**Stroke Disease Prediction**

**A PROJECT REPORT**

***Submitted by***

**Patel Bhargav Rajeshbhai**

**(181200107026)**

***In partial fulfilment for the award of the degree of***

**BACHELOR OF ENGINEERING**

**In**

**Information Technology**

**Aditya Silver Oak Institute of Technology, Ahmedabad**

**Logo, company name

Description automatically generated Logo

Description automatically generated**

**Gujarat Technological University, Ahmedabad**

**April,2022**

**Logo, company name

Description automatically generated Logo

Description automatically generated**

**Aditya Silver Oak Institute of Technology**

**Nr. Bhavik Publications, Opp. Bhagwat Vidyapith, S.G. Road, Gota Cross Road, Gota, Ahmedabad- 382481 – Gujarat.**

**CERTIFICATE**

This is to certify that the project report submitted along with the project entitled **Stroke Disease Prediction** has been carried out by **Patel Bhargav Rajeshbhai** under my guidance in partial fulfillment for the degree of Bachelor of Engineering in Information Technology, 8th Semester of Gujarat Technological University, Ahmadabad during the academic year 2021-22.

\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Prof . Manish Singh Prof. Jalpa C. Shah

Internal Guide Head of Department

**Logo, company name

Description automatically generated Logo

Description automatically generated**

**Aditya Silver Oak Institute of Technology**

**Nr. Bhavik Publications, Opp. Bhagwat Vidyapith, S.G. Road, Gota Cross Road, Gota, Ahmedabad- 382481 – Gujarat.**

**DECLARATION**

We hereby declare that the Internship / Project report submitted along with the Internship / Project entitled Stroke Disease Prediction submitted in partial fulfillment for the degree of Bachelor of Engineering in Computer Engineering to Gujarat Technological University, Ahmedabad, is a bonafide record of original project work carried out by me at Inexture Solutions LLP under the supervision of Pritesh Thaker and that no part of this report has been directly copied from any students’ reports or taken from any other source, without providing due reference.

Sr. no. Name of Student Sign of Student

1. Patel Bhargav Rajeshbhai \_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Industry Certificate**



**Logo, company name

Description automatically generated Logo

Description automatically generated**

**Aditya Silver Oak Institute of Technology**

**Nr. Bhavik Publications, Opp. Bhagwat Vidyapith, S.G. Road, Gota Cross Road, Gota, Ahmedabad- 382481 – Gujarat.**

**ACKNOWLEDGEMENT**

I would start by thanking my honourable faculty Pro. Manish Singh, who has provided me with the necessary guidance and information needed to complete this project report. Also, he provided me such task which helped me to start with the right topics to think on and provided me such environment which given me a correct direction to work on my weakness. Thanks for giving me technical and non-technical knowledge and guideline during activation period, and the entire team for being helpful and support.

Your sincerely,

Bhargav Patel

(181200107026)

**ABSTRACT**

*A stroke is a medical condition in which the blood arteries in the brain rupture, leading brain damage. Symptoms may appear if the brain's flow of blood and other nutrients is disrupted. Stroke is the leading cause of death and disability worldwide, according to the World Health Organization (WHO). Early awareness of the numerous stroke warning symptoms can assist to lessen the severity of the stroke. To forecast the chance of a stroke happening in the brain, many machine learning (ML) models have been created. Random forest, support vector machines, decision trees and classifiers, and neural networks were the most often utilised approaches.*

**List of Figure**

Fig 4.2.1 Gender data Visulization………………………………………………………….…6

Fig 4.2.2 Age data Visulization………………………………………………….…………….6

Fig 4.2.3 Hypertension data visulization………………………………………………………7

Fig 4.2.4 Heart\_disease data visualization………………….…………………………………7

Fig 4.2.5 Ever Married data visualization……………………………………………………..8

Fig 4.2.6 Work\_type data visualization………………………………………………………..8

Fig 4.2.7 Residence\_type Data Visulization…………………………………………………..9

Fig 4.2.8 Avg\_Glicose\_level Data visualization ……………………………………………...9

Fig 4.2.9 Bmi data visualization ………………………………………………………………9

Fig 4.2.10 smoking\_status data visualization………………..……………………………….10

Fig 4.2.11 Feature Compering ……………………………….………………………………10

Fig 5.1.1 Smote technique data Before Balancing ……………………………….………….13

Fig 5.1.2 Smote technique data After Balancing ………………………………….…………13

Fig 5.1.3 Performance Evaluation of various model (Smote Technique)……………………14

Fig 5.1.4 Performance of RandomeForestClassifier………………………………………....14

Fig 5.1.5 Learning Curve for Nueral Network ………………………………………………15

Fig 5.1.6 Neural Network Classification Report …………………………………………….16

Fig 5.1.7 Confusion Matrix for Neural Network ……………………………………………16

Fig 5.2.1 NearMiss Data Before Balancing …………………………………………………17

Fig 5.2.2 NearMiss Data After Balancing …...………………………………………………17

Fig 5.2.3 Performance of various model(Near-miss technique)…………………………..…18

Fig 5.2.4 Performance of RandomforestClassification ……………………………………...18

Fig 5.2.5 Evaluation of Neural Network …………………………………………………….18

Fig 5.2.6 Learning Curve (Loss over Epoch) : ………………………………………………19

Fig 5.2.7 Learning Curve (Accuracy over Epochs)…………………………………………..19

Fig 5.2.8 Confusion Matrics of nueral Network(NearMiss)…………………………………19

Fig 5.3.1 Recall values Comparision ………………………………………………………..20

Fig 5.3.2 Precision Values Comparision …………………………………………………….20

Fig 5.3.2 Accuracy values Comparision …………………………………………………….21

Fig 5.3.2 F1 Score Comparision …………………………………………………………….21

Fig 7.1 Flow chart …………………………………………………………………………...23

Fig 7.2 Login Page …………………………………………………………………………..24

Fig 7.3 Home page …………………………………………………………………………..24

**List of Tables**

Table 4.1 Data Desciption ……………………………………………...............5

Table 4.3 Feature Analysis Result…………………………………………..…11

**Table of Content**

Declaration…………......................................................................................................................i

Industy Certificate…………………………………………………………………………………………………………………ii

Acknowledgement ...................................................................................................................iii

Abstract....................................................................................................................................iv

List of Figures.............................................................................................................................v

List of Tables ...........................................................................................................................vii

Table of Contents....................................................................................................................viii

