CSE3013 ARTIFICIAL INTELLIGENCE

REVIEW 3 PROJECT REPORT

on

BRAIN STROKE PREDICTION AND ANALYSIS.

Prepared by

Harsh Rajpal - 20BCI0271
Rohan Gupta – 20BCI0260
Mrinal Sharma – 20BCI0247
Sasmita Singh- 20BCE2226

Under the supervision of

Dr. Annapurna Jonnalagadda

VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

School of Computer Science and Engineering
Vellore Institute of Technology, Vellore.

November 17, 2022

Table of Contents

1. Problem Statement

In a medical condition known as a stroke, the blood vessels in the brain are ruptured, harming the brain. Symptoms may appear when the brain's flow of blood and other nutrients is disrupted. The World Health Organization (WHO) claims that stroke is the leading cause of death and disability worldwide. Early detection of the numerous stroke warning symptoms can lessen the stroke's severity.

2. Introduction

2.1 Motivation:

The World Health Organization (WHO) claims that Stroke is the leading cause of death and disability worldwide. Many patients die on reaching the hospital, and those who reach the hospital have significantly less time left. Doctors can't predict Stroke for every patient in a given time; hence an ML Model will help them predict it sooner. We will analyse various algorithms and find the best-performing algorithm for this task with a high accuracy rate.

2.2 Significance:

We will use the open-access Stroke Prediction dataset to find humans' most important reasons for stroke. We will test our dataset using a range of physiological parameters and machine learning algorithms to discover which model shows the highest precision in stroke prediction.

2.3 Scope and applications:

Over 13 million people worldwide experience a stroke each year, and 5.5 million die from one, with these figures rising sharply each year. Being able to detect stroke might alter everything. Since more information about additional factors that contribute to stroke would be available with access to this data, the models may be trained more effectively. Fast initial response will contribute to lowering the mortality rate in low- and middle-income nations, where it's estimated that two out of every three citizens experience a stroke.

