International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 05 | May - 2023 SJIF 2023: 8.176 ISSN: 2582-3930

Early Detection of Brain Stroke using Machine Learning Techniques

Bhoomika R S[1], Aishwarya J Annigeri[2], Himaja Desiraju[3], Harsh Kumar Goswami[4], Prof. Vani K.A [5]

Department of Information Science and Engineering, Dayananda Sagar College of Engineering

[1]123rsbhoomika@gmail.com
[2]aishwaryajannigeri@gmail.com
[3]himajadesiraju@gmail.com
[4]harshgoswami38@gmail.com
[5]vanika-ise@dayanandasagar.edu

Abstract - As in line with the WHO, stroke is the principle source of death and incapability round the world . It is a clinical condition that causes brain damage by way of tearing blood vessels.It is able to likewise take place whilst the development of blood and unique dietary supplements to the cerebrum is disturbed. Most of exploration has targeted on foreseeing breathing screw ups, however not very many examinations have focused on the gamble of a mind assault. Numerous device learning fashions are being advanced with this in mind and are expecting the possibility of a mind stroke. Various physiological elements were used on this paper to educate various fashions for correct prediction using 10 machine studying algorithms inclusive of logistic regression, decision tree type, k-nearest acquaintances, random forest classifier, svm ,adaboost classifier,bernoullinb,mlp,bagging classifier and lgbm.

***Key Words*:** stroke,logisticregression,decisiontree,knn,svm,randomforest,adaboost,bernoullinb,mlp,bagging,lgbm,flask

1.INTRODUCTION

Stroke takes place while the blood circulation to exclusive regions of the cerebrum is disturbed or faded, prompting the cells in the ones areas of the thoughts to kick the bucket because of an absence of dietary supplements and oxygen. Stroke is the second maximum common reason of loss of life and person disabilities round the arena, with four hundred–800 strokes consistent with 100,000 human beings, 15 million new acute strokes each 12 months, 28,500,000 disability healthy life years, and 28–30-day case fatality fees ranging from 17% to 35%. With the amount of passes from coronary contamination and stroke anticipated to ascend to 5,000,000 each 2020 from three,000,000 out of 1998, the weight of stroke might be going to become worse. This could occur because of the continuous phase and health changes so as to spark off an expansion within the gamble elements for vascular illness and the extra installed populace. Emerging economies are chargeable for 85% of all stroke deaths worldwide.

Normal stroke threat factors-stroke can take place in all people, regardless of their race, sex, or progress in years; however all matters being the same, the probability of having a stroke increases assuming a character has any gamble factors that could set off a stroke.

Spotting individual playing and how to oversee it is the pleasant method for protecting oneself as well as different people.As per research, eighty% of strokes can be saved faraway from as such. Modifiable and non-modifiable stroke threat attributes are outstanding. Modifiable hazard attributes have been in addition classified as lifestyle danger attributes and clinical chance attributes. Way of life threat credits, for instance, tobacco usage, substance use, absence of hobby, and excessive bmi can regularly be modified, while scientific gamble ascribes, as an example, hypertension, intrinsic coronary illness, diabetes, and high blood ldl cholesterol can usually be handled. Non-modifiable risk attributes, on the other hand, whilst uncontrollable, are useful resources within the identity of individuals prone to stroke.

A solid, adjusted manner of existence that rejects risky propensities like smoking and ingesting, maintains a valid weight record (bmi), everyday blood glucose tiers, and high-quality coronary heart and kidney capability can assist with forestalling stroke. Foreseeing a stroke is basic, and it ought to be dealt with fast to live away from irreversible damage or demise with the development of medical innovation, it's miles presently viable to utilize ml techniques to foresee the beginning of a stroke. Logistic regression, decision tree class, k-nearest neighbors, random forest classifier, svm and different ML algorithms are beneficial because they permit for correct prediction and proper analysis.

The closing part of the studies article is based as follows. We analyze some present literature of new findings in segment 2. Phase 3 includes an assertion of the methodological approaches. The results and analysis are represented in section 4, and the performance evaluation and correlation effects are defined in elements. Finally, in segment 5 the conclusion is defined.

