Brain Stroke Prediction Using Machine Learning Models

1st Emil Bluemax
CSE
PES UNIVERSITY
Bangalore, India
emil.bluemax@gmail.com

2nd J.P.DANIEL CHRISTOPHER
CSE
PES UNIVERSITY
Bangalore City, India
danielchristopher513@gmail.com

3rd Aditya Rajendra Khot
CSE
PES UNIVERSITY
Bngalore City, India
adityakhot55@gmail.com

***Abstract*—**
Stroke is the sudden death of brain cells due to inadequate blood flow and oxygen resulting from a blood clot obstructing an artery in the brain or a blood vessel rupturing. Most of the times early prediction of stroke can prevent save the person's life.

The NA values in the dataset have been imputed with median values. Correlation matrix was plotted among the attributes and there wasn't any significant correlations between the attributes. Hypertension,heart disease and stroke attributes were converted into string type to use get_dummy function to get multiple attribute splits in one hot encoded format. Random oversampling is done to fix the target under-sampling issue. Standard scaler is used to scale all attributes to common scale. The dataset is then split into test-train dataset in a 80-20 ratio. Multiple models (KNN,DT,RFC,XGB) are fitted and accuracy are calculated for evaluation. The proposed study has an accuracy of:

A Decision Tree Classifier (DTC):97.32%, XGBoost (XGB):98.04%,Random Forest classifier (RFC):99.33%, KNN Classifier :97.42%

***Index Terms*—Stroke, Machine learning, Classification, Data pre-processing, Confusion matrix,k-fold Cross-Validation**

I. Introduction

Stroke is a disease that affects the arteries leading to and within the brain. A stroke occurs when a blood vessel that carries oxygen and nutrients to the brain is either blocked by a clot or ruptures. According to the WHO, stroke is the 2nd leading cause of death worldwide.

Globally, 3% of the population are affected by sub-arachnoid hemorrhage, 10% with intracerebral hemorrhage, and the majority of 87% with ischemic stroke. 80% of the time these strokes can be prevented

Early detection of stroke is a crucial step for efficient treatment to be able to do that, Machine Learning (ML) is an ultimate technology which can help health professionals make clinical decisions and predictions.

Our specific Goal is to predict whether a person will incur brain stroke based on the entered parameters. These entered parameters are analysed by the ML model to predict the target.

II. Previous Work

Based on the literature review conducted (All the papers were using the same dataset we are using) Various pre-processing methodologies were seen such as imputing the N/A values using the mean and 0.
For handling the categorical values Label-Encoding was done. Some scaling and normalizing was done for numeric attributes Various graphs and metrics were displayed which helps us better comprehend the relationship amongst the variables
Data oversampling was done due to the huge imbalance in the target attribute between minority class and majority class This was handled using Synthetic Minority Oversampling Technique (SMOTE).
One paper focused on boosting the present prediction model using Ensemble learning and had developed a new algorithm focused on brain stroke prediction combining the results of the multiple models in the ensemble
The accuracy of the previous works were low. We aim to increase the accuracy of the prediction model. We also aim to bring in a Web UI based front end to receive the parameters from the user and show the predictions on the website

The limitation's we identified was that the target attribute in the dataset was highly under sampled. There were some N/A values and some outliers. There was very less correlation amongst the attributes in the dataset.

III. Pre-Processed Solution

A. Data Dictionary

dataset consist of 5110 people's information and now all the attributes are described:
age: This attribute means a person's age. It's numerical data.
gender: This attribute means a person's gender. It's categorical data.
hypertension: This attribute means that this person is hypertensive or not. It's numerical data.
work-type: This attribute represents the person work scenario.

