IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 12, Issue 4, April 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204301 |

Stroke Prediction and Analysis by Using Machine Learning Models

R.Keerthivasan, M.Harrishwar, Dr. D. C. Joy Winnie Wise

UG Students, Dept. of Computer Science And Engineering. Rajalakshmi Institute Of Technology, Chennai, Tamil Nadu, India

Professor, Dept. of Computer Science And Engineering, Rajalakshmi Institute Of Technology, Chennai, Tamil Nadu, India

ABSTRACT: Low blood and oxygen levels can cause brain cells to suddenly die, a condition known as a stroke. There are two potential explanations for this: a blood clot obstructing a cerebral artery or a blood vessel rupture. Early diagnosis is crucial to preventing stroke deaths. To fill up the gaps in the dataset, the median values were added. There were no discernible links between the qualities in the correlation matrix that was constructed between them. To facilitate modeling, attributes such as heart disease, stroke, and hypertension were converted into string types and stored in a one-hot encoded manner using the get dummy function. The undersampling of the target was made up for by random oversampling.

To ensure uniformity, standard scaling was applied to all attributes. The dataset was then split into a training set and a test set in an 80-20 ratio. Various models, including K-Nearest Neighbors (KNN), Decision Tree Classifier (DTC), Random Forest Classifier (RFC), and XGBoost (XGB), were fitted to the data. The accuracy of each model was calculated for evaluation.

The results of the study indicate high accuracy rates for the models: A Decision Tree Classifier (DTC):97.32%, XGBoost (XGB):98.04%,Random Forest classifier (RFC):99.33%, KNN
Classifier :97.42%

KEYWORDS: Stroke, Machine learning, Classification, Data pre-processing, Confusion matrix,k-fold Cross-Validation.

I. INTRODUCTION

Stroke is a disease that affects both the brain and the arteries[1] that supply it. A stroke may occur from a blood clot that bursts or becomes blocked in a blood vessel supplying the brain with nutrition and oxygen. According to the World Health Organization, stroke is the second most prevalent cause of death worldwide.Worldwide, 3% of persons suffer from subarachnoid hemorrhage, 10% from intracerebral hemorrhage, and 87% from ischemic stroke. In eighty percent of cases, these strokes can be prevented.Early detection is critical for the effective treatment of stroke. Machine learning (ML) is one of the best technologies accessible to medical experts for this goal, allowing them to make clinical judgments and predictions.Our specific goal is to predict the probability of a brain stroke based on the entered data. To forecast the target[2], the ML model examines these entered factors.

II. PREVIOUS WORK

According to the results of the literature review (every paper used the same dataset that we do the literature review that was done revealed different pre-processing techniques[2] (all the studies used the same dataset we are utilizing), such as imputing the N/A values using the mean and 0.

In order to handle the categorical data, label encoding was done. Some scaling and normalizing was done for the numerical properties. To help us comprehend how the variables relate to one another[3], a variety of graphs and metrics were displayed. Oversampling of the data was required due to the significant difference in the target attribute between the minority class and the majority class[4]. This was handled using the Synthetic Minority Oversampling Technique (SMOTE).

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204301 |

Using ensemble learning, a novel algorithm for brain stroke prediction was developed in one paper by combining the results of several models in the ensemble[2]. This new method was then used to enhance the existing prediction model. The accuracy of the earlier works was lacking. Our objective is to increase the prediction model's accuracy. We also intend to incorporate a Web UI front end so that the predictions can be shown on the website and user parameters can be gathered.

A constraint identified in the dataset pertained to the notable undersampling of the target attribute. There were both N/A values and outliers. There was not much of a relationship between the attributes in the dataset[1].

