International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT217656

Predicting stroke risk by Migraine using AI

Anchal Singh*[1], Dr. Surabhi Thorat[2]
*[1]Computer Science Department, S. K. Somaiya College, Somaiya Vidyavihar University, Mumbai, India
[2]Professor, Department of Computer Science, S. K. Somaiya College, Somaiya Vidyavihar University, Mumbai, India

Article Info
Volume 7, Issue 6
Page Number: 285-290

Publication Issue :
November-December-2021

Article History
Accepted : 05 Dec 2021
Published : 15 Dec 2021

ABSTRACT

Stroke is a blood clot or bleeds in the brain, which can make permanent damage that has an effect on mobility, cognition, sight or communication. It is the second leading cause of death worldwide and one of the most life- threatening diseases for persons above 65 years. It damages the brain like “heart attack” which damages the heart. Every 4 minutes someone dies of stroke, but up to 80% of stroke can be prevented if we can identify or predict the occurrence of stroke in its early stage. In this paper, I used different types of machine learning algorithms for stroke prediction on the Healthcare Dataset Stroke data. Four types of machine learning classification algorithms were applied; Linear Regression, Confusion matrices, Random Forest Classifier, and Logistic Regression were used to build the stroke prediction model. Support, Precision, Recall, and F1-score were used to calculate performance measures of machine learning models. The results showed that Random Forest Classifier has achieved the best accuracy at 94 % [1].

Keywords : Stroke, Logistic regression, Random Forest classifier and Machine learning algorithm.

I. INTRODUCTION

Migraine is the name given to the condition that people who experience these headaches often have. Migraine can range from mild to extremely debilitating and last for just a few hours but some cases can be more severe and last for days. It is one of the most common ailments in the united states, with over 36 million people experiencing them each year. Some people experience when they are stressed, like when they are under pressure at work or during an exam, or in bad weather. Yet others suffer from them seasonally or all year round. Not every migraine sufferer shows symptoms of sensitivity to light, sound, or smell. Some may show sensitivity to all three while others experience only one or two of these symptoms. According to headache disease in many countries takes place due to excess work, mental stress and many more reasons supporting it. Diagnosis is complex and major task that needs to be executed precisely and instinctively. The diagnosis is usually made based on doctor’s practice and knowledge. This guides to undesired results and extreme medical loss of treatments provided to patients. Hence, the prediction of different types of headache, blood pressure, blood sugar.

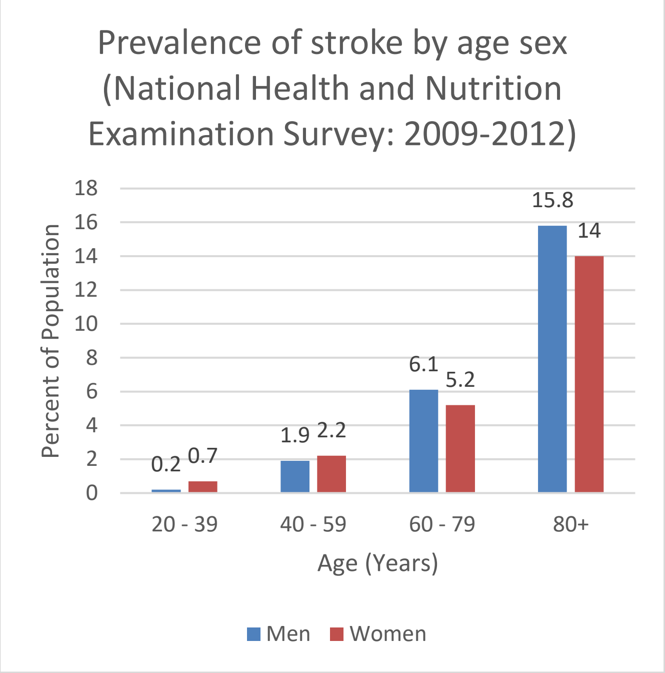
Early prediction of stroke diseases is useful for the prevention or for early treatment mediation. Machine learning and data mining are playing key roles in predicting stroke. For example, support vector machine, logistic regression, random forest classifier and linear regression. Machine learning is a type of artificial intelligence that targets to design a computer with human thinking capability. The aim of machine learning grants computers to make a specific task depending on patterns and interference without using clear instruction [2].

What is Brain stroke?

A stroke is an interruption in the blood supply to the brain, caused by the rupture of one or more small blood vessels. The interruption deprives essential oxygen and nutrients, which can result in neuronal death. A stroke is a brain condition in which part or all of the brain dies. The most common type of stroke, called ischemic stroke, occurs when an artery that supplies blood to the brain becomes blocked by a blood clot, usually due to smoking tobacco and high cholesterol. Many people believe that only older people develop strokes because they are more prone to heart disease and diabetes. The main aim of this study was to explore the relation between migraine and ischaemic stroke in people.

