iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.52600

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

Analysis and Prediction of Stroke Using Machine Learning Algorithms

Dr.Priti Mishra[1], Chitra A[2], Keerthana D[3], Pruthvika S[4]
[1, 2, 3, 4]*CSE Department East West College of Engineering Bangalore*

Abstract: ***Stroke is the second most common cause of mortality and are associated with substantial, protracted disability. Stroke is the sudden death of brain cells due to a lack of oxygen, which is brought on by a blood vessel blockage or disruption of a supply line to the brain. The World Health Organisation predicts that the death rate from stroke will continue to rise in the future year. Numerous studies have been conducted to find stroke illnesses. a deep learning-based artificial intelligence method for forecasting different forms of stroke. Ischemic stroke, hemorrhagic stroke, and transient ischemic attack are the three types. We used data gathered from a medical research institute in our work. The preprocessing technique eliminates duplicate records, missing data, and contradictory data. Deep learning is employed to forecast if the patient is experiencing stroke illness or not, and the principle component analysis computation is used to reduce estimations. It actualizes deep learning classification to predict the stroke sickness. When a patient's information is entered, a trained model and forecasts of various stroke types are checked. The major goal of this research is to improve stroke prediction and identify distinct types of stroke.***
Keywords: ML: Machine Learning, CSV: Comma- Separated Values, Random Forest algorithm, Decision Tree Classifier, XGBBoost, KNN Classifier.

I. INTRODUCTION

The burden of stroke is a significant global health issue. Stroke is the second leading cause of death worldwide and the leading cause of adult disability. Each year, there are an estimated 15 million new acute strokes, resulting in 28,500,000 disability-adjusted life-years. The 28-30-day case fatality rate for stroke ranges from 17% to 35%.The projected increase in stroke and heart disease-related deaths, from three million in 1998 to five million in 2020, indicates that the burden of stroke is expected to worsen. This increase is attributed to ongoing health and demographic transitions, leading to a rise in vascular disease risk factors and an aging population. Developing countries bear the majority of the stroke burden, accounting for 85% of global stroke deaths. The social and economic consequences of stroke are substantial. In the United States, the cost of stroke was estimated to be as high as $49.4 billion in 2002. Additionally, post-discharge costs were estimated to amount to 2.9 billion Euros in France.

In the United States, stroke is the fifth-leading cause of death, responsible for approximately 11% of total deaths. Over 795,000 individuals in the United States suffer from stroke. In India, stroke is the fourth major cause of death.Machine learning algorithms have shown promise in predicting the occurrence of strokes. While previous studies have primarily focused on heart stroke prediction, this paper aims to predict brain stroke using machine learning. The study utilizes five different classification algorithms and finds that XGBoost performs the best, achieving higher accuracy. It's important to note that the model in this paper is trained on textual data rather than real-time brain images. The dataset used for this task is obtained from Kaggle and contains various physiological traits as attributes for prediction.The paper describes the process of data preprocessing, including handling null values, label encoding, and one-hot encoding if necessary. The dataset is then split into train and test data, and a model is built using various classification algorithms. The accuracy of each algorithm is calculated and compared to identify the best-trained model for stroke prediction. To facilitate user interaction, the paper develops an HTML page and a Flask application. The web application allows users to enter values for prediction, while the Flask application connects the trained model with the web interface.In conclusion, this paper highlights the importance of machine learning in predicting the occurrence of stroke, specifically brain stroke. By utilizing various classification algorithms, the study determines the best-performing algorithm and demonstrates its potential for stroke prediction.

However, it acknowledges the limitation of training on textual data rather than real-time brain images. The paper suggests further research and extension of the study to include all current machine learning algorithms. The causes of mortality from stroke are often related to comorbidities and complications that arise during different time periods. The most critical period for survival and the highest number of fatalities occur within the first month following the onset of stroke symptoms, with the first week being particularly crucial.

