

STROKE PREDICTION USING MACHINE LEARNING

Mrs. Suchetha N V,

Abstract

Stroke is a clinical status where improper supply of blood to the brain cells or internal bleeding result to the death of a patient. This work mainly focuses on usage of machine learning approaches for the early prediction of stroke by considering attributes such as age, glucose level, body mass index, previous medical records and many more. A countable number of works have been carried out for predicting the same using many machine learning algorithms. In the proposed system the model prediction will be based on supervised algorithms such as Decision tree, Random Forest, and K-nearest neighbor.

Introduction

1.1 Project Introduction

Stroke is a kind of medical emergency which occurs because of a bleeding or blockage in blood vessels. This blockage in blood vessels might occur due to several reasons such as Stress, Hypertension, and many other factors. The chances of blood vessel blockage can be much observed much more in age old people. Senior individuals need extra attention because it is more deadly for the ageing population. A condition like stroke necessitates ongoing observation and monitoring. Due to stress, inactivity, drug use, and poor eating habits, the number of stroke patients is rising daily. Because of decrease in the blood flow, brain eventually runs out of oxygen and nutrients which causes the brain cells to die. Because of the narrowing of blood vessels, the blood pressure occurring between the blood vessels connecting the brain and the heart starts to reduce. This reduce in blood pressure causes less blood transfer from heart to brain which indirectly reduces the amount of oxygen being supplied to the brain. The reduction in oxygen supply to the brain causes symptomatic diseases such as dizziness, headache and many other. Severe dizziness and headaches are the major symptoms that indicate a chance of stroke being occur able.

Another possibility is such that because of narrowing of blood vessels, the pressure being applied on the walls of blood vessels might increase and cause the blood vessels to burst which in turn result to Ischemic Stroke. A series of methods can be employed to predict out the occurrence of Ischemic Stroke in an individual such as physical examination, blood tests, computerized tomography scan and magnetic resonance imaging, where blood test happens to be the most utilized method. While many strokes are treatable, others can result in permanent disability or death. Stroke causes dysfunction in certain areas of the brain, resulting in difficulties with the brain blood arteries. Early discovery can stop serious brain-related illnesses, impairments, and even fatalities. The incidence of strokes has also increased due to COVID-19. The noninvasive methods have been extremely important in identifying strokes in COVID-19 patients. A doctor can treat a stroke patient in the most appropriate manner with minimal brain damage by adopting the rapid diagnosis method.

1.2 Problem Description

Stroke being mainly classified into two different types, both Hemorrhagic and Ischemic strokes require different therapies since they have different bodily effects and causes. The treatment may begin with tracking of drugs and breaking down the clots and preventing others from forming. This type of treatment may take much amount of time which may put the life of the patient in danger as it will be consuming a considerable amount of time and even other kind of treatments include dilating a balloon inside a blocked artery using a medical device called catheter which may be a kind of allergic for some people. Stroke can be prevented by following some traditional methods such as brief walking for every day and yoga proves to enrich the blood flow to brain and other parts of the body reducing the stroke impact. To get past these methods our project which will be predicting out whether an individual is having risk of stroke within minutes, without any physical test or medical examination.

Literature Review

2.1 General Introduction

Literature Survey is an important activity, which we must do while gathering information about a particular topic. It will help us to get required information or ideas to do work. The following paragraphs discuss the related work and issues in the area Prediction and Analysis of stroke using specific algorithms.

2.2 Literature Survey

In the paper titled “Detection of Stroke Disease using Machine Learning Algorithms” [1], the authors Romana Rahman Ema, Tasfia Ismail Shoily, Sharmin Akter Tanna, Sumaiya Jannat, Tajul Islam and Taslima Mostafa Alifa have used K Nearest Neighbour, Naive Bayes, Random Forest and J48 model to train the data. They collected stroke patients data from multiple sources and constructed a data set of 1057 records. After the performance analysis they got the accuracy of 85.6% for Naive Bayes and 99.8% for K Nearest Neighbour, J48, Random Forest.

In the paper titled “Performance analysis of machine learning approaches in stroke prediction” [2], the authors Maria Sultana Keya, Minhaz Uddin Emon, Tamara Islam Meghla, Shanim Kaiser, M Shamim Mamun and Md. Mahfujur Rahman conducted research to predict the chances of getting stroke. They trained 10 different classifiers such as K Nearest Neighbours, Adaboost classifier, Logistic Regression, Decision Tree, Stochastic Gradient Descent, XGBoost classifier, Quadratic Discriminant Analysis, Gaussian classifier, Multilayer Perceptron and Gradient Boosting Classifier. When they compared the performance of these base classifiers with weighted voting approach, it has achieved 97% accuracy.

In the paper titled “Stroke Prediction using Machine Learning Algorithms” [3], the authors Vijayaganth V, Revanth S, Sanjay N and Sanjay S compared Machine Learning Algorithms performance on Prediction of Stroke. In their study, they compared the performance of Multilayer perceptron, Support vector machine, Random Forest, and Decision tree. After the study it was stated that Support vector Machine gives an accuracy

of 98.99%. So, they created a model which was mainly based on Support Vector Machine Algorithm for the entire project. Results shown that, this method has provided higher accuracy than other well-known methods. In the project, the authors provided input to the multi-layer perceptron based on the input weights obtained from the previous nodes.