A. Factor responsible for stroke

Strokes are caused by clogging in the brain's arteries that carry oxygen and nutrients to the brain. This malfunctioning of the arteries can be caused by other factors, like smoking, diabetes, high blood pressure, cholesterol and heart disease. It shows differences based ontheir age and gender.



In this graph, Percent of Population versus Age (Years) shows the frequency of stroke by age and sex according to National Health and Nutrition Examination Survey: 2009 – 2012. In which, blue graph indicates men and orange graph indicates women. In this age of 80 and above had more pervasiveness of stroke to others in which stroke occurrence percent in men was higher than women and 20 to 39 age group had fewer than the other ages group in which stroke occurrence percent in women was higher than men.

The organization of this document is as follows. In Section 2 (Methods and Material), I'll give detail of any modifications to equipment or equipment constructed specifically for the study and, if pertinent, provide illustrations of the modifications. In Section 3 (Result and Discussion), present your research findings and your analysis of those findings. Discussed in Section 4(Conclusion) a conclusion is the last part of something, its end or result.

II. METHOD AND MATERIAL

This research paper explained the idea of establishing a large cohort of migraine patients affected by stroke in this world.

3.1 Database

I used healthcare care dataset which is a secondary data in which both categorical and numerical features were identified. It was used to train and test models for predicting stroke disease. This dataset includes of 10 independent variables as features and one dependent variable as the class label that is used to predict stroke disease. The Variables name is like gender, age, hypertension, heart_disease, ever_married, work_type, residence_type, avg_glucose_level, bmi, smoking status and stroke. There were two values for class label which is: 0 for absence of stroke; another is 1 for presence of stroke.

A. Data Pre-Processing:

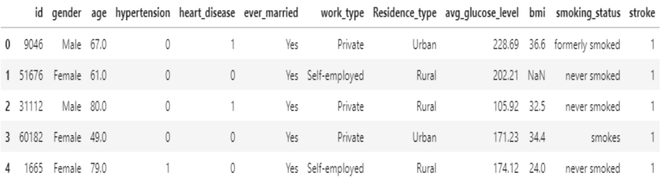
Data pre-processing is the main step for sufficiently illustrating the data for the machine learning algorithm. It is playing an important role in enhancing the performance results of machine learning. In this stage, some steps are applied.

1. BMI feature has many missing values. Mean or median is applied to fill missing values.
2. Both categorical and numerical features were identified. Converting categorical features into numerical data, we are using LabelEncoder.
3. The database is imbalanced data. Imbalanced data means there is an unequal ratio of values for each class label. We handle imbalanced data by using SMOTE algorithm.

In this step, we will pre-prepare the data so that we can use it in our code effectively. The code is given below:

1. #Data Pre-processing Step
2. # importing libraries
3. Import numpy as nm
4. import matplotlib.pyplot as mtp
5. import pandas as pd
6. #importing datasets
7. data_set= pd.read_csv('healthcare_dataset_stroke_data.csv')

By executing the above code, we will get the dataset as the output. Consider the given image:



	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1

B. Machine Learning Algorithm:

In this stage, four types of machine learning were used: Logistic regression (LR), Random Forest classifier (RF), Confusion Metrices and Linear Regression for both RF and LR.

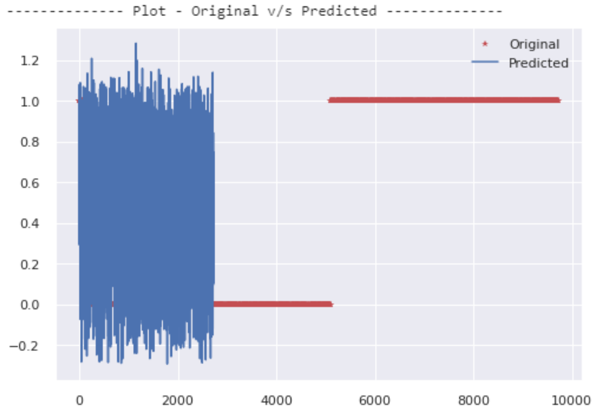
• Logistic regression algorithm is used to find the relation between the target variables and predictive variables. The target value is lie between 0 and 1. The values for logistic regression is 0.814 and for Random Forest is 0.943.

• Random forest is a very popular machine learning classifier for developing prediction models in many research settings. It is a collection of trees which are constructed using randomly selected training datasets and random subsets of predictor variables for modelling outcomes. Random forest usually gives higher accuracy compared to a single decision tree model.

