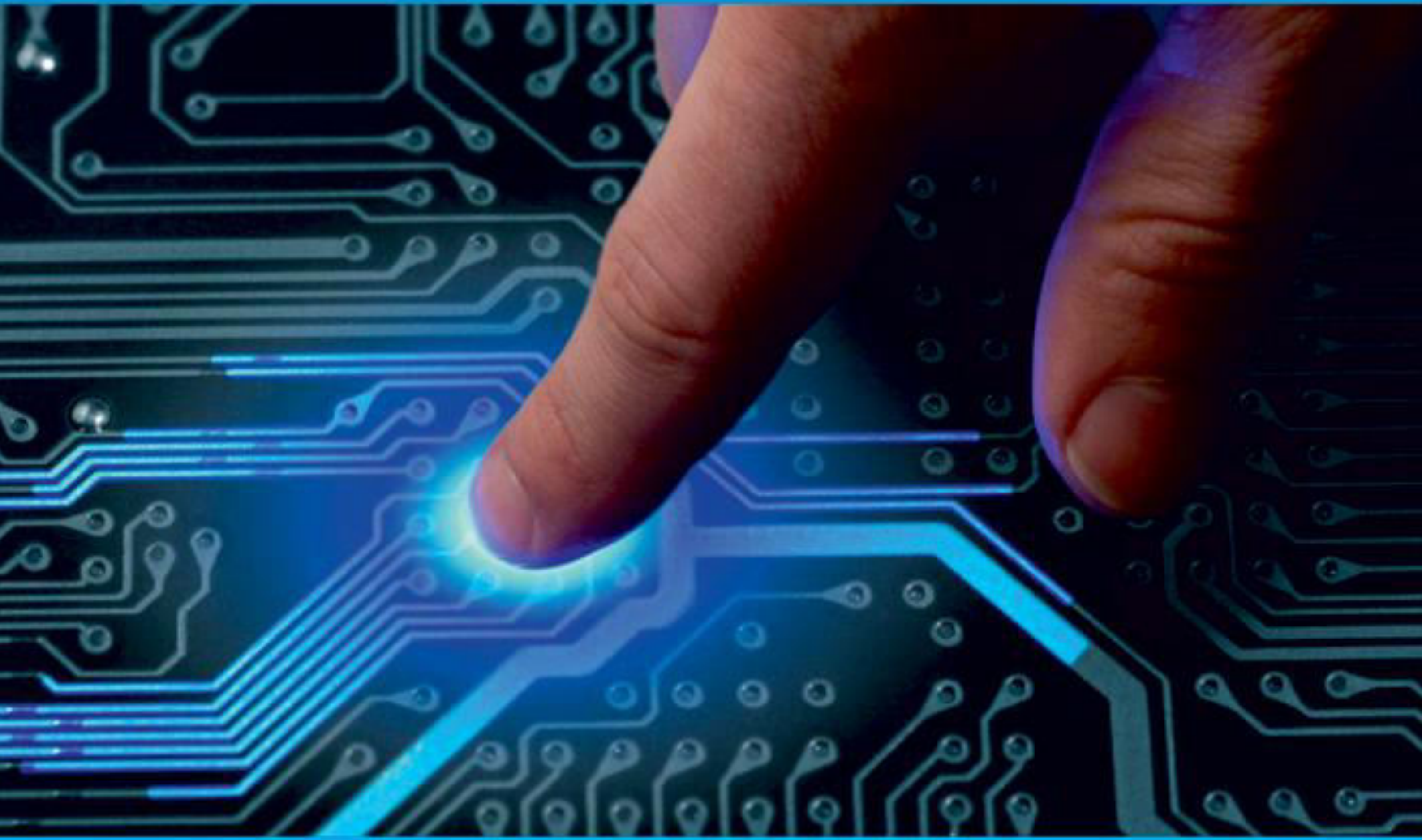
IJIRCCE

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

Stroke Prediction System Using M.L

Prof. Vinod Bhamare[1], Saishashank N. Petkar[2], Renuka D. Joshi[3], Aditya M. Nikam[4], Sonali S. Bagul[5]

Assistant Professor, Dept. of Computer Engineering, Sandip Institute of Technology and Research Centre

Nashik, Maharashtra, India[1]

UG Student, Dept. of Computer Engineering, Sandip Institute of Technology and Research Centre

Nashik, Maharashtra, India[2,3,4,5]

ABSTRACT: Stroke, a cerebrovascular malady, stands as a noticeable donor to worldwide mortality, forcing significant wellbeing and monetary challenges on people and healthcare frameworks alike. Strikingly, health-related conduct has developed as a vital determinant of stroke chance, picking up expanded consideration in preventive endeavours. Different machine learning models have been formulated to estimate stroke hazard or encourage robotized stroke conclusion, leveraging way of life components and radiological imaging indicators. Shockingly, there is a outstanding nonattendance of models utilizing information from research facility tests. As the moment driving cause of passing around the world, stroke continues as a critical wellbeing burden. This extend utilizes machine learning standards on broad existing datasets to foresee stroke hazard based on possibly modifiable variables. The ensuing objective includes creating an application to convey personalized notices agreeing to person stroke chance levels, went with by way of life redress messages tending to particular stroke chance variables. Stroke, a critical wellbeing concern universally, requires convenient recognizable proof and mediation to avoid serious results. Avoiding strokes includes advancing wellbeing and raising mindfulness of chance variables.

KEYWORDS: stroke, research facility tests, machine learning innovation, prescient analytics.

I. INTRODUCTION

A neurological deficit called a stroke frequently consequences from acute damage to the relevant worried machine because of a vascular problem. Globally, it stands as a first-rate contributor to disability and mortality. within the US, the prevalence of stroke is anticipated at 2.5%, impacting over 7 million humans aged 20 and above. The circumstance considerably diminishes sufferers' health and normal pleasant lifestyles, additionally straining sanatorium services and mattress availability. The economic toll in the United States of America alone is expected to be around US $351.2 billion between 2014 and 2015. There are essential forms of stroke: ischemic and haemorrhagic. Haemorrhagic strokes stem from ruptured vessels inflicting bleeding within the brain, even as ischemic strokes result from arterial blockages inside the mind, constituting 85% to 90% of all strokes.[1]

Preventative measures, inclusive of the merchandising of fitness and attention to danger factors, play an important role in mitigating the incidence of stroke. way of life factors including obesity, nutritional habits, alcohol consumption, and bodily inactivity contribute to the threat. Moreover, underlying conditions like diabetes, high blood pressure, and cardiovascular illnesses are connected to stroke. thus, powerful self-management of these situations and the adoption of a wholesome way of life can function as preventive measures.

Stroke represents a debilitating and probably lifestyles-threatening medical circumstance, providing a giant international health mission. in keeping with the sector health enterprise (WHO), strokes account for about 11% of all international deaths, rating as the second leading cause of mortality. moreover, strokes frequently result in extreme disabilities, placing a big burden on healthcare systems and adversely affecting the fine of life for the ones affected.