In the paper titled “Prediction and control of stroke by data mining” [4], the authors N. Toghianfar, L. Amini, M. T. Farzadfar, R. Azarpazhouh, S. A. Mousavi and R. Norouzi. collected 807 records and conducted research to predict stroke. According to their study Decision tree algorithm is the best model to predict stroke with 95% accuracy and K Nearest Classifier was giving 94% accuracy. And considering the accuracy score provided by both the both the algorithms, they created two different models with each model based on different algorithms. The training and testing of the model with similar and dissimilar inputs provided conflicting results which were further analyzed to arrive at a conclusion.

In the paper titled “Classification of Ischemic Stroke using Machine Learning Algorithms” [5], the authors Selma Yaheya Adam, Mohammed Bekri Bashir and Adil Yousif developed a Machine Learning model for predicting the ischemic stroke using K Nearest Neighbour classifier and Decision Tree algorithms. They collected data from various hospitals and prepared data set of 400 patients. After the experiment, it is found that decision tree classification is better than the K Nearest Neighbour for the prediction of stroke. Even the dataset was well balanced to arrive at a clear and much more in-depth conclusion. The model when provided with any input would provide a precise result of nearly complete accurate.

In the paper titled "Stroke Prediction using Artificial Intelligence" [6], the authors M. Sheethal singh and Prakash Choudhary compared different techniques with their proposed method for the prediction of stroke using Cardiovascular Health Study. In the proposed method, for feature selection Decision Tree was used, for dimension reduction Principal Component Analysis was used and for the classification purpose Back propagation Neural Network was used. Results shown that, this method has provided higher accuracy than other well-known methods.

In the paper titled "Stroke Prediction using Machine Learning Algorithms" [7], the authors Gunjan Gupta, Harshitha K, Prajna K B, Harshitha P and Vaishak P compared

five different algorithms of machine learning for the stroke prediction. Authors used K Nearest neighbor, Support vector classifier, Decision tree, Logistic regression, and Random Forest to assess the stroke risk. Among them Random Forest was giving the better accuracy of 95%. The training and testing of the model with similar and dissimilar inputs provided conflicting results which were further analyzed to arrive at a conclusion.

In the paper titled “Predicting Stroke Risk with an Interpretable Classifier” [8], the authors Jose a. Pino, Sergio penafiel, Horacio sanson and Nelson baloian developed a model using Dempster-Shafer theory. They collected data from a hospital in Japan and constructed a dataset. According to the study, their method performed better prediction compared to other commonly used machine learning techniques. The training and testing of the model with similar and dissimilar inputs provided conflicting results which were further analyzed to arrive at a conclusion.

In the paper titled “A Survey on Stroke Disease Classification and Prediction using Machine Learning Algorithms” [9], the authors Veena Potdar, Yashu Raj CY and Lavanya Santhosh reviewed different methods for the prediction of stroke by considering various previous works. They concluded that, Random Forest can be used for the stroke prediction as it provides the better results. Also, they mentioned that selection of the method should be based on problem type, data analysis and other factors. Both training and testing of the model with similar and dissimilar inputs provided conflicting results which were further analyzed to arrive at a conclusion. Results shown that, this method has provided higher accuracy than other well-known methods.

In the paper titled “A Comprehensive Method for Identification of Stroke using Deep Learning” [10], the authors Surya. S, B. Yamini , T. Rajendran and K.E. Narayanan discussed different deep learning techniques for the identification of stroke. The training and testing of the model with similar and dissimilar inputs provided conflicting results which were further analyzed to arrive at a conclusion. . Results proved that this method has provided higher accuracy than other well-known methods. Various works which used Computed Tomography Scan, Magnetic Resonance Imaging and Electroencephalogram images for the identification of stroke in the brain are discussed in this paper.

2.3 Summary

Considering literature survey as an important aspect we went through many papers, and we have arrived at a conclusion that a perfect solution cannot be obtained from just analysis of physical aspects, so we are just after maximum accuracy which can be obtained. We went through many of the algorithms which are much suited and found out that the following algorithms K Nearest Neighbor, Decision Tree and Random Forest are much more suitable for our project. Considering the accuracy level of each algorithm we arrive at a final algorithm for our project.

Problem Formulation

3.1 General

Before attempting to solve a problem, we need to formulate the problem. It is important to exactly define the problem which is being solved. Proper definition of the problem being solved provides a clear intention and provides ideal ways of solving it. Once the problem is clearly defined, then a search for an ideal way of solving it begins. Problem formulation is the process of defining a problem, identifying the cause of the problem, and determining the solution. At first, a proper idea defining the problem statement is identified, then the cause and solutions are found.

3.2 Problem Statement

In recent days consumption of packaged and unhealthy foods is increasing at an exponential manner. Because of increase in consumption of unhygienic foods bodily factors such as high cholesterol, uncontrolled blood pressure, diabetes and modern lifestyle are causing stroke in individuals of all ages. But many patients in whom stroke can be observed are relatively older people. Early prediction of stroke can be helpful in minimizing casualty caused due to the stroke. The proposed project helps in predicting on whether an individual has any immediate risk of developing stroke or not.

