

**A STUDY FOR PREDICTING CEREBRAL STROKE USING DIFFERENT  
KIND OF MACHINE LEARNING TECHNIQUES**

By

**Yeasmin Hena Sathi**

**ID: 171-35-213**

Under the supervision of

**Dr. Imran Mahmud**

Associate Professor & Head  
Department of Software Engineering  
Faculty of Science and Information Technology  
Daffodil International University



**Department of Software Engineering  
DAFFODIL INTERNATIONAL UNIVERSITY  
5 JUNE 2021**

## **APPROVAL**

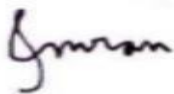
This thesis titled as “A STUDY FOR PREDICTING CEREBRAL STROKE USING DIFFERENT KIND OF MACHINE LEARNING TECHNIQUES”, submitted by Yeasmin Hena Sathi, ID: 171-35-213 to software Engineering Department, Daffodil International University has been accepted qualify the prerequisites for authorization of graduating B.Sc in Software Engineering and approved as to its style and contents.

## **BOARD OF EXAMINERS**

## THESIS DECLARATION

I do hereby declare that this report was written by me, Yeasmin Hena Sathi, under the supervision of Dr. Imran Mahmud, Associate Professor & Head, Dept. of Software Engineering in Daffodil International University. This thesis is submitted in partial fulfillment of the requirement for the degree of B.Sc. in Software Engineering. This thesis neither in whole nor in part has been previously submitted for any degree.

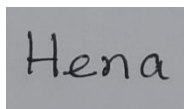
**Supervised By,**



**Dr. Imran Mahmud**

Associate Professor & Head  
Department of Software Engineering  
Faculty of Science and Information Technology  
Daffodil International University

**Submitted By,**



Yeasmin Hena Sathi

ID: 171-35-213

Department of Software Engineering  
Faculty of Science and Information Technology  
Daffodil International University

## **ACKNOWLEDGEMENT**

First of all, I would like to thank my Allah for enabling me to work this proposal work. I would remain forever obliged to the Department of Software Engineering, Daffodil International University, Bangladesh for giving me the scope to carry out the thesis. I would like to express my cordial gratitude to my supervisor, Dr. Imran Mahmud, Associate Professor & Head, Department of Software Engineering. His guidance helped in all the time of research and writing of this thesis. I also want to give exceptional appreciation to MD. Rajib Mia, Lecturer, Department of Software Engineering and Shariful Islam, Lecturer, Department of Software Engineering. Their legitimate course and direction assist me with setting up this proposal work with no trouble. I also grateful to the faculties of the Software Engineering Department for offering their helpful hands during this thesis, Lastly ,I want to acknowledge the contribution of my parents , family members and my friends for their constant and never ending motivation. Their hopefulness and support have permitted them to defeat any obstruction at any stage.

<b>Contents</b>	<b>Page</b>
<b>APPROVAL .....</b>	<b>ii</b>
<b>THESIS DECLARATION .....</b>	<b>iii</b>
<b>ACKNOWLEDGEMENT .....</b>	<b>iv</b>
<b>LIST OF FIGURES .....</b>	<b>vi</b>
<b>LIST OF TABLE .....</b>	<b>vi</b>
<b>LIST OF ABBREVIATIONS .....</b>	<b>vii</b>
<b>CHAPTER 1: INTRODUCTION.....</b>	<b>2</b>
<b>1.1 Background .....</b>	<b>2</b>
<b>1.2 Motivation of Research.....</b>	<b>3</b>
<b>1.3 Problem Statement: .....</b>	<b>3</b>
<b>1.4 Research Question: .....</b>	<b>3</b>
<b>1.5 Research Objectives: .....</b>	<b>4</b>
<b>1.6 Research Scope: .....</b>	<b>4</b>
<b>1.7 Thesis Organization.....</b>	<b>4</b>
<b>CHAPTER 2: LITERATURE REVIEW.....</b>	<b>5</b>
<b>CHAPTER 3: METHODOLOGY.....</b>	<b>6</b>
<b>3.1 Dataset Description.....</b>	<b>6</b>
<b>3.2 Data Preprocessing.....</b>	<b>8</b>
<b>3.2.1 Data cleaning .....</b>	<b>8</b>
<b>3.2.2 Data Encoding .....</b>	<b>8</b>
<b>3.2.3 Applying SMOTE .....</b>	<b>8</b>
<b>3.3 Hyperparameter Tuning .....</b>	<b>9</b>
<b>3.3.2 Manual Search .....</b>	<b>10</b>
<b>3.3.3 Random Search .....</b>	<b>10</b>
<b>3.4 Machine Learning Model .....</b>	<b>11</b>
<b>3.4.1 Random Forest .....</b>	<b>11</b>
<b>3.4.2 K-Nearest Neighbors.....</b>	<b>12</b>
<b>3.4.3 Logistic Regression .....</b>	<b>12</b>
<b>CHAPTER 4: RESULT AND DISCUSSION.....</b>	<b>13</b>
<b>REFERENCES.....</b>	<b>18</b>

## LIST OF FIGURES

Figure no	Figure Caption	Page no
Fig. 3.1.	Pie chart ‘smoking status’ attributes that shows the attributes data is deficient.	07
Fig. 3.2.	Pie chart of ‘ bmi’ attributes where shows ‘bmi’ attributes has 3% missing values.	07
Fig. 3.3.	Imbalanced class before using SMOTE & Balanced class after using SMOTE.	09
Fig. 3.4.	Workflow which describes the process of predicting cerebral stroke using stroke prediction dataset.	11
Fig. 4.1.	Correlation matrix for stroke prediction dataset attributes which shows the correlated value between the featured of cerebral stroke prediction dataset.	14
Fig4.2.	Precision-Recall curve for machine learning techniques which describes the best techniques for cerebral stroke prediction.	15
Fig.4.3.	ROC curve for ML techniques which describes the best techniques for cerebral stroke prediction.	16

