IN PARTNERSHIP WITH PLYMOUTH UNIVERSITY

Name: Lanka Pathmakumara
Student Reference Number:

Module Code: PUSL 2078	Module Name: Statistics For Data Science
Coursework Title: Final Report	
Deadline Date:09/04/2024	Member of staff responsible for coursework: Ms. Kavishka Rajapaksha
Programme: BSc (Hons) Data Science	

Please note that University Academic Regulations are available under Rules and Regulations on the University website www.plymouth.ac.uk/studenthandbook.

Group work: please list all names of all participants formally associated with this work and state whether the work was undertaken alone or as part of a team. Please note you may be required to identify individual responsibility for component parts.

Lanka Pathmakumara -10899186
Gabbalage Dilshan - 10899287
Ponnahennadige Dias - 10899285
Bathala Wicramasinhe - 10899497
Balasuriya Balasuriya - 10899180

We confirm that we have read and understood the Plymouth University regulations relating to Assessment Offences and that we are aware of the possible penalties for any breach of these regulations. We confirm that this is the independent work of the group.

Signed on behalf of the group:

Individual assignment: ***I confirm that I have read and understood the Plymouth University regulations relating to Assessment Offences and that I am aware of the possible penalties for any breach of these regulations. I confirm that this is my own independent work.***

Signed :

Use of translation software: failure to declare that translation software or a similar writing aid has been used will be treated as an assessment offence.

I *have used/not used translation software.

If used, please state name of software..

Overall mark % Assessors Initials Date

*Please delete as appropriate Sci/ps/d:/students/cwkfrontcover/2013/14

Statistics For Data Science

PUSL2078

Coursework 2023–2024

Stroke Risk Prediction

Group D

Contents

1.Introduction

It is estimated by the World Stroke Organization (Anon, n.d.) that 13 million individuals globally experience a stroke every year, with 5.5 million of those deaths occurring as a result. It has an enormous effect on every aspect of life because it is the leading cause of death and disability in the world. The victim of a stroke affects their social environment, employment, and family. Furthermore, it can affect anyone, at any age, regardless of gender or physical condition, against popular perception (Elloker and Rhoda, 2018).

A stroke is characterized as an acute neurological condition of the brain's blood arteries that happens when blood flow to a part of the brain is cut off, depriving the brain's cells of oxygen. There are two types of strokes. They are ischemic and hemorrhagic strokes. It can cause either temporary or permanent harm, ranging from minor to very severe. Hemorrhages are uncommon and are caused by a blood artery burst, which can cause brain hemorrhage. The most frequent type of strokes, ischemic strokes, occur when an artery narrows or becomes blocked, causing blood flow to stop to a specific part of the brain (Katan and Luft, 2018; Bustamante *et al.*, 2021).

The following factors increase the risk of stroke: age (since stroke can strike anyone at any age, even children) plus hypertension, atherosclerosis-related carotid stenosis, smoking, high blood cholesterol, diabetes, obesity, sedentary lifestyle, alcohol consumption, blood clotting disorders, estrogen therapy, and use of euphoric substances like cocaine and amphetamines (Anon, n.d.), (Boehme *et al.*, 2017).

Stroke also advances quickly and has a wide range of symptoms. Sometimes symptoms appear gradually, and other times they appear suddenly. It's also possible for symptoms to awaken a person while they're asleep. The abrupt onset of one or more symptoms is indicative of a stroke. The most common ones are paralysis (typically on one side of the body) of the arms or legs, numbness in the arms or legs or on the face, trouble speaking, difficulty walking, dizziness, headache, vomiting, and a reduction in the mouth's angle (crooked mouth). Ultimately, a patient suffering from a massive stroke goes unconscious and enters a coma (Mosley *et al.*, 2007; Anon, n.d.).

Patients who have had a stroke are immediately diagnosed with a computed tomography scan. Patients suffering from ischemic stroke can be accurately diagnosed by magnetic resonance imaging (MRI). Strokes come in two varieties: severe and moderate. Most of the time, the first twenty-four hours are crucial. The route of treatment, which mostly comprises of medicine and occasionally surgery, will be highlighted by the diagnosis. Intubation and mechanical breathing must be performed by the intensive care unit when a patient goes into a coma (crossref, n.d.),(Anon, n.d.).

The majority of stroke survivors have ongoing difficulties long after their recovery, depending on the severity of their stroke. These difficulties can include memory, concentration, and attention problems; difficulty speaking or understanding speech; emotional problems such as depression; loss of balance or walking ability; loss of sensation on one side of the body; and difficulty swallowing food [9], [10].

Following a stroke, recovery aids in regaining lost function. With the help of neurologists, kinesiotherapists, and speech therapists, an appropriate plan is developed to guarantee the patient's quick social and psychological rehabilitation (crossref, n.d.), (Anon, n.d.). Regular blood pressure checks, continuous physical exercise, maintaining a healthy weight, quitting alcohol and tobacco use, and adhering to a nutritious diet reduced in fat and salt are all recommended to lower the risk of stroke (Pandian *et al.*, 2018).