3.3 Objectives of the Present Study

The objectives of the proposed project are as follows:

1. To create a machine learning model which can predict risk of Brain Stroke in an individual at early stage
2. To train the data set using the ML algorithms, namely K Nearest Neighbour, Decision Tree and Random Forest.
3. To provide a web interface where users can input their data and predict the risk of having stroke.

3.4 Summary

Convolutional methods of treating stroke may include medical examination and intake of some kinds of drugs, but to overcome these kinds of methods we have proposed a system where in just by filling out a simple form, it can be predicted out whether an individual will be having stroke or not. Our system does not require any kind of medical examination or scanning which makes it much more economical.

Requirements and Methodology

4.1 Requirements

The proposed project consists of following requirements:

4.1.1 Hardware requirements

4.1.2 Software requirements

4.1.1 Hardware Requirements

The hardware requirements for the proposed project are depicted in the table below:

Table 4.1: Hardware requirements

| Sl. No | Hardware / Equipment | Specification |
|--------|----------------------|---------------|
| 1 | Processor | i3 or higher |
| 2 | RAM | 4 GB or more |

4.1.2 Software Requirements

The software requirements for the proposed project are depicted in the table below:

Table 4.2: Software requirements

| Sl. No | Software | Specification |
|--------|------------------|----------------------------|
| 1 | Anaconda | Anaconda 64 - bit |
| 2 | Python | Python version 3 or higher |
| 3 | Django framework | Version 3.2 or higher |

4.2 Methodology Used

- 1) **Gathering data:** Initially Stroke Prediction Dataset is collected from Kaggle website. This dataset consists of 5110 patient records with 12 attributes such as gender, age, hypertension, work type, average glucose level etc.
- 2) **Data pre-processing:** Then the next step is data pre-processing. It is the process of converting raw data into a clean data. In this process missing values, noisy and inconsistent data in the dataset are handled.
- 3) **Training and testing of models:** Here data is split for training and testing purpose. 70% of the data is used to train the model and remaining 30% of the data is used for

testing. Then dataset will be trained using K Nearest Neighbor, Decision Tree, Random Forest, and K-Nearest Neighbor algorithms.

- 4) **Evaluation:** After the training and testing of the models, confusion matrix is plotted, and accuracy score is computed for each algorithm. Then based on the accuracy score best suited algorithm for the prediction of stroke is identified.

System Design

5.1 System Design

System design is a one of the critical stages in system or software development. System design can be defined as a process of identifying all modules that are needed to achieve the objectives of the system or software.

5.1.1 Architecture of proposed system

System design represents the overall architecture of the proposed stroke prediction system, which is depicted in figure 5.1 shown below.

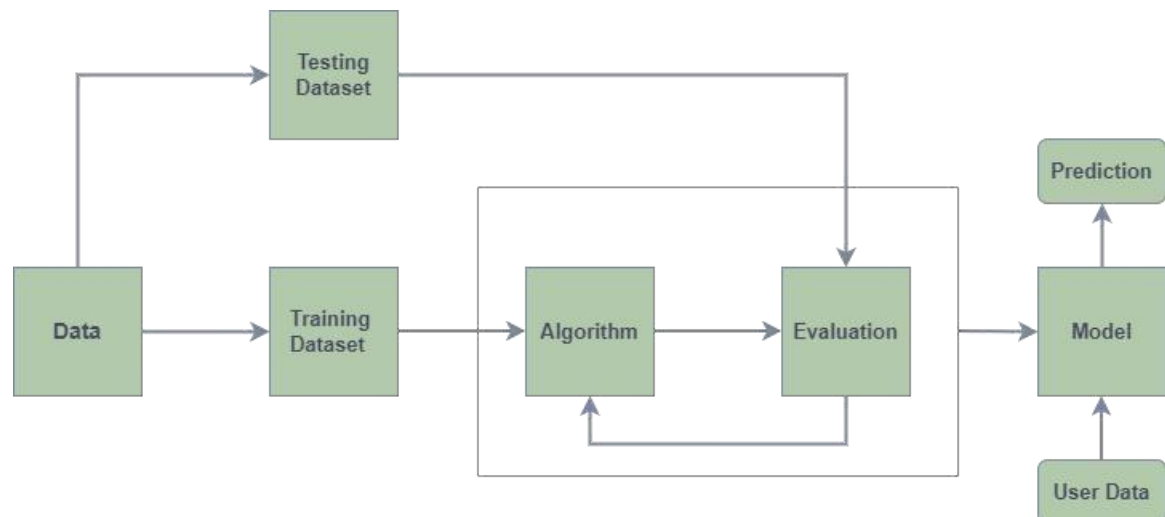


Figure 5.1: Architecture of the proposed stroke prediction system

In the above depicted architecture design of the proposed project, major stages involved in the stroke risk prediction are represented. Initially, available data is divided into 2 parts as training data and testing data. After that training data is used for training 3 different models. Then using different metrics such as accuracy, precision performance of the models is evaluated. Based on these values best performing model is selected among the 3 algorithms. Finally, best model is integrated with the user interface, and it can be used by the users for the stroke risk prediction by entering their data.

