Birmingham City University

SCHOOL OF COMPUTING AND DIGITAL TECHNOLOGY

MSc Advanced Computer Science

**STROKE PREDICTION**

**Advanced Data Science - CMP7167**

**Module Coordinator**

**Nouh Sabri Elmitwally**

**Submitted by,**

**Mehak Qasim – 21159589**

**DATE: 9 May, 2022**

**Word Count: 3000**

**TABLE OF CONTENT**

Abstract………………………………………………………………………………………………3

Keywords……………………………………………………………………………………………3

Introduction. ……………………………………………………………………………………..4

Background…………………………………………………………………………………………4

Aims and Objectives…………………………………………………………………………….5

Dataset Description…………..…………………………………………………………………6

Questions to be addressed…………………………………………………………………..7

Machine Learning Model…………………………………………………………..…………8

Accessing and Importing the Dataset…………………………………….…….8

Data Preprocessing……………………………………………………………………..9

Data Visualization and EDA…………………………………….………………….13

Training Model…………………………………………………………….……………18

Applying K-Fold Cross Validation……………………………………………….21

Conclusion: ………………………………………………………………………………………..21

Reference:………………………………………………………………………………………….22

**ABSTRACT:**

A medical disorder which effects the blood arteries of brain and makes them cracked is referred to as Stroke in medical terminology. It normally happens because of the interruption in the blood flow to the brain or because of the supply of other nutrients to the blood being disturbed. When this all happens, there are very high chances that symptoms will be developed. Stroke is one of the major cause for Disability and death around the globe as per World Health Organization (WHO). If symptoms are recognized at the early stage it can help to reduce the intensity of Stroke for sure. To understand and predict the chances of having stroke in the brain, so many different types of Machine Learning models have already been developed. In this study we have used different types of general and medical parameters and machine learning algorithms, name as Decision Tree (DT) Classification GaussianNaiveBayes, Random Forest (RF) Classification, LGBMClassifier, and KNeighbor. These all are used to train the models for trustworthy and reliable prediction. Among all of these algorithms, KNeighbor had an accuracy of around 95.45 percent and it is the top performing algorithm for this job. We have taken Open-access Stroke Prediction dataset for development of the method. If we compare the accuracy of our models which we have used in this task with the accuracy of the previous studies, this study models accuracy is higher than the previously performed models which shows that models which we have used in this investigation are more dependable.

**Keywords:**

Disorder, Stroke, Prediction, Machine Learning, Algorithm, neurological damage, chronically disabled, pandemic,Stroke survivors, supervised learning.

**INTRODUCTION:**

We all are very well aware of how dangerous and deadly Stroke is and it becomes more deadly especially when you don’t have quick access to first aid. Stroke can cause anyone serious health issues or even it could become a reason of someone’s death. Knowing this, we will try to minimize the risk of getting stroke and this research is going to help people know the reason behind having strokes which will enable them to do the necessary measures and take steps that could help them not to get stroke and reducing the chance of disabilities and serious health condition which a person could face after having stroke or even causing the death.

There are so many different kinds of Strokes but the main reason stroke occurs is because of the Blood blockage or bleed in the brain and it could cause the person neurological damage, health complications for the rest of life or even death in some cases.

Stroke is something which could happen to anyone and anytime regardless of their age but definitely it’s very much because of their health condition. And health condition is something which we can easily understand by different methods and procedures. So, we are going to know the relation between health condition and having stroke and this will help us understand how we can reduce the risk of having those health conditions and avoid having stroke. Through this research we will try to minimize the chances of having strokes as much as possible.

**BACKGROUND:**

There is a very high chance that many of us must might know someone around us who is affected by stroke. Stroke is one of leading causes of death globally and it stands on the second position for the cause of death. Around 50% of the survivors from stroke are chronically disabled and struggle in having a normal life after stroke. This can cause real difficulties in someone’s life and it could be very challenging for anyone to get back to normal life after having a stroke, also it creates an impact on your immediate family because you struggle with so many things afterwards and you need to be very careful to avoid any further stroke as it could be more dangerous than the one you had before.