**Chapter 1 Company Overview ……….………………………………..……………………1**

1.1 HOSTORY…….………….………………………………..……………...……....1

1.2 PRODUCT ………..………………………………………..………..……...…….1

1.3 CAPACITY ………..………………………………………..……..……...………1

**Chapter 2 Project Introduction ……….………………………..……………………..……2**

2.1 Overview ………….………………………………………..……...……...............2

2.2 Porpuse ………………………………………………..……...…………………...2

2.3 Technology ……………………………………………..……...………………….3

2.4 Project Planing…………………………………………….…................................3

**Chapter 3 SYSTEM ANALYSIS ……………………………………………………...……4**

3.1 Study of current syste…...….……………………………………………...……....4

3.2 Proposed System………………………………………………..……...………….4

**Chapter 4 Implementation …………….……………………………………………..……..5**

4.1 Introduction to data…… .………………………………………………………....5

4.2 Exploratory Data Analysis….……………………………………………………..6

4.3 Feature Analysis Result …….…………………..………………………………..11

4.4 Observation about data features…… .…………………………………………...12

4.5 Missing Values Handling ….…………………………………………………….12

4.6 Label Encoding……………..…………………..………………………………..12

4.7 Standardizing and splitting.....…………………..………………………………..12

**Chapter 5 Model Testing………………………………………………….………..............13**

5.1 SMOTE Technique…………………………………………………………........13

5.2 NearMiss Technique……………………………………………………………..17

5.3 Performance Evaluation of Neural Network With NearMiss Technique ……...20

**Chapter 6 H5 Model and pickle files………………………………………………………22**

6.1 Introduction to h5 and pickle files………………………………..………………22

6.2 Introduction to Pickle module…………………………………….………….…..22

6.3 Introduction to Keras Module ………………………………………….………..22

**Chapter 7 Designing User Interface……………………………………………….………23**

7.1 Flow Chart …………….…………………………………………………………23

7.2 Login Page …………….…………………………………………………………24

7.3 Home Page..………………………………………………………………….…..24

7.4 Conclusion ………………………………………………………….……………25

7.5 Future Work ………………………………………………………….………….25

**References………………………….……………………………………………..…………26**

**CHAPTER 1- Company Overview**

**1.1 HISTORY**

Inexture Solutions LLP is established 10 years ago by the intellectuals to provide such software solutions that are functional, reliable, maintainable and cost-friendly to our existing and growing client and customer base. To consistently cater to their growing needs for an optimal solution, ensuring excellent support and service platform to give a hassle-free experience in achieving their dreams.

**1.2 PRODUCT**

Our company provide variety of services Software development, Web-Portal development, Website Designing, E-commerce development, SEO, Customized App Development, Data Management Software with cloud hosting facility, etc.

**1.3 CAPACITY**

There are currently 90+ employees are working in this company and there is different capacity of each department. Web development department have capacity of developing around 50 big full-fledged websites a year. App development department have capacity of developing around 30 big full-fledged mobile applications a year. Software development department have capacity of developing around 25 big full-fledged software a year.

**CHAPTER 2- Project Introduction**

**2.1 Overview**

When blood flow to different parts of the brain is interrupted or reduced, the cells in those areas of the brain don't get the nutrients and oxygen they need, and they die. A stroke is a life-threatening medical condition that need immediate medical intervention. To avoid additional damage to the afflicted area of the brain, as well as effects in other parts of the body, early identification and effective therapy are necessary. According to the World Health Organization (WHO), fifteen million people worldwide suffer from strokes each year, with one person dying every four to five minutes. According to the Centers for Disease Control and Prevention(CDC), stroke is the sixth greatest cause of death in the United States . Approximately 795,000 persons in the United States suffer from the devastating effects of strokes on a regular basis. It is the fourth largest major cause of death in India. Ischemic and hemorrhagic strokes are the two types of strokes. Clots hinder drainage in a chemical stroke, whereas a weak blood artery breaks and bleeds into the brain in a hemorrhagic stroke. Stroke can be avoided by living a healthy and balanced lifestyle that includes quitting smoking and drinking, maintaining a reasonable BMI and glucose level, and having great heart and kidney function.

**2.2 Purpose**

A stroke is a medical condition in which the blood arteries in the brain rupture, leading brain damage. Symptoms may appear if the brain's flow of blood and other nutrients is disrupted. Stroke is the leading cause of death and disability worldwide, according to the World Health Organization (WHO). Early awareness of the numerous stroke warning symptoms can assist to lessen the severity of the stroke. To forecast the chance of a stroke happening in the brain, many machine learning (ML) models have been created.

**2.3 Technology**

Technologies used for the project are listed below:

* Machine Learning Algorithms
* NearMiss Technique
* SMOTE Technique
* Neural Network
* Front-End: HTML5, CSS3
* Back-End: Django
* Database: Postgresql
* Version Control: Git

**2.4 Project Planing**

The steps for basic project planning was:

* Requirement gathering
* Data Collection
* Exploratory Data Analysis
* Training
* Testing
* Deployment

**CHAPTER 3- SYSTEM ANALYSIS**

**3.1 Study Of Current System**

In all current Systems for balancing a imblanced data they have used a SMOTE Technique which Add duplicate values to minority class**.** SMOTE first selects a minority class instance a at random and finds its k nearest minority class neighbors. The synthetic instance is then created by choosing one of the k nearest neighbors b at random and connecting a and b to form a line segment in the feature space. The synthetic instances are generated as a convex combination of the two chosen instances a and b.

**3.2 Proposed System**

In new system will use NearMiss Undersampling Technique to balance the imbalance data. NearMiss Technique remove the values form majority class.

Steps of NearMiss Technique :

1. The algorithm first calculates the distance between all the points in the larger class with the points in the smaller class. This can make the process of undersampling easier.
2. Select instaces of the larger class that have the shortest distance with the smaller class. These n classes need to be stored for elimination.
3. If there are m instances of the smaller class then the algorithm will return m\*n instances of the larger class.

**CHAPTER 4- Implementation**

**4.1 Introduction to data**

The investigation was conducted using a globally accessible stroke prediction dataset. This dataset has 5110 rows and 12 columns. The output column stroke value is either 1 or 0. A score of 0 indicates that no stroke risk has been identified, whereas a value of 1 indicates that a risk of stroke has been identified. In this dataset, the chance of 0 in the output column (stroke) is greater than the likelihood of 1 in the same column.

|  |  |  |
| --- | --- | --- |
| **Attribute Name** | **Type**  ***(Values)*** | **Description** |
| 1. id | Integer | A unique integer value for  patients |
| 2. gender | String literal *(Male, Female, Other)* | Tells the gender of the patient |
| 3. age | Integer | Age of the Patient |
| 4. hypertension | Integer  *(1, 0)* | Tells whether the patient has hypertension or not |
| 5. heart\_disease | Integer  *(1, 0)* | Tells whether the patient has heart disease or not |
| 6. ever\_married | String literal  *(Yes, No)* | It tells whether the patient is married or not |
| 7. work\_type | String literal *(children, Govt\_job, Never\_worked, Private, Self- employed)* | It gives different categories for work |
| 8. Residence\_type | String literal  *(Urban, Rural)* | The patient's residence type is stored |
| 9.avg\_glucose\_level | Floating point number | Gives the value of average glucose level in blood |
| 10. bmi | Floating point number | Gives the value of the patient's Body Mass Index |
| 11. smoking\_status | String literal *(formerly smoked, never smoked, smokes, unknown)* | It gives the smoking status of the patient |
| 12. stroke | Integer  *(1, 0)* | Output column that gives the stroke status |