It's categorical data.
residence-type: This attribute represents the person living scenario. It's categorical data.
heart-disease: This attribute means whether this person has a heart disease person or not. It's numerical data.
avg_glucose_1eve1: This attribute means what was the level of a person's glucose condition. It's numerical data.
bmi: This attribute means body mass index of a person. It's numerical data.
ever-married: This attribute represents a person's married status. It's categorical data.
smoking-Status: This attribute means a person's smoking condition. It's categorical data.
stroke: This attribute means a person previously had a stroke or not. It's numerical data.

Stroke is the target attribute and rest of the attribute are used as response class attributes.

B. Pre Processing

- The data pre-processing step deals with the missing values in the dataset, converts the numeric values into string type so that we can perform one hot encoding and also we handle the under sampling of the target attribute.
- we observe that there is very low correlation among the attributes, the highest correlation observed was between age and BMI with a value of 0.32 all other correlation value's were less than 0.3
- These NAN values had been substituted by the median value of the BMI which was 36.6.
- Dummies functions were added using the get_dummies() method present in pandas library, It converts categorical data into dummy or indicator variables. A dummy variable is a numeric variable that encodes categorical information similar to one hot encoding
  In a dummy variable:
  A 1 encodes the presence of a category A 0 encodes the absence of a category If there more than one category it creates a separate column
- During EDA, it was found that the dataset used in this paper had only 249 entries of people who suffered stroke and 4861 people didn't have a stroke, making the dataset highly imbalanced with only about 4.8% of the total entries of minority class stroke. If machine learning algorithms are applied on such dataset it would have resulted in poor performance on minority class whose performance is the most important
- In order to overcome this problem of imbalanced classes, Random Oversampling is used.
- Random Oversampling includes selecting random examples from the minority class with replacement and supplementing the training data with multiple copies of this instance, hence it is possible that a single instance may be selected multiple times.