## LIST OF TABLE

Table No	Table Caption	Page no:
Table 1	Dataset Attributes representation which is used for predicting cerebral stroke.	06
Table 2	Confusion matrix which describes the TP, TN.FP, FN of all the three Machine learning Techniques.	15
Table 3	Values of Precision, Recall, F1- Score of all the three Machine Learning Techniques.	15

## **LIST OF ABBREVIATIONS**

ML	Machine Learning
SVM	Support Vector Machine
IST	International Stroke Trial
KNN	K-Nearest Neighbor
SMOTE	Synthetic Minority Oversampling Technique
RF	Random Forest
LR	Logistic Regression
ROC	Receiver Operating Characteristic
AUC	Area under the ROC Curve
TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative

## **ABSTRACT**

An interruption or reduction in the supply of blood to the cerebrum produces a cerebral stroke. This supply shortage leads in a deficit of oxygen or vital nutrients, which causes brain cells to die. Stroke occurs mostly as a result of people's lifestyle choices in the advanced time-changing factors, for example, excessive glucose levels, heart disease, stoutness, and diabetes. Developing countries account for 85 percent of all stroke deaths worldwide. The early termination of a cerebral stroke is critical for effective counteraction and therapy. The best way to deal with this risk is to prevent it from happening in the first place by managing the relevant metabolic factors. Nonetheless, it is difficult for clinical workers to determine how much additional safety precautions are necessary for an expected patient based only on the examination of physiological indicators unless they are plainly abnormal. Examination reveals that behaviors extricated from various hazard limits transmit critical information for the prediction of stroke. The data was obtained from the Harvard Dataverse Repository and was properly prepared and tested using machine learning techniques such as RF, LR and KNN. We implemented the RF, LR, and KNN algorithms with hyperparameter tuning in this study and determined the best method among them. For performance evaluation, we use the AUC-ROC curve, the Precision-Recall curve, and the F1-score, and all reports reveal the best strategies. This strategy may be seen as a different option, with a low cost and a constant analytic technique that can get exact stroke prediction.



## CHAPTER 1: INTRODUCTION

### 1.1 Background

Stroke is a neurological condition in which regions of the brain do not get blood, ending in a stroke that can be deadly at times (Rajora M., 2021). With an estimated 795,000 stroke events in the United States each year, stroke is now the leading cause of long-term adult human disorder and the fifth most cause of death in the United States. (Roger VL, 2011). Due to a lack of sufficient therapy, it is doubtful that stroke victims will be entirely healed. Even if the patient is recovered, they will confront harsh realities such as lifetime disability, inability, limited social functions, and so on. As a result, the condition places a considerable strain on individuals, the medical community, and civilization. (Liu Y, Yin B, Cong Y. , 2020). Predicting stroke risk will make a major impact in stroke detection and control healing (Rajora M., 2021).

Cerebral stroke is a major cause of death and the primary cause of adult long-term impairment in developed nations. The majority of cerebral strokes (85%) are ischemic in nature, produced by the obstruction of a substantial cerebral vein by a blood vessel or an effusion, causing a lack of blood flow and tissue loss in the afflicted area (Gibson CL. , 2013). According to the Health ministry and Welfare's data on the top ten causes of death for Taiwanese in 2018, cerebrovascular disease is the fourth leading cause of death, taking around 10,000 lives a year (C. -C. Peng, S. -H. Wang, S. -J. Liu, Y. -K. Yang and B. -H. Liao, 2020). While the pathogenesis of stroke is still unknown, it is widely accepted that stroke is intimately linked to abnormal metabolic markers for both hemorrhagic and ischemic stroke, and that more than 90% of metabolic syndrome points for this scenario are repressible, so additional focus should be given on reduction (Liu T, Fan W, Wu C., 2019) . Early stroke detection is a source of concern and is desperately required in the field of medicine. In terms of assessing the risk of a stroke, ML approaches are likely worth investigating.

In this study, we will try to showcase the best ML approaches based on classification algorithms for estimating cerebral stroke for medical evaluation based on clinical data with inadequacy and quantity disproportion in our work.

### **1.2 Motivation of Research**

The early termination of a cerebral stroke is critical for effective counteraction and therapy. The best way to deal with this risk is to prevent it from happening in the first place by managing the relevant metabolic factors. Nonetheless, it is difficult for clinical workers to determine how much additional safety precautions are necessary for an expected patient based only on the examination of physiological indicators unless they are plainly abnormal. Examination reveals that behaviors extricated from various hazard limits transmit critical information for the prediction of stroke.

### **1.3 Problem Statement:**

Since finding out about past comparative works, I've seen a few shortcomings, which may take this to a higher level of the examination.

- ✓ No balancing data process for imbalanced dataset.
- ✓ Did not work with one hot encoding.
- ✓ Did not work with hyperparameter tuning.
- ✓ Performance evaluation using ROC, Precision-Recall curve.

### **1.4 Research Question:**

The fundamental steps of research is to investigate some answerable queries of an issue. That is called research questions. The list is given bellow.

- ✓ How to gather the informational index for chosen disease?
- ✓ Why delete data from dataset?
- ✓ How to convert data and why this process is important?
- ✓ Why data need to be balanced?

