Stroke Prediction using Machine Learning Algorithms

Aditi Agarwal

BCA Student, IITM

Abstract

The study "Stroke Prediction Using Machine Learning Algorithms" seeks to investigate the use of various machine learning algorithms in predicting the risk of stroke incidence. The study made use of a dataset that included patient demographics, medical history, lifestyle variables, and health indicators including blood pressure and glucose level.

To determine the best-performing method for stroke prediction, numerous machine learning techniques are used, including logistic regression, decision trees, XG Boost and KNN and Ensemble learning techniques like Max Voting and Averaging. The algorithms were evaluated using several measures such as accuracy, sensitivity, specificity, and F1 score.

According to the study's findings, the XG Boost algorithm outperforms all others in stroke prediction, with an accuracy of 94% and F1 score of 0.94. Age, hypertension, diabetes, smoking status, and prior stroke history were also revealed as significant predictors of stroke in the research.

Overall, the work demonstrates the potential of machine learning algorithms for stroke prediction and sheds light on the significance of various patient demographics, medical history, and lifestyle variables in predicting stroke incidence. The study's findings may help to build more accurate stroke prediction models and assist healthcare practitioners in identifying persons at high risk of stroke for prompt intervention and prevention.

Keywords: stroke, machine learning, logistic regression, random forest, XG Boost, KNN, Ensemble Learning

Introduction

A brain stroke, also known as a cerebrovascular accident, is a dangerous medical illness that happens when blood flow to the brain is disrupted by a clogged or burst blood vessel. A shortage of blood supply to the brain can result in brain cell death and lasting brain damage. Brain strokes can happen quickly and without warning, resulting in a variety of symptoms such as paralysis, trouble speaking, and cognitive impairment.

According to World Health Organization, “In low-and middle-income countries, which include those of the WHO South-East Asia Region, over 11 million strokes occur every year. This causes

4 million deaths annually, and leaves approximately 30% of survivors seriously disabled. For the 70% of survivors who recover, the likelihood of suffering further strokes is greatly increased."

Hence predicting the chances of stroke can not only avoid disability but death as well. The healthcare facilities can also provide better treatment.

Machine Learning can be used to forecast the incidence of a stroke as technology advances in the medical industry. Machine Learning algorithms are helpful in creating accurate predictions and providing proper analyses. Previous stroke research has mostly focused on heart stroke prediction. Brain stroke has received very little attention. This work is centered on utilizing Machine Learning to forecast the onset of a brain stroke. The essential components of the techniques utilized and the findings achieved are that XG Boost fared the best among the various classification algorithms tested, attaining a higher accuracy score.

To begin with this paper, a dataset with several physiological features as attributes is taken from Kaggle. These characteristics are then analyzed and utilized to make the final forecast. The dataset is initially cleansed and prepared for understanding by the machine learning model. This is known as Data Preprocessing. The dataset is examined for null values and filled accordingly. Data Visualization is done after that, using various types of graphs and finding correlation between features using a heatmap. Then, if necessary, label encoding is used to convert string values to integers, followed by one-hot encoding.

The dataset is divided into train and test data after Data Preprocessing. utilizing multiple Classification Algorithms, a model is then constructed utilizing this new data. The accuracy of each method is calculated and compared to choose the best-trained model for prediction.

Literature Survey

Stroke Prediction is an important field of study to avoid the number of deaths and disability amongst people. Both traditional risk prediction models and machine learning algorithms can be useful in identifying persons at high risk of stroke. However, studies suggest that these models require more validation and refining to increase their accuracy and usability in clinical practice.

Rizos T. et al. (2014) published "Predicting the risk of stroke in patients with atrial fibrillation: A systematic review" The CHA2DS2-VASc score was shown to be the most reliable for predicting stroke risk in individuals with atrial fibrillation in this study, which analyzed multiple risk assessment methods for stroke.

Hilkens N. A. et al. (2013) published "Prediction models for stroke: a systematic review" in 2013. This systematic research identified 39 distinct stroke prediction models and evaluated their accuracy and reliability. The scientists discovered that established risk indicators, such as age and blood pressure, showed a moderate to strong predictive value.

Nannoni S. et al. (2020) published "Machine learning for predicting stroke: a systematic review." The application of machine learning algorithms in stroke prediction was studied in this study, and

it was discovered that they can increase the accuracy of stroke risk assessment when compared to traditional models. However, the authors stressed that additional validation of machine learning models utilizing vast and varied datasets is required.

