Stroke Disease Detection and Prediction

Shehla Mushtaq1

1Sukkur IBA University, Sindh, Pakistan

shehlamushtaq63@gmail.com

ABSTRACT

Since stroke illness commonly leads to death or significant disability, active primary prevention and early detection of prognostic indicators are essential. Both ischemic and hemorrhagic strokes should be treated immediately with the proper thrombolytic or anticoagulant medicines. It is essential to identify the stroke's early warning signals in real time, which differ from person to person. Prior studies, however, have focused more on developing recommendations for clinical or acute therapy after the start of stroke than on detecting its warning signs. The aim of this research study is to use 6 different machine learning (ML) algorithms such as Decision Tree, Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, Support Vector Machine (SVM), and Naive Bayes to predict the likelihood of a stroke occurring in the brain. The results showed that Random Forest and Decision Tree performed better than the other four algorithms in classifying people as having a stroke illness or not, with accuracy rates of 90% and 87.8%, respectively. By assisting physicians and other healthcare professionals in taking preventative steps and making tailored suggestions to high-risk patients, such as lifestyle changes and medication, the proposed system would play a vital role in the medical industry.

**Keywords**: stroke disease prediction, stroke disease detection, hypertension, stroke

I.INTRODUCTION

Stroke is a disease in which blood vessels in the brain burst, causing brain damage. If there is a problem with the flow of blood and other nutrients to the brain, symptoms might develop. When a blood clot obstructs blood flow, this disease develops.[[1](#link1)] and the cause of this clotting is any brain injury. In this case, cells of the brain become damaged or die. Brain stroke can be occurred due to many other factors. Age, gender, and inherited characteristics are risk factors. Modifiable risk factors include high blood pressure, high blood pressure, smoking, heart disease, diabetes, low cholesterol, physical inactivity, and obesity. After a stroke, difficulties can include weakness, paralysis, exhaustion, headaches, dizziness, and even death. There are two classifications for strokes:

1. Ischemic Strokes:

Ischemic stroke is the most common type, accounting for approximately 87% of all strokes. It occurs as a result of blockage of the vessels that supply blood to the brain.

1. Hemorrhagic Strokes:

In this kind, a brain blood artery bursts or spills. High blood pressure has been associated with hemorrhagic strokes.

II. RELATED WORK

Deep learning has been used in many studies to solve various problems. Predicting Stroke Mortality Using Deep Learning uses machine learning models and techniques to predict stroke patient deaths [[4].](#link4) Various types of machine learning have been developed and used, such as Support Vector Machine (SVM), Random Forest (RF), Artificial Neural Network (ANN), and Linear Regression (LR) to determine the risk of osteoporosis in postmenopausal women. They are able to do this using medical data from studies such as the South Korean Health and Welfare Survey.While the accuracy is sufficient, very little data is used (1000 subjects for training and 674 tests). [[5]](#link5)

III. DATASET DISCUSSION

This information is used to determine and predict whether a patient is more likely to have a stroke, based on factors such as gender, age, various diseases, and smoking. Each column contains information about the patient.[[1]](#link1) The dataset includes 12 features and 5110 observations. Each attribute is defined below in a Table 1.

Table 1: Dataset Properties and Description

|  |  |  |
| --- | --- | --- |
| S.No | Attribute | Description |
| 1. | Id | Unique Identifier |
| 2. | Gender | Male/Female /Other |
| 3. | Age | Patient Age |
| 4. | Hypertension | 0/1 |
| 5. | Heart\_Disease | 0/1 |
| 6. | Ever\_Married | Yes/No |
| 7. | Work\_Type | Govt./Private/Self Employed |
| 8. | Residence\_Type | Rural/Urban |
| 9. | Avg\_Glocose\_level | In blood |
| 10. | BMI | Body Mass Index |
| 11. | Smoking\_Status | Smoke/Unknown |
| 12. | Stroke | 0/1 |

**\*Note: "** **The word "unknown" in the smoking\_status field indicates that the patient's information is not accessible. And 1 indicates yes while 0 indicates no.**

IV. METHODOLOGY

A. System Overview

Based on input characteristics like gender, age, the existence of different diseases, and smoking status, the system is intended to estimate the chance of a patient experiencing a stroke. The 5110 analysis results in each row containing patient-related information as well as data used to train and evaluate the classifier. Identity, Gender, Age, Blood Pressure, Heart Disease, Marriage, Occupation, Type of Residence, Average Blood Pressure Level, BMI, Stroke and Smoking are the 12 characteristics included in this document. Following six machine learning classifiers have been applied to the dataset.

