IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⋇ Impact Factor 7.105 ⋇ Vol. 9, Issue 6, June 2022

DOI: 10.17148/IARJSET.2022.9620

STROKE PREDICTION USING MACHINE LEARNING

Sathya Sundaram .M[1], Pavithra.K[2] , Poojasree.V[3] , Priyadharshini.S[4]

[1]Assistant Professor, Computer Science Engineering,Paavai Engineering College, Namakkal, Tamilnadu

[2] UG - Computer Science Engineering,Paavai Engineering College, Namakkal, Tamilnadu

[3] UG - Computer Science Engineering,Paavai Engineering College, Namakkal, Tamilnadu

[4] UG - Computer Science Engineering,Paavai Engineering College, Namakkal, Tamilnadu

Abstract: A Stroke is a health condition that causes damage by tearing the blood vessels in the brain. It can also occur when there is a halt in the blood flow and other nutrients to the brain. According to the World Health Organization (WHO), stroke is the leading cause of death and disability globally. Earlyrecognition of the various warning signs of a stroke can help reduce the severityof the stroke. Different machine learning (ML) models have been developed to predict the likelihood of a stroke occurring in the brain. This research uses a range of physiological parameters and machine learning algorithms, such as Logistic Regression (LR), Decision Tree (DT) Classification, Random Forest (RF) Classification, and Voting Classifier, to train four different models for reliable prediction. Random Forest was the best performing algorithm for this task with an accuracy of approximately 96 percent. The dataset used in the development of the method was the open-access Stroke Prediction dataset. The accuracy percentage of the models used in this investigation is significantly higher than that of previous studies, indicating that the models used in this investigation are more reliable. Numerous model comparisons have established their robustness, and the scheme can be deduced from the study analysis.

Keywords: Stroke; machine learning; logistic regression; decision treeclassification; random forest classification; k-nearest neighbors; support vector machine; Naïve Bayes classification.

1.INTRODUCTION

According to the Centers for Disease Control and Prevention (CDC), stroke is the fifth-leading cause of death in the United States. Stroke is a non-communicable infection that is liable for around 11% of total deaths. Consistently, over 795,000 individuals in the United States experience the ill effects of a stroke . It is the fourth significant reason for death in India. With the advancement of technology in the medical field, predicting the occurrence ofa stroke can be made using Machine Learning. The algorithms present in Machine Learning are constructive in making an accurate prediction and give correct analysis. The workspreviously performed on stroke mostly include the ones on Heart stroke prediction. Very less works have been performed on Brain stroke. This paper is based on predicting the occurrenceof a brain stroke using Machine Learning. The key components of the approaches used and results obtained are that among the five different classification algorithms used Naïve Bayes has best performed obtaining a higher accuracy metric. The limitation with this model is thatit is being trained on textual data and not on real time brain images. The paper shows the implementation of six Machine Learning classification algorithms. This paper can be further extended to implementing all the current machine learning algorithms. A dataset is chosen from Kaggle with various physiological traits as its attributes to proceed with this task.

These traits are later analyzed and used for the final prediction. The dataset is initially cleaned and made ready for the machine learning model to understand. This step is called Data Preprocessing. For this, the dataset is checked for null values and fill them. Then Label encoding is performed to convert string values into integers followed by one-hot encoding, if necessary. After Data Preprocessing, the dataset is split into train and test data. A model is then built using this new data using various Classification Algorithms. Accuracy is calculatedfor all these algorithms and compared to get the best-trained model for prediction. After training the model and calculating the accuracy, an HTML page and a Flask application are developed. The web application is for the user to enter the values for prediction. The flask application is a framework that connects the trained model and the web application. After proper analysis, the paper concludes which algorithm is most appropriate for the prediction ofstroke.

DOI: 10.17148/IARJSET.2022.9620

OBJECTIVE
Describe a method for determining if a person is having a stroke.

- The prime objective of this project is to construct a prediction model for predicting stroke using machine learning algorithms.
- The dataset was obtained from Kaggle website "Healthcare dataset stroke data".Categorical features, numerical features and multi collinearity analysis will be carried on for better understanding of the data.
- Five different models - SVM, Decision tree, Random Forest, K-nearest neighbor, Logistic regression are considered.
- Finally, better performing algorithm will be chose to predict stroke and a simple Graphical User Interface is created using tkinter.

2.EXPERIMENTAL METHODS OR METHODOLOGY

- There is limited previous work on utilizing machine learning algorithms to estimate perfusion parameters. In this work, we present a novel bi-input convolutional neural network (bi-CNN) to approximate four perfusion parameters without using an explicitdeconvolution method.
- These bi-CNNs produced good approximations for all four parameters, with relative average root Mean-Square Errors (MSE) and Mean Absolute Error (MAE) less than equal of the maximum values.
- These results show that machine learning techniques area promising tool for perfusionparameter estimation without requiring a standard deconvolution process.

ADVANTAGE

- Early prediction of stroke can be done.
- The cost of medication will be minimized.
- Accuracy rate will be high
- High performance.