In the Recent years especially after the rise of Covid, healthcare system has been under so much attention as whole world faced the challenges caused because of COVID-19. Rapid Technological innovations and advancements have played a vital role in justifying the effects of the pandemic.

The main challenge which is faced is the consideration of medtech advancements and innovations as very expensive and difficult to apply. So, it is very important that the one making decisions should make sure that which technology will be beneficial for operations, heath results and they have a long lasting effect.

One specific dominant example is stroke diagnosis and care system application because stroke is considered as the World’s biggest killers and it’s also considered to be a leading reason for serious long-term disability in the United States which results in spending billions of dollars per year on it. So we are doing to study the reasons behind the stroke and health conditions which are likely to result in stroke and it will help us create a system where we can identify risk for potential stroke. This will definitely help us preventing people from having stroke which can result in permanent disabilities if the person survives the stroke and sometimes if it gets worse it can cause death as well.

**AIM AND OBJECTIVE:**

The Aim for this project is to calculate and identify the risk of having stroke and to make sure that necessary steps are taken to reduce the chances of having Stroke and saving the Human lives or the difficulties that are expected to be faced by the Stroke survivors which make their life very difficult. This research will help in understanding the situation of any specific human body and how likely they have chances of having stroke, this will be done by taking data at multiple times at a set time interval for the readings.

In this project we will collect the data set from the Body of any human being and then analyzing the data set by using machine learning algorithm and predicting the chances of having stroke or not. Results will be based on the collected data set from the human body and it will only show if there is a chance or not and not the level of having chances but just Yes or No. This result will be enough for a person to know if they need to take any action or medication to avoid having stroke or they are fine without any actions to be made.

Date collected to predict the chances of having stroke will also consider the below all points to give as accurate results as possible.

* How strong and healthy a body is to survive the abnormal health conditions
* Is there any other disease which is already impacting the body and making it week and can result in multiplying the factors which cause the Stroke
* How much mentally strong a person is to absorb the stress which could be caused because of any unwanted/abnormal condition
* Is the body underweight with respect to its height and gender and considering other factors

**DATASET DESCRIPTION:**

The dataset for this project is taken from kaggle website. This dataset has 11 clinical features and the data contains 5110 observations with 12 attributes in total. There are **8** categorical variables and **3** numerical variables in the dataset. This dataset is used to predict whether a patient is likely to get stroke based on the input parameters like married or not, gender, various diseases, bmi, age, work type and smoking status. Each row in the data provides the person’s or patient’s information.

Following are the features in dataset:

1. id: unique identifier
2. gender: "Male", "Female" or "Other"
3. age: age of the patient
4. hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
5. heart\_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
6. ever\_married: "No" or "Yes"
7. work\_type: "children", "Govt\_jov", "Never\_worked", "Private" or "Self-employed"
8. Residence\_type: "Rural" or "Urban"
9. avg\_glucose\_level: average glucose level in blood
10. bmi: body mass index
11. smoking\_status: "formerly smoked", "never smoked", "smokes" or "Unknown"
12. stroke: 1 if the patient had a stroke or 0 if not \*Note: "Unknown" in smoking\_status means that the information is unavailable for this patient

**Questions to be addressed**

**The problem to tackle**

The following study will use different categorical and numerical features for detecting the estimated risk of Stroke in different age of people.

**Importance of predicting Stroke chances**

According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, approximately 11% of total deaths of different ages in a year is caused by Stroke. And in the present world it is a 3rd leading cause of death and disability together.

Annually, around 15 million people worldwide suffer a stroke. Of these, 5 million die and another 5 million are left permanently disabled, placing a burden on family and community.

**Role of Machine Learning in tackling this problem**

Early prediction of stroke by using artificial intelligence can decrease the chances of sudden stroke because treatments can be done for making the situation better.

**Type of machine learning task to be implemented**

This machine learning task is a classification problem which requires classification algorithm to solve the problem because in this dataset we have categorical data (whether the person having chances of stroke or not) thus the target can be 0 or 1 only.