Liu M. et al. (2020) published "Comparison of risk prediction scoring systems for stroke: a systematic review." The Atherosclerotic Cardiovascular Disease Risk Score (ASCVD) and the Framingham Risk Score (FRS) were shown to be the most reliable tools for stroke risk prediction in this systematic analysis of the performance of several risk scoring systems.

Li Y. et al. (2019) published "Prediction of stroke risk in patients with atrial fibrillation using machine learning." The use of machine learning algorithms in predicting stroke risk in patients with atrial fibrillation was investigated in this study, and it was discovered that they beat standard risk prediction models. The authors suggested that machine learning might be a valuable tool in clinical practice for assessing stroke risk.

Stroke is a medical condition that causes brain damage by tearing blood vessels. The majority of research has focused on the prediction of heart attacks, but very few studies have focused on the risk of a brain attack. The papers used machine learning algorithms to train models for accurate prediction based on various physiological factors. Naïve Bayes, Random Forest and Multi-Layer Perceptron (MLP) classifier are some of the algorithms that gave the best results. Stroke is an acute neurological dysfunction caused by a central nervous system injury caused by decreased blood flow to the brain.

There is a pressing need to model the effect of various risk factors on stroke occurrence, and artificial intelligence (AI) appears to be the right tool.

System Methodology

A comprehensive literature search was performed using various databases, including PubMed, Scopus, and Web of Science. The search terms included "stroke prediction," "machine learning," "artificial intelligence," and "neural network." Studies published in the past five years were included in this review.

Database

The database is taken from Kaggle and it is used to predict whether the patient is likely to get a stroke or not based on various input parameters such as, age, gender, smoking status etc. In the current dataset, there are 11 features and one binary target which are considered as attributes for studying .

Attribute Information

1) id: unique identifier

2) gender: "Male", "Female" or "Other"

3) age: age of the patient

4) hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension

5) heart_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease

6) ever_married: "No" or "Yes"

7) work_type: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"

8) Residence_type: "Rural" or "Urban"

9) avg_glucose_level: average glucose level in blood

10) bmi: body mass index

11) smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"*

12) stroke: 1 if the patient had a stroke or 0 if not

Proposed System of study

The following diagram shows the proposed system of the study. The architecture contains the following stages :

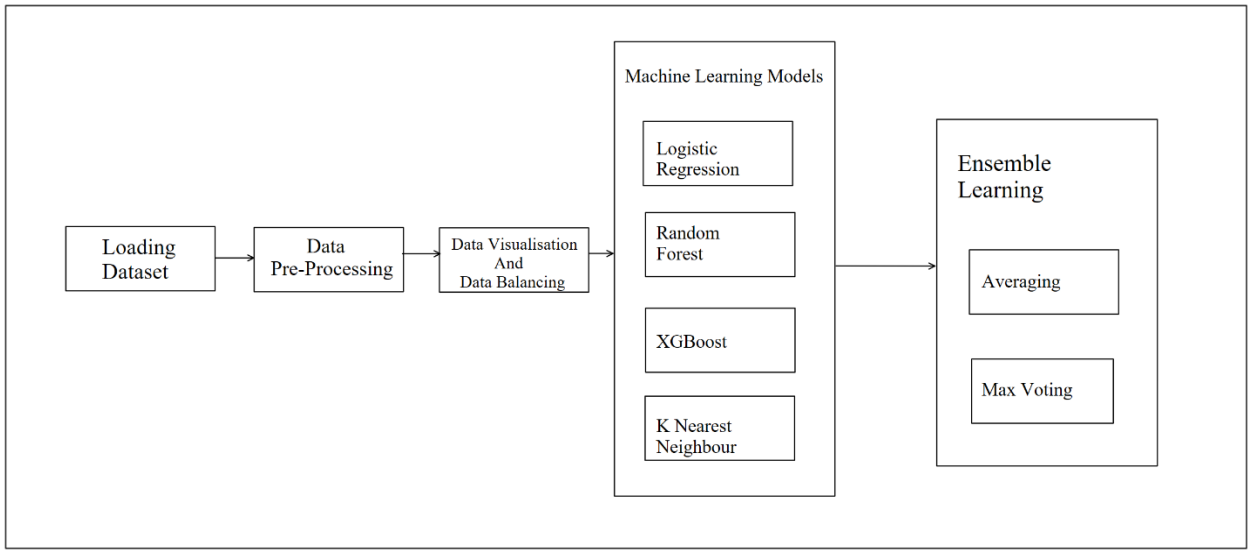

Figure 1: Proposed study

Data Pre-Processing

Data pre-processing is an important stage in the examination of research papers. It entails cleaning and translating raw data into an analysis-ready format. This procedure entails deleting unnecessary data, dealing with missing values, dealing with outliers, and normalizing the data.