- ✓ What is hyperparameter tuning and why hyperparameter tuning is important
- ✓ What type of parameter used in this examination?
- ✓ How to examine the performance appraisalment?
- ✓ How to add references?

### **1.5 Research Objectives:**

The key goals of this thesis are given below:

- ✓ To propose the best model to achieve a better result with data imbalanced and incompleteness.
- ✓ For early diagnosis, this approach can acquire precise stroke prediction.
- ✓ Stroke risk prediction make a significant difference in stroke prevention and early recovery.

### **1.6 Research Scope:**

Research scopes describes the area which the researchers have been analyzed. Here is our research scopes.

- ✓ Balancing data
- ✓ Converting categorical data to numerical data
- ✓ Using hyperparameter tuning for best model performance.
- ✓ Performance measurement using ROC, Precision- Recall curve and AUC score.

### **1.7 Thesis Organization**

The paper has been organized with five sections which are described underneath:

Part 1: In this section, the examination foundation, inspiration, issue articulation and targets are given.

Part 2: This section describes a conversation of the current related work.

Part 3: This section contains the examination technique and approaches as follows for the

Examination.

Part 4: This section contrasts the tested outcomes and existing methodologies.

Part 5: The examination result and the constraint of this investigation are introduced here and the heading of things to come work of exploration has likewise been guided.

## **CHAPTER 2: LITERATURE REVIEW**

In the meanwhile, much research has been conducted using various ML approaches to predict cerebral stroke. Researchers have been working on the rapid detection of neurological illnesses, which can be difficult to obtain. Their contributions and accompanying work are described here to provide context.

M. S. Singh et al. (M. S. Singh and P. Choudhary , 2017) with this research, they investigated alternative techniques for stroke prediction using the Cardiovascular Health Study (CHS) dataset. The decision tree algorithm is used to pick attributes, the principle component analysis technique is used to minimize dimension, and the backpropagation neural network classification technique is used to construct a classification model. This study revealed the best prediction model for stroke illness after investigating and analyzing classification efficiency with various methodologies and variance prototype accuracy using AI.

R. S. Jeena et al. (R. S. Jeena and S. Kumar, 2016) revealed that measures generated from several risk variables give helpful information for stroke prediction. This researchers focused at the many physiological characteristics that act as risk factors for stroke outcome. To analyze and evaluate results from the IST database, the SVM was employed. In this research, the author used SVM with several kernel functions and observed that the linear kernel had 90% efficiency.

Selma Yahiya Adam et al. (Selma Yahiya Adam, Adil Yousif and Mohammed Bakri Bashir, 2016) authors employed the KNN and decision tree algorithms in their investigation to classify ischemic stroke. In their investigation, medical practitioners that used the decision tree method to identify stroke had higher accuracy.

