International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 07 | Apr 2021 www.irjet.net p-ISSN: 2395-0072

CEREBROVASCULAR ACCIDENT PREDICTION USING LOGISTIC REGRESSION ALGORITHM

M.GEETHA[1], P.RADHIKA[2], S.DHARANI PRIYA[3], PURANAM RAVIKUMAR SAI LATHA[4]

Assistant Professor of M.E., Computer Science Department,
Panimalar Institute of Technology, Chennai. [1]
Students of B.E., Computer Science Department,
Panimalar Institute of Technology, Chennai. [2, 3, 4]

geetha114new@gmail.com[1], radhikapadhu97@gmail.com[2],
dharanisakthi11@gmail.com[3],
prsailatha2000@gmail.com[4]

Abstract - *The modern world is getting affected by cerebrovascular accident. This disease affects a person in such a way so that the patients can't be cured as easily as possible. So, the main objective of this research project is predicting the brain stroke risk level of a patient using Big Data Analytics and Machine Learning. The prediction of stroke disease is based on the dataset attributes. It is proposed to develop a model which can predict the vulnerability of a stroke given basic symptoms like age, sex, hypertension etc. The results give the accuracy of the logistic regression algorithm based on the dataset that have been taken. Performance validation is done using Receiver Operating Characteristics (ROC) curve. And also the co-efficient and confusion matrix for the regression output is found.*

***Key Words*: Machine Learning, Big Data Analytics, Receiver Operating Characteristics(ROC), Logistic Regression Algorithm.**

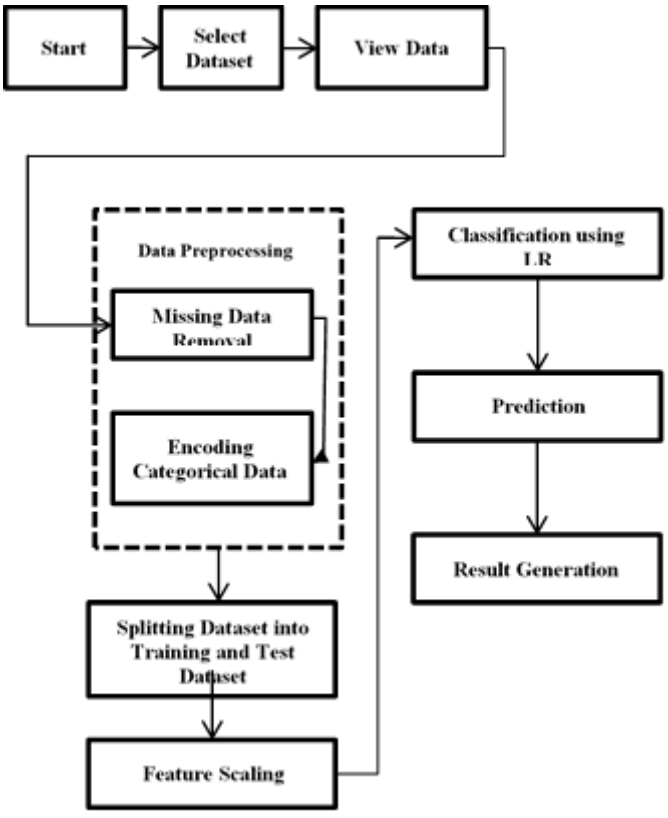
1.INTRODUCTION

Stroke is the one of the most common disease. This disease is quite common now-a-days and has long term disabilities which if predicted can be prevented. Different attributes are taken which can relate to this disease well to find the risk factors for prediction of stroke. Logistic regression is used on dataset based on risk factors. The prediction of stroke disease is based on the dataset attributes.

The results give the accuracy of the logistic regression algorithm based on the dataset we have taken. Performance validation is done using Receiver Operating Characteristics (ROC) curve. And also the co-efficient and confusion matrix for the regression output is found. Logistic Regression (LR)[12][13][14] is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis.