2. LITERATURE SURVEY

In the paper[1], a weighted voting classifier is proposed for prognosticating stroke using conditions attributes like high blood pressure, bmi, heart grievance, common glucose role, smoking reputation, former stroke, and age. The proposed weighted vote casting classifier's overall performance is as compared to that of nation- of- the- art classifiers similar as logistics retrogression(lr), Decision Tree classifier(dtc), stochastic gradient decomposition(sgd), k neighbors, and others with rigor 78, 91, 65, 87 independently

In the paper[2], 3 models — naive bayes, decision tree, and random forest — might be used within the take a look at. Every model's vaccination uses the case's medical history as a trait. As in step with the paper, the random forest ranks first in terms of accuracy with 94.781 accuracy status. It is observed by the decision tree, which has 91.906 accuracy status, and the

naive bayes technique, which has an 89.976 accuracy standing.

In the paper[3], the dataset was changed to received from the website of the International Stroke Trial(IST). It incorporated the coexisting general information age, exposure, the time between the morning of the circumstance and randomization, whether atrial fibrillation(af) became available, whether or not headache drug changed into wished in commodity like 3 days of the randomization, systolic palpitation at randomization, function of mindfulness, and neurological insufficiency. For factor evaluation and vaccination, random forest, multinomial naive bayes, and adaboost classifiers had been used on this paper

In paper[4], they sought to produce models for stroke vaccination in an elderly Chinese population with imbalanced statistics. Statistics have been accumulated from a potential cohort in 2012 and 2014 that had 1131 actors(56 stroke instances and 1075non-stroke subjects). To prognosticate stroke using demographic, life, and medical characteristics, machine literacy ways like formalized logistic regression(rlr), support vector machine(svm), and arbitrary timber(rf) have been carried out.

In the paper[5], the main donation of their exploration is crowds they conduct a methodical analysis of hazard elements for stroke vaticination; and that they deliver a widespread performance for stroke vaticination using slicepart ml algorithms comparable as decision tree algorithm with 74.31 accuracy, random forest with 74.53 accuracy and neural network device with 75.02 accuracy.

In the paper[6], a mongrel system learning method to prognosticate cerebral stroke for clinical opinion grounded at the cerebral statistics with space and imbalance changed is taken into consideration. Two methods were involved inside the entire manner in the beginning, random forest method turned into espoused to impute lacking values earlier than brackets. Secondly, an automated hyperparameter optimization(autohop) grounded on deep neural networks(DNN) is applied for stroke validation on an imbalanced dataset.

In the paper[7], they evaluate their machine for prognosticating strokes to different procedures using the cardiovascular health study(chs) dataset. Then, the pinnacle detail analysis set of rules is used to lessen the size, the decision tree algorithm is used to pick out the features, and lower back propagation neural network bracket algorithm is used to make a bracket model. A prophetic model for the stroke complaint with 97.7accuracy became established after assaying and comparing bracket edges with colorful styles and variant fashions.

In the paper[8], they compared DNN to a few other ML tactics for prognosticating 5-time stroke instances the usage of a massive populace- grounded emc database of more or less 800,000 instances. The outcomes show that dnn and grade boosting decision tree(gbdt) methods also can produce excessive vaticination rigor that outperform logistic regression(lr) and support vector machine(svm).

3. METHODOLOGY

In this layout, we advocate using different machine learning techniques to know methods to prognosticate stroke lawsuits based on high blood pressure, bmi function, coronary heart complaint, average glucose position, smoking reputation, former stroke, and age. The usage of these excessive- price features, ten exceptional classifiers had been trained, along with Logistic regression,K-Nearest Neighbors, Random Forest Classifier, Decision Tree Classifier, and others. Following that, the results of the bottom classifiers were added up using the weighted vote casting technique to achieve the loftiest accuracy.Every step might be made experience of within the following sub-regions.The figure below shows the proposed methodology.

