Brain Stroke Prediction Using Machine Learning Approach

DR. AMOL K. KADAM[1], PRIYANKA AGARWAL[2], NISHTHA[3], MUDIT KHANDELWAL[4]

[1] *Professor, Department of Computer Engineering, Bharati Vidyapeeth (Deemed to beUniversity) College of Engineering, Pune, Maharashtra, India*

[2, 3, 4] *Student, Department of Computer Engineering, Bharati Vidyapeeth (Deemed to be University) College of Engineering, Pune, Maharashtra, India*

Abstract- A Stroke is an ailment that causes harm by tearing the veins in the mind. Stroke may likewise happen when there is a stop in the blood stream and different supplements to the mind. As per the WHO, the World Health Organization, stroke is one of the world's driving reasons for death and incapacity. The majority of the work has been completed on heart stroke forecast however not many works show the gamble of a cerebrum stroke. Subsequently, the AI models are worked to foresee the chance of cerebrum stroke. The project is pointed towards distinguishing the familiarity with being in danger of stroke and its determinant factors amongst victims. The research has taken numerous factors and utilized ML calculations like Logistic Regression, Decision Tree Classification, Random Forest Classification, KNN, and SVM for accurate prediction.

Indexed Terms- Machine learning; logistics regression; decision tree classification; random forest classification; k-nearest neighbor; support vector machine.

I. INTRODUCTION

As per the Centers for Disease Control and Prevention (CDC), stroke is the fifth Leading demise reason [1] in the US. Stroke is an infection that is responsible for around eleven percent of complete passings. Reliably, north of 795,000 people in the USA experience the impacts of a stroke [2]. It is the fourth main cause for demises in India. With the cutting-edge innovation in clinical science, foreseeing the event of a stroke can be made utilizing ML algorithms. The Machine learning calculations are valuable in making exact forecasts and can give right examination. The works recently performed on stroke generally remember the ones for Heart stroke expectation. Not much of work has been performed on Brain stroke. The study Centers around foreseeing cerebrum stroke event utilizing Machine Learning. The key methodologies were utilized, and results are gotten with five distinct grouping calculations. The disadvantage to this model is that it is being prepared on text-based information and not on constant cerebrum pictures. The paper shows the execution of 5 Machine Learning methodologies. This paper can be additionally reached out to execute all the ongoing AI calculations. A dataset is browsed Kaggle [3] with different qualities as its credits to continue further. A huge subject of AI in drug is used in this undertaking. An AI model would take the patient's data and propose a lot of reasonable Expectations. The system can eliminate hid data from a chronicled clinical informational index and can expect patients with contamination and use the clinical profiles like Age, circulatory strain, Glucose, etc it can predict the likelihood of patients getting a sickness. Gathering computations are used with the quantity of properties for the assumption for sickness [4].

II. RELATED WORK

A great deal of work has been finished in the part of stroke expectation. Jeena et al. give an investigation of different gamble elements to comprehend the likelihood of stroke [5].

Govindarajan [6] managed the data assembled from Sugam Multispeciality Hospital. The dataset contained more than 500 records of patients and many fascinating class names of two huge Stroke types. They applied Support Vector Machine (SVM), Artificial Neural Network (ANN), Logistic Regression, Decision Tree, Bagging, and Boosting.

Among the above Machine Learning Algorithm, they got the highest accuracy using ANN Algorithm with ~95%.

Sung et al [7] dealt with clinical information which contained data about the ischemic stroke of 739 patients. This information contains 17 clinical factors which incorporated the historical backdrop of past TIA, a gamble factor for vascular illnesses, patient's segment data, stroke subtypes, neuroimaging boundaries, and so on and this will be utilized for working out the precision of an AI calculation in foreseeing END. They checked with 4 Machine Learning Algorithms: - Deep brain organization, Boosted Trees, Logistic Regression, and Bootstrap choice forest.0.966, 0.966, 0.966, and 0.946 are the exactness score got from the model. Among every one of the calculations the most elevated region under the bend worth of 0.934 and precision of 0.966 is accomplished by Boosted Tree calculation.

Choudhary and Singh [8] chipped away at information gathered from the workforce of Physical treatment. It contains data about cardiovascular wellbeing studies. The dataset comprises of more than 5,800 examples. They isolated the dataset into three distinct clinical phrasings: stroke and claudication, stroke and TIA, stroke and Angioplasty. It additionally incorporates in excess of 600 characteristics. They involved head part examination for dimensionality reduction.C4.5 calculation is utilized for include choice. By ANN execution they achieved 95%, 95.2%, and 97.7% Accuracy.