Complications of stroke can include various medical conditions such as hyperglycemia, hypoglycemia, hypoglycemia, hypertension, hypotension, fever, infarct extension or rebreeding, cerebral edema, herniation, coning, aspiration, aspiration pneumonia, urinary tract infection, cardiac dysrhythmia, deep venous thrombosis, and pulmonary embolism, among others. During the first week after stroke onset, death is often due to transtentorial, herniation and hemorrhage. Hemorrhagic deaths typically occur within the first three days, while deaths due to cerebral infarction usually happen between the third and sixth day. After the first week, death is usually a result of complications resulting from relative immobility, such as pneumonia, sepsis, and pulmonary embolism. There are traditional risk factors associated with stroke, and understanding and managing these risk factors can help prevent strokes. These risk factors can be divided into modifiable and non-modifiable categories. Modifiable risk factors include lifestyle factors such as smoking, alcohol use, physical inactivity, and obesity, as well as medical factors like high blood pressure, atrial fibrillation, diabetes mellitus, and high cholesterol. Non-modifiable risk factors, such as age, gender, and family history, cannot be controlled but can help identify individuals at risk for stroke. Prevention of stroke is crucial, especially since more than 70% of strokes are first events. Primary stroke prevention focuses on behavior modification and requires information about baseline perceptions, knowledge, and prevalence of risk factors in specific populations. In a related study, the review focuses on acute ischemic stroke, which affects over 700,000 individuals annually in the United States. The study emphasizes the importance of early recognition and aggressive treatment protocols in the emergency department to optimize outcomes. Collaboration among healthcare professionals is crucial for identifying patients within the therapeutic time window for thrombolytic and neuroprotective treatments. A shift in approach is necessary, with healthcare professionals striving for better outcomes by being knowledgeable about early and aggressive evaluation and treatment recommendations for patients with acute ischemic stroke. Additionally, healthcare professionals aim to educate patients and their families about stroke prevention. Understanding the acute and post-acute settings is essential for improving patient outcomes and initiating appropriate rehabilitation and prevention strategies.

II. RELATED STUDY

This study highlights the significant number of individuals, over 700,000, who experience a stroke each year in the United States. In the past, there may have been a skeptical approach to the management of stroke, but there have been considerable changes in the understanding and approach to stroke in the last decade. The concept of "time is brain" emphasizes the importance of prompt action and collaboration among healthcare professionals to initiate acute stroke protocols in emergency departments and identify patients within the therapeutic time window for thrombolytic and neuroprotective treatments. Healthcare professionals aim to achieve the best possible outcomes for individuals who have had a stroke. The skeptical approach towards patients with acute ischemic stroke is no longer suitable.

Present-day healthcare experts recognize the importance of interdisciplinary collaboration to achieve better outcomes by being knowledgeable about early and aggressive evaluation and treatment recommendations. In addition to acute care, healthcare professionals also focus on educating patients and their families about stroke prevention. Understanding stroke starts in the acute setting and continues in the outpatient setting, home, or rehabilitation, with various important aspects of care needing attention.Overall, the study emphasizes the need for a coordinated and proactive approach to stroke management. It underscores the importance of timely intervention, collaboration among healthcare professionals, patient education, and ongoing care to improve outcomes for individuals who have experienced an acute ischemic stroke.

III. PROBLEM STATEMENT

Stroke is the second leading cause of death worldwide and remains an important health burden both for the individuals and for the national healthcare systems. Potentially, modifiable risk factors for stroke include hypertension, cardiac disease, diabetes, and dysregulation of glucose metabolism, atrial fibrillation, and lifestyle factors. Therefore, the goal of our project is to apply principles of machine learning over large existing data sets to effectively predict the stroke based on potentially modifiable risk factors. Then it intended to develop the application to provide a personalized warning on the basis of each user's level of stroke risk and a lifestyle correction message about the stroke risk factors.