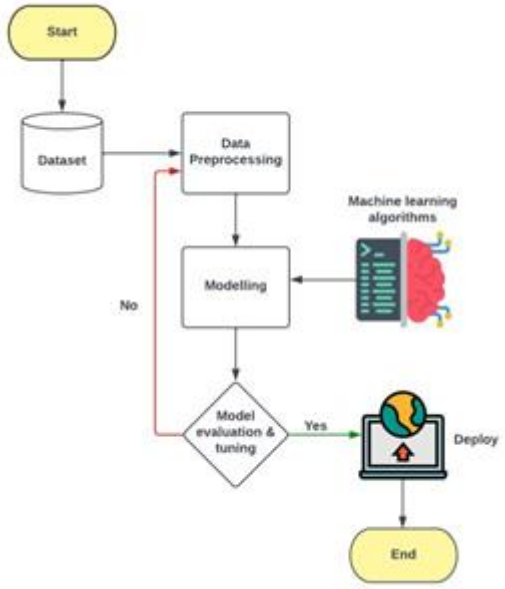

Fig -1: Proposed system methodology

A. Dataset: The preliminary step consists of accumulating a whole lot of records related to stroke sufferers, for instance, scientific records, real side effects, lab take a look at outcomes, and scientific imaging statistics. The dataset taken into consideration for our design is the healthcare-datasetstroke-facts[12].This dataset is used to prognosticate whether a case is likely to get stroke grounded on the input parameters like gender, age, colorful conditions, and smoking status. Every row within the information affords applicable records about the patient.This data could be used to teach the machine mastering set of rules to fetch patterns and perceive the early signs and symptoms of stroke. The dataset discussed above is explained in element in table 1.

Table -1: Stroke Dataset

Sl no.	*Attribute name*	*Type(values)*	*Description*
1	ID	Integer	Patient identification code

2	Gender	String literal(Male,Female, other)	Tells the sex of the patient
3	Age	Integer	Specifies the age of the patient
4	Hypertension	Integer(0,1)	Considers 0 for no & 1 for yes
5	Heart disease	Integer(0,1)	Considers 0 for no & 1 for yes
6	Ever married	String Literal(Yes,No)	Specifies if the patient is married or not
7	Work type	String literal(never_worked ,govt job,private,self employed etc)	Specifies work type of the patient
8	Residence type	String literal(Rural,Urban)	Specifies the type of residency of the patient
9	Average glucose level	Float	Specifies the avg glucose level in blood
10	BMI	Float	Index used to specify if the patient is obese or not
11	Smoking status	String literal (formerly smoked, never smoked, smokes, unknown)	Specifies the smoking status of the patient
12	Stroke	Integer(0,1)	Specifies the status of stroke

B. Data Preprocessing: Data has to be preprocessed after it's been collected to make sure that it's far suitable for machine learning algorithms. This can involve drawing the records to do away with missing or inconsistent information, homogenizing the statistics to insure that it is on a harmonious scale, and changing the facts into a numerical layout that may be reused by way of the set of rules.The dataset taken has 12 attributes, as stated in table 1. At first, the column' id' is dropped due to the fact it would not make an essential difference in model structure. Also the dataset is checked for null values and filled if any installation. In this case, the column' bmi' has null values full of the mean of the column information and the usage of a knn imputer.

C. Modeling: A model may be produced with the aid of education and a set of rules on a dataset after the information has been reused in conjunction with the rudiments. There are multitudinous styles of system literacy algorithms. We employ ten algorithms, every of that is defined underneath.

1) Logistic regression: Logistic regression is a supervised ML algorithm of a set of rules that is used to forecast the opportunity of a double outgrowth(i.e., yes or no,1/0, real/ fake). It really works through fitting a logistic characteristic to a fixed of enter capabilities and markers. The logistic feature maps any realvalued center to a price among 0 and 1, which represents the opportunity of a superb outgrowth. Logistic retrogression is usually utilized in operations similar to credit score scoring, unsolicited mail filtering, and scientific opinion.