After importing and reading the dataset, we check for null values amongst the 12 attributes. The only null values were present in the BMI column and hence we use mean values to fix it.

Furthermore, data pre-processing might include feature selection and feature engineering, which serve to minimize data dimensionality and increase analysis performance.

With the help of feature engineering, we make initial insights about the dataset:

- Mean age of people is around 43
- Mean BMI is more than normal
- Both Categorical and numerical features are present.

  Categorical Features: gender, ever_married, work_type, Residence_type, smoking_status

  Binary Numerical Features: hypertension,heart_disease, stroke

  Continuous Numerical Features: age, avg_glucose_level, bmi

- Most of the data is categorical which need a special attention for visualization.

Overall, data pre-processing is an important step in ensuring the correctness and dependability of study findings.

Data Visualization

Data visualization enables researchers to find patterns, trends, and correlations in data that might otherwise be invisible from raw data. It can also assist to emphasize crucial results and ideas, making it simpler to communicate complicated material to a larger audience.

Seaborn is an open-source Python library built on top of Matplotlib. It is used for visualization and exploratory data analysis. By importing seaborn library, various graphs such as countplot, distplot and scatterplots can be used which can be used to compare different attributes.

A correlation heatmap depicts the relationship between variables in a dataset graphically. The heatmap illustrates the correlation coefficients between each pair of variables, with the color of the cell indicating the strength of the correlation.

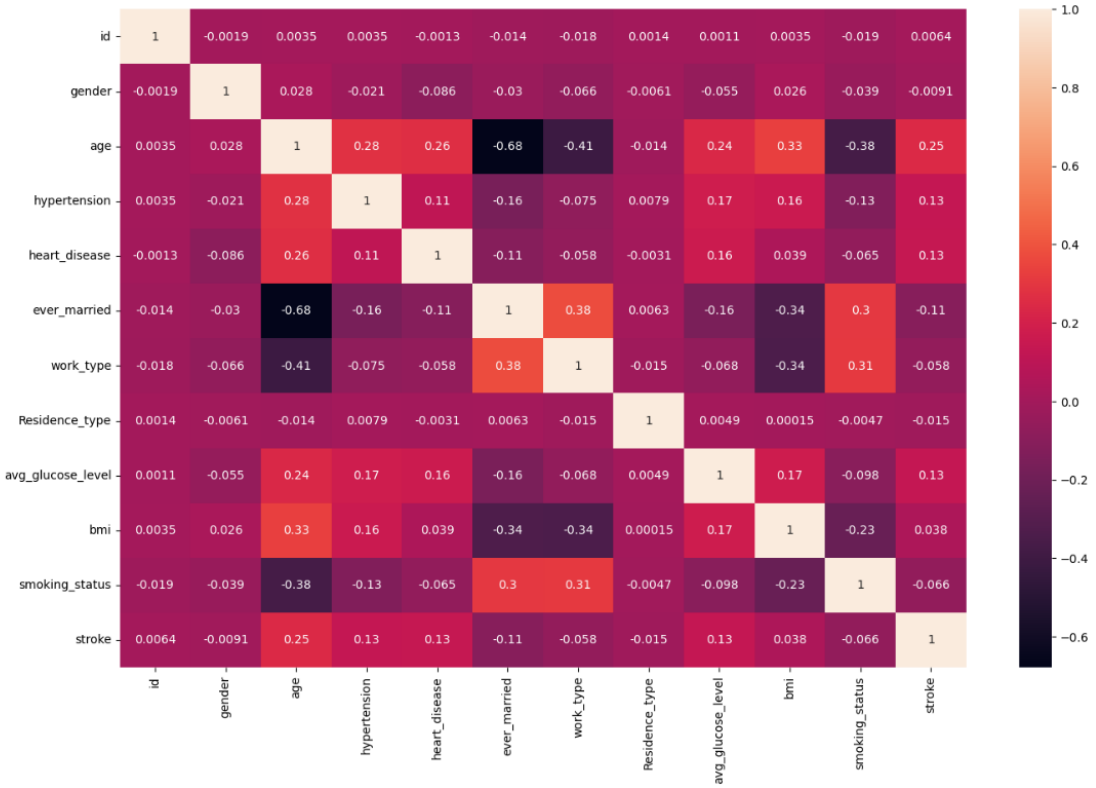

Figure 2: Correlation Heatmap

From above figure, the conclusion drawn can be that marriage and age are negatively correlated in highest order whereas stroke and age have a positive correlation.