1. Decision Tree
2. Logistic Regression
3. K-Nearest Neighbors (KNN)
4. Random Forest
5. Support Vector Machine (SVM)
6. Naive Bayes
7. Also useful is the area in the receiver operating characteristic curve (AUC ROC curve), which is used to graphically evaluate the performance of the classifier. During the model training phase, hyperparameter changes were used to improve the performance of the model and the accuracy reached 80%. Using these accurate measurements, the System can accurately predict whether a patient will suffer a stroke based on the device. It is important to note that the performance of the model will vary depending on the accuracy and representativeness of the data and the specific features and isolated items selected. To increase the accuracy and predictability of the model in estimating stroke incidence, the model can be updated regularly and improved as new information becomes available.

B. System Implementation

The proposed system is implemented in various machine learning phases such as at first phase, we preprocessed the dataset performed some operations such as missing data analysis, handling in balanced data, and label encoding.

At second phase we have trained our dataset on six different machine learning algorithms such as Support Vector Machine (SVM), Neighbors (KNN), K-Nearest Random Forest, Decision Tree, Logistic Regression, and Naive Bayes to predict the likelihood of a stroke occurring in the brain.

At last phase, we have evaluated machine learning performance of models by giving test data in order to compare which algorithm outperform and give optimum accuracy among all six. The results show that the accuracy of the random forest and decision tree among the four matches is 90% and 87.8%, respectively.

The flowchart of proposed system is as follows:

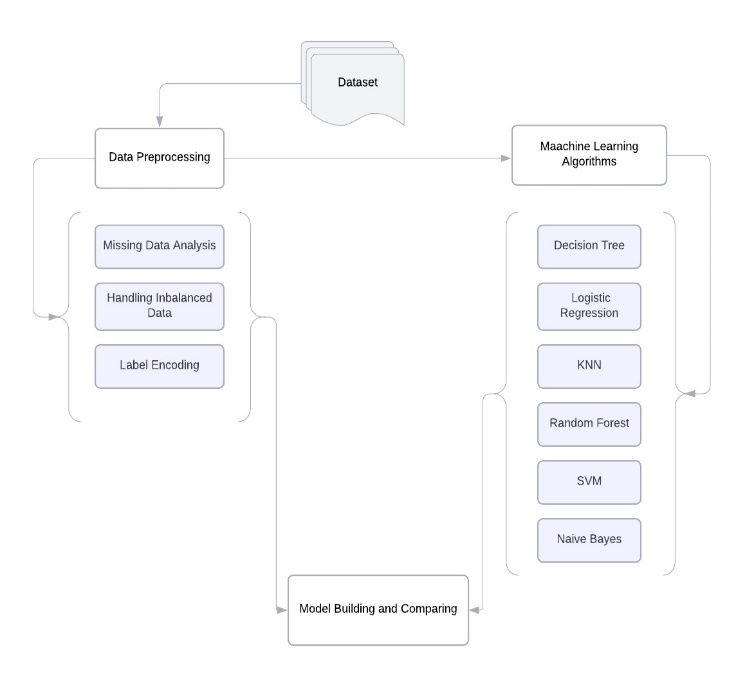
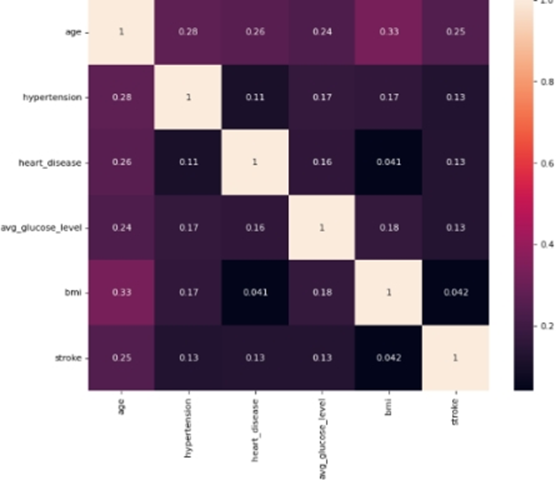


Figure 1: Flowchart of Stroke Detection and Prediction:

* Correlation:

Showing correlation matrix using heat-map. It shows the relationship between pairs of variables, in this data it shows t features are not correlated to each other.