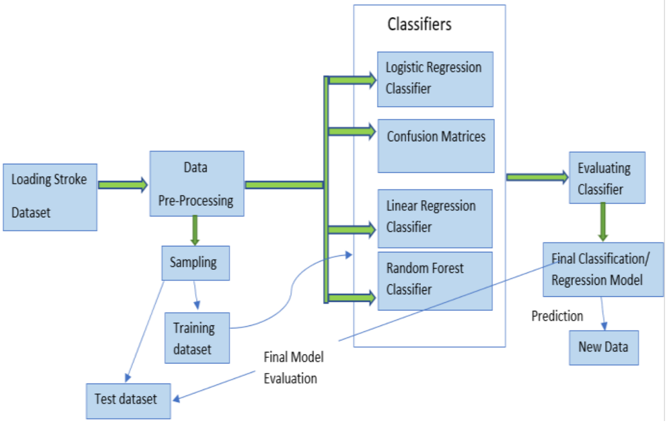
• Confusion metrices is a table that is used to describe the performance of a classifier on a set of test data and visualize the predictive analysis like accuracy, recall and specificity. They give direct comparison of values like true and false positive, negative. It helps to plotted a heat map for Random Forest and Logistic Regression.

• Linear regression demonstrates the connection between the two variables by proper a linear equation to discover the information. One variable is considered as an explanatory variable and the other is consider as a dependent variable. It is used to evaluate real values based on continuous variables. Here, we show relationship between independent and dependent variables by fitting a best line. This best fit line is called as regression line and it represented by a linear equation i.e. Y= a *X + b. In this equation 'y' represents dependent variable, 'a' represents slope, 'X' represents independent variable and 'b' represents intercept. It is probably one of the most popular algorithms in statistics.

Figure shows the plot between original and predicted data. In this, Linear regression performance on the test dataset in which value of Test MAE is 0.294 and value of Test R2 is 0.45 are given below.



C. Flow chart of Methodology:



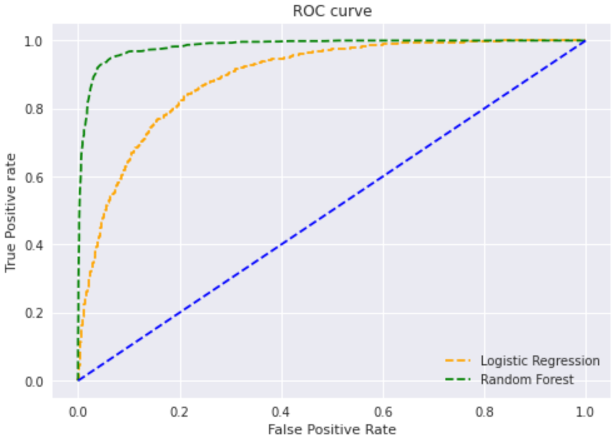
III. RELATED WORK

In this paper authors [1], worked is accomplished by a big data platform that is Apache Spark. MLib library is an API integrated with Spark to provide machine learning algorithms. In this, Decision Tree, Support Vector Machine, Random Forest Classifier and Logistic Regression were used to build the stroke prediction model. The hyperparameter tuning and cross-validation were involved with machine learning algorithms to enhance results. The proposed stroke prediction system is developed on Apache Spark. The results showed that random forest classifier achieved the best accuracy result. In this paper [3], Find of the 49,711 patients are hospitalized for first stroke, 1084 were migraineurs by using triptans. These data include age, sex, stroke on admission uniformed on the Scandinavian stroke scale, stroke subtype and cardiovascular profile. Ischemic stroke was respected from hemorrhagic stroke by computed tomography. They used dataset table such as Basic characteristics of the Danish population aged 25–80 years in the period 2003–2011 by triptan use and Relative risks and 95% confidence intervals for triptan users compared to non-users by stroke type in this data. In this study, migraine is associated with an etiology of stroke different from thromboembolism this may have clinical suggestions for this part of the stroke population in regard to both acute treatment and secondary elimination. In this Paper authors [4], came up with four classification technique K-NN, SVM, Random Forest, & Naïve Bays. In this database, with 114 different data having six different categories have been observed to manage the experimentation with proposed migraine headache classification technique. The results clearly reveal performance improvement with proposed migraine headache classification compare to intensity of pain, environment factors and their associated symptoms which is the best results Naïve Bays Classification technique among four. They used four classifiers such as KNN, Naïve Bays, Random Forest and support vector machine, the results of accuracy score of the naïve bays is 0.475 AUC and Precision is 0.905 so, the Naïve bays is the best classifier out of these. In this paper [5], They included 107 acute anterior ischemic stroke patients were medicated by the cardiovascular method. They have used algorithms such as MATLAB, SPSS, artificial neural networks & support vector to form a supervised machine which is able to classify the above predictors. It showed a positive accuracy closer to 70% of predictive output using supervised machine learning. In this experiment conducted on probably collected database of acute ischaemic stroke managed by endovascular intervention. The final confusion matrix of the neural network, demonstrated an overall accordant of 80% between the target and output classes. But the support vector machine had better performance, with a root mean squared error of 2.064. In this paper authors [6], used the Data Mining techniques like classification, logistic regression in health domain to predict the Ischemic Stroke. They suggested a model for predicting Ischemic stroke using Data mining Techniques which were classification, logistic regression. They used data software, algorithm and logistic regression for pre-processing, cleaning and analyzing the data. Experiments have been conducted on medical database of LGPM, training set consists all the information regarding patient. However, in the confusion matrix, they concluded the prediction model achieves 28 + 37 = 65 bad predictions. The error rate is 19.7% for any algorithm its accuracy and performance of greater significance.