IV. LITERATURE SURVEY

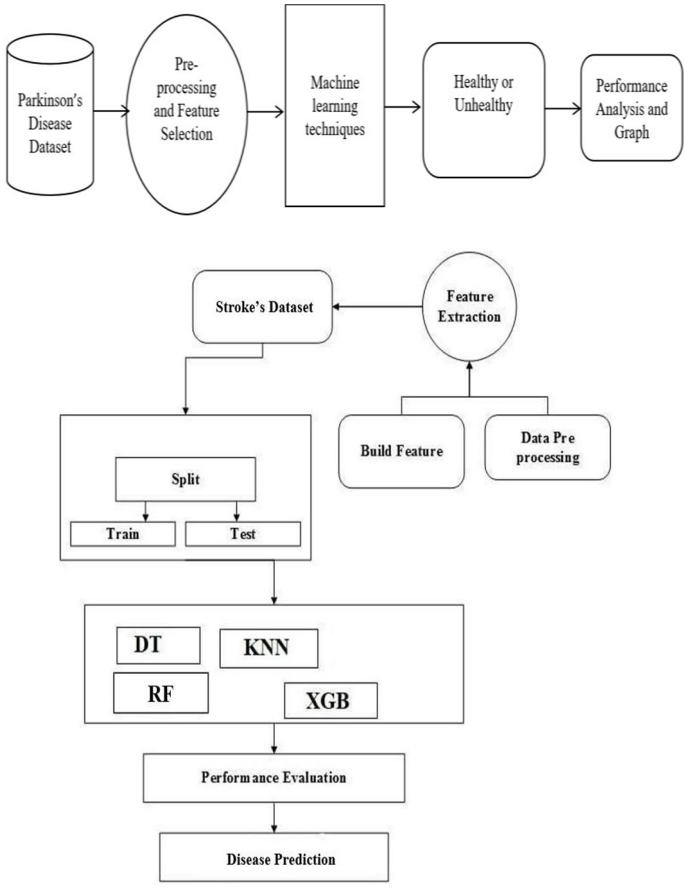
1) "Stroke prediction using artificial intelligence"- M. Sheetal Singh, Prakash Choudhary - In this paper, here, decision tree algorithm is used for feature selection process, principal component analysis algorithm is used for reducing the dimension and adopted back propagation neural network classification algorithm, to construct a classification model.

2) "Stroke Prediction Using Machine Learning Based On Artificial Intelligence"- Youngkeun Choi and Jae Won Choi -In this paper they have worked on datasets of 43,400 patients and 11 predicted attributes from a Kaggle websites. Published in the year 2020.
3) "Machine Learning For Predicting Ischemic Stroke"- Assit.Prof.Pathanjali C, Priya T, Monisha G, Samyuktha Bhaskar, Ruchita Sudarshan- Worked on clinical data contents information about ischemic stroke by using ML algorithms like SVM and random forest classifier and their accuracy is predicted.
4) "Prediction Of Brain Stroke Severity Using Machine Learning Algorithms"- Vamsi Bandi, Debnath Bhatacharyya, Divya Midhun, Chakkravarthy - Worked on medical records based on NIHSS (National Institute of Health Stroke Scale) tool by using ML algorithms like Decision tree, Random forest, Linear SVM, Logistic Regression, Gaussian Naïve Bayer, Poly SVM, RBG SVM, AdaBoost with SGD and found that Random Forest was the best.
5) "Prediction Of Stroke Using Machine Learning"- Akash Mahesh, Shashank H N, Shrikanth S, Thejas A M - Worked on dataset to predict the stroke using ML Algorithms like ANN, Decision Tree, Navie Bayer and they founded that ANN is the best performing algorithm.

V. METHODOLOGY

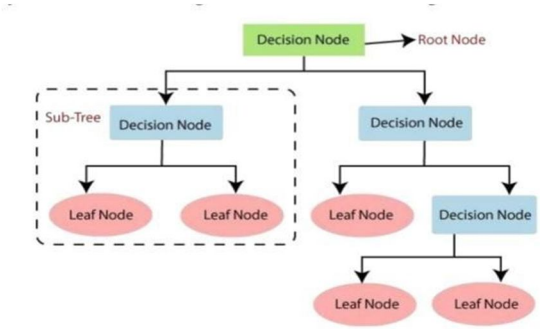
In this project, this section represents the Methods of the Project including machine learning techniques. Stroke Disease data sets have been considered in this project.

System Design:



1) *Data Collection:* The dataset used in this study was collected from a medical college institute. It consists of 1500 samples, with 1000 male and 500 female patients. The dataset includes 30 features related to patient history, hospital details, risk factors, and symptoms. Risk factors such as age, gender, blood pressure, chest pain, alcohol consumption, diabetes, headache, cigarette smoking, family history, hypertension, cholesterol levels, heart rate, face deficit, arm/hand deficit, leg/foot deficit, visuospatial disorder, and blood vessels were analyzed carefully for stroke prediction.

2) *Data Preprocessing:* The dataset presented some challenges due to a significant number of missing values and a large number of features. To improve the classification performance, several preprocessing techniques were applied. This included removing duplicate records, handling inconsistent and noisy data, and dealing with missing data. Features with missing values were removed from the database. The dataset was carefully selected to reduce complexity and improve the accuracy of the classification.
3) *Dimension Reduction:* Dimension reduction aims to reduce the dimensionality of the dataset by obtaining a set of principal factors that capture the most important features for classification. In this study, Principal Component Analysis (PCA) was used for dimension reduction. PCA helps to reduce the complexity associated with high-dimensional data, allowing for more efficient and effective classification.