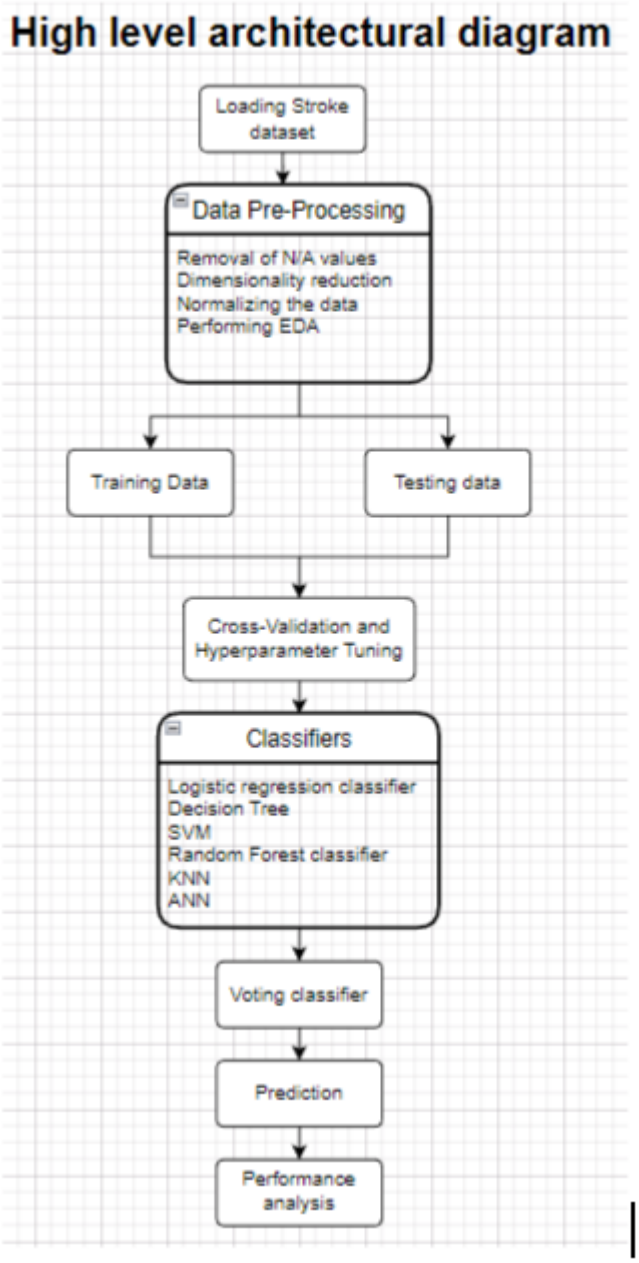

Fig. 1. Architectural diagram

C. Building a model - Classifier's used

The classifiers used were Decision Tree Classifier, K-Nearest Neighbours, XGBoost Classifier, and Random Forest Classifier.
All these classifiers are well known, and we can compare our results with the previous works.
We use 80% of the instances of our dataset to train our algorithm, and the remaining 20% of the data is considered to assess the trained model.
For ML model validation, we use k-fold cross validation process. In the k-fold cross validation technique, total dataset

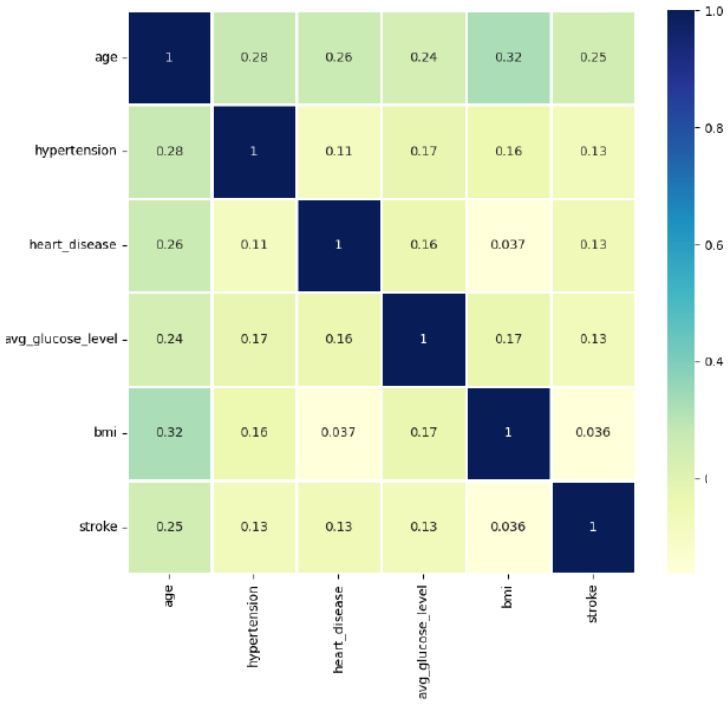

Fig. 2. Correlation Matrix

is employed to training and testing the classification process. The dataset is split into k parts aka folds. In training procedure, this process uses k-1 folds to train ML model and one-fold is employed to test model. This process is repeated k times, and every fold can be considered as test data-set. Advantage of this technique, that the all samples within the data-set are used for train and test, which removes the high variance.
Confusion matrices are used to evaluate the model's performance by calculating accuracy, Area Under the Curve, precision, f-1 score and Receiver Operator Curve . Analyzing these values, we find out the best model to predict stroke.

D. K-Nearest Neighbours Algorithm

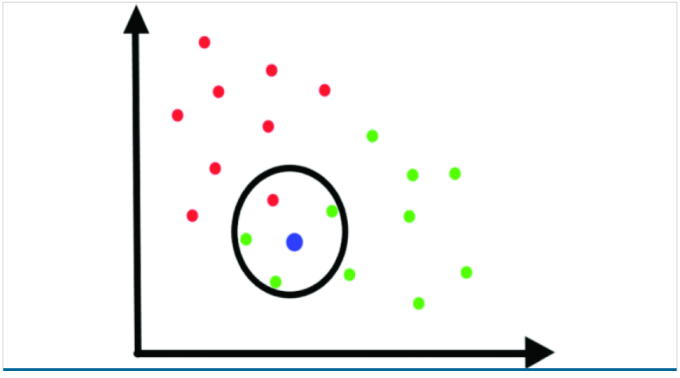

Fig. 3. Simple KNN

KNN checks the classes of a chosen number of training data samples which surround a test data sample, so as to make a prediction on which class the test data sample belongs to. k denotes the number of the nearest data samples, i.e. the neighbors.