**Importance of Pre-Processing Stage**

When the dataset has different types of anomalies like null values (missing data), imbalance data, useless columns (which are not required for the target), values in different scales, then pre-processing stage is necessary.

**Machine Learning Model:**

There are two different types of Machine Leaning techniques which are supervised learning and unsupervised learning. For this project we will be using supervised learning technique because the dataset for this project is based on known inputs and outputs. The purpose of this machine learning model is to predict the output value which is “Stroke”, by training the model using known inputs like gender, age, heart disease, married or not, bmi, smoking status etc.

Multiple models like SVM, Decision Tree Classifier, Random Forest Classifier, KNeighbors Classifier, LGBM Classifier, XGB Classifier have been used for training the dataset and getting the high accuracy.

**Accessing and Importing the Dataset**

For gathering data or creating a dataframe we are using Panda’s library, dataframe consists of a table that contains data where we can load our dataset into.



Figure 1. Importing dataset into dataframe

Once our dataset is imported into dataframe then we can have access of this dataset. We can get data of all columns using pd.options.display.max\_columns = None

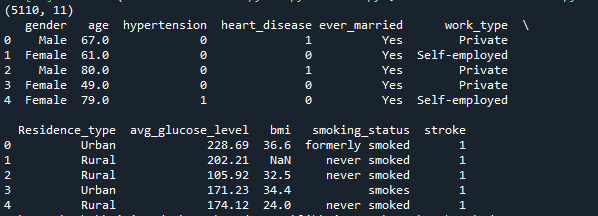


Figure 2. All columns of dataset

For getting statistical analysis of the dataset we can use different functions like describe and shape from Panda library:

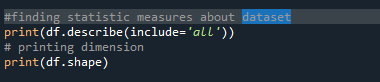


Fig. 3 viewing dataset info

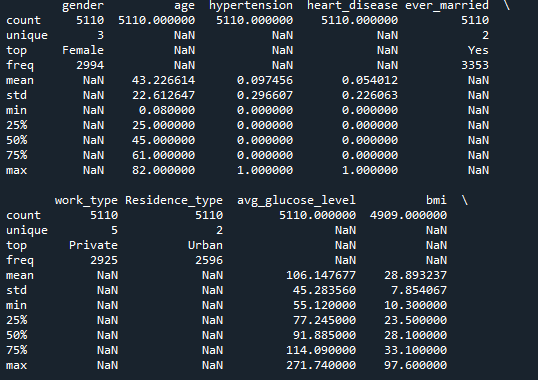
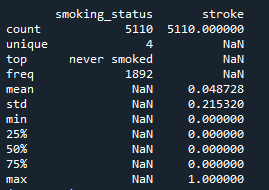


Fig.4 Output of df.describe()



Fig.5 Output of df.shape()

**Data pre-processing**

In order to get the meaningful understanding from the data, Data preprocessing in Machine Leaning is used and it is a very important process which helps to improve the data quality resulting in better, meaningful and understandable information from the data. Data preprocessing stage in Machine Learning is a process which directs towards the method of making (Organizing and Filtering) raw data into a meaningful data to make it ready for constructing and training Machine Learning models. If we make it short and brief to understand the data preprocessing method, it is data mining technique to change raw data into a reasonable, readable and clear presentation.

Data Preprocessing helps with everything that might stop model to function efficiently. After the relevant dataset collection, data needs to be filtered and organized for model development. In the very start of the process, column Id is lost as it has no impact on model construction. Then the data is examined to find and fill the null values if any detected. String Literals in the Dataset are converted into integer values using the Label encoding so computer can understand. Because computer can normally understand numbers, strings needs to be converted into integers. There are five columns for the data type string in the collected dataset. While doing Label encoding, all the strings encoded and complete dataset is converted into numbers. The data is very imbalanced which is used for stroke prediction. Total number of rows are 5110, out of these rows 249 rows indicate the stroke possibility and 4861 confirms the stroke absence.