4) *Classification Algorithm:* In this study we employed machine learning techniques for stroke prediction. Random Forest model is and it acts as a concept extractor for the input dataset. The model parameters are automatically adjusted using the generalized. When user details are entered, the trained model is used to predict whether the person is getting stroke or not. Overall, this study combines data collection, preprocessing, dimension reduction, and the application of deep learning techniques to predict strokes based on various risk factors and symptoms.

5) *Attributes used for prediction:*

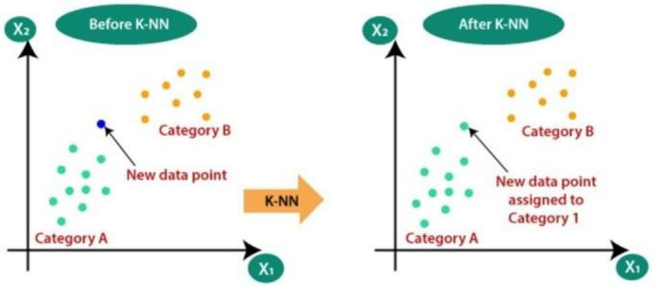
```
gender
age
hypertension
heart_disease
ever_married
work_type
Residence_type
avg_glucose_level
bmi
smoking_status
stroke
dtype: int64
```

6) *First five rows of the dataset:*

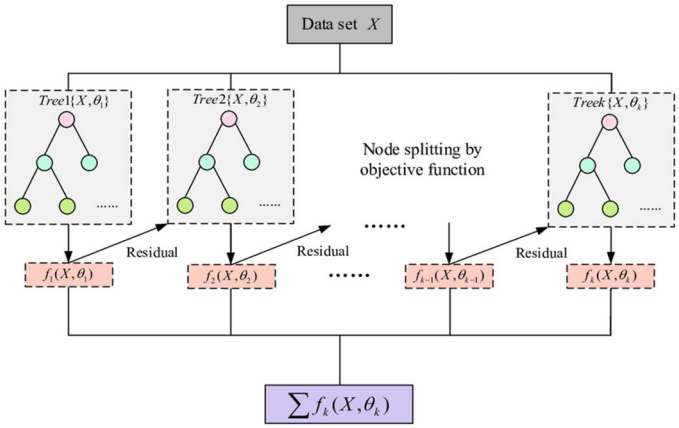
	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1

VI. MACHINE LEARNING TECHNIQUES

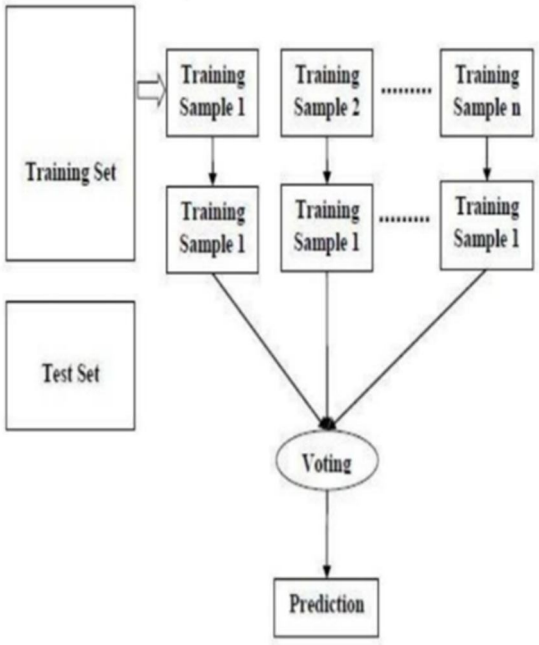
1) *Decision Tree Classifier:* Both regression and classification concerns are addressed using classification with DT. Furthermore, as the input variables already have a related output variable, this methodology is a supervised learning model. It resembles a tree the data is constantly segmented according to a specific parameter in this method. The decision node and the leaf node are the two parts of a decision tree. At the former node, the data is divided, and the latter is the node that produces the result. It may be very beneficial in resolving issues with decision-making.
2) *KNN Classifier:* K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.