3. Literature Study

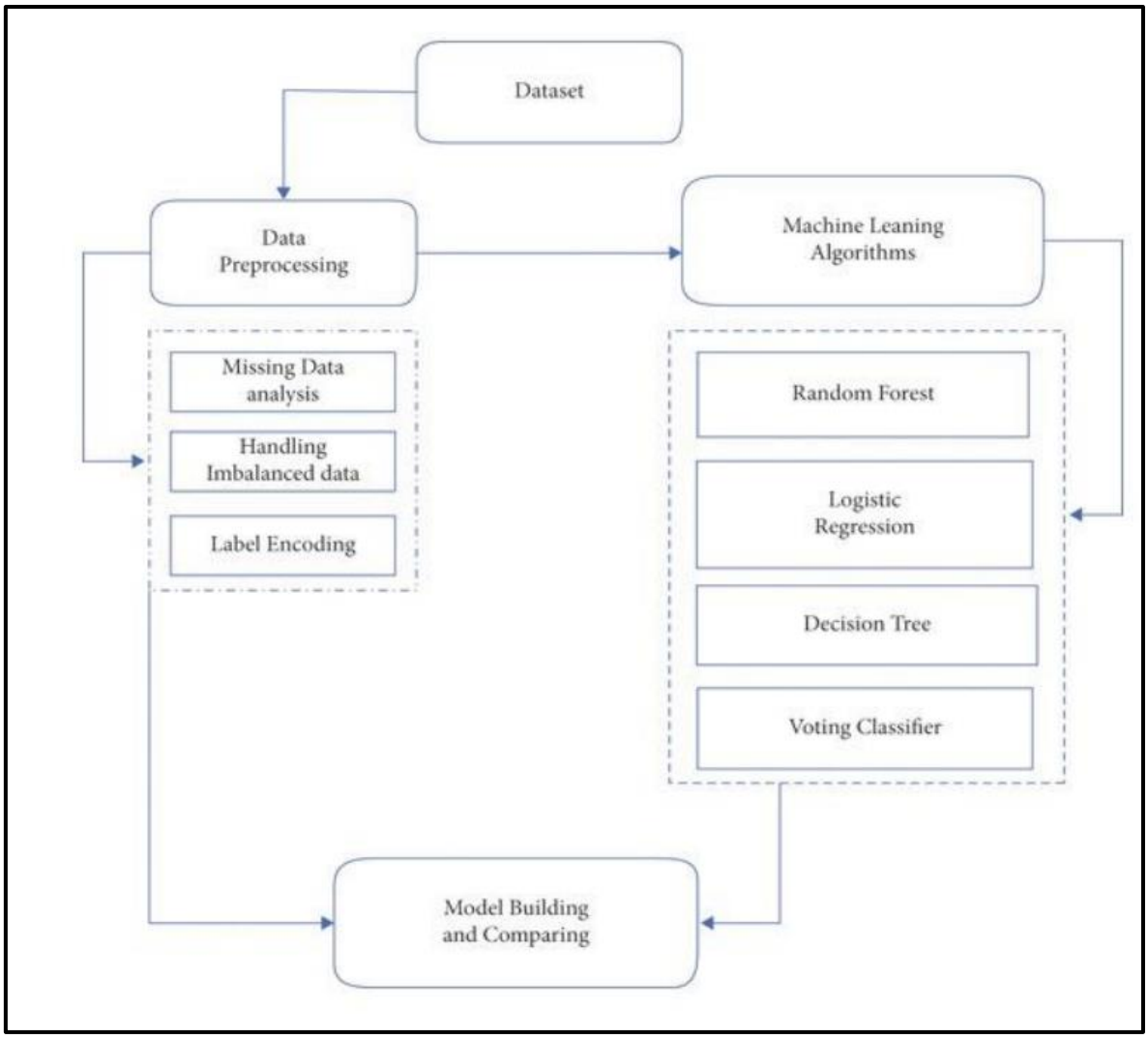
S.NO	Title of Paper	Summary	Gaps Identified in Paper
1	P. Garg, P. Kumar, K. Shakya, D. Khurana and S. Roy Chowdhury, “Detection of Brain Stroke using Electroencephalography (EEG)”	This paper focuses on identifying an approach to identify stroke non-invasively and cost-effectively using Electroencephalography (EEG).	The spatial resolution of EEG is a disadvantage since it is challenging to determine whether a signal originated close to the surface of the brain (in the cortex) or from a deeper location because the electrodes record electrical activity at the brain's surface.
2.	K. Sudharani, T. C. Sarma and K. Satya Prasad, "Brain stroke detection using K-Nearest Neighbor and Minimum Mean Distance technique”	In this paper the authors have proposed novel algorithm employing LabVIEW software and estimated the Identification score and Classification score and also the stroke area.	The fact that LabVIEW is a proprietary language is its main drawback.
3.	M. S. Hossain, S. Saha, L. C. Paul, R. Azim and A. Al Suman, "Ischemic Brain Stroke Detection from MRI Image using Logistic Regression Classifier"	In this research, a logistic regression classifier-based machine learning strategy for effective ischemic brain stroke diagnosis using magnetic resonance imaging (MRI) is provided.	The major limitation of Logistic Regression is the assumption of linearity between the dependent variable and the independent variables.
4.	P. Han, Z. Liu, Q. Li, M. Yu and R. Zhao, "Development of Realistic Numerical Brain Model Based on MRIs for Microwave Brain Stroke Detection,"	In this paper, a two-dimensional truth-oriented brain model is proposed for imaging simulation of wearable stroke detection devices. Stroke detection based on ultra-wideband microwave imaging is a new detection method, which is portable, fast and safe.	Since cellular activities in two-dimensional models take place on a flat monolayer surface rather than the three-dimensional orientation of the brain, they are unable to accurately represent the in vivo microenvironment of the human brain.

5.	I. Bisio, A. Fedeli, F. Lavagetto, M. Pastorino, A. Randazzo and A. Sciarrone, "Brain stroke detection by means of complex dielectric permittivity reconstruction at microwaves"	In this work, we propose a quantitative method for detecting haemorrhagic brain strokes through the reconstruction of the dielectric permittivity distribution inside head. The associated nonlinear inverse scattering problem is iteratively solved by means of an inexact-Newton method.	Experimental validation of the proposed method has not been done yet.
6.	A. Zamani, A. T. Mobashsher, B. J. Mohammed and A. M. Abbosh, "Microwave imaging using frequency domain method for brain stroke detection"	The scattered electric field and power inside the imaging zone with an elliptical shape are predicted using an image reconstruction technique based on the Mathieu function. In this paper, the system configuration, imaging technique, and results are given.	Mathieu functions take significant energy and computations to solve.
7.	M. A. Shokry and A. M. M. A. Allam, "Planar spiral antenna for brain stroke detection"	Spiral antenna is designed to detect brain stroke. It operates in the MedRadio spectrum. The antenna is simulated using CST microwave studio and fabricated on Rogers 4350 of thickness 1.524 mm, relative permittivity of 3.66 loss tangent of 0.04 S/m and operates at 426.6 MHz. It is placed on the external surface of the human's head to detect brain stroke. It is measured on a real human's head using the network analyzer.	Reduced aperture efficiency, which results in lower gain for the specified aperture size, is one well-known restriction. This is because spiral antennas are inherently frequency-independent. The electrical size of the aperture rises proportionally with frequency.
8.	V. K. Dubey, M. Raj, A. K. Sahwal, K. Murari and M. K. Saha, "Brain Hemorrhagic Stroke Detection by Image Processing"	The proposed method consists of several steps which include image pre-processing and segmentation, feature extraction, and classification	These include issues such as the handling of image uncertainties that cannot be otherwise eliminated, including various sorts of information that is incomplete, noisy, imprecise, fragmentary, not fully reliable, vague, contradictory, deficient, and overloading

