International Journal of Innovative Research in Computerand Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.194 |

||Volume 8, Issue 4, April 2020||

Stroke Prediction using Logistic Regression Algorithm

Parag B. Shah[1], Dr. B.H. Chandrashekar[2]

Student, Department of MCA, RV College of Engineering, Bengaluru, Karnataka, India[1]

Associate Professor, Department of MCA, RV College of Engineering, Bengaluru, Karnataka, India[2]

ABSTRACT: In clinical perspectives, the prediction of the event of heart disease is significant to determine not if the patient has a coronary illness. Specific observable research methodologies and data mining methodologies are now being used daily to enhance the accuracy of disease analysis in the medication sector. One of the most well-known problems of clinical focus is that all specialists do not have equal knowledge and experience to treat their patients, they offer their own choice that can result in poor results and lead to death for the patients. Stroke is major cause of death and one of the main drivers of the planet's severe long haul deficiencies. In this paper, a study on Logistic Regression is carried out on the Healthcare Dataset for stroke expectations. Performance metrics were computed to evaluate the assessment model.

KEYWORDS: Stroke Prediction, Logistic Regression

I. INTRODUCTION

Stroke is that the most generally recognized precarious neurological issue in huge numbers of the developing countries. Opportune forecast of stroke is before a test and is especially noteworthy within the field of biomedical research. Examinations show that different risk factors convey important data about the event of Stroke. Way of life leads to an increase in the chance of stroke with components that incorporate eating routine, smoking, weight, physical inactivity, cigarette and liquor utilization, heredity factors, age, sex, and medicate use [1].

Early detection of stroke problems is important for prevention. Machine Learning helps to mine and predicting stroke. Support Vector Machine, Logistic Regression, Random Forest Classifier, and AI is a kind of computerized thinking designed to render a computers with the ability to think like a human. AI's purpose allows computers to make a particular undertaking without using guidance, based on examples and impedance [2].

Numerous systematic experiments and AI techniques be used to accomplish stroke discovery. In addition, the discernible connection with a risk factor can be attempted with the result utilizing different models of regression. Composing uncovers that in the course of recent decade's strategic regression has been conducted in expected sickness regions.

II. RELATED WORK

In [3] authorsexamines the utilizes of Mapreduce calculation with the aid of contrasting meta-heuristic methods alongside organized persevering fuzzy neural gadgets on UCI AI to keep the dataset for awaiting coronary illness. The consequences of this examination carried out 98.12 % precision for the 45 events of testing set, while contrasted and meta-heuristic technique alongside the organized neural machine and everyday fuzzy faux neural gadget. The yield exactness of proposed Mapreduce calculation is unmistakably better, due to the dynamic diagram and immediately scaling. Here Hbase is utilized for putting away resultant records.

The Azam et al. [4] The paper depicts programmed determination of coronary vein disease (CAD) patients using enhanced SVM, optimized in these SVM parameters to increase forecast accuracy, which gives a 99.2 percent accuracy using cross-approval k-overlap. The paper serves at the beginning of a period to determine illness and to decrease expenses. The acquired accuracy is a great idea for foreseeing whether or not the individual has coronary disease.

The paper [5] centers on Bone Tumor Diagnosis Using a Naïve Bayesian Segment Model and Radiographic Featuresthe Curtis Langlotz et al. The paper is based on two techniques of Naïve Bayes precision, one is an integral model of Naïve Bayes and the other is differential accuracy of Naïve Bayes Using Naïve Bayes Primary accuracy 62% accuracy is achieved and using Naïve Bayes Differential accuracy 80% accuracy is achieved. The paper's impediment is that the precision isn't enough for a superior option.

Min et al. [6] concluded a model scenario using regressionwith modifiable risk factors for the conclusion of the stroke.In this study, a benchmark sample of 500 patients had no history of strokeand 367 patients with stroke were

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.194 |

||Volume 8, Issue 4, April 2020||

considered from Korea's National Health Insurance Company Social Insurance database.Risk models for male and femaleclassification were not developed separately.After regression analysis, 6 indicator factors were selected andthe model gave a very low accuracy of 64.70 %.