Information and communication technologies (ICTs), especially artificial intelligence (AI) and machine learning (ML), are playing a bigger role in the early detection of many diseases, such as diabetes, high blood pressure, cholesterol, COVID-19, sleep disorders, hepatitis C, chronic kidney disease (CKD), and others. The stroke will be of particular concern to us during this inquiry. For this specific condition, a number of research projects have employed machine learning models.

This work offers a method for developing machine learning models for the occurrence of stroke that are effective in binary classification. The class balancing methodology, known as the synthetic minority over-sampling technique (SMOTE) (Maldonado *et al.*, 2019), was employed since it is crucial for developing efficient algorithms for stroke prediction. After then, a variety of models are developed, assembled, and assessed with the balanced dataset.

Logistic regression was assessed for our needs. Next, we created a web application to estimate the risk of stroke.

This is how the remainder of the paper is structured. The pertinent works concerning the topic under discussion are described in Section 2. Next, a description of the dataset and an analysis of the employed approach are presented in Section 3. Furthermore, we outline the experimental design and go over the obtained research findings in Section 4. Section 5 concludes with an overview of future paths and conclusions.

2.Related Work

The scientific community has shown a keen interest in developing tools and plans for monitoring and predicting a broad range of diseases that have a substantial impact on human health. This section will include the most recent research that use machine learning techniques to predict the risk of stroke.
First, the authors in (Shoily *et al.*, 2019) employed four machine learning algorithms—naive Bayes, J48, K-nearest neighbor, and random forest—to correctly identify a stroke. The accuracy of the naive Bayes classifier was only 85.6%, compared to 99.8% for the J48, K-nearest neighbor, and random forest classifiers.

Furthermore, to categorize stroke risk levels, (Anon, n.d.) used logistic regression, naive Bayes, Bayesian network, decision tree, neural network, random forest, bagged decision tree, voting, and boosting model with decision trees. According to the trial results, the random forest produced the highest precision (97.33%), while the boosting model with decision trees achieved the highest recall (99.94%).

Furthermore, (Sailasya and Kumari, 2021) applies the Kaggle dataset (Anon, n.d.). Our research work proposes developing a Web app and implementing logistic regression as a machine learning technique for stroke prediction.

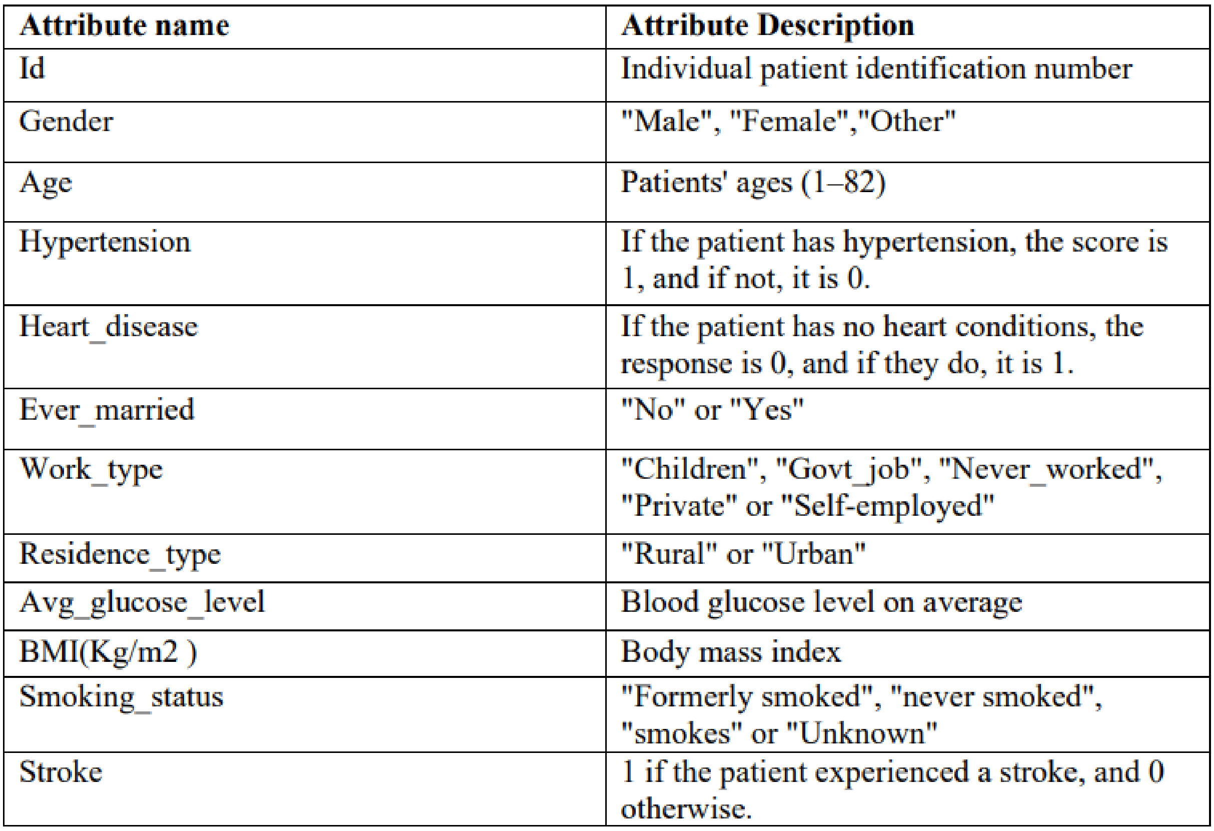
In conclusion, current research has demonstrated the ability of machine learning algorithms to precisely and accurately forecast the risk of stroke. Numerous models, such as naive Bayes, logistic regression, decision trees, neural networks, and ensemble techniques like boosting and random forest, have been investigated; nonetheless, each has demonstrated unique performance traits. These research collectively demonstrate the promise role of machine learning in improving stroke risk assessment, despite variations in techniques and datasets. Building on these findings, our research aims to further this field by creating a Web application that uses logistic regression to predict stroke, thereby offering a useful tool for proactive risk mitigation and health management to both individuals and healthcare professionals.