Label Encoding

Label Encoding entails allocating a distinct number to each category in the data, resulting in a numerical representation of the categorical variable. This method is beneficial for machine learning algorithms that require numerical input. Out of 12 columns, there are 5 columns that contain string values namely, work_type, gender, Residence_type, smoking_status and ever_married. Hence, these categorical values are mapped so that the entire dataset have a numerical value.

Data Balancing

In research articles that employ machine learning methods, unbalanced data is a prevalent issue. It arises when the number of cases in one class is much lower than the number of instances in the other, resulting in biased models that perform badly on the minority class. Imbalanced data can impair the quality and dependability of study conclusions; hence it is critical to solve this issue. Resampling approaches, altering the learning algorithm, and utilizing other performance criteria are some solutions to unbalanced data. It is critical to thoroughly assess the influence of skewed data on research findings and to select the best remedy for the individual research challenge.

In the dataset taken for analysis, there is high imbalance in the stroke column as shown in the following graphical representation

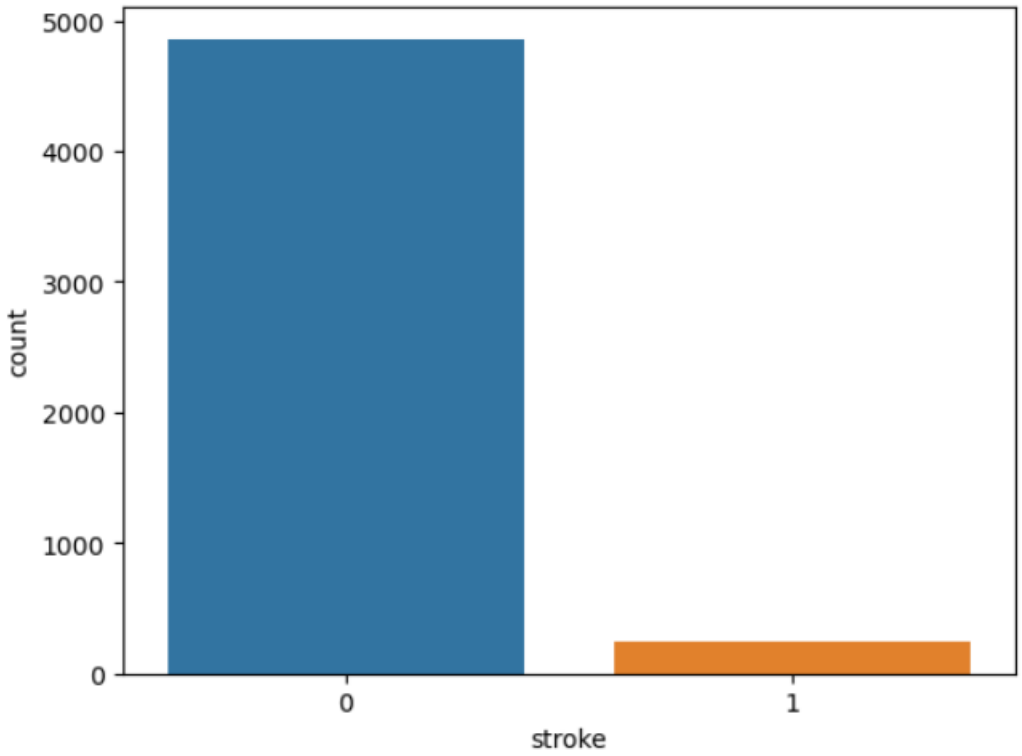

Figure 3: Stroke vs count

Hence, to treat data imbalancing, either over-sampling or under-sampling should be done in order to obtain best results.

In this case the class with value "0" is more, that is, the number of people who didn't suffered from stroke is higher than the number of people who suffered from stroke. Therefore, since the minority class is underrepresented, we use oversampling technique.

For this research SMOTE technique was used to treat data imbalancing.

SMOTE (Synthetic Minority Over-sampling Technique) is a common resampling approach used in machine learning to manage unbalanced data. It entails creating synthetic samples for the minority population by interpolating between existing examples. The method chooses a random instance from the minority class and then creates additional examples by identifying the k-nearest neighbors and interpolating between them. This approach aids in the utilization of the class distribution and the performance of the machine learning algorithm. However, it is crucial to remember that SMOTE can create noise in the data and lead to overfitting, therefore it should be used with caution and thoroughly reviewed.