.

4. Implementation

4.1 Architecture:



4.2 Algorithm:

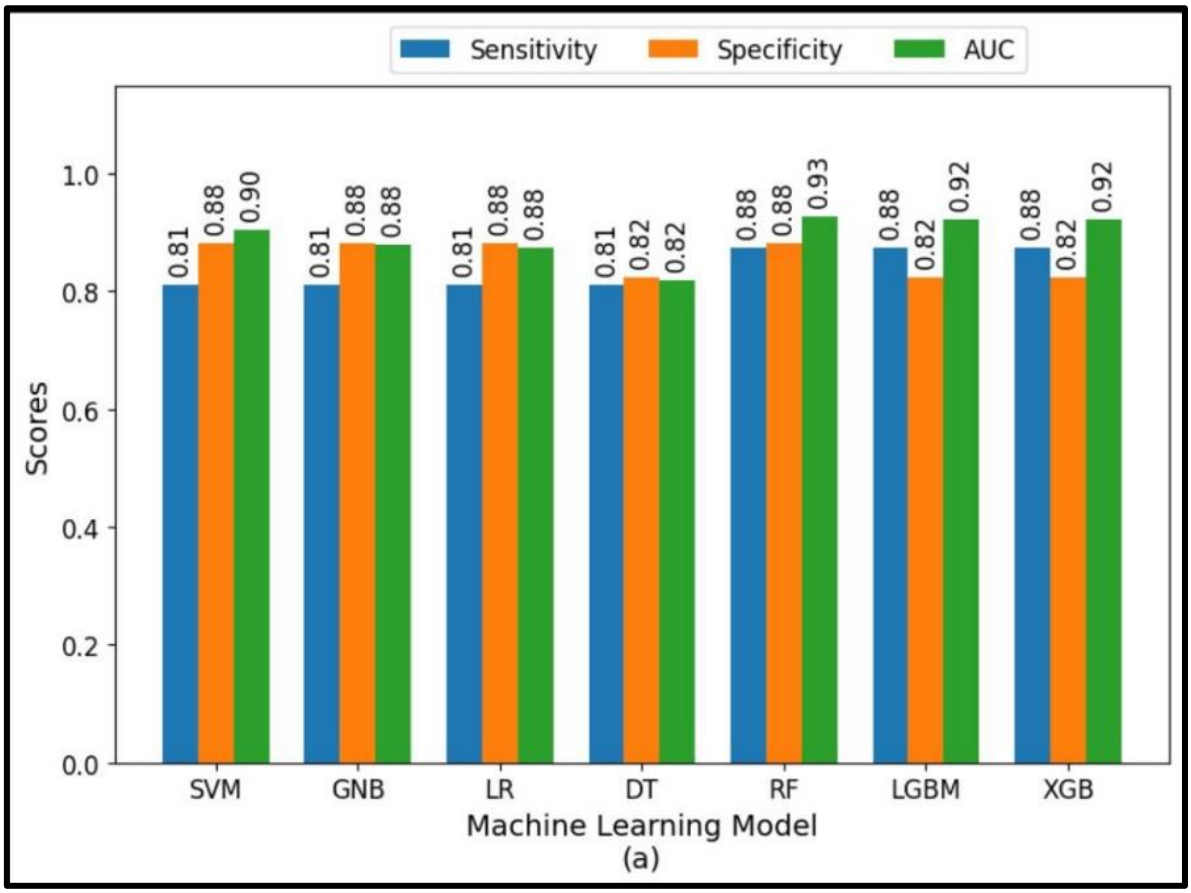
The dataset present in Kaggle for stroke expectation was exceptionally imbalanced. The dataset has a sum of 5110 lines, with 249 lines showing the chance of a stroke and 4861 columns affirming the absence of a stroke. While utilizing such information to prepare a machine-level model might bring about precision, other exactness measures, for example, accuracy and review are lacking.