II. LITERATURE SURVEY

In research [2] focuses on the use of machine learning algorithms for stroke prediction, a crucial medical condition that requires early diagnosis. They apply various classification algorithms, including Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbours, Support Vector Machine, and Naïve Bayes, to predict stroke risk. The authors compare their performance metrics, such as accuracy, precision, recall, and F1 score, to identify the most effective algorithm for stroke prediction. This could lead to the development of machine learning models with high accuracy, enabling early diagnosis and improved patient outcomes.

The research [3] propose an explainable ML model which is Random Forest for stroke prediction, which not only predicts stroke risk but also provides insights into the factors influencing it. This interpretability is crucial for medical professionals, as it allows for more targeted preventive measures and personalized treatment plans. The model can also be used by medical professionals or patients to assess stroke risk, potentially promoting early intervention and improving overall stroke care. This approach could be beneficial for both medical professionals and patients.

The research [4] highlights the potential of machine learning model that is Logistic Regression in stroke prediction, arguing that understanding the reasoning behind predictions is equally important. They propose an ML system that prioritizes both accuracy and explain ability, incorporating explainable AI techniques to predict stroke risk and reveal factors influencing those predictions. This interpretability is valuable for medical professionals, enabling targeted preventive measures and personalized treatment plans, potentially leading to improved patient care and better stroke outcomes.

The research [5] acknowledge the potential of the given 5 Machine Learning algorithm given a highest accuracy of 98.56%. for stroke prediction but highlight a limitation in current research. They focus on core risk factors like age, gender, underlying medical conditions, and smoking history, aiming to achieve accurate prediction and enhance model interpretability. This approach may make predictions more transparent to medical professionals, enabling them to make informed decisions about patient care. This could potentially improve the effectiveness of ML-based stroke prediction in clinical settings. Stroke is a major cause of death worldwide, and timely intervention is crucial to minimize disability and improve patient prognosis.

The research [6] explores the effectiveness of proposed methods which are compared with several existing state-of-the-art ML algorithms, that is, RF, decision tree, naive Bayes, and ST algorithms in predicting stroke risk early. They acknowledge the growing body of research on ML for stroke prediction, but a critical gap exists in comparing the performance of different algorithms. The study aims to address this gap by conducting a comparative analysis of ML algorithms' ability to predict stroke risk effectively. This can help identify the most suitable ones for early stroke prediction, improving the accuracy and efficiency of ML-based systems in clinical settings. This analysis could lead to faster diagnosis and improved patient outcomes.

The research [7] proposes an improvised Random Forest for predicting stroke severity, addressing the research gap in stroke prediction. The authors explore the use of various ML algorithms to analyse patient data, including demographics, medical history, imaging scans, and neurological assessments. By training the model on data with known stroke severity levels, they aim to achieve accurate predictions for new patients. This information can be invaluable for medical professionals, allowing them to tailor treatment plans and prioritize patients. Early assessment of stroke severity can also help allocate resources efficiently within healthcare systems. The study aims to address the research gap in stroke prediction and improve patient outcomes.

The research [8] focuses on the use of machine learning (ML) for stroke detection and prediction. They argue that early detection is crucial for minimizing brain damage and improving patient outcomes. ML offers a promising alternative by analysing patient data and predicting stroke risk. Previous research on ML for stroke prediction has focused on high accuracy, but these studies often lack robustness, making them less reliable in real-world applications. This research aims to develop robust and reliable ML-based tools for stroke detection and prediction, potentially leading to earlier diagnosis, improved treatment decisions, and better patient outcomes.

The research [9] explores the use of machine learning algorithms for developing a stroke prediction system, highlighting the importance of early detection for minimizing brain damage and improving patient outcomes. The model uses patient data, including demographics, medical history, lifestyle habits, and biological markers, to learn patterns and relationships between these factors and stroke occurrence. Once trained, the model can predict stroke risk in new patients, helping medical professionals take preventive measures and personalize treatment plans. However, further research is needed to understand the specific algorithms used and their effectiveness in predicting stroke risk.

The research [10] explores the use of prediction models, Decision Tree, Naive Bayes and Neural Network, showed acceptable accuracy in identifying stroke-prone patients for stroke prediction, a major global health concern. The authors propose using ML algorithms to analyse patient data and predict stroke risk. The models rely on factors like demographics, medical history, lifestyle habits, and biological markers to learn patterns and relationships. Once trained, the model can predict stroke risk in new patients based on individual data, enabling early intervention and personalized treatment plans. However, further research is needed to understand the specific ML algorithms used and their effectiveness in predicting stroke risk.

The paper [11] explores the use of machine learning (ML) algorithms for stroke prediction. The authors suggest that ML models use patient datasets containing demographics, medical history, lifestyle habits, and biological markers to learn patterns and relationships between these factors and stroke occurrence. Once trained, these models can predict stroke risk in new patients based on individual data, helping medical professionals take preventive measures and personalize treatment plans, leading to improved patient outcomes. However, further research is needed to understand the specific ML algorithms used and their effectiveness in predicting stroke risk.

III. SYSTEM METHODOLOGY

Our system for stroke prediction using machine learning follows a series of steps to ensure data quality, model effectiveness, and user-friendly deployment.

First, we curated an appropriate dataset from Kaggle, a platform for open-source data. Data pre-processing was then performed to clean and organize the data for machine learning algorithms. This included handling missing values and transforming categorical variables into numerical representations suitable for the models.

Next, we built five machine learning models: Logistic Regression, Decision Tree Classification, Random Forest Classification, K-Nearest Neighbours, and Support Vector Classification. These models were trained on the pre-processed data to learn the relationships between various factors and stroke risk.

To evaluate the effectiveness of each model, we employed five different accuracy metrics. These metrics included Accuracy Score (overall correctness), Precision Score (proportion of true positives among predicted positives), Recall Score (proportion of true positives identified), F1 Score (harmonic mean of precision and recall), and Receiver Operating Characteristic (ROC) curve (visual representation of model performance). By comparing these metrics across all five models, we were able to identify the model that yielded the most accurate and reliable stroke risk predictions.

The chosen model was deployed via a web application, allowing user interaction through an HTML page. Flask, a Python framework, was used to connect the web application to the machine learning model. The model analyses user data and predicts stroke risk, which is displayed on the web page. This user-friendly interface facilitates easy access to the stroke prediction model, potentially aiding in early stroke detection.

The system follows a workflow outlined in Fig. 1[18] for data preparation, model building, evaluation, and deployment.

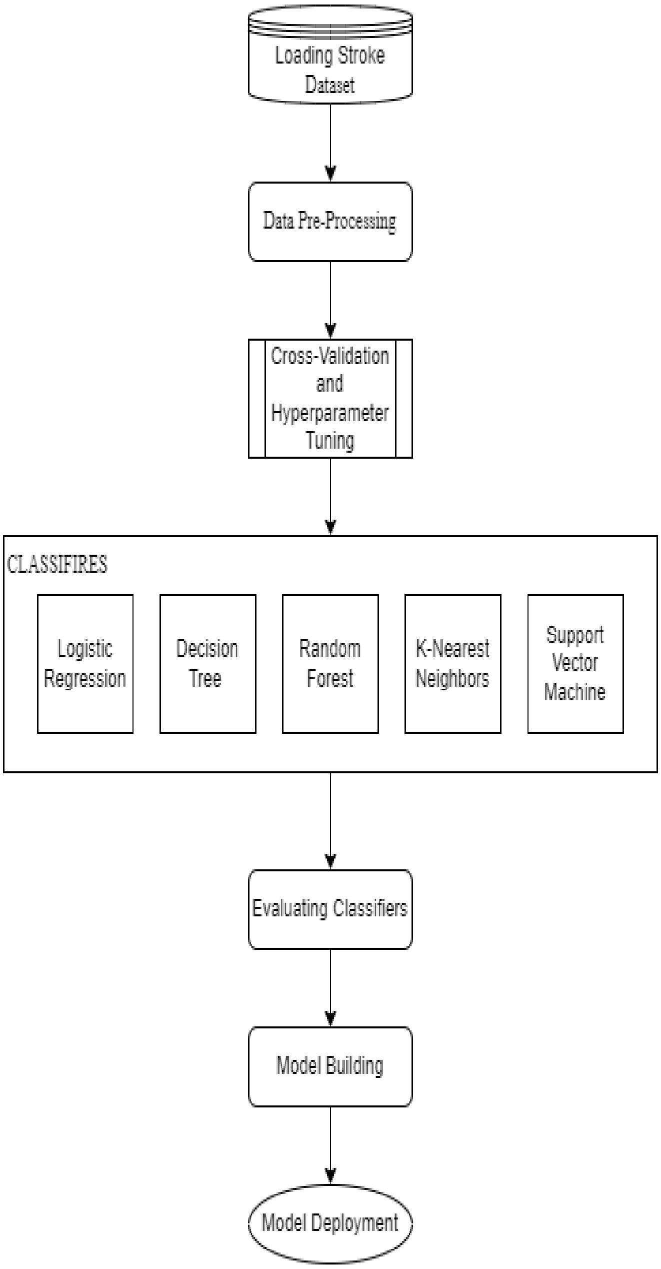


Fig. 1: Proposed System

IV. IMPLEMENTATION

The implementation of our project is as follows:

1. Dataset Selected:
   The dataset selected from Kaggle for stroke pre-diction [12]. This dataset has 4909 rows and 12 columns. The columns have 'id', 'gender', 'age', 'hypertension', 'heart_disease', 'ever_married' , 'work_type', 'Residence_type' , 'avg_glucose_level' , 'bmi' , 'smoking_status' and 'stroke' as the main attributes. The output column 'stroke' has the value as either '1' or '0'. The value '0' indicates no stroke risk detected, whereas the value '1' indicates a possible risk of stroke. The dataset discussed above is summarized in Table 1.

Table 1: Stroke Dataset

Attribute	Description
id	Unique identifier for each patient
gender	"Male" or "Female"
age	Patient's age in years
hypertensive	Whether the patient has hypertension (0 - No, 1 - Yes)
heart_disease	Whether the patient has heart disease (0 - No, 1 - Yes)
ever_married	Marital status ("No" or "Yes")
work_type	Type of work the patient does (e.g., "Government", "Private", "Self-employed")
residence_type	Where the patient lives ("Urban" or "Rural")
avg_glucose_level	Average blood glucose level in mg/dL
bmi	Body mass index
smoking_status	Smoking habits ("formerly smoked", "never smoked", "smokes")
stroke	Whether the patient experienced a stroke (0 - No, 1 - Yes)

2. Data Preprocessing:
Before developing a model, data preprocessing is necessary to eliminate undesired noise and outliers from the dataset that could cause the model to deviate from appropriate training. This step handles anything that prevents the model from operating as efficiently as possible. The next stage is to clean the data and make sure it is prepared for model construction. Table 1 lists the 12 attributes of the dataset that are used. First off, since the column "id" is not really important for model creation, it is removed. The dataset is then examined for null values, and any discovered are filled in. Label encoding is the following task.

3. Label Encoding:
To make the dataset's string literals understand-able to machines, label encoding converts them into integer values. It is necessary to transform the strings into integers since the computer is often taught using numbers. In the gathered dataset, strings are the data type for five of the columns. Label encoding converts all of the strings into a combination of numbers, encoding the whole dataset.

V. MODEL BUILDING

1. Splitting the Data:
The next stage after preparing the data is to develop the model. To improve accuracy and efficiency for this work, the data is divided into training and testing sets, with the ratio remaining at 80% training and 20% testing. The model is trained using a variety of classification techniques as follows. Random Forest Classification, K-Nearest Neighbors Classification, Decision Tree Classification, Support Vector Machine, and Logistic Regression are the classification techniques utilized for this purpose.

2. Classification Algorithms:

   A. Decision Tree Classification: Regression and classification issues are resolved by decision tree classification [14]. This approach functions as a supervised learning technique in which the output variables are pre-correlated with the input variables. Its structure resembles a tree. The data in this method is continuously divided based on a specific parameter. The decision tree classification method produced an accuracy of 94.7% in this instance of stroke prediction. The precision score is 89.3% recall scores is 87.8%. This algorithm yielded an F1 Score of 88.5%. Fig. 2 displays the receiver operating characteristic (ROC) curve for decision tree classification, which is 92.3%.

B. Logistic regression: A supervised learning technique called logistic regression is utilized to forecast the likelihood of the output variable [13]. When the output variable includes binary values, this technique fits the data the best (0 or 1). Since there are only two potential values for the output attribute in the dataset, logistic regression is used. The accuracy attained after applying this technique on the dataset is 94.7%. Other accuracy measures, such as precision score which is 89.3%, recall score which is 87.8%, can also be used to determine the algorithm's efficiency. This method yielded an F1 Score of 88.5%. As seen in Fig. 3, the Receiver Operating Characteristic (ROC) curve for Logistic Regression is 95.8%.

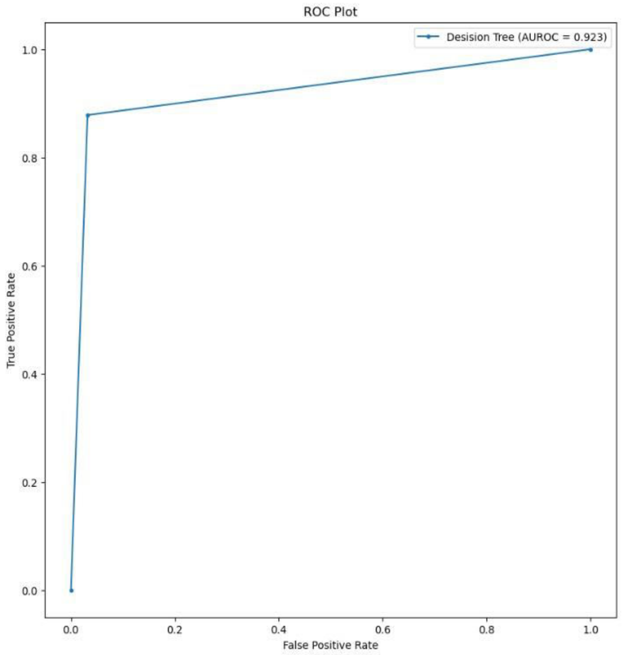


Fig 2: Decision Tree Classification ROC Curve.

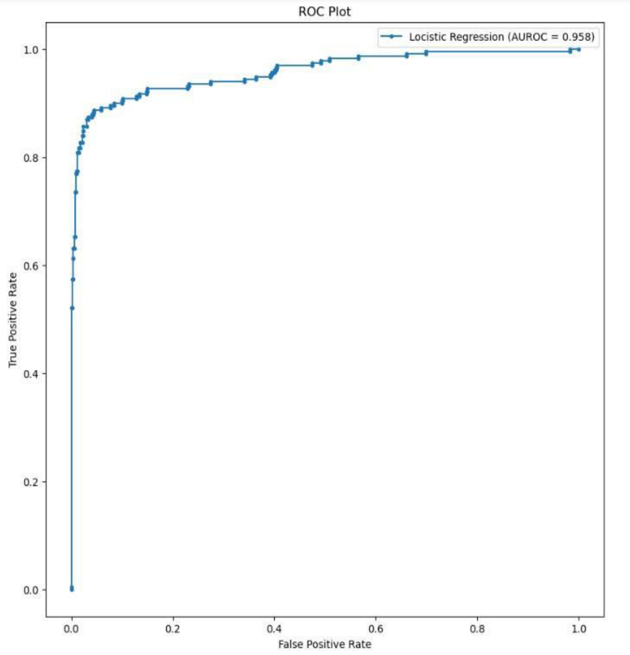
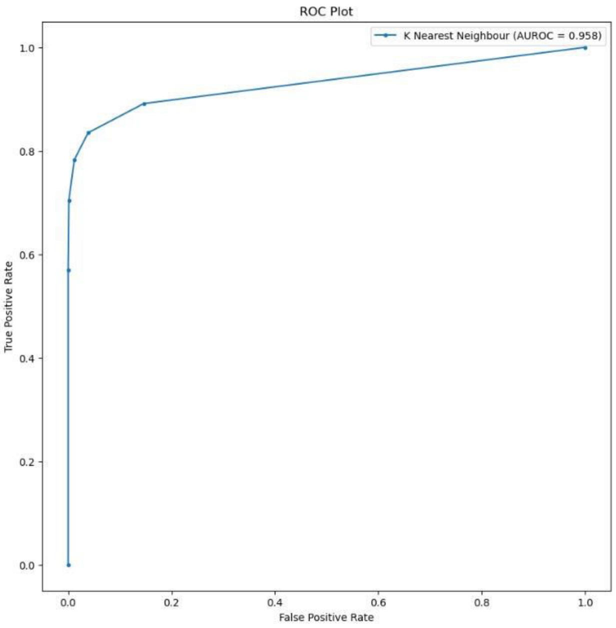


Fig. 3: Logistic Regression ROC Curve.

C. K-nearest neighbors classification: K-Nearest Neighbors (KNN) is the classification algorithm that is employed. It is a method of supervised learning as well. The lazy algorithm KNN [15] would not train right away after receiving the dataset. Rather, it operates on the dataset after storing it for the purpose of categorization. Finding similarities between the new instance (or data) and the existing data is the fundamental idea behind KNN. KNN produced a accuracy of 93.9%. The precision score is 95.2% recall scores is 78.2%. This algorithm yielded an F1 Score of 85.9%. Fig. 4 displays the receiver operating characteristic (ROC) curve for KNN, which is 95.8%.

D. Random Forest classification: Multiple independent decision trees that have each been trained separately on a different subset of data make up Random Forests [16]. The outputs are derived from every decision tree, which are created throughout the training process. There is a process known as "voting" for the algorithm's final prediction. By using this strategy, votes for an output class—in this case, "stroke" and "no stroke"—are cast for each decision tree. The class with the most votes is selected by the random forest as the winning guess. Random Forest produced an accuracy of 97%. The precision score is 99.9% recall scores is 87.3%. This algorithm yielded an F1 Score of 93.2%. Fig. 5 displays the receiver operating characteristic (ROC) curve for Random Forest, which is 97%.



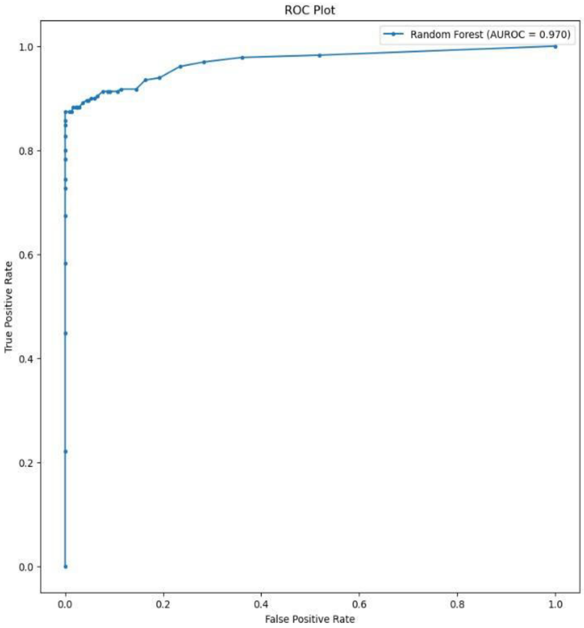


Fig. 4: KNN classification ROC Curve.

Fig. 5: Random Forest classification ROC Curve.

E. Support vector machine: It is a method of supervised learning that may be used with learning algorithms to examine data for regression and classification. Support Vector Machine (SVM) [17] scales to high-dimensional data quite effectively. Support vector machine produced a accuracy of 94.8%. The precision score is 95.4% recall scores is 81.7%. This algorithm yielded an F1 Score of 88%. Fig. 6 displays the receiver operating characteristic (ROC) curve for Support vector machine, which is 96.7%.

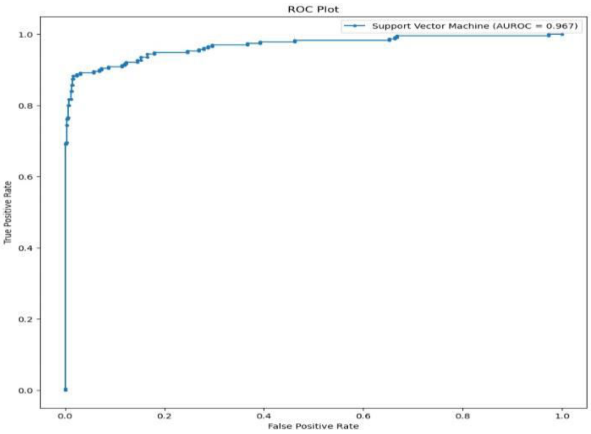
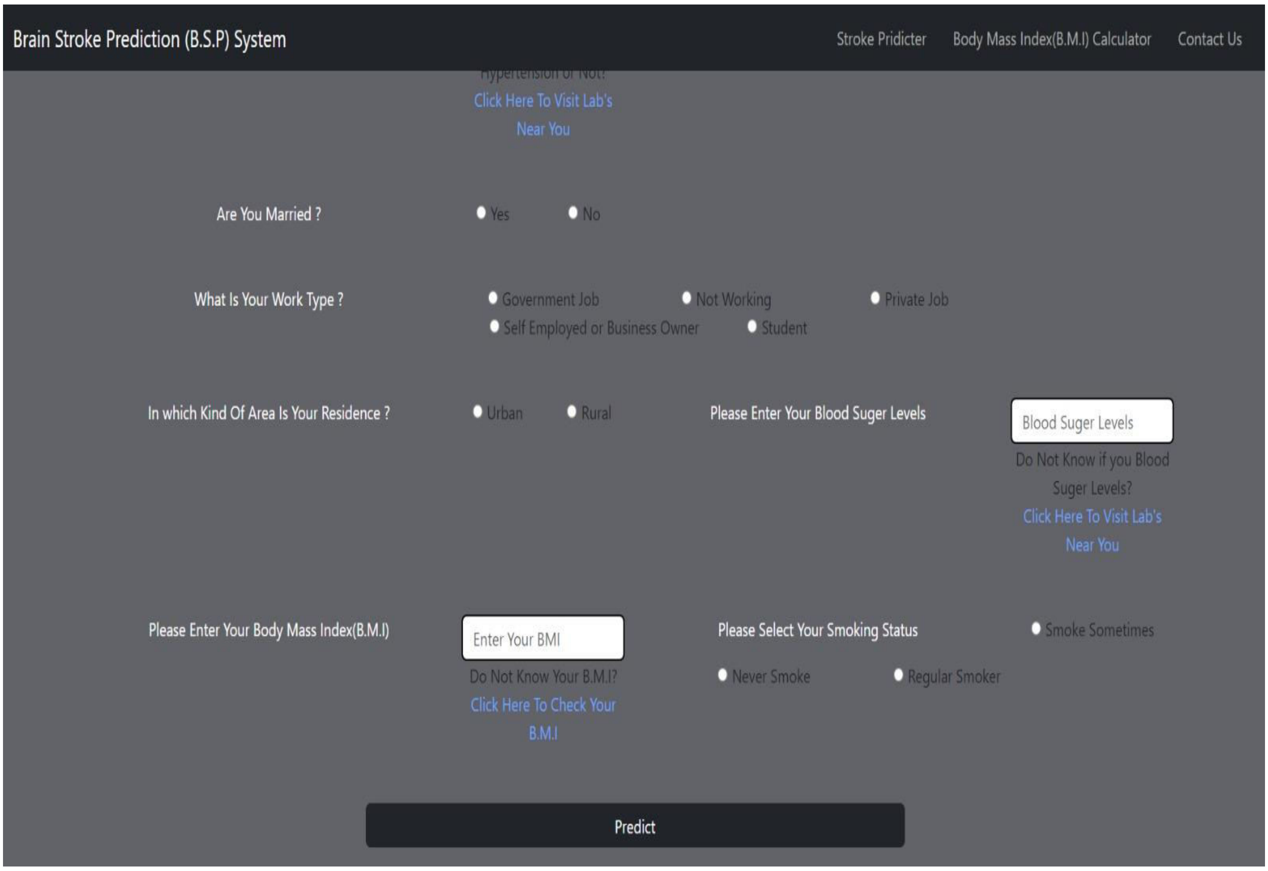


Fig. 6: Support vector machine ROC Curve.

VI. DEPLOYMENT

As we have completed the model building phase. Let’s look at the next step that is project deployment. As we have observed in the model building phase that “Random Forest classification” algorithm outperforms all other algorithms. So, the model trained using Random Forest classification is dumped using pickle. The next task is to develop a web application using flask framework to enter the input parameters.

The web page is built using HTML, CSS, Bootstrap, JavaScript code. This application features an input form that uses the user's input values to forecast the likelihood of a stroke occurring. When the user clicks on the 'Predict' button, the entered parameters are passed to the model using flask framework. Snaps of input form are shown in Fig. 7.



Brain Stroke Prediction (B.S.P) System

Stroke Pridicter | Body Mass Index(B.M.I) Calculator | Contact Us

Hypertension or Not?
Click Here To Visit Lab's Near You

Are You Married ? ○ Yes ○ No

What Is Your Work Type ? ○ Government Job ○ Not Working ○ Private Job ○ Self Employed or Business Owner ○ Student

In which Kind Of Area Is Your Residence ? ○ Urban ○ Rural

Please Enter Your Blood Suger Levels: Blood Suger Levels
Do Not Know if you Blood Suger Levels?
Click Here To Visit Lab's Near You

Please Enter Your Body Mass Index(B.M.I): Enter Your BMI
Do Not Know Your B.M.I?
Click Here To Check Your B.M.I

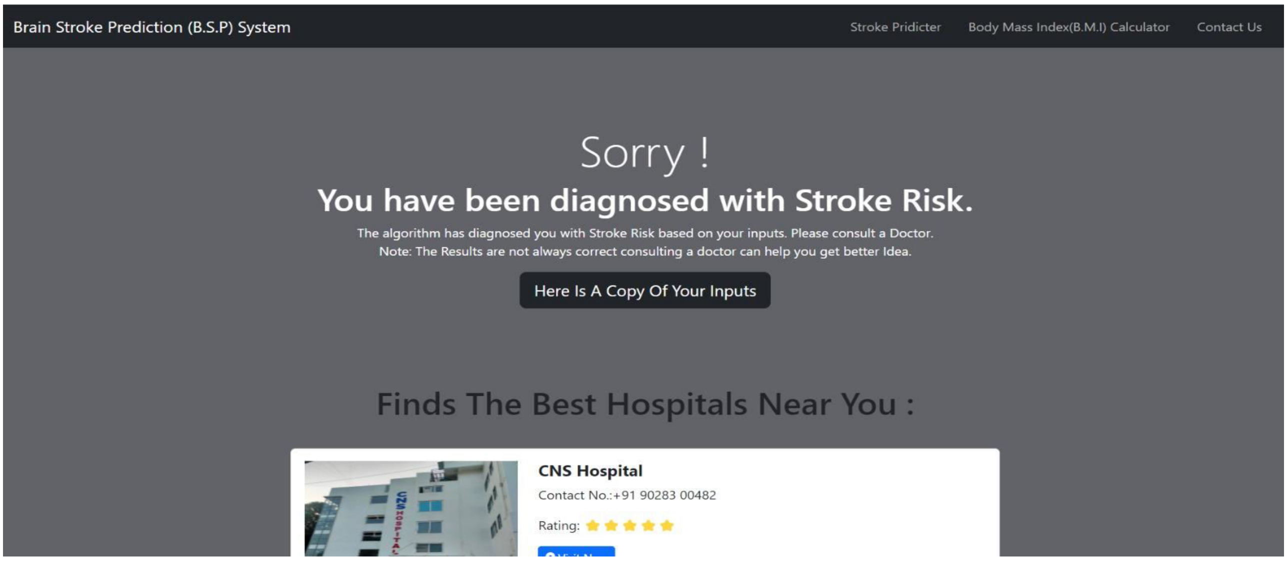
Please Select Your Smoking Status ○ Smoke Sometimes ○ Never Smoke ○ Regular Smoker

Predict

Fig 7: Input Form in HTML, CSS, Bootstrap, JavaScript code

The flask framework acts as a bridge between the web page and the trained machine learning model. The input values are sent to the flask framework which sends the values to the model for prediction.

The machine learning model predicts the outcome when it has the input parameters which it receives from the flask framework. The user may then view the forecast on the webpage as seen in Fig.8 and Fig.9.



Brain Stroke Prediction (B.S.P) System
Stroke Pridicter
Body Mass Index(B.M.I) Calculator
Contact Us
Here Is A Copy Of Your Inputs
Visit N...

Sorry !

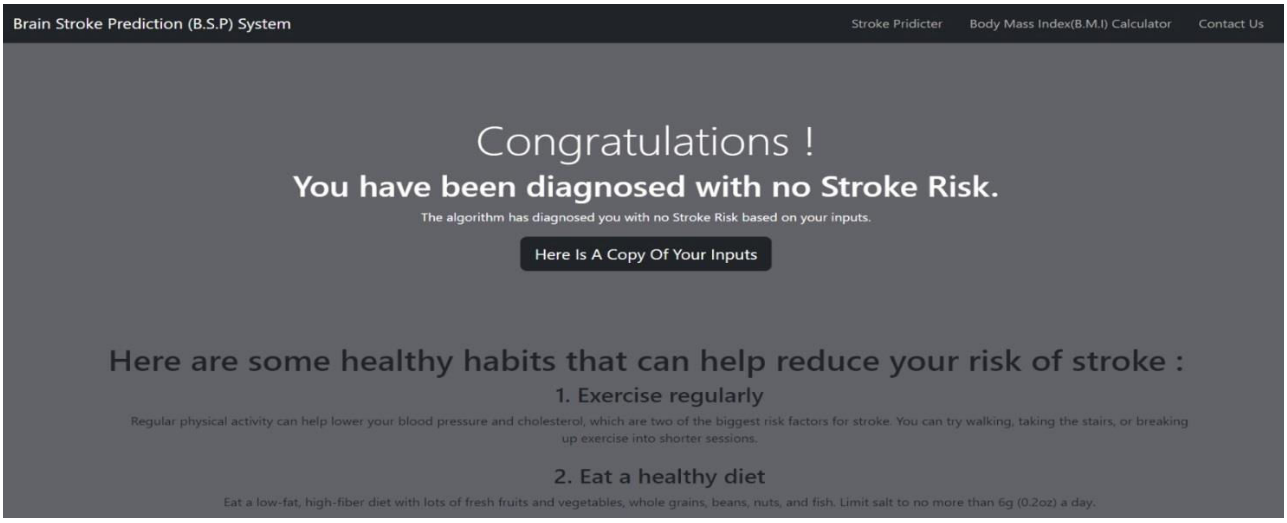
You have been diagnosed with Stroke Risk.

The algorithm has diagnosed you with Stroke Risk based on your inputs. Please consult a Doctor.
Note: The Results are not always correct consulting a doctor can help you get better Idea.

Finds The Best Hospitals Near You :

CNS Hospital
Contact No.:+91 90283 00482
Rating: ★★★★★

Fig. 8: Output With Stroke Risk



Brain Stroke Prediction (B.S.P) System
Stroke Pridicter
Body Mass Index(B.M.I) Calculator
Contact Us
Here Is A Copy Of Your Inputs

Congratulations !

You have been diagnosed with no Stroke Risk.

The algorithm has diagnosed you with no Stroke Risk based on your inputs.

Here are some healthy habits that can help reduce your risk of stroke :

1. Exercise regularly

Regular physical activity can help lower your blood pressure and cholesterol, which are two of the biggest risk factors for stroke. You can try walking, taking the stairs, or breaking up exercise into shorter sessions.

2. Eat a healthy diet

Eat a low-fat, high-fiber diet with lots of fresh fruits and vegetables, whole grains, beans, nuts, and fish. Limit salt to no more than 6g (0.2oz) a day.

Fig. 9: Output With No Stroke Risk

VII. Conclusion and Future Work

Stroke is a significant medical illness that requires prompt treatment to prevent further complications. Developing a machine learning model can aid in predicting strokes early on, reducing the severity of future consequences. This research demonstrates how machine learning systems may accurately predict stroke using many physiological variables. With an accuracy of 97%, Random Forest classification outperforms all the other models. Fig. 10 compares the accuracy scores of different algorithms. Random Forest classification also outperforms other models in terms of precision, recall, and F1 scores. The comparison of the F1 score, precision score, and recall score of different algorithms is displayed in Fig. 11, 12, and 13 respectively.

The project has enormous room to expand and have an influence, and it has the ability to take many important turns. Integration of data from several sources, such as wearable technology and genetics, promises a more thorough evaluation of risk. The model is guaranteed to stay flexible and adaptable to changing health patterns through ongoing learning processes.

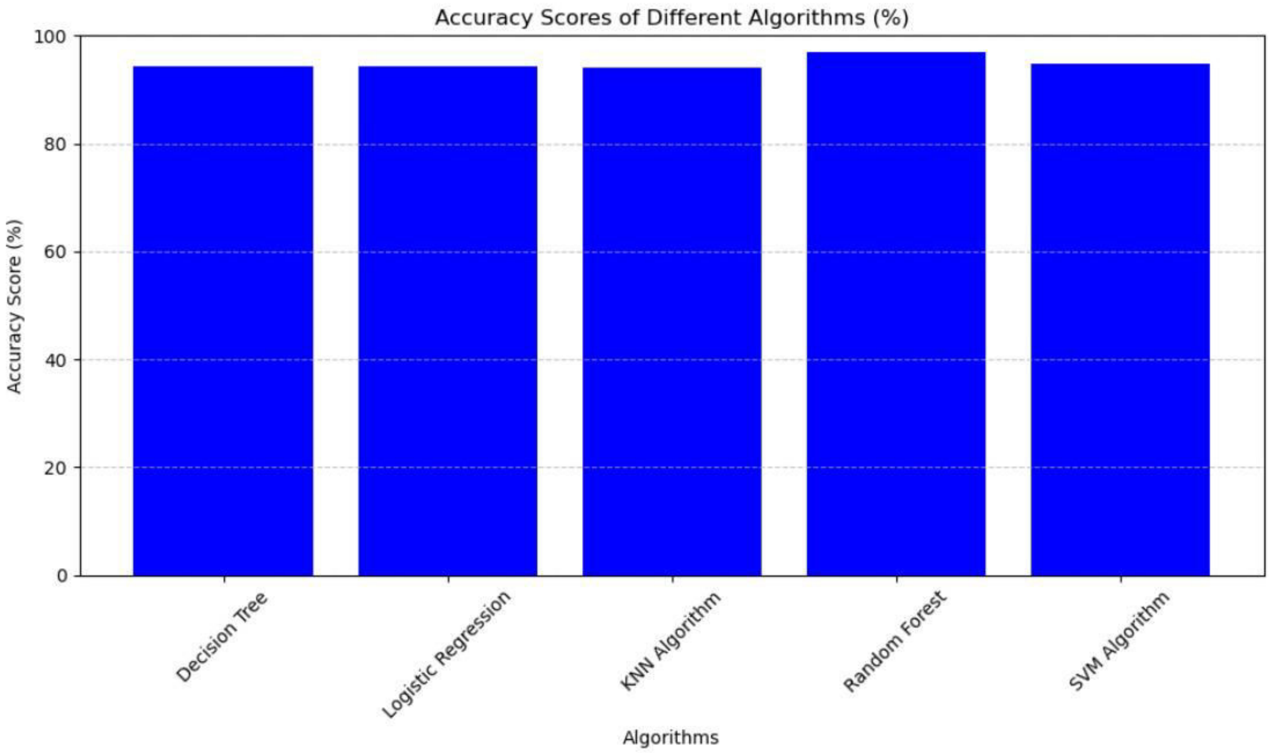


Fig. 10: Comparing the Accuracies

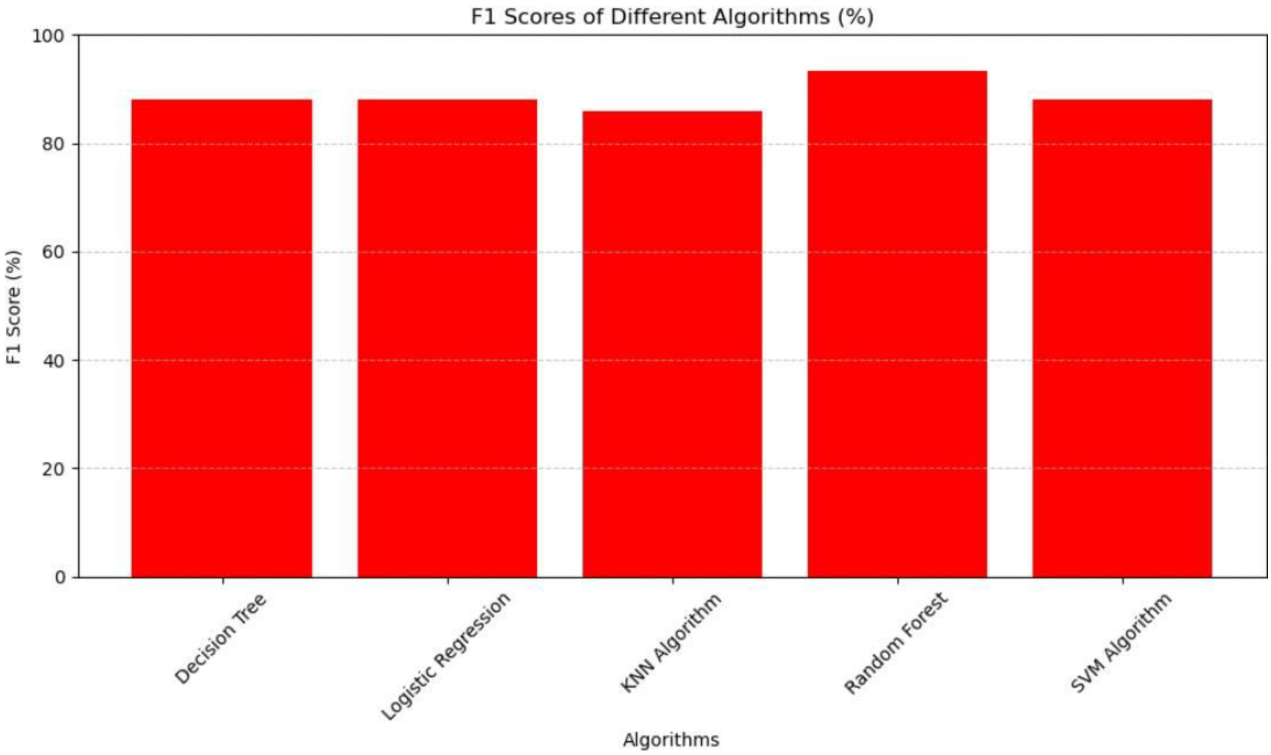


Fig. 11: Comparing the F1 Scores

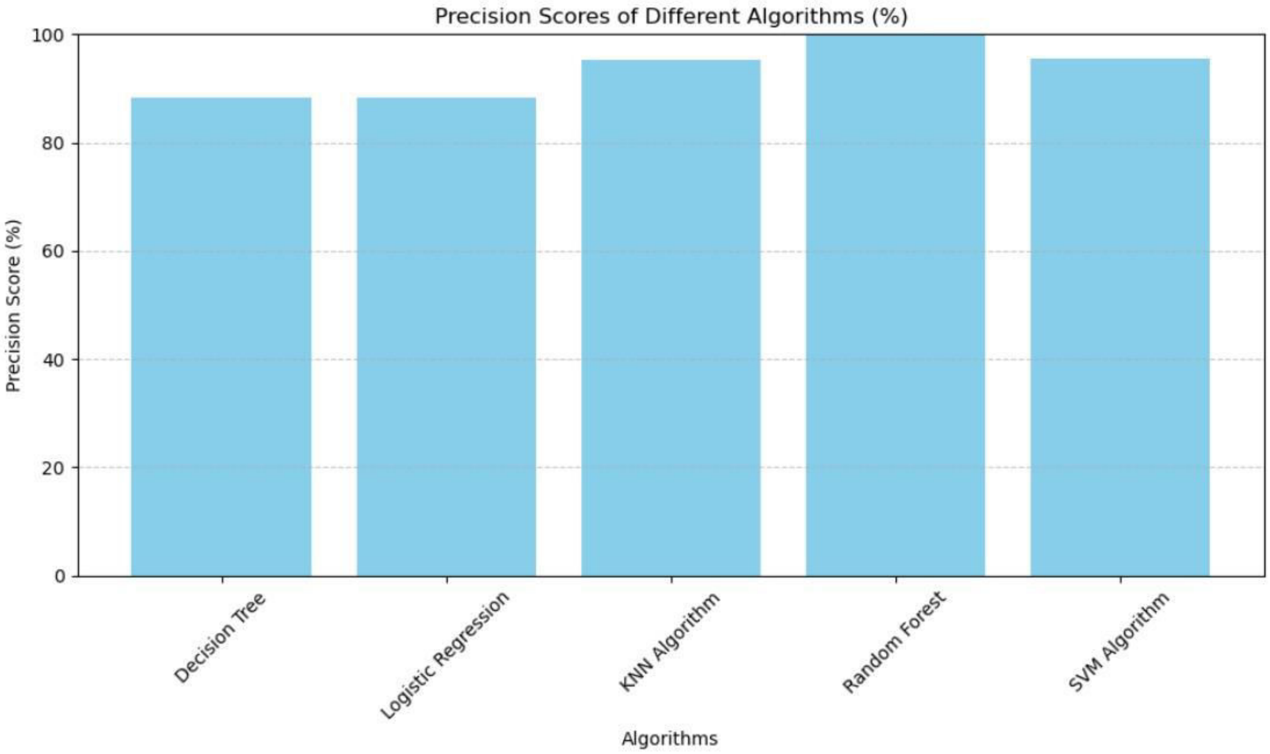


Fig. 12: Comparing the Precision Scores

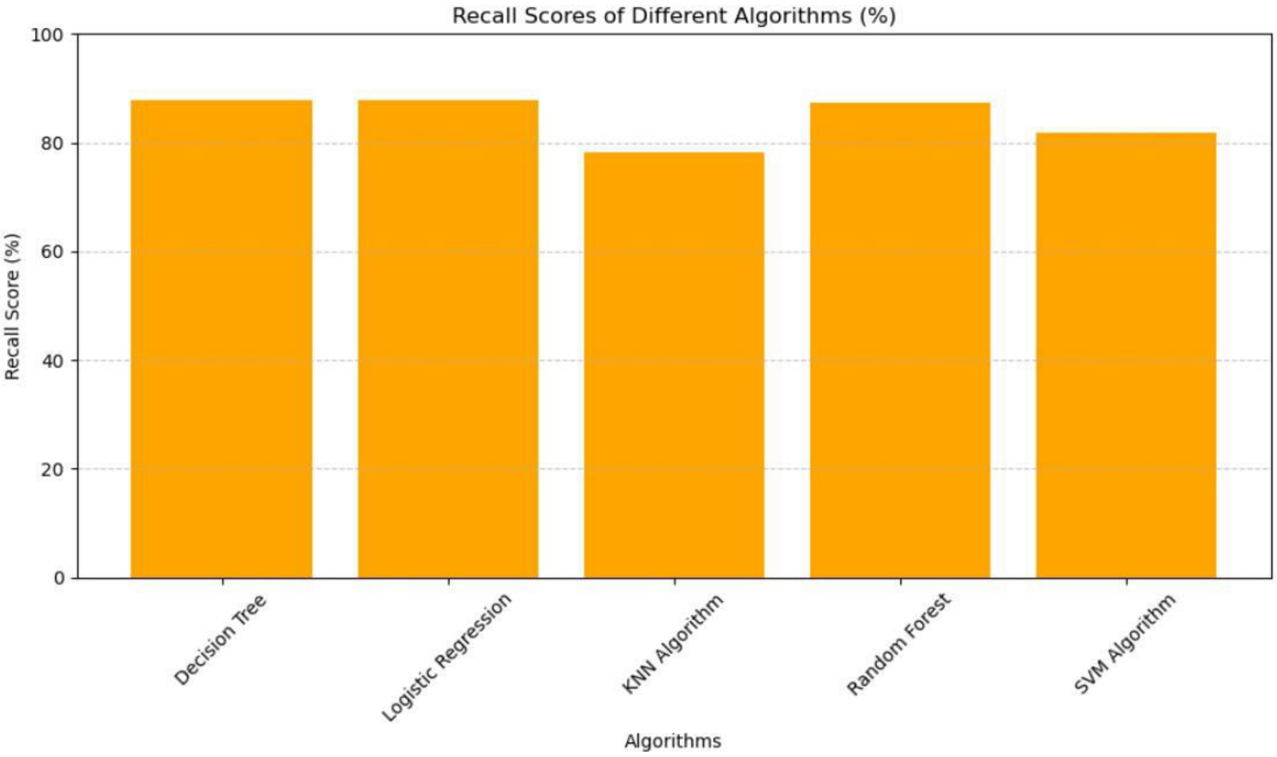


Fig. 13: Comparing the Recall Scores

REFERENCES

1. The National Center for Biotechnology Information (NCBI)
2. Sailasya, Gangavarapu, and Gorli L. Aruna Kumari. "Analyzing the performance of stroke prediction using ML classification algorithms." International Journal of Advanced Computer Science and Applications, 12 Jun 2021.
3. Mridha K, Ghimire S, Shin J, Aran A, Uddin MM, Mridha MF. “Automated Stroke Prediction Using Machine Learning: An Explainable and Exploratory Study with a Web Application for Early Intervention.” IEEE Access, 22 May 2023.
4. Agarwal P., Khandelwal M., Nishtha K. "Brain Stroke Prediction Using Machine Learning Approach." Iconic Research and Engineering Journals,13 Jul 2022.
5. Saxena, Neha "BrainOK: Brain Stroke Prediction using Machine Learning." Journal of Emerging Technologies and Innovative Research (JETIR), 4 April 2022.
6. Padimi V, Telu VS, Ningombam DD. “Performance analysis and comparison of various machine learning algorithms for early stroke prediction.” ETRI Journal, 29 Dec 2022.
7. Bandi V, Bhattacharyya D, Midhunchakkravarthy D. “Pre-diction of Brain Stroke Severity Using Machine Learning.” IIETA Journal 31 Dec 2020.
8. Tazin T, Alam MN, Dola NN, Bari MS, Bourouis S, Khan MM. “Stroke disease detection and prediction using robust learning approaches.” Journal of healthcare engineering. 26 Nov 2021.
9. Sudhahar, T. N., Meera, S., Saraswathi, K., Catherine, S. S., & Priya, B. S. “Stroke Prediction System Using Machine Learning Algorithms.” International Journal of Scientific Development and Research (IJSDR). Jun 2022.
10. Kunder, A. M., Shashank, H. N., Srikanth, S., & Thejas, A. M. “Prediction of Stroke Using Machine Learning”. ResearchGate Publication. Jun 2020.
11. Kv, H., P, H., Gupta, G., P, V., & Kb, P. “Stroke Prediction Using Machine Learning Algorithms.” International Journal of Innovative Research in Engineering & Management (IJIREM). Jul 2021.
12. Dataset named ‘Stroke Prediction Dataset’ from Kaggle:https://www.kaggle.com/datasets/meddata/stroke-prediction-dataset
13. Scikit-learn.org's Logistic Regression documentation.
14. Scikit-learn.org's Decision Tree Classification documentation.
15. Scikit-learn.org's K-nearest neighbors classification docu-mentation.
16. Scikit-learn.org's Random Forest classification documenta-tion.
17. Scikit-learn.org's Support vector machine documentation.
18. Khalane S, Petkar SN, Joshi RD, Nikam AM, Bagul SS. “Stroke Prediction Using Machine Learning.” International Journal for Research in Applied Science and Engineering Technology (IJRASET). 11 Nov 2023.

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