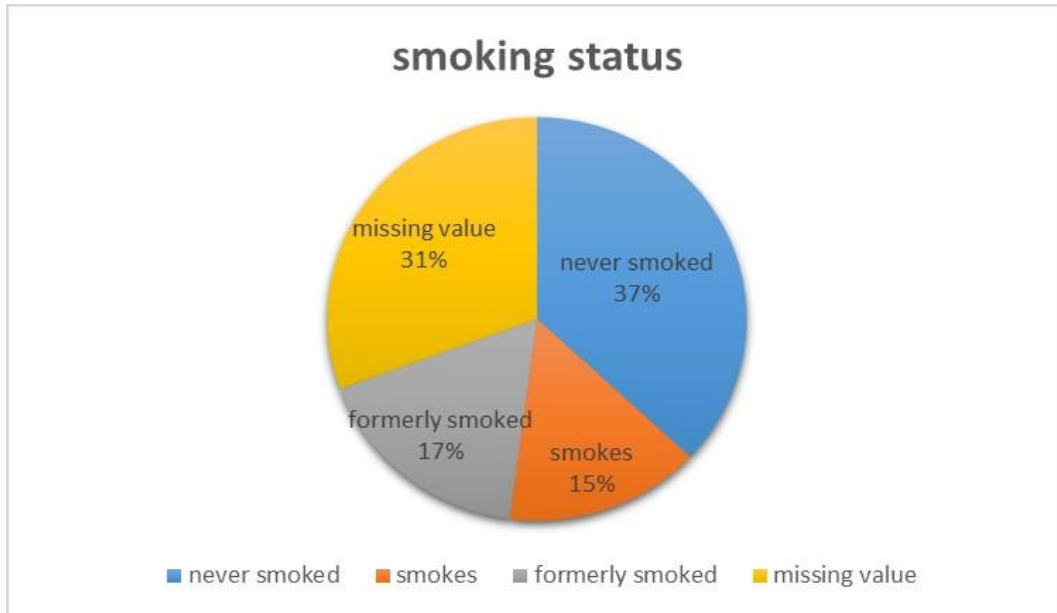
## CHAPTER 3: METHODOLOGY

### 3.1 Dataset Description

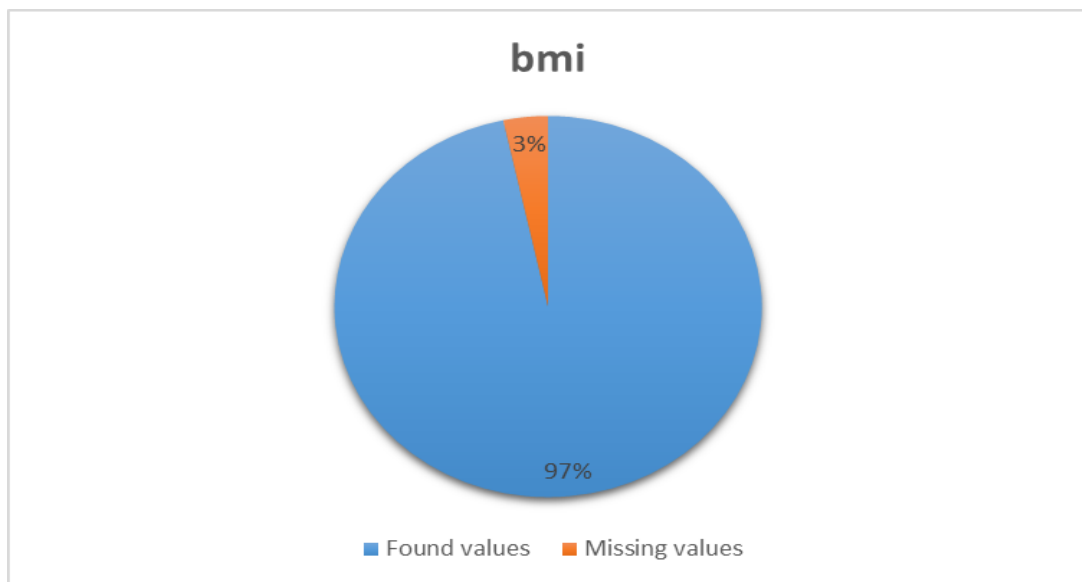
This study's dataset is taken from the Harvard Dataverse Repository (M, Mark, 2021). It is primarily focused on the patient's basic physiological records, prior medical history, and living situation. The dataset is a level unbalanced sort dataset with 12 attributes, comprising 783 stroke events in 43,400 samples gathered. Furthermore, the dataset is insufficient, with 30% of smoking status data and 3% of body mass index (BMI) data missing. Risk factors in this dataset include age, gender, hypertension, heart\_disease, BMI, smoking\_status, glucose\_level, ever\_married, work\_type, and residence\_type. Table 1 shows dataset attributes type and their values information and Figure 3.1 describes condition of smoking\_status in dataset.

**Table 1.** Dataset Attributes representation which is used for predicting cerebral stroke.

Attributes Name	Value	Attributes Name	Value	Attributes Name	Value
gender	Male, Female, Others	ever_married	Yes, No	bmi	Numeric data
age	Numeric Data	work_type	Children, Private, Govt_job, Never_worked, self-employed	Smoking status	Formerly smoked, never smoked, smokes.
hypertension	Numeric data	Residence_type	Urban, Rural		
heart_disease	Numeric data	avg_glucose_level	Numeric Data		



**Fig.3.1.** Pie chart ‘smoking status’ attributes that shows the attributes data is deficient. Figure 3.2 describes the condition of bmi has 3% of missing information which need to replaced or removed for better execution-assessment.



**Fig 3.2.** Pie chart of ‘ bmi’ attributes where shows ‘bmi’ attributes has 3% missing values.

## **3.2 Data Preprocessing**

### **3.2.1 Data cleaning**

There are some missing values for the BMI and smoking status characteristics in the dataset.

Missing values, on the other hand, are addressed by removing missing value rows. After the data was eliminated, 29072 samples were selected.

### **3.2.2 Data Encoding**

Absolute factors are commonly used in both regression and classification arrangement analysis.

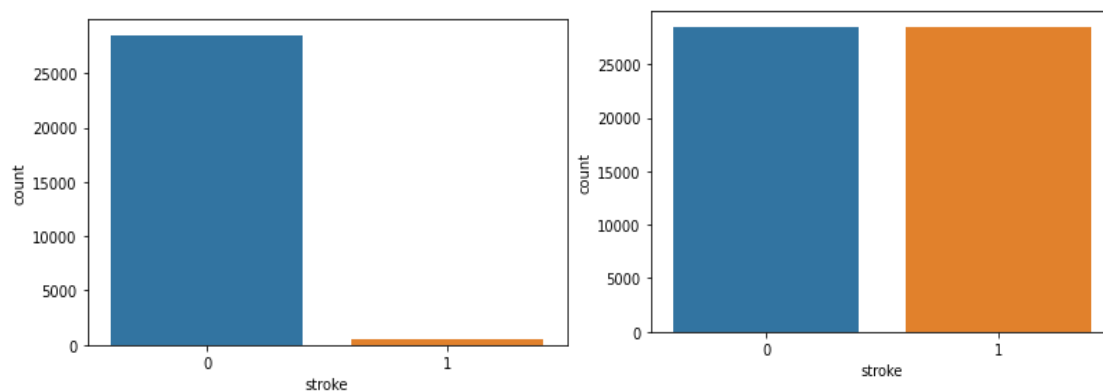
Nonetheless, ML computations only recognize numeric characteristics as information. To use unmodified data for ML purposes, the data should be encoded into numeric characteristics to the point that each all-out highlight is addressed with a number (Carey, 2003) . The most often used encoding scheme is One Hot Encoding. It compares each level of the all-out factor to a set reference level. One hot encoding converts a single parameter with n perceptions and d specific attributes to d double factors with n perceptions each. Every perception that demonstrates the presence (1) or absence (0) of the categorical-paired component (Kedar Potdar, Taher S Pardawala and Chinmay D Pai, 2017). To transform categorical information such as gender, ever married, employment type, dwelling type, and smoking status into numeric data, one-hot encoding is employed in this study article.