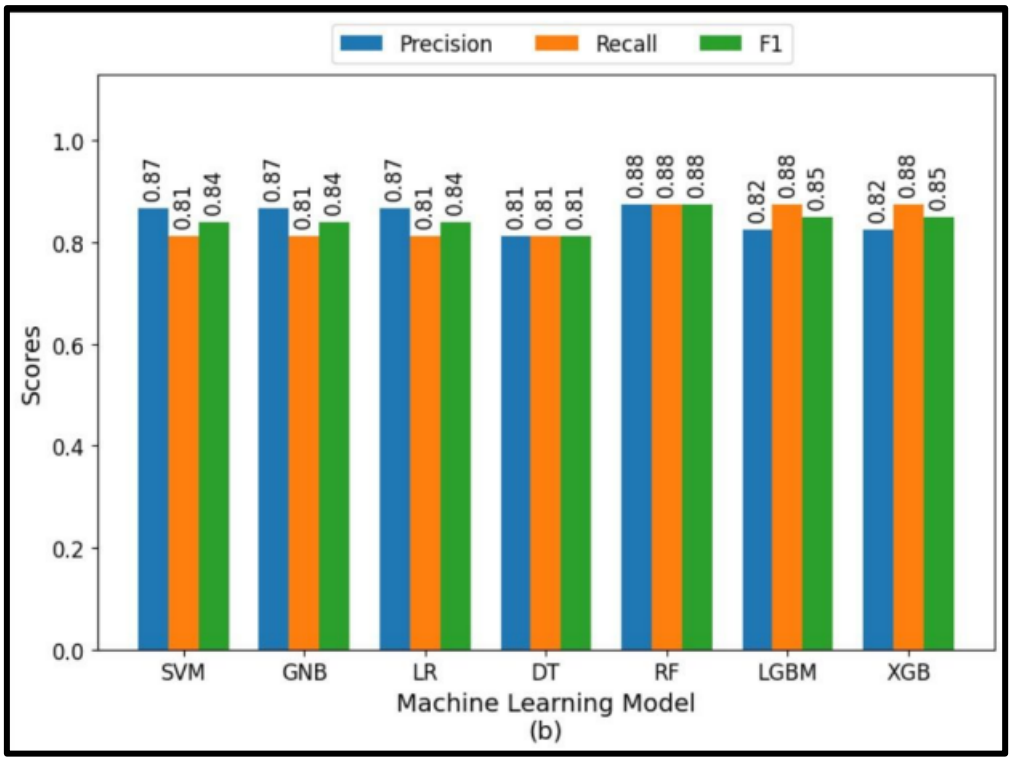
To prepare our calculations in an effective way we followed the methodology.

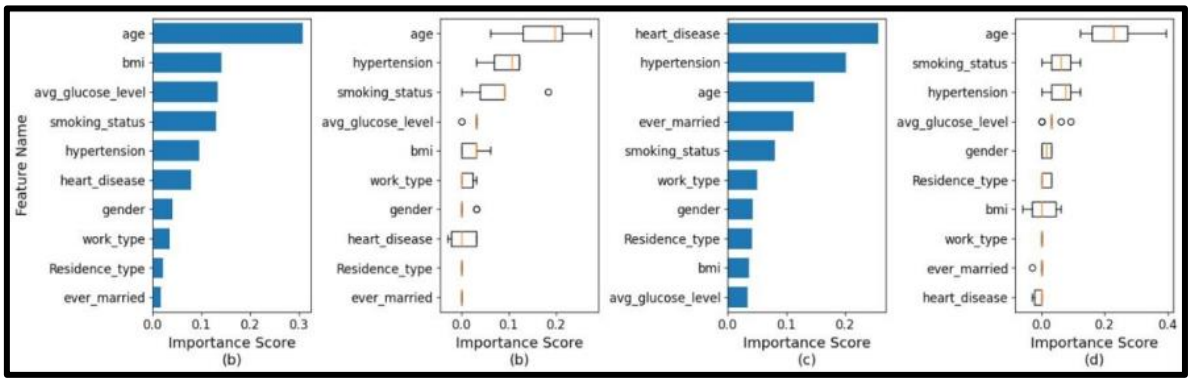
- Divide the information in 6:4, 60 % Train Information and 40 % Test Information.
- Train Information is then up sampled involving up sample capability in CARET library.