Table : Data Discriptio

**4.2 Exploratory Data Analysisi**

**4.2.1 Gender**

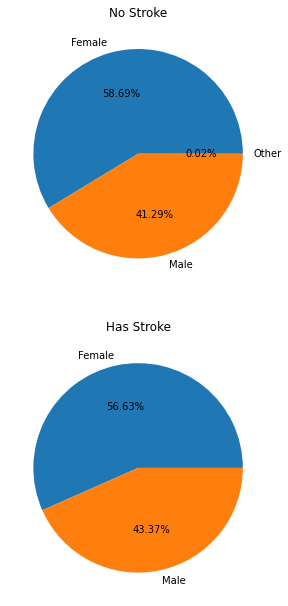
****

Figure 1 : Gender data Visulization

**4.2.2 Age**

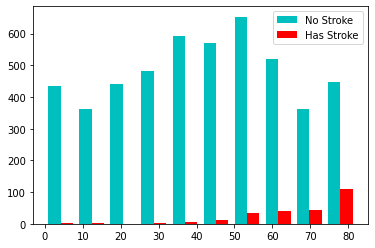
****

Figure 2 : Age Data Visulization

**4.2.3 Hypertension**

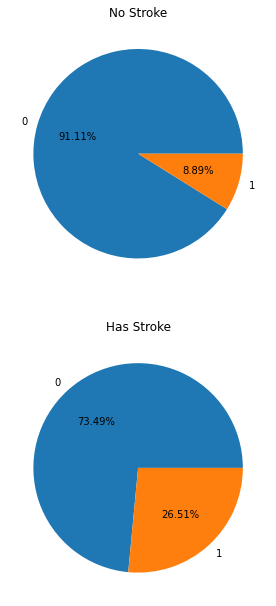
****

Figure 3 : Hypertension data visulization

**4.2.4 Heart\_disease**

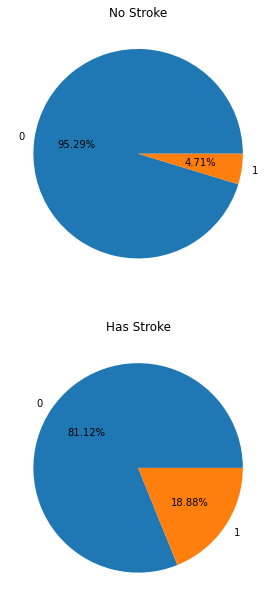
****

Figure 4 : Heart\_disease Data Visulization

**4.2.5 Ever\_Married**

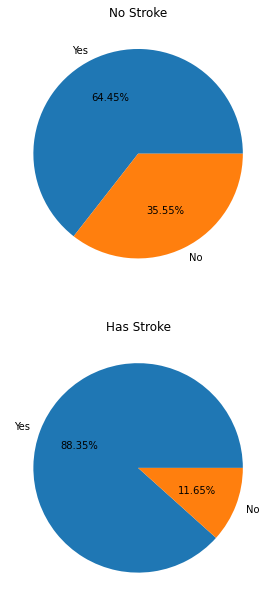
****

Figure 5 : Ever\_Married Data Visulization

**4.2.6 Work\_type**

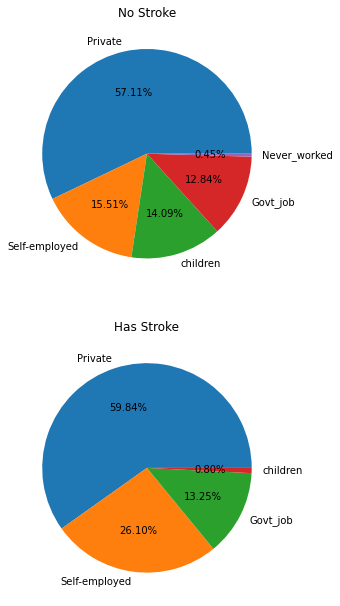
****

Figure 6 : Work\_Type data visulization

**4.2.7 Residence\_type**

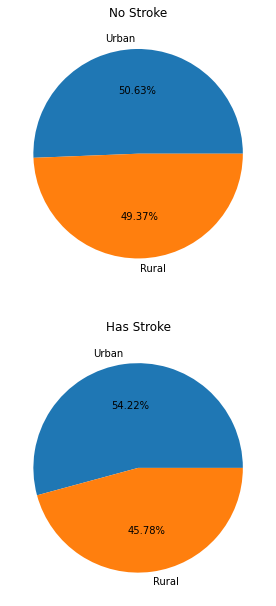
****

Figure 7 : Residence\_type Data Visulization

**4.2.8 Avg\_glucose\_level**

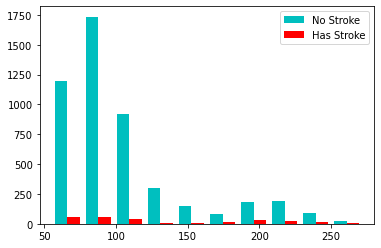
****

Figure 8 : Avg\_glucose\_level Data Visulization

**4.2.9 Bmi**

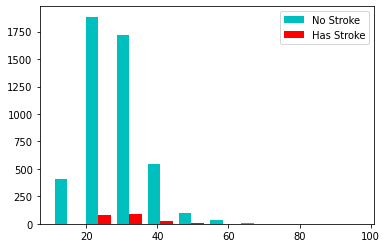
****

Figure 9 : Bmi Dadta Visulization

**4.2.10 Smoking\_status**

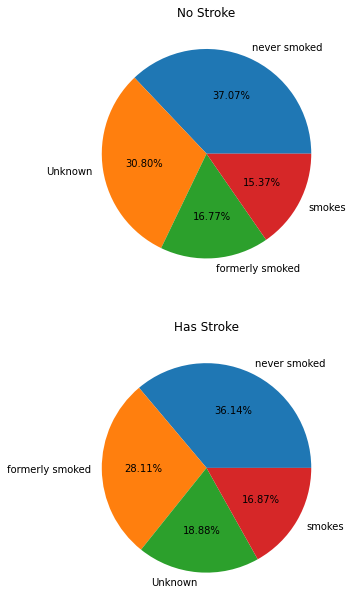
****

Figure 8 :Smoking\_status Data visualization

**4.2.10 Feature Comparision**

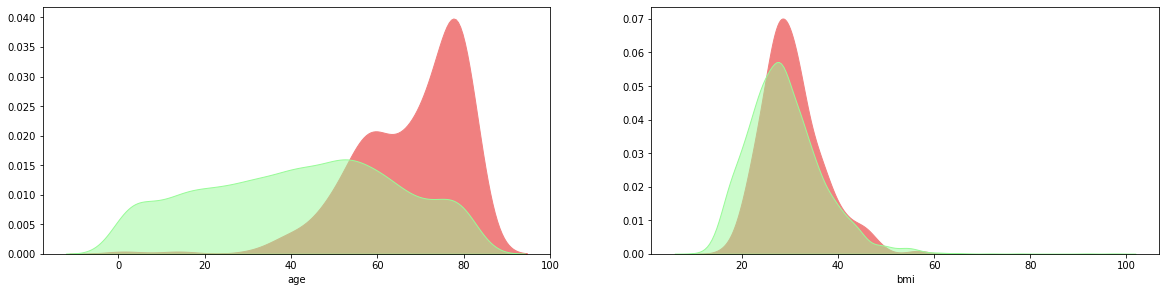
****

Figure 9 : feature Comparison

**4.3 Feature Analysis Result**

|  | **No Stroke** | **Has Stroke** | **Note** |
| --- | --- | --- | --- |
| gender (Most) | Female | Female | No Clear Difference also Other catagory can be ignore |
| age (Median) | 43 | 71 | the median age of stroke patients is higher than patient with no stroke |
| hypertension (Most) | 0 | 0 | the patient who has hypertension from stroke patient is 18 % higher than the patient with no stroke |
| heart\_disease (Most) | 0 | 0 | the patient who has heart disease from stroke patient is 14 % higher than the patient with no stroke |
| ever\_married (Most) | Yes | Yes | the patient who ever married from stroke patient is 24 % higher than the patient with no stroke |
| work\_type (Most) | Private | Private | the patient who work as self-employed from stroke patient is 11.4% higher than the patient with no stroke |
| Residence\_type (Most) | Urban | Urban | No Clear Difference |
| avg\_glucose\_level (Median) | 91.5 | 105.2 | the median of avg\_glucose\_level from Stroke Patient is higher than the Patient with no Stroke |
| bmi (Median) | 28.3 | 30.5 | the median of bmi from Stroke Patient is little higher than the Patient with no Stroke |
| smoking\_status (Most) | never smoked | never smoked | The patient who smokes or formerly smoked from is 13% higher than the patient with no stroke |
| Whole Dataset | 95.13% | 4.87% | The Data Is Imbalanced |

Table 2: Feature Analysis Result

**4.4 Observations about data features**

* People with age 65-85 have high chances of getting stroke.
* bmi can't distinguish stroke patterns and also have 4% missing values, Hence It can be drop.
* there are higher samples of no stroke (stroke=0) as compared to the other class. Hence it is a Highly Imbalanced dataset
* Others category in 'gender' can be ignored
* Type of Residence either Urban or Rural has no effect on having stroke. This feature can also be dropped.
* Dataset is Imbalanced.

