International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 05 | May - 2023 SJIF 2023: 8.176 ISSN: 2582-3930

BRAIN STROKE PREDICTION

P.Logeswari, Julee Kumari, Shivani Kumari, Satish Kumar Choudhary, Kumar Ritu Raj

Research Scholar School of Computer Application, Lovely ProfessionalUniversity, Jalandhar, 144411.

School of Computer Application, Lovely Professional University, Jalandhar144411.

Abstract: Brain stroke is a leading cause of death and disability worldwide, and early detection and treatment are critical to improving patient outcomes. In recent years, there has been increasing interest in using predictive models to identify individuals at high risk of stroke, allowing for targeted interventions and preventive measures. Traditional risk factors such as age, blood pressure, and smoking status have been used in tools like the Framingham Stroke Risk Profile, but newer approaches such as machine learning and genetic analysis offer the potential for more accurate and personalized risk prediction. Studies have explored the use of electronic health records and imaging biomarkers to develop predictive models, as well as identify specific genetic variants associated with increased stroke risk. As these methods continue to improve, they may ultimately help to reduce the burden of stroke on individuals and healthcare systems.

1. INTRODUCTION

Stroke is a major public health concern worldwide, causing significant morbidity, mortality, and disability. Early detection and prevention of stroke can be crucial in reducing the burden of the condition on individuals, healthcare systems, and society. Brain stroke is caused by the interruption of blood supply to the brain, leading to damage to brain cells and, in severe cases, permanent disability or death. There are several risk factors associated with brain stroke, including age, gender, hypertension, diabetes, smoking, and family history. Early prediction of stroke can help identify high-risk individuals and implement preventive measures.

A stroke occurs when the blood flow to various areas of the brain is disrupted or diminished, resulting in the cells in those areas of the brain not receiving the nutrients and oxygen they require and dying. A stroke is a medical emergency that requires urgent medical attention. Early detection and appropriate management are required to prevent further damage to the affected area of the brain and other complications in other parts of the body.

With the development of technology in the medical sector, it is now possible to anticipate the onset of a stroke by utilizing ML techniques. We use hear six ML techniques Radial Support Vector Machines (Linear and rbf), Decision Trees, K-Nearest Neighbours, Gaussian Naive Bayes, Random Forests, and Logistic Regression. We figure out which model gives the best accuracy then we implement it for the user. The implementation of these ML classification methods is shown in this paper.

2. EXPERIMENTAL PART

2.1. Dataset Overview

```
Data columns (total 8 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   gender             4981 non-null   object
 1   age                4981 non-null   float64
 2   hypertension       4981 non-null   int64
 3   heart_disease      4981 non-null   int64
 4   avg_glucose_level  4981 non-null   float64
 5   bmi                4981 non-null   float64
 6   smoking_status     4981 non-null   object
 7   stroke             4981 non-null   int64
dtypes: float64(3), int64(3), object(2)
```

Fig.2.1.1: Information of Dataset

There are a total of 4981 rows and 8 columns "Gender", "Age", "Hypertension", "Heart_disease", "Avg_glucose_level", "Bmi", "Smoking_status", and "Stroke". Where three float data types, three int data types, and two object data types.

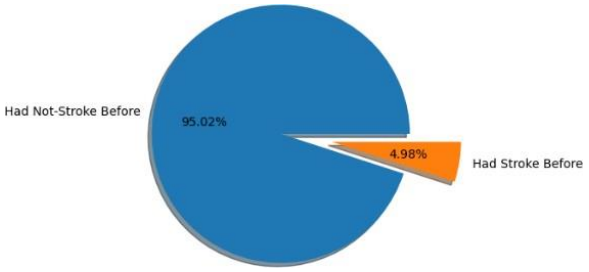


Fig.2.1.2: Stroke or Not Stroke

In this table, there are 95.02% of persons had Not Stroke before and 4.98% person had a stroke before.