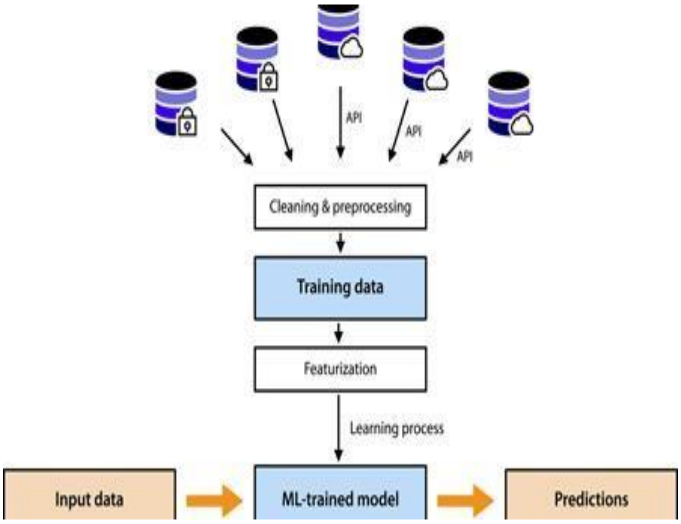


figure 1 Processing Methods

1) Random forest
2) Decision tree
3) Logistic regression

1.RANDOM FOREST.

The classification algorithm chosen was RF classification . RFs are composed of numerous independent decision trees that were trained individually on a random sample of data. -ese trees are created during training, and the decision trees’ outputs are collected. A process termed voting is used to determine the final forecast

DOI: 10.17148/IARJSET.2022.9620

made by this algorithm. EachDT in this method must vote for one of the two output classes (in this case, stroke or no stroke). The final prediction is determined by the RF method, which chooses the class wit the most votes. It may be utilized for relapse detection and grouping tasks, and the overall weighting given to information characteristics is readily apparent. Additionally, it is a beneficial approach since the default hyperparameters it employs often give unambiguous expectations. Understanding the hyperparameters is critical since there are relatively few of them, to begin with. Overfitting is a wellknown problem in machine learning, although it occurs seldom with the arbitrary random forest classifier. If there are sufficient trees in the forest, the classifier will not overfit the model.

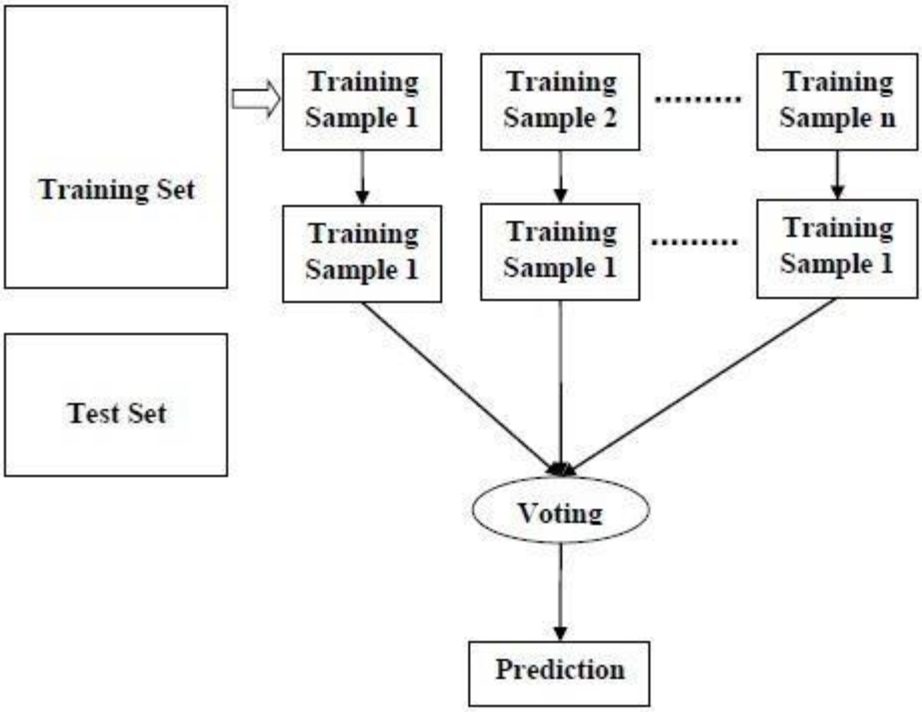


figure 2 Random Forest Classifier

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.

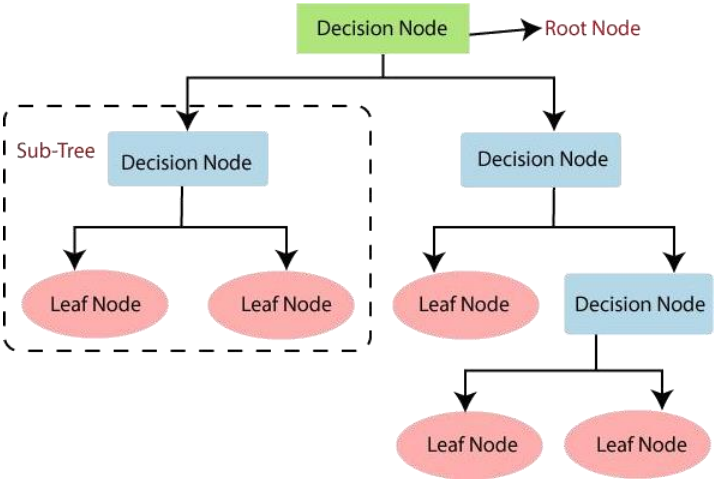


figure 3 decision tree classifier

DOI: 10.17148/IARJSET.2022.9620

LOGISTIC REGRESSION.