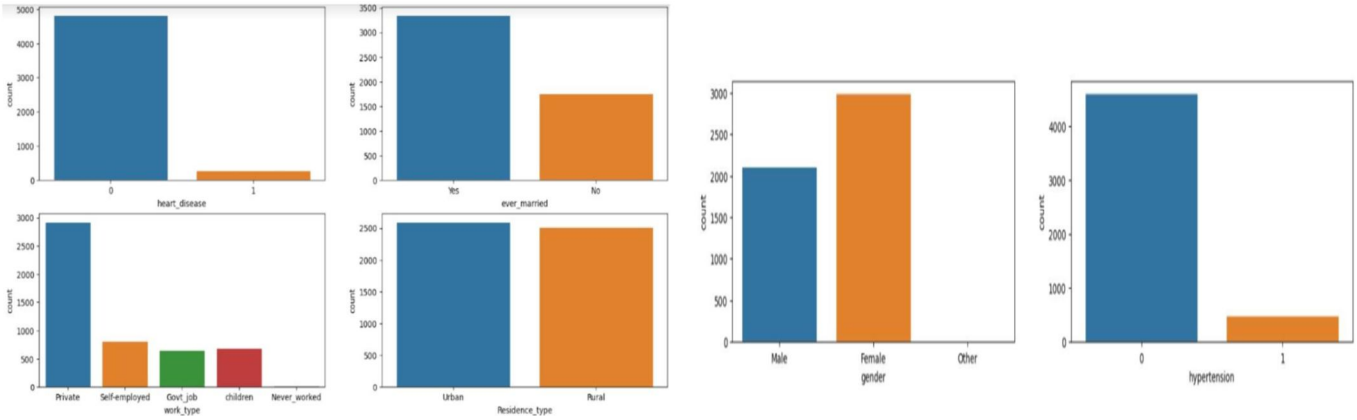
3) *XGBoost:* The XGBoost (eXtreme Gradient Boosting) is a popular and efficient open-source implementation of the gradient boosted trees algorithm. Gradient boosting is a supervised learning algorithm that attempts to accurately predict a target variable by combining an ensemble of estimates from a set of simpler and weaker models. XGBoost is an optimized distributed gradient boosting library designed for efficient and scalable training of machine learning models. It is an ensemble learning method that combines the predictions of multiple weak models to produce a stronger prediction. XGBoost stands for "Extreme Gradient Boosting" and it has become one of the most popular and widely used machine learning algorithms due to its ability to handle large datasets and its ability to achieve state-of-the-art performance in many machine learning tasks such as classification and regression.