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 characteristic indicates whether or not the individual has hypertension. It's numerical data.
work-type: This attribute represents the person work scenario.
residence-type: This attribute represents the person living scenario. The data are categorical.
heart-disease: This attribute means whether this person has a heart disease person or not. It's numerical data.
avg glucose level: 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. The data are categorical.
smoking-Status: This attribute means a person's smoking condition. The data are categorical.
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:

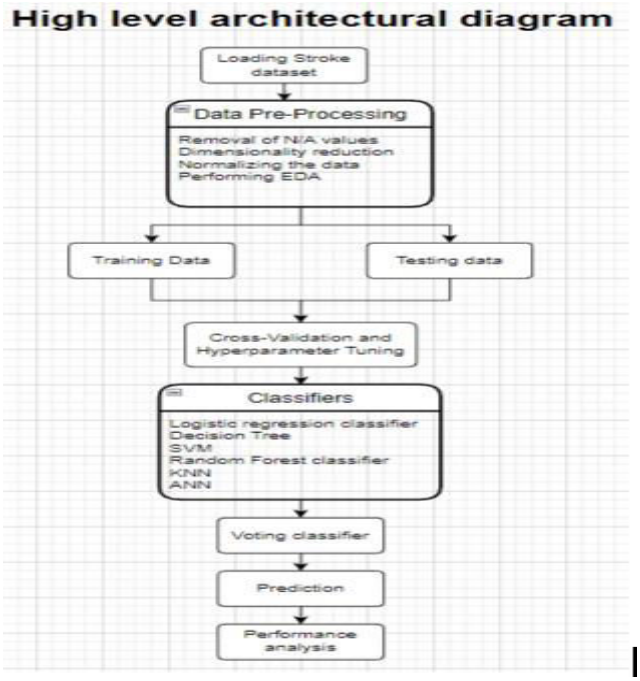
- The pre-processing step of the data handles the missing values in the dataset, converts numeric values into string form so that a single hot encoding may be performed[5], and corrects undersampling of the target property.
- We discover that there is only a small correlation (correlation coefficient of 0.32) between the variables; the only correlation that is higher than 0.3 is between age and BMI.
- These NAN values were substituted with the median BMI of 36.6.
- To add dummies functions, the acquire dummies() function from the pandas package was utilized. This function converts categorical data into dummy or indicator variables[4]. A dummy variable is a numerical variable that encodes categorical data, much like one hot encoding.
- With just 249 entries representing stroke patients and 4861 entries representing non-stroke patients[5], it was found through EDA that the dataset used in this work was highly skewed. This indicates that stroke cases in the minority class accounted for just 4.8% of all dataset items[6]. Using machine learning algorithms on this dataset would have resulted in low performance for the minority class, whose performance is the most important.
- It is conceivable for a single instance to be chosen more
- than once because Random Oversampling involves replacing random examples from the minority class with duplicates.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204301 |



C. Building a model – Classifier's used

The ones used were the Random Forest Classifier, XGBoost Classifier, K-Nearest Neighbors, and Decision Tree Classifier. Given the popularity of these classifiers, we can compare our results to those of previous studies. We employ 80% of the dataset's instances to train our method, and the remaining 20% is used to assess the trained model[8]. For ML model validation, we use the k-fold cross validation process. The kfold cross validation method uses the full dataset.

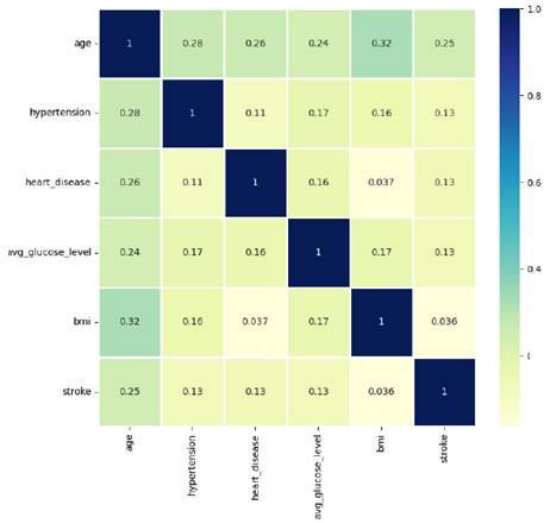


Fig. 2. Correlation Matrix

is used to test and train the classification procedure. The dataset is divided into k folds, or segments. The ML model is trained using k-1 folds in the training step, and the model is tested using one fold. Every fold can be thought of as a test data set, and this procedure is done k times[9]. One benefit of this strategy is that it eliminates[5] high variation because all samples in the data set are used for both training and testing. Confusion matrices are used to compute the accuracy, precision, f-1 score, area under the curve, and receiver operator curve in order to assess the performance of the model[4]. We determine the most accurate model for stroke prediction by analyzing these values.