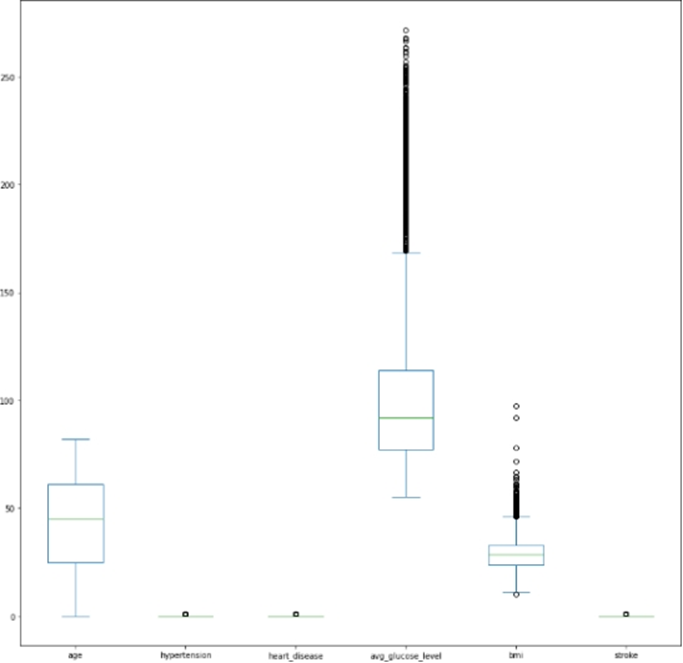


* Outliers:

Outliers in the "average glucose level" characteristic were found using a boxplot. We can compute the mean, median, minimum, maximum, and quartiles to comprehend the statistics of this data.

1. Gender and Stroke
2. Age and Stroke
3. Hypertension and Stroke
4. Heart Disease and Stroke
5. Ever Married and Stroke
6. Average Glucose Level and Stroke
7. BMI and Stroke

We can learn more about the connections between certain characteristics and the risk of having a stroke by looking at these associations. Boxplots can be a useful tool for visualizing and identifying outliers, while statistics give important details about the distribution and central tendency of the data.



* The Decision Tree (DT):

Based on input parameters, the system uses the Decision Tree classifier, which has a 79% accuracy rate, to predict the chance of stroke. Its interpretability increases its value in activities requiring decision-making. The confusion matrix provides an overview of correct and incorrect predictions to evaluate the performance of the model. Metrics such as True Positives, True Negatives, False Positives, and False Negatives are used to measure the effectiveness of the model.

The technology shows promise with a 79% accuracy rate and is helpful in medical decision-making. With the addition of fresh data, ongoing monitoring and adjustments are made to improve predictive skills. The system's capacity to discern between cases of stroke and non-stroke is crucial for well-informed medical actions. The user-friendly interface makes it simple to enter patient data for the evaluation of stroke risk. The model is kept current with the most recent medical discoveries through regular upgrades.

* Logistic Regression:

A popular machine learning approach for categorical predictions is logistic regression (LR). To forecast a categorical dependent variable, it makes use of a number of independent factors. Logistic regression produces discrete results, often reflecting binary options like true or false, 0 or 1, and yes or no. Additionally, the model offers probabilities ranging from 0 to 1 for each prediction.

Using the input features provided, this system uses logistic regression to forecast the chance of stroke based on the occurrence of a category event. On the presented dataset, the model's accuracy was 78%, demonstrating its ability to consistently produce accurate predictions. logistic regression probability scores are useful because they let us gauge how definite each prediction is. We can be more certain that a stroke will occur when the probability is close to 1, whereas a probability number close to 0 implies that there is little chance that a stroke will occur.

Overall, the popularity of logistic regression and its 78% accuracy in this system make it a useful tool for predicting the occurrence of strokes and supporting clinical decision-making. The performance and dependability of the model can be further improved with frequent updates and data improvements.

* KNN Classifier:

Both classification and regression issues are solved using K-Nearest Neighbors (KNN), a fundamental machine learning technique. Based on a set of input values, it anticipates output values and classifies a data point according to the categorization of its closest neighbors.

Based on the input features given, this system uses KNN to forecast the chance of stroke. The model's 79% accuracy rate shows that it can correctly anticipate outcomes in a sizable number of situations. KNN is a well-liked option for a variety of classification jobs due to its simplicity and efficacy. It can dynamically acclimatize to new data points and doesn't need intentional training. The model places a data point in the majority class among its K closest neighbors by comparing it to those neighbors.

The KNN algorithm shows its value in forecasting the occurrence of strokes and aids in medical decision-making with a 79% accuracy rate. The model's performance can be further optimized, and the addition of additional data can improve its ability to forecast the future.