**4.5 Missing Value Handling**

1. Deleting Rows
2. Replacing with Mean/Median/Mode
3. Assigning An Unique Category
4. Predicting the missing values

After doing exploratory data analysis, exclude the 'bmi', 'id', and 'Residence type' columns, as well as the row data for the Gender type 'other'.

**4.6 Label Encoding**

Label encoding converts a string data column to an integer type, allowing the model to better grasp the data pattern. Label Encoding is required for the 'gender','smoking status', 'work type', and 'ever married' columns.

**4.7 Standardizing and splitting**

Standardizing data helps model to tarin and test model faster on given standardized data. After Standardinzing the data splite the data in training and testing part For training on different models and evaluate performance.

**CHAPTER 5 – Model Testing**

**5.1 SMOTE Technique**

SMOTE Technique is a Over-sampling technique.

Working -*SMOTE first selects a minority class instance a at random and finds its k nearest minority class neighbors. The synthetic instance is then created by choosing one of the k nearest neighbors b at random and connecting a and b to form a line segment in the feature space. The synthetic instances are generated as a convex combination of the two chosen instances a and b.*

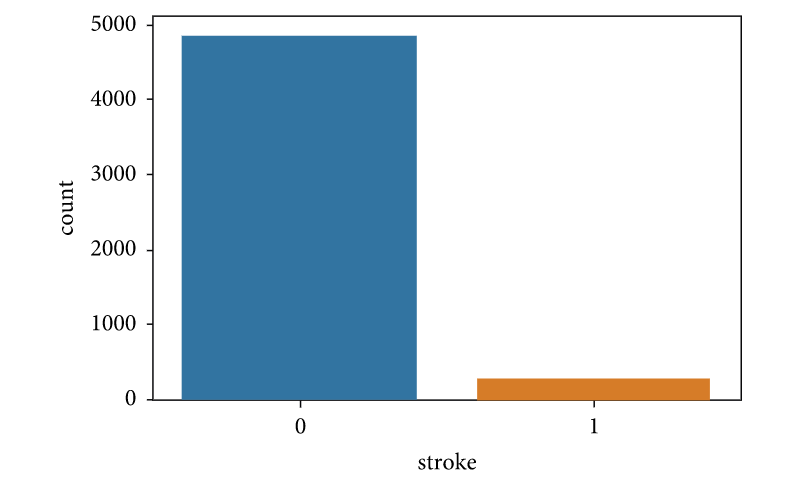


Figure 10 : Data Before Balancing

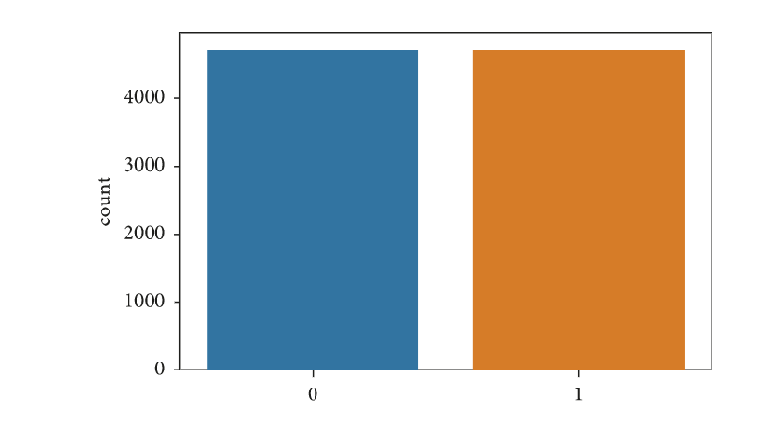


Figure 11 : Data After Balancing

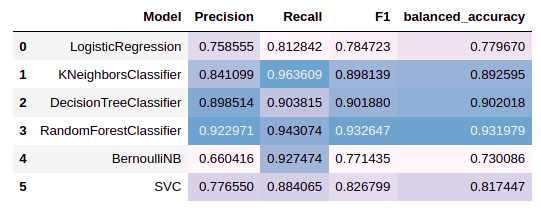


Figure 12 : Performance Evaluation of Various Model (SMOTE Technique)

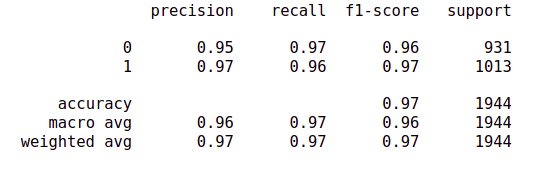
****

Figure 12 : Performance of RandomForestClassifier

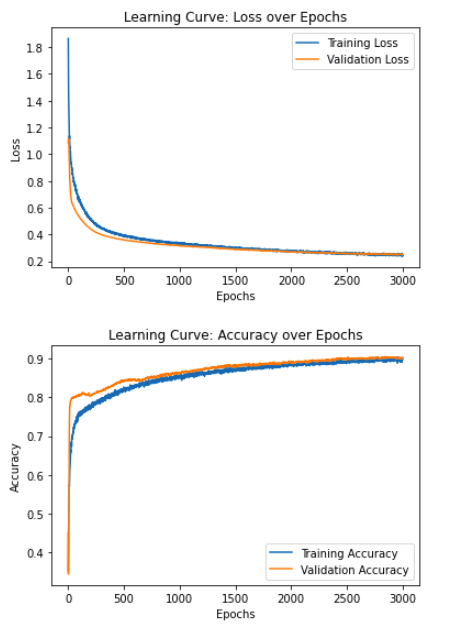


Figure 13 : Learning Curve For Neural Network

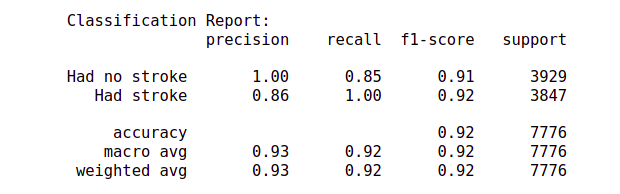
****

Figure 14 : Neural Network Classification Report

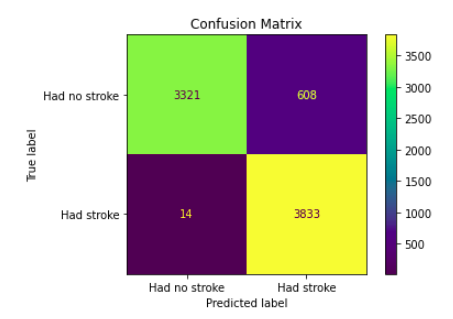


Figure 15 : Confusion Matrix For Neural Network Model

**5.2 NearMiss (Under- Sampling) Technique**

NearMiss (Under-Sampling) technique Used For Balancing the data which removes data from majority class. For this technique We need large set of data.