3.Materials and Methods

3.1. Dataset Description

Our study was built upon a Kaggle dataset (Anon, n.d.). There were 5110 participants, and the following is a description of each attribute:

Table 1 - DATASET ATTRIBUTE NAMES AND DESCRIPTION



Attribute name	Attribute Description
Id	Individual patient identification number
Gender	"Male", "Female","Other"
Age	Patients' ages (1–82)
Hypertension	If the patient has hypertension, the score is 1, and if not, it is 0.
Heart_disease	If the patient has no heart conditions, the response is 0, and if they do, it is 1.
Ever_married	"No" or "Yes"
Work_type	"Children", "Govt_job", "Never_worked", "Private" or "Self-employed"
Residence_type	"Rural" or "Urban"
Avg_glucose_level	Blood glucose level on average
BMI(Kg/m2)	Body mass index
Smoking_status	"Formerly smoked", "never smoked", "smokes" or "Unknown"
Stroke	1 if the patient experienced a stroke, and 0 otherwise.

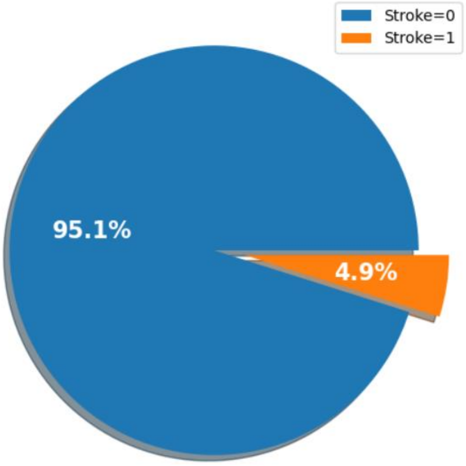


Figure 1-The visualizing count of classes (stroke and non-stroke) along with the percentage.

Count plots were used to graphically compare distributions in our analysis of stroke occurrences across various demographic and health-related parameters. The gender distribution of stroke cases is seen in Figure 2, with 108 strokes occurring in men and 141 in women. Figure 3 shows the correlation between the type of residence and the incidence of strokes: 135 strokes were reported among urban residence and 114 among rural people. The association between marital status and stroke incidence is depicted in Figure 4, which shows that there are 220 strokes among married people and 29 strokes among single people. Moreover, Figure 5 illustrates the relationship between heart disease and strokes by showing that patients with heart disease had 47 strokes whereas those without had 202 strokes. The distribution of strokes by status of hypertension is seen in Figure 6, with 183 strokes among people without hypertension and 432 among those who do. Finally, Figure 7 looks into how smoking status affects the incidence of strokes and shows that different smoking groups have variable numbers of strokes. Insights into the correlation between the dataset's demographic and health variables and the incidence of strokes are produced by these visualizations, which contribute to a better comprehension of stroke risk factors and possible interventions.

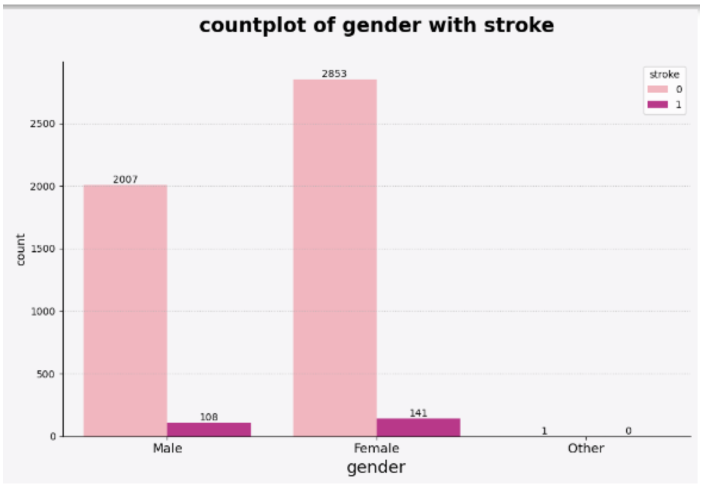


Figure 3- gender distribution of stroke.