5.1.2 System Flowchart

The flowchart of the proposed system is shown in below figure 5.2 :

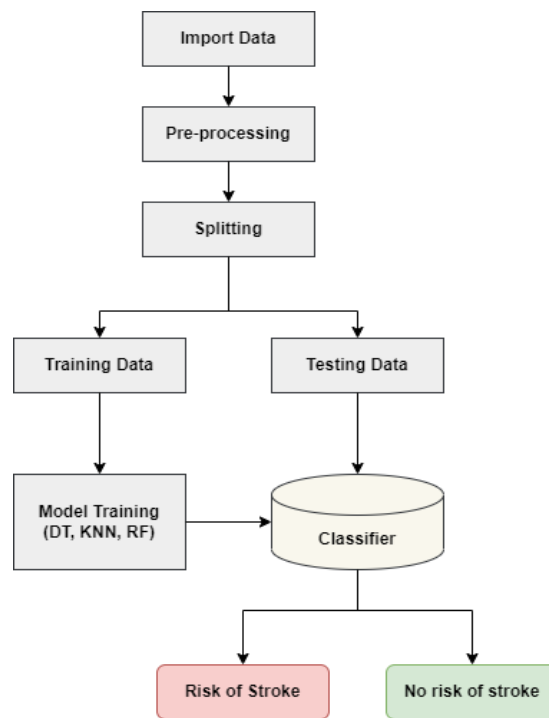


Figure 5.2: Flowchart of the proposed system

A system flowchart shows how the data is transported within the system. Once the data is imported from the data set then it can be pre-processed so that it is ready to use by a machine learning model. After the pre-processing steps data can be split for training and testing. Decision Tree, K Nearest Neighbor and Random Forest models are trained and then the testing data is used for evaluating their performance. Model takes the data of a person as the input and classifies him/her as having risk of stroke or as not having risk of stroke.

6.1 Pseudo code

Pseudo code is the step by step written procedure to achieve the proposed outcome, which can be then converted to the programming language code.

// Pseudo code for stroke risk prediction

1. Read data from user
2. Compute Result by providing input data to the trained model
3. if Result is equal to 1:
 print " The patient has a risk of stroke being occurred "
 else :
 print " The patient has no risk of stroke"
4. End if

System Testing, Results and Discussion

7.1 System Testing

Error detection is the main objective of system testing. It is the process of looking for any flaws or weaknesses in a product. It offers a way of testing whether parts, sub-assemblies, assemblies, and finished product performs its functions properly or not. It is the process of testing software to make sure that it satisfies user expectations and meets requirements without failing in any circumstances. Also, it is used to determine whether the created system is operating in accordance with the initial goals and specifications. Software testing process begins once after the program is written and its documentation, associated data structures are designed. Software testing is crucial for fixing errors in the system, without such assistance the project cannot be finished.

Table 7.1 : Unit Test Cases

| Test case number | Input | Stage | Expected behavior | Observed behavior | Status P=Pass F=Fail |
|-------------------------|---|-----------------------------|--|--------------------------|-------------------------------------|
| 1 | Clicking links in navigation bar | User Interface design stage | Corresponding page for that link must be loaded | As expected | P |
| 2 | Height and Weight in the BMI calculator | User Interface design stage | BMI must be calculated and displayed | As expected | P |
| 3 | Sample data from test set with stroke value = 0 | User Interface design stage | Result should be displayed as No Risk of Stroke | As expected | P |
| 4 | Sample data from test set with stroke value = 1 | User Interface design stage | Result should be displayed as patient has Risk of Stroke | As expected | P |

7.2 Result Analysis

The main aim of the project was to predict the stroke risk for a given individual. Table 7.1 shows the analysis which was performed on three algorithms considering their accuracy score. It was found that Random Forest provided the better accuracy than the other two algorithms.

Table 7.2: Performance Analysis of Three Algorithms

| Algorithms | Accuracy |
|---------------------|----------|
| Decision Tree | 70.6% |
| K-Nearest Neighbour | 76% |
| Random Forest | 80% |

Home Page

The figure 7.1 represents the homepage of the web application which includes an image carousel. Also, brief information about the stroke and need for its prediction is given in the home page.