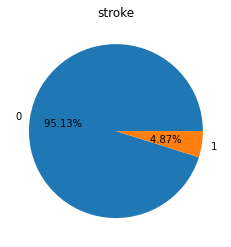


Figure 16 : Data before Balancing

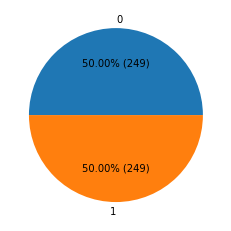


Figure 17 : After data Balancing

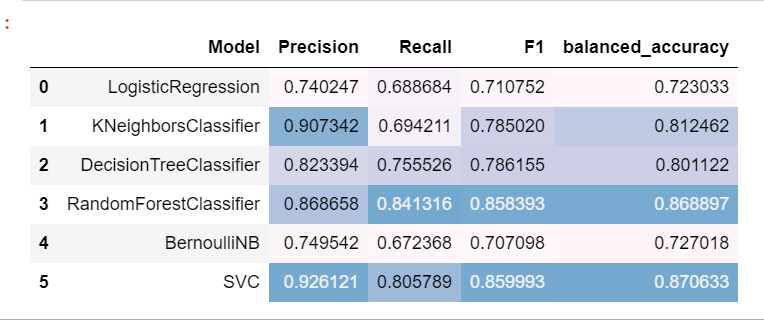


Figure 18 : Performance of various model(Near-Miss Technique)

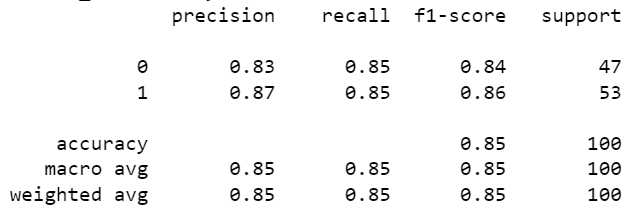


Figure 19 : Performance of RandmForestCLassfication

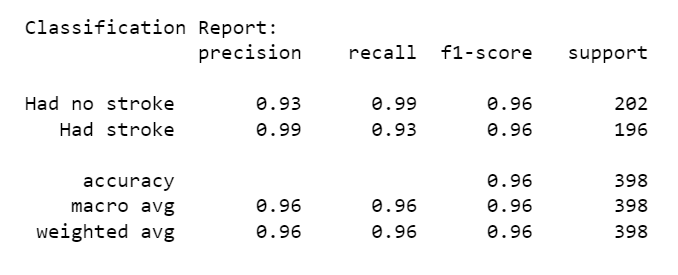


Figure 20 : Evaluation of Neural Network

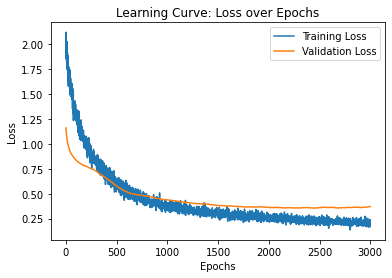


Figure 21 : Learning Curve (Loss over Epochs) : For Nueral network

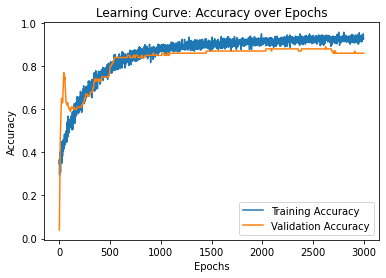


Figure 22 : Learning Curve (Accuracy Over Epochs) : For Nueral Network

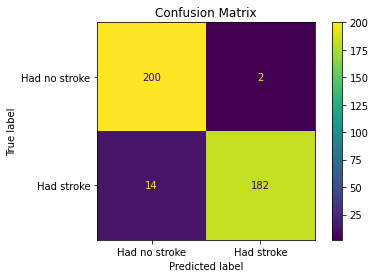


Figure 23 : Confusion Matrics of Neural Network

**5.3 Performance Evaluation of Models With Nearmiss Technique**

The 'Neural Network' algorithm performs the best out of all the algorithms tested, with a 96 percent accuracy rate. The following graph shows a comparison of accuracies achieved from various methods. 'Neural Network' outperformed the others in terms of accuracy, recall, and F1 scores.

