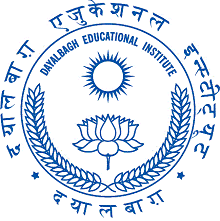
**DAYALBAGH EDUCATIONAL INSTITUTE**

****

**Product Manufacturing Project**

**Report (EGC381)**

**BRAIN STROKE PREDICTION**

**Submitted by: Submitted to:**

**Arpita Satsangi(2104225) Dr. A. Charan Kumari**

**B. Tech 3rd year**

**Branch –Electrical**

CERTIFICATE

This is to certify that the project work entitled is a bonafide work of the following student, under my guidance.

Arpita Satsangi 2104225

**SUPERVISOR:**

**Dr. A. Charan Kumari**

Electrical Engineering Department

Faculty of Engineering

Dayalbagh Educational Institute

Dayalbagh, Agra-282005

Signature of the Supervisor:

Date:

**ACKNOWLEDEMENTS**

I would like to express my gratitude to the project supervisor

**Dr. A. Charan Kumari**, Department of Electrical Engineering, Faculty of Engineering, Dayalbagh Educational Institute (D.E.I), Agra for her pain staking and whole hearted efforts and the valuable guidance at each stage of the development of this project. She is very supportive, highly concerned and accessible and has spent a lot of her time in fruitful discussions and in reviewing the work periodically. Her constant encouragement and zealous support made it possible to complete the work.

Also, our institution Dayalbagh Educational Institute was very generous for providing us this opportunity of working on this project. We would also like to thank our families and friends for their extreme support and belief in us at all points of life.

**ABSTRACT**

Stroke is a medical disorder in which the blood arteries in the brain are ruptured, causing damage to the brain. When the supply of blood and other nutrients to the brain is interrupted, symptoms might develop. According to the World Health Organization (WHO), stroke is the greatest cause of death and disability globally. Early recognition of the various warning signs of a stroke can help reduce the severity of the stroke. Different machine learning (ML) models have been developed to predict the likelihood of a stroke occurring in the brain.

The dataset used in the development of the method was the open-access Stroke Prediction dataset. It is used to predict whether a patient is likely to get stroke based on the input parameters like age, various diseases, bmi, average glucose level and smoking status.

K-nearest neighbor and random forest algorithm are used in the dataset. The accuracy of both the algorithms are 95.62% and 95.21% respectively.

**TABLE OF CONTENT**

**CERTIFICATE**…………………………………………………….....02

**ACKNOWLEDGEMENT**...…....…………………………………....03

**ABSTRACT**…………………………...….…...……………………....04

**1.Introduction** .....……………..................…………….….…….........06

**2.The dataset**.....................…..................................……………….…..08

**3.Technology used**….…........................……...…..……………….….10

**4.Methodology**….....................................................……………….….12

**5. Implemented Algorithms**….………...…..…………………...........14

**6.** **Experimental Results**……..................................……………….....17

**7.Conclusion**……....................................................……………….….17

**REFERENCES**……...……………………...…………………...…....18

**1. INTRODUCTION**

Stroke occurs when the blood flow to various areas of the brain is disrupted or diminished, resulting in the cells in those areas of the brain not receiving the nutrients and oxygen they require and dying. A stroke is a medical emergency that requires urgent medical attention. Early detection and appropriate management are required to prevent further damage to the affected area of the brain and other complications in other parts of the body.

According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths.

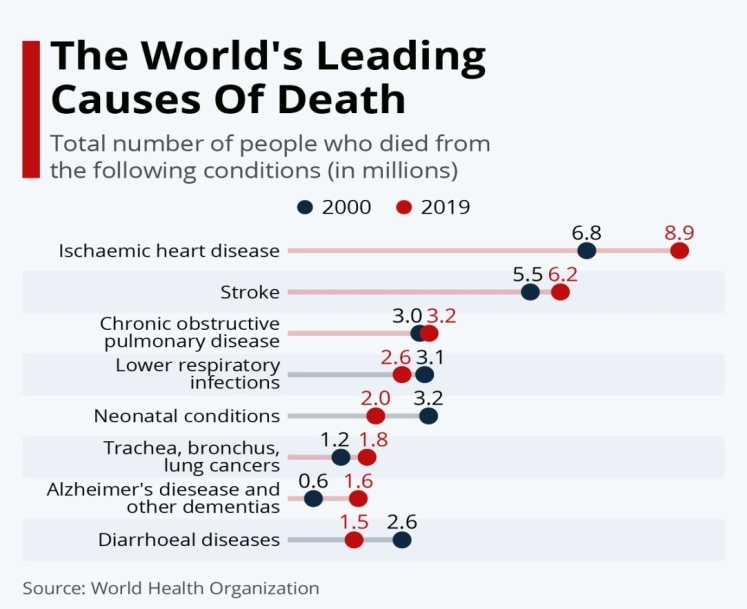


Fig: The world's leading causes of death (source: WHO)

**Brain Stroke** happens when there is a blockage in the blood circulation in the brain or when a blood vessel in the brain breaks and leaks. The burst or blockage prevents blood and oxygen reaching the brain tissue. Without oxygen the tissues and cells in the brain are damaged and die in no time leading to many symptoms.  
  