5. Result Analysis

According to empirical findings, the XGB model performed the best, followed by the RF model, as shown in figure 1. Figure 2 displays the feature significance scores for these two top-performing models. Age was consistently the most crucial characteristic for prediction across various models, and feature significance approaches. After "age," there were differences in the relevance of the features. Figures 2(a) and 2(c) show that the permutation-based feature importance calculation from the test data and the tree-based feature importance calculation from the training data both provided higher degrees of difference. The greater degree of divergence between XGB and RF algorithms' basic principles (behavior) and tree-based feature importance may be the cause. These findings suggest that a more reliable strategy for identifying the most beneficial or informative features uses permutation-based feature significance.



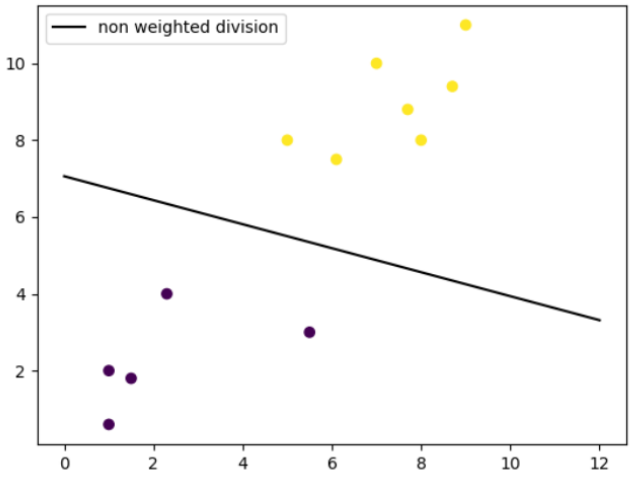

Precision
Recall
F1
Scores
1.0
0.8
0.6
0.4
0.2
0.0
0.87 0.81 0.84
0.87 0.81 0.84
0.87 0.81 0.84
0.81 0.81 0.81
0.88 0.88 0.88
0.82 0.88 0.85
0.82 0.88 0.85
SVM
GNB
LR
DT
RF
LGBM
XGB
Machine Learning Model
(b)


Feature Name
age
bmi
avg_glucose_level
smoking_status
hypertension
heart_disease
gender
work_type
Residence_type
ever_married
0.0 0.1 0.2 0.3
Importance Score
(b)
age
hypertension
smoking_status
avg_glucose_level
bmi
work_type
gender
heart_disease
Residence_type
ever_married
0.0 0.1 0.2
Importance Score
(b)
heart_disease
hypertension
age
ever_married
smoking_status
work_type
gender
Residence_type
bmi
avg_glucose_level
0.0 0.1 0.2
Importance Score
(c)
age
smoking_status
hypertension
avg_glucose_level
gender
Residence_type
bmi
work_type
ever_married
heart_disease
0.0 0.2 0.4
Importance Score
(d)

Machine Learning Model mechanics:

1. Singular vector machine

It can handle both classification and regression on linear and non-linear data. Making a straight line between two classes is how a straightforward linear SVM classifier functions. In other words, the data points on one side of the line will all be assigned to one category, while the data points on the other side of the line will be assigned to a different category. Making sense of all the machine learning lingo is made easier by using a 2-D illustration. In essence, you have a grid with some data points on it. You're attempting to categorise these data points, but you don't want to include any data in the incorrect category. In other words, you're looking for the line connecting the two points that are closest to one another while keeping the other data points apart. The support vectors you'll employ to locate that line are therefore provided by the two nearest data points. The decision boundary is the name of that line.