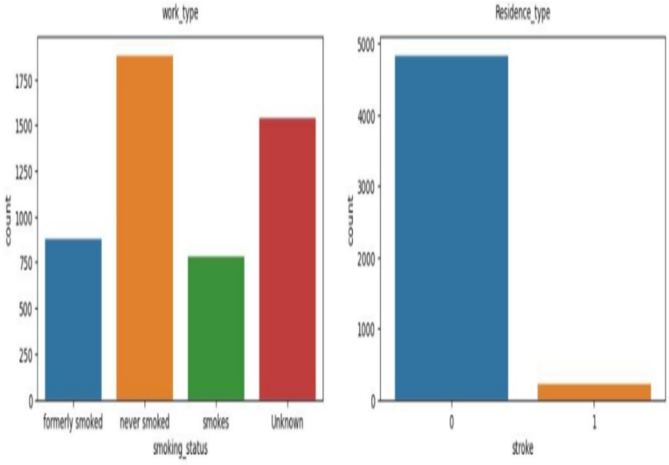


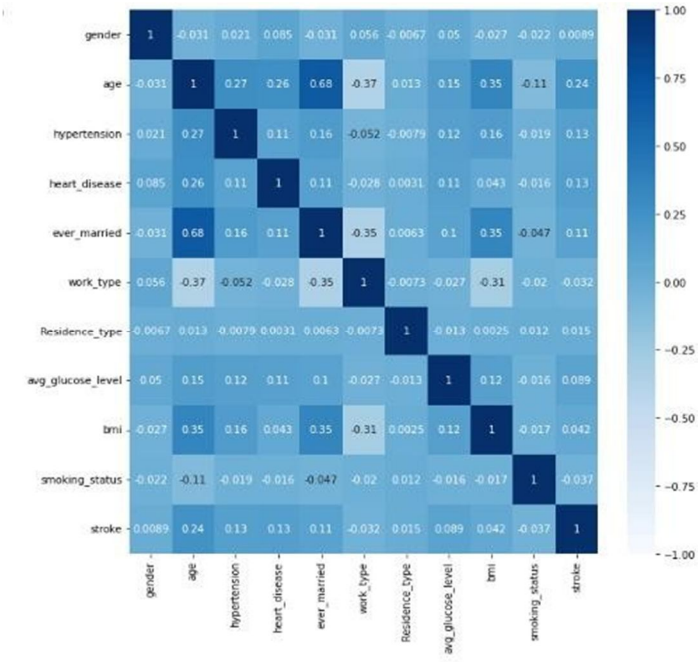
4) *Random Forest:* Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. "The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting. It takes less training time as compared to other algorithms. It predicts output with high accuracy, even for the large dataset it runs efficiently. It can also maintain accuracy when a large proportion of data is missing. Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees.



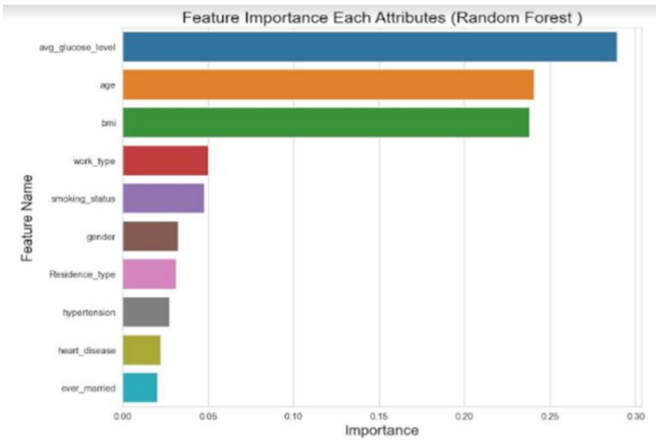
VII. DATA ANALYSIS

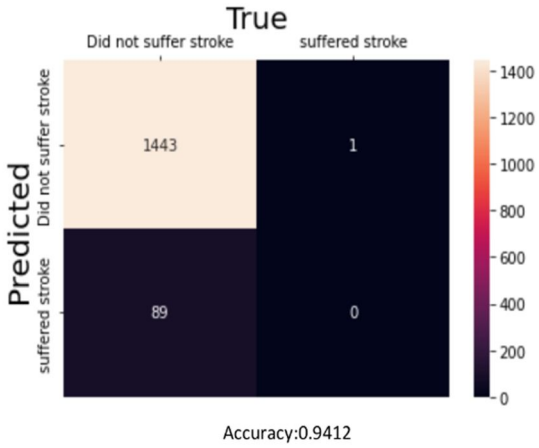






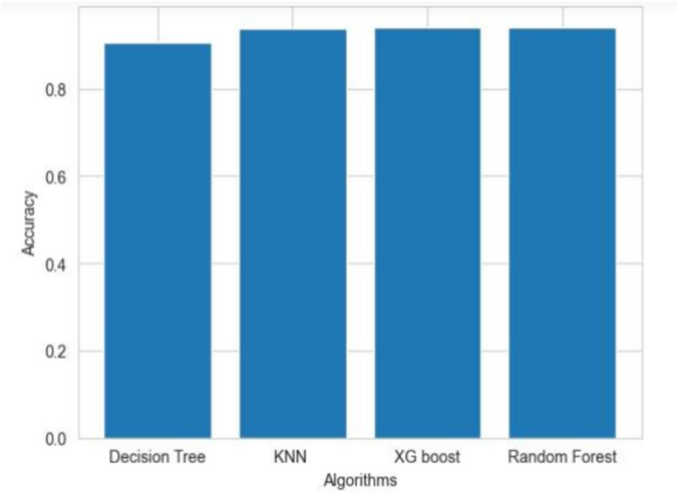
A. Correlation Matrix





Confusion matrix of Random Forest Algorithm

VIII. RESULT ANALYSIS



These are the variation of accuracy of all the four algorithms.

IX. PERFORMANCE EVALUATION MEASURE

Various evaluation matrices were used for checking the performance of the classifier. For this purpose, the confusion matrix was used. It is a 2∗2 matrix due to two classes in the dataset. The confusion matrix gives two types of correct prediction of the classifier and two types of incorrect prediction of the classifier. The confusion matrix is presented above.