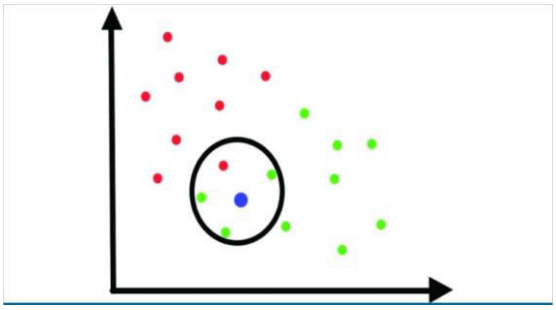
International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204301 |

D. K-Nearest Neighbours Algorithm



In order to predict which class a test data sample belongs to, KNN examines the classes of a certain number of training data samples that surround it. The number of the closest data samples[3], or neighbors, is indicated by the letter k.

KNN uses neighboring class membership to classify the newly unlabeled data. This idea is used by the KNN method in its computation.

When KNN encounters a new instance, it does two operations. 1. determine which K points are closest to the new data point. 2. KNN identifies which class the new data belongs in by utilizing the neighboring classes.
It is necessary to compute the Euclidean distance between the test sample and the designated training samples[6].

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

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.

The largest benefit value among the various attributes is utilized to select one attribute as the root. Equation 1 below provides the formula

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

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 is a collection of decision trees where each tree is based on a randomly distributed vector value that is individually sampled for every tree in the forest. Every tree allots a voting unit to the input's most popular class[9]. 1. Random Forests Come Together More accurate results can be obtained by centralizing random forests and figuring out the margin function. Given a random training set drawn from the random vector distribution Y,X, the classification ensemble consists of h1(x), h2(x),..., hk(x)[8]. The following formula can be used to calculate the margin:

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

Fig. 7. Random forest eq-1

I() is the indicator function. The average number of votes in Y,X for a class is greater than the average vote for other classes, and this difference is measured using the margin function[8]. The findings of the categorization are more accurate the greater the margin obtained.

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204301 |

Intenseness and Association For generalization error, the upper bound on the random forest can be obtained by

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

Fig. 8. Random forest eq-2

G. XGboost Algorithm

An algorithm for supervised learning based on semble trees is called XGBoost. Its goal is to optimize a cost objective function made up of a regularization term () and a loss function(d):

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

Fig. 9. XGBoost eq-1

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

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

Fig. 10. XGBoost eq-2

where is the weight associated with each leaf, is a regularization term on the weight, and is the minimum split loss reduction. A hedonistic method is employed to choose the split that maximizes the profit[7].

H. Classification Metrics

The teaching assistants requested that a comparison of the different classifiers utilized to determine[6] the classification model to select be presented. As a result, we have shown the various models' performance metrics. This aids in our selection of the most accurate model.

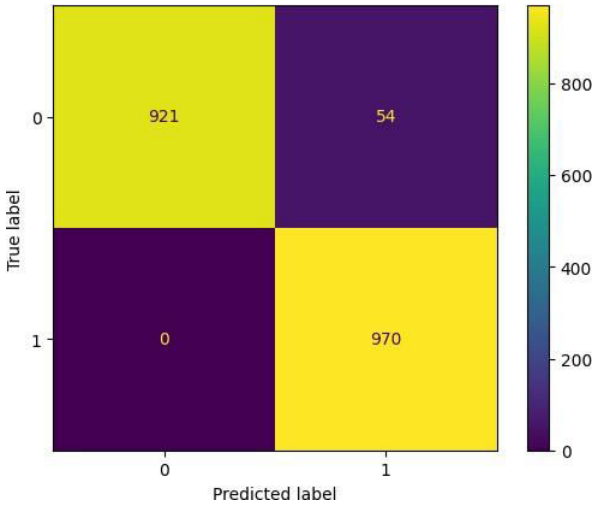


Fig. 11. K-Nearest neighbours Classifier

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204301 |

The k-nearest neighbors (k-NN) algorithm is a simple yet powerful supervised learning algorithm used for classification and regression tasks. It's based on the idea that data points with similar features tend to belong to the same class or have similar output values.