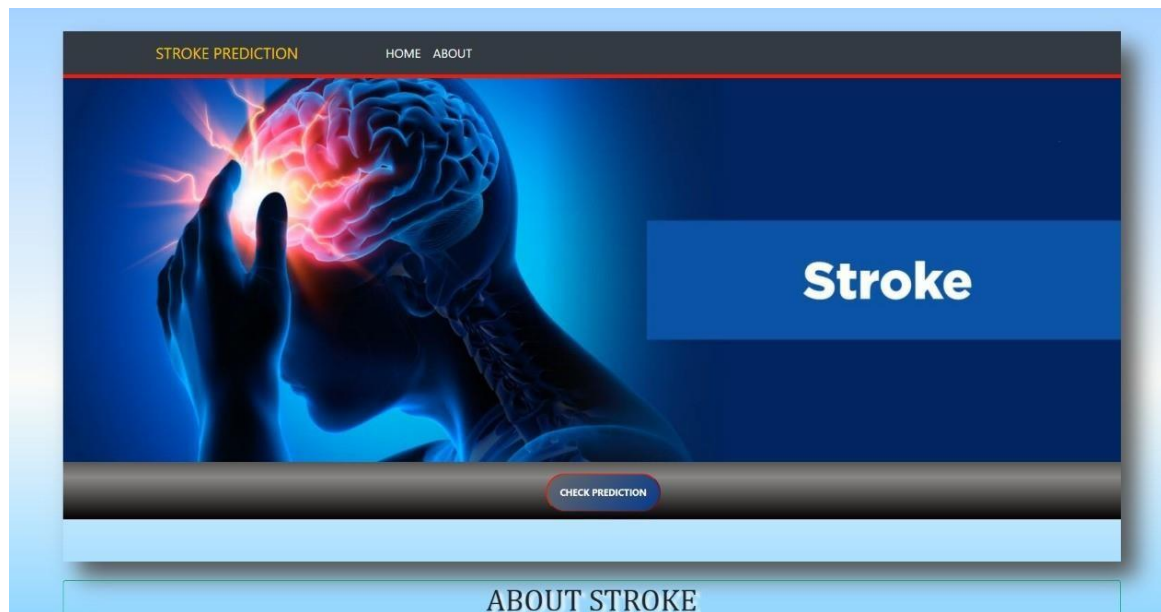


Figure 7.1: Home Page

User Form Page

Figure 7.2 shows the form which must be filled by the users for predicting the stroke risk.

The screenshot shows a web page with a dark header containing 'STROKE PREDICTION' and 'HOME'. The main content area has a blue-to-purple gradient background. A white form box is centered, titled 'PLEASE FILL THIS FORM TO PREDICT STROKE RISK'. The form contains the following fields:

- Gender: -- Select your gender --
- Age: -- Enter your Age --
- Hypertension: -- Do you have Hypertension? --
- Heart Disease: -- Do you have any Heart Disease? --
- Marital Status: -- Are you married? --
- Work Type: -- Select your work type --
- Residence Type: -- Select your residence type --
- Average Glucose Level: -- Enter Average Glucose level --
- Body Mass Index (BMI): -- Enter Body Mass Index --
- Smoking Status: -- Select smoking status --

A blue 'Submit' button is located at the bottom of the form. Below the button is a link: 'To Know Your BMI Click here'.

Figure 7.2: User Form Page

Figure 7.3 represents the stroke prediction form which is filled with details of a person. After entering their details users can click the submit button to see the prediction result.

The screenshot shows the same web page as Figure 7.2, but the form is now filled with data. The fields contain the following values:

- Gender: Male
- Age: 30
- Hypertension: Yes
- Heart Disease: Yes
- Marital Status: Yes
- Work Type: Self employed
- Residence Type: Urban
- Average Glucose Level: 108
- Body Mass Index (BMI): 35
- Smoking Status: formerly smoked

The blue 'Submit' button and the link 'To Know Your BMI Click here' are still visible at the bottom of the form.

Figure 7.3: Filled User Form

Results Page

Figure 7.4 represents the result page where the prediction result is displayed to the user. This figure shows result of a person who has no risk of stroke.

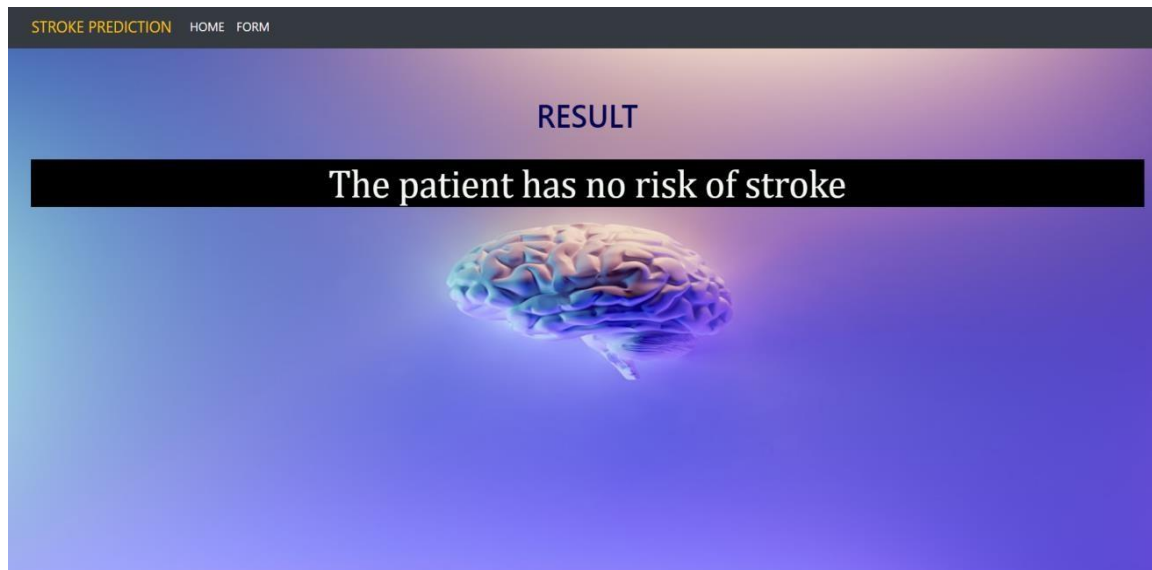


Figure 7.4: Result Page with result as No Risk of Stroke

Figure 7.5 represents the result page where the prediction result is displayed to the user. The following figure shows result of a person who has a risk of stroke being occurred.

