

Ensemble-Learning based framework for Brain Stroke Prediction

Major project report submitted in partial fulfilment of the requirement for the degree of Bachelor of Technology

in

Computer Science and Engineering

By

Mohammad Kounen Khan (181277)

UNDER THE SUPERVISION OF

Dr Aman Sharma

(Assistant Professor (SG), CSE & IT).



**Department of Computer Science & Engineering and Information
Technology**

**Jaypee University of Information Technology, Wagnaghat,
173234, Himachal Pradesh, INDIA**

Certificate

Candidate's Declaration

We hereby declare that the work presented in this report entitled “**Brain Stroke Prediction**” in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology in Computer Science and Engineering** submitted in the Department of Computer Science & Engineering and Information Technology, Jaypee University of Information Technology Waknaghat is an authentic record of our own work carried out over a period from August 2021 to December 2021 under the supervision of **Dr Aman Sharma** (Assistant Professor (SG) – CSE & IT). The matter embodied in the report has not been submitted for the award of any other degree or diploma.

Mohammad Kounen Khan (181277)

This is to certify that the above statement made by the candidate is true to the best of my knowledge.

Dr Aman Sharma

Assistant Professor (Grade-II)

Department of Computer Science and Engineering & Information Technology

ACKNOWLEDGEMENT

We accept this open door to offer our significant thanks and profound respect towards our guide Dr Aman Sharma (Assistant Professor (SG), CSE and IT) for his excellent direction, checking and consistent support over the span of this undertaking. The gift, help and direction given by him from time to time will take us far in the excursion of life on which we are going to set out.

We are additionally obliged to staff individuals from JUIT College, for the significant data given by them in their separate fields. We are thankful for their collaboration during the time of our task.

Finally, we thank all-powerful, our folks and our colleagues for their consistent support without which this task would not have been conceivable.

TABLE OF CONTENT

Content	Page No.
Declaration by Candidate	I
Certificate by Supervisor	II
Abstract	III
Introduction	6
Related work	8
Materials and Requirements	10
About the data	11
Data pre-processing	16
Proposed framework	24
Language & Technology Used	29
Requirements	32
Classification of Model	33
Proposed methodology	35
Experiment evaluation & Results	38
comparison with existing models	39
Application & Limitation	41
Future work	43
Conclusion	44
References	45

ABSTRACT

In our project entitled “Brain Stroke Prediction”, The main aim is basically to build a system to predict whether the patient will get a stroke or not according to the given data and past medical records as well. 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. The Dataset Includes Unique Id, gender, age, hypertension, history of heart disease, marital status- whether they are married or unmarried, type of work- what kind of work they do whether fieldwork, office work etc., residence type- whether they live in a colony or flat, or own home or on rent etc., average glucose level, BMI, smoking status- whether they smoke or not, and stroke.

Chapter 01: INTRODUCTION

Introduction

Machine learning (ML), a subset of AI that focuses on developing algorithms that decide how to generate predictions based on data, is becoming more popular in the field of bioinformatics. Bioinformatics is a computer science subject that analyzes and manipulates natural data with computational tools and numerical tools. Prior to the introduction of AI calculations, bioinformatics calculations had to be specifically designed by hand, for example, brain stroke prediction, which proved to be quite difficult. Machine learning methods and algorithms such as regression models, ensemble models etc. enable the calculation to be programmed including realizing, which means the calculation can figure out how to consolidate different highlights of the information into a more dynamic arrangement of highlights from which to lead to further learning based solely on the dataset. This diverse technique to manage learning plans in the information permits such systems to make extremely muddled figures when applied to tremendous datasets. The quantity and number of normal datasets accessible have recently increased, allowing bioinformatics specialists to apply these Machine learning algorithms. Six normal areas have been subjected to AI. Artificial intelligence (AI) systems that analyze neuroimaging data are being utilized to aid in the investigation of strokes. Because of a lack of analytic devices, as well as a dearth of experts and a variety of other assets that affect adequate expectation and prescription of heart patients, analysis and treatment of stroke infection are extremely difficult in emerging nations. PC innovation and AI procedures have recently been more prevalent in this concern, with the goal of improving the framework to assist experts in the early stages of making decisions concerning the disease. Furthermore, the treatment costs can be significantly reduced as a result of this. Embracing protected, reasonable methods and using current innovation can decrease the requirement for guardians while additionally bringing down by and large medical services costs. A few lives could be saved if shrewd dynamic techniques and advancements were developed [1]. Stroke is the third driving reason for death and the chief reason for genuine long haul incapacity in the United States. A precise forecast of stroke is profoundly significant for early mediation and treatment [2]. AI (ML) is perhaps the most generally involved strategy for rapidly preparing machines and creating prescient models for better direction. By examining the patient's condition, and auditing past clinical records of a patient, AI helps with the early discovery of mind stroke and decides its tendency of it. AI strategies are the most well-known techniques for accomplishing great outcomes in arrangement and expectation issues. Early detection can greatly improve the prognosis. The main goal of this paper is to propose an ensemble learning-based framework for brain stroke prediction. This paper examines existing stroke prediction models in depth and reports on the highly accurate and efficient results.

About Stroke:

A stroke happens when the arteries in our blood that transport oxygen and nutrients to the brain become blocked or break. When this occurs, a region of the brain is deprived of blood as well as brain cells. Eighty per cent of strokes are preventable. A stroke can be caused by any of the following factors: a blood clot in the brain that prevents blood flow. The medical name for this is an ischemic stroke. When a blood artery bursts, it cuts off blood flow to the brain, resulting in a hemorrhagic stroke. A mini-stroke is possible when a blood clot causes a TIA (transient ischemic attack). the human brain is a very complex and sophisticated organ that is used to control our body, our actions and physical functions. When the blood supply is being disturbed the certain function of our brain does not operate properly.

“Brain Stroke Prediction”, The main aim is basically to create a system which gives us the chances/prediction of whether the patient will get a stroke or not according to available data and past medical records as well. The Dataset Includes Unique Id, patient’s gender, age of the patient, hypertension activity, history of heart-related issues (if any), ever married- whether they are married or unmarried, type of work- what kind of work they do whether fieldwork, office work etc., residence type- whether they live in a colony or flat, or own home or on rent etc., average glucose level, BMI, smoking status- whether they smoke or not, and stroke.

Related Work

[1] This research proposes a crossover strategy that combines missing value imputation with an auto Hpo-based DNN prediction model to enhance medical prediction based on physiological signs of possible stroke patients. Our model yielded a false negative rate of only 19.1 percent and an overall accuracy of 71.6 percent. The false negative rate decreases by 51.5 percent and the total error increases by 1.7 percent when compared to the mean discoveries of other regularly utilized techniques. As a result of these adjustments, our method may significantly lower the false negative rate without sacrificing overall accuracy. As a result, the crossbreed machine learning technique utilized in this work for stroke prediction is effective and reliable. Furthermore, this method can dynamically optimize the hyperparameter without the need for manual selection, and it takes into account multi-factor correlation, which is more sophisticated than the single-factor analysis often employed in traditional medicine.

[2] The conservative mean feature selection performs exceptionally well for the CHS dataset, as demonstrated in this study. However, because it examines the effectiveness of each feature separately, we recognize that our feature selection approach may not function well in other datasets with strongly correlated features.

To solve this problem, we might prune the features using an L1 regularized feature selection approach (e.g., L1 regularized logistic regression) before fine-tuning using conservative mean feature selection.

[3] The proposed research work has employed ten classifiers to find out the performance of stroke occurrence in a person. To predict stroke, the recommended weighted voting classifier took into account orientation, age, hypertension, heart disease, average glucose level, BMI, and smoking status feature characteristics. In comparison to the regularly used other machine learning algorithms, weighted voting delivered the highest accuracy of around 97%, according to the performance evaluation.

[4] The notion of feature selection is based on an algorithm. The dimension is reduced using the component analysis technique. Backpropagation neural network classification was used. To build a classification model, you'll need an algorithm. After doing some research, by comparing the classification efficiency of various approaches. Our work is the most accurate in terms of various models and accuracy. A predictive model for the stroke disease with 97.7% accuracy. For feature selection, the proposed technique employs the Decision Tree algorithm, Pca for dimension reduction, and ANN for classification. The experimental discoveries recommend that the proposed strategy outperforms other well-known strategies in similar situations

[5] Stroke is a serious medical ailment that should be addressed as soon as possible. The development of a machine learning model can aid in the early detection of stroke and diminish the seriousness of future consequences. The efficacy of several machine learning algorithms incorrectly predicting stroke based on multiple physiological parameters is demonstrated in this research. With an accuracy of 82 per cent, the Naive Bayes Classification method outperforms the others. Nave Bayes outperformed all other models in terms of accuracy, recall, and F1 score.

[6] With the rising number of people dying from heart attacks, it's become critical to design a system that can efficiently and precisely forecast heart attacks. The goal of the research was to discover the best effective machine learning system for detecting cardiac stroke. Utilizing the Kaggle dataset, this study analyses the accuracy scores of Random Forest, Decision Tree, and KNN algorithms for predicting cardiac stroke. According to the discoveries of this review, the Random Forest algorithm is the most effective algorithm for predicting heart stroke, with an accuracy score of 99.17 per cent.

[7] This study demonstrated that using Data Science and Machine Learning algorithms, it was feasible to predict the result of a stroke based on existing information about the individual. In addition, the CRISP-DM approach served as a guide through the data analysis, making the process considerably easier and efficient without losing sight of the business challenge and making the appropriate decisions based on it.

[8] A sufficiently big dataset of stroke victims has been properly categorised. For stroke illness identification, Naive Bayes, J48, k-NN, and Random Forest were utilised. We can observe from the performance study that Naive Bayes outperforms other approaches. Collecting this information and preparing it for usage with WEKA is the innovation and key contribution of our study. The model can assist those who have received a warning sign that they may be suffering from a stroke. Even a field expert finds it challenging to link vast volumes of complicated data regarding patients, hospitals, illness diagnosis, electronic patient records, medical equipment, and other topics in the healthcare industry. It will assist the doctor in gaining a better understanding of the condition. The dataset is not entirely symmetrical, which is one of our method's constraints.