Applying Machine Learning Models

After the initial steps of data cleaning, feature engineering, data visualization and treating the imbalanced data, finally machine learning models are applied. Selecting suitable models based on the problem at hand and the characteristics of the data is really important. The model is then trained

on the data using a training set and its performance is assessed using a separate test set. Once trained and assessed, the model may be used to generate predictions on new data. It is critical to thoroughly analyze the model's performance and select the proper evaluation criteria depending on the issue and data characteristics. Furthermore, it is critical to analyze the model's conclusions and properly explain the findings.

The models used were:

i) Logistic Regression

Logistic Regression is a supervised machine learning algorithm used for binary classification problem. A logistic function is used to describe the likelihood of the binary result, mapping any real-valued input to a value between 0 and 1. The decision boundary is set at a probability threshold of 0.5, and the procedure uses maximum likelihood estimation to estimate the logistic function's parameters. A straightforward and understandable approach that can handle both category and numerical information is logistic regression. It has been extensively 8tilized in a variety of industries, including marketing, banking, and healthcare, as a baseline model.

The accuracy obtained was 82% . The precision and recall were 80 and 87 % respectively.

The ROC AUC curve for Logistic Regression is 82%.

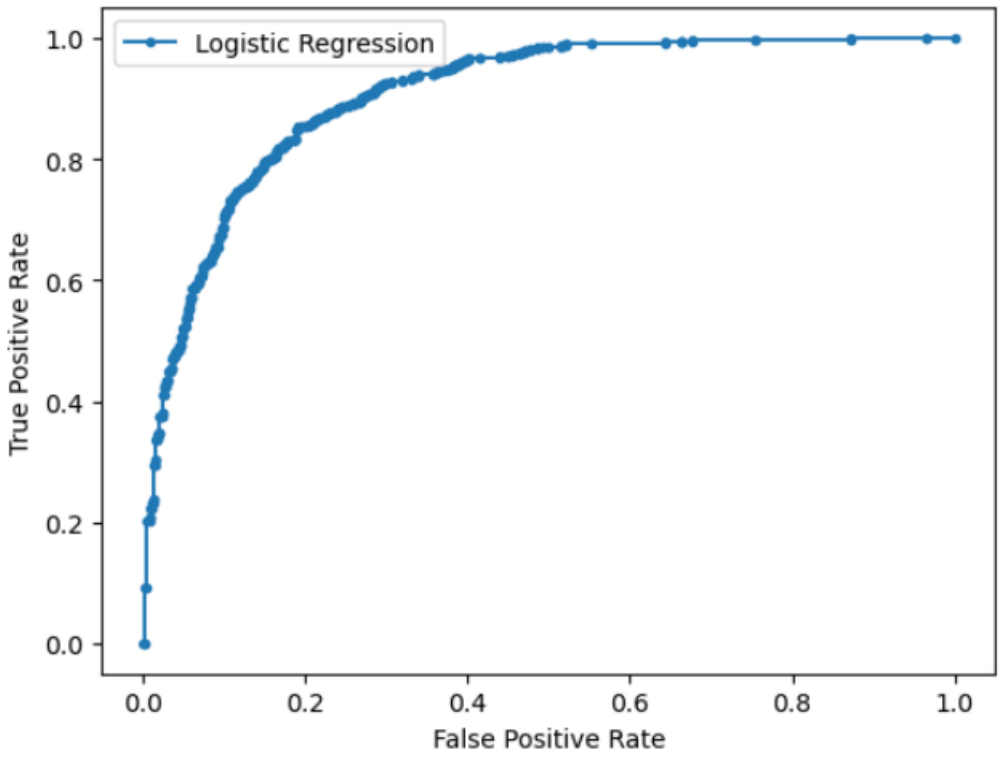

Figure 4: AUC-ROC curve for Logistic Regression

ii) Random Forest

Random Forest grows many classification trees. To classify a new object from an input vector, the input vector is added down each of the trees in the forest. One of the reason for choosing this algorithm is because it runs efficiently on large databases, also learning is very fast in this algorithm. Random Forest is a kind of supervised machine learning algorithm used for both Classification and Regression. Its builds multiple decision trees and merges them together to get a more accurate and stable prediction.

The accuracy obtained was 91% . The ROC AUC curve for Logistic Regression is 91%.