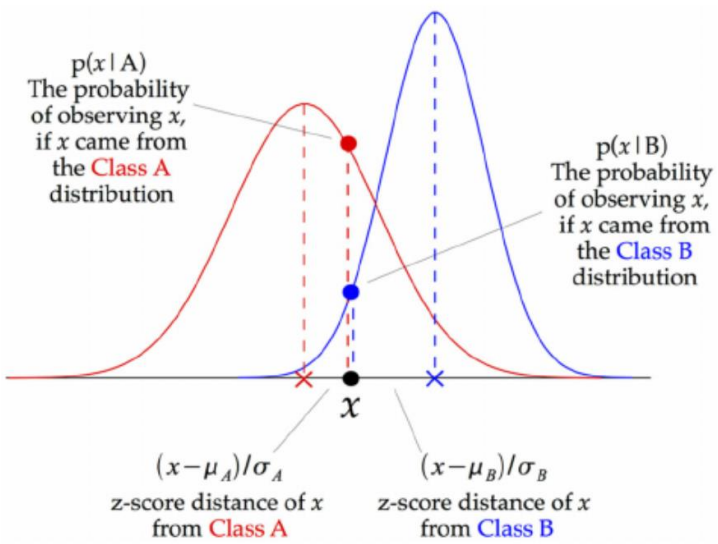


linear SVM

The boundaries of the decision can take any form. Because you may locate the decision boundary using more than two features, it is also known as a hyperplane.

2. Gaussian Naïve Bayes

A group of classification algorithms built on the Bayes' Theorem are known as naive Bayes classifiers. It is a family of algorithms rather than a single method, and they are all based on the idea that every pair of features being classified is independent of the other. Assuming that the data is described by a Gaussian distribution with no covariance (independent dimensions) between dimensions is one method for building a straightforward model. Finding the mean and standard deviation of the points within each label, which is all that is required to define such a distribution, will allow this model to be fit.



The Gaussian Naive Bayes (GNB) classifier is demonstrated in the image above. Every data point's z-score distance from each class mean, which is the distance from the class mean divided by the class's standard deviation, is calculated. As a result, we can see that the Gaussian Naive Bayes has a slightly different methodology and is effective.

3. Logistic Regression model

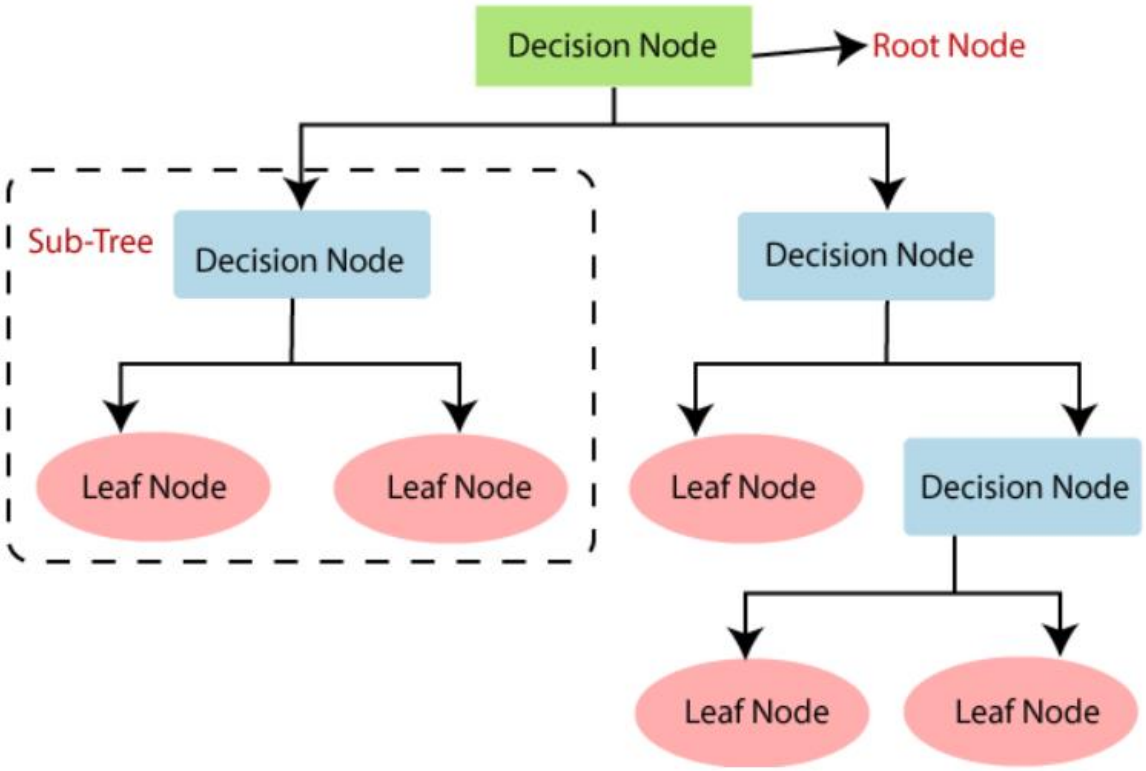
Machine learning uses the categorization method known as logistic regression. The dependent variable is modelled using a logistic function. Due to the dichotomous structure of the dependent variable, there are only two viable classes. This method is therefore employed while working with binary data. The sigmoid function is used in logistic regression to convert predicted values to probabilities. Any real value can be transformed into a value between 0 and 1 with this function. This function has exactly one inflection point and a non-negative derivative at each point.