Selma gathered a dataset from a few emergency clinics and clinical Centers. The clinic report incorporates the patient serial number, CT, age of patient, gender, MRI analyse, and different factors for all patients hospitalized in the medical clinic. The dataset contained around 410 patients, whose age is mostly somewhere in the range of 48 and 87 years. A couple of cases in the age of 32 years and the vast majority of them are male. The presentation of Decision tree characterization is superior to the exhibition of the KNN calculation. Clinical experts utilized a decision tree calculation to order and analyze ischemic stroke patients.

III. ANALYSIS DATASET

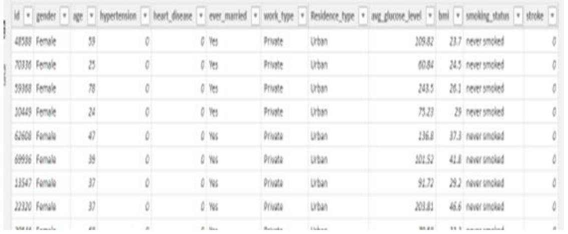
We have a given dataset for stroke prediction. This particular dataset has 12 columns and 5110 rows. The columns and rows have information about different individuals in different datatypes. They are as follows:

i. Patient ID
ii. Gender of the individual
iii. Age
iv. Information about prior occurance of Hypertention
v. Previous heart Diseases
vi. Marital Status
vii. Work Status
viii. Residential Type
ix. Glucose level of different individuals
x. BMI value
xi. Smoking status

Based on the attributes mentioned above we find out the probability of a future stroke risk using the system we have developed. The output that we have is in binary form. '0' indicates no stroke risk detected, and '1' indicates a possible risk of stroke.

TOTAL NUMBER OF PEOPLE WITH STROKES
STROKE
249

This dataset has a total number of 249 individuals with a possible future stroke risk. These individuals are then alerted using the system to consult a medical professional for further follow-up.



id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
48588	Female	59	0	0	Yes	Private	Urban	109.82	23.7	never smoked	0
70336	Female	25	0	0	Yes	Private	Urban	60.84	24.5	never smoked	0
59368	Female	78	0	0	Yes	Private	Urban	243.5	26.1	never smoked	0
10449	Female	24	0	0	Yes	Private	Urban	75.23	29	never smoked	0
62608	Female	47	0	0	Yes	Private	Urban	136.8	37.3	never smoked	0
69936	Female	39	0	0	Yes	Private	Urban	101.52	41.8	never smoked	0
13547	Female	37	0	0	Yes	Private	Urban	91.72	29.2	never smoked	0
22320	Female	37	0	0	Yes	Private	Urban	203.81	46.6	never smoked	0

The dataset discussed above is summarized in Table:

TABLE I. STROKE DATASET

Attribute Name	Type (Values)	Description
1. id	Integer	A unique integer value for patients
2. gender	String literal *(Male, Female, Other)*	Tells the gender of the patient
3. age	Integer	Age of the Patient
4. hypertension	Integer *(1, 0)*	Tells whether the patient has hypertension or not
5. heart_disease	Integer *(1, 0)*	Tells whether the patient has heart disease or not
6. ever_married	String literal *(Yes, No)*	It tells whether the patient is married or not
7. work_type	String literal *(children, Govt_job, Never_worked, Private, Self-employed)*	It gives different categories for work
8. Residence_type	String literal *(Urban, Rural)*	The patient's residence type is stored
9. avg_glucose_level	Floating point number	Gives the value of average glucose level in blood
10. bmi	Floating point number	Gives the value of the patient's Body Mass Index
11. smoking_status	String literal *(formerly smoked, never smoked, smokes, unknown)*	It gives the smoking status of the patient
12. stroke	Integer *(1, 0)*	Output column that gives the stroke status

Numerous inferences could be drawn out from the said dataset. To represent the entire dataset, the following BI Dashboard has been created:

IV. METHODOLOGY

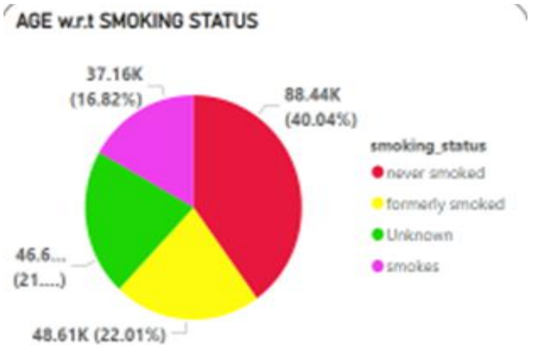
We first Import the following libraries:

1) Pandas
2) Numpy
3) Matplotlib
4) seaborn

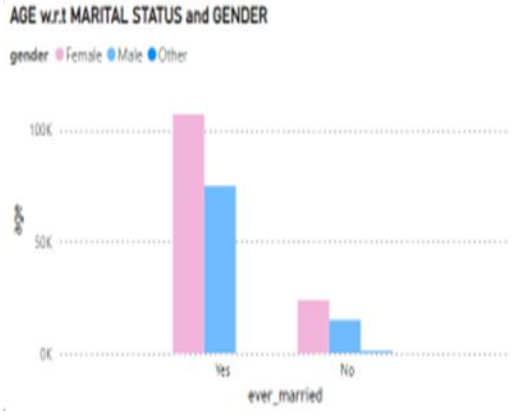
- Cleaning: Primarily as a prerequisite for the dataset, we first checked for null values and replaced them by the average values.

- Data Analysis: We then checked for outliers in the dataset and analysed it. Data in the form of 'yes/no', 'male/female' are converted to numeric form.

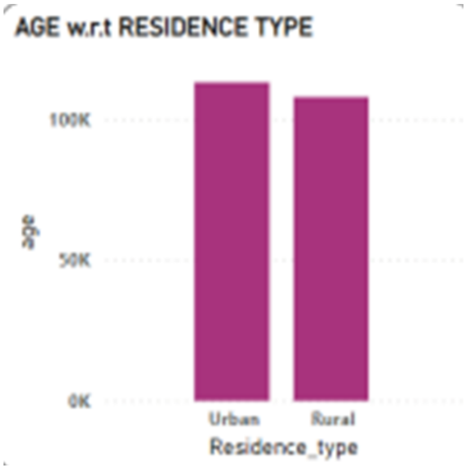
a) Here we observe a visual representation for the smoking habits of different age groups classifying our dataset into five separate classes.



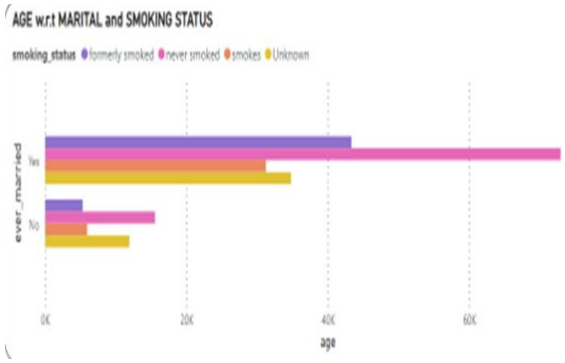
b) We also analyse the gender bifurcation of our dataset.



c) Our analysis also depends upon the residence type of the patient leading to Urban and Rural classes.



d) Marital status also affects our final outcome so we analyse whether our dataset has higher number of married or unmarried people.

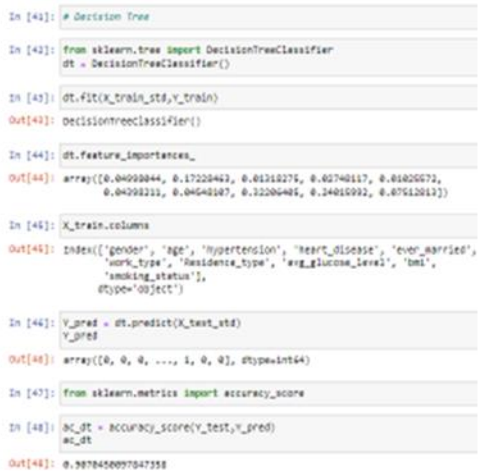


V. IMPLEMENTATION

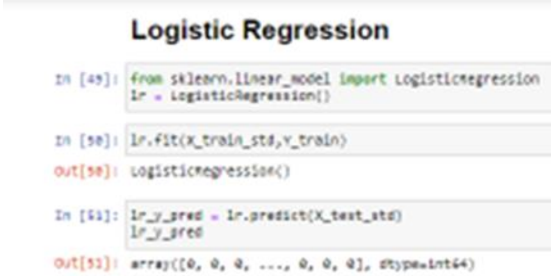
The following algorithms have been used for the brain stroke detection system that we have created:

1) Decision Tree
2) Logistic Regression
3) Random Forest
4) Support Vector Machine
5) K Nearest Neighbour

All these are used to predict the possibility of stroke in a person.

