However other machine learning algorithms may be implemented either as single or hybrid model on various standard dataset to further improve stroke prediction.

DISCLAIMER (ARTIFICIAL INTELLIGENCE)

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc) and text-to-image generators have been used during writing or editing of manuscripts.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Maqbool M, Toor UUR, Nahra SF. Stroke, a foremost cause for disability and functional impairment, Indo American Journal of Pharmaceutical Sciences. 2019; 6(3):5403-5409.
2. Ohoud A, Riyad A. Prediction of stroke using data mining classification techniques. International Journal of Advanced Computer Science and Applications. 2018;9:457-460.
3. Maren E, Shipe ME, Stephen AD, Farhood F, Eric LG. Developing prediction models for clinical use using logistic regression: An overview. Journal of Thoracic Disease. 2019;11(4):576-584.
4. Ogbu HN, Agana MA. Intranet Security Using a LAN Packet Sniffer to Monitor Traffic. *In* Natarajan M. (*Eds*) CCSIT, NCWMC, DaKM. 2019;9(8):57-68.
5. Mateen BA, Liley J, Denniston AK, Holmes CC, Vollmer SJ. Improving the quality of machine learning in health applications and clinical research. Nature Machine Intelligence. 2020;2:554-556
6. Habehh H, Gohel S. Machine Learning in Healthcare, Current Genomics. 2021; 22: 291-300.
7. Mohammed GM. Detection and analysis of diabetes by using logistic regression (LR). International Research Journal of Modernization in Engineering Technology and Science. 2023;5(1):609-613.
8. Barbosa C. Prediction Model of Heart Disease With Logistic Regression; 2020. Available:https://medium.com/analytics-vidhya/prediction-model-of-heart-disease-with-logistic-regression-62a461e5474, accessed on 18th January, 2023
9. Babatola TB. Heart disease prediction : A logistzic regression implementation from python scikit-learn; 2020. Available:https://bimie.medium.com/heart-disease-prediction-a-logistic-regression-implementation-from-python-scikit-learn-c4eb391a873f, accessed on 18th January, 2023
10. Ciu T, Oetama RS. Logistic regression prediction model for cardiovascular disease, International Journal of New Media Technology. 2020;VII(1):33-38.
11. Zhang Y, Diao L, Ma L. Logistic regression models in predicting heart disease. Journal of Physics: Conference Series. 2021;1-5.
12. Eleftherakou O. Stroke prediction: Logistic Regression with Julia; 2022. Available:https://medium.com/mlearning-ai/stroke-prediction-logistic-regression-with-julia-523f90eb5ae, accessed on 15th December, 2022

13. Shayesteh SP, Shiri I, Karami AH, Hashemian R, Kooranifar S, Ghaznavi H, Shakeri-Zadeh A. Predicting lung cancer patients' survival time via logistic regression based models in a quantitative radiomic framework. Journal of Biomed Phys Eng. 2020;10(4):479-492.
14. Ambrish G, Ganesh B, Anitha G, Chetana S, Kiran M. Logistic regression technique for prediction of cardiovascular disease. Journal of Global Transitions Proceedings. 2022;3:127–130
15. Ram DJ, Chandra KD. Predicting type 2 diabetes using logistic regression and machine learning approaches. International Journal of Environmental Research and Public Health. 2021; 1-17.
16. Nopour R, Shanbehzadeh M, Kazemi-Arpanahi H. Using logistic regression to develop a diagnostic model for COVID-19: A single-center study, Journal of Education and Health Promotion. 2022;11:1-6.

© Copyright (2024): Author(s). The licensee is the journal publisher. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Peer-review history:
The peer review history for this paper can be accessed here:
https://prh.globalpresshub.com/review-history/1348