After completion of data preparation and organizing the imbalanced dataset, next stage will be constructing the model. The data needs to be divided into two different roles, one is training and other one is testing. 80 Percent data will be for training and 20 percent will be for testing. After we split the data, variety of classification methods are used to train the model. In this study we have used classification algorithms named as Decision Tree classification method, Random Forest, Gaussian Naïve Bayes, KNeighbors and LGBMClassifier.

**Dropping Column**

In order to prepare data for training we have to preprocess it. As we can see in fig 2 we have a column “id” which is not useful for training the model so we will drop it using the following code.

df.drop("id", axis=1, inplace=True)

Also remove the “Other” value from the column “gender”. So that we will have only 2 values Male and Female.

df.drop(df[df.gender == 'Other'].index, axis = 0, inplace=True)

**Dummy Encoding**

After analyzing the dataframe I found out that columns like gender, ever\_married and residence type have categorical data so I need to do dummy encoding (transform data into binary variables)

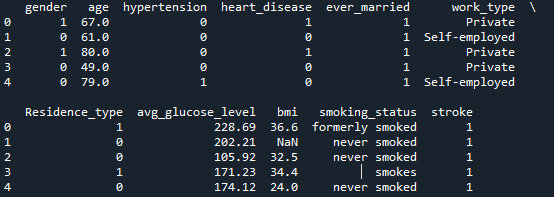


Figure 6. After dummy encoding

**Frequency Encoding**

We have two columns work\_type and smoking\_status with different values so I have to do frequency encoding.

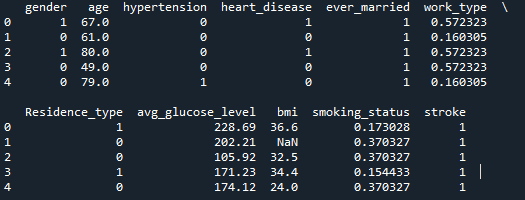


Figure 7. After frequency encoding

**Replacing null value with mean value**

Next step is to find if any column have any null or missing value in it then replace that value with the mean value.

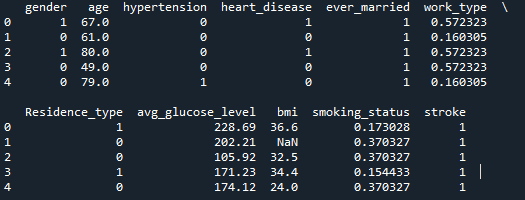


Figure 8. bmi has a null value

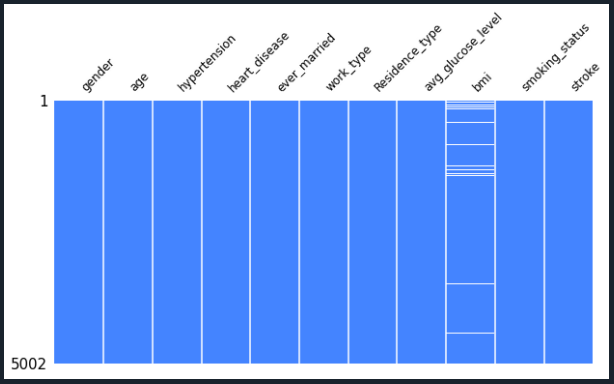


Fig 9. Null value in bmi

Figure 4 showing that bmi column have nul or missing values. So I have to find the mean and fill these null values with the mean value.

df=df.fillna(np.mean(df['bmi']))