The Rajendra Acharya et al.[7] Explain PC helped to evaluate the diabetic topic by using pulse fluctuation signals using a discrete wavelet shift technique using different classifiers like Decision Tree (DT), K-Nearest Neighbor (KNN), Naïve Bayes (NB), and Support Vector Machine (SVM). The standard acquired accuracy is 92.02 % by using DT for cross-approval within 10 times. Registered accuracy is important for forecasting, although varying is not sufficient.

Additionally, Cheng et al. [8] took a shot in predicting ischemic stroke by using two Artificial Neural Network (ANN) models on the Sugam Multispecialty Hospital dataset, Kumbakonam located in Tamil Nadu, India. Moreover, the specialists believed that the accuracy levels were 79.20% and 95.10%.

The Cemil et al. [9] Suggest the use of information-finding processes on the expectations of stroke patients based on the ANN and SVM, which provided 81.82% and 80.38% accuracy for ANN and 85.9% and 84.26% for ANN and SVMrespectively for the preparation of information index. ANN tends to be a more reliable result or proposed research than the SVM. The precision the paper acquires isn't enough to anticipate patients with stroke.

The Ashok Kumar Dwivedi. [10] The paper aims to evaluate the efficiency of various machine learning methods for heart disease predictions using ten-fold cross-validation. The paper uses various algorithms such as Naïve Bayes, Classification Tree, KNN, Logistic Regression, SVM, and ANN, giving 83%, 77%, 80%, 85%, 82%, 82% and 84% respectively. Compared to other algorithms the Logistic Regression provides better accuracy.

Table 1. Algorithm Comparisons used in Related Work

Year	Authors	Techniques Used	Accuracy
2019	CM Wu, M Badshah, V Bhagwat [3]	Mapreduce, Hbase	98.12%
2017	Azam et al. [4]	Optimized SVM	99.2%
2017	Curtis Langlotz etal.[5]	Naïve Bayesian	80%
2017	Min et al. [6]	Regression Analysis	64.70%
2015	Rajendra Acharya et al.[7]	Decision Tree	92.02% obtained by DT
		KNN	
		Naïve Bayes (NB)	
		SVM	
2014	Cheng et al. [8]	ANN	79.20% and 95.10%
2015	Cemil et al. [9]	ANN	Training Dataset: 81.82%
			Test Dataset: 85.9%
		SVM	Training Dataset: 80.38%
			Test Dataset: 84.26%
2018	Ashok Kumar Dwivedi. [10]	Naïve Bayes	83%
		Classification Tree	77%
		KNN	80%
		Logistic Regression	85%
		SVM	82%
		ANN	84%

III. DATASET

The dataset[11] consists of independent variables like gender, age, hypertension, avg_glucose_level, BMI, Heart Diseases, Ever_Married, Residence_Type,Smoking_Statusand a class label which represents 0 denoting no stroke and 1 or the presence of stroke. This dataset is used to make the modellearn and test the dataset to foresee whether a person may suffer stroke or not.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.194 |

||Volume 8, Issue 4, April 2020||

Table 2. Attributes and Description of the Stroke Dataset

Variable	Description
Id	Patient ID
Age	Age
Gender	Patient's Gender
Hypertension	0 - No Hypertension, 1 - Hypertension
Heart Disease	0 -No Heart Problem 1 – Heart Problem
Ever Married	Yes / No
Work Type	Type of occupation
Residence Type	Rural / Urban
Average Glucose Level	Average glucose level
BMI	Patient's Body Mass Index
Smoking Status	Patient's Smoking Status

IV. PROPOSED METHODOLOGY

A. *Data Preprocessing*

A few information processing strategies can be implemented to enhance the accuracy of the information and to reduce the preparation time. Through the investigation, unimportant qualities such as number, name, and address will be discarded which are inconsistent and have missing details. The standardization of knowledge standardizes all numerical features in the dataset given.

B. *Principal Component Analysis:*

Principal Component Analysis Algorithm is utilized for lessening the measurements and it decides the traits including more towards the forecast of stroke infection. The assessment metric will be a score from AUC-ROC.