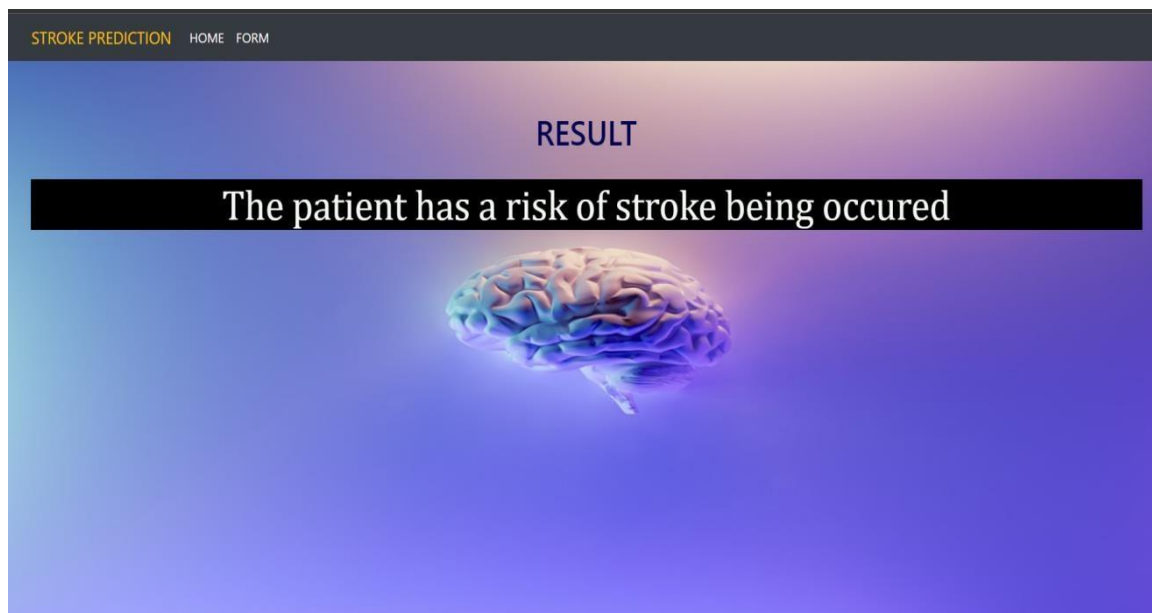
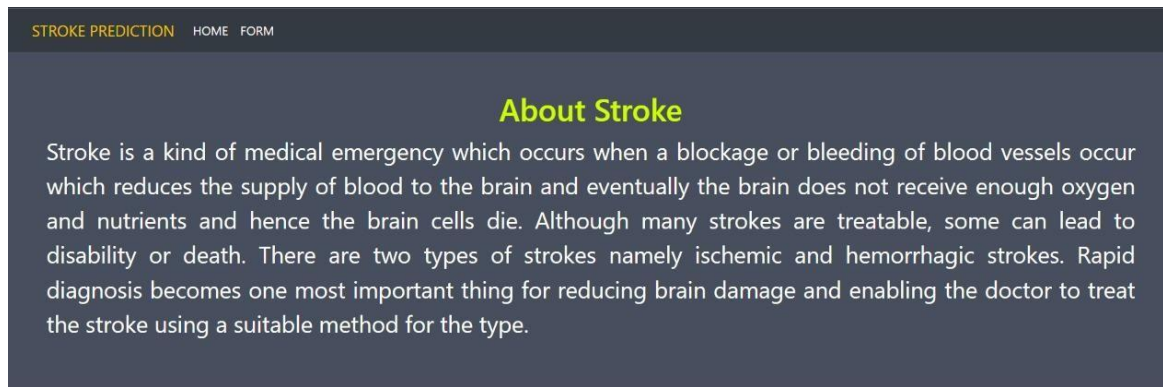


Figure 7.5: Result Page with result as Risk of Stroke

About Page

Figure 7.6 represents the result page where the prediction result is displayed to the user. This figure shows result of a person who has no risk of stroke.



What are the types of stroke?

Stroke can be caused either by a clot obstructing the flow of blood to the brain (called an ischemic stroke) or by a blood vessel rupturing and preventing blood flow to the brain (called a

Figure 7.6: About Page with descriptive information on stroke

Figure 7.7 depicts the information about the types of brain stroke with a brief description about each stroke.