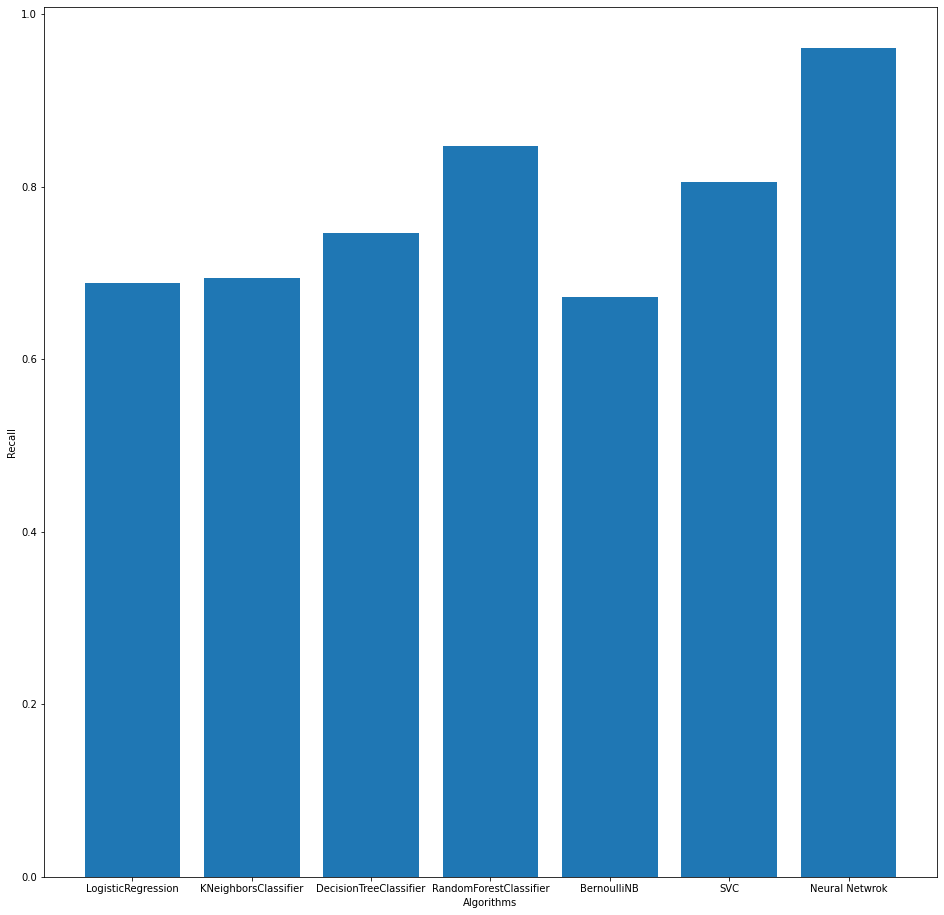


Figure 24 : Recall values Comparision

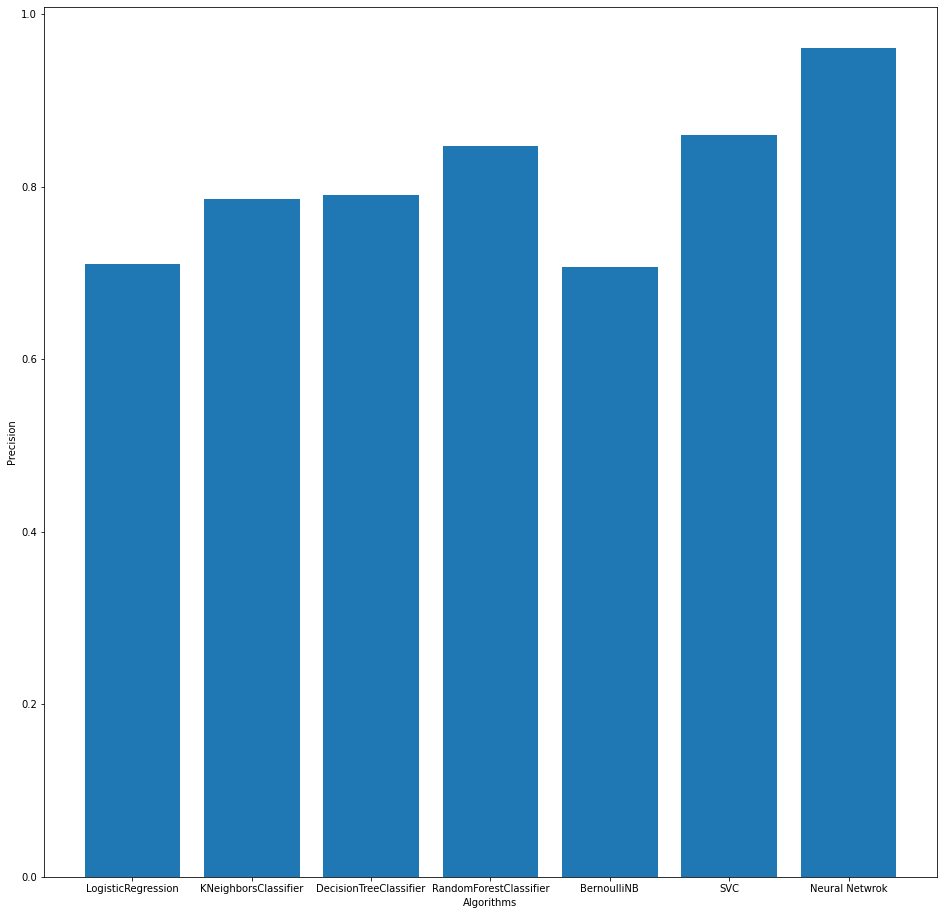


Figure 25 : Precision Values Comparision

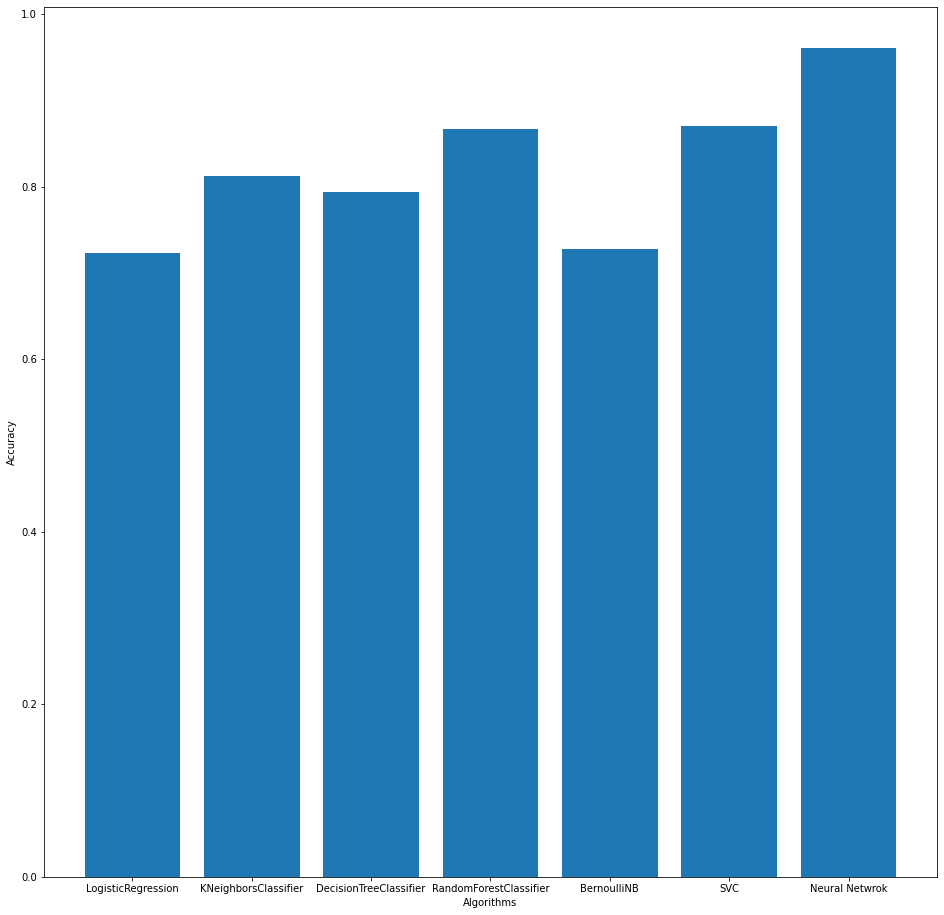


Figure 26 : Accuracy Values Comparision

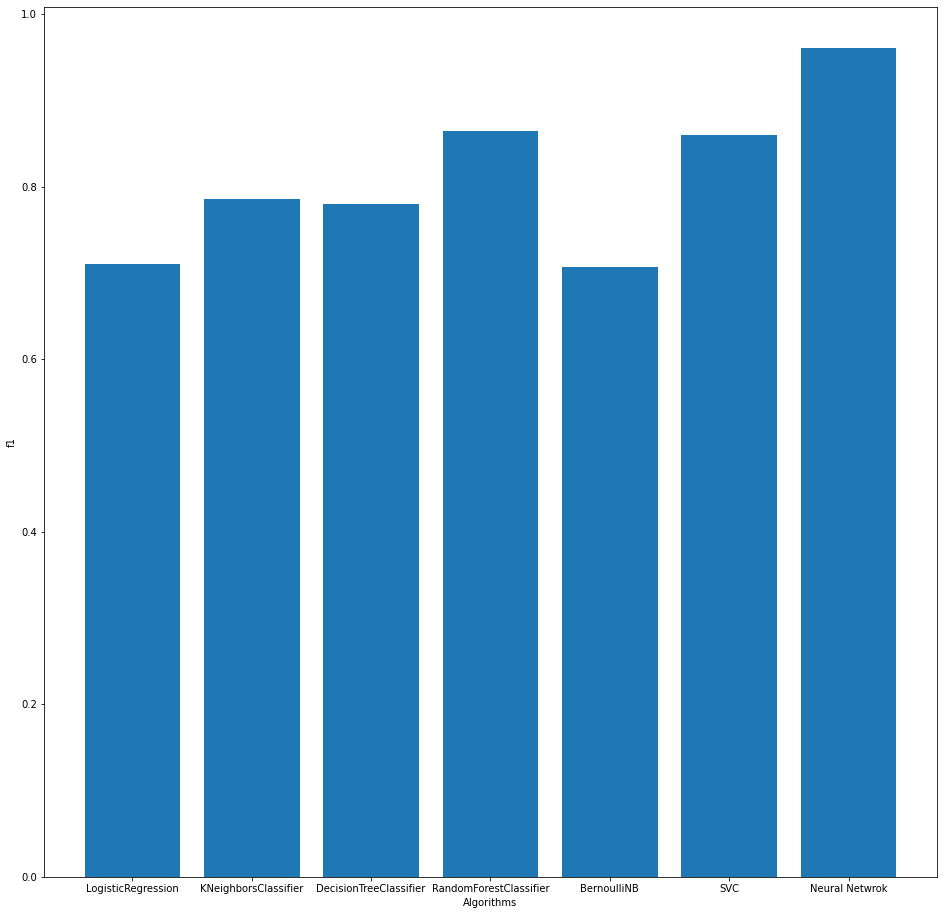


Figure 27 : F1 Score Comparision

**CHAPTER 6 - H5 And Pickle files**

**6.1 H5 file**

Using H5 file we can store scaler file which will be used to ssclae our data in when we connect our model to frontend.