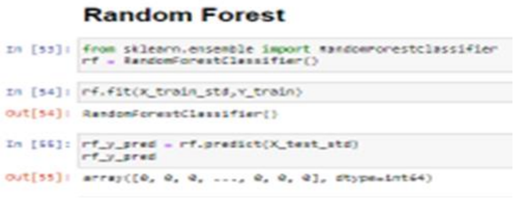


```
In [41]: # Decision Tree

In [42]: from sklearn.tree import DecisionTreeClassifier
         dt = DecisionTreeClassifier()

In [43]: dt.fit(X_train_std,Y_train)
Out[43]: DecisionTreeClassifier()

In [44]: dt.feature_importances_
Out[44]: array([0.04099044, 0.17228463, 0.01318275, 0.02740117, 0.01026672,
                0.04398211, 0.04648107, 0.32206485, 0.24016932, 0.07512013])

In [45]: X_train.columns
Out[45]: Index(['gender', 'age', 'hypertension', 'heart_disease', 'ever_married',
                'work_type', 'Residence_type', 'avg_glucose_level', 'bmi',
                'smoking_status'],
               dtype='object')

In [46]: Y_pred = dt.predict(X_test_std)
         Y_pred
Out[46]: array([0, 0, 0, ..., 1, 0, 0], dtype=int64)

In [47]: from sklearn.metrics import accuracy_score

In [48]: ac_dt = accuracy_score(Y_test,Y_pred)
         ac_dt
Out[48]: 0.9078450097847358
```



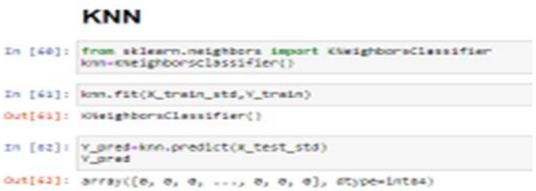
Logistic Regression

```
In [49]: from sklearn.linear_model import LogisticRegression
         lr = LogisticRegression()

In [50]: lr.fit(X_train_std,Y_train)
Out[50]: LogisticRegression()

In [51]: lr_y_pred = lr.predict(X_test_std)
         lr_y_pred
Out[51]: array([0, 0, 0, ..., 0, 0, 0], dtype=int64)
```



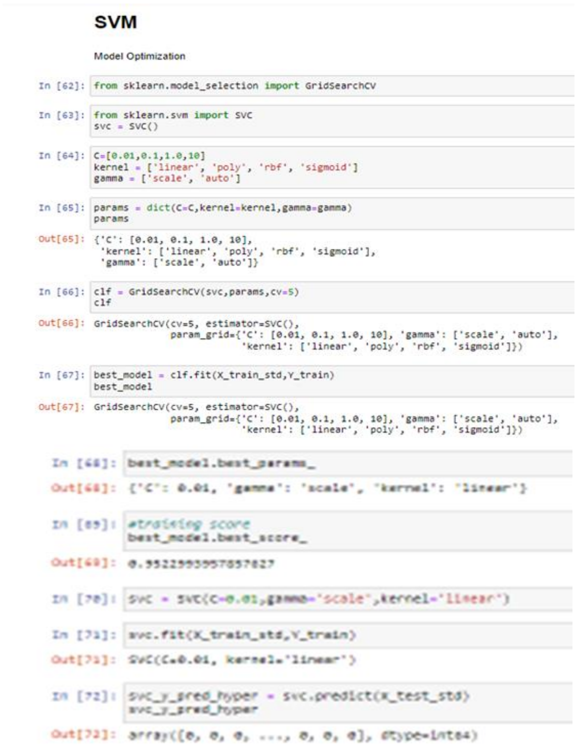
Random Forest

```
In [53]: from sklearn.ensemble import RandomForestClassifier
         rf = RandomForestClassifier()

In [54]: rf.fit(X_train_std,Y_train)
Out[54]: RandomForestClassifier()

In [66]: rf_y_pred = rf.predict(X_test_std)
         rf_y_pred
Out[55]: array([0, 0, 0, ..., 0, 0, 0], dtype=int64)
```



KNN

```
In [60]: from sklearn.neighbors import KNeighborsClassifier
         knn=KNeighborsClassifier()

In [61]: knn.fit(X_train_std,Y_train)
Out[61]: KNeighborsClassifier()

In [62]: Y_pred=knn.predict(X_test_std)
         Y_pred
Out[62]: array([0, 0, 0, ..., 0, 0, 0], dtype=int64)
```



