

Brain Stroke Prediction Portal Using Machine Learning

Atharva Kshirsagar, Student, Mumbai, India, atharvaksh@gmail.com

Harsh Goyal, Student, Mumbai, India, harshgoyal1216@gmail.com

Shubham Loya, Student, Mumbai, India, shubhamloya156@gmail.com

Anindita Khade, Assistant Professor, Mumbai, India, anindita.khade@siesgst.ac.in

Abstract: A stroke occurs when the blood supply to part of your brain is interrupted or reduced, preventing brain tissue from getting oxygen and nutrients. Brain cells begin to die in minutes. A stroke is a medical emergency, and prompt treatment is crucial. Early action can reduce brain damage and other complications. In 2015, there were about 42.4 million people who had previously had a stroke and were still alive. In 2015, stroke was the second most frequent cause of death after coronary artery disease, accounting for 6.3 million deaths. About 3.0 million deaths resulted from ischemic stroke while 3.3 million deaths resulted from hemorrhagic stroke. Hence, correct detection and finding presence of stroke inside a human becomes essential. There are various medical instruments available in the market for predicting brain stroke but they are very much expensive and they are not efficient enough to be able to calculate the chance of having a brain stroke. So, there is a need to find better and efficient approach to diagnose brain strokes at an early stage

Keywords -- Brain Stroke; Random Forest (RF); Extreme Gradient Boosting (XGB); K Nearest Neighbors (KNN); Machine Learning (ML); Prediction; Support Vector Machines (SVM).

I. INTRODUCTION

In recent times, the stress levels in individuals are at an all time high. This increases the chances of strokes in individuals. In 2015, stroke was the second most frequent cause of death after coronary artery disease, accounting for 6.3 million deaths. About 3.0 million deaths resulted from ischemic stroke while 3.3 million deaths resulted from hemorrhagic stroke. The figures today are much higher than this. An exhaustive and easy to use tool is much needed for detection of strokes. With advancement of computer science in different research areas including medical sciences, this has been made possible. A machine-learning system is trained rather than explicitly programmed, as it provides a better choice for achieving high accuracy for detection of heart diseases. Medical organizations, all around the world, collect data on various health related issues. These data can be exploited using various machine learning techniques to gain useful insights. But the data collected is very massive and, many times, this data can be very noisy. These datasets, which are too overwhelming for human minds to comprehend, can be easily explored using various machine learning techniques. Thus, these algorithms have become very useful, in recent times, to predict the presence or absence of heart related diseases accurately. Stroke is the second leading cause of death worldwide and remains an important health burden both for the individuals and for the national healthcare systems. The main aim of this project is to build an efficient prediction

model and deploy for prediction of disease. Machine Learning is a faster-emerging technology of Artificial Intelligence that contributes various algorithms like Logistic Regression, SVM, Random Forests and many more which is effective in making decisions and predictions from the large quantity of data produced by the healthcare industry. Based on the proposed problem, ML provides different classification algorithms to divine the probability of a patient having a Brain Stroke.

II. LITERATURE SURVEY

In order to get required knowledge about various concepts related to the present analysis, existing literature was studied. Some of the important conclusions were made through those are listed below.

[1] "Computer Methods and Programs in Biomedicine" - Jae-woo Lee, Hyunsun Lim, Dong-wook Kim, Soon-ae Shin, Jinkwon Kim, Bora Yoo, Kyunghye Cho - The Purpose of this paper was Calculation of 10-year stroke prediction probability and classifying the user's individual probability of stroke into five categories.

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

[3] "Deep learning algorithms for detection of critical

findings in head CT scans: a retrospective study” - Rohit Ghosh, Swetha Tanamala, Mustafa Biviji, Norbert G Campeau, Vasantha Kumar Venugopal - In this paper Non-contrast head CT scan is the current standard for initial imaging of patients with head trauma or stroke symptoms. This article aimed to develop and validate a set of deep learning algorithms for automated detection.

[4] “Prediction of stroke thrombolysis outcome using CT brain machine learning” - Paul Bentley, JebanGanesalingam, AnomaLalani, CarltonJones, KateMahady, SarahEpton, PaulRinne, PankajSharma, OmidHalse, AmrishMehta, DanielRueckert - Clinical records and CT brains of 116 acute ischemic stroke patients treated with intravenous thrombolysis were collected retrospectively (including 16 who developed SICH). The sample was split into training (n = 106) and test sets (n = 10), repeatedly for 1760 different combinations. CT brain images acted as inputs into a support vector machine (SVM), along with clinical severity. Predictive performance, assessed as area under receiver-operating-characteristic curve (AUC), of the SVM (0.744) compared favourably with that of prognostic scores (original and adapted versions: 0.626-0.720; $p < 0.01$).

[5] “Probability of Stroke: A Risk Profile from the Framingham Study” - Philip A. Wolf, MD; Ralph B. D'Agostino, PhD, Albert J. Belanger, MA; and William B. Kannel, MD - In this paper, A health risk appraisal function has been developed for the prediction of stroke using the Framingham Study cohort.

[6] “Development of an Algorithm for Stroke Prediction: A National Health Insurance Database Study” -Min SN, Park SJ, Kim DJ,

Subramaniyam M, Lee KS – In this research, this paper aimed to derive a model equation for developing a stroke pre- diagnosis algorithm with the potentially modifiable risk factors.

[7] “Medical software user interfaces, stroke MD application design (IEEE)” Elena Zamsa-The article presents the design of an application interface for associated medical data visualization and management for neurologists in a stroke clustering and prediction system called Stroke MD.

[8] “Focus on stroke: Predicting and preventing stroke” Michael Regnier- This paper focuses on cutting-edge prevention of stroke.

[9] “Effective Analysis and Predictive Model of Stroke Disease using Classification Methods”-A.Sudha, P.Gayathri, N.Jaisankar- This paper, principal component analysis algorithm is used for reducing the dimensions and it determines the attributes involving more towards the prediction of stroke disease and predicts whether the patient is suffering from stroke disease or not.

[10] “Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study” - Rohit Ghosh, Swetha Tanamala, Mustafa Biviji, Norbert G

Campeau, Vasantha Kumar Venugopal - In this paper Non-contrast head CT scan is the current standard for initial imaging of patients with head trauma or stroke symptoms. This article aimed to develop and validate a set of deep learning algorithms for automated detection

III. METHODOLOGY

The methodology we are proposing is that we can create a web application wherein people can put in their data and the proposed system can easily identify and classify people with brain stroke from healthy people.

i) Algorithms:

The various algorithms tested were:

1. Logistic Regression
2. Support Vector Machine
3. KNN: Accuracy
4. Random Forest
5. XG Boost

Logistic Regression:

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labeled "0" and "1". In the logistic model, the log-odds (the logarithm of the odds) for the value labeled "1" is a linear combination of one or more independent variables ("predictors"); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value).

Support Vector Machine (SVM):

These classifiers depend on primary danger minimization head and factual learning hypothesis with a point of deciding the hyperplanes (choice limits) that produce the effective division of classes. The basic calculation is Support Vector Classification (SVC) and it spins around the impression of a "margin"- on one or the other side of a hyperplane that partitions two information classes.

Augmenting the margin makes the biggest conceivable distance among the hyperplane and the occurrences on one or the other side of the hyperplane decrease an upper bound on the expected speculation blunder. It chips away at two kinds of information i.e., straightly divisible information and directly Non-distinguishable information.

K-Nearest Neighbors:

The k-nearest neighbors classifier is amongst the simplest of all machine learning algorithms. It is based on the principle that the samples that are similar lie in close proximity. It

classifies the test objects on the basis of the number of closest training examples. It is also termed as a lazy-learning algorithm. KNN is a non-parametric algorithm which means that it does not assume anything on the underlying data distribution.

In this, the Euclidean distance is calculated between the test data and every sample in the training data followed by classifying the test data into a class in which most of k-closest neighbor's of training data belong to. K is usually a very small positive integer. As the value of K increases it becomes difficult to distinguish between the various classes. Cross-validation along with other heuristic techniques are used to choose an optimal value of K.

Random Forest:

Random Forest is a mainstream ML calculation that has a place with the administered learning procedure. It tends to be utilized for both Arrangement and Relapse issues in ML. It depends on the idea of ensemble learning, which is a cycle of joining different classifiers to take care of a complex problem,

Extreme Gradient Boosting Classifier:

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework. In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks. However, when it comes to small-to-medium structured/tabular data, decision tree based algorithms are considered best-in-class right now. Please see the chart below for the evolution of tree-based algorithms over the years.

ii) System Design

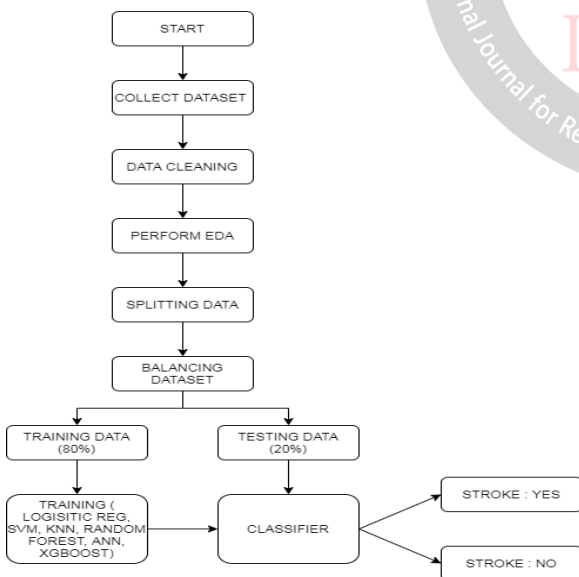


Figure No.1 - Project Flow

The above figure shows the steps involved in executing the project.

The Dataset:

It is a comprehensive dataset of about 5110 to analyze and

rightly predict our results. The dataset was divided in 80% and 20% parts for the training and testing set respectively.

```
data.shape
(5110, 12)
```

Figure No.2 - 5110 Data Points; 12 attributes

The parameters were:

1. Age
2. Gender
3. Hypertension (Yes or No)
4. Heart Disease (Yes or No)
5. Ever Married (Yes or No)
6. Work Type
7. Residence Type
8. Average Glucose Level
9. BMI
10. Smoking Status

Balancing Dataset:

The data acquired was highly imbalanced with the number of data points with 'stroke'=True less than the data points with 'stroke'=False.

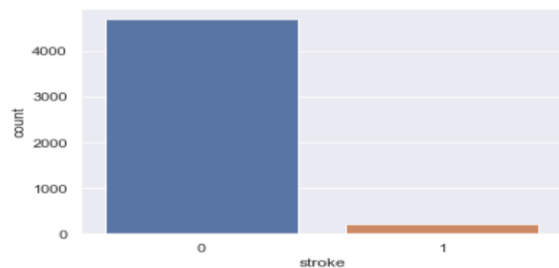


Figure No.3 - Data Imbalance

SMOTE Algorithm: Synthetic Minority Oversampling Technique

SMOTE is an oversampling technique where the synthetic samples are generated for the minority class. This algorithm helps to overcome the overfitting problem posed by random oversampling. It focuses on the feature space to generate new instances with the help of interpolation between the positive instances that lie together.

After applying SMOTE Algorithm:

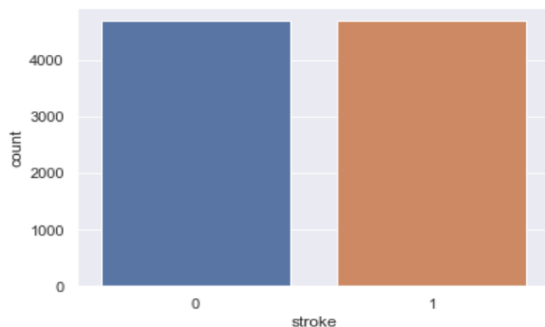


Figure No.4 - Balanced Dataset after application of 'SMOTE'

Correlation Matrix:

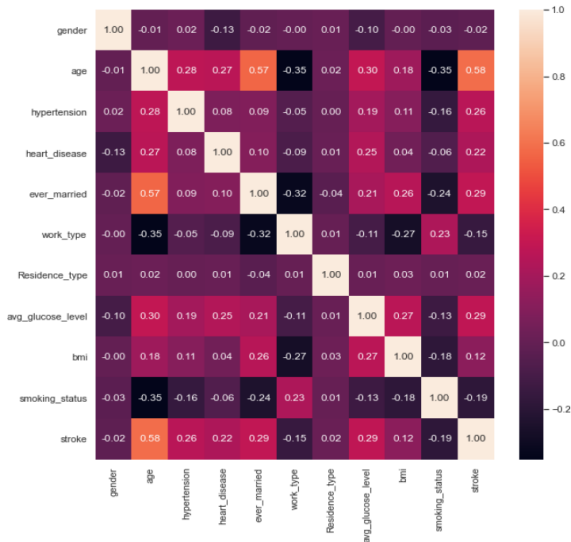


Figure No.5 - Correlation between attributes

In the above heatmap, we can see that there is no multicollinearity present and 'Age' and 'Glucose Level' are some of the highest correlated features with 'Stroke'.

Best Features using Chi-Square Test

	Features	Score	p-value
1	age	28769.699489	0.000000e+00
7	avg_glucose_level	20217.338216	0.000000e+00
2	hypertension	530.644495	2.045199e-117
3	heart_disease	418.051068	6.481220e-93
5	work_type	301.858134	1.297032e-67
9	smoking_status	258.941225	2.919480e-58
8	bmi	239.738634	4.484304e-54
4	ever_married	184.627713	4.732296e-42
0	gender	1.583776	2.082176e-01
6	Residence_type	1.421878	2.330945e-01

Figure No.6 - Best Features

In the above table, we can see that **Age**, **Average Glucose Level** and **Hypertension** are the top 3 features having maximum impact on output 'Stroke'. Chi Square Test is used to find out this result.

IV. EXPERIMENTAL RESULTS

The results of various algorithms implemented are as under:

i) Logistic Regression

Accuracy: **81.18%**

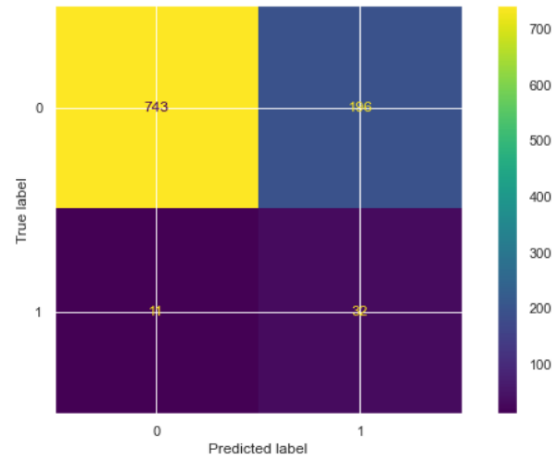


Figure No.7 - Confusion Matrix for Logistic Regression

While the number of true negatives is high for Logistic Regression, the number of Data Points classified as False Negative are substantially high.

ii) Support Vector Machine (SVM)

Accuracy: **81.107%**

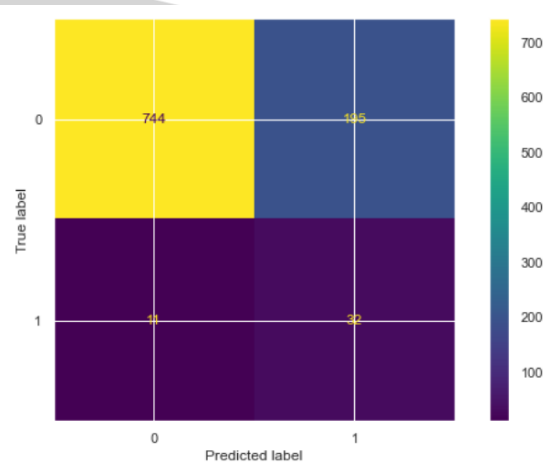


Figure No.8 - Confusion Matrix for SVM

Similarly to Logistic Regression, the False Negative for SVM Model is also very high.

iii) K Nearest Neighbors (KNN)

Accuracy: **88.69%**

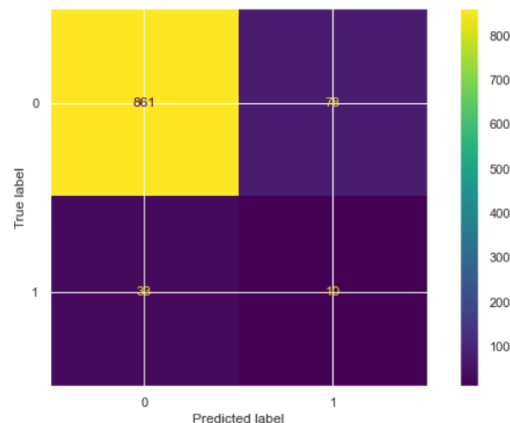


Figure No.9 - Confusion Matrix for KNN

In KNN Model, the number of false positives drastically decreases but the number of false negatives increases.

iv) Random Forest

Accuracy: **91.14%**

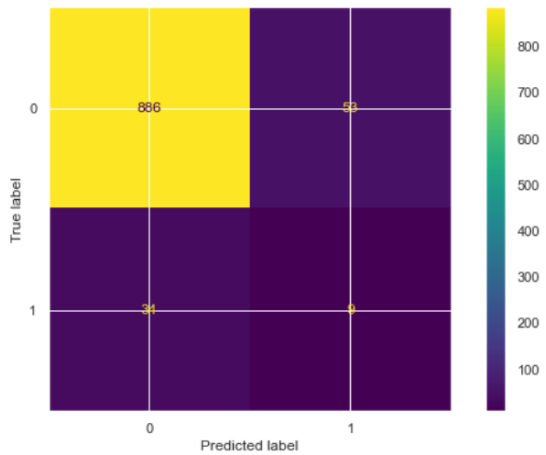


Figure No.10 - Confusion Matrix for Random Forest

In the model using Random Forest, the accuracy obtained is the highest. Similar to KNN, the number of false negatives has increased.

v) XGBoost

Accuracy: **85.03%**

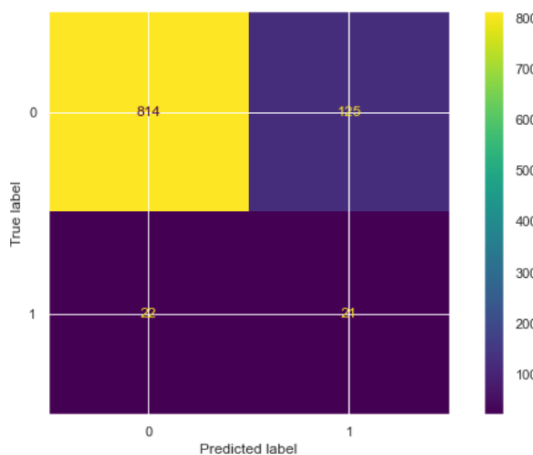


Figure No.11 - Confusion Matrix for XGBoost

In the XGBoost model, although the accuracy is low, after hyperparameter tuning, it is anticipated to give promising results.

Hyperparameter Tuning

We tune the models created for getting an increased accuracy.

Two ways of hyperparameter tuning:

1. GridSearchCV
2. RandomizedSearchCV

Algorithm used:

RandomizedSearchCV on XGBoost Model:

Accuracy after hyperparameter tuning: **93.68%**

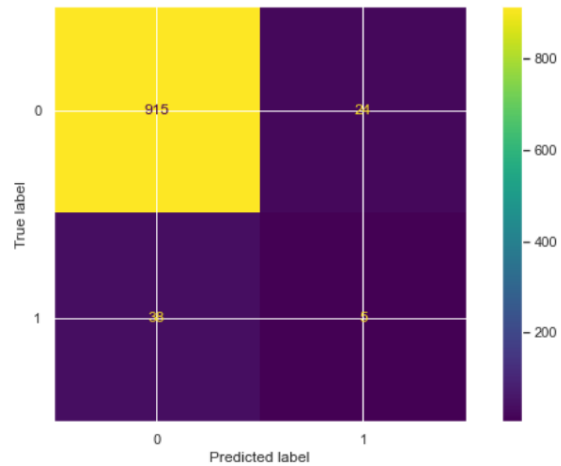


Figure No.12 - Confusion Matrix for tuned XGBoost

As we can observe, after tuning parameters of the XGBoost model using RandomizedSearchCV algorithm, the accuracy jumps to 93.68% which is the highest achieved so far. The number of false positives have also **reduced drastically** which helps alleviate the problem when solving a problem statement in the medical domain.

Summary:

Sr No.	Algorithm	Accuracy (%)
1.	XGBoost	93.68
2.	Random Forest	91.14
3.	KNN	88.69
4.	Logistic Regression	81.18
5.	SVM	81.10

Figure No.13 - Summary of all the models used

In the above table, it can be observed that a total of 5 Machine Learning Algorithms were tested among which, XGBoost gave the maximum accuracy of 93.68%.

Why XGBOOST?

XG Boost is an optimized gradient boosting algorithm. It provides parallel processing, tree-pruning, handling missing values, and regularization to avoid over fitting or bias.

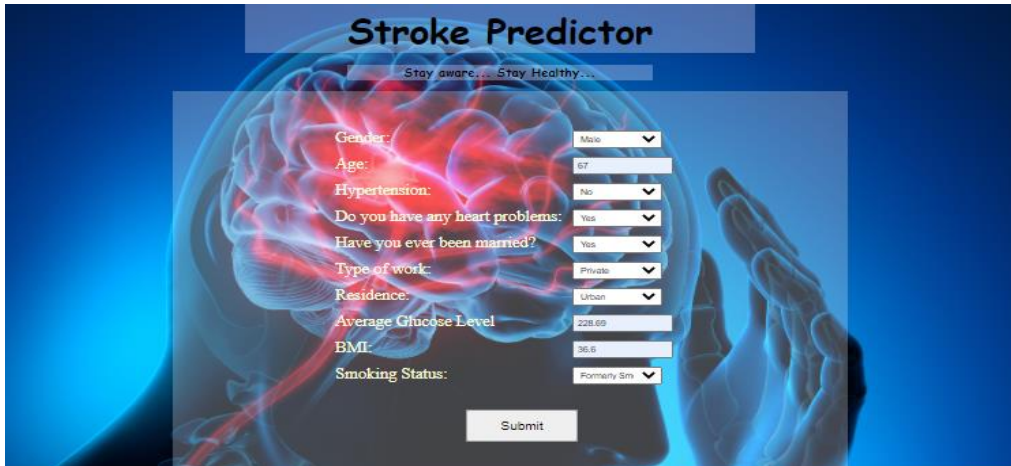
Hence, the algorithm used for this problem statement is Extreme Gradient Boosting Algorithm. It is further optimized using RandomisedSearchCV algorithm and has given an accuracy of 93.68%.

V. Deployment on Web Application

Heroku for hosting the web application.

The XGBoost model is deployed as a web application using Python's Flask framework. Live deployment is done using

VI. PORTAL FOR BRAIN STROKE PREDICTION



Stroke Predictor

Stay aware... Stay Healthy...

Gender: Male
Age: 67
Hypertension: No
Do you have any heart problems: Yes
Have you ever been married? Yes
Type of work: Private
Residence: Urban
Average Glucose Level: 228.69
BMI: 36.6
Smoking Status: Formerly Sm

Submit

Prediction: You have a chance of suffering from a stroke. Please, consult a doctor immediately.

[Click here](#) to book an appointment with Doctor

Figure No.14 - Output for 'Stroke' = True

In this GUI, a HTML form is presented to the user wherein the user can enter the required attributes for predicting whether the user is at risk of suffering from a brain stroke or not. In the above example, the output can be seen to be 'True', i.e, the user is identified to be at a risk of suffering from a stroke.



Stroke Predictor

Stay aware... Stay Healthy...

Gender: Female
Age: 58
Hypertension: Yes
Do you have any heart problems: No
Have you ever been married? Yes
Type of work: Self-Employed
Residence: Urban
Average Glucose Level: 222.74
BMI: 34.2
Smoking Status: Never Smoked

Submit

Prediction: You do not have a chance of suffering from a stroke. But, for a safer side you may consult a doctor within 15-20 days.

[Click here](#) to book an appointment with Doctor

Figure No.15 - Output for 'Stroke' = False

In the above example, the output shows that the user is not at a risk of suffering from a stroke and the output is 'False'. A link is provided in the footer which redirects the user to Apollo Hospital for a precautionary checkup nevertheless.

Website Link:

[StrokePredictor \(brain-stroke-prediction.herokuapp.com\)](http://StrokePredictor(brain-stroke-prediction.herokuapp.com))

VII. Conclusion & Future Scope

The importance of knowing and understanding the risks of brain stroke is very much in these trying times. The model predicts the probability of brain stroke on the basis of very

trivial day-to-day and known to all parameters. This makes this project highly relevant and of need to society. The objective of implementing the project on a web platform was to reach as many individuals as possible. The early warning can save someone's life who might have a probability of a stroke.

Therefore, in conclusion this project helps us predict the patients who are diagnosed with brain stroke by cleaning the dataset and applying XGBoost Model to get an accuracy of an average of 93.68%.

Outcomes:

1. The Brain stroke detection model has been tested using 5 ML classification modeling techniques. The accuracy of our selected model i.e **XGBoost Model** is **93.68%**.
2. The number of false positives in the XGBoost Model is very low which helps solve the problem of false positives in the medical domain.
3. Due to the dearth of data wherein Stroke was found to be 'True', we used 'SMOTE' algorithm to produce synthetic data points, which exacted a toll on the accuracy of our model.
4. To avoid data leakage, SMOTE was applied after splitting of dataset to the training dataset and as a result, the model encountered an imbalanced testing dataset. If more data is available, the model is anticipated to achieve better accuracy.
5. The top 3 features having the maximum impact on chances of getting a Stroke are identified as '**Age**', '**Average Glucose Level**' and '**Hypertension**'.

VIII. References

- [1] "Computer Methods and Programs in Biomedicine" - Jae-woo Lee, Hyunsun Lim, Dong-wook Kim, Soon-ae Shin, Jinkwon Kim, Bora Yoo, Kyunghee Cho.
- [2] "Stroke prediction using artificial intelligence"- M. Sheetal Singh, Prakash Choudhary. (IEEE - 2017)
- [3] "Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study" - Rohit Ghosh, Swetha Tanamala, Mustafa Biviji, Norbert G Campeau, Vasantha Kumar Venugopal.
- [4] "Prediction of stroke thrombolysis outcome using CT brain machine learning" - Paul Bentley, JebanGanesalingam, AnomaLalani, CarltonJones, KateMahady, SarahEpton, PaulRinne, PankajSharma, OmidHalse, AmrishMehta, DanielRueckert
- [5] "Probability of Stroke: A Risk Profile from the Framingham Study" - Philip A. Wolf, MD; Ralph B. D'Agostino, PhD, Albert J. Belanger, MA; and William B. Kannel, MD.
- [6] "Development of an Algorithm for Stroke Prediction: A National Health Insurance Database Study" - Min SN, Park SJ, KimDJ, Subramaniam M, Lee KS
- [7]. "Medical software user interfaces, stroke MD application design (IEEE)" -ElenaZamsa
- [8] "Focus on stroke: Predicting and preventing stroke" - Michael Regnier

[9] "Effective Analysis and Predictive Model of Stroke Disease using Classification Methods"-A.Sudha, P.Gayathri, N.Jaisankar

[10] "Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study" -Rohit Ghosh, Swetha Tanamala, Mustafa Biviji, Norbert G Campeau, Vasantha Kumar Venugopal