KNN classifies the new unlabeled data by determining which classes its neighbors belong to. KNN algorithm utilizes this concept in its calculation.

When a new instance is encountered, KNN does two operations. 1 : finds K points closest to the new data point. 2 : using the neighbors classes, KNN determines as to which class the new data should be classified into.

the Euclidean distance needs to be calculated between the test sample and the specified training samples.

$$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Fig. 4. Euclidean distance

E. Decision Tree Algorithm

In general, the steps of the CART algorithm in making a decision tree are:

- Choose an attribute as the root.
- Divide cases into branches
- Repeat the process on each branch until all cases in the branch have the same class.

To select an attribute as the root, the highest gain value is based on the existing attributes. entropy can be calculated using the formula as shown in equation 1 below.

$$Entropy(s) = \sum_{i=1}^{n} P_i * \log_2 Pi$$

Fig. 5. Decision tree eq-1

$$Gain(S, A) = Entropy\ (S) - \sum_{i=1}^{n} \frac{|S_i|}{|s|} * Entropy\ (S_i)$$

Fig. 6. decision tree eq-2

F. Random Forest Algorithm

A Random forest consists of a combination of decision trees in which each tree rests on a random vector value sampled independently and with the same distribution for all trees in the forest. each tree assigns a voting unit to the most popular class in input . The 1. **Random Forests Converge** Centralizing random forests by determining the margin function can determine more accurate results. If the classification ensemble is h1(x),h2(x),. . . ,hk(x) with a random training set from the random vector distribution Y,X. The margin can be determined by the following equation:

$$mg(X,Y) = av_k I\left(h_k(X) = Y\right) - \max av_k I\left(h_k(X) = j\right).$$

Fig. 7. Random forest eq-1

The indicator function is I(). The margin function is used to measure how far the average number of votes in Y,X for a class exceeds the average vote for other classes. The larger the margin obtained, the more accurate the classification results.

2. **Strength and Correlation** The upper bound on the random forest can be derived for generalization error by

$$PE^* \leq \bar{P}\left(1 - s^2\right)/s^2$$

Fig. 8. Random forest eq-2

G. XGboost Algorithm

XGBoost is a supervised learning algorithm based on ensemble trees. It aims at optimising a cost objective function composed of a loss function (d) and a regularization term ():

$$\Omega(\theta) = \underbrace{\sum_{i=1}^{n} d\left(y_i, \hat{y_i}\right)}_{Loss} + \underbrace{\sum_{k=1}^{K} \beta\left(f_k\right)}_{regularization},$$

Fig. 9. XGBoost eq-1

where yiˆ is the predictive value, n the number of instances in the training set, K is the number of trees to be generated and fk is a tree from the ensemble trees. The regularization term is defined as:

$$\beta(f_t) = \gamma T + \frac{1}{2}\left[\alpha \sum_{j=1}^{T} |c_j| + \lambda \sum_{j=1}^{T} c_j^2\right],$$

Fig. 10. XGBoost eq-2

where is the minimum split loss reduction, is a regularization term on the weight and c is the weight associated to each leaf. A greedy approach is performed to select the split that increases the most the gain.

H. Classification Metrics

IV. Questions from peer review

The TA's had asked to present the comparison between various classifiers used to come to the conclusion of choosing a certain classification model.
Hence we have presented the performance metrics of various models. This helps us choose the model with the highest accuracy .