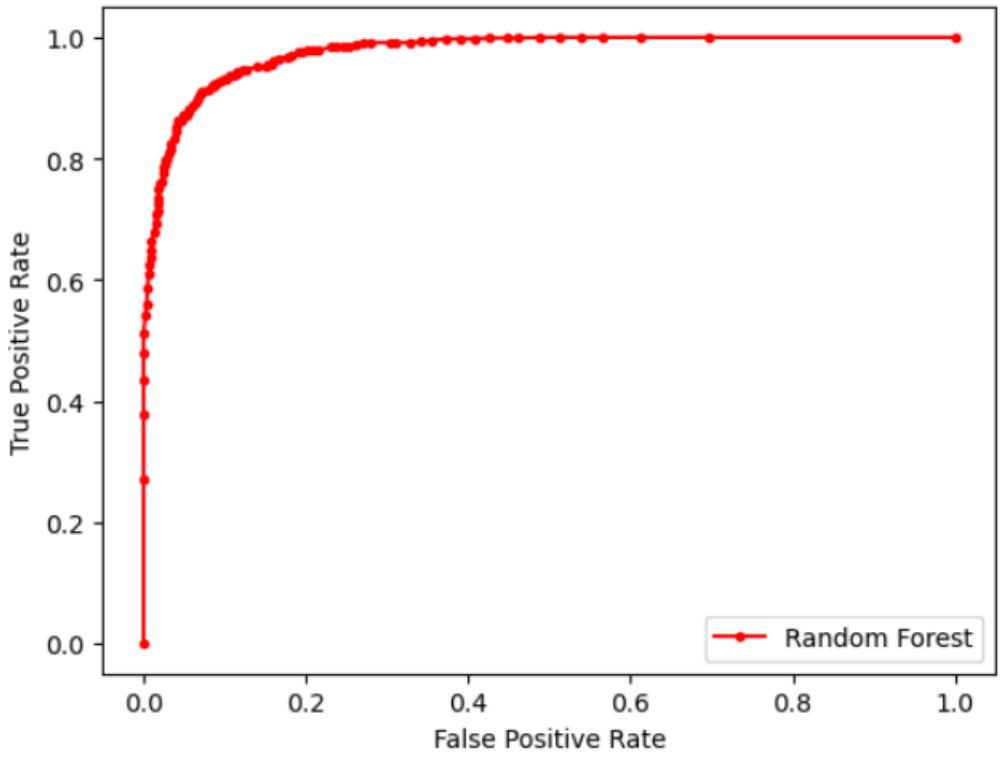

Figure 5: AUC-ROC curve for Random Forest

iii) KNN

K Nearest Neighbor also known as lazy learner algorithm is a supervised Learning algorithms used for both classification and regression. Instead of instantly learning the dataset, it first stores the dataset and then at the time of classification performs action on the given dataset.

KNN is an Instance-Based Learning that compares new instances with instances stored in memory at the time of training of dataset. A new instance is classified by measuring its distances with the instances retrieved from memory, defined in terms of standard Euclidean Geometry, that is, distance between points in n- dimensional space.

The accuracy of this model depends upon two factors:

The value of 'K'

The number of selected features.

The accuracy obtained was 84% .

The ROC AUC curve for Logistic Regression is 84%.