1) *Confusion Matrix Description TP:* True Positive means output as positive such that predicted result is correctly classified. TN: True Negative means output as negative such that predicted result is correctly classified. FP: False Positive means output as positive such that predicted result is incorrectly classified. FN: False Negative means output as negative such that predicted result is incorrectly classified.

	Positive (1)	Negative (0)
Positive (1)	TP	FN
Negative (0)	FP	TN

The dataset is spitted into 2 parts: training data and testing data. 70% of the data was taken for training and 30% for testing.

2) *Classification Accuracy:* Classification accuracy shows the correct rate of prediction results. It computes from the confusion matrix. The classification accuracy is found by equation 2
3) *Classification Error:* Classification error shows the incorrect rate of prediction results. It computes from the confusion matrix. The classification error is found by equation 3:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} * 100$$
$$Error = \frac{TP + TN + FP + FN}{TP + TN + FP + FN} * 100$$

4) *Precision:* Precision is an important model performance evaluation matrix. It is the fraction of related instances among the total retrieved instances. It is a positive predicted value. The precision is calculated as follows in equation 4:

$$Precision = \frac{TP}{TP + FP} * 100$$

$$Recall = \frac{TP}{TP + FN} * 100$$

5) *Recall:* Recall is also an important model performance evaluation matrix. It is the fraction of related instances among the total number of retrieved instances. The recall is calculated as follows in equation 5:

6) *F-Measure* It is also known as F Score. F-measure is calculated so as to measure the accuracy of test. It is calculated from the precision and recall by equation 6:

$$F - Measure = 2^* \frac{Precision * Recall}{Precision + Recall}$$

7) *ROC and AUC:* The performance of the classification model is measured from the Receiver operating a characteristic curve (ROC). ROC is a graph that is created for true positive rate vs. false positive rate at different classifications threshold. The entire area under the ROC curve is known as area of the curve (AOC). It gives a collective measure of performance across all achievable classification's threshold.
8) *GINI Coefficient:* It is also known as GINI index. It is a measure of statistical distribution. It is used to measure the inequality amongst values of attributes. It is also possible to say that it calculates the impurity of a particular attribute in the form of degree or probability.

X. CONCLUSION

This article objects to predict stroke disease based on full features and important features of stroke dataset's. For feature selection co-relation based feature selection technique is used. In this perception core machine learning algorithms were appeared like Decision tree Classifier, KNN Classifier, Random Forest and XGBoost. For each algorithm the results were computed based on selected features. It is observed that Random Forest Algorithm has highest accuracy of 94419 among all the other algorithms considered. The stroke prediction using machine learning method can be used to find whether the patient having stroke or not. The early warning can save someone's life who might have a probability of a stroke. This project by using Random Forest Algorithm predicts the probability of stroke on the basis of very trivial day-to-day and known to all parameters which makes this project highly relevant and of need to society.

	Metric	DT	KNN	XGB	RF
0	Accuracy	0.906067	0.938030	0.940639	0.941292
1	F1-Score	0.217391	0.020619	0.165138	0.021739
2	Recall	0.224719	0.011236	0.101124	0.011236
3	Precision	0.210526	0.125000	0.450000	0.333333