### **3.2.3 Applying SMOTE**

A large number of datasets are required in ML. Most learning algorithms struggle with class imbalance because they are biased toward learning and detecting dominant groups. As a result, minority cases are frequently incorrectly labelled (T. Maciejewski and J. Stefanowski, 2011). Because certain machine learning models are prone to unbalanced data, the data must be pre-processed before the models can be constructed. To begin, we utilize the SMOTE technique to oversample minority groups in order to increase the amount of data in such classes. It generates

simulated records for the minority class by employing linear interpolation. These synthetic records are created for each example in the minority group by randomly picking one or more of the KNN. (T. Maciejewski and J. Stefanowski, 2011). During the oversampling step, data is rebuilt, and several models can be applied to the processed data.

Data which is used for this study was imbalanced class type. For getting better quality and performance we apply smote and made data balanced class type. In Figure 3.4. the first diagram show that the imbalanced value of class 0 and 1 before using SMOTE. And second diagram clearly show that the balanced value of class 0 and 1 after using SMOTE.



**Fig. 3.3.** Imbalanced class before using SMOTE & Balanced class after using SMOTE.

### 3.3 Hyperparameter Tuning

Because hyperparameters affect all activities of efficient processes and have a significant influence on model success, they are required for ML approaches. Modern supervised ML algorithms feature hyperparameters that must be configured before they can be executed. The user can establish hyperparameters using the tool model's default settings, manually configure them, or tweak them for maximum prediction efficiency using a tuning procedure (Probst, P., A. Boulesteix and B. Bischl, 2019). Hyperparameters that cannot be obtained directly via the standard prepping procedure. They are usually established before the actual planning method begins. These



constraints determine important aspects of the algorithm, such as its complexity or how quickly it should learn. There are three most generally utilized hyperparameter determination techniques 1) Grid Search 2) Manual Search 3) Random Search (D. C. R. T. P. K. S.-H. L. R. M. P. Steven R. Young, 2015) .

### **3.3.1 Grid Search**

Grid search is a very well and widely used approach for matching hyperparameters. This is used to examine an explicitly defined sample of AI computation hyperparameters. (Putatunda, Sayan & Kiran R, Dr. ).

Furthermore, Grid search is quick to use and does similar functions without issue. However, as the number of hyperparameters increases, conducting Grid search becomes computationally expensive. (Putatunda, Sayan & Kiran R, Dr. )

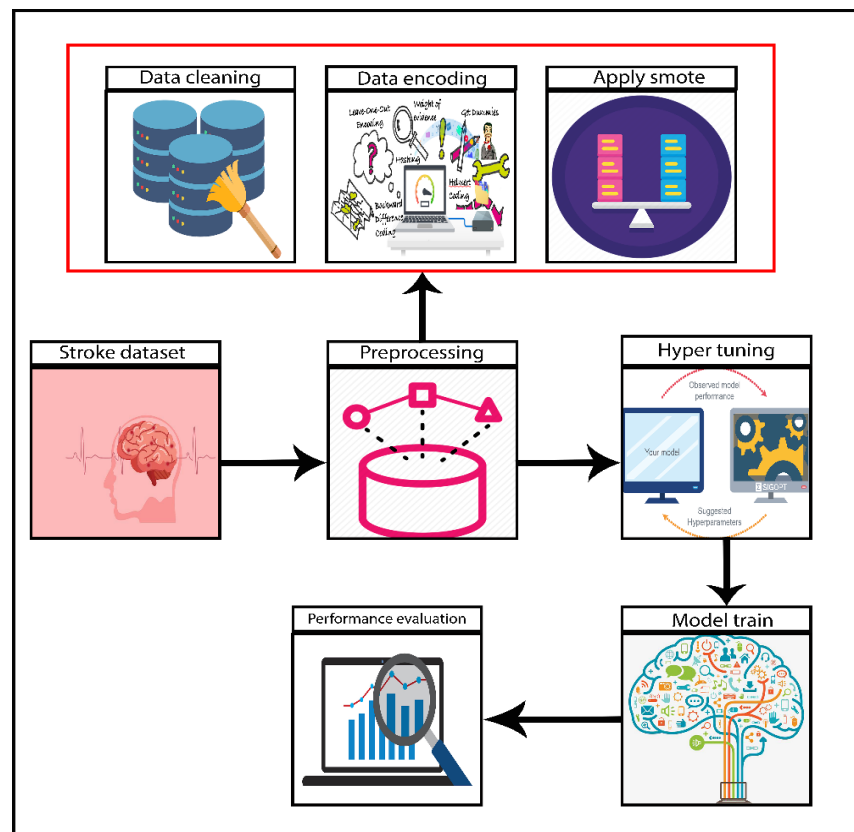
### **3.3.2 Manual Search**

The specialist physically selects a collection of hyperparameters to evaluate. This can sometimes result in a quick solution, although prior area knowledge and engagement with hyperparameter selection based on that space are necessary. In any event, a developing space master will have difficulty in selecting hyperparameter. Furthermore, the identical hyperparameter cannot be applied to a different dataset. A similar manual method should be used with the new dataset.

### **3.3.3 Random Search**

Random search is the most efficient way of exploring the hyper-parameter space. It addresses difficulties caused by under-sampling of critical dimensions, and random search draws from the same phase space as a grid search at a known thickness (Panda, Balaram, 2019). This means that it selects a survey method of hyperparameter values that accurately represents the grid's community.

We will rely on random search in this study. In high-dimensional spaces, Random search surpasses Grid search, and in general, Random search surpasses Grid search in the majority of circumstances (J. Bergstra and Y. Bengio, 2012). When compared to other hyperparameter approaches, the randomized hyperparameter approach requires the least amount of time to execute (Putatunda, Sayan & Kiran R, Dr. ).



**Fig. 3.4.** Workflow for the process of predicting cerebral stroke using stroke prediction dataset.

## 3.4 Machine Learning Model

### 3.4.1 Random Forest

Random forests, also known as random decision forests, are a group learning approach for arrangement, relapse, as well as other activities that function by building a massive number of possible trees during the preparation work and offering the group that is the process of the groups (characterization) or average assessment (relapse) of the leaf nodes. RF, as the name implies, is made up of a large set of individual decision trees that operate together as an ensemble. Every

individual tree in the RF lets out a class forecast and the class with most votes turns into our model's anticipation (Revanth S, Sanjay S, Sanjay N, Vijayaganth V, 2020). The principal concept of conducting RF seems to be a basic and amazing thing collective insight. With predictive analytics, the interpretation which the RF functions admirably is innumerable typically uncorrelated models (trees) act as a board of trustees which outperforms each of the individual material methods.

### **3.4.2 K-Nearest Neighbors**

K-Nearest Neighbors is a case-based learning technique, which saves all the preparation information for order. Being a sluggish learning strategy precludes it in numerous applications, for example, dynamic web digging for an enormous archive. One approach to improve its effectiveness is to discover a few delegates to address the entire preparing information for grouping. Building an inductive taking in model from the preparation dataset and utilizing this model (representatives) for order. (Guo, Gongde & Wang, Hui & Bell, David & Bi, Yaxin., 2004). The KNN arrangement was created from the need to perform discriminant examination when stable modulation appraisals of sample values are misleading or difficult to decide (Leif E. Peterson, 2009).

### **3.4.3 Logistic Regression**

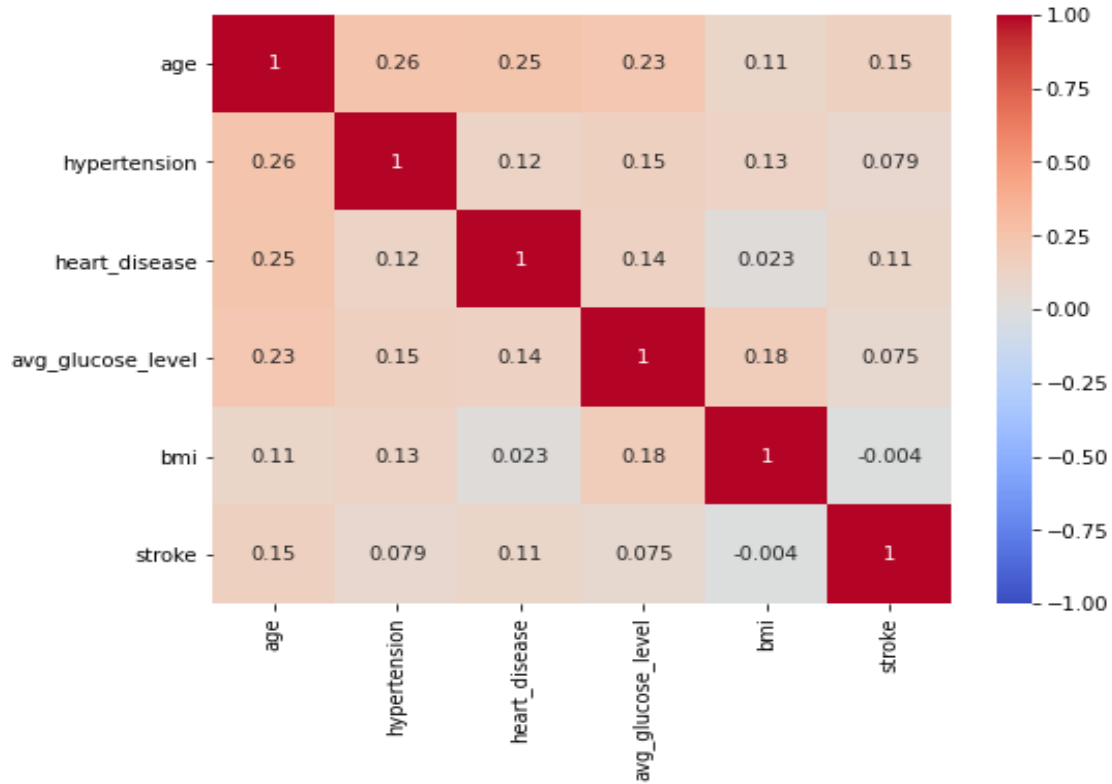
When the dependent variable is dichotomous, logistic regression is the proper regression strategy to use (binary). The logistic regression, like other regression studies, is a predictive analysis. A true model is LR, at its core, relies on a defined capacity to demonstrate a paired dependent variable, but it may be extended in a variety of ways. (FRED C. PAMPAL, 2000). In regression study, LR is assessing the actual boundaries of a strategic method. A parallel logistic model features a dependent variable with two alternative characteristics, similar to the true/false up short catered to by a marker variable, where the two qualities are designated '0' and '1' in mathematics (Garg, Shruti., 2019). The calculated sigmoid capacity is frequently indicated as m:

$$m = \frac{1}{1+e^{-n}} \quad (1)$$

Where m is the target value that is dependent on the selected values n. And if the performance is greater than 0.5, the result is 1, otherwise it is 0.