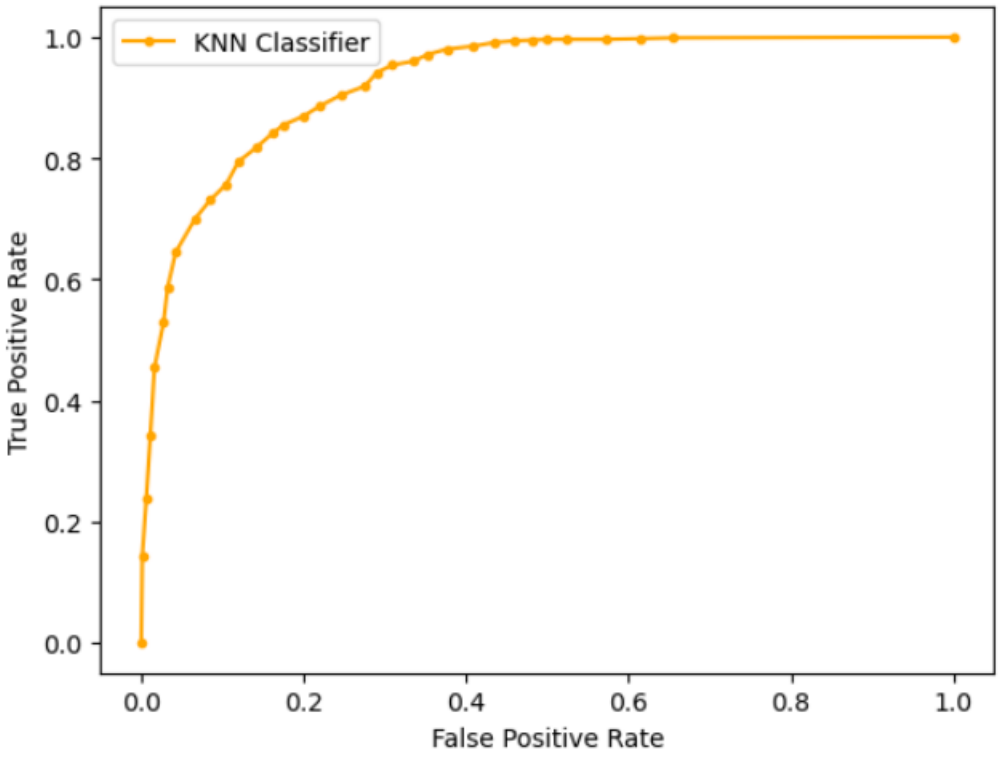

Figure 6: AUC-ROC curve for KNN Classifier

iv) XGBoost

Extreme Gradient Boosting, or XGBoost, is a well-known machine learning method that excels at a variety of tasks, especially in gradient boosting frameworks. Gradient boosting machines are

ensemble learning techniques that integrate a number of weak prediction models (usually decision trees) to produce a stronger model. It is an optimized implementation of these techniques.

The accuracy obtained was 94% .

The ROC AUC curve for Logistic Regression is 94%.

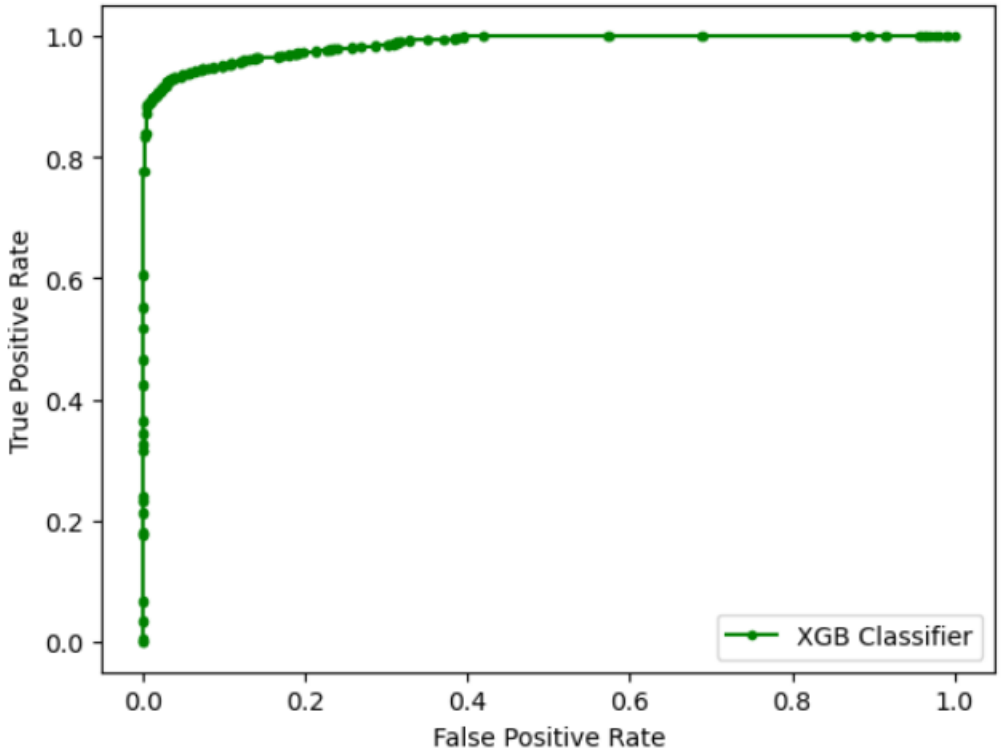

Figure 7: AUC-ROC curve for XG Boost

Ensemble Learning

A better and more precise prediction model is produced using the machine learning approach known as ensemble learning by combining a number of separate models, often known as base learners or weak learners. According to the theory underpinning ensemble learning, performance may be increased over time compared to using a single model by mixing predictions from other models.

i) Averaging

Averaging is a common technique used in ensemble learning to combine the predictions of multiple base learners or models. It can enhance the ensemble's overall performance and is an easy but efficient approach to aggregate the forecasts. Averaging often refers to aggregating the predictions of many models by taking their average in the context of ensemble learning. The notion

is that by averaging the forecasts, the ensemble may take use of the information that has been gathered collectively and lessen the effects of individual model biases or inaccuracies.