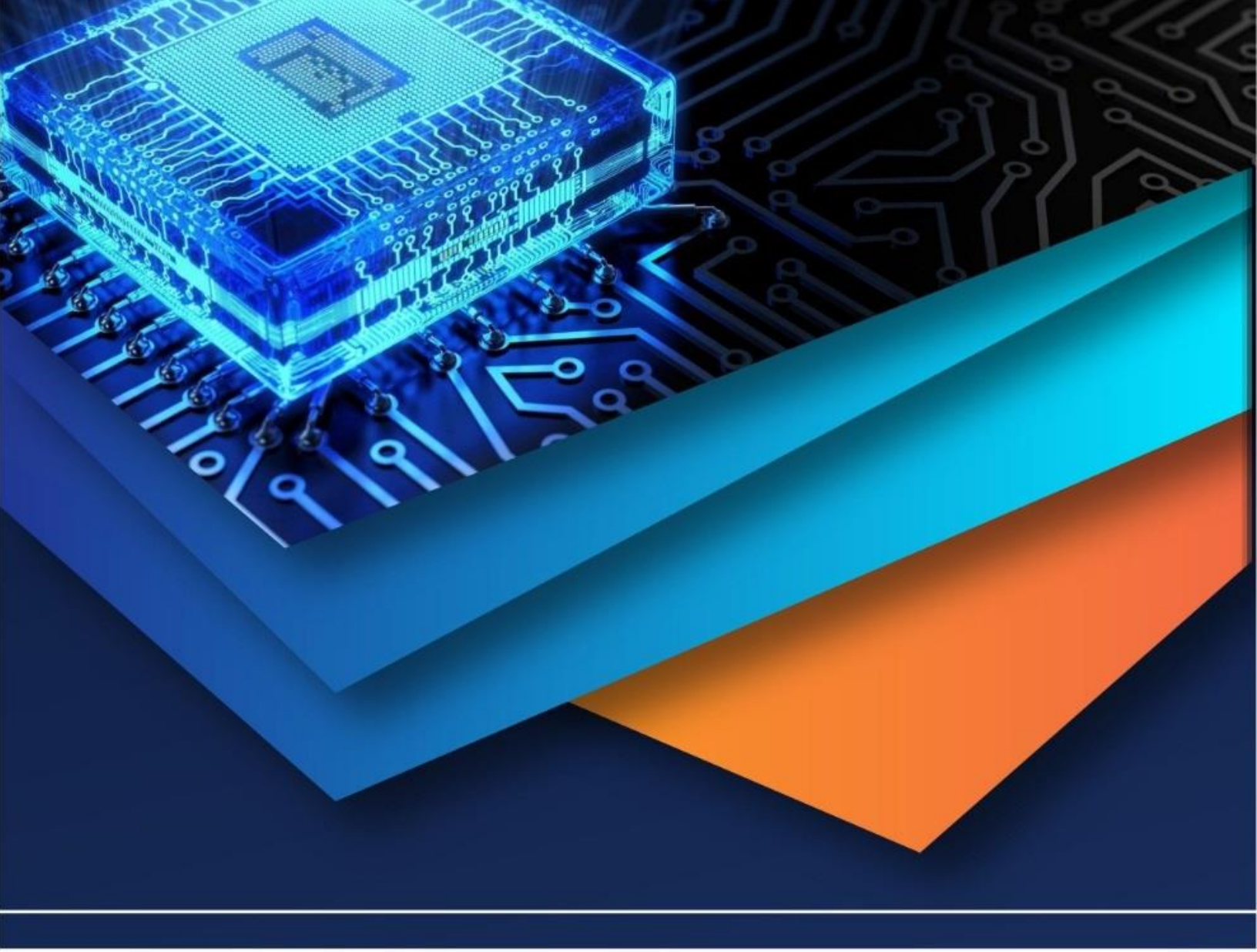
ALGORITHMS	ACCURACY
Decision Tree	0.906
KNN Classifier	0.938
Random Forest	0.941
XGBoost	0.940

XI. FUTURE WORK

This project helps to predict the stroke risk using prediction model in older people and for people who are addicted to the risk factors as mentioned in the project. In future, the same project can be extended to give the stroke percentage using the output of current project. This project can also be used to find the stroke probabilities in young people and underage people by collecting respective risk factor information's and doctors consulting.

REFERENCES

[1] https://www.researchgate.net/publication/342437236_Pred iction_of_Stroke_Using_Machine_Learning
[2] https://ijirem.org/DOC/2-stroke-prediction-using- machine-learning-algorithms.pdf
[3] https://thesai.org/Downloads/Volume12No6/Paper_62Ana lyzing_the_Performance_of_Stroke_Prediction.pdf
[4] "Stroke prediction using artificial intelligence"- M. Sheetal Singh, Prakash Choudhary
[5] "Prediction Of Stroke Using Machine Learning"- Akash Mahesh, Shashank H N, Shrikanth S, Thejas A M
[6] "Machine Learning For Predicting Ischemic Stroke"- Assit.Prof.Pathanjali C, Priya T, Monisha G, Samyuktha Bhaskar, Ruchita Sudarshan
[7] "Stroke Prediction Using Machine Learning Based On Artificial Intelligence"- Youngkeun Choi and Jae Won Ch
[8] Stroke Prediction Dataset | Kaggle
[9] Introduction to Logistic Regression | by Ayush Pant | Towards Data Science
[10] Machine Learning Tutorial (tutorialspoint.com)
[11] Machine Learning - GeeksforGeeks
[12] Stack Overflow



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY