JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Well watch: Nurturing health with explainable stroke insights.

Chippala Naveen, Sarang Khajuria, Harsh Kumar
Student
Lovely Professional University

Under the Guidance of
Jasleen Kaur Paintal

School Of Computer Applications

Abstract:

The majority of strokes happen as a result of an unanticipated blockage in the heart and brain's pathways. Stroke can be minimized by being aware of the various warning symptoms of the disease in advance. Using various machine learning techniques in conjunction with the presence of hypertension, body mass index, heart disease, average glucose level, smoking status, history of stroke, and age, this research work suggests an early prediction of stroke disorders. Six distinct classifiers—Logistics Regression, Decision Tree Classifier, KNeighbors Classifier, Naïve Bayes classifier, Support Vector Machine and Random Forest . The majority of research has been done on heart stroke prediction, but relatively little has been done on brain stroke risk. In light of this, numerous machine learning models are developed to forecast the likelihood of a brain stroke. This study uses machine learning methods such as Naïve Bayes, Support Vector Machine, K-Nearest Neighbors, Decision Tree Classification, Random Forest Classification, and Logistic Regression to analyze a variety of physiological parameters.

Keywords:

Machine learning Approaches, Hypertension, Decision tree Classifier, Kneighbors Classifier, Support Vector Machine.

Introduction:

When the blood flow to different parts of the brain is interrupted or reduced, the cells in those areas of the brain do not receive enough nutrients and oxygen, and they begin to die. This is

known as a stroke. A stroke is a medical emergency that has to be treated right away. To prevent more harm to the damaged area of the brain and potential complications in other body parts, early detection and appropriate management are essential. The World Health Organization (WHO) estimates that 15 million people worldwide suffer from strokes each year, with one victim dying every four to five minutes.

Hemorrhagic and ischaemic strokes are the two types that occur. Clots obstruct the drainage in an ischemical stroke, whereas a weak blood artery bursts and bleeds into the brain in a hemorrhagic stroke. A healthy, balanced lifestyle that eliminates unhealthy habits like drinking and smoking, manages body mass index (BMI) and average blood sugar, and preserves kidney and heart health can avoid stroke. In order to avoid irreversible harm or death, stroke prediction is essential and must be addressed. This study examined cardiac disease, average blood sugar, BMI, and hypertension as risk factors for stroke. Furthermore, the suggested prediction system's decision-making procedures may benefit greatly from machine learning.

Very few documented research studies have employed machine learning models to predict stroke in the literature. Artificial neural networks (ANN), stochastic gradient descent, decision trees, k-nearest neighbor (kNN), principal component analysis (PCA), convolutional neural networks (CNN), naive bayes, and other machine learning techniques are examples. There exists a correlation between some diseases/attributes, including heart disease with stroke, average glucose level, BMI, and hypertension.

We provide the following contribution to this paper:

• Based on factors including age, smoking status, body mass index, heart disease, hypertension, average blood sugar, and prior strokes, a weighted voting classifier is suggested for the prediction of stroke.

• The suggested weighted voting classifier's performance is contrasted with that of the most advanced classifiers, including KNeighbors, Decision Tree Classifier (DTC), Logistics Regression (LR), Random Forest(RF), Support Vector Machine(SVM) and Naïve Bayes(NB).

Research Objectives:

This research aims to achieve the following objectives:

- **Evaluation of Accuracy**: Determine how well various machine learning methods predict the occurrence of strokes. Comparing the effectiveness of different models, including decision trees, random forests, support vector machines, neural networks, etc., could be one way to do this.

- **Feature Importance Analysis:** Analyze the significance of various input variables in the machine learning models to identify the most pertinent features or risk factors for stroke prediction. This may facilitate comprehension of the fundamental causes of stroke incidence.

- **Model Generalization:** Evaluate the machine learning models' capacity for generalization by putting them to the test on different datasets or using cross-validation methods. By achieving this goal, it will be ensured that the models are not overfitting to the training dataset and perform well on unseen data.

Methodology:

The methodology entails gathering and preprocessing clinical data, training and assessing different machine learning models for the prediction of stroke using pertinent metrics and techniques, assessing generalization and robustness, comparing with baseline methods, analyzing feature importance and model interpretability, clinically validating the models, and documenting findings for possible implementation.

Significance:

The significance derives from the possibility of employing cutting-edge machine learning techniques to anticipate strokes with accuracy. This can enhance patient outcomes, support medical personnel in making decisions, and maximize the use of resources in clinical settings.

Review of Literature:

Physicians have historically distinguished between hemorrhagic stroke, which results from internal bleeding in the brain, and ischemic stroke, which is caused by blood clots. More data points, such as imaging scans, genetic information, and medical histories, can be analyzed by machine learning algorithms to more accurately predict the particular type of stroke.

Targeted Therapy: Different strategies are needed to treat hemorrhagic and ischemic strokes. By accurately identifying the type of stroke early on, medical professionals can start the best possible treatment strategy, which may improve patient outcomes.

Clinical studies: Machine learning can assist in patient selection for stroke treatment studies by identifying the precise type of stroke a patient is likely to have. More specialized treatments may result from this.

Risk Stratification and Early Intervention: Large-scale datasets can be analyzed using machine learning techniques to identify people who are at high risk of stroke but have not yet shown symptoms. This makes it possible to proactively use preventative treatments, such as medication or lifestyle modifications, which may completely avoid strokes.

Personalized Treatment Plans: To generate customized treatment plans, machine learning algorithms can examine a patient's genetic information, particular medical history, and kind of stroke. To maximize recovery and reduce long-term consequences, this may entail adjusting prescriptions, rehabilitation techniques, and even dietary guidelines.

Real-time Monitoring and Intervention: Wearable technology and sensors can be combined with machine learning algorithms to remotely monitor stroke victims. Via the analysis of brain activity patterns and vital signs, these devices could identify early indicators of stroke recurrence and prompt prompt medical attention.

Research Gap:

The current body of literature on machine learning techniques for stroke prediction has a research gap that needs to be filled. This gap is caused by the need for additional studies to be conducted in order to address potential biases, improve the interpretability and generalization abilities of models, and validate model performance in actual clinical settings. Further study in these areas is necessary because there are insufficient thorough studies assessing the clinical usefulness and long-term predictive accuracy of machine learning-based stroke prediction systems.

SYSTEM METHODOLOGY :

Various Kaggle datasets were taken into consideration in order to move on with the implementation. Among all the datasets that were already available, one that was suitable for model construction was selected.
The next stage after gathering the dataset is to prepare it so that the computer can understand it more clearly. We refer to this stage as "data preprocessing." This stage involves applying label encoding that is unique to this dataset, addressing imbalanced data, and handling missing values.
The preprocessed data is now prepared for model creation. Preprocessed datasets and machine learning techniques are needed for model construction. The Random Forest Classification method, K-Nearest Neighbor algorithm, Support Vector Classification, Naïve Bayes Classification algorithm, Decision Tree Classification algorithm, and Logistic Regression are utilized. The accuracy measures Accuracy Score, Precision Score, Recall Score, F1 Score, and Receiver Operating Characteristic (ROC) curve are used to compare the six distinct models that have been built. The model that performs the best in terms of accuracy metrics is determined by comparing the models before moving on to the deployment step. An HTML page is created for the model's deployment that makes it simple for the user to enter the input parameters and receive the output. Using a flask application—basically, a Python framework—the user's parameters are delivered to the model, connecting the web application and the model. The flask application receives the outcome of the model's prediction of the output based on the input parameters. This flask will now show the outcome on the webpage so that the visitor can review it. Fig. 1 shows the flow chart for the technique of the suggested system.

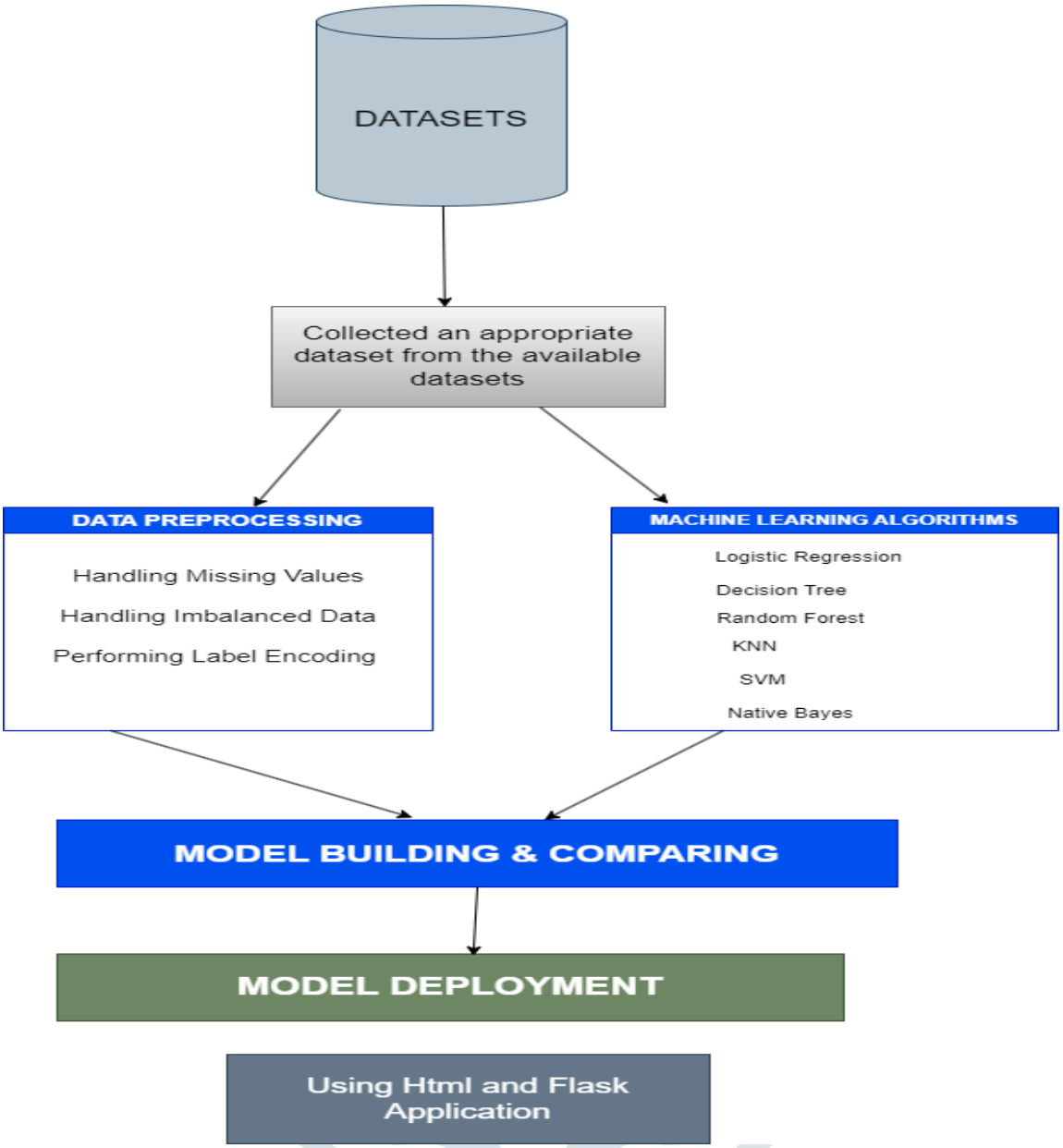

Fig 1:- Procedure for stroke prediction.

Research Methodology:

This section is divided into three parts, these are: Data description, machine learning classifiers & evaluation matrices, implementation procedures. These three processes are described below:

A. Data Description

The 5109 entries in Kaggle's "Healthcare-dataset-stroke data," a publicly accessible repository, served as the dataset's original source. This dataset is a useful tool for performing in-depth analyses and modelling since it comprises 2115 male and 2994 female subjects across 12 different variables.

Description of Data:

- ID: A distinct identification assigned to every dataset entry.
- Gender: A categorical variable (e.g., Male, Female, Other) that indicates a person's gender.

- Age: A numerical variable that expresses a person's age in years.
- Hypertension: A binary variable that indicates if a person has high blood pressure (1) or not (0).
- Heart Disease: A binary variable denoting the presence or absence of heart disease in the individual (1) or (0).
- Ever Married: A categorical variable with the options Yes and No that represent marital status.
- Work Type: A categorical variable (such as Private, Self-employed, Government job, children, and Never worked) that indicates the kind of employment.
- Residence: A categorical variable that denotes the kind of home (rural, urban, etc.).
- Average Glucose Level: A numerical variable that expresses the blood's average glucose content.
- Body Mass Index (BMI): A numerical statistic that indicates the BMI of the participants.
- Smoking Status : This feature captures the participant's smoking status.
- Stroke**:** This feature represents if the participant previously had a stroke or not.

B. Preprocessing Data:

Prior to developing a model, data preprocessing is necessary to eliminate undesirable noise and outliers from the dataset.

leading to a departure from appropriate instruction. This step addresses anything that prevents the model from operating as efficiently as possible. The following stage after gathering the relevant dataset is to clean the data and ensure that it is prepared for model creation. Table I lists the 12 attributes of the dataset that was collected. 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. The column 'bmi' in this instance has null values that are filled in with the column data mean.

C. Label Encoding

Encoding the string literals in the dataset is known as label encoding. into integer values so that the computer can comprehend them. The strings must be transformed into integers since the computer is often trained on numbers. The gathered dataset contains five columns with strings as the data type. All of the strings are encoded during label encoding, converting the dataset as a whole into a collection of numerals.

D. Managing Data Inequalities

The dataset selected for the stroke prediction task is really unbalanced. 5109 rows make up the whole dataset, of which 249 rows indicate a possible stroke and 4861 rows show no possibility of a stroke. Fig. 2 is a graphic representation of the imbalance. With such data, a machine-level model might be trained with accuracy; but, other accuracy metrics, such as precision and recall, are not very useful. The prediction is inefficient and the results inaccurate if such unbalanced data is not handled. As a result, handling this unbalanced data is necessary before creating an effective model. The undersampling method is applied for this reason. In order to match the majority class, undersampling balances the data.

```
import seaborn as sns
sns.countplot(x='stroke', data=dataset)
```

```
9]: <matplotlib.axes._subplots.AxesSubplot at 0x16d6edb5648>
```

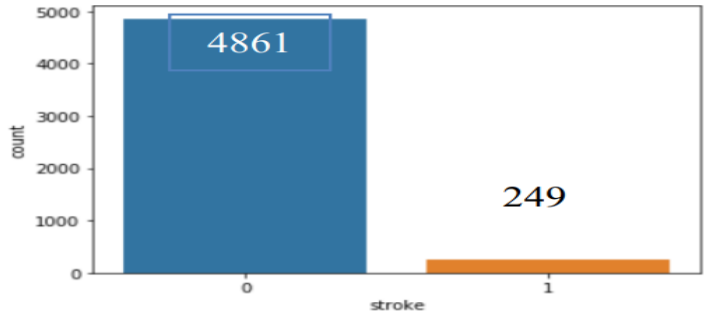

Fig 2: Before Under Sampling.

```
import seaborn as sns
sns.countplot(x='stroke', data=under_sample)
```

```
]: <matplotlib.axes._subplots.AxesSubplot at 0x16d6ee28508>
```

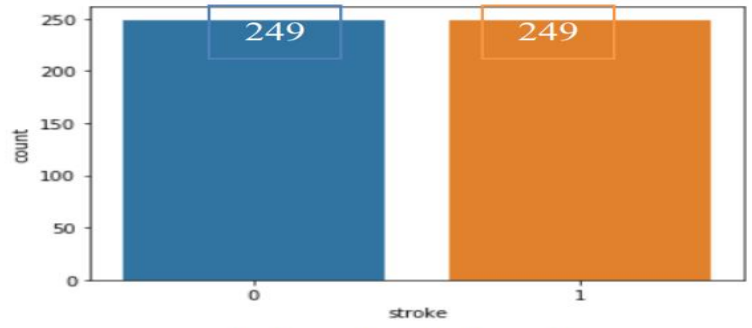

Fig 3:- After Under Sampling.

Model Building:

Six different algorithms for the comparative analysis were included following a thorough review of the literature. The selection of these algorithms—Decision Tree, Logistic Regression, Random Forest, Support Vector Machine, Naïve Bayes and K Nearest Neighbor—was based on their recognition in the machine learning community and their proven performance in a range of scenarios. Algorithms can be effectively employed to tackle diverse classification and regression jobs due to their distinct features and benefits. By use of an extensive comparison analysis, the objective is to assess the efficacy, merits, and drawbacks of every algorithm, ultimately pinpointing the optimal strategy for the particular issue area under consideration.

1.Decision Tree : The Decision Tree algorithm is a flexible and user-friendly technique for machine learning applications involving regression and classification. Recursively dividing the dataset into subsets according to the values of attributes is how it works. In the end, it forms a structure like a tree, with each internal node denoting a choice made in response to a specific feature. As a result of these choices, leaf nodes that match the expected result form. Decision trees are preferred because they are transparent and easy to read, enabling people to have an intuitive understanding of the decision-making process. They might, however, experience overfitting, particularly when working with intricate datasets.

2) Logistic Regression : For binary classification tasks, the basic approach utilised is called logistic regression. It is not a regression model, despite its name; it is a linear model for classification. Using a logistic (or sigmoid) function, logistic regression models the likelihood that a given input belongs to a specific class. Because of its ease of use, effectiveness, and interpretability, it is extensively utilised. Furthermore, it offers information on how specific features affect the outcome variable, which makes it a useful tool for exploratory data research. Nevertheless, the usefulness of logistic regression may be limited in scenarios with complicated interactions due to its assumption of linearity between characteristics and the response variable's log-odds.

3) Random Forest : Using the combined strength of several decision trees, Random Forest is an ensemble learning technique that enhances prediction robustness and accuracy. During training, a large number of decision trees are built, each based on a random subset of the features and a random portion of the training data. The final forecast is produced by combining the forecasts from each individual tree, usually by average or voting. High-dimensional datasets and nonlinear interactions are areas in which Random Forest shines. Through ensemble averaging, it also effectively reduces overfitting. However, in comparison to individual decision trees, its interpretability might be impaired.

4) SVM(support vector machine):

SVMs are yet another effective instrument in the toolbox of machine learning. They do well on tasks involving regression and classification. Consider that you have a dataset that has colored dots for each class. SVMs seek to isolate these dots with the greatest margin by creating a dividing line (or a hyperplane in higher dimensions). The data points known as support vectors that are closest to the line define this margin. By employing kernel functions, which basically project the data into a higher-dimensional space where linear separation becomes easier, SVMs are able to handle complex data. Even though SVMs have a reputation for being quite accurate—especially when applied to well-separated data—they can be computationally costly when applied to huge datasets. It might also be difficult to understand how they make decisions, thus they need meticulous parameter adjustment for best results.

5) Naive Bayes: Naive Bayes could be a wise option. It employs a Bayesian probabilistic methodology. This is the main concept: Suppose you wish to determine whether a fresh fruit in your basket is an orange or an apple. Naive Bayes considers each feature of the fruit (like color, size, etc.) independently and calculates the probability of it being an apple or an orange based on those individual features. Naive Bayes predicts by averaging these probabilities. Naive Bayes is rapid to train and analyse, making it a valuable tool for exploratory data analysis. But in practical situations, the premise that features are independent can occasionally provide unreliable outcomes. Additionally, it may have trouble with highly connected features or Continuos data.

6) K Nearest Neighbour (KNN): Based on the proximity principle, K Nearest Neighbour is a straightforward yet powerful method for classification and regression problems. By examining the average value or majority class of its k nearest neighbours in the feature space, a KNN may predict the class or value of a new data point. Since it is non-parametric, it does

not rely on any presumptions about how the data will be distributed. KNN can adapt to complex decision boundaries and is simple to implement. Its effectiveness, however, can depend on the distance measure and k value that are chosen, and it might be computationally costly for big datasets.

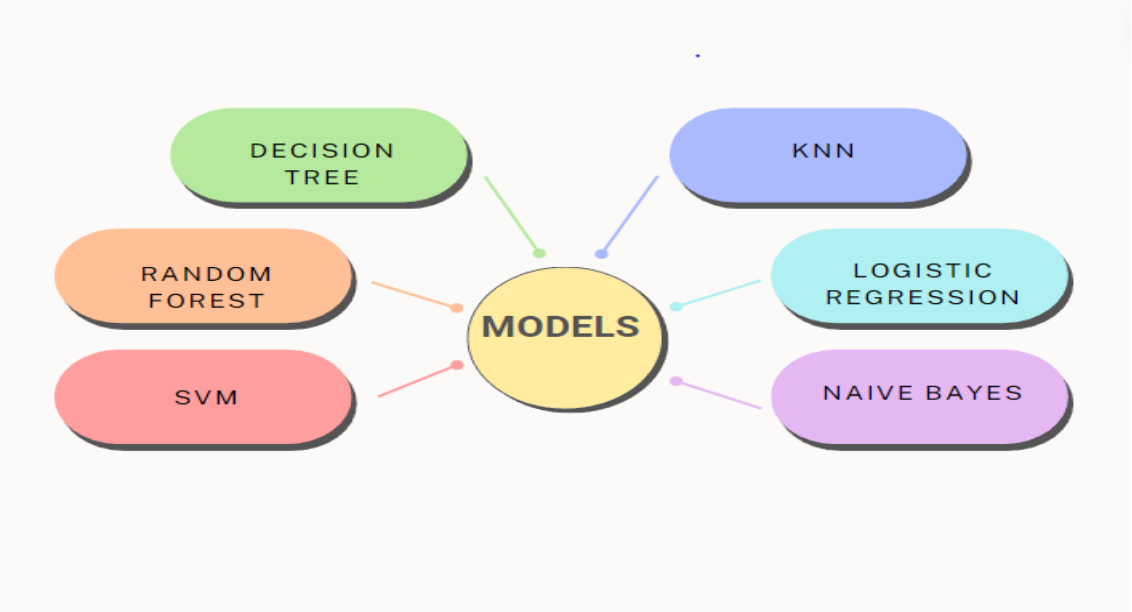

Fig 4:- model building

Following the construction of the model, it can be said that Naïve Bayes has outperformed other algorithms. Thus, pickle is used to dump the model that was trained with Naïve Bayes classification. Creating a flask application and a web application to enter the input parameters is the next step. Simple HTML code is used in the website's construction. This program features an input form that asks the user to enter input values in order to forecast the likelihood of a stroke. The flask application receives the input parameters when the user clicks the "Check Here" button.

Implementation :

1.Data Source :

- Primary Data Initial data compilation involved gathering reference materials and identifying essential features through collaborative consultations with neurologists and cardiologists. These experts provided valuable insights into the key attributes relevant to stroke analysis, ensuring a comprehensive dataset.
- Secondary Data Supplementary data was sourced from the Kaggle platform, specifically the "Healthcare-dataset-stroke data" repository. This publicly available dataset comprises 5109 entries, with a gender distribution of 2994 females and 2115 males. It encompasses 12 features, offering a rich source of information for analysis and model development.

2.Imported Libraries:

- NumPy , a foundational Python library, facilitated array manipulation for scientific computations. Its functionality extended to linear algebra, matrix operations, and Fourier transformations, enabling robust data processing and analysis.

- Pandas , a versatile library, was instrumental in data analysis tasks. It provided extensive support for handling various file formats such as SQL, JSON, and Excel, along with essential operations like data merging, selection, reshaping, and cleaning, thus ensuring data integrity and usability.
- Matplotlib Pyplot module served as a powerful tool for data visualization, offering a plethora of functions akin to MATLAB. It facilitated the creation of visually appealing plots and graphs, essential for elucidating insights from the dataset and conveying analytical findings effectively.
- Seaborn Building upon Matplotlib's capabilities, Seaborn enhanced data visualization with its high-level interface. It streamlined the process of generating informative and visually striking graphs, thereby augmenting the interpretability and aesthetic appeal of the visualizations.

3. Data Cleaning:

- Data cleaning procedures focused on addressing missing values and null entries, ensuring the dataset's quality and reliability. Missing data was handled by removing rows with null values or redundant entries, thereby enhancing the dataset's suitability for subsequent analysis and modeling tasks.

4.Data Analysis:

- An extensive data analysis regimen encompassed categorical and numerical feature analyses, along with multicollinearity assessments. These analyses aimed to uncover hidden relationships and intrinsic attributes within the dataset, thereby enhancing the performance and interpretability of machine learning models.

5. Algorithm Implementation:

- Following a thorough literature survey, six distinct algorithms—Decision Tree, Logistic Regression, Random Forest, Support Vector Machine, Naïve Bayes and K Nearest Neighbor—were selected for comparative analysis. This comparative study aimed to evaluate their efficacy in stroke prediction, thereby informing the selection of the most suitable algorithm for the task.

6. Cross-Validation:

- To mitigate overfitting risks and ensure model robustness, cross-validation techniques were employed to assess the effectiveness of all models. This involved evaluating model performance on independent test datasets, providing insights into their generalizability and reliability.

7. Graphical User Interface (GUI) Development :

- The development of a graphical user interface (GUI) entailed designing an intuitive interface for user interaction.

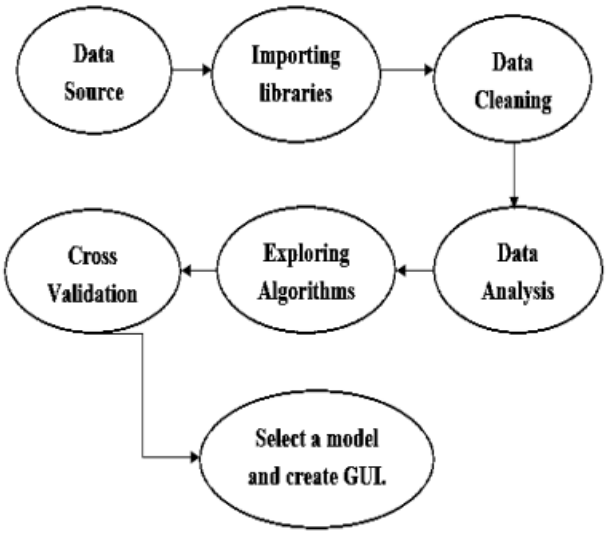

Fig 5. graphical representation of procedure for predicting stroke of different algorithms in step by step.

Results & Discussion:

Stroke is a serious medical illness that needs to be treated right away to prevent complications. Developing a machine learning model can lessen the severe effects of stroke in the future and aid in its early prediction. This study demonstrates how well different machine learning algorithms perform in terms of accurately predicting stroke based on a variety of physiological characteristics. With an accuracy of 82%, Naïve Bayes Classification outperforms all the other algorithms. Fig. 12 compares the accuracy values derived from different approaches.
When it comes to precision, recall, and F1 scores, Naïve Bayes has done better than the others. The comparison of the F1 score and accuracy is shown below.

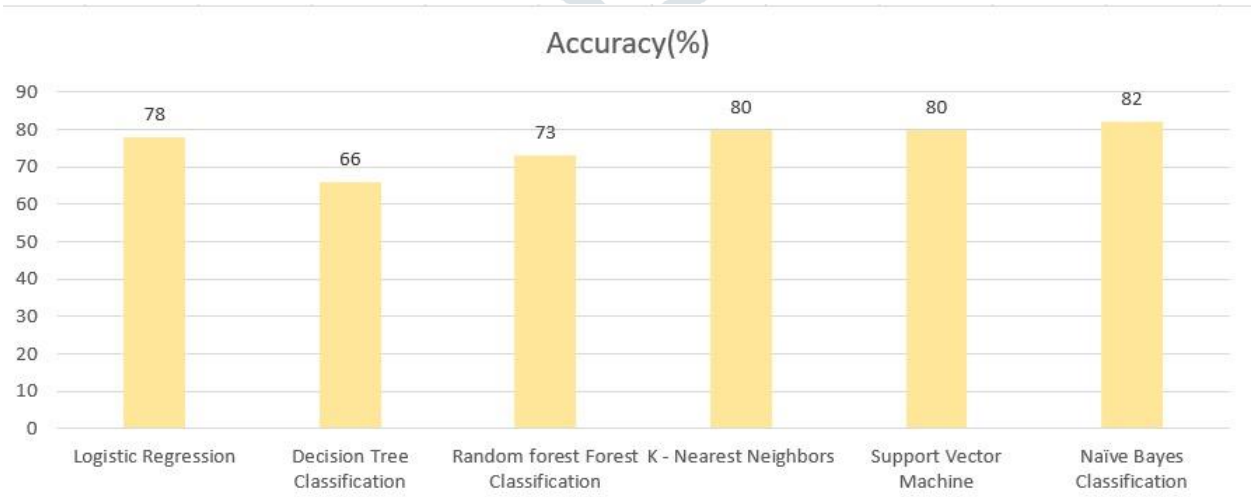

Fig 6 . Comparing the Accuracy of ML Algorithms.

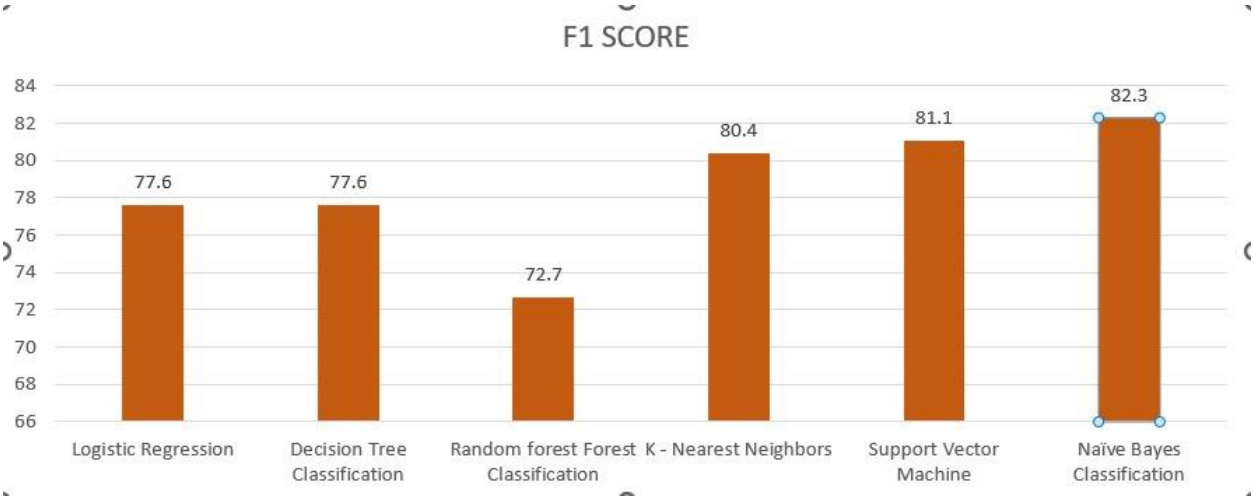

Fig 7. Comparing the F1 Scores of ML Algorithms.

This research proposes to apply multiple machine learning techniques to the selected dataset. This undertaking can be increased even further by employing neural networks to train the model. More accuracy criteria can be taken into account while comparing the performance. The only data used in this experiment is text, which may not always be reliable for predicting strokes. It would be more effective in the future to gather a dataset made up of images, such as brain CT scans, in order to forecast the likelihood of stroke.

Conclusion:

The study "Well watch: Nurturing health with explainable stroke insights" concludes by providing a thorough examination of machine learning-based stroke prediction. Given the damaging effects that strokes can have on a person's health, the study explores the importance of early stroke detection and therapy. The work highlights a research need in machine learning-based stroke prediction by a thorough evaluation of the literature, opening the door for the suggested methodology.

The precise definition of the research objectives centers on the assessment of model generalization, feature importance analysis, and accuracy. The study's methodology includes preprocessing the data, developing a model with six different machine learning methods, cross-validation, and creating a graphical user interface for user interaction.

With an accuracy of 82%, the results and explanations demonstrate how successful Naïve Bayes Classification is compared to other methods. The superiority of Naïve Bayes in stroke prediction is further supported by precision, recall, and F1 scores. A clear comparison of algorithm performance is made easier with the use of graphic representations.

In order to improve prediction accuracy, the study's conclusion makes recommendations for future research directions, such as using neural networks and a variety of datasets, including brain CT scans. In general, the study advances methods for predicting stroke, which may enhance patient outcomes and the process of making healthcare decisions.

REFERENCES

1. Learn about Stroke. Available online: https://www.world-stroke.org/world-stroke-day-campaign/why-stroke-matters/learnabout-stroke (accessed on 25 Jan 2024).
2. Kaggle's Heart Stroke Dataset: Avalaible online: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset (accessed on 26 Jan 2024).
3. Statistics of Stroke by Centers for Disease Control and Prevention.
4. Knowledge of Algorithms: https://www.javatpoint.com/machine-learning-algorithms
5. knowledge of designs: https://www.w3schools.com/python/matplotlib_intro.asp.