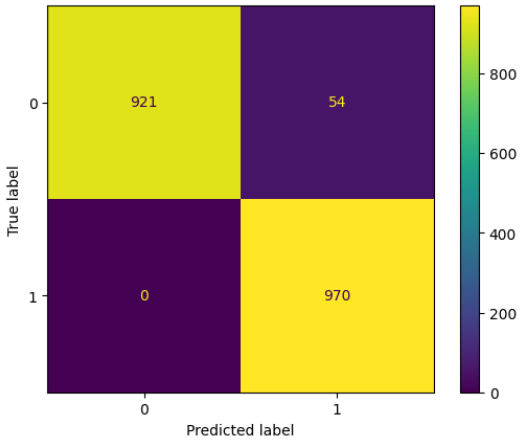

Fig. 11. K-Nearest neighbours Classifier

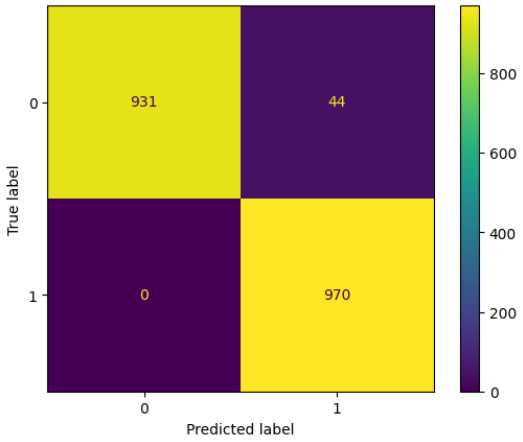

Fig. 12. XG boost confusion matrix

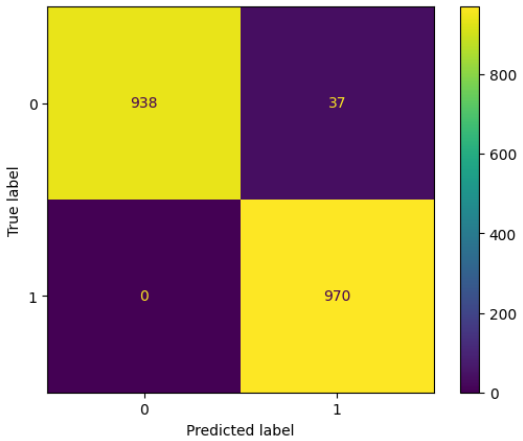

Fig. 13. Decision tree confusion matrix

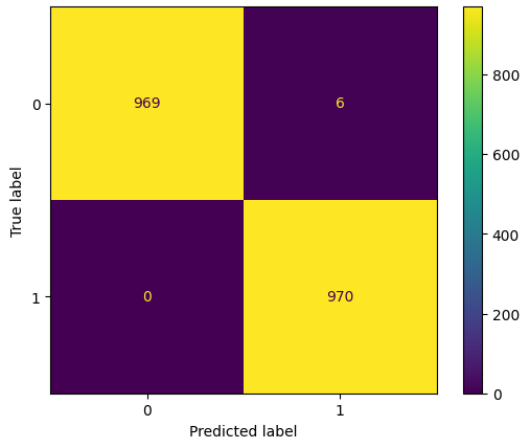

Fig. 14. Random forest confusion matrix

	Accuracy	AUC	Precision	ROC	F1 Score
KNN	97.42%	0.9743	0.9687	0.9743	0.96
XGBoost	98.04%	0.9983	0.9623	0.9983	0.98
Decision Tree	97.32%	0.974	0.94	0.974	0.94
Random Forest	99.33%	1.00	0.9500	1.00	0.97

Fig. 15. Performance Metrics

V. Experimental result and future scope

Random forest was chosen as the model used in the future predictions due to the high accuracy and high performance in other performance metrics.