SVM

Model Optimization

```
In [62]: from sklearn.model_selection import GridSearchCV

In [63]: from sklearn.svm import SVC
         svc = SVC()

In [64]: C=[0.01,0.1,1.0,10]
         kernel = ['linear', 'poly', 'rbf', 'sigmoid']
         gamma = ['scale', 'auto']

In [65]: params = dict(C=C,kernel=kernel,gamma=gamma)
         params

Out[65]: {'C': [0.01, 0.1, 1.0, 10],
          'kernel': ['linear', 'poly', 'rbf', 'sigmoid'],
          'gamma': ['scale', 'auto']}

In [66]: clf = GridSearchCV(svc,params,cv=5)
         clf

Out[66]: GridSearchCV(cv=5, estimator=SVC(),
                      param_grid={'C': [0.01, 0.1, 1.0, 10], 'gamma': ['scale', 'auto'],
                                  'kernel': ['linear', 'poly', 'rbf', 'sigmoid']})

In [67]: best_model = clf.fit(X_train_std,Y_train)
         best_model

Out[67]: GridSearchCV(cv=5, estimator=SVC(),
                      param_grid={'C': [0.01, 0.1, 1.0, 10], 'gamma': ['scale', 'auto'],
                                  'kernel': ['linear', 'poly', 'rbf', 'sigmoid']})

In [68]: best_model.best_params_

Out[68]: {'C': 0.01, 'gamma': 'scale', 'kernel': 'linear'}

In [69]: #training score
         best_model.best_score_

Out[69]: 0.9522993957857827

In [70]: svc = SVC(C=0.01,gamma='scale',kernel='linear')

In [71]: svc.fit(X_train_std,Y_train)

Out[71]: SVC(C=0.01, kernel='linear')

In [72]: svc_y_pred_hyper = svc.predict(X_test_std)
         svc_y_pred_hyper

Out[72]: array([0, 0, 0, ..., 0, 0, 0], dtype=int64)
```

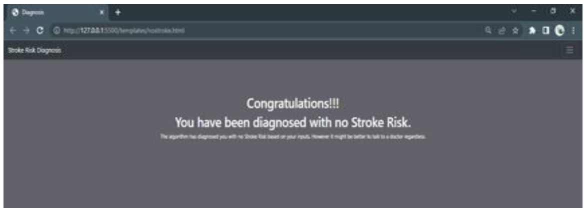
ACCURACY OF EACH ALGORITHM

ALGORITHM	PERCENTAGE ACCURACY
Logistic Regression	95.71%
Decision Tree	90.21%
KNN	94.52%
SVM	94.71%
Random Forest	94.52%

CONCLUSION

To conclude the paper, a machine learning system has been created which would alert the person using about a probable future brain stroke and further suggests to consult a medical professional. The GUI is made using HTML, CSS, Flask. We get a total accuracy of 97%.

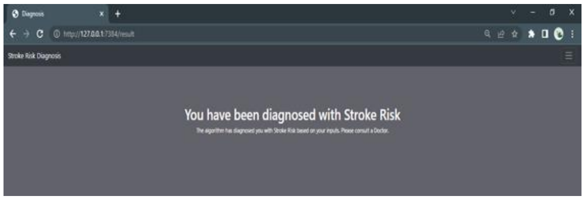
SCREENSHOT OF UI



Congratulations!!!

You have been diagnosed with no Stroke Risk.

The algorithm has diagnosed you with no Stroke Risk based on your inputs. However it might be better to talk to a doctor regardless.



You have been diagnosed with Stroke Risk

The algorithm has diagnosed you with Stroke Risk based on your inputs. Please consult a Doctor.

REFERENCES

[1] Concept of Stroke by Healthline.

[2] Stroke by Center for Disease Control and Prevention.

[3] Dataset named ‘Stroke Prediction Dataset’ from Kaggle: https://www.kaggle.com/fedesoriano/stroke-prediction-dataset.

[4] Stroke Prediction Using Machine Learning Algorithms: https://ijirem.org/DOC/2-stroke-prediction-using-machine-learning-algorithms.pdf

[5] Stroke prediction using SVM R S Jeena; Sukesh Kumar https://ieeexplore.ieee.org/document/7988020

[6] P. Govindarajan, R. K. Soundarapandian, A. H. Gandomi, R. Patan, P. Jayaraman, and R. Manikandan, stroke disease classification using machine learning algorithms,” Neural Computing and Applications in 2019.

[7] Sung, S.M., Kang, Y.J., Cho, H.J., Kim, N.R., Lee, S.M., Choi, B.K., Cho, G. (2020). early neurological prediction deterioration by machine learning algorithms. CN and Neurosurgery.

[8] M. S. Singh and P. Choudhary prediction of stroke using artificial intelligence IN AIA. Electromechanical Engineering Conference (IEMECON).