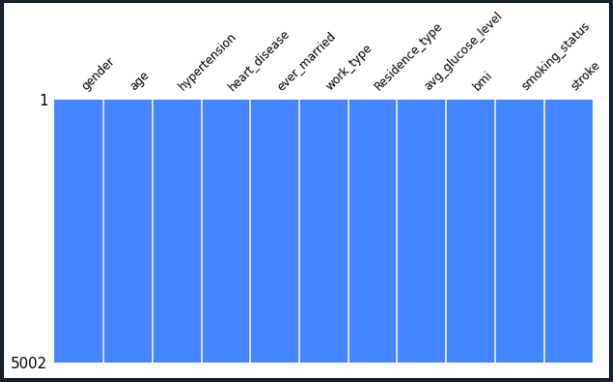
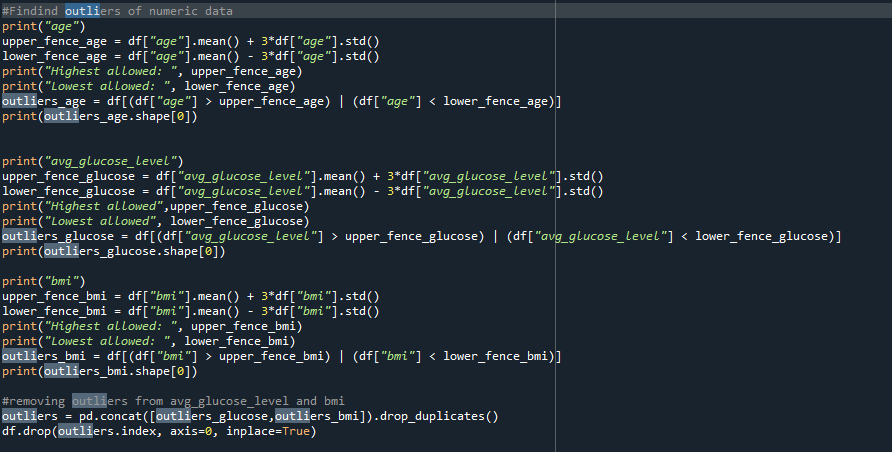


Fig 10. Null value removed

Above bar plot showing that now bmi has no missing or null value

**Outliers**

There are columns avg\_glucose\_level and bmi in our dataframe which have numeric data and these numeric data have some values which are noticeably different form the other values. So, I will have to find outliers in these columns using Z-Score and then remove them.



**Data visualization and EDA**

Count of all categorical features ("gender", "ever\_married", "work\_type", "Residence\_type", "smoking\_status")