A mathematical technique known as a cost function is used to calculate the difference between expected and forecasted values. A cost function is a way to quantify how inaccurate the model is in estimating the relationship between x and y. The cost, loss, or error terms are used to describe the value that the cost function returns. The cost function for logistic regression is represented by the following equation:

$$\text{Cost}(h_\theta(x), Y(\text{actual})) = -\log(h_\theta(x)) \text{ if } y=1$$
$$-\log(1-h_\theta(x)) \text{ if } y=0$$

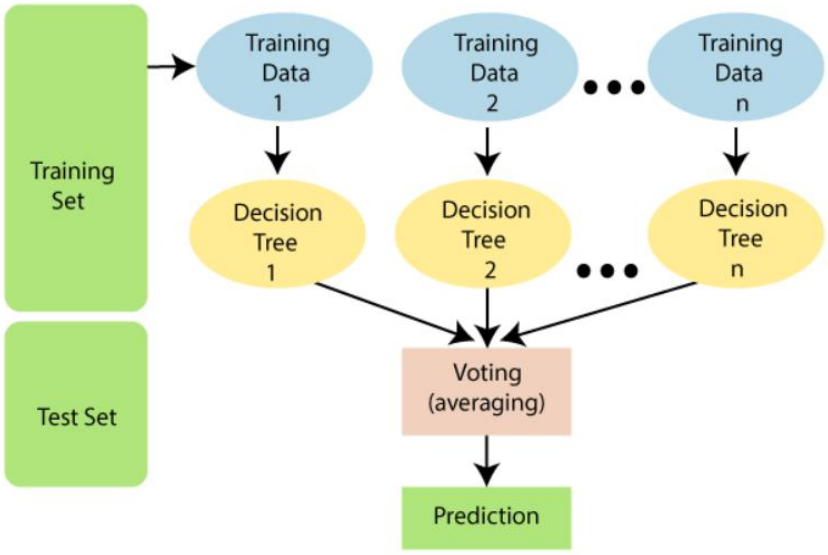
4. Decision Tree model

A supervised learning method called a decision tree can be used to solve classification and regression problems, but it is typically favoured for doing so. It is a tree-structured classifier, where internal nodes stand in for a dataset's features, branches for the decision-making process, and each leaf node for the classification result. It is a graphical representation for obtaining all feasible answers to a decision or problem based on predetermined conditions.



5. Random Forest model

Random Forest is a classifier that uses many decision trees on different subsets of the provided dataset and averages the results to increase the dataset's predicted accuracy. Instead, than depending on a single decision tree, the random forest uses forecasts from each tree and predicts the result based on the votes of the majority of predictions. It is based on the idea of ensemble learning, which is a method of combining various classifiers to address complex issues and enhance model performance. Higher accuracy and overfitting are prevented by the larger number of trees in the forest.



6. LightGBM

The documentation for LightGBM lists a lengthy set of parameters for this relatively new technique.

The dataset is getting bigger and bigger. Giving reliable results using conventional data science methods has gotten exceedingly challenging. Due of its rapid speed, Light GBM is preceded with Light. Large amounts of data can be handled using Light GBM, which requires less memory to operate.

The popularity of Light GBM is also attributed to its emphasis on precise outcomes. Data scientists frequently use LGBM to build data science applications since it also enables GPU learning.

LGBM should not be applied to tiny datasets. Light GBM is easily able to overfit little data and is susceptible to overfitting.

7. XGBoost

A particularly efficient and precise ML technique is XGBoost. But now LightGBM has challenged it, running even quicker with equivalent model accuracy and more user-tunable hyperparameters.The main reason for the speed difference is that LightGBM splits the tree nodes one node at a time, whereas XGBoost splits them one level at a time.

As a result, XGBoost's algorithmic engineers eventually made advancements to catch up to LightGBM, enabling users to run XGBoost in split-by-leaf mode (grow policy = "lossguide") as well. With this increase, XGBoost is now much faster, but LightGBM is still roughly 1.3–1.5 times as fast as XGB.

Another distinction between XGBoost and LightGBM is that XGBoost has monotonic constraint, whereas LightGBM does not. The model interpretability may be enhanced, but some model accuracy will be lost and training time will increase.