C. *Logistic Regression Algorithm:*

Logistic Regression is utilized to foresee whether a patient can have stroke or not. The relationship between independent variable and dependent output is defined by a logistic regression function given by eq. (1)

$$\Pi(X) = \frac{e^{\beta 0+\beta 1X1+\beta 2X2+...+\beta pXp}}{1+e^{\beta 0+\beta 1X1+\beta 2X2+...+\beta pXp}} \quad (1)$$

Where X1, X2...Xp are the probabilities of outcomes.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.194 |

||Volume 8, Issue 4, April 2020||

D. *Computing Accuracy and Evaluating Classification:*

Confusion Matrix helps to assess the evaluation of models. Confusion Matrix describes the performance measurements of the logistic regression in terms of Accuracy, Precision [12], F Measure and Recall [13].

$$\text{Accuracy} = (TP+TN) / (TP+TN+FN+FP) \quad (2)$$

$$\text{Precision} = TP/ (TP+FP) \quad (3)$$

$$\text{F Measure} = (2*TP) / (2TP+FP+FN) \quad (4)$$

$$\text{Recall} = TP/ (TP+FN) \quad (5)$$

(TP –True Positive, TN – True Negative, FP - False Positive, FN- False Negative)

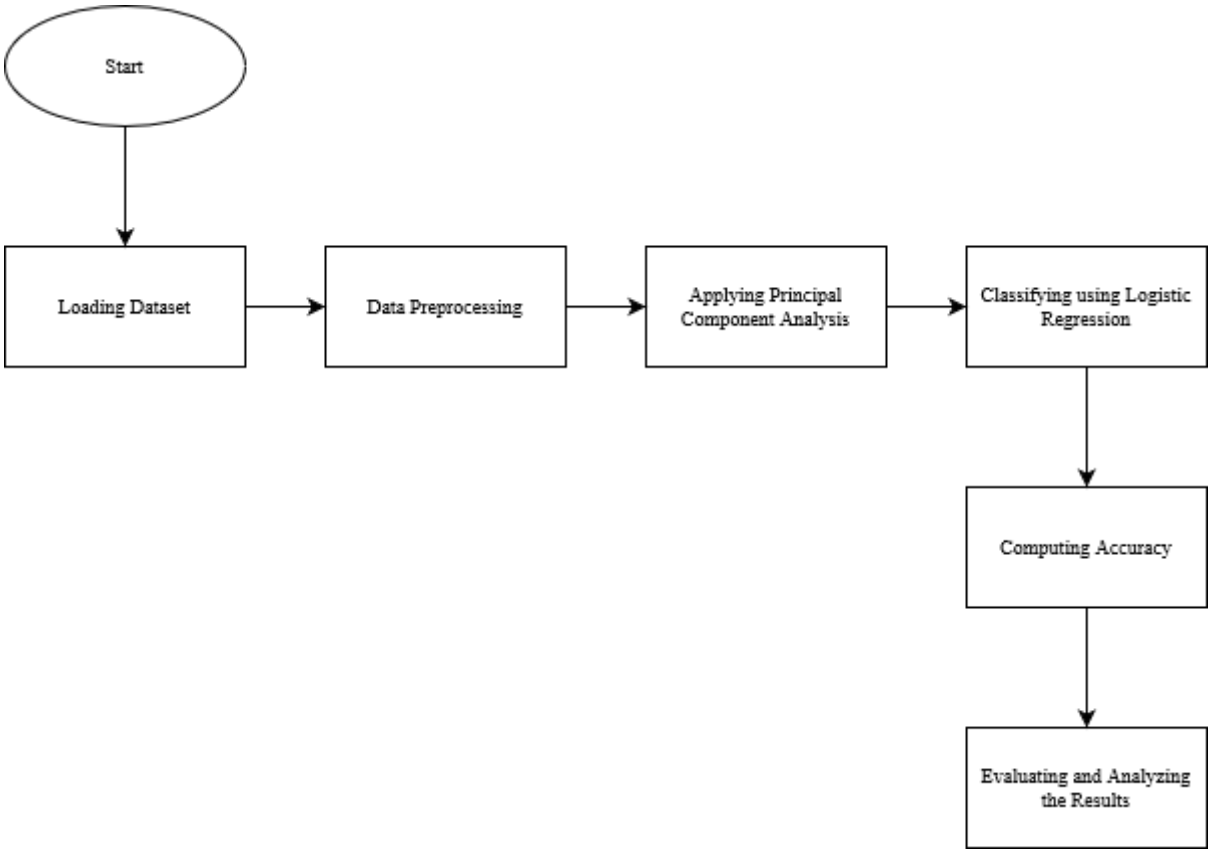

Fig. 1. Flow Chart of Proposed Methodology

V. EXPERIMENTAL RESULTS

Table 3 displays the metrics of logistic regression for respective classes.For class 0, precision had the highest percentage at 75 percent, while recall had the lowest amount at 72 percent.For class 0, f1-score received the maximum ranking of 76 percent.