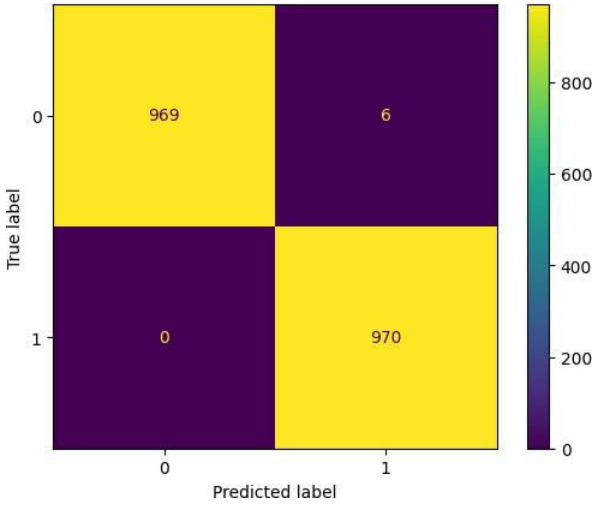


Fig. 12. XG boost confusion matrix

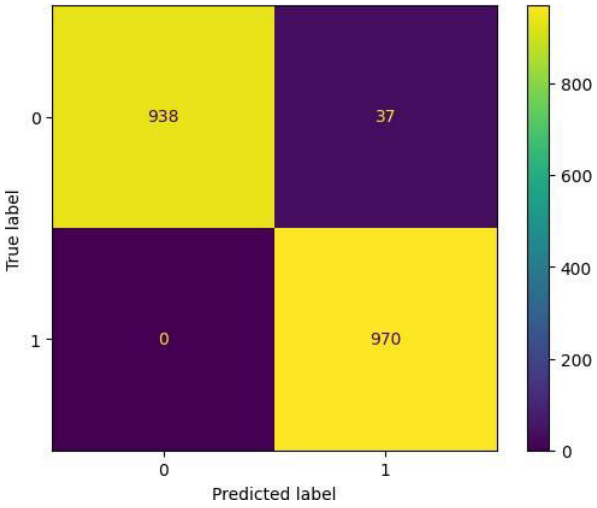


Fig. 13. Random forest confusion matrix

	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. 14. Performance Metrics

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 4, April 2024 ||

| DOI: 10.15680/IJIRCCE.2024.1204301 |

IV. EXPERIMENTAL RESULT ANF FUTURE SCOPE

Because of its excellent accuracy and performance in other performance parameters, random forest was selected as the model to be used in the upcoming forecasts. With a score of 99.33%, Random Forest achieved the greatest accuracy value. Random Forest has an advantage in data classification since it can handle huge sample sizes and[8] works well with partial attribute data. Additionally, it received an F1 score of 0.97, ROC and AUC scores of 1, and precision of 0.9500.When incorrect (semantically) values are input, such as age, BMI, or average blood sugar levels, the model breaks down. Although adding more trees usually improves the model's performance, doing so makes it slower and more difficult to operate in real time. Due to the need to construct numerous decision trees and run/evaluate them concurrently[9], the model's execution also demands a high processing capacity. When the attribute values are entered inside their valid ranges, the model functions as intended.Additional Deep Learning Models can also be utilized to more accurately forecast the risk of stroke without the need for clinical testing.Other classification models, such as the voting classifier, may be used. However, due to the extreme [7]imbalance between the proportion of positive and negative brain stroke cases, various adjustments and boosting must be performed to provide equal representation.Provide a framework to advance the state of the art that combines deep learning with brain magnetic resonance imaging (MRI)[9]. It takes advantage of deep learning breakthroughs to further enhance brain stroke prediction accuracy

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.5660,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. 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 Inter- national Conference on Computing, Communication and Networking Technologies (ICCCNT) DOI: 10.1109/ICCCNT45670.2019.894468910.1109/ICOSEC51865.2021.9591840.
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. *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.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details