#### **CHAPTER 4: RESULT AND DISCUSSION**

Correlation is utilized to assess the strength of the link between two parameters. Coefficient of correlation quantifies overall degree of similarity between the variables (Senthilnathan, Samithamby, 2019). In most applications, two correlation coefficients are applied: Pearson's Product Moment Correlation Coefficient and Spearman's Rank Correlation Coefficient. In this study focuses on how Pearson's Simple Linear Correlation is used to investigate the relationship between variables. If indeed the concept of one attribute seems to be favorable as well as pretty identical to either the concept of some other attribute, there seems to be a potential of positive connection between them, which can produce a good pearson correlation; but even if concept of one parameter is optimistic while the trend of another variable is practically negative, there is a potential that the two variables have a negative relation with each other (Senthilnathan, Samithamby, 2019).In Figure 4.1 , the highest notable correlation is 0.26, which is between age and bmi .



**Fig. 4.1.** Correlation matrix for stroke prediction dataset attributes which shows the correlated value between the featured of cerebral stroke prediction dataset.

Here the stroke boundary has been used as a predictor parameter, while the remainder including its boundaries has been used as response factors. Stroke boundary only accepts double qualities, where 0 addresses non-stroke and 1 addresses stroke. To train the dataset, the whole dataset has separated in the proportion of 80:20 for the training set as well as test set individually. Three ML algorithms RF, KNN, LR has implemented to the training data to anticipate the test data outcomes giving throughout the confusion matrix as demonstrated in Table 1. For performance evaluation, some formulas has followed (Garg, Shruti., 2019).

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN} \quad [2]$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad [3]$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad [4]$$

$$\text{F1 score} = 2 * \frac{(\text{Recall} * \text{Precision})}{(\text{Recall} + \text{Precision})} \quad [5]$$

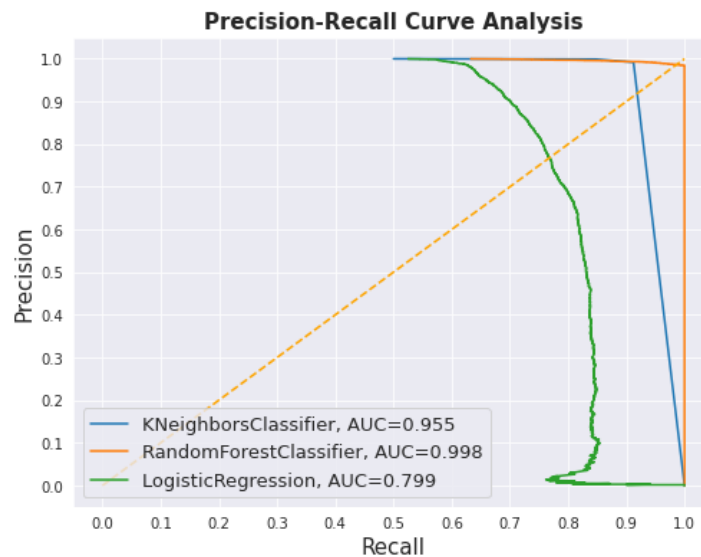
**Table 2.** Confusion matrix which describes the TP, TN,FP, FN of all the three Machine learning Techniques.

Random Forest	KNN	Logistic Regression
[[5702 3] [ 91 5614]]	[[5161 544] [ 46 5659]]	[[4225 1480] [1129 4576]]

**Table 3.** Values of Precision, Recall, F1- Score of all the three Machine Learning Techniques.

Name	Random Forest	KNN	Logistic Regression
Precision	0.99	0.95	0.77
Recall	0.99	0.95	0.77
F1- Score	0.99	0.95	0.77

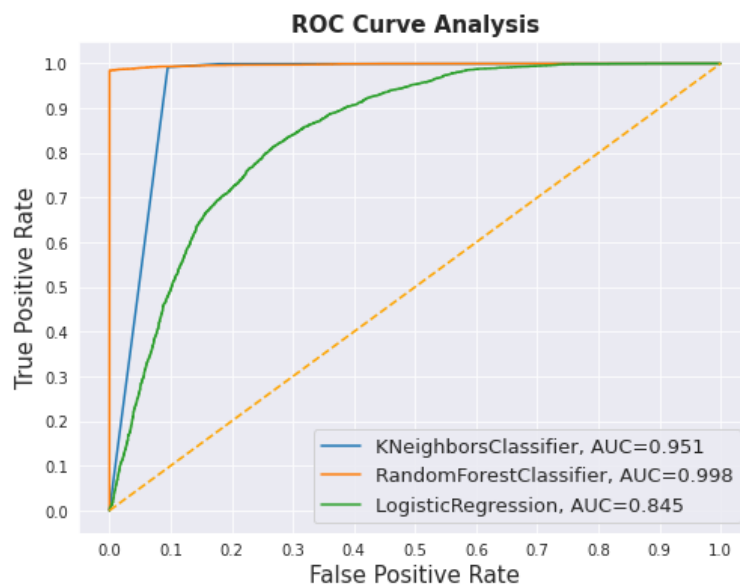
Table 1 and Table 2 describes the confusion matrix value (TP,TN,FP,FN) and different measure for three machine learning techniques, which is used for the prediction of cerebral stroke.



**Fig. 4.2.** Precision-Recall curve for machine learning techniques which describes the best techniques for cerebral stroke prediction.

Here also examine the utilization of space under the receiver operating characteristic (ROC) curve (AUC) and Precision- Recall curve as a performance appraisal for machine learning techniques. ROC analysis is used in clinical epidemiology to quantify how accurately medical diagnostic tests (or systems) can discriminate between two states, typically referred to as "diseased" and "non diseased". An excellent model has AUC near to the 1 which means it has a good measure of separability. A poor model has AUC near to the 0 which means it has the worst measure of separability. In Fig. 4.2 Precision- Recall curve analysis shows that RF gives AUC= 0.998, KNN gives AUC= 0.955, LR gives AUC =0.799 score. Also in 4.3 ROC curve analysis shows that RF gives AUC= 0.998, KNN gives AUC= 0.951, LR gives AUC =0.845 score.