[9] In comparison to the existing approach, the suggested model improves the accuracy of IgG and IgA antibody prediction. To develop a multilevel ensemble model in this work, seven models were used: decision tree, ELM, RF, neural network, SVM, Avnet, and RRF. With variable-length epitopes, a unique multilevel ensemble model is created for prediction, and it delivers good accuracy, Gini, AUC, specificity, and sensitivity. There are three steps to the multilevel ensemble model. True and false predictions are employed in this strategy to obtain an accurate recommended model. The advantage of utilising correct prediction as an input to other models is that false-positive outcomes may be avoided. The data is sent through seven models, each of which learns the data correctly to produce dependable and accurate results.

Chapter 02

MATERIALS & REQUIREMENTS

Data Collection

The practice/method of collecting, acquiring and analyzing data from a variety of sources is known as data collection. Data should be collected and kept in a form that makes sense for the business challenge at hand in order to use it to assemble viable artificial intelligence (AI) and machine learning solutions .

The speed of decision-making is substantially increased when judgments are based on data and facts. The process of taking decisions becomes rapid and dependable as we can make confident conclusions using real-time data and prior data trends.

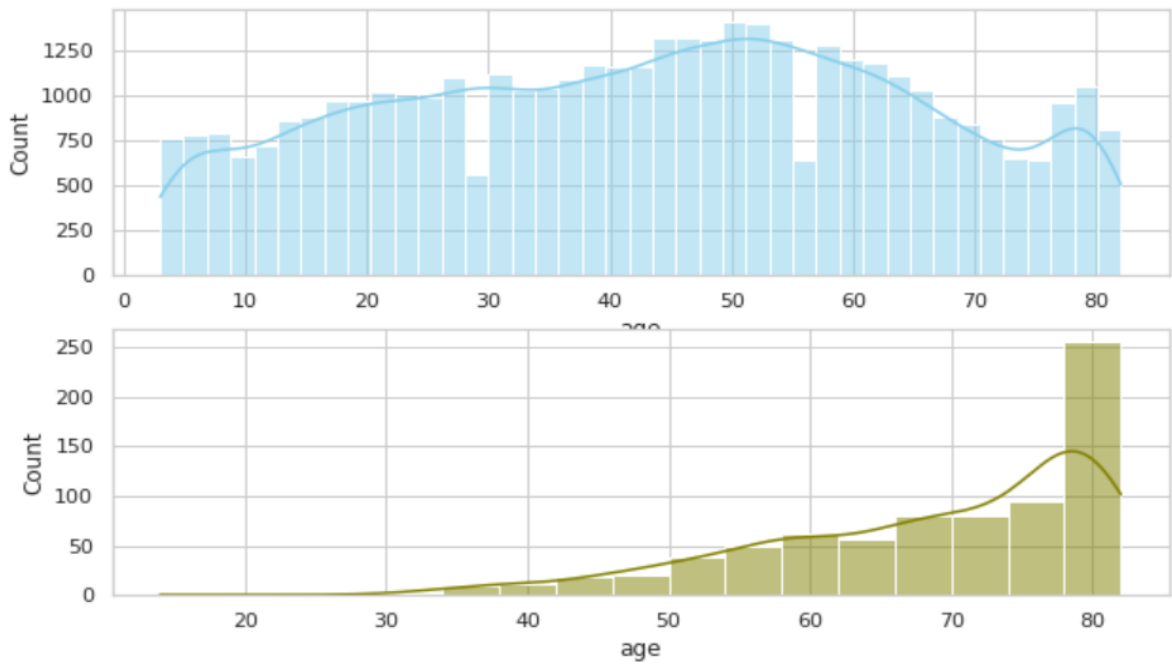
In this Project following data is being used to generate the desired results:

1. Unique Id
2. Gender
3. Age
4. Hypertension
5. history of heart disease
6. marital status
7. Work Type
8. residence type
9. average glucose level
10. BMI
11. smoking status
12. stroke

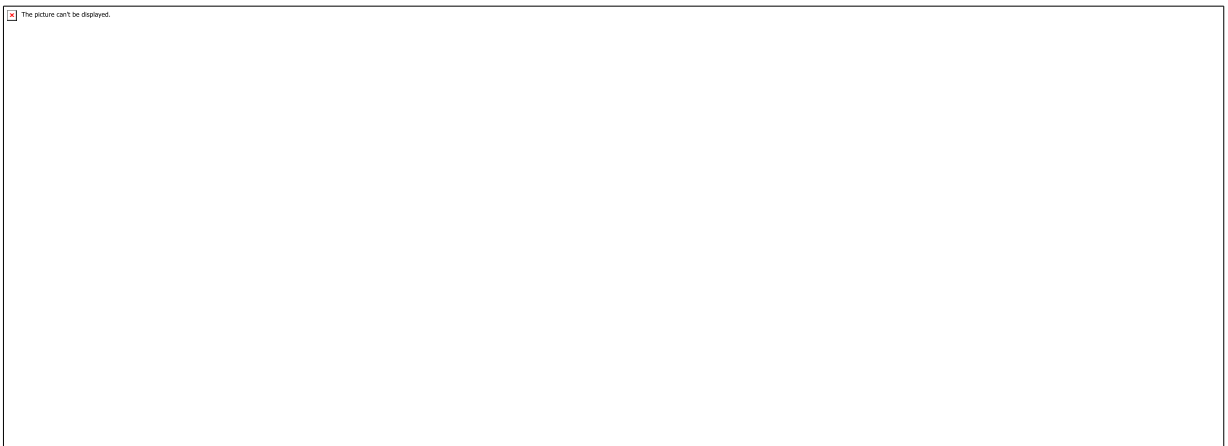
About the Data

Here are the things that affects stroke in the patient:

Gender: Women are generally have greater impact of stroke than males because they experience more occurrences and are less likely to recover. Men have greater age-specific stroke rates than women, although women experience more stroke events than men due to their longer life expectancy and significantly higher occurrence at older ages. There is no evidence of sex differences in stroke subtype or severity, with the exception of subarachnoid haemorrhage. Although some studies have revealed that women are less likely than males to get certain in-hospital therapies, most of the discrepancies vanish after age and comorbidities are taken into consideration.

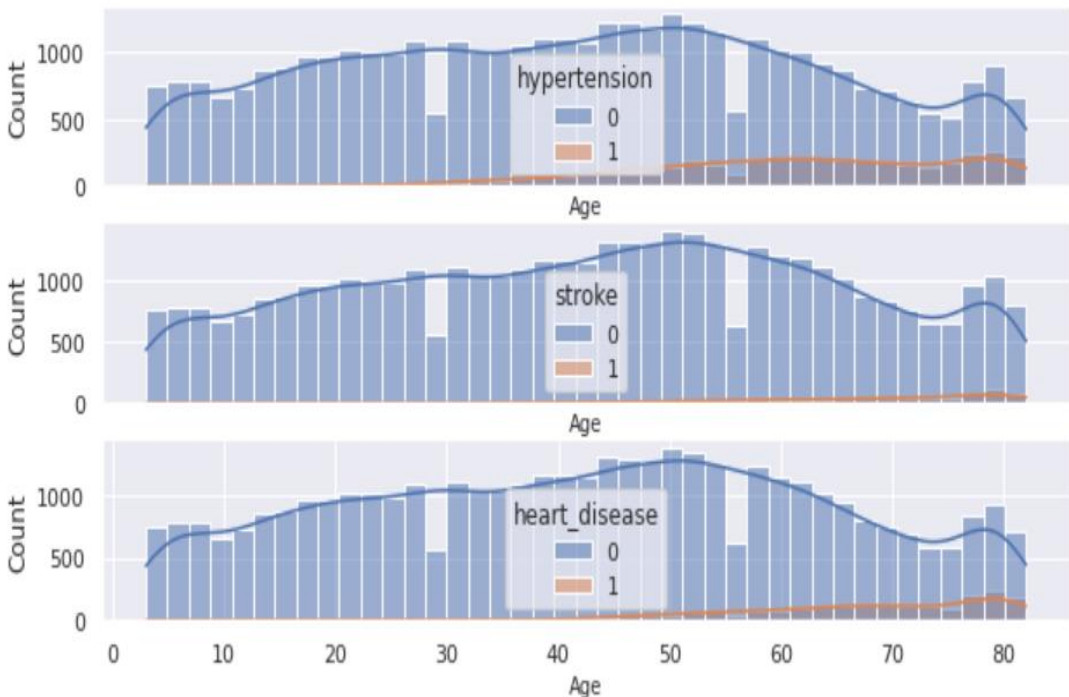


Age: Clinical stroke is characterized as rapidly arising indications of focal neurologic disruption caused by a vascular source that lasts longer than 24 hours. The gamble of stroke grows with age, with the incidence doubling each decade beyond the age of 45, and over 70% of all strokes occurring after the age of 65. According to research published in the journal Stroke in February 2020, between 10% to 15% of strokes occur in adults aged 18 to 50.

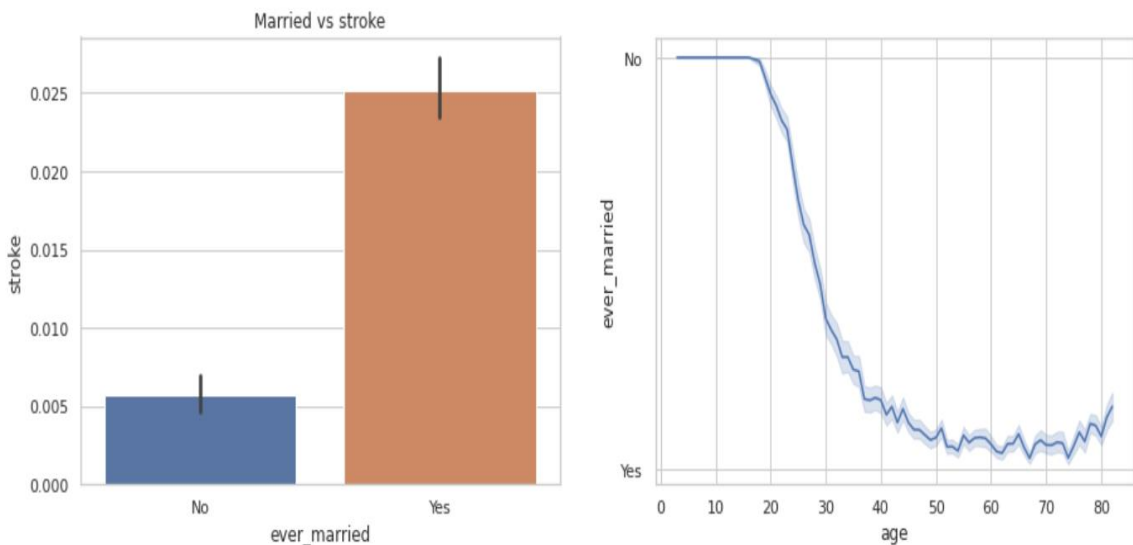


Hypertension: High blood pressure damages blood vessels, causing them to constrict, burst, or leak. Blood clots can develop in the arteries leading to your brain, obstructing blood flow and potentially causing a stroke if your blood pressure is too high.

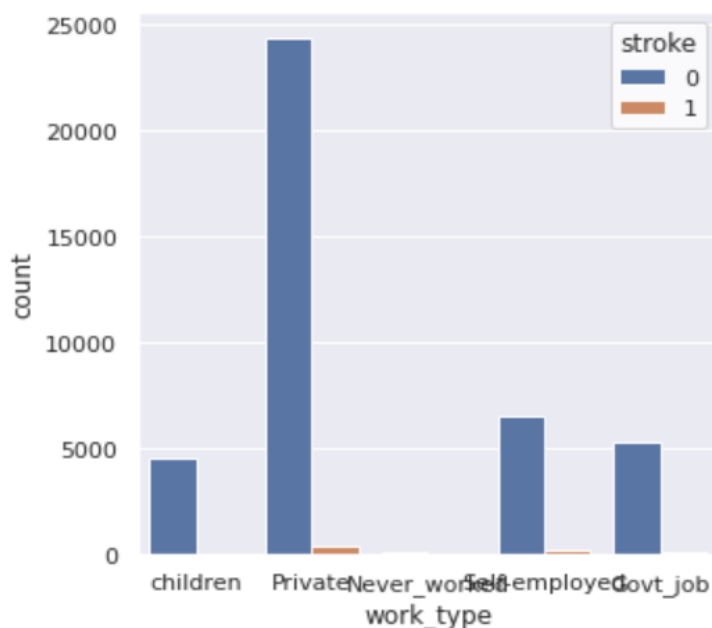
Heart Disease: The gradual occlusion of blood arteries due to the accumulation of fatty substances and cellular debris (plaques). Angina or a heart attack can be caused by plaque buildup in the arteries delivering blood to your heart muscle. A stroke is caused by plaque buildup and blood clots in the arteries delivering blood to the brain.



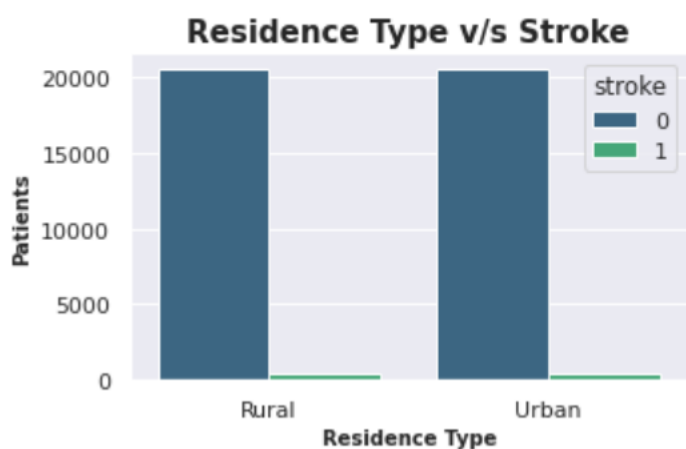
Marital Status: According to a recent study, social isolation, rather than marital status, was linked to all-cause mortality following a stroke. Only 33% of the 655 stroke patients in the research were married, making it a small study. Furthermore, patients with a pre-stroke impairment who received home care were more likely to be labelled as socially isolated, which might lead to misclassification bias.



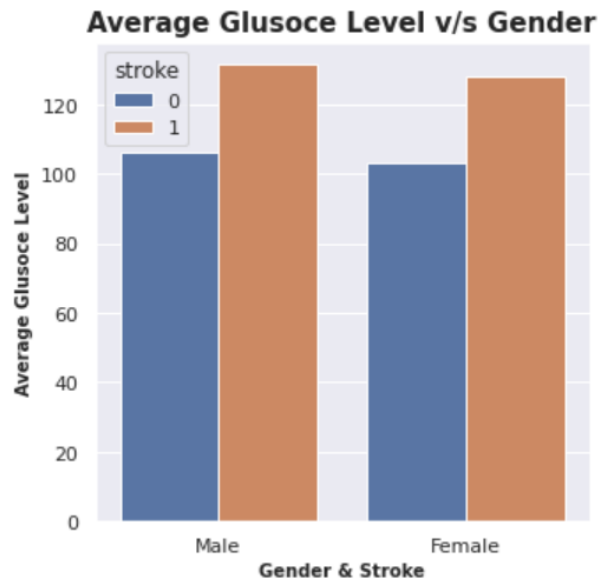
Work Type: People who worked in high-stress positions had a 22% greater risk of stroke than those who worked in low-stress professions, according to the study. Women who worked in high-stress employment had a 33% greater risk of stroke than women who worked in low-stress positions. People who worked in high-stress employment were 58 per cent more likely than those who worked in low-stress positions to suffer an ischemic stroke.



Residence Type: Risk factors were more prevalent yet more uncertain to be controlled in rural than in urban inhabitants without prior stroke in a population-based study of over 6 million people with universal access to physician and hospital services, whereas risk factor prevalence and treatment were similar in those with prior stroke. Even after controlling for risk variables, the rural location was linked to a higher risk of stroke and mortality.

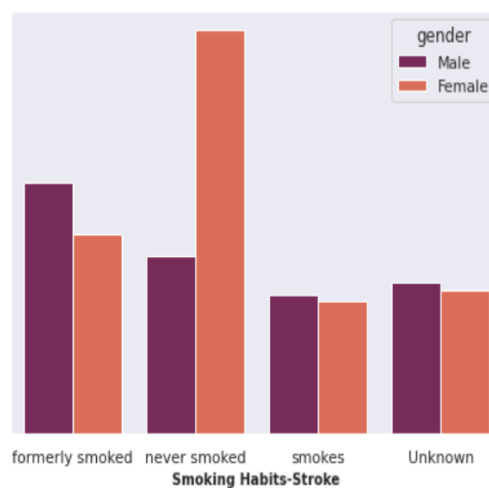
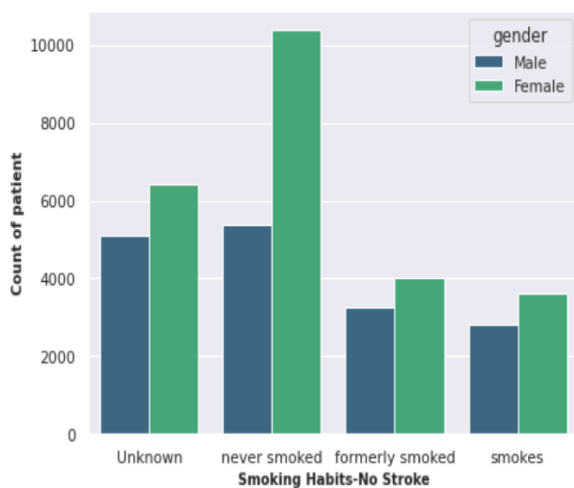


Average Glucose Level: In acute type of stroke, the blood glucose levels are frequently been raised, and higher incoming/admission glucose levels are linked to bigger lesions, higher mortality, and a worse functional prognosis. Hyperglycemia is linked to an increased risk of infarct hemorrhagic transformation in individuals receiving thrombolysis. Hyperglycemia, characterized as a blood glucose level of more than 6.0 mmol/L (108 mg/dL), was found in two-thirds of all ischemic stroke subtypes on admission, and in at least half of each subtype, including lacunar strokes.



BMI: A stroke affects around 75% of adults aged 65 and over. However, studies suggest that having a higher BMI at any age increases your risk of stroke. Obesity increases high blood pressure, which is one of the primary causes of stroke, according to medical researchers.

Smoking Status: Nicotine makes your heart beat quicker and boosts your blood pressure, while carbon monoxide limits the quantity of oxygen in your blood. This raises your chances of having a stroke. Smoking can also cause atrial fibrillation, a heart disorder that increases the gamble of stroke..



Data Pre-Processing

In ML, getting ready information is an essential advance that assists us with working on the nature of information and works with the most common way of separating helpful data from information Data preprocessing in Machine Learning alludes to the most common way of cleaning and putting together crude information so it very well might be utilized to make and prepare Machine Learning models. Information preprocessing is an information mining method utilized in Machine Learning that changes over crude information into a decipherable and justifiable configuration. Prior to fostering a model, information preprocessing is fundamental to take out the unfortunate commotion and exceptions from the dataset, which could cause a difference from typical preparation. This progression deals with all that keeps the model from proceeding as effectively as could be expected. Subsequent to the social event the fundamental information, the accompanying advance is to clean it and guarantee that it is prepared for model development.

Missing Value Treatment:

Missing Value Treatment is critical because the data insights or performance of your predictive model may be harmed if missing values are not handled properly.

We have used the following ways for data pre-processing and treating missing values:

- I) **KNN Imputation (k-Nearest Neighbor Imputation):** Regression models may be used to forecast if the input variables are numeric, which is a typical scenario. "Nearest Neighbor Imputation," or "KNN imputation," is the process of using a KNN model to forecast or fill in missing data.
- II) **Simple Imputer:** Simple Imputer is a scikit-learn class that may aid with missing data in predictive model datasets. It's used to impute/replace missing numerical or categorical data for one or more characteristics with acceptable values.
- III) **The mean value Imputer:** In mean value imputation we treat missing values by replacing it with by either mean, median or mode of the given dataset to minimize the error and get the desired results.

By using the mean value Imputer, we got the best performance.

Data Binning:

Data binning, also known as bucketing, is a data pre-processing method for reducing the impact of minor observation mistakes. The original data values that fall inside a particular narrow interval, called a receptacle, are replaced by a value representative of that interval, which is usually the centre value.

Binning statistical data is a method of dividing a large number of more or less continuous values into a smaller number of "bins."

We have binned a few Numerical variables into small bins.

Label Encoding:

We replace the categorical value with a numeric value between 0 and the number of classes minus 1 in label encoding. We utilize if the category variable value has five unique classes (0, 1, 2, 3, and 4).

Label encoder can be utilized to normalize labels.

It may also be utilized to convert non-numerical labels to numerical labels (as long as they are hashable and comparable).

We dealt with categorical variables in our dataset using Label encoding.

Balancing Dataset:

Every output class (or goal class) is represented by the same number of input samples in a balanced dataset. Over-sampling or under-sampling are two forms/strategies that can be used to achieve the balancing.

We have used SMOTE Library for balancing the data.

The synthetic Minority Oversampling Approach (SMOTE) is an oversampling technique that balances a dataset by creating synthetic minority class data points.

To generate synthetic data points, SMOTE employs a k-nearest neighbour technique. The SMOTE algorithm has the following steps:

1. Finding the vector of the minority class.
2. Determining the number of closest numbers (k) to take into account.
3. Create a synthetic point by drawing a line between the minority data points and any of their neighbors.
4. Step 3 should be repeated until the data is balanced for all minority data points and their k neighbors.

Chapter 03

LITERATURE SURVEY

Pipeline:

A ML pipeline is a way of codifying and automating the creation of a machine learning model. Data extraction, preprocessing, model training, and deployment are all handled by machine learning pipelines.

Standard scalers:

The main logic behind Standard Scaler is that it will turn your data into distribution with a mean of 0 and a standard deviation of 1.

This is done feature-by-feature in the case of multivariate data.

Given the data distribution, each value in the dataset will be subtracted from the mean and then divided by the overall dataset's standard deviation (or feature in the multivariate case).

Remove the mean and scale to unit variance to standardize characteristics.

A sample x 's standard score is determined as follows:

$$(x - u) / s = z$$

where u is the training samples' mean, or zero if with `mean=False`, and s is the training samples' standard deviation, or one if with `std=False`.

By computing the necessary statistics on the samples in the training set, each feature is individually centred and scaled. The mean and standard deviation are then saved and used to convert later data.

Logistic Regression

logistic regression (LR), it is a type of classification model rather than a regression. Strategic relapse is a basic and successful methodology for double and direct order issues. It's a clear characterization model that produces extraordinary outcomes with straightly distinct classes.

It is an as often as possible utilized ordering approach in the modern world. A calculated relapse model, as Adaline and perceptron, is a factual strategy for double grouping that might be applied to multiclass order. It's a strategy for assessing the probability of a discrete result given an assortment of information factors.

The most well-known strategic relapse models give a double result, like valid or misleading, yes or no, etc. Demonstrating situations with multiple discrete results with multinomial strategic relapse is conceivable.

The Mathematical expression of Logistic Regression is given by:

$$P = \frac{e^{a+bX}}{1 + e^{a+bX}}$$

Where P is the Probability.

a & b are parameters of the model.

Decision Tree

Decision tree classifiers have numerous applications. Their capacity to catch elucidating decision-production data from the information given is their most fundamental characteristic. Preparing sets can be utilized to produce decision trees. A straightforward model for ordering models is a decision tree. It is administered by AI, where the information is disintegrated and persistently founded on certain boundaries. The methodology for deciding the class of a given informational index in a decision tree begins at the root hub of the tree. This calculation checks the upsides of the first quality against the upsides of the record trait (the genuine informational collection), then, at that point, follows the branch and moves to the following hub in light of the correlation. The calculation contrasts the property estimation and other kid hubs and continues on toward the following hub. It rehashes the entire interaction until it arrives at the leaf hub of the tree.

In the decision Tree, we use the following mathematical formulas:

1. Entropy: The amount of some set of information from the given data is required to describe accurately is known as Entropy.

$$Entropy = - \sum_{i=1}^n p_i * \log(p_i)$$

2. Gini Index: It is the measure of the Inequality of the given data. The value ranges between 0 and 1. If the value of the Gini Index is 0 that means the given data is perfectly homogeneous and if it's 1 that is it have maximal inequality between the elements.

$$Gini\ index = 1 - \sum_{i=1}^n p_i^2$$

Where i is the number of classes.

Algorithm:

Given a preparation set $T=(a_i, b_i)/b_i=0$ or $b_i=1, i \in [1, N]$ with the end goal that $a_i \in R^n$, the errand of the choice tree is to recursively segment the element space to such an extent that examples having a similar mark are assembled together. May there be D information at every hub p .

The competitor split $\phi=(j, k_p)$ is utilized to divide the information into subsets $D_{left}(\phi)$ and $D_{right}(\phi)$. Where j characterizes the quantity of highlights and k_p is the limit.

$$D_{left}(\phi) = (a, b)/a_j \leq k_p \quad (1)$$

$$D_{right}(\phi) = D/D_{left}(\phi) \quad (2)$$

The mistake at every hub p is determined utilizing capacity $E()$. The capacity definition can change contingent upon the errand (Classification/Regression) of the choice tree.

$G(D, \phi) = n_{left} N_p E(D_{left}(\phi)) + n_{right} N_p E(D_{right}(\phi)) \quad (3)$ Our goal is to choose ideal boundaries which limit the mistake.

$$D^* = \operatorname{argmin}_{\phi} G(D, \phi) \quad (4)$$

Recursively, rehash the above advance for every subset $D_{left}(\varphi^*)$ and $D_{right}(\varphi^*)$ for the greatest profundity of the tree, $N_p < \min - \text{tests}$ or $N_p = 1$

Table 1: Symbols used in Algorithm

S. No.	Symbols	
1.	kn	Used to split the data
2.	kp	threshold
3.	E()	Function
4.	D _{left}	Left subset of data
5.	D _{right}	Right subset of data
6.	D	Number of attributes in the dataset at each node
7.	j	Number of features
8.	φ	Function variable

Ensemble Learning

Ensemble is the art of bringing together a diverse group of learners (individual models) to improve the model's stability and predictive power. Ensemble Learning is the process of combining all of the predictions. The ensemble members are models that learn and contribute to the ensemble. They may be of the same kind or different types, and they may or may not have been trained on the same training data that is available. The ensemble members' predictions can be aggregated using statistics like the mode or mean, or more advanced approaches that learn how much and under what conditions to trust each member.

Based on a sequence of questions and situations, a Decision Tree evaluates the prediction value. The tree considers numerous weather parameters and makes a choice or asks another question based on each element. With the same framework, Decision Trees may also handle quantitative issues.

Commonly used ensembling techniques are-

1. Bagging: Bagging tries to implement comparable learners on tiny sample populations and then averages the results. You can employ different learners on various populations in generalized bagging. As you may assume, this aids in the reduction of variance error.

2. Boosting: Boosting is a method of iterative strategy that is repeating the method for adjusting an observation's weight that relies on the previous categorization. It seeks to raise the weight of observation if it was classified erroneously, and vice versa. Boosting reduces bias error and produces good prediction models in general. They may, however, overfit the training data on occasion.

3. Stacking: This is an intriguing method of mixing models. A learner is used to integrate the output of multiple learners. Depending on the combining learner we select, this can result in a reduction in either bias or variance error.

Voting Classifier: A voting ensemble is a machine learning model that combines predictions from many models

into one. It's a technique for improving model performance with the objective of exceeding any individual model in the ensemble.

In a voting ensemble, the predictions from many models are integrated. It can be used for classification or regression. This implies calculating the average of the model's predictions in the case of regression. When categorizing, the predictions for each label are tallied together, and the label with the most votes is picked.

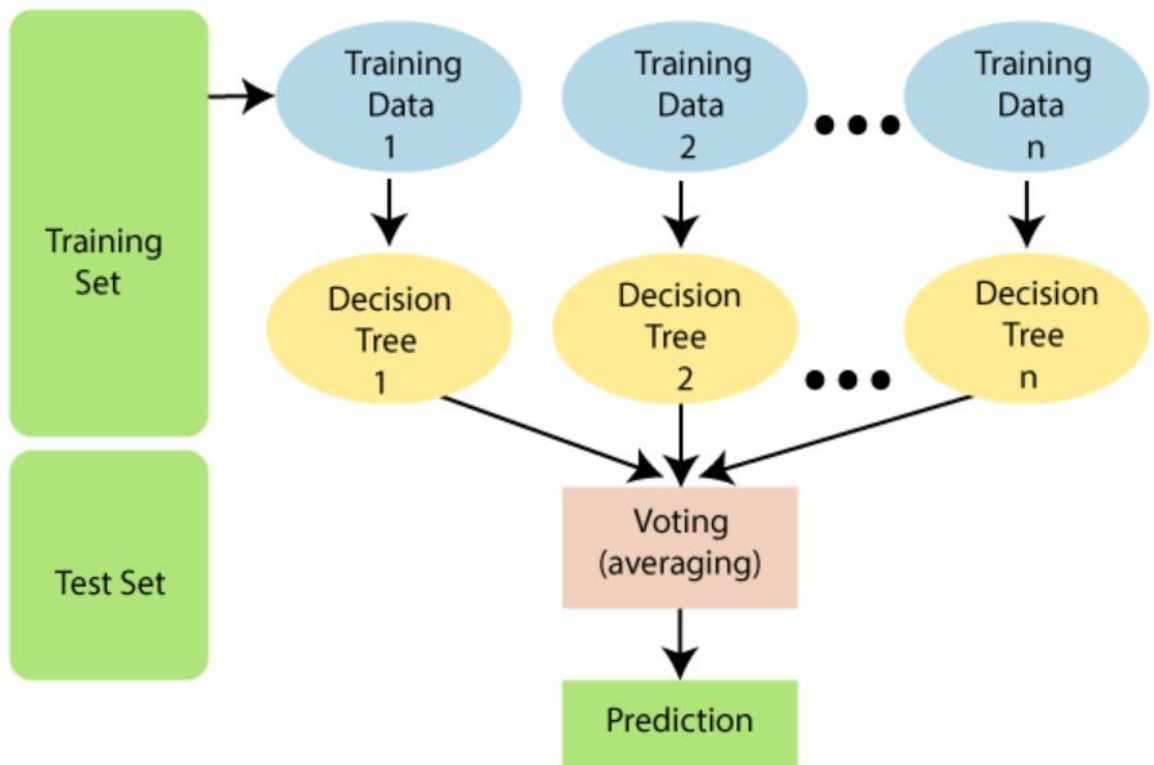
Random Forest

The decision tree is the groundwork of arbitrary backwood classifiers. A decision tree is a progressive design worked from the properties of an information assortment (or free factors). An action combined with a subset of the elements separates the decision tree into hubs. The arbitrary woods is an assortment of decision trees that are connected to an assortment of bootstrap tests produced from the first informational index. To parcel the hubs, the entropy (or Gini list) of a subset of the properties is used. The bootstrapped subsets of the first informational collection have a similar size to the first informational collection. Breiman's articles on irregular wood classifiers are significant (Breiman, 1996, 2001). According to Suthaharan, the bootstrapping method works within the development of arbitrary timberlands with the required number of decision trees to increment arrangement exactness through the idea of cross-over diminishing (2015). The best trees are then picked by a democratic cycle and a bagging method (bootstrap total). This average arbitrary timberland method is utilized in the proposed mental registering design.

The bootstrapped subsets of the first informational collection have a similar size to the first informational collection. Breiman's articles on irregular wood classifiers are significant (Breiman, 1996, 2001). According to Suthaharan, the bootstrapping method works within the development of arbitrary timberlands with the required number of decision trees to increment arrangement exactness through the idea of cross-over diminishing (2015). The best trees are then picked by a democratic cycle and a bagging method (bootstrap total). This average arbitrary timberland method is utilized in the proposed mental registering design.

The bigger is the number of trees in the forest, the more precise and accurate it is and the problem of overfitting is avoided.

Following is the diagram to explain the functionality of Random Forest:



ref. [Machine Learning Random Forest Algorithm - Javatpoint](#)

Chapter 4 - Proposed Framework:

In this section, we explained our model selection criteria and parameter setting of different algorithms used in building the framework. The experimental setup & proposed methodology has been explained in detail.

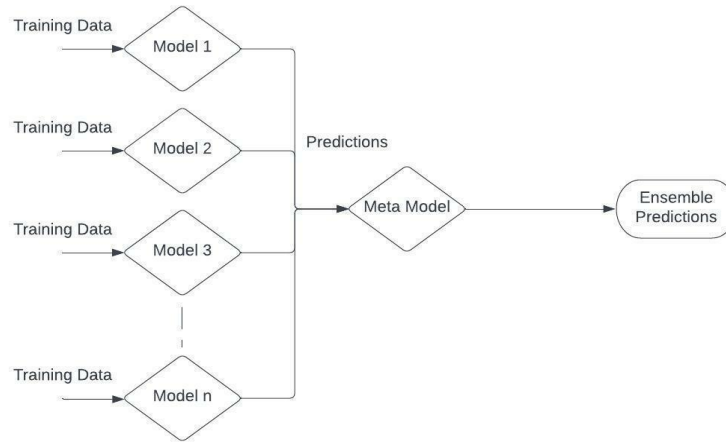


Figure 2: The proposed ensemble framework for brain stroke prediction.

4.1 Model Selection

Figure 2 comprehensively shows a proposed framework for brain stroke prediction. Firstly, we have taken Data from Mendeley [11]. After deeply analysing, and cleaning the data by finding correlations among other attributes we have removed outliers. Then split our data into two partitions: Training Dataset & Testing Dataset.

After we build our model, we performed an ensembling of algorithms and formed five different ensemble models to predict the stroke from our framework to find the optimal combination. After which we input our pre-processed data into the base learners to generate ensemble predictions. According to the exhibition of various standard models on cross-approval precision, we select the best performing models for the stacked gathering so their group will create better execution in contrast with individual AI models. At long last, we contrast our outcomes and the base students and different existing ongoing methodologies for coronary illness expectations.

4.2 Parameter Setting:

In this section, we have defined diverse parameters used to increase accuracy across our Ensemble model. We have tried a few combinations of ensemble learning along with the data pre-processing tools. We achieved an average accuracy of 96.01% and an F-1 score of 95.99 when we used Stratified K-fold for cross-validation.

4.3 Experimental Setup

4.3.1 Data set

The Mendeley dataset was used to retrieve the brain stroke dataset [11]. The dataset was compiled and cleaned by different data pre-processing methods and outliers were treated. The dataset comprises 12 features of 45000+ instances. The attributes in the dataset are id(in numeric), gender, age(in years), hypertension, a record of heart disease, marital status, work type, residence type, average glucose level, BMI level, smoking status, and stroke. The description of a few attributes is shown in Table 3.

Table 2: Description of few Attributes

Attribute	Description
Gender	Gender of patients (Male = male, Female= female)
hypertension	1= Yes, 0= No
Heart disease	1= Yes, 0= No
Marital status (ever married or not)	Yes=yes, No=no
Residence type	Rural, Urban
Work type	Private, Never_worked, children
Smoking status	Formerly smoked, Unknown, never smoked, smokes.

Training Dataset

The first data needed to train machine learning models is known as training data (or a training dataset). Machine learning algorithms are taught how to make predictions or perform a task using training datasets .

If you're developing a sentiment analysis model, keep these tips in mind (that analyses text for opinion polarity: positive, negative, and neutral).

Whether you're utilizing supervised or unsupervised learning, the data for AI training will differ .

Unsupervised learning makes use of data that hasn't been tagged. To form inferences and reach conclusions, models must detect patterns (or similarities and variances) in the data .

In supervised learning, people must tag, label, or annotate data according to their criteria in order to train the model to get the intended result (output) .

Machine learning techniques allow machines to solve issues based on previous observations. Machine learning models have the advantage of improving over time when they are exposed to more relevant training data.

Let's break down the data training procedure into three simple steps:

1. Provide training data to a machine learning model.
2. Assign a desired output to the training data. The training data is transformed into text vectors, which are

integers that represent data characteristics.

3. Put your model to the test by giving it test (or unknown) data. On the basis of manually tagged samples, algorithms are trained to correlate feature vectors with tags, and subsequently learn to generate predictions when processing unseen data.

Five Characteristics of Good Training Data

1. **Relevant:** You'll need data that are relevant to the work at hand or the problem you're attempting to address, of course. If you want to automate customer support operations, you'll need to employ a dataset of real customer support data, or the results will be biased. You'll need a dataset from Twitter, Facebook, Instagram, or whichever social media site you'll be examining if you're training a model to evaluate social media data.
2. **Uniform:** The data should all originate from the same source and have the same properties.
3. **Representative:** The data points and variables in your training data must match those in the data you'll be evaluating.
4. **Comprehensive:** Your training dataset should be large enough to meet your goals and have the right breadth and range to cover all of the model's intended use cases.
5. **Diverse:** Otherwise, the findings would be biased since the dataset does not reflect the training and user community. Check for hidden biases among persons in charge of training the model, or hire a third party to examine the criteria.

F1- Score:

The F-score, also known as the F1-score, is a metric for how accurate a model is on a given dataset. It's used to assess binary classification systems that divide examples into 'positive' and 'negative' categories.

The F-score, which is defined as the harmonic mean of the model's accuracy and recall, is a technique of combining the model's precision and recall.

The F-score is a popular metric for assessing information retrieval systems like search engines, as well as a variety of machine learning models, particularly in natural language processing.

It's possible to tweak the F-score such that accuracy takes precedence over recall, or vice versa. The F0.5-score and the F2-score, as well as the conventional F1-score, are common modified F-scores.

Formula of F1 score is:

$$F_1 = \frac{2}{\frac{1}{\text{recall}} \times \frac{1}{\text{precision}}} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

$$= \frac{\text{tp}}{\text{tp} + \frac{1}{2}(\text{fp} + \text{fn})}$$

Where, Tp = number of true positives,
 Fp = number of false positives &
 Fn = number of false negatives.

4.3.3 Data visualization & Correlation of data attributes

Table 3: Correlation with Target Diagnosis (stroke)

age	0.16
hypertension	0.075
Heart disease	0.11
Average glucose level	0.079
BMI	0.018

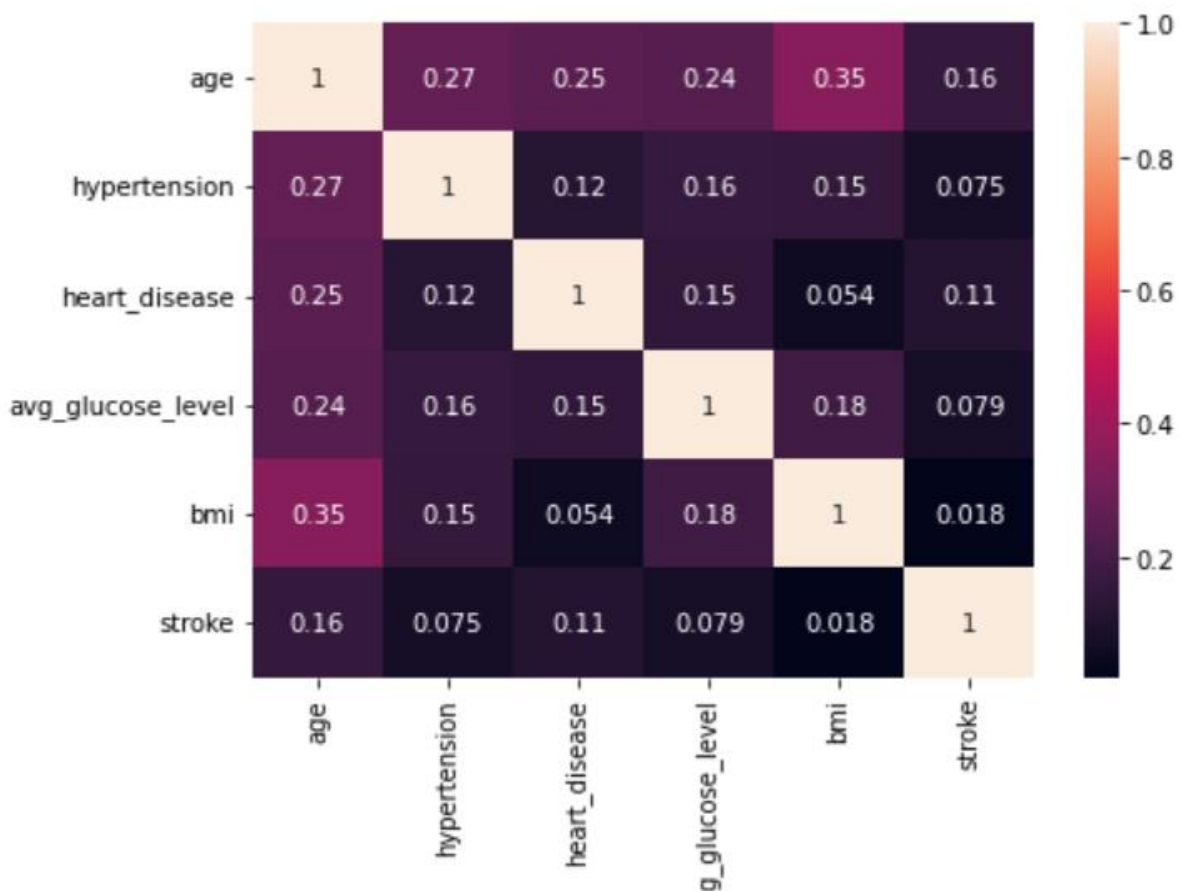


Figure 1: Cross-correlation values through Heat map

age: This represents the age of a patient. It is represented in numerical data.

gender: This represents how the patient identifies himself. It's a categorical kind of data.

hypertension: it represents whether the patient is hypertensive or not. It is numerical data.

work type: it means what kind of scenario in which the patient works (eg: private, government etc.). It's categorical data.

residence type: it means where the patient lives. It comes under categorical data.

heart disease: this tells us whether a patient has a heart attack history or not. It's a kind of numerical data.

Average glucose level: The average glucose level of a patient is represented by this variable. It's numerical data.

BMI: this variable represents the body to mass ratio or index of a patient. It is numerical data.

ever married: it represents the marital status of the patient i.e ever married or not. It's categorical data.

smoking Status: it tells us about the smoking habit or pattern of the patient. It is categorical data.

stroke: This represents whether a patient had a stroke before or not. It is numerical data.

In this, stroke is the decision class attribute and the rest of the remaining attribute is the response class which helps us to identify the stroke in a patient.

Language Used

Python:

Python is a undeniable level, general-purpose programming language that is interpreted. The utilization of considerable indentation in its plan philosophy promotes code readability. Its language components and object-oriented approach are aimed at assisting programmers recorded as a hard copy clear, logical code for both small and large-scale projects .

Created by: Python Software Foundation

Day for kickoff: 20 Feb. 1991

Made by: Guido van Rossum

Stages on which it upholds: Windows, Linux, macOS and so forth.

Most recent Version: 3.10.0

System Requirements:

- Windows 7 or above.
- Minimum 3 GB RAM (8 GB recommended)
- Minimum 2 GB of memory space.
- Intel Core i5 or above processor.

Technology Used

Google Colaboratory:

Google Colaboratory (or known as Colab notebooks) let you mix executable code and rich text, as well as graphics, HTML, LaTeX, and more, in a single document. Your Colab notebooks are saved in your Google Drive account when you create them. You can quickly share your Colab notebooks with coworkers or acquaintances, allowing them to provide comments or even make changes.

Machine Learning:

Machine learning (ML) is a sort of artificial intelligence (AI) that allows software programmes to improve their prediction accuracy without being expressly planned to do so. In order to forecast new output values, machine learning algorithms use past data as input.

Machine learning is a crucial part of the rapidly expanding discipline of data science. Algorithms are trained to generate classifications or predictions utilizing statistical approaches, revealing crucial bits of knowledge in data mining initiatives. Following that, these bits of knowledge drive decision-making inside applications and enterprises, with the goal of influencing important growth KPIs. As huge data expands and grows, the demand for data scientists will rise, necessitating their assistance in distinguishing the most important business questions and, as a result, the data required to answer them.

Scikit learn:

In Python, Scikit-learn (Sklearn) is the most usable and robust machine learning package. It utilizes a Python consistency interface to give a set of fast tools for machine learning and statistical modelling, such as classification, regression, clustering, and dimensionality reduction. Rather than importing, altering, and summarising data, the Scikit-learn toolkit concentrates on data modelling. Scikit-learn is a free Python machine learning package. It supports Python numerical and scientific libraries like Numpy and Scipy, as well as techniques like support vector machine, random forests, and k-neighbors.

Numpy:

Numpy (also known as Numerical Python) is a library that consists of multidimensional array objects and a set of functions for manipulating them. Numpy allows you to conduct mathematical and logical operations on arrays. Numpy is a Python scripting language. 'Numerical Python' is what it stands for. It is the most important Python module for scientific computing.

Pandas:

Pandas is a Python library that provides quick, versatile, and expressive data structures for working with "relational" or "labelled" data. Its goal is to serve as the foundation for undertaking realistic, real-world data analysis in Python. Furthermore, it aspires to be the most powerful and versatile open-source data analysis and manipulation tool accessible in any language. It is well on its way to achieving this aim.

Seaborn:

Seaborn is a Python package based on matplotlib that is open-source. It's utilized for exploratory data analysis and data visualisation. With data frames and the Pandas library, Seaborn is a breeze to use. The graphs that are generated may also be readily altered. A couple of the upsides of information perception are recorded underneath.

1. Charts might help with the revelation of information patterns, which is gainful in any AI or estimating exertion.
2. Non-specialized people will comprehend your information better assuming that you use charts.
3. Introductions and reports with outwardly satisfying diagrams might be fundamentally more captivating to the peruser.

Matplotlib:

Matplotlib is a data visualisation and graphical plotting package for Python and its numerical extension Numpy that runs on all platforms. As a result, it provides an open-source alternative to MATLAB. Matplotlib's APIs (Application Programming Interfaces) may also be utilized to incorporate charts in graphical client interfaces.

In most cases, a Python matplotlib script is constructed so that only a hardly any lines of code are necessary to create a visual data plot. Two APIs are overlaid by the matplotlib scripting layer:

The pyplot API is a tree of Python code objects, with matplotlib at the top.

pyplot is a set of Oo (Object-Oriented) API objects that may be constructed with more freedom than pyplot. This API allows you to utilize Matplotlib's backend layers directly.

Chapter 04

Requirements on Major Project

DataSet:

Simply described, a dataset in machine learning is a collection of data bits that may be considered as a single unit by a computer for analytic and prediction purposes. This means that the data gathered should be homogeneous and intelligible to a computer that does not see data in the same manner that people do. The dataset is used to train the machine using different models and to use it so that the machine can make a step and predict the next move to do.

Knowledge of Machine Learning Algorithms:

The Knowledge of Machine Learning algorithms, pre-processing of data techniques and different models is a necessity for this project as we have used it a lot in the whole project.

Classification Model

Data Collection:

First we have imported the required file and then taken the Data to see[12].

Data Pre-processing:

We have done Pre-processing of data in the following ways:

First we treated null values present in our data. we have used mean value imputer, simple Imputer and Knn imputation method for replacing the null values present in it.

For the data, The average for both the genders male and female was taken separately and then missing values were replaced by their averages according to the gender.

There were only a few other genders present other than male and female so we have dropped them so as to make our model more predictable and easy.

We have dropped the id column from our data as it was not giving any meaningful information and was not helping to separate or detect or make any difference in the prediction model.

Smoking status: the missing value in smoking status were replaced with unknown, as unknown was already mentioned for few patients in that column.

Data Binning:

We have binned the continuous data which were present in the dataset which allow us to deal with the data and treat them easily.

The binned variables are (x2, y2).

Data label encoding:

There are few categorical variables were present in the data set to deal with them easily we have labelly encoded them.

Balancing The Dataset:

The dataset which we have was imbalanced so we make sure to balance them so as to make it easily understandable and get our desired results and make our model more smooth and easy to predict.

Models Used:

i) **Logistic Regression:** logistic regression (LR), it is a type of classification model rather than a regression. Strategic relapse is a basic and successful methodology for double and direct order issues. It's a clear characterization model that produces extraordinary outcomes with straightly distinct classes.