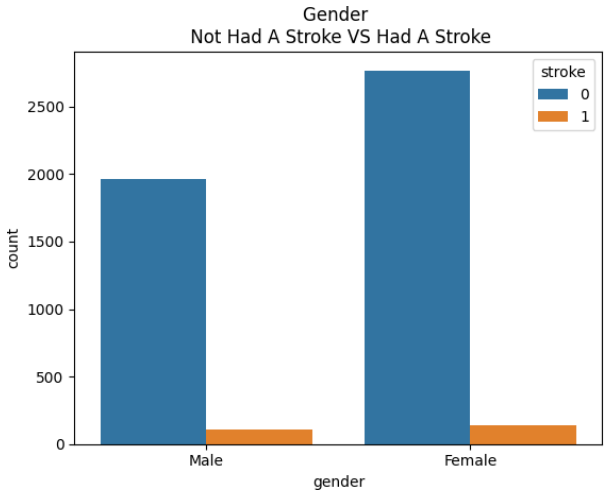


Fig.2.1.3: Gender with Stroke

This visualization bar graph shows us the number of females is greater than the number of males in both of the cases Not had a stroke and had a stroke.

stroke	0	1	All
smoking_status			
Unknown	1453	47	1500
formerly smoked	797	70	867
never smoked	1749	89	1838
smokes	734	42	776
All	4733	248	4981

Fig.2.1.4: Smoking_Status with Stroke

This graph represents the smoking status of different people.

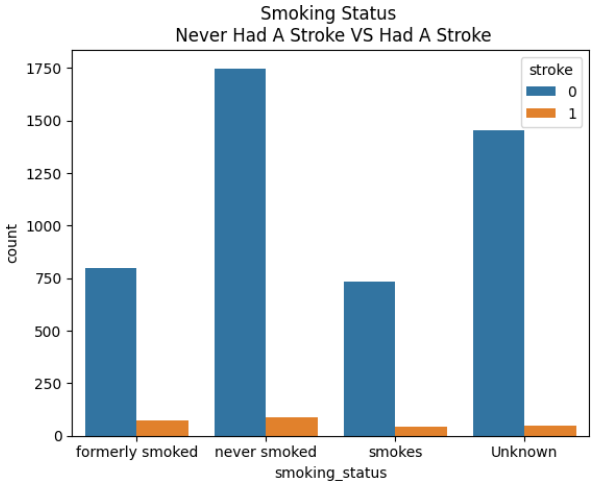


Fig.2.1.5: Stroke with Smoking_Status

In this graph, we get that the number of never smokers is less % of stroke compared to other people.

gender	smoking_status / stroke	Unknown	formerly smoked	never smoked	smokes	All
Female	0	783	430	1132	422	2767
	1	25	34	62	19	140
Male	0	670	367	617	312	1966
	1	22	36	27	23	108
All		1500	867	1838	776	4981

Fig.2.1.6: Stroke with Gender

This graph shows the stroke of Females and males compared to the different types of Smoking status.

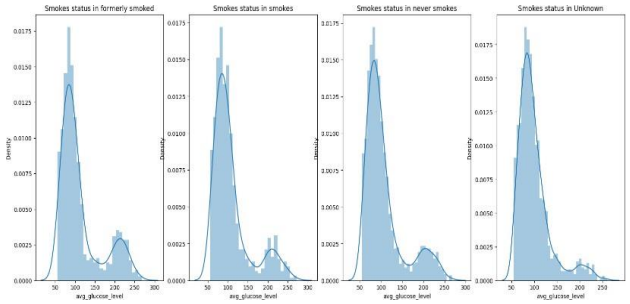


Fig.2.1.7: Smoking_status with Average Glucose Level

This graph represents the Different types of smokers with respect to average glucose levels like formerly smoked, Smokes, Never Smokes, and Unknown.

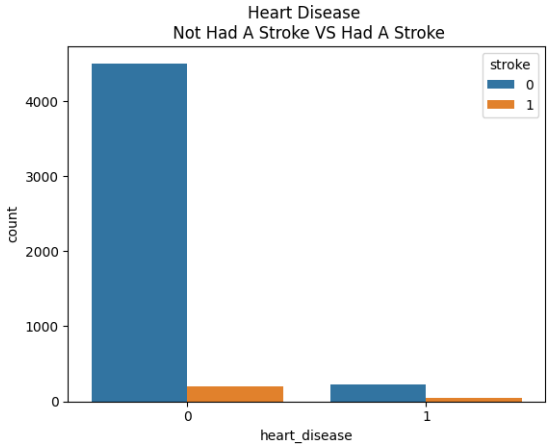


Fig.2.1.9: Stroke with Heart Disease

This graph showed us that a person who has heart disease they have a chance of being a Stroke.