Table 3. Logistic Regression Results

Class	Precision	Recall	F1 Score	Support
0	0.75	0.72	0.73	5930
1	0.73	0.78	0.76	5858

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.194 |

||Volume 8, Issue 4, April 2020||

VI. CONCLUSION AND FUTURE WORK

The main aim of this work was to analyze the elements that lead to updating the Stroke threat altogether. In this study, Logistic Regression helped to evaluate the significance of indicator factors and construct a model for evaluating stroke risk. The new system of clinical assessment can help doctors make quick clinical decisions that were impossible with traditional systems all the more precisely. The benefits of overview articles is to develop the current practice for better dynamics by using specific equations and stressing techniques of choice.

Determine the predictive performance of each algorithm and apply the proposed system to the area it needed. Use more than that relevant method of selection of features to improve accuracy algorithm performance. Make stakeholders use this the proposed methodology for creating an attractive working environment a condition that helps to make good decisions.

REFERENCES

1. Jeena, R. S., &Sukeshkumar, A. (2019, October). Development of a Stroke Risk Assessment Model for a Small Population in South Kerala using Logistic Regression. In TENCON 2019-2019 IEEE Region 10 Conference (TENCON) (pp. 350-355). IEEE.
2. Ali, A. A. (2019). Stroke Prediction using Distributed Machine Learning Based on Apache Spark. Stroke, 28(15), 89-97.
3. Wu, C. S. M., Badshah, M., & Bhagwat, V. (2019, July). Heart Disease Prediction Using Data Mining Techniques. In Proceedings of the 2019 2nd International Conference on Data Science and Information Technology (pp. 7-11).
4. Dolatabadi, A. D., Khadem, S. E. Z., &Asl, B. M. (2017). Automated diagnosis of coronary artery disease (CAD) patients using optimized SVM. Computer methods and programs in biomedicine, 138, 117-126.
5. Do, B. H., Langlotz, C., & Beaulieu, C. F. (2017). Bone tumor diagnosis using a naïve Bayesian model of demographic and radiographic features. Journal of digital imaging, 30(5), 640-647.
6. Min, S. N., Lee, K. S., Park, S. J., Subramaniyam, M., & Kim, D. J. (2017, July). Development of Stroke Diagnosis Algorithm Through Logistic Regression Analysis with National Health Insurance Database. In International Conference on Applied Human Factors and Ergonomics (pp. 364-366). Springer, Cham.
7. Acharya, U. R., Vidya, K. S., Ghista, D. N., Lim, W. J. E., Molinari, F., &Sankaranarayanan, M. (2015). Computer-aided diagnosis of diabetic subjects by heart rate variability signals using discrete wavelet transform method. Knowledge-based systems, 81, 56-64.
8. Cheng, C. A., Lin, Y. C., & Chiu, H. W. (2014, July). Prediction of the prognosis of ischemic stroke patients after intravenous thrombolysis using artificial neural networks. In ICIMTH (pp. 115-118).
9. Colak, C., Karaman, E., &Turtay, M. G. (2015). Application of knowledge discovery process on the prediction of stroke. Computer methods and programs in biomedicine, 119(3), 181-185.
10. Dwivedi, A. K. (2018). Analysis of computational intelligence techniques for diabetes mellitus prediction. Neural Computing and Applications, 30(12), 3837-3845.
11. Healthcare dataset stroke data. [Cited 2019; Available from: https://www.kaggle.com/asaumya/healthcare-dataset-stroke-data.
12. Davis, J., &Goadrich, M. (2006, June). The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd international conference on Machine learning (pp. 233-240).
13. Chai, K. M. A. (2005, August). Expectation of F-measures: Tractable exact computation and some empirical observations of its properties. In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 593-594).