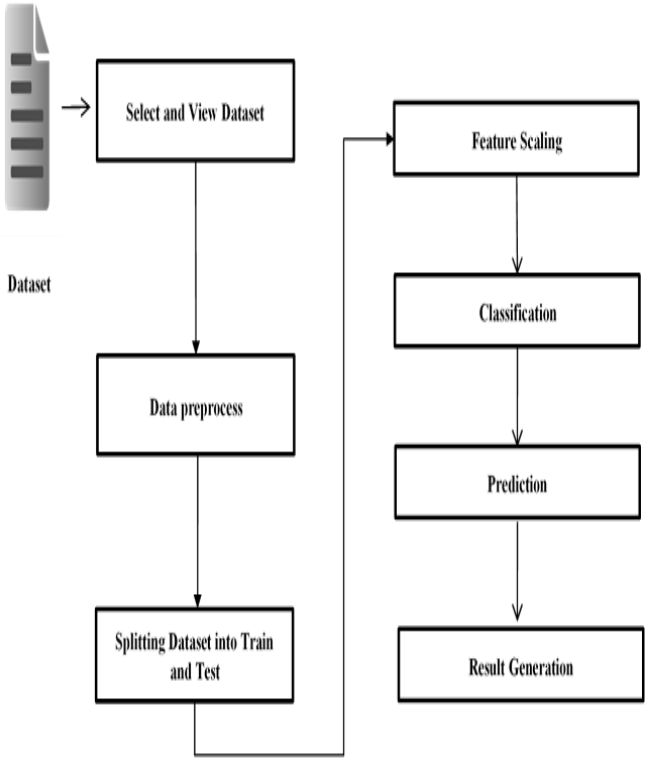
Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

2.SYSTEM ARCHITECTURE



In today's world, we could see that majority of the people face the risk of congestive heart failure which gives a sudden impingement to an individual that sometimes one lacks time to get treated immediately. Hence it is quintessential that timely and early diagnosis is performed which being quiet challenging concern for the medical association. This work utilized the logistic regression algorithm for machine learning which is very exact in providing the results when compared to the other algorithms.

3.METHODOLOGY



A.DATA SELECTION AND LOADING

The data selection is the process of selecting the data for detecting the attacks. In this project, the Heart disease dataset is used for detecting disease. The dataset which contains the information about gender,age,hypertension,heart_disease,ever_married, work_type,Residence_type,avg_glucose_level,bmi,smoking_status,stroke.

B.DATA PREPROCESSING

Data pre-processing is the process of removing the unwanted data from the dataset. Missing data removal: In this process, the null values such as missing values are removed using imputer library. Encoding Categorical data: That categorical data is defined as variables with a finite set of label values. That most machine learning algorithms require numerical input and output variables. That an integer and one hot encoding is used to convert categorical data to integer data.

C.SPLITTING DATASET INTO TRAIN AND TEST DATA

Data splitting is the act of partitioning available data into two portions, usually for cross-validator purposes. One portion of the data is used to develop a predictive model. And the other to evaluate the model's performance. Separating data into training and testing sets is an important part of evaluating data mining models. Typically, when you separate a data set into a training set and testing set, most of the data is used for training, and a smaller portion of the data is used for testing.

D.FEATURE EXTRACTION

Feature scaling. Feature scaling is a method used to standardize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data pre-processing step. Feature Scaling or Standardization: It is a step of Data Pre Processing which is applied to independent variables or features of data. It basically helps to normalize the data within a particular range. Sometimes, it also helps in speeding up the calculations in an algorithm.

E.CLASSIFICATION

Logistic regression is another technique borrowed by machine learning from the field of statistics. It is the go-to method for binary classification problems (problems with two class values). In this post you will discover the logistic regression algorithm for machine learning. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

F.PREDICTION

It's a process of predicting the attacks in the network from the dataset. This project will effectively predict the data from dataset by enhancing the performance of the overall prediction results. And also find the accuracy of the prediction and generate confusion matrix for the output.

G.EXPLORATORY ANALYSIS

Exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. In this phase, we take each and every variables like age, BMI, average glucose level and do exploratory analysis.

H.VALIDATION OF PERFORMANCE

A confusion matrix is a summary of prediction results on a classification problem. The number of correct and incorrect predictions are summarized with count values and broken down by each class. This is the key to the confusion matrix. The confusion matrix shows the ways in which your classification model is confused when it makes predictions. It gives us insight not only into the errors being made by a classifier but more importantly the types of errors that are being made.