2) Decision Tree Classifier: Decision tree classifier is a supervised system algorithm that is used to classify cases grounded on a fixed set of guidelines deduced from a tree-suchlike version. The set of rules works via recursively unyoking the input records into subsets grounded at the values of enter features, until a stopping criterion is met. The appearing tree can be used to classify new instances with the aid of following the route from the root knot to a splint knot that corresponds to the prognosticate class. Decision tree brackets are normally utilized in operations much like fraud discovery, client segmentation, and medical opinion.

3)K-Nearest Neighbour: K-Nearest Neighbors(KNN) is a supervised algorithm that is used for bracket and regression. The set of rules works via changing the k-nearest neighbors to a new enter case grounded on a distance metric(e.g., euclidean distance) among the input features. The prognosticate affair is also grounded at the maturity class(for bracket) or the average cost(for regression) of the nearest neighbors. Knn is normally used in operations much like recommender structures, photograph recognition, and anomaly discovery.

4)Random Forest Classification: Random forest classifier is a supervised machine studying a set of rules it truly is used for division. It works via constructing more than one selection bushes and adding up their prognostications to advantage a final validation. Each tree is built by way of aimlessly opting for a subset of enter capabilities and a subset of education cases. The performing ensemble of bushes is suitable to seize complex non-linear connections between enter features and affair lessons. Random forest classifier is usually utilized in operations just like credit score scoring, photograph bracket, and fraud discovery.

5)Support Vector Machine: Support vector machine(SVM) is a supervised system getting to know the set of rules it's used for bracket and regression. It really works by way of converting the hyperplane that maximizes the outer edge between lessons of entered instances. The periphery is defined as the distance between the hyperplane and the nearest instances from each elegance. Svm can take care of each direct and non-linear bracket problem and the usage of kernel functions that transfigure the input area into a complicateddimensional factor area. Svm is typically utilized in operations similar to textbook bracket, picture recognition, and bioinformatics.

6)AdaBoost: Adaboost(adaptive boosting) is an ML algorithm that is used for brackets. It works via combining multiple weak classifiers to produce a robust classifier. The vulnerable classifiers are trained on exclusive subsets of the training information and assigned weights grounded on their performance. The very last validation is grounded on the weighted sum of the weak classifier prognostications. Adaboost is normally used in operations much like face discovery, speech popularity, and bioinformatics.

7)BernoulliNB: BernoulliNB(bernoulli naive bayes) is an ML algorithm used for brackets. It is a version of the naive bayes set of rules that assumes the enter capabilities are double(i.e., yes/no,1/0). Bernoullinb calculates the tentative possibility of a class given the input functions the use of

bayes' theorem. It's normally utilized in textbook brackets, junk mail filtering, and sentiment analysis.

8)MLP: Mlp(multi-layer perceptron) is a type of artificial neural network used for bracket and retrogression. It consists of multiple layers of linked bumps(neurons) that carry out nonlinear metamorphoses at the center. The very last validation is grounded at the affair of the remaining subcaste. Mlp can handle complicated nonlinear connections between enter capabilities and affair lessons. It is typically utilized in operations just like photograph reputation, speech reputation, and natural language processing.

9)Bagging Classification: Bagging (bootstrap aggregating) is a machine mastering algorithm that is used for bracket and regression. It really works through constructing multiple models on unique bootstrap samples of the education data and adding up their prognostications to advantage a very last validation. The acting ensemble of models is appropriate to lessen friction and ameliorate performance as compared to a single version. Bagging classifier is commonly used in operations much like credit scoring, client churn vaticination, and fraud discovery.

10)LGBM: LGBM (Light Gradient Boosting Machine)is an ML algorithm that is used for bracket and regression. It's a variant of the grade boosting set of rules that makes use of a tree- grounded version to suit the facts. Lgbm is designed to be speedily and extra memory-effective than different grade boosting algorithms, making it appropriate for huge-scale datasets. Lgbm is normally used in operations similar to online advertising and marketing, advice structures, and fiscal modeling.