Once brain cells die, they generally do not regenerate and devastating damage may occur, sometimes resulting in physical, cognitive and mental disabilities. It is crucial that proper blood flow and oxygen be restored to the brain as soon as possible.  
  
 Worldwide, brain stroke is the second leading cause of death and third leading cause of disability. In some cases, the warning signs of a stroke can be obvious but what’s going on inside the body is incredibly complex. 80% of strokes are preventable. But once you’ve had a stroke, the chances you have another one are greater.

Stroke may be avoided by leading a healthy and balanced lifestyle that includes abstaining from unhealthy behaviors, such as smoking and drinking, keeping a healthy body mass index (BMI) and an average glucose level, and maintaining an excellent heart and kidney function. Stroke prediction is essential and must be treated promptly to avoid irreversible damage or death. With the development of technology in the medical sector, it is now possible to anticipate the onset of a stroke by utilizing ML techniques. The algorithms included in ML are beneficial as they allow for accurate prediction and proper analysis.

**2. THE DATASET**

**Attribute Information about the 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(high blood pressure)  
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", "Government job", "Never worked", "Private" or "Self-employed"  
8) Residence type: "Rural" or "Urban"  
9) Average glucose level: average glucose level in blood  
10) BMI: body mass index  
11) Smoking status: "formerly smoked", "never smoked", "smokes" or "Unknown"(Note: "Unknown" means that the information is unavailable for this patient)  
12) Stroke: 1 if the patient had a stroke or 0 if not

The stroke prediction dataset was used to perform the study. The data contains 5110 observations with 12 attributes. This dataset is used to predict whether a patient is likely to get stroke based on the input parameters like gender, age, various diseases, and smoking status. Each row in the data provides relevant information about the patient.

The output column 'stroke' has the value as either '1' or '0'. The value '0' indicates no stroke risk detected, whereas the value '1' indicates a possible risk of stroke. This dataset is highly imbalanced as the possibility of '0' in the output column ('stroke') outweighs that of '1' in the same column. Only 249 rows have the value '1' whereas 4861 rows with the value '0' in the stroke column. For better accuracy, data pre-processing is performed to balance the data.

**3. TECHNOLOGY USED**

**SOFTWARE**

**1. JUPYTER NOTEBOOK**- The Jupyter Notebook is an open source web application that you can use to create and share documents that contain live code, equations, visualizations, and text. Jupyter Notebook is maintained by the people at Project Jupyter .

**2. PYTHON**- Python is a high-level, interpreted, interactive and object-oriented scripting language. Python is designed to be highly readable. It uses English keywords frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages. Python can be used for developing complex scientific and numeric applications. Python is designed with features to facilitate data analysis and visualization.

**3. LIBRARIES-** The libraries used are as follow:

* **Pandas** is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language. It is built on top of the NumPy library. Pandas is fast and it has high performance & productivity for users.
* **NumPy** stands for Numerical Python. It is a Python library used for working with arrays. It also has functions for working in domain of linear algebra, fourier transform, and matrices. It is an open source project and you can use it freely.
* **SKlearn** stands for scikit-learn is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python.
* **Matplotlib** is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits

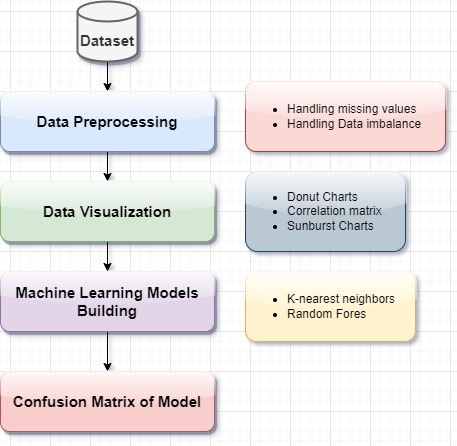
**4. METHODOLOGY**

Fig: Methodology

First the dataset is downloaded from the kaggle website. The necessary libraries like Pandas, NumPy, etc. are imported. Then the data is read using pandas.