6. Future Work

- Over 13 million people worldwide experience a stroke each year, and 5.5 million die from one, with these figures rising sharply each year.
- Being able to detect stroke might alter everything.
- Other risk factors for stroke include the use of cigarettes, physical inactivity, poor nutrition, dangerous alcohol consumption, atrial fibrillation, elevated blood lipid levels, genetic predisposition, and psychological factors.
- Since more information about additional factors that contribute to stroke would be available with access to this data, the models may be trained more effectively.
- A reliable data collection in this area would enable researchers to improve their algorithms and provide a live list of people who could be at high risk of suffering a stroke.
- Fast initial response will contribute to lowering the mortality rate in low- and middle-income nations, where it's estimated that two out of every three citizens experience a stroke.

7. References

[1] P. Garg, P. Kumar, K. Shakya, D. Khurana and S. Roy Chowdhury, "Detection of Brain Stroke using Electroencephalography (EEG)," 2019 13th International Conference on Sensing Technology (ICST), 2019, pp. 1-6, doi: 10.1109/ICST46873.2019.9047678

[2] K. Sudharani, T. C. Sarma and K. Satya Prasad, "Brain stroke detection using K-Nearest Neighbor and Minimum Mean Distance technique," 2015 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2015, pp. 770-776, doi: 10.1109/ICCICCT.2015.7475383.

[3] M. S. Hossain, S. Saha, L. C. Paul, R. Azim and A. Al Suman, "Ischemic Brain Stroke Detection from MRI Image using Logistic Regression Classifier," 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), 2021, pp. 763-767, doi: 10.1109/ICREST51555.2021.9331090.

[4] P. Han, Z. Liu, Q. Li, M. Yu and R. Zhao, "Development of Realistic Numerical Brain Model Based on MRIs for Microwave Brain Stroke Detection," 2021 IEEE 15th International Conference on Electronic Measurement & Instruments (ICEMI), 2021, pp. 212-216, doi: 10.1109/ICEMI52946.2021.9679514.

[5] I. Bisio, A. Fedeli, F. Lavagetto, M. Pastorino, A. Randazzo and A. Sciarrone, "Brain stroke detection by means of complex dielectric permittivity reconstruction at microwaves," 2017 IEEE MTT-S International Microwave Workshop Series on Advanced Materials and Processes for RF and THz Applications (IMWS-AMP), 2017, pp. 1-3, doi: 10.1109/IMWS-AMP.2017.8247391

[6] A. Zamani, A. T. Mobashsher, B. J. Mohammed and A. M. Abbosh, "Microwave imaging using frequency domain method for brain stroke detection," 2014 IEEE MTT-S International Microwave Workshop Series on RF and Wireless Technologies for Biomedical and Healthcare Applications (IMWS-Bio2014), 2014, pp. 1-3, doi: 10.1109/IMWS-BIO.2014.7032452.

[7] M. A. Shokry and A. M. M. A. Allam, "Planar spiral antenna for brain stroke detection," 2015 9th European Conference on Antennas and Propagation (EuCAP), 2015, pp. 1-4.

[8] V. K. Dubey, M. Raj, A. K. Sahwal, K. Murari and M. K. Saha, "Brain Hemorrhagic Stroke Detection by Image Processing," 2021 4th International Conference on Recent Trends in Computer Science and Technology (ICRTCST), 2022, pp. 359-363, doi: 10.1109/ICRTCST54752.2022.9781835.

[9] V. Krishna, J. Sasi Kiran, P. Prasada Rao, G. Charles Babu and G. John Babu, "Early Detection of Brain Stroke using Machine Learning Techniques," 2021 2nd International Conference on Smart Electronics and Communication (ICOSEC), 2021, pp. 1489-1495, doi: 10.1109/ICOSEC51865.2021.9591840.

[10] M. Wang, X. Yao and Y. Chen, "An Imbalanced-Data Processing Algorithm for the Prediction of Heart Attack in Stroke Patients," in IEEE Access, vol. 9, pp. 25394-25404, 2021, doi: 10.1109/ACCESS.2021.3057693.

[11] Tahia Tazin, Md Nur Alam, Nahian Nakiba Dola, Mohammad Sajibul Bari, Sami Bourouis, Mohammad Monirujjaman Khan, "Stroke Disease Detection and Prediction Using Robust Learning Approaches", Journal of Healthcare Engineering, vol. 2021, Article ID 7633381, 12 pages, 2021. https://doi.org/10.1155/2021/7633381