D. Model evaluation and tuning: This step entails assessing the performance of the machine learning methods by the usage of criteria comparable as sharpness, explicitness, and precision. This step helps to decide the effectiveness of the set of rules in detecting the early symptoms of stroke and pick out regions for enhancement. The performance of the model is predicted on a separate subset of the information that the algorithm has not seen beforehand. And additionally model tuning is completed to optimize the overall performance of the extraordinary gadget studying algorithms used for stroke vaticination. It is an essential step within the machine studying channel that could help to acquire the trendy viable performance of the models at the given dataset.

E. Model Deployment: In this step we deploy our model to a web site using the flask framework.Flask is a famous framework for constructing net operations using python. It is a light framework that doesn't endure specific gear or libraries, making it clean to set up and use.Model deployment with flask includes integrating a skilled version into a product terrain for making prognostications on new information. The method consists of saving the version in a format that may be loaded by way of flask, developing a flask operation, defining a validation characteristic, and putting in url routes for the operation. As soon as the operation is configured, it can be stationed to a product terrain similar to an internet garçon. Through planting the model with the use of flasks, our layout can give healthcare specialists and cases an easy- to- use internet interface for stroke risk prediction.

4. OUTCOMES AND DISCUSSION

The results and model comparison are given in Table2. The outcome and model contrast are given in Table 2. Accuracy is given in terms of percentage, and we've considered accuracy till the third decimal point. The(0,1) stated in precision,recall & f1-score corresponds to(no-stroke,stroke). We have taken constant support of(1591,96) independently.

Table -2: Model Comparison

Model	*Accuracy*	*Precision(0,1)*	*Recall(0,1)*	*f1-score(0,1)*
Logistic regression	94.309	(0.94,0)	(1,0)	(0.97,0)
Decision Tree	90.456	(0.95,0.15)	(0.95,0.15)	(0.95,0.15)
KNN	94.190	(0.94,0.25)	(1,0.01)	(0.97,0.02)
Random Forest	94.250	(0.94,0)	(1,0)	(0.97,0)
SVM	94.309	(0.94,0)	(1,0)	(0.97,0)
AdaBoost	94.072	(0.94,0.17)	(1,0.01)	(0.97,0.02)
Bernoulli NB	93.775	(0.94,0.24)	(0.99,0.04)	(0.97,0.07)
MLP	94.309	(0.94,0.00)	(1,0)	(0.97,0)
Bagging	93.953	(0.94,0.25)	(0.99,0.03)	(0.97,0.06)
LGBM	94.368	(0.95,0.55)	(1,0.06)	(0.97,0.11)

As indicated by using the above table, the lgbm model is the maximum reliable and right model for our facts. As a result, we utilize the flask framework to install this model to our designed web page to foresee stroke probabilities. We input data into the web page through an input form structure (fig 2), that's then used to interrupt down stroke chance making use of the lgbm version. Due to the end result, it'll set off a stroke page (fig 3) or no stroke page (fig 4).