Since there is no as such requirement for IDs hence its column is removed. By using the function dataframe.isnull().sum(), it returns the number of missing (non-zero) values in the dataset.

It is observed that only in one column of BMI, 201 values are missing. Also in gender only one value is stated as Other while the rest are only male or female, hence for simplicity it is removed from the dataframe.

Then Donut Charts, Correlation matrix, Sunburst charts are made to visualize the data.

After Data Preprocessing, the dataset is split into train and test data(train-3926 , test-982). A model is then built using this new data using two Classification Algorithms. Accuracy is calculated for all these algorithms and compared to get the best-trained model for prediction.

**5. IMPLEMENTED ALGORITHMS**

The most common disease identified in the medical field is stroke, which is on the rise year after year. Using the publicly accessible stroke prediction dataset, it measured two commonly used machine learning methods for predicting brain stroke recurrence, which are as follows:(i)Random forest (ii)K-Nearest neighbors.

**1) RANDOM FOREST:**

The classification algorithm chosen was RF classification. RFs are composed of numerous independent decision trees that were trained individually on a random sample of data. These trees are created during training, and the decision trees’ outputs are collected. A process termed voting is used to determine the final forecast made by this algorithm. Each DT in this method must vote for one of the two output classes (in this case, stroke or no stroke). The final prediction is determined by the RF method, which chooses the class with the most votes.