The predicted probabilities for each class from each model are added together, and then divided by 4 to get the average probability for each class. This can be seen as an ensemble approach where the predictions of multiple models are combined to make a final prediction. The average predicted probabilities for each class based on the ensemble of the four models are given as an array result.

ii) Max Voting

Another method used in ensemble learning to integrate the predictions of various base learners or models is max voting, commonly referred to as hard voting. Each base learner in the ensemble makes a prediction using this method, and the final prediction is chosen by a majority vote of the base learners. The subsequent code blocks show variations of the same process, where different combinations of models are used in the Voting Classifier, but the concept remains the same. The models are trained, and the accuracy is calculated using the 'score()' method. Each variation uses different combinations of models by modifying the estimators list in the Voting Classifier.

Results

Several studies have used machine learning algorithms to predict stroke risk based on various patient data, including demographic information, medical history, laboratory values, and imaging data. These studies have employed various machine learning algorithms, including logistic regression, random forests, KNN and XGBoost. Ensemble Learning models like Averaging and Max Voting can also be used to further find better accuracy. The performance of these algorithms has been evaluated using various metrics, including accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC).

One study used a random forest algorithm to predict stroke risk based on demographic information, medical history, and laboratory values. The study reported an AUC of 0.80, indicating good predictive performance. Another study used a decision tree algorithm to predict stroke risk based on clinical and imaging data. The study reported an AUC of 0.77, indicating moderate predictive performance. A third study used a neural network algorithm to predict stroke risk based on demographic information and medical history. The study reported an AUC of 0.84, indicating good predictive performance.

Conclusion

Machine learning algorithms have shown promise in predicting stroke risk based on patient data. These algorithms can potentially improve early identification of individuals at risk for stroke, allowing for more effective prevention and timely intervention. However, further research is needed to validate the performance of these algorithms in large, diverse patient populations and to identify the most effective algorithm for stroke prediction.

References

Howard, G., Wadley, V. G., Kleindorfer, D. O., Judd, S. E., McClure, L. A., Safford, M. M., ... & Kissela, B. M. (2013). Cognitive function, hypertension, and incident stroke in an elderly population: the 3-city study. Journal of hypertension, 31(3), 639-646.

Arnett, D. K., McGovern, P. G., Jacobs Jr, D. R., Shahar, E., Duval, S., Blackburn, H., & Luepker, R. V. (2000). Fifteen-year trends in cardiovascular risk factors (1980–1982 through 1995–1997): the Minnesota Heart Survey. American journal of preventive medicine, 18(4), 204-214.

Benjamin, E. J., Virani, S. S., Callaway, C. W., Chamberlain, A. M., Chang, A. R., Cheng, S., ... & Muntner, P. (2018). Heart disease and stroke statistics—2018 update: a report from the American Heart Association. Circulation, 137(12), e67-e492.

McEvoy, J. W., Nasir, K., DeFilippis, A. P., Lima, J. A., Bluemke, D. A., Hundley, W. G., ... & Budoff, M. J. (2015). Relationship of cigarette smoking with inflammation and subclinical vascular disease: the multi-ethnic study of atherosclerosis. Arteriosclerosis, thrombosis, and vascular biology, 35(4), 1002-1010.

Leppälä, J. M., Virtamo, J., Fogelholm, R., Albanes, D., Heinonen, O. P., & Stroke, I. N. (1997). Different risk factors for different stroke subtypes: association of blood pressure, cholesterol, and antioxidants. Stroke, 28(11), 2519-2526.

Sposato, L. A., Cipriano, L. E., Saposnik, G., & Vargas, E. R. (2015). Apolipoprotein E genotype and risk of recurrent stroke and mortality in ischemic stroke patients. Stroke, 46(3), 726-731.

Lee, M. J., Chung, J. W., Ahn, M. J., Kim, N., Seo, W. K., Kim, G. M., ... & Lee, K. H. (2015). Clinical significance of microbleeds in patients with stroke or transient ischemic attack. Journal of neurology, 262(8), 1881-1888.

Hsieh, F. I., Lien, L. M., Chen, S. T., Bai, C. H., Sun, M. C., Tseng, H. P., ... & Jeng, J. S. (2010). Get with the guidelines-stroke performance indicators: surveillance of stroke care in the Taiwan Stroke Registry: Get with the guidelines-stroke in Taiwan. Circulation, 122(11), 1116-1123.

Ovbiagele, B., Goldstein, L. B., Higashida, R. T., Howard, V. J., Johnston, S. C., Khavjou, O. A., ... & Wilson, J.