IV. RESULT AND DISCUSSION

In this section showing the performance of two algorithms are Logistic regression and Random Forest. Higher the AUC, better the model is at differentiating between patient with disease and no disease. After plotting AUC-ROC curve we can observed that Random Forest curve is higher than the Logistic Regression ROC curve. Therefore, Random Forest did a better task of classifying the positive class in the dataset. The ROC curve is plotted with True Positive Rate (TPR) against the False Positive Rate (FPR) where, TPR represented on y-axis and FPR represented on the x-axis. After this, F1 score is applied on this data which is the average of precision and recall.

In this, accuracy score of applying Random Forest and Logistic regression. The Random Forest has the highest accuracy score at 98% and the Logistic regression has accuracy score at 89%.



V. CONCLUSION

We begin with reading data and then classified categorical features and numerical features. After this we deal with null values in BMI feature. Then we perform EDA on features. We accomplish that we have imbalance data that is negative examples is greater that positive class. After visualization we handle imbalance data. After this we move to most important part of building model. Before starting to train model, we split our data into train data (testing purpose) and test data (validation purpose) and perform Feature scaling. Random Forest Classifier and Logistic Regression models were tried in this. To check which model performs best plot ROC-AUC curves along with classification report and confusion matrices. While Random Forest win the race. I therefore selected the Random Forest as my model.

VI. REFERENCES

[1]. Hager Ahmed, Sara F Abd-el Ghany, Eman Younis, Nahla Omran “Stroke Prediction using Distributed Machine Learning Based on Apache Spark” International Journal of Advanced Science and Technology Vol. 28, No. 15, (2019), pp. 89-97

[2]. Kunder Akash Mahesh, Shashank H N, Srikanth S, Thejas A M “Prediction of Stroke

Using Machine Learning" researchagte.net publication, June 2020

[3]. Vanna Albieri, Tom Skyhøj Olsen, Klaus Kaae Andersen "Risk of Stroke in Migraineurs Using Triptans. Associations with Age, Sex, Stroke Severity and Subtype" EBioMedicine 6(2016) 199-205

[4]. Rahul Deo Sah, Dr. Jitendra Sheetlani, Dharam Raj Kumar, and Indra Nath Sahu "Migraine (Headaches) Disease Data Classification Using Data Mining Classifiers" Quest Journals Journal of Research in Environmental and Earth Science Volume 3~ Issue 4 (2017) pp: 10-16

[5]. Hamed Asadi, Richard Dowling and Bernard Yan, "Machine Learning for outcome prediction of acute ischemic stroke", PLOSONE Vol.9 Issue2, Feb2014

[6]. Balar Khalid and Naji Abdelwahab, "A model for predicting Ischemic stroke using Data Mining algorithms", IJISET, Vol. 2 Issue 11, Nov 2015, ISSN: 2348-7968.

[7]. Mrs. Veena Potdar, Mrs. Lavanya Santhosh, Yashu Raj Gowda CY "A Survey on Stroke Disease Classification and Prediction using Machine Learning Algorithms" International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 IJERTV10IS080219 Vol. 10 Issue 08, August-2021

Cite this article as :

Anchal Singh, Dr. Surabhi Thorat, "Predicting stroke risk by Migraine using AI", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7 Issue 6, pp. 285-290, November-December 2021. Available at doi : https://doi.org/10.32628/CSEIT217656
Journal URL : https://ijsrcseit.com/CSEIT217656