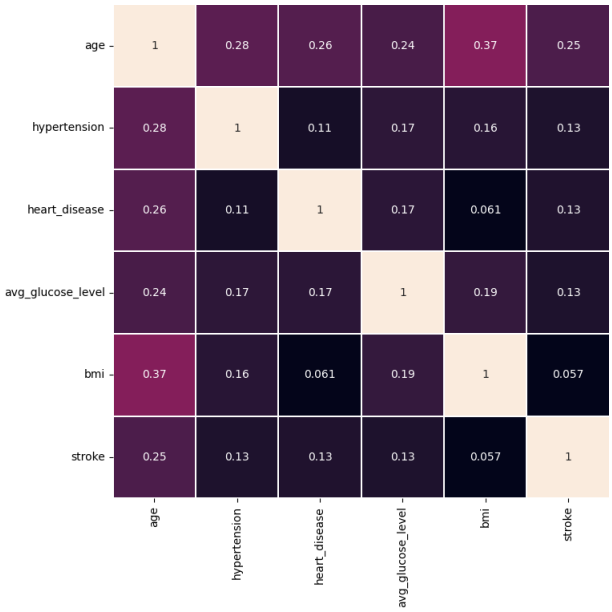
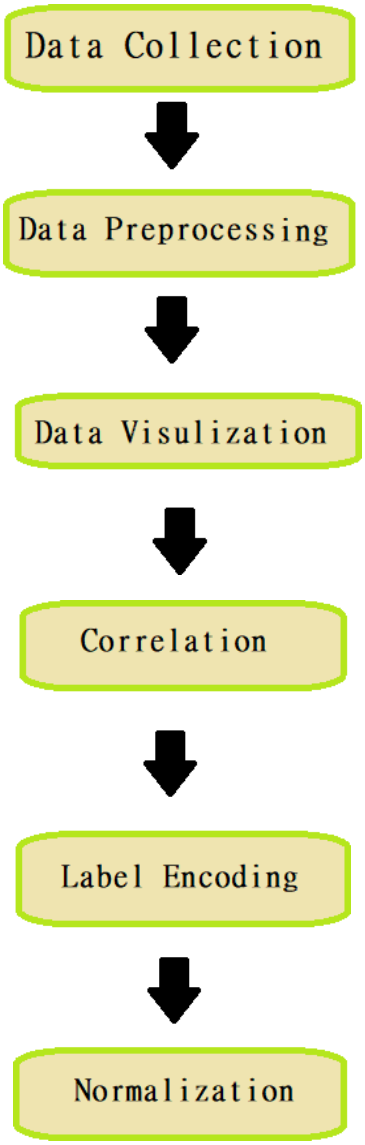
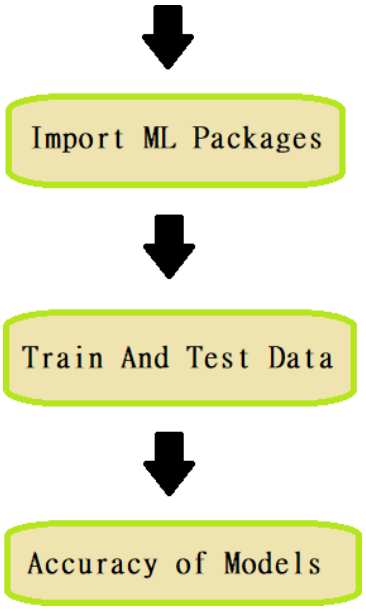


Fig.2.1.10: Correlation of data

This is a correlation graph where the light color represents the high correlation and the dark color represents the less correlation of data. In this table, the second highest correlation is between age and Bmi. The fewer correlation data is Bmi and Stroke.

2.2. Working Process on Dataset





2.3. Model Explanation

2.3.1 : Radial Support Vector Machines

Radial Support Vector Machines (SVM) are a type of machine learning algorithm used for classification and regression analysis. In SVM, the goal is to find a hyperplane that maximally separates data points into different classes. The radial basis function (RBF) is a popular kernel function used in SVM that allows for non-linear classification by transforming the data into a higher dimensional space.

Accuracy:
50% Train and 50% Test Data = 95%
60% Train and 40% Test Data = 95%
70% Train and 30% Test Data = 95%

2.3.2 : Decision Trees

Decision Trees are a popular machine learning algorithm used for both classification and regression tasks. It works by partitioning the data into smaller subsets based on the values of input features until a stopping criterion is met.