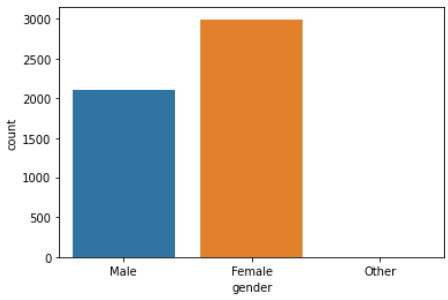


Fig 11. Count plot for Gender

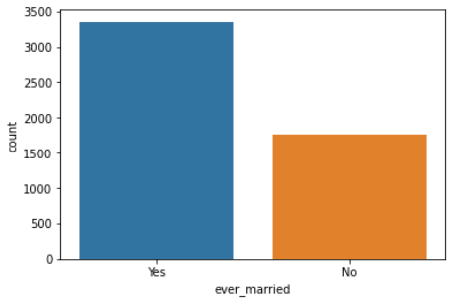


Fig 12. Count plot for ever\_married

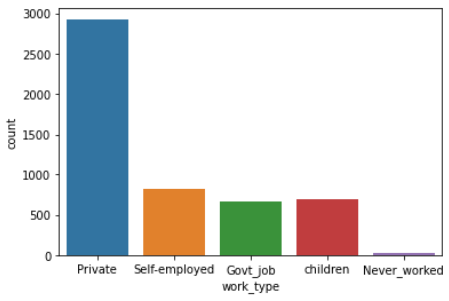


Fig 13. Count plot for work\_type

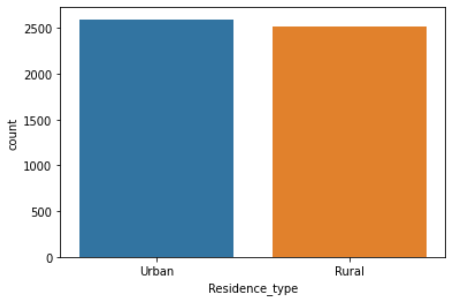


Fig 14. Count plot for Residence\_type

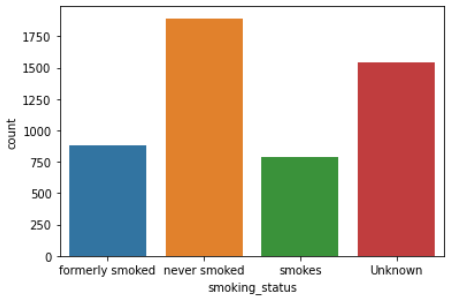


Fig 15. Count plot for smoking\_status

**Visualizing the relationship between smoking\_status and stroke**

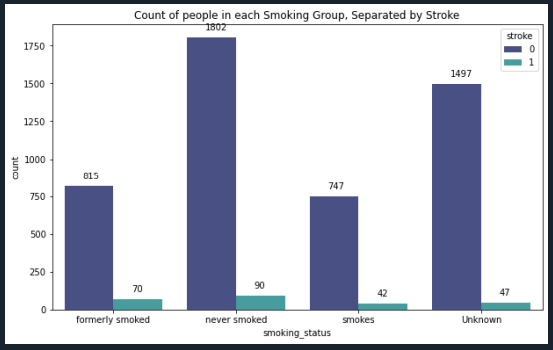


Fig. 16

**Visualizing the relationship between work type and stroke**

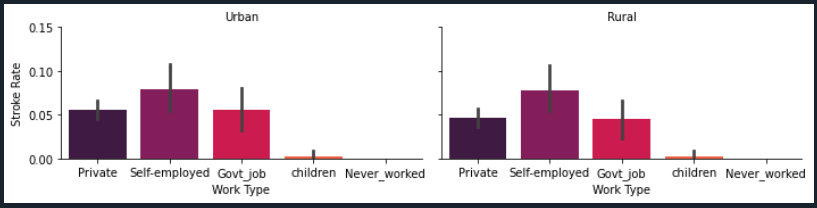


Fig. 17

**Correlation between variables**

To visualize the relationship between features, heatmap can be implemented which we can also called as correlation matrix.

Heatmap is showing that:

* 'age' and 'bmi' are correlated fairly high. This shows that the bmi and age are directly proportional to each other means if age is increasing then the value of bmi is also increasing.
* 'age' and 'hypertension' are also correlated. This means that in our society elders or senior citizens mental wellbeing is not good and this is quite dangerous for the society.