It is as often as possible utilized ordering approach in the modern world. A calculated relapse model, as Adaline and perceptron, is a factual strategy for double grouping that might be applied to multiclass order. It's a strategy for assessing the probability of a discrete result given an assortment of information factors.

ii) **Decision Tree:** Decision tree classifiers have numerous applications. Their capacity to catch elucidating decision-production data from the information given is their most fundamental characteristic. Preparing sets can be utilized to produce decision trees. A straightforward model for ordering models is a decision tree. It is administered by AI, where the information is disintegrated and persistently founded on certain boundaries. The methodology for deciding the class of a given informational index in a decision tree begins at the root hub of the tree. This calculation checks the upsides of the first quality against the upsides of the record trait (the genuine informational collection), then, at that point, follows the branch and moves to the following hub in light of the correlation. The calculation contrasts the property estimation and other kid hubs and continues on toward the following hub. It rehashes the entire interaction until it arrives at the leaf hub of the tree.

iii) **Random Forest:** The decision tree is the groundwork of arbitrary backwood classifiers. A decision tree is a progressive design worked from the properties of an information assortment (or free factors). An action combined with a subset of the elements separates the decision tree into hubs. The arbitrary woods is an assortment of decision trees that are connected to an assortment of bootstrap tests produced from the first informational index. To parcel the hubs, the entropy (or Gini list) of a subset of the properties is used. The bootstrapped subsets of the first informational collection have a similar size to the first informational collection. Breiman's articles on irregular wood classifiers are significant (Breiman, 1996, 2001). According to Suthaharan, the bootstrapping method works within the development of arbitrary timberlands with the required number of decision trees to increment arrangement exactness through the idea of cross-over diminishing (2015). The best trees are then picked by a democratic cycle and a bagging method (bootstrap total). This average arbitrary timberland method is utilized in the proposed mental registering design.

iv) **Custom Ensemble Models** - We have trained our data on custom ensemble models made with the help of Voting Classifiers. The Ensemble models combine various models like Logistic Regression and Decision Tree to train and test.

4.3.4 Proposed Methodology

Our Proposed Methodology comprises 5 Stage Process. At First, we will stack the information from the Dataset Collected from Mendeley, For the creation of the required model for Stroke prediction, it is important to use the right data for this as well as include important features for doing so later we will Preprocess the Data, Required Data Transformation Techniques and utilization of Various ML Algorithms and Resulting of different standard Metrics.

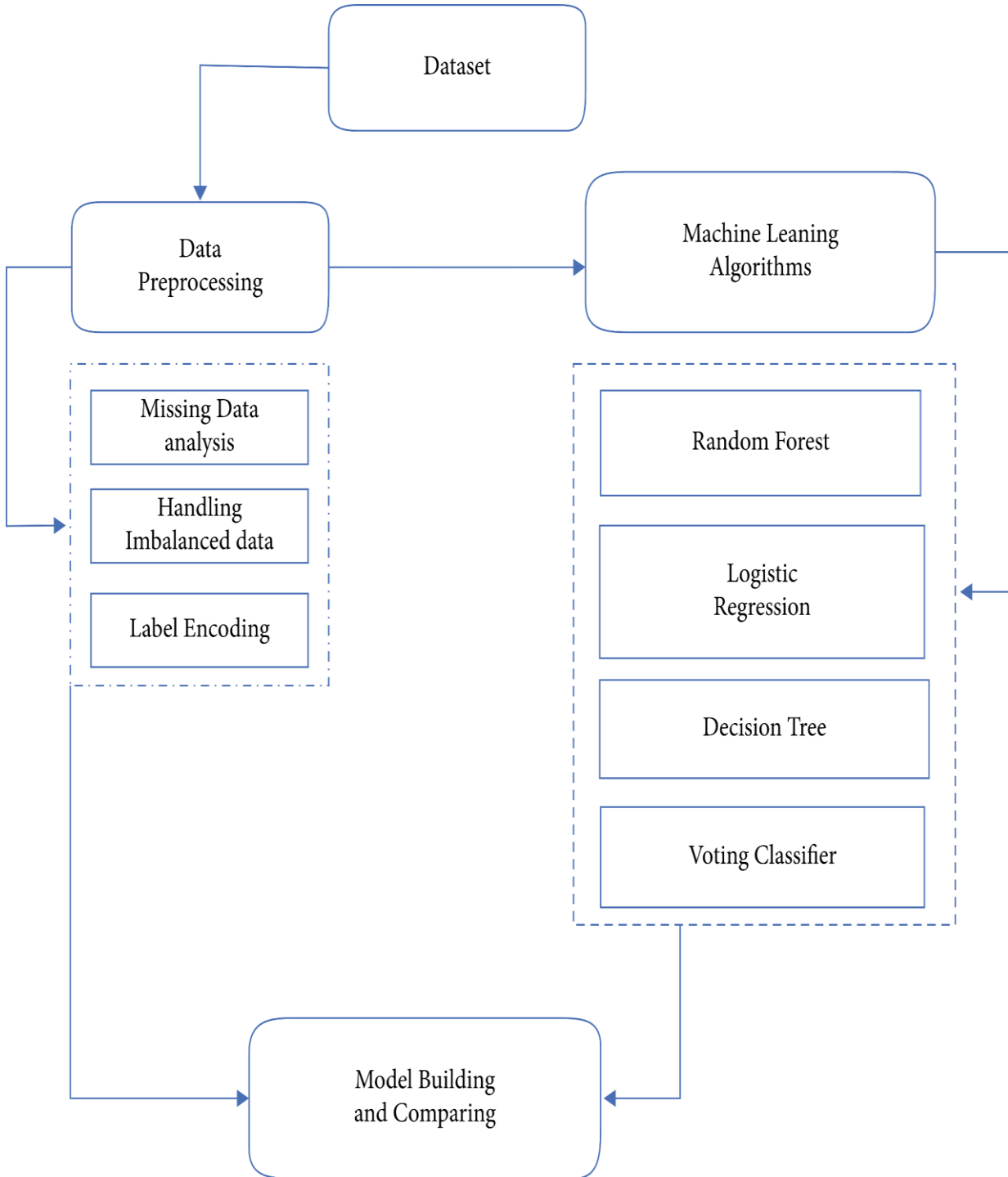


Figure 11: The architecture of the proposed Ensemble Model

TRAINING THE MODELS

i) Imbalanced data:

When the data were not balanced and we performed different methods we got the following results:

Accuracy: 96.66%

but F1 score = 4.12

	precision	recall	f1-score	support		precision	recall	f1-score	support
0	0.98	0.98	0.98	8199	0	0.98	0.98	0.98	8199
1	0.05	0.04	0.04	158	1	0.02	0.02	0.02	158
accuracy			0.97	8357	accuracy			0.97	8357
macro avg	0.51	0.51	0.51	8357	macro avg	0.50	0.50	0.50	8357
weighted avg	0.96	0.97	0.97	8357	weighted avg	0.96	0.97	0.96	8357
Accuracy Score:	0.9666148139284432				Accuracy Score:	0.965657532607395			
F1 Score:	0.041237113402061855				F1 Score:	0.020477815699658702			

In this case F1 score was too bad, so we have moved to next model

ii) KNN Imputed and Balanced Data:

Normal data:

Logistic Regression Test Accuracy: 0.7452139982928911

Decision Tree Test Accuracy: 0.690830386538227

RandomForest Test Accuracy: 0.7422875259114742

Binned Data:

	precision	recall	f1-score	support		precision	recall	f1-score	support
0	0.98	0.89	0.93	8159	0	0.98	0.92	0.95	8159
1	0.90	0.99	0.94	8243	1	0.92	0.99	0.95	8243
accuracy			0.94	16402	accuracy			0.95	16402
macro avg	0.94	0.94	0.94	16402	macro avg	0.95	0.95	0.95	16402
weighted avg	0.94	0.94	0.94	16402	weighted avg	0.95	0.95	0.95	16402
Accuracy Score:	0.9375685891964395				Accuracy Score:	0.9521399829289111			
F1 Score:	0.9407132931912923				F1 Score:	0.9539075802947566			
					Confusion Matrix :	[[7494 665] [120 8123]]			

The best accuracy achieved in this model is 95.21% using Random Forest and Binned data.

iii) Using Ensembling Techniques:

Ensemble Techniques:

- 1) Using Logistic regression, Decision Tree, Random forest and SVC:
Accuracy: 93.04%
- 2) Using Decision Tree, Random forest and SVC:
Accuracy: 84%
- 3) Random forest, SVC and Logistic Regression:
Accuracy: 87%
- 4) Decision Tree, Random Forest, Logistic Regression:
Accuracy: 94.88%

The best accuracy we got here is 94.88%.

iv) Final Model:

DataSet: Imputation (gender wise mean value) and binned numerical variables.

Model Used: Random Forest

The following given below is the classification report for the Random Forest model.

	precision	recall	f1-score	support
0	0.98	0.93	0.96	8385
1	0.94	0.98	0.96	8658
accuracy			0.96	17043
macro avg	0.96	0.96	0.96	17043
weighted avg	0.96	0.96	0.96	17043

Accuracy Score: 0.9583406677228188

F1 Score: 0.959990983883692

When we used Stratified K-fold for cross-validation,

we achieved an average accuracy of **96.01%**

And, the minimum accuracy was **95.44%**.

And Maximum accuracy was **96.50%**.

Chapter 05

Experimental Evaluation and results:

In this Section, we have discussed the result obtained and analysis of our proposed framework. In addition, we compared our model to various current models in terms of accuracy, precision, sensitivity, precision, F1 Score, etc.

5.1 Performance Metrics:

True positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs) are the performance measures mentioned here, as stated below:

True positive rate (TPR) or sensitivity: When the disease is present, it describes the chance of a classifier correctly anticipating a positive result. The formula is as follows:

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad \text{-----Eqn1[4]}$$

Specificity or True negative rate (TNR): It is a classifier's likelihood of predicting a poor outcome when there is no sickness. The formula is as follows:

$$\text{Specificity} = \frac{TN}{TN+FP} \quad \text{-----Eqn2[7]}$$

Accuracy: It is one of the most widely used metrics for assessing the performance of a classifier. It's stated as: It's calculated as a percentage of correctly identified samples.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{-----Eqn3[8]}$$

AU-ROC (area under the receiver operating characteristic curve): It is also a helpful and extensively used performance statistic for classification issues. TPR vs FPR at various threshold values are plotted. The AU-ROC is an excellent metric for performance comparison because it evaluates performance across a wide range of class distributions and error levels. This is how it's defined:

$$\text{AU-ROC} = 1/2 \left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP} \right) \quad \text{-----Eqn4[3]}$$

F1 score: It is defined as the weighted average of precision and recall (or harmonic mean). A score of 1 is considered the best, while a score of 0 is regarded as the worst. In F-measures, the TNs are not taken into account. The following formula can be used to compute the F1 score:

$$\text{F1 score} = 2 \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad \text{-----Eqn5[6]}$$

5.2 Comparison with ML Models

We built numerous baseline models and used Stratified K-fold for cross-validation. Models with high accuracies are used in the stacking approach. The accuracy of the baseline models is shown in Table 5. As we have observed from Table 5, the best performing algorithm is using Stratified K Fold. As shown in Table 5, we have tested the algorithms on the basis of accuracy. From Table 5, the Stratified K fold classifier has a greater classification accuracy of 96.01 per cent than the other classifiers. Hence we can conclude that the algorithms are chosen by us to ensemble our model gives the highest accuracy of the 9 algorithms we have compared it with.

Reasons for using Stratified K Fold: By using this method, It maintains the class ratio same throughout the K folds as the ratio in the original dataset. It boosts the accuracy and F score of the model and can produce great results as we can see in Table 5. It has outperformed other algorithms and combinations of ensemble models. It uses the method of splitting data into folds that follows the criteria such as each fold having the same proportion of observations. It is in particular the most commonly used method in machine learning.

Table 5: Comparison of ML algorithms & their respective accuracies

S. No.	Algorithm	Accuracy
1.	Logistic regression	69.57%
2.	Decision Tree	94.26%
3.	Random Forest	95.26%
4.	Decision Tree Using Stratified K fold	96.01%
5.	Ensemble model of - decision tree, random forest, SVM and logistic regression with voting classifier with binned data	92.73%
6.	ENSEMBLE MODEL OF - DECISION TREE, RANDOM FOREST AND SVM COMBINED WITH VOTING CLASSIFIER	84.32%
7.	ENSEMBLE MODEL OF - RANDOM FOREST, SVM AND LOGISTIC REGRESSION COMBINED WITH VOTING CLASSIFIER WITH BINNED DATA	87.8%
8.	ENSEMBLE MODEL OF - DECISION TREE, RANDOM FOREST AND LOGISTIC REGRESSION COMBINED WITH VOTING CLASSIFIER WITH BINNED DATA	94.8%
9.	ENSEMBLE MODEL OF - DECISION TREE, RANDOM FOREST AND SVM COMBINED WITH VOTING CLASSIFIER WITH BINNED DATA	94.7%

Table 6: Comparison of proposed Framework with existing ML Models

Model	Accuracy
Proposed Approach	96.01%
Deep neural network	71.6%
Support vector machine	Outperforms cox model
Naive Bayes	95%
Decision tree, principle component analysis	94.7%
Logistic Regression, Random Forest, Decision Tree, KNN, SVM and Naïve Bayes	82%
KNN, Random forest	90.15%
Random Forest	91.7%
Machine Learning, Decision Tree, Naive Bayes, Random Forest, Machine Learning	94.24%
Decision Tree, Logistic Regression(LR), Random Forest, Voting Classifier	95.8%
Logistic Regression, SVM	92%

RESULTS:

After performing all the models and combinations of ensemble models, The best Accuracy we got is from Random Forest Which was of 96.01% and an F1 score of 95.99. This shows us that we were able to predict the stroke in the patient pretty accurate after performing all the possible models on the dataset. We were able to train, treat and generate output from the dataset.

Application

- 1) This can be applied in the medical field to predict the possibility of Brain Stroke in a patient following by their data which includes their age, work type, residence type, smoking status, medical history etc.
- 2) It can also be used to make the medicines and to find the cure for the brain stroke depending on the condition of a particular patient, age and past experiences.

Limitations:

There are a few limitations of our project:

1. The chances of stroke below the age of 40 is very less so our model may not be able to predict if someone of less age develops symptoms.
2. The numerical output of a nomogram, or prediction system included in an electronic health record, may provide the impression of scientific legitimacy, which may be harmful to patient care.
3. Also, including simply those who are expected to have a certain result presupposes that our therapies will be equally effective in those who are expected to have a high or low risk of that event—this may not be the case.
4. Prediction of Stroke might create panic in a patient for a short duration but it will be beneficial and can be treatable early as well.

Future Work:

We have built an amazing brain stroke prediction system and were successful so far but we have our thoughts for future implementations as well. we have thought to further improve its accuracy and write a research paper on it and publish it in a trusted space so that it can be recognized and accessible to all we have thought to make it more robust, easy and simplified. We are also planning to make a web application on it and deploy it online as well in future.

Conclusion:

We built a machine learning model to predict whether someone has a risk of encountering stroke. For this, we used the patient's past health records and daily habits like smoking and living conditions as dataset. For the best predictions, the dataset is trained using a random forest and the accuracy achieved is 96.01% and the F1 score is 95.99.

To Conclude, I would like to say it's been a rollercoaster journey so far in this project with so many hurdles and much efforts. Furthermore, we shall gladly accept the feedback and add new features to it in the future to make it more robust and easily reachable.

References

- [1] Tianyu Liu, Wenhui Fan, Cheng Wu, A hybrid machine learning approach to cerebral stroke prediction based on the imbalanced medical dataset, *Artificial Intelligence in Medicine*, Volume 101, 2019, 101723, ISSN 0933-3657.
- [2] [KDD '10: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining](#) July 2010.
- [3] M. U. Emon, M. S. Keya, T. I. Meghla, M. M. Rahman, M. S. A. Mamun and M. S. Kaiser, "Performance Analysis of Machine Learning Approaches in Stroke Prediction," 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2020, pp. 1464-1469, doi: 10.1109/ICECA49313.2020.9297525.
- [4] M. S. Singh and P. Choudhary, "Stroke prediction using artificial intelligence," 2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON), 2017, pp. 158-161, doi: 10.1109/IEMECON.2017.8079581.
- [5] Analyzing the Performance of Stroke Prediction using ML Classification Algorithms Gangavarapu Sailasya¹, Gorli L Aruna Kumari² Department of Computer Science and Engineering GITAM Institute of Technology, GITAM (Deemed to be University).
- [6] Heart Stroke Prediction using Machine Learning B.P. Deepak Kumar^{*1} , Sagar Yellaram^{*2} , Sumanth kothamasu^{*3} , Surendhar Reddy Puchakayala^{*4} ^{*1}Assistant Professor, Department of Computer Science and Engineering, CMR Technical Campus, Medchal, Telangana, India ^{*2} JNTUH, Computer Science and Engineering, CMR Technical Campus, Medchal, Telangana, India ^{*3} JNTUH, Computer Science and Engineering, CMR Technical Campus, Medchal, Telangana, India ^{*4} JNTUH, Computer Science and Engineering, CMR Technical Campus, Medchal, Telangana, India.
- [7] Stroke prediction through Data Science and Machine Learning Algorithms Preprint · June 2021 DOI: 10.13140/RG.2.2.33027.43040, Jose-A Tavares.
- [8] Shoily, Tasfia & Islam, Tajul & Jannat, Sumaiya & Tanna, Sharmin & Alif, Taslima & Ema, Romana. (2019). Detection of Stroke Disease using Machine Learning Algorithms. 1-6. 10.1109/ICCCNT45670.2019.8944689.
- [9] 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 Khan¹ ¹Department of Electrical and Computer Engineering, North South University, Bashundhara, Dhaka 1229, Bangladesh, ²Department of Information Technology, College of Computers and Information Technology, Taif University, P.O. Box 11099, Taif 21944, Saudi Arabia
- [10] Machine Learning in Action: Stroke Diagnosis and Outcome Prediction. Shraddha Mainali¹, Marin E. Darsie^{2,3}, and Keaton S. Smetana⁴
- [11] Divya Khanna, Prashant Singh Rana, Multilevel Ensemble Model for Prediction of IgA and IgG antibodies, *Immunology Letters* (2017), <http://dx.doi.org/10.1016/j.imlet.2017.01.017>
- [12] Benjamin Letham. Cynthia Rudin. Tyler H. McCormick. David Madigan. "Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model." *Ann. Appl. Stat.* 9 (3) 1350 - 1371, September 2015. <https://doi.org/10.1214/15-AOAS848>

Dataset:

[13] <https://data.mendeley.com/datasets/x8ygrw87jw/1>

Others:

[14] [Stroke - Symptoms and causes - Mayo Clinic](#)

[15] <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>

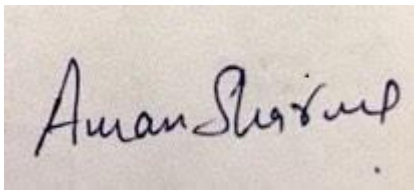
[16] [sklearn.ensemble.VotingClassifier — scikit-learn 1.0.1 documentation](#)

[17] <https://scikit-learn.org/stable/modules/preprocessing.html#preprocessing>

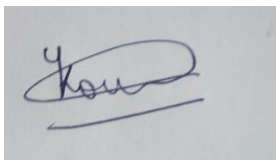
[18] https://scikit-learn.org/stable/modules/cross_validation.html#cross-validation-iterators

[19] https://scikit-learn.org/stable/modules/model_evaluation.html#classification-metrics

[20] [Stroke \(Brain Attack\) - Conditions & Treatments - UCLA Radiology, Los Angeles, Westwood, Manhattan Beach, Santa Monica, CA \(uclahealth.org\)](#)



Aman Singh



Kou