The decision tree starts with a root node, which represents the entire dataset, and then recursively splits the data into smaller subsets based on the values of a selected feature. The splitting process continues until the stopping criterion is met, which could be a maximum depth of the tree, a minimum number of data points in a node, or a minimum reduction in impurity measure such as entropy or Gini index.

Accuracy:
50% Train and 50% Test Data = 90%
60% Train and 40% Test Data = 90%
70% Train and 30% Test Data = 90%

2.3.3 : K-Nearest Neighbours

K-Nearest Neighbors (K-NN) is a non-parametric machine learning algorithm used for both classification and regression tasks. The algorithm works by finding the k closest training examples in the feature space to the new input data point and then using a majority vote (for classification) or an average (for regression) of their target values as the predicted output.K-NN is a lazy algorithm, meaning thatit does not learn a specific model from the training data but instead memorizes the training data to perform classification or regression at prediction time. The algorithm is based on the assumption that data points that are close to each other in the feature space are likely to belong to the same class or have similar target values.

Accuracy:
50% Train and 50% Test Data = 94%
60% Train and 40% Test Data = 94%
70% Train and 30% Test Data = 94%

2.3.4 : Gaussian Naive Bayes

Gaussian Naive Bayes (GNB) is a probabilistic machine learning algorithm used for classification tasks. It is based on Bayes' theorem, which states that the probability of a hypothesis (class label) given the observed evidence (input features) is proportional to the probability of the evidence given the hypothesis, multiplied by the prior probability of the hypothesis. In GNB, it is assumed that the input features are normally distributed and independent of each other, hence the term "naive". The algorithm estimates the mean and variance of each feature for each class label from the training data and uses these estimates

to compute the conditional probability of each feature given a specific class label.
Accuracy:
50% Train and 50% Test Data = 87%
60% Train and 40% Test Data = 86%
70% Train and 30% Test Data = 86%

2.3.5 : Random Forests

Random Forests is an ensemble learning method for classification, regression, and other machine learning tasks. It combines multiple decision trees trained on different subsets of the training data and features to reduce overfitting and improve accuracy.

The algorithm works by first creating a set of decision trees, each trained on a random subset of the training data and a random subset of the input features. Each tree independently produces a prediction, and the final prediction is obtained by aggregating the individual predictions, typically by taking a majority vote for classification tasks or a weighted average for regression tasks.
Accuracy:
50% Train and 50% Test Data = 94%
60% Train and 40% Test Data = 94%
70% Train and 30% Test Data = 94%

2.3.6 : Logistic Regression

Logistic Regression is a statistical machine learning algorithm used for binary classification tasks. It models the probability of a binary output variable (usually labeled as 0 or 1) as a function of one or more input features. The output of the algorithm is the predicted probability of the positive class, which can be thresholded to make binary predictions. Logistic Regression works by first estimating the parameters of a logistic function, which maps the input features to a probability score between 0 and 1. The logistic function is a sigmoid curve that asymptotically approaches 0 for negative inputs and 1 for positive inputs, with a transition region that corresponds to the decision boundary.
Accuracy:
50% Train and 50% Test Data = 95%
60% Train and 40% Test Data = 95%
70% Train and 30% Test Data = 95%

3. CONCLUSION

In summary, Stroke is an illness that should be treated as soon as possible to avoid further complications. The development of an ML model could aid in the early detection of stroke and the subsequent mitigation of its severe consequences. Our review highlights the potential of machine learning techniques for stroke risk prediction.

In these six ML techniques Radial Support Vector Machines(Linear and rbf), Decision Trees, K-Nearest Neighbours, Gaussian Naive Bayes, Random Forests, and Logistic Regression provides different accuracy. In this project, we use Logistic Regression with the highest accuracy 95.05% to predict the stroke. and robust machine learning models for stroke risk prediction.

4. REFERENCES

JAEHAK YU1 , SEJIN PARK 2 , SOON-HYUN KWON1 , KANG-HEE CHO3 , AND HANSUNG LEE 4 , “ AI-Based Stroke Disease Prediction System Using Real-Time ElectromyographySignals”,IEEEAccess,volume10,2022.https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9761215

H. Mcheick, H. Nasser, M. Dbouk and A. Nasser, "Stroke Prediction Context-Aware Health Care System," 2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE), 2016, pp. 30-35, doi: 10.1109/CHASE.2016.49.https://ieeexplore.ieee.org/document/7545809