4.RESULTS

With the increase in the stroke rates at juvenile ages, there is a need to put a system in place to be able to detect the symptoms of a stroke at an early stage and thus prevent it. Thus it is proposed to develop a model which can predict the vulnerability of a stroke given basic symptoms like age, sex, hypertension etc. The machine learning algorithm **logistic regression** has proven to be the most accurate and reliable with the accuracy of 98.25%.

S. No	No. of Techniques	Accuracy (%)	Time(s)
1	Sequential Minimal Optimization (SMO)	84.07	0.02
2	Bayes Net (BN)	81.11	0.02
3	Multi-Layer Perception (MLP)	77.4	0.75
4	Navies Bayesian (NB)	89.77	0.01
5	Logistic regression	98.25	0.01

From the above table, we can inference the proposed classification techniques performance which is compared with prevailing techniques of SMO (Sequential Minimal Optimization), Bayes Net and MLP (Multi-Layer Perception), Navies Bayesian. Effective outcome is exhibited by the proposed Logistic regression with greater performance in contrast to rest of the techniques.

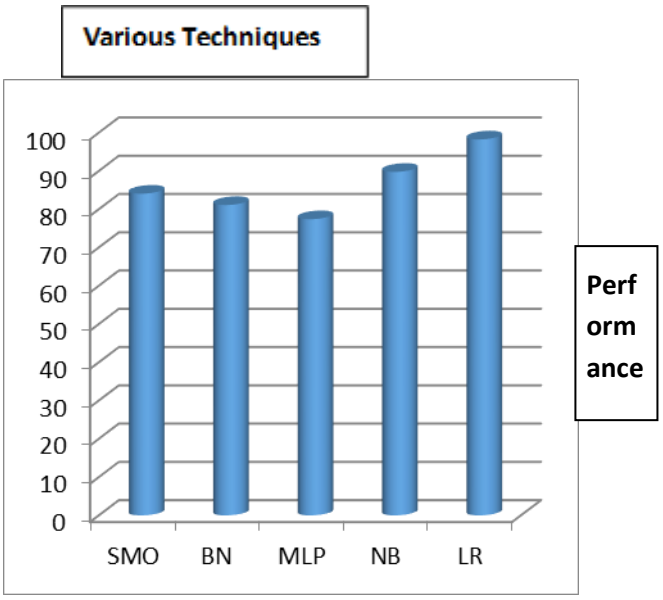


Fig: Performance of Heart disease classification techniques

The above graph displays the comparsion between Logical Regression classification techniques to different techniques like SMO (Sequential Minimal Optimization), Bayes Net and MLP (Multi-Layer Perception) , Navies Bayesian. Effective performance is reported with the proposed heart diseases classification technique than the remaining ones.

In Anaconda Spyder, the data can be downloaded from the internet and read from the storage and used within the IDE. Various operations like data cleaning, checking

missing data, missing data removal, can be done on the data.