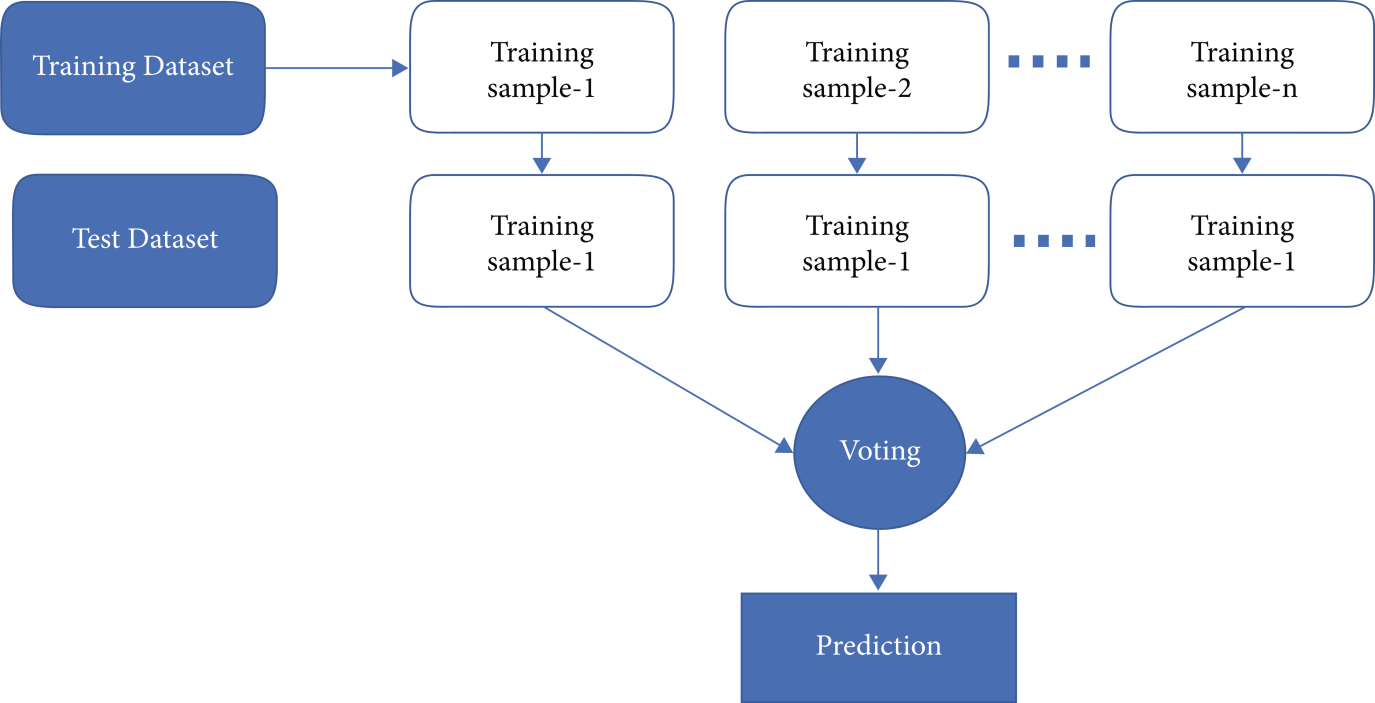


Fig: Random forest classification

**2) K-NEAREST NEIGHBORS:**

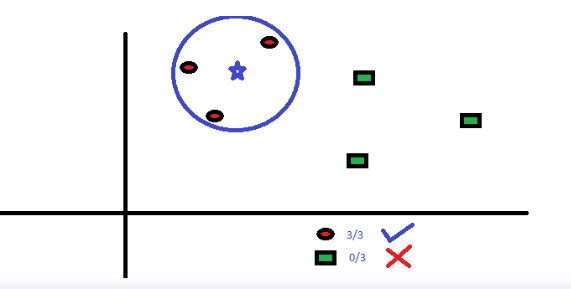
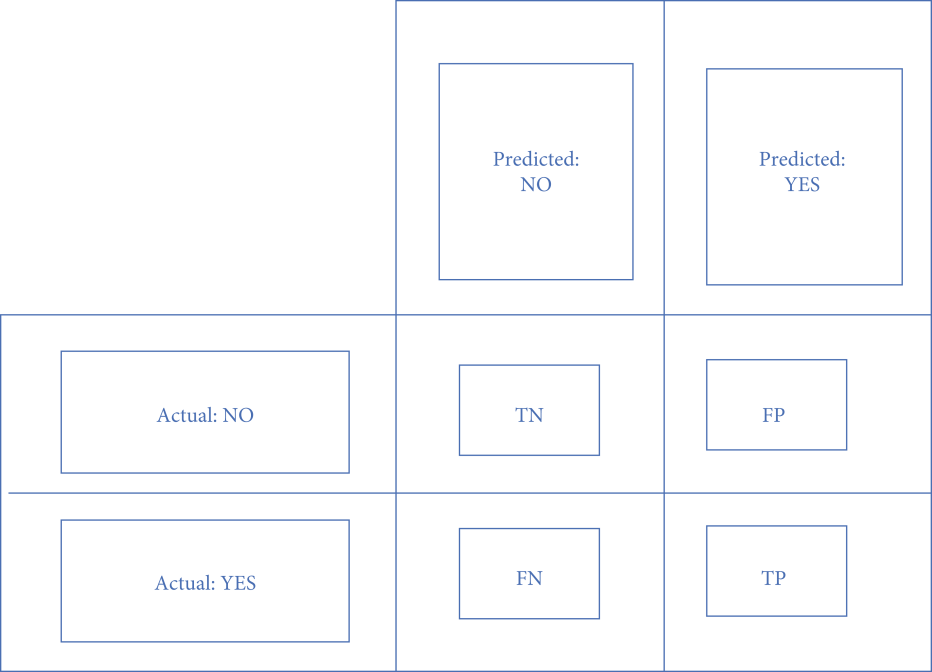
K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using it.

Fig: K-Nearest neighbors

It is also called a lazy learneralgorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

**CONFUSION MATRIX:**

Figure depicts the confusion matrix or evaluation matrix. The confusion matrix is a tool for evaluating the performance of machine learning classification algorithms. The confusion matrix has been used to test the efficiency of all models created. The confusion matrix illustrates how often our models forecast correctly and how often they estimate incorrectly. False positives and false negatives have been allocated to badly predicted values, whereas true positives and true negatives were assigned to properly anticipated values. The model’s accuracy, precision-recall trade-off, and AUC were utilized to assess its performance after grouping all predicted values in the matrix.

Fig: Confusion matrix

**EXPERIMENTAL RESULTS**

* The Accuracy of K Nearest Neighbors Classifier is 95.01%
* The Accuracy of Random Forest Classifier is 95.21%.

**CONCLUSION**

Stroke is a critical medical condition that should be treated before it worsens. Building a machine learning model can help in the early prediction of stroke and reduce the severe impact of the future. This model shows the performance of two machine learning algorithms in successfully predicting stroke based on multiple physiological attributes. Between the algorithms chosen, K Nearest neigbors Classification performs best with an accuracy of 95.62%.

This model suggests the implementation of various Machine learning algorithms on the dataset taken. This project can be further extended by training the model using Neural Networks. The comparison of the performance can be done by taking more algorithms and more accuracy metrics into consideration. This work is limited to textual data, which might not always be accurate for stroke prediction. Collecting a dataset consisting of images such as Brain CT scans to predict the possibility of stroke would be more efficient in the future.

**REFERENCES**

* Concept of Stroke by Healthline.
* Dataset named ‘Stroke Prediction Dataset’ from Kaggle:

https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8686476/