The flowchart for the logistic regression model . In the supervised learning approach, LR is one of the most commonly used ML algorithms . It is a forecasting method that uses a collection of independent factors to predict a categorical dependent variable. Utilizing logisticregression, the output of a categorical dependent variable is predicted. As a result, the output must be discrete or categorical in nature. It may be yes or no, 0 or 1, true or false, etc., but probability values between 0 and 1 are given. The classification problems are addressed with LR, and the regression problems are addressed using linear regression. Instead of a regression line, we usean S-shaped logistic function that predicts the two maximum values (0 or 1).

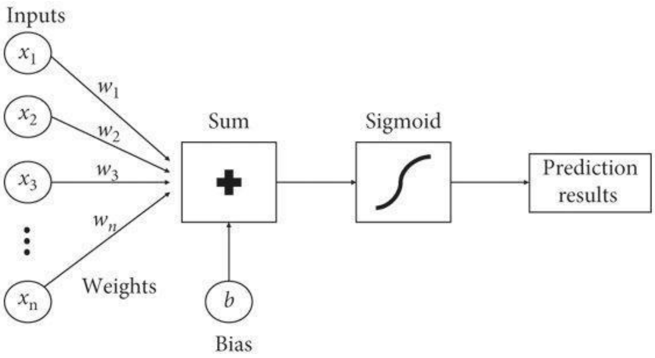
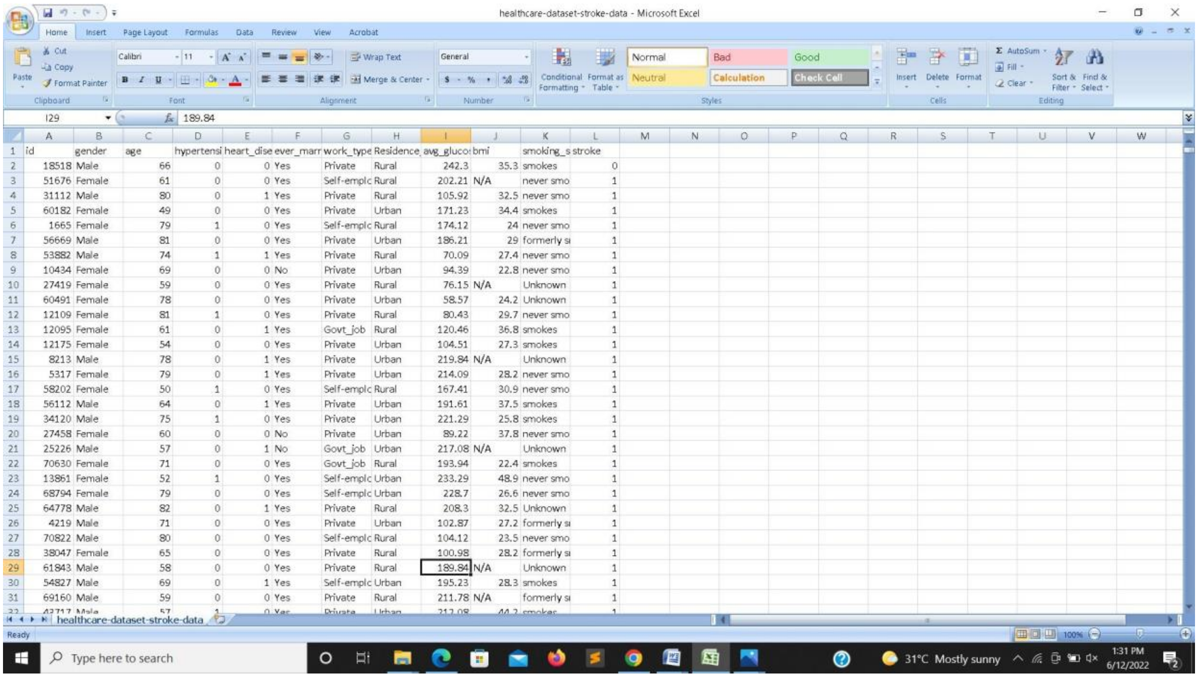