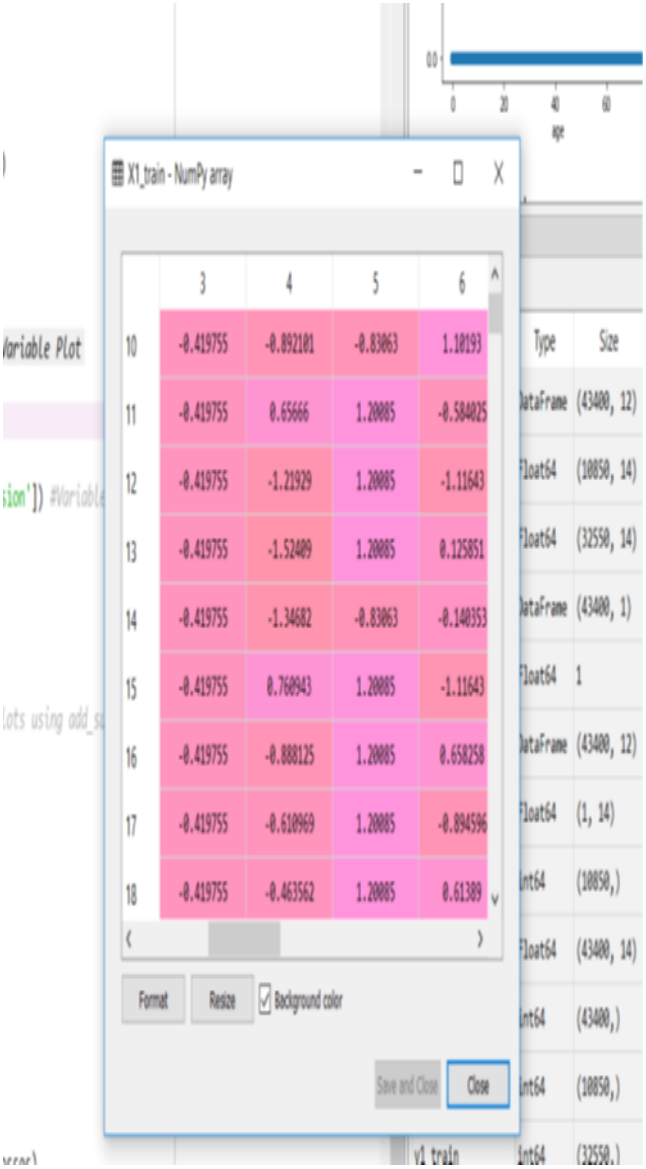


Fig: Standardized Dataset

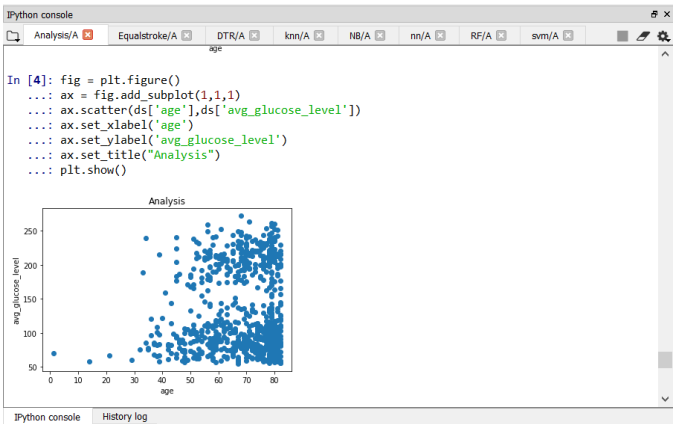


Fig: Exploratory Analysis

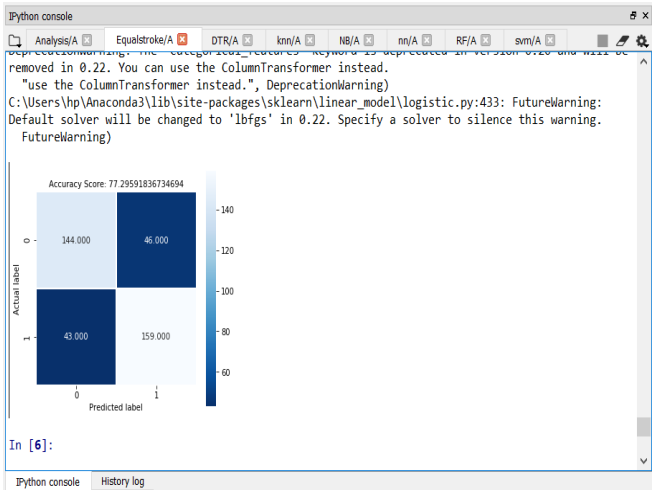


Fig Confusion Matrix

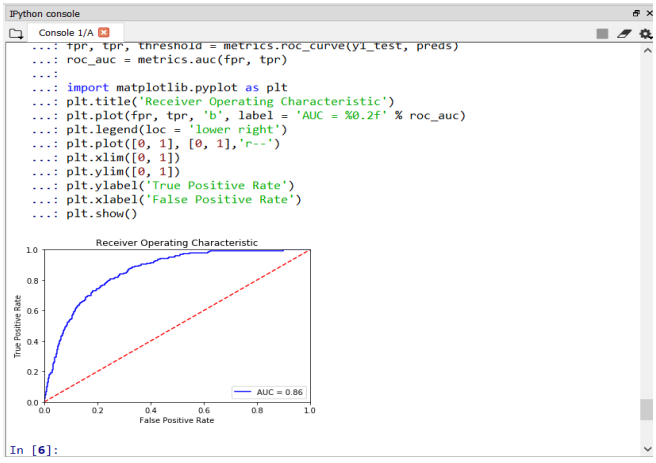


Fig.ROC Curve

5. CONCLUSIONS

Thus, through this project a comprehensive analysis of data using machine learning was implemented on a common dataset. Various operations are done on the dataset, the system developed is trained using the dataset. The results were acquired and interactive analysis is made. A confusion matrix is used for summarizing the performance of the algorithm. The true positive rate and false positive rate are plotted in the ROC curve. Performance of the algorithm is tested using ROC curve.

6.FUTURE ENHANCEMENT

As far as future work, other disease such as Heart diseases, kidney, diabetes etc., can be enhanced. We can also make use of other variables such as sugar level, blood pressure level, and cholesterol level for predicting exposure to stroke. The True Prediction Rate (TPR) and false prediction Rate (FPR) of the Receiver Operating Characteristics (ROC) curve can be enhanced further more.

7.REFERENCES

[1] Elena V Kuklina, "Assessing and Managing Risk for Cardiovascular Disease", Apr 2010, Vol 3, No.2.
[2] G.G. Bennett, S.J. Herring, E. Puleo, E.K. Stein, K.M. Emmons, M.W.Gillman, "Web-based weight loss in primary care: a randomized controlled trial, Obesity18 (2010) 308–313.
[3] John T. Behrens, "Principles and Procedures of Exploratory Data Analysis", The American Psychological Association, Inc., Vol. 2, No. 2, 131-160 1082-989X.
[4] K G M Moons, M L Bots, J T Salonen, P C Elwood, A Freire de Concalves,Y Nikitin, J Sivenius, D Inzitari, V Benetou, J Tuomilehto, P J Koudstaal, D E Grobbee," Prediction of stroke in the general population in Europe (EUROSTROKE)", pp i30–i36 2005.