Fig -2: Input form structure

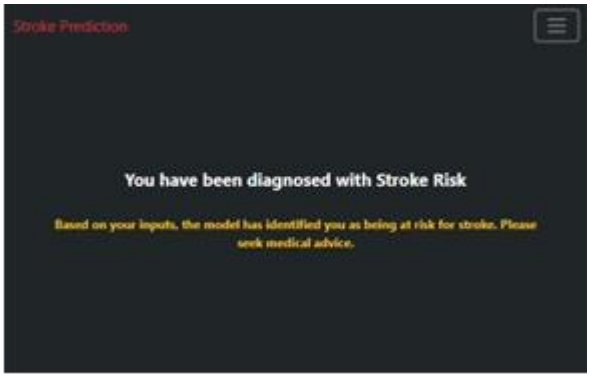

Fig -3: Stroke Page

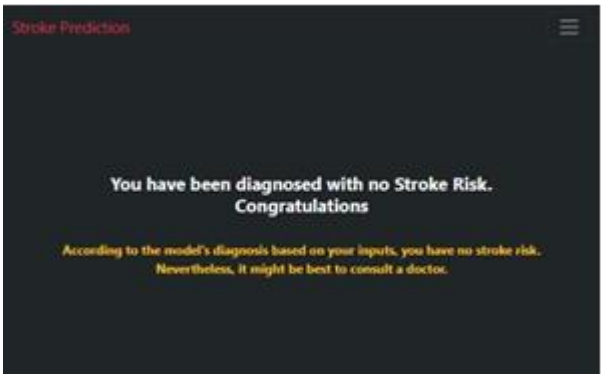

Fig -4: No Stroke Page

5. CONCLUSIONS

Several machine learning algorithms were used in the project, including Logistic Regression, Decision Tree Classification, K-Nearest Neighbors, Random Forest Classifier, Support Vector Machine, AdaBoost, Bagging classifier, MLP, BernoulliNB, and LGBM.It was discovered that lgbm produced the maximum accurate effects for the dataset used. As a way to make the challenge available and simple to apply for sufferers and healthcare experts, flask changed into extensively utilized for version deployment.

Future advancements and tendencies within the area have a number of abilities. To increase the effectiveness and generalizability of the fashions, the dataset used on this venture can be broadened and different. Deep studying and other modern gadget mastering techniques will also be investigated for the early detection of strokes. For the reason of enabling real-time monitoring and stroke detection, the fashions may also be included with different healthcare systems

REFERENCES

1.Minhaz Uddin Emon,Maria Sultana Keya,Tamara Islam Meghla, Md. Mahfujur Rahman,M Shami Al Mamun and M Shami Kaiserk - Performance Analysis of Machine Learning Approaches in Stroke Prediction,IEEE(2021)
2.Nugroho Sinung Adi, Richas Farhany, Rafidah Ghinaa, Herlina Napitupulu - Stroke Risk Prediction model using Machine learning, IEEE (2021)
3.Gang Fanga,Wenbin Liua, Lixin Wangb - A machine learning approach to select features important to stroke prognosis.ELSEVIER (2020)
4.Yafei Wu and Ya Fang-Stroke Prediction with Machine Learning Methods among Older Chinese,International Journal of Environmental research and Public health.(2020)
5.Chidozie Shamrock Nwosu, Soumyabrata Dev, Peru Bhardwaj, Bharadwaj Veeravalli and Deepu John - Predicting Stroke from Electronic Health Records,IEEE(2019).
6.Tianyu Liu, Wenhui Fan, Cheng Wu-A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset,Elsevier(2019)
7.M.Sheetal Singh,Prakash Choudhary-Stroke Prediction using ArtificialIntelligence,IEEE(2017)
8.Chen-Ying Hung, Wei-Chen Chen, Po-Tsun Lai, Ching-Heng Lin, and Chi-Chun Lee- Comparing Deep Neural Network and Other Machine Learning Algorithms for Stroke Prediction in a Large-Scale Population-Based Electronic Medical Claims Database,IEEE(2017)
9.SolmazNorouzi,RamazanFallah,AhmadPourdarvish,Seyed Morteza Shamshirgaran-Survival Analysis of Patients with Brain Stroke in the Presence of Competing Risks: A Weibull Parametric Model,Part of Journal of Biostatistics and Epidemiology 7(3):204-212,October 2021
10. Valery L. Feigin, Bo Norrving, George A. Mensah-Global Burden of Stroke ,Available:https://www.ahajournals.org/doi/10.1161/CIRCRESAHA.116.308413 Volume 120,Issue 33 February 2017[Accessed: January 20th,2023]
11.Amelia K. Boehme, Charles Esenwa, Mitchell S.V. Elkind - Stroke Risk Factors, Genetics, and Prevention, Available:https://www.ahajournals.org/doi/10.1161/CIRCRESAHA.116.308398 Volume 120, Issue 33 February 2017[Accessed: January 20th,2023]
12.Fedesoriano,"Stroke Prediction Dataset", Kaggle[online dataset],Available:https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset [Accessed: March 10th,2023]