* Random Forest:

The algorithm employed for this system is classification using Random Forest (RF). Multiple independent decision trees that were each trained using a random sample of the data make up RFs. These trees are built during training, and predictions are made using an aggregation of their outputs.

The adaptability of Random Forest is one of its best qualities. It may be applied to a variety of tasks, including grouping and relapse detection, and it gives a clear explanation of the weights assigned to various data elements. Random Forest attained a remarkable accuracy of 90% in this system. With such great accuracy, the model must correctly anticipate the vast majority of data instances.

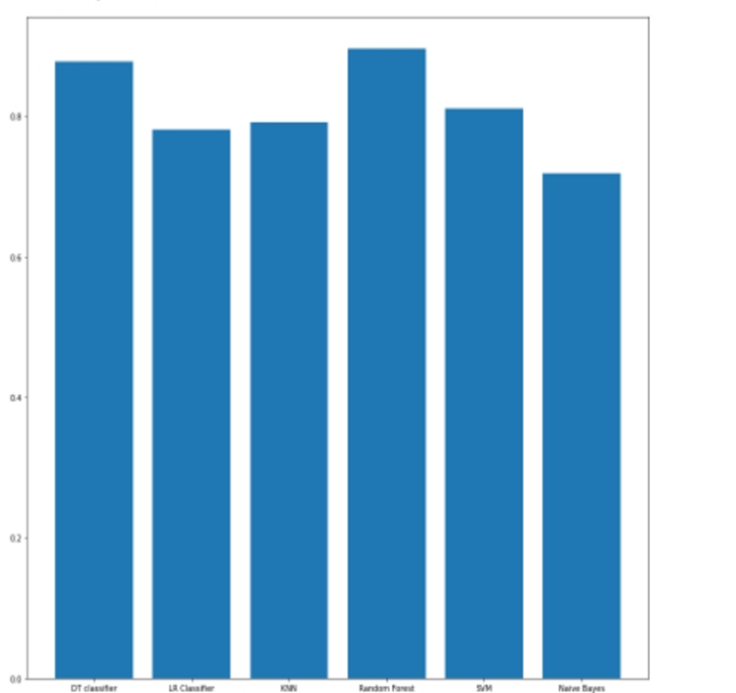
Random Forest demonstrates to be a significant tool for forecasting the occurrence of strokes and assisting medical decision-making due to its capacity for handling difficult tasks and excellent accuracy. The model's performance will be further improved, and ongoing enhancements will increase its accuracy in predicting outcomes of strokes.

* Support vector machine (SVM):

SVMs are flexible algorithms that can be utilized for regression and classification tasks on both linear and non-linear data. SVM was able to predict stroke occurrences in this system with 81% accuracy, supporting clinical judgment effectively. With frequent updates and tuning, the model's performance can be improved with fresh data.

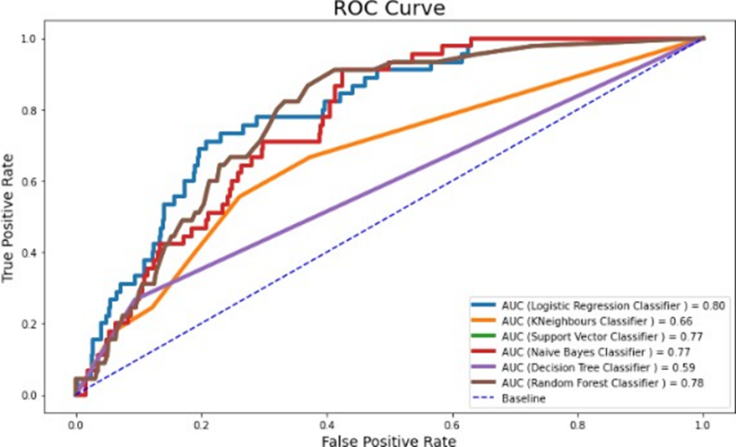
* Naïve Bayes:

Naive Bayes, is renowned for processing continuous and discrete data with equal proficiency. In comparison to other approaches, it needs comparatively fewer training data. It is appropriate for large datasets due to its scalability with regard to the number of predictors and data points. Because of the algorithm's speed, real-time predictions are possible, making it useful in situations when timing is crucial. Naive Bayes attained an accuracy of 72% in this system, proving its capacity to reasonably anticipate the occurrence of strokes. Its performance can be enhanced even further by routine updates and data cleaning.