[5] Mai Shouman, Tim Turner, Rob Stocker, "Integrating Clustering with different Data Mining Techniquies in the Diagnosis of Heart Disease", Journal of Computer Science and Engineering, volume 20, issue 1, august 2013.
[6] Mohammad Shafenoor Amina, Yin Kia Chiama, Kasturi Dewi Varathan,"Identification of significant features and data mining techniques in predicting heart disease", Volume 36, Pg 82-93, March 2019.
[7] Mohsin Raza, "Repository Evaluation, Selection, and Coverage Policies",Thomson Reuters, S SR 1206 119, 2012.
[8] Ohoud Almadani, KSA Riyad Alshammari, "Prediction of Stroke using Data Mining Classification Techniques", Vol. 9, No. 1, 2018.
[9] Tom Fawcett, "An introduction to ROC analysis", Institute for the Study of Learning and Expertise, Pattern Recognition Letters 27, 861–874, 2006.
[10] Tom M. Mitchell, "Generative and Discriminative Classifiers: Naive bayes and Logistic Regression", Machine Learning", Chapter 3, February 2016.
[11] X.-F. Zhang, J. Attia, C. D'Este, X.-H. Yu, X.-G. Wu, A" risk score predicted coronary heart disease and stroke in a Chinese cohort", J. Clin.Epidemiol. 58 (2005)951–958.
[12]. DR.S.Hemalatha "Early Detection of Dementia by Observing Change in the Driving Pattern of a Person using Smart Phone Sensors and DTW Algorithm" International Journal for Research in Applied Science & Engineering 2016 IC Value: 13.98.
[13].Dr.S.Hemalatha, "Driving style acknowledgement based permit issuance utilizing cell phone sensors and DTW calculation "International Journal for Research in Applied Science & Engineering Technology (IJRASET) 2016 IC Value: 13.98.
[14]. P. Sheela Rani, P. Subhashree , N. Sankari Devi "Computer vision based gaze tracking for accident prevention " 2016 World Conference on Futuristic Trends in Research and Innovation for Social Welfare. DOI: 10.1109/STARTUP.2016.7583976 IEEE.

.