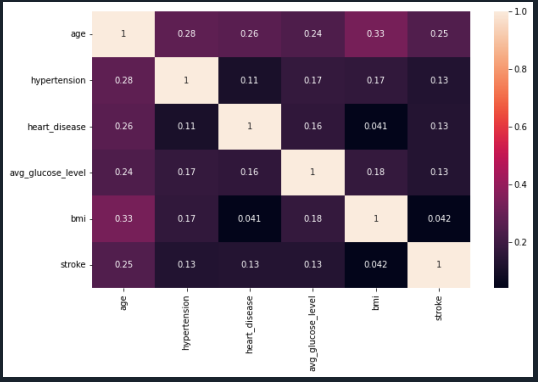


Fig. 18

**Training Model:**

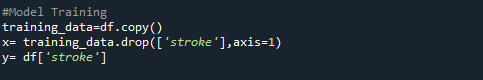


Fig. 19

### Splitting into Training and Test set

### 

Fig. 20

### Feature Scaling

### 

Fig. 21

### Training Model using Decision Tree Classifier

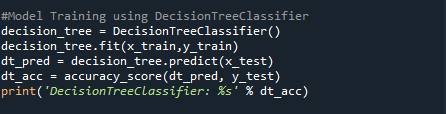


Fig. 22

**Training Model using Random Forest Classifier**

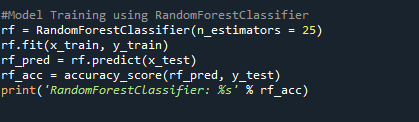


Fig. 23

**Training Model using K Neighbors Classifier**

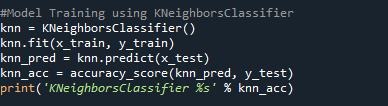


Fig. 24

**Training Model using LGBM Classifier**

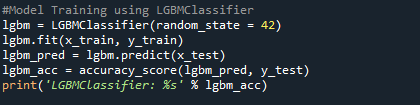


Fig. 25

**Training Model using XGB Classifier**

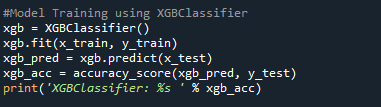


Fig. 26

**Training Model using Gaussian Naïve Bayes**

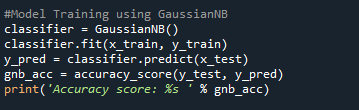


Fig. 27

**Comparing Accuracy of different Classifier Models**

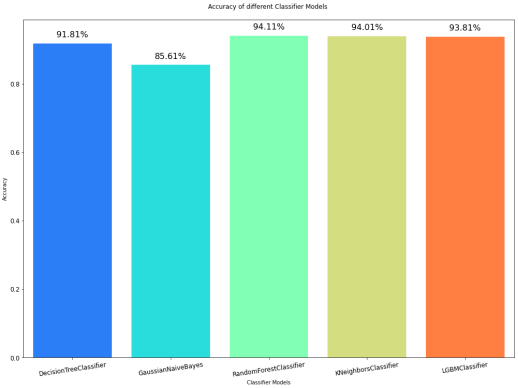


Fig. 28 visualizing different classifier model’s accuracy score

**Making confusion matrix**



Fig. 29



Fig. 30

**Applying K-Fold Cross Validation**

Moreover, cross validation is implemented to test and train model on different portions of data ‘k’ number of iterations. The goal is to flag problems like overfitting and selection bias by testing the model to predict new data that was not used in estimating it (A. Ramezan, A. Warner and E. Maxwell, 2019).

There are many techniques for cross-validation to test models ability to predict unseen data like K-Fold Cross Validation. In order to implement cross validation k-fold method, first we need to import its relevant library.

**Tuning KNeighbors Classifier with Cross Validation**

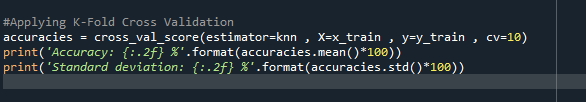


Fig. 31



Fig. 32

**Conclusion:**

Stroke is very dangerous and deadly medical disease which could happen to anyone and anytime if they don’t take good care of themselves and ignore the symptoms. Stroke is a disease which should be treated at earliest possible without any delay to avoid any ongoing complications. Even if someone survives a stroke, there are very high chances of having disabilities for the rest of life if not treated properly. Through the Development of Machine Learning Model, it could help in detecting stroke at early stage. In this study we have used different physiological variables for the investigation to increase the effectiveness of Machine Learning Algorithm will help in predicting stroke. Among all of the Testing Methods, KNeighbors Classifier has the highest accuracy with 95.45% and is considered as most reliable to predict stroke. When using Cross-Validation metrics for forecasting brain stroke, KNeighbors Classifier outperformed the other processes. Future implications for this study is to use larger dataset and advance machine learning models. In return of providing basic information, it will help general public in knowing the chances of having stroke in a patient. In best case scenario, it would benefit the patients in getting early or pre-treatment for stokes and recover their health issues after the stroke.

**References**

Ramezan, C., A. Warner, T. and E. Maxwell, A. (2019) ‘Evaluation of Sampling and CrossValidation Tuning Strategies for Regional-Scale Machine Learning Classification’, Remote Sensing, 11(2), p. 185. doi:10.3390/rs11020185.

Li, J. (2017) ‘Data Visualisation’, Information Systems, p. 14.

WHO, “The top 10 causes of death.” https://www.who.int/newsroom/fact-sheets/detail/the-top-10-causes-of-death (accessed Jun. 09, 2021).

A. Dey, “Machine Learning Algorithms: A Review,” Int. J. Comput. Sci. Inf. Technol., vol. 7, no. 3, pp. 1174–1179, 2016, [Online]. Available: [www.ijcsit.com](http://www.ijcsit.com).

C Sundstro 'm, Machine learning algorithms for stroke diagnostics, 2014.