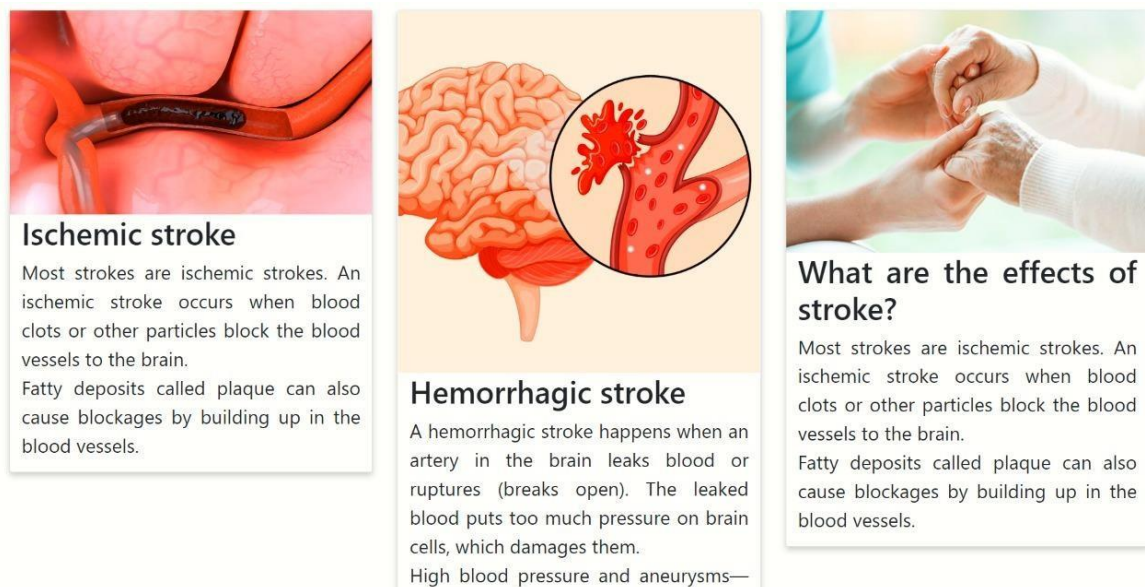
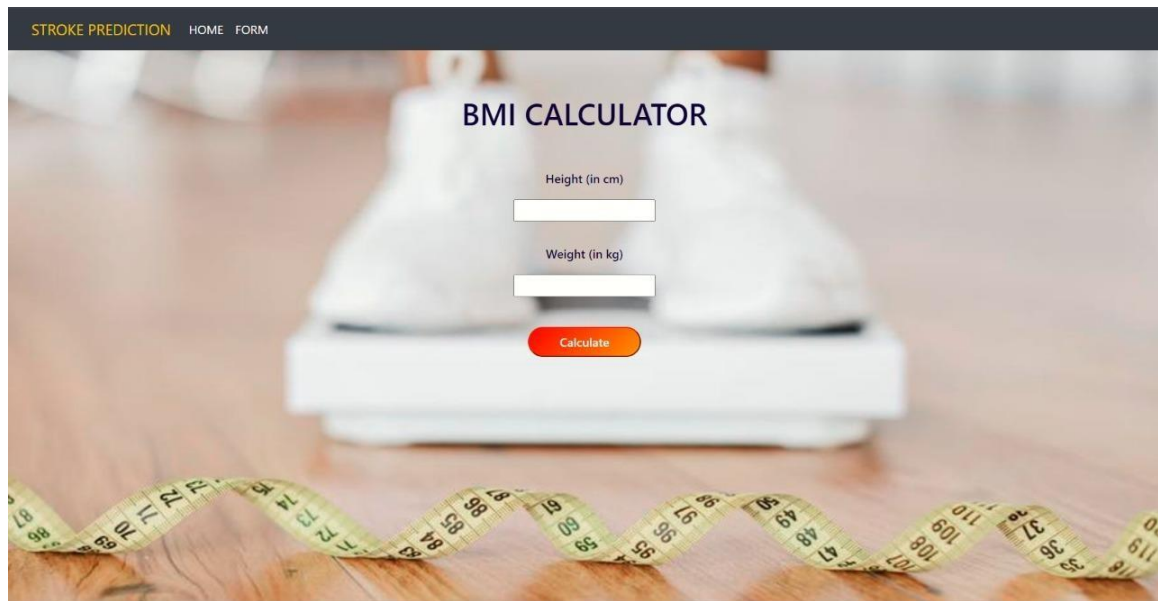


Figure 7.7: Referential page containing more information about types of strokes

BMI Calculator

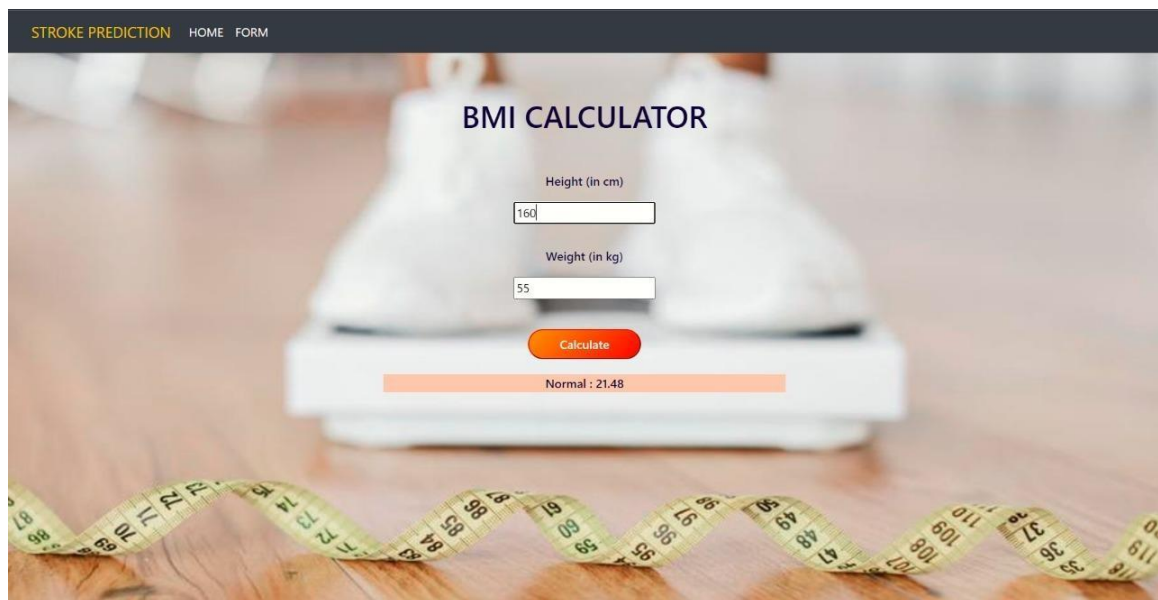
Figure 7.8 represents a page consisting of Body Mass Index calculator with fields such as height and weight to be entered.