For predicting cerebral stroke, Random Forest gives better results according this research .By following this process, cerebral stroke can be predicted better with Random Forest ML techniques.



**Fig. 4.3.** ROC curve for ML techniques which describes the best techniques for cerebral stroke prediction.

## CHAPTER 5: CONCLUSION

In this research paper, we have introduced a machine learning proceed toward consolidating the components of missing information Cleaning, Balancing Data, Encoding absolute information to numeric information, Hyperparameter Tuning, and execution assessment. Here we use to prepare dataset RF, KNN, LR algorithm. as we have seen in this research, the RF model performs very well for the cerebral stroke forecast dataset .Contrasted with past investigations, our ML classifiers present generally elite (surmised 0.99 AUC) in the prediction of stroke. There are a few components viewed as causally connected with our better. The principal factor is the information amount which is significant for the ML algorithms. To optimize its model, supervised ML requires a large amount of labeled data. With 43,400 samples, our dataset is significantly larger than the study mentioned above. The second factor is the quality of the data. The inherent requirement for large training datasets may have an impact on the accuracy of ML algorithms in studies. Poor data quality, such as incorrect labeling or conflicting data, can lead to machine learning errors. We used a data removal process to prevent this. We hope that this paper will encourage the utilization of machine learning methods in the analysis of healthcare data.



## REFERENCES

- [1] R. M. N. N. Rajora M., "Stroke Prediction Using Machine Learning in a Distributed Environment.," *Springer*, 2021.
- [2] G. A. L.-J. D. A. R. B. J. B. T. C. M. D. S. d. S. G. F. E. F. C. F. H. G. C. G. K. H. S. H. J. H. P. H. V. K. B. K. S. L. D. L. J. L. L. Roger VL, "American Heart Association Statistics Committee and Stroke Statistics Subcommittee," *Heart disease and stroke statistics--2011 update: a report from the American Heart Association*, 2011.
- [3] Liu Y, Yin B, Cong Y. , "The Probability of Ischaemic Stroke Prediction with a Multi-Neural-Network Model," *Sensors* , 2020.
- [4] Gibson CL. , "Cerebral Ischemic Stroke: is Gender Important?," *Journal of Cerebral Blood Flow & Metabolism*, 2013.
- [5] C. -C. Peng, S. -H. Wang, S. -J. Liu, Y. -K. Yang and B. -H. Liao, "Artificial Neural Network Application to the Stroke Prediction," *IEEE 2nd Eurasia Conference on Biomedical Engineering, Healthcare and Sustainability (ECBIOS)*, 2020.
- [6] Liu T, Fan W, Wu C., "A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset.," *Artif Intell Med*, 2019.
- [7] M. S. Singh and P. Choudhary , "Stroke prediction using artificial intelligence," *IEMECON*, 2017.
- [8] R. S. Jeena and S. Kumar, "Stroke prediction using SVM," *ICCICCT*, 2016.
- [9] Selma Yahiya Adam, Adil Yousif and Mohammed Bakri Bashir, "Classification of Ischemic Stroke using Machine Learning Algorithms," *International Journal of Computer Applications* , 2016.
- [10] M, Mark, "Replication Data for: Prediction of Cerebral Stroke," *Harvard Dataverse*, 2021.
- [11] G. Carey, " Coding Categorical Variables," *PSYC5741*, 2003.
- [12] Kedar Potdar, Taher S Pardawala and Chinmay D Pai, " A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers," *International Journal of Computer Applications* , 2017.
- [13] T. Maciejewski and J. Stefanowski, "Local neighbourhood extension of SMOTE for mining imbalanced data," *IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, 2011.
- [14] Probst, P., A. Boulesteix and B. Bischl, "Tunability: Importance of Hyperparameters of Machine Learning Algorithms," *J. Mach. Learn. Res.*, 2019.
- [15] D. C. R. T. P. K. S.-H. L. R. M. P. Steven R. Young, "Optimizing deep learning hyper-parameters through an evolutionary algorithm," *Workshop on Machine Learning in High-Performance Computing Environments*, 2015.
- [16] Putatunda, Sayan & Kiran R, Dr. , "A Modified Bayesian Optimization based Hyper-Parameter Tuning Approach for Extreme Gradient Boosting," 2020.
- [17] Panda, Balaram, "Hyperparameter Tuning," 2019.
- [18] J. Bergstra and Y. Bengio, "Random Search for Hyper-Parameter Optimization," *Journal of Machine Learning Research*, 2012.
- [19] Revanth S, Sanjay S, Sanjay N, Vijayaganth V, "Stroke Prediction using Machine Learning Algorithms," *International Journal of Disaster Recovery and Business Continuity*, 2020.

- [20] Guo, Gongde & Wang, Hui & Bell, David & Bi, Yaxin., "KNN Model-Based Approach in Classification," 2004.
- [21] Leif E. Peterson, "K-nearest neighbor," *Scholarpedia*, 2009.
- [22] FRED C. PAMPAL, "LOGISTIC REGRESSION: A Primer," 2000.
- [23] Garg, Shruti., "Prediction of Type 2 Diabetes using Machine Learning Classification Methods," *Procedia Computer Science*, 2019.
- [24] Senthilnathan, Samithamby, "Usefulness of Correlation Analysis," *SSRN Electronic Journal*, 2019.