**6.2 INTRODUCTION TO PICKLE MODULE**

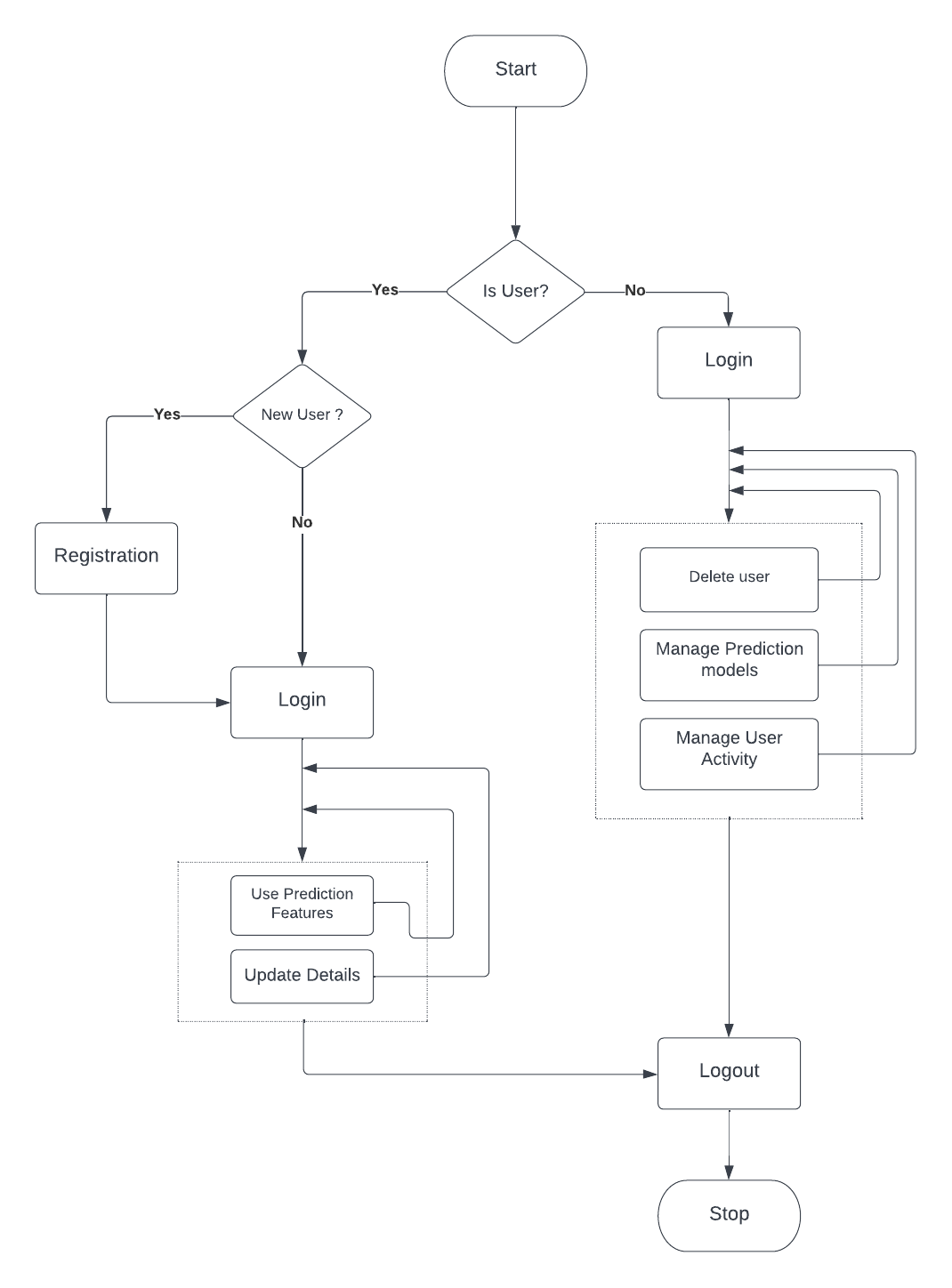
By python pickle module we can load the Machine learning model gathered in pickle file as above mentioned. And after opening the pickle file using open function of file module we can give the opened file to pickle.load(openFileObject). Then we can call this pickle file function called predict inside our callback function of DASH in this way we can link the pickle file to website.