* AUC ROC CURVE:

The area under the curve (AUC) is an important consideration in binary classification tasks such as the stroke predictor system. It evaluates the classifier's ability to discriminate between good and negative events. The AUC-ROC curve provides insight into the performance of a classifier by comparing different levels of true positives to false positives.

A higher value of AUC indicates better discriminating power, which demonstrates that the classifier is more accurate in distinguishing between stroke and non-stroke cases. AUC of 0.5 indicates guessing at random, whereas AUC of 1 indicates a flawless classifier. This system's AUC ROC curve provides valuable data on how well each classifier predicts strokes, facilitating the choice of the most reliable model for medical decision-making.

* Hyperparameter Tuning:

A learning algorithm learns model parameters based on the parameters whose values influence the learning process. Therefore, we have defined three parameters for logistic regression: penalty, C, and solver. Based on these settings, any model will yield a good accuracy score.

## Hyperparameter Tuning on Logistic Regression:

## The model was trained, and it achieved an initial accuracy of 77%. To discover the optimal combination of hyperparameters, hyperparameter optimization utilizing Grid Search was used. This optimization resulted in a Logistic Regression model accuracy score of 80%, indicating an improvement over the baseline accuracy.

## Hyperparameter Tuning on Random Forest:

## We've specified four critical parameters for the random forest model that will help us achieve a high degree of accuracy. These parameters are 'n\_estimators','max\_features','max\_depth', and 'criterion'. These parameters have values of [200, 500] for 'n\_estimators', [auto','sqrt', 'log2'] for'max\_features', [4, 5, 6, 7, 8] for'max\_depth', and ['gini', 'entropy'] for 'criterion'. We were able to attain a precision score of 85% using the Random Forest model by using these settings.

## Hyper parameter Tuning on Naive Bayes:

We have set the option "Var smoothing" for Naive Bayes to be np.logspace (0,-9, number=100). Here, we arrive at a Naive Bayes accuracy of 70%. After applying grid search to various models, including

We observed that the Random Forest model has the greatest score for accuracy when compared to logistic Regression and naive Bayes.

VI. RESULTS

The most accurate classifier in the system at hand is Random Forest with accuracy of 90%, while Naive Bayes performs the least well of the other classifiers with accuracy of 87.8%. This demonstrates how well Random Forest predicts the occurrence of strokes, while Naive Bayes performs considerably less well in this situation.

|  |  |  |
| --- | --- | --- |
| S.No | Classifier | Accuracy |
| 1. | Decision Tree | 87.8% |
| 2. | Logistic Regression | 78% |
| 3. | KNN | 79% |
| 4. | Random Forest | 90% |
| 5. | SVM | 81% |
| 6. | Naïve Bayes | 72% |

VII. FUTURE WORK

Increasing data collecting, researching sophisticated machine learning models, increasing model interpretability, and integrating with healthcare systems are all aspects of stroke disease detection and prediction. For better patient outcomes and public awareness, real-time prediction, remote health monitoring, personalized risk assessment, and clinical trials are also essential.

VIII. CONCLUSION

Stroke is a possibly fatal medical illness that need quick treatment to avert further consequences.[[1](#link1)] Creating a machine learning model might help in recognizing strokes more quickly, thereby reducing their serious effects. Many machine learning algorithms demonstrate accuracy in predicting strokes using the aforementioned parameters, which include smoking, hypertension, and heart disease.[[3]](#link3) We will be able to use the tools and techniques covered in the Machine Learning through our work on this dataset.

IX. REFERENCES

1. Stroke disease detection and prediction. (2021, February). Hindawi. Retrieved December 7, 2022, from <https://www.hindawi.com/journals/jhe/2021/7633381/>
2. Machine learning for brain stroke. (2020, October 10). ScienceDirect. Retrieved December 7, 2022, from <https://www.sciencedirect.com/science/article/pii/>
3. A. K. Uttam, "Analysis of Uneven Stroke Prediction Dataset using Machine Learning," 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS), 2022, pp. 1209-1213, doi: 10.1109/ICICCS53718.2022.9788309

[4] Int. J. Environ. Res. Public Health 2019, 16(11),

1876; <https://doi.org/10.3390/ijerph16111876>

Received: 30 April 2019 / Revised: 23 May 2019 / Accepted: 24 May 2019 / Published: 28 May 2019

[5] Kim, S.K.; Yoo, T.K.; Kim, D.W. Osteoporosis risk prediction using machine learning and conventional methods. In Proceedings of the 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Osaka, Japan, 3–7 July 2013; IEEE: Piscataway, NJ, USA, 2013; pp. 188–191]