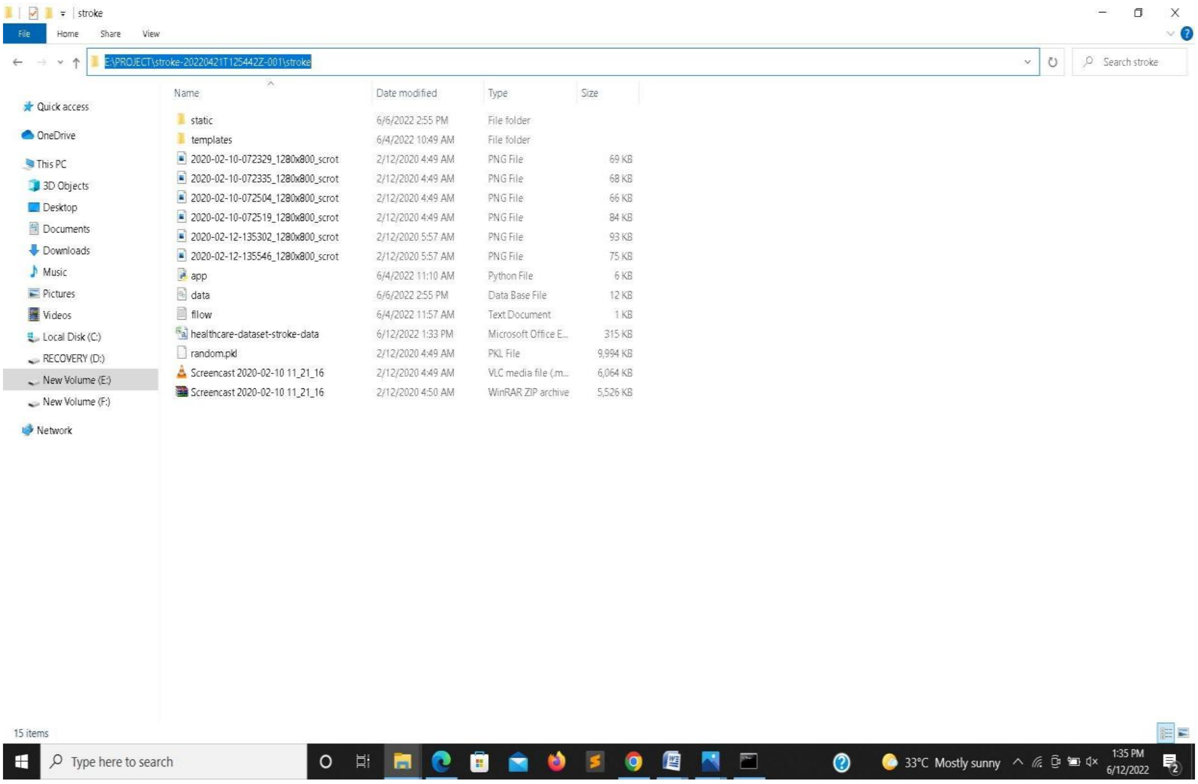
figure 4 logistic regression

RESULTS AND DISCUSSION SCREENSHOTS

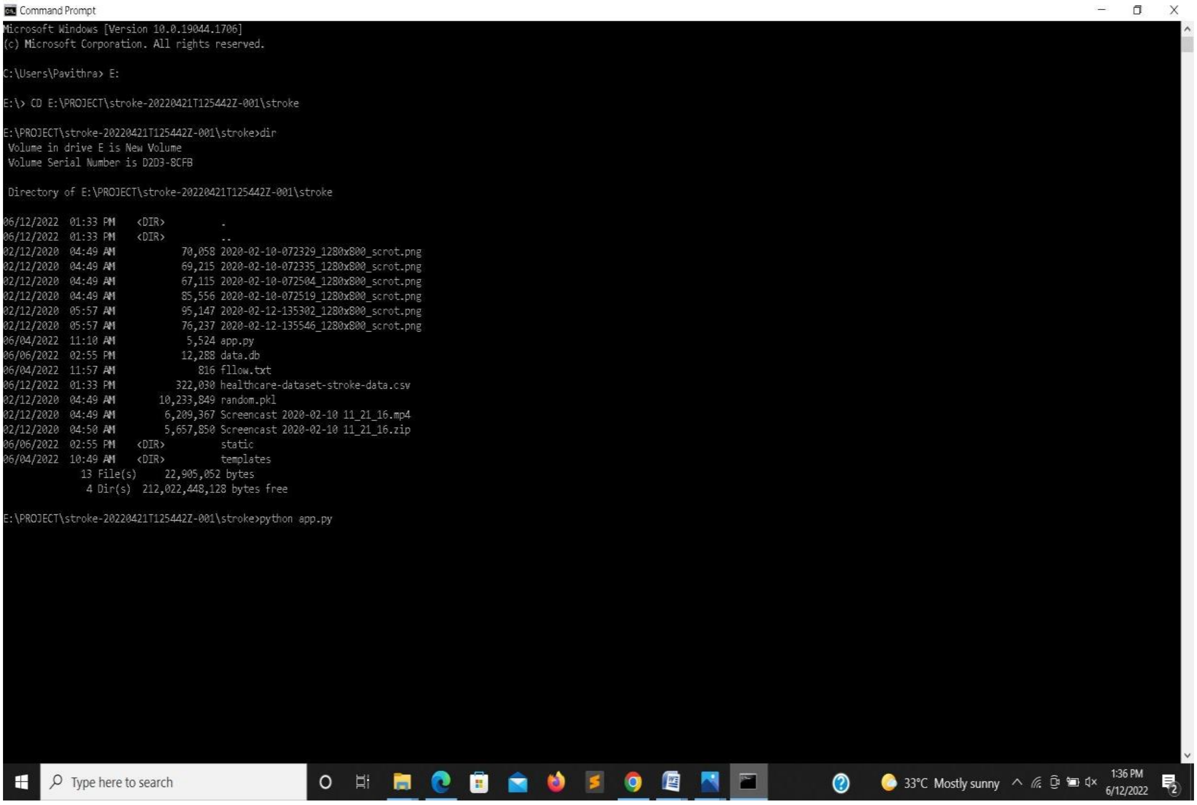


	A	B	C	D	E	F	G	H	I	J	K	L
1	id	gender	age	hypertensi	heart_dise	ever_marr	work_type	Residence_	avg_gluco	bmi	smoking_s	stroke
2	18518	Male	66	0	0	Yes	Private	Rural	242.3	35.3	smokes	0
3	51676	Female	61	0	0	Yes	Self-emplc	Rural	202.21	N/A	never smo	1
4	31112	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smo	1
5	60182	Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
6	1665	Female	79	1	0	Yes	Self-emplc	Rural	174.12	24	never smo	1
7	56669	Male	81	0	0	Yes	Private	Urban	186.21	29	formerly sı	1
8	53882	Male	74	1	1	Yes	Private	Rural	70.09	27.4	never smo	1
9	10434	Female	69	0	0	No	Private	Urban	94.39	22.8	never smo	1
10	27419	Female	59	0	0	Yes	Private	Rural	76.15	N/A	Unknown	1
11	60491	Female	78	0	0	Yes	Private	Urban	58.57	24.2	Unknown	1
12	12109	Female	81	1	0	Yes	Private	Rural	80.43	29.7	never smo	1
13	12095	Female	61	0	1	Yes	Govt_job	Rural	120.46	36.8	smokes	1
14	12175	Female	54	0	0	Yes	Private	Urban	104.51	27.3	smokes	1
15	8213	Male	78	0	1	Yes	Private	Urban	219.84	N/A	Unknown	1
16	5317	Female	79	0	1	Yes	Private	Urban	214.09	28.2	never smo	1
17	58202	Female	50	1	0	Yes	Self-emplc	Rural	167.41	30.9	never smo	1
18	56112	Male	64	0	1	Yes	Private	Urban	191.61	37.5	smokes	1
19	34120	Male	75	1	0	Yes	Private	Urban	221.29	25.8	smokes	1
20	27458	Female	60	0	0	No	Private	Urban	89.22	37.8	never smo	1
21	25226	Male	57	0	1	No	Govt_job	Urban	217.08	N/A	Unknown	1
22	70630	Female	71	0	0	Yes	Govt_job	Rural	193.94	22.4	smokes	1
23	13861	Female	52	1	0	Yes	Self-emplc	Urban	233.29	48.9	never smo	1
24	68794	Female	79	0	0	Yes	Self-emplc	Urban	228.7	26.6	never smo	1
25	64778	Male	82	0	1	Yes	Private	Rural	208.3	32.5	Unknown	1
26	4219	Male	71	0	0	Yes	Private	Urban	102.87	27.2	formerly sı	1
27	70822	Male	80	0	0	Yes	Self-emplc	Rural	104.12	23.5	never smo	1
28	38047	Female	65	0	0	Yes	Private	Rural	100.98	28.2	formerly sı	1
29	61843	Male	58	0	0	Yes	Private	Rural	189.84	N/A	Unknown	1
30	54827	Male	69	0	1	Yes	Self-emplc	Urban	195.23	28.3	smokes	1
31	69160	Male	59	0	0	Yes	Private	Rural	211.78	N/A	formerly sı	1
32	43717	Male	57	1	0	Yes	Private	Urban	217.08	44.2	smokes	1

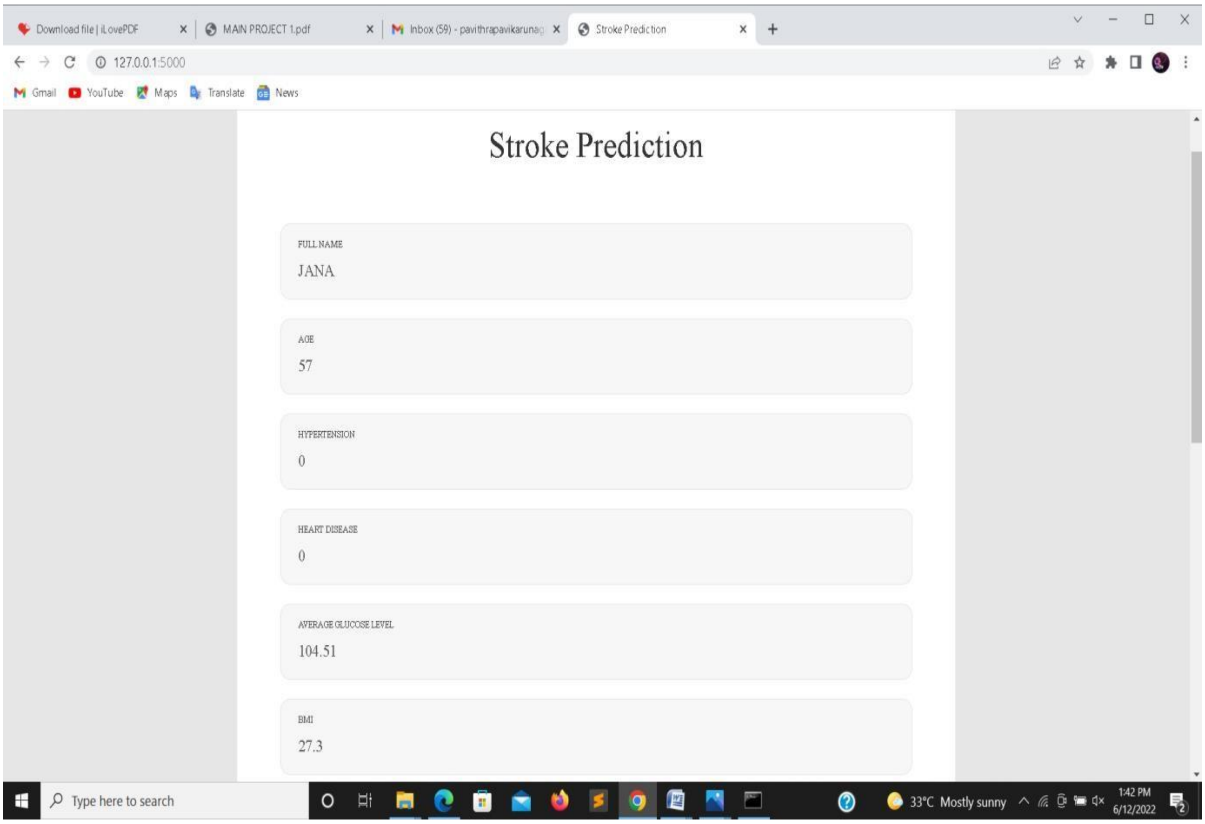
DOI: 10.17148/IARJSET.2022.9620



Name	Date modified	Type	Size
static	6/6/2022 2:55 PM	File folder	
templates	6/4/2022 10:49 AM	File folder	
2020-02-10-072329_1280x800_scrot	2/12/2020 4:49 AM	PNG File	69 KB
2020-02-10-072335_1280x800_scrot	2/12/2020 4:49 AM	PNG File	68 KB
2020-02-10-072504_1280x800_scrot	2/12/2020 4:49 AM	PNG File	66 KB
2020-02-10-072519_1280x800_scrot	2/12/2020 4:49 AM	PNG File	84 KB
2020-02-12-135302_1280x800_scrot	2/12/2020 5:57 AM	PNG File	93 KB
2020-02-12-135546_1280x800_scrot	2/12/2020 5:57 AM	PNG File	75 KB
app	6/4/2022 11:10 AM	Python File	6 KB
data	6/6/2022 2:55 PM	Data Base File	12 KB
fllow	6/4/2022 11:57 AM	Text Document	1 KB
healthcare-dataset-stroke-data	6/12/2022 1:33 PM	Microsoft Office E...	315 KB
random.pkl	2/12/2020 4:49 AM	PKL File	9,994 KB
Screencast 2020-02-10 11_21_16	2/12/2020 4:49 AM	VLC media file (.m...	6,064 KB
Screencast 2020-02-10 11_21_16	2/12/2020 4:50 AM	WinRAR ZIP archive	5,526 KB



```
Microsoft Windows [Version 10.0.19044.1706]
(c) Microsoft Corporation. All rights reserved.

C:\Users\Pavithra> E:

E:\> CD E:\PROJECT\stroke-20220421T125442Z-001\stroke

E:\PROJECT\stroke-20220421T125442Z-001\stroke>dir
 Volume in drive E is New Volume
 Volume Serial Number is D2D3-8CFB

 Directory of E:\PROJECT\stroke-20220421T125442Z-001\stroke

06/12/2022  01:33 PM    <DIR>          .
06/12/2022  01:33 PM    <DIR>          ..
02/12/2020  04:49 AM            70,058 2020-02-10-072329_1280x800_scrot.png
02/12/2020  04:49 AM            69,215 2020-02-10-072335_1280x800_scrot.png
02/12/2020  04:49 AM            67,115 2020-02-10-072504_1280x800_scrot.png
02/12/2020  04:49 AM            85,556 2020-02-10-072519_1280x800_scrot.png
02/12/2020  05:57 AM            95,147 2020-02-12-135302_1280x800_scrot.png
02/12/2020  05:57 AM            76,237 2020-02-12-135546_1280x800_scrot.png
06/04/2022  11:10 AM             5,524 app.py
06/06/2022  02:55 PM            12,288 data.db
06/04/2022  11:57 AM               816 fllow.txt
06/12/2022  01:33 PM           322,030 healthcare-dataset-stroke-data.csv
02/12/2020  04:49 AM        10,233,849 random.pkl
02/12/2020  04:49 AM         6,209,367 Screencast 2020-02-10 11_21_16.mp4
02/12/2020  04:50 AM         5,657,850 Screencast 2020-02-10 11_21_16.zip
06/06/2022  02:55 PM    <DIR>          static
06/04/2022  10:49 AM    <DIR>          templates
              13 File(s)     22,905,052 bytes
               4 Dir(s)  212,022,448,128 bytes free

E:\PROJECT\stroke-20220421T125442Z-001\stroke>python app.py
```



Stroke Prediction

FULL NAME
JANA

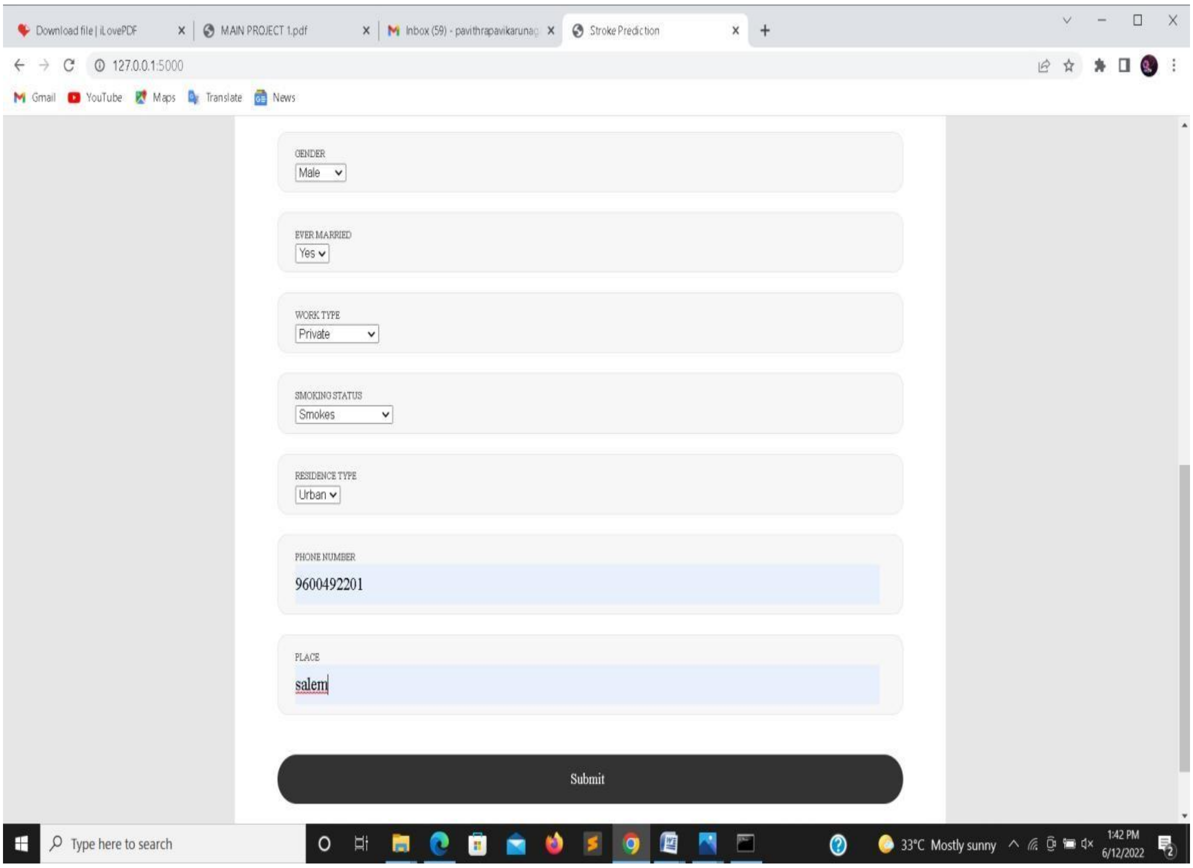
AGE
57

HYPERTENSION
0

HEART DISEASE
0

AVERAGE GLUCOSE LEVEL
104.51

BMI
27.3



GENDER
Male

EVER MARRIED
Yes

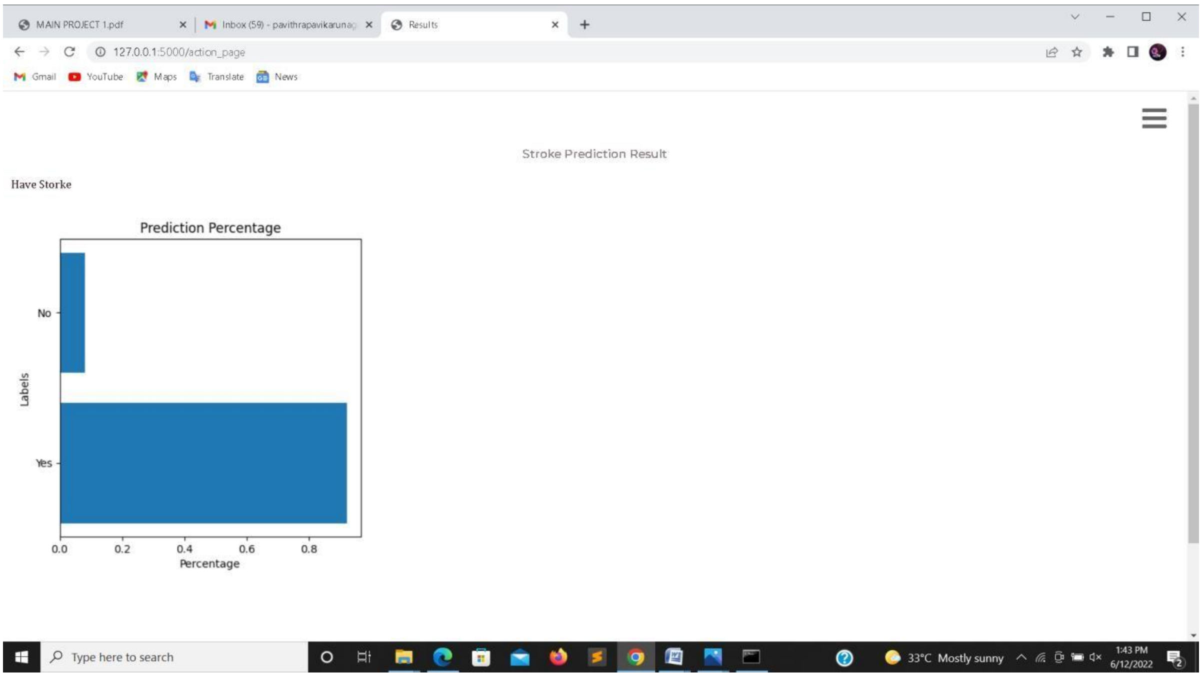
WORK TYPE
Private

SMOKING STATUS
Smokes

RESIDENCE TYPE
Urban

PHONE NUMBER
9600492201

PLACE
salem



Stroke Prediction Result

Have Storke

CONCLUSION

By doing so, it urges medical users to strengthen the motivation of health management and induce changes in their health behaviors. A model for predicting stroke using machine learning algorithms. After, thoroughly reviewing various IEEE papers we selected five different models such as decision tree, random forest and logistic regression . Key attributes/features were selected under the guidance of medical practitioners. Visualizing health data allows professionals to present key/common trends and information via graphs, charts and visuals that helps even a data analysts understand the dataset. Hence, data visualization was our main objective. Used libraries like pandas, matplotlib, seaborn and Pywaffle for informative and attractive representation of data. Predictive analytics is a popular business intelligence trend. They help doctors make data driven decisions in no time which can even predict and prevent deadly diseases. In this project, we have carried on categorical feature analysis, numerical feature analysis and multicollinearity successfully. Applied different model on the dataset. A comparative study amongst the five different models showed that random forest, logistic regression and K nearest neighbor has an accuracy of 95.5%, whereas decision tree was 91.13% accurate and support vector machine exhibited accuracy of 92.43%. Finally, Random Forest was chosen as the best model with high accuracy and less false negative. To facilitate seamless use of the application, a Graphical User Interface (GUI) was created using tkinter.

FUTURE SCOPE

Stroke is dependent on a lifestyle attributes as well as past medical history. Here inthis paper, we have considered seven lifestyle attributes and three medical conditions. In the future, or better performance of the model more medical attributes can be considered suchhas Systolic blood pressure, diastolic blood pressure, pulse pressure, mean blood pressure, The min, max and mean value of a pulse. Also, mRS score, NIHSS score, CHADS2 score can be added to get a more accurate and precise output.

REFERENCES

1.A predictive analytics approach for stroke prediction using machine learning and neural network soumyddbrata Dev a,b , Hewei Wang c,d , Chidozie Shamrock Nwosu , Nishtha Jain , Bharadwaj Veeravalli , Deepu John Healthcare Analytics 2 (2022) 100032.

2.Analyzing the Performance of Stroke Prediction using ML Classification Algorithms Gangavarapu Sailasya1 , Gorli L Aruna Kumari2 (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 12, No. 6, 2021.

3.Stroke Prediction Using Machine Learning Algorithms, Gangavarapu Sailasya , Gorli L Aruna Kumari, International Journal of Innovative Research in Engineering & Management (IJIREM) ISSN: 2350-0557, Volume-8, Issue-4, July

2021 https://doi.org/10.21276/ijirem.2021.8.4.2 www.ijirem.org.

4.Stroke Disease Detection and Prediction Using Robust Learning Approaches Tahia Tazin , 1 Md Nur Alam,1 Nahian Nakiba Dola,1 Mohammad Sajibul Bari,1 Sami Bourouis , 2 and Mohammad Monirujjaman Khan, Hindawi Journal of Healthcare Engineering Volume 2021, Article ID 7633381, 12 pages https://doi.org/10.1155/2021/7633381.

5.Pradeepa, S., Manjula, K. R., Vimal, S., Khan, M. S., Chilamkurti, N., & Luhach, A. K.: DRFS: Detecting Risk Factor of Stroke Disease from Social Media Using Machine Learning Techniques. In Springer (2020).

6.Vamsi Bandi, Debnath Bhattacharyya, Divya Midhunchakkravarthy: Prediction of Brain Stroke Severity Using Machine Learning. In: International Information and Engineering Technology Association (2020).

7.Nwosu, C.S., Dev, S., Bhardwaj, P., Veeravalli, B., John, D.: Predicting stroke from electronic health records. In: 41st Annual International Conference of the IEEE Engineering

8.Fahd Saleh Alotaibi: Implementation of Machine Learning Model to Predict Heart Failure Disease. In: International Journal of Advanced Computer Science and Applications (IJACSA) (2019).

9.Ohoud Almadani, Riyad Alshammari: Prediction of Stroke using Data Mining Classification Techniques. In: International Journal of Advanced Computer Science and Applications (IJACSA) (2018)

10.Jeena R.S and Dr.Sukesh Kumar “Stroke prediction using SVM”, International Conference on Computing, Communication and Networking Technologies (ICCCNT), IEEE, 2016.