**6.3 INTRODUCTION TO KERAS MODULE**

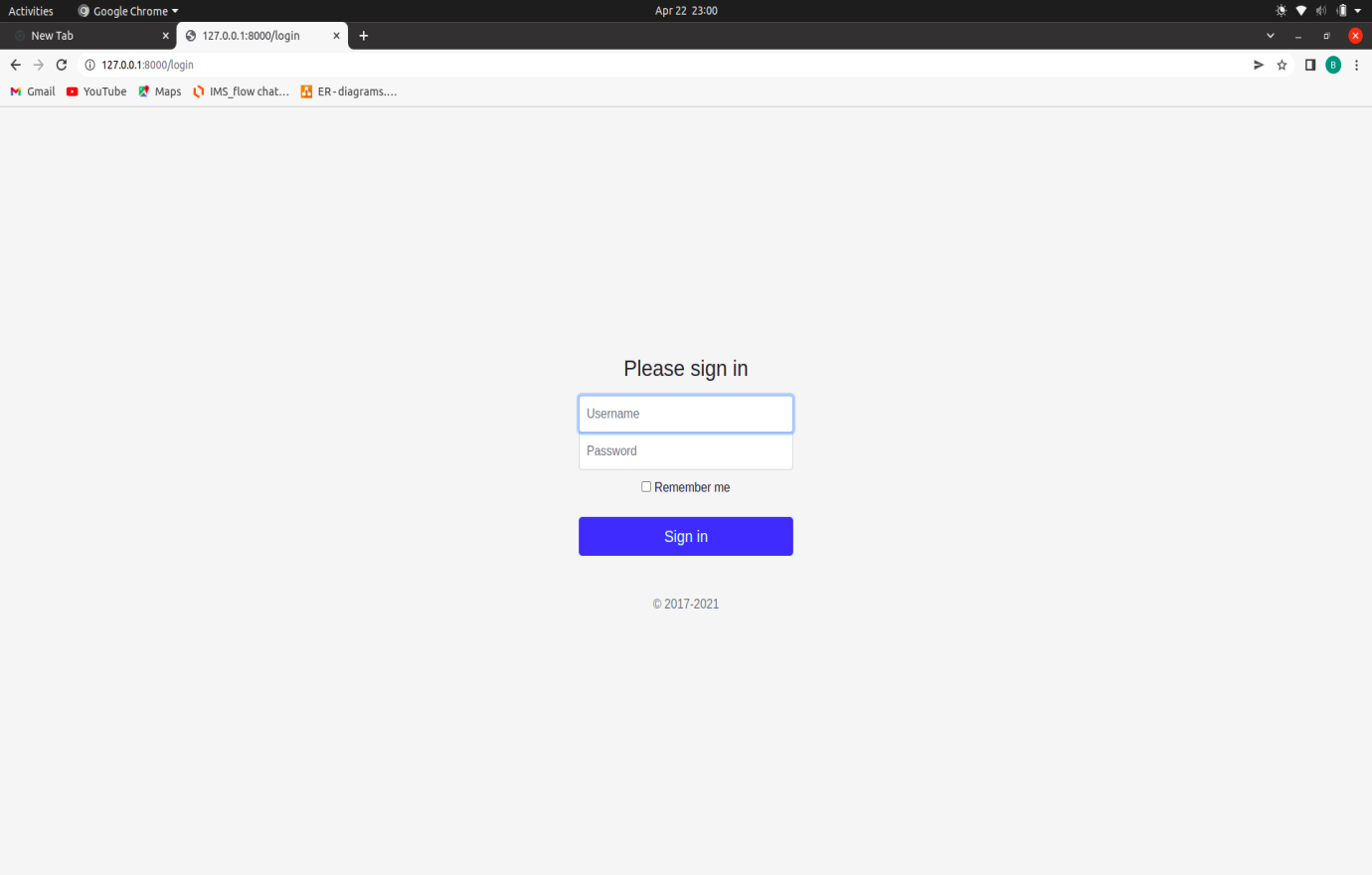
Keras is the main module of python in which the models module is located in which there is a function names load\_model which is used to load the h5 files as mentioned above. But h5 files need to have the preprocessed data because these models take data of some period of regular time interval and predicts the data of one time interval so numpy and pandas one of the famous python modules come into the picture of this model, after predicting the values from the learned data set the callback function of DASH use these created predicts functions by providing some set of inputs to it.

**CHAPTER 7 – User Interface Design**

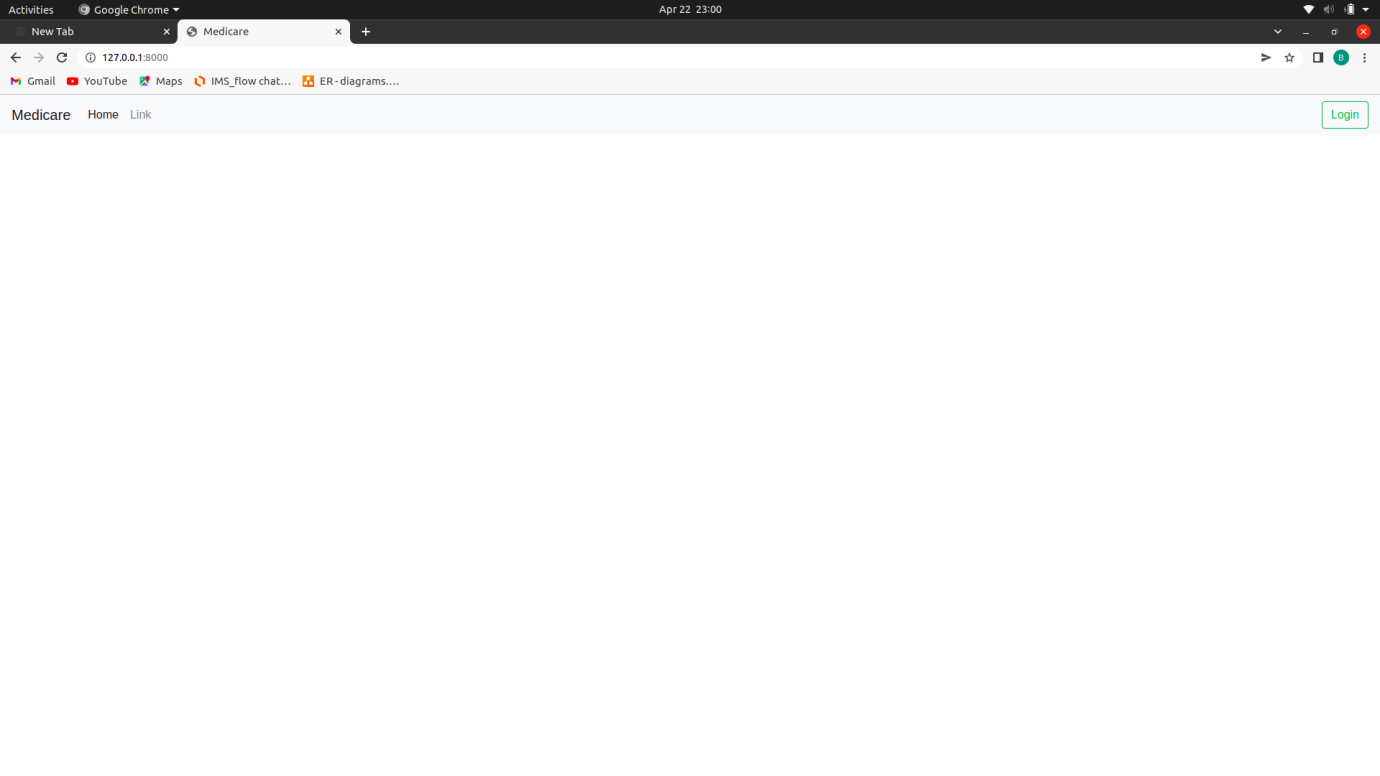
**7.1 Flow Chart**

****

**7.2 Login page**

****

**7.3 Home page**

****

**7.4 Conclusion**

With NearMiss imbalanced data managing technique The 'Neural Network' algorithm performs the best out of all the algorithms tested, with a 96 percent accuracy rate. The following graph shows a comparison of accuracies achieved from various methods. 'Neural Network' outperformed the others in terms of accuracy, recall, and F1 scores.

Delivering additional data as an input-set to neural networks and giving Brain CT Scan Image as input can increase neural network accuracy.

**7.5 Future Work**

Complete the user interface design and integret it with h5 and pickle file model using keras model to make end-to-end system.

**REFERENCES**

[1] Nueral Network

[2] Dataset named ‘Stroke Prediction Dataset’ from Kaggle

[3] Singh, M.S., Choudhary, P., Thongam, K.: A comparative analysis for various stroke prediction techniques. In: Springer, Singapore (2020).

[4] 7 Techniques to Handle Imbalanced Data – Kdnuggets.

[5] Under sampling Techniques -https://machinelearningmastery.com/undersampling-algorithms-for-imbalanced-classification/

[6] Documentation for Logistic Regression from Scikit-learn.org.

[7] Documentation for Decision Tree Classification from Scikit-learn.org.

[8] Documentation for Random Forest Classification from Scikit-learn.org.

[9] Documentation for K-Nearest Neighbor from Scikit-learn.org.

[10] Documentation for SVM from Scikit-learn.org.