The screenshot shows a web page titled "BMI CALCULATOR". At the top, there is a navigation bar with "STROKE PREDICTION", "HOME", and "FORM". The main content area features a white keyboard on a wooden surface with a yellow measuring tape in the foreground. The calculator form includes two input fields: "Height (in cm)" and "Weight (in kg)", both currently empty. Below the fields is a red "Calculate" button.

Figure 7.8: Body Mass Index Calculator Page

Figure 7.9 illustrates the output of the Body Mass Index page with the body type being described along with the value.



The screenshot shows the same BMI Calculator page as Figure 7.8, but with sample input. The "Height (in cm)" field contains the value "160" and the "Weight (in kg)" field contains "55". The red "Calculate" button is now highlighted. Below the button, a horizontal orange bar indicates the "Normal : 21.48" range.

Figure 7.9: Body Mass Index calculated for a sample input

7.3 Summary

The Stroke Prediction application was created using HTML, CSS, and Django framework. The user can manually enter the fields included on the website. The attributes considered during prediction are like Age, health issues records, working type and many such attributes. The snapshots included in the previous section indicates the outcome of the model designed.

Conclusion and Scope for Future Work

8.1 Conclusion

The project proposes a way of constructing an easier and simpler model for the prediction of stroke for an individual just based on some physical entities such as age, gender, work type, and many other. Proposed system tends out to be much more economical and less time consuming in predicting out whether an individual will be having stroke or not. The system developed will be simple and easy to use which makes it much more versatile.

8.2 Scope for Future Work

The project can be further improvised by considering real time datasets taken from the hospital using doctor's consent. This method further provides much scope for easier usage and much more economical when compared to the traditional medical methods. The project can be further utilized by providing a front-end window with a device separately for its prediction.

References

- [1] Tasfia Ismail Shoily, Sharmin Akter Tanna, Tajul Islam, Taslima Mostafa Alifa, Sumaiya Jannat and Romana Rahman Ema “Detection of Stroke Disease using Machine Learning Algorithms”, IEEE-2019.
- [2] Minhaz Uddin Emon, M Shamim Kaiser, Maria Sultana Keya, Md. Mahfujur Rahman and Tamara Islam Meghla “Performance analysis of machine learning approaches in stroke prediction” Fourth International Conference on Electronics, Communication and Aerospace Technology, IEEE Xplore.
- [3] Revanth S, Vijayaganth V, Sanjay N and Sanjay S “Stroke Prediction using Machine Learning Algorithms” International Journal of Disaster Recovery and Business Continuity, 2020 Volume 11, No. 1
- [4] L. Amini, N. Toghianfar, R. Azarpazhouh, S. A. Mousavi, M. T. Farzadfar, S. A. Mousavi and F. Jazaieri “Prediction and control of stroke by data mining,” International Journal of Preventive Medicine, volume 4, Supply 2, May 2013.
- [5] Selma Yahiya Adam, Mohammed Bakri Bashir and Adil Yousif “Classification of Ischemic Stroke using Machine Learning Algorithms”. International Journal of Computer Applications (0975 – 8887) Volume 149 – No.10.
- [6] M. Sheetal Singh, Prakash Choudhary "Stroke Prediction using Artificial Intelligence " 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON), 2017.
- [7] Harshitha K V, Prajna K B, Harshitha P, Vaishak P and Gunjan Gupta “Stroke Prediction Using Machine Learning Algorithms”. International Journal of Innovative Research in Engineering & Management, Volume 8, Issue 4, 2021.
- [8] Sergio penafiel , Jose a. Pino, Horacio sanson and Nelson baloian “Predicting Stroke Risk with an Interpretable Classifier”. IEEE Access Electronic ISSN: 2169-3536 DOI: 10.1109/ACCESS.2020.3047195.

- [9] Veena Potdar, Yashu Raj Gowda CY and Lavanya Santhosh “A Survey on Stroke Disease Classification and Prediction using Machine Learning Algorithms”. International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Volume 10 Issue 08, August 2021.
- [10] Surya.S, K.E. Narayanan, B. Yamini and T. Rajendran “A Comprehensive Method for Identification of Stroke using Deep Learning”. Turkish Journal of Computer and Mathematics Education Volume 12 No.7 , 2021.