The highest accuracy value was obtained by Random Forest with a result of 99.33%, Random Forest has the advantage in classifying data because it works for data that has incomplete attributes, and is good for handling large sample data. It also obtained ROC and AUC score of 1 with Precision of 0.9500 and F1 Score : 0.97

The model fails when invalid (semantically) values are entered eg. for age,bmi,avg glucose levels. Increasing the number of trees usually increase the performance of the model but it becomes quite slow making it a little hard to use in real time. The model to run also requires a lot computational powers as it requires to build many decision trees and run/evaluate them parallely. The model works well when the values for the attributes are entered in their valid ranges.

Further Deep Learning Models can also Be used in order to predict the risk of stroke more effectively without performing clinical test.

We can try using other classifying models like voting classifier, Since the proportion of positive and negative brain stroke cases are highly imbalanced some transformations and boosting needs to be done to represent equal representation

Propose a framework that uses brain Magnetic Resonance Imaging (MRI) with deep learning to improve the state of the art. It exploits advancements in deep learning to improve brain stroke prediction performance further.

VI. Contributions

Emil Bluemax : Model building and selection, Introduction

J P Daniel Christopher: Pre-Processing,Feature Extraction and Splitting into test and train.

Aditya Rajendra Khot : Abstract, Performance metrics

References

[1] S. Gupta and S. Raheja, "Stroke Prediction using Machine Learning Methods," 2022 12th International Conference on Cloud Computing, Data Science Engineering (Confluence), 2022, pp. 553-558, doi: 10.1109/Confluence52989.2022.9734197.

[2] N. S. Adi, R. Farhany, R. Ghina and H. Napitupulu, "Stroke Risk Prediction Model Using Machine Learning," 2021 International Conference on Artificial Intelligence and Big Data Analytics, 2021, pp. 56-60, doi: 10.1109/ICAIBDA53487.2021.9689740.

[3] M. U. Emon, M. S. Keya, T. I. Meghla, M. M. Rahman, M. S. A. Mamun and M. S. Kaiser, "Performance Analysis of Machine Learning Approaches in Stroke Prediction," 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020, pp. 1464-1469, doi: 10.1109/ICECA49313.2020.9297525.

[4] R. Islam, S. Debnath and T. I. Palash, "Predictive Analysis for Risk of Stroke Using Machine Learning Techniques," 2021 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2), 2021, pp. 1-4, doi: 10.1109/IC4ME253898.2021.9768524.

[5] A. Devaki and C. V. G. Rao, "An Ensemble Framework for Improving Brain Stroke Prediction Performance," 2022 First International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), 2022, pp. 1-7, doi: 10.1109/ICEEICT53079.2022.9768579.

[6] 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.

[7] M. Sheetal singh, Prakash choudhary, " Stroke Prediction using Artificial Intelligence ", 8th Annual Industrial Automation and Electromechanical Engineering conference(IEMECON) 2017 DOI: 10.1109/IEMECON.2017.8079581.

[8] Tasfia Ismail Shoily, Tajul Islam, , Sumaiya Jannat, Sharmin Akter Tanna,Taslima Mostafa Alif, Romana Rahman Ema. " Detection of Stroke disease using Machine Learning Algorithms " 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT) DOI: 10.1109/ICCCNT45670.2019.8944689

[9] V. J. Jayalaxmi, V geetha, M. Ijaz, " Analysis and Prediction of Stroke using Machine Learning Algorithms " 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA) — 978-1-6654-2829-3/21/$31.00 ©2021 IEEE — DOI: 10.1109/ICAECA52838.2021.9675545

[10] I. L. Cherif and A. Kortebi, "On using eXtreme Gradient Boosting (XGBoost) Machine Learning algorithm for Home Network Traffic Classification," 2019 Wireless Days (WD), 2019, pp. 1-6, doi: 10.1109/WD.2019.8734193.

[11] . Taunk, S. De, S. Verma and A. Swetapadma, "A Brief Review of Nearest Neighbor Algorithm for Learning and Classification," 2019 International Conference on Intelligent Computing and Control Systems (ICCS), 2019, pp. 1255-1260, doi: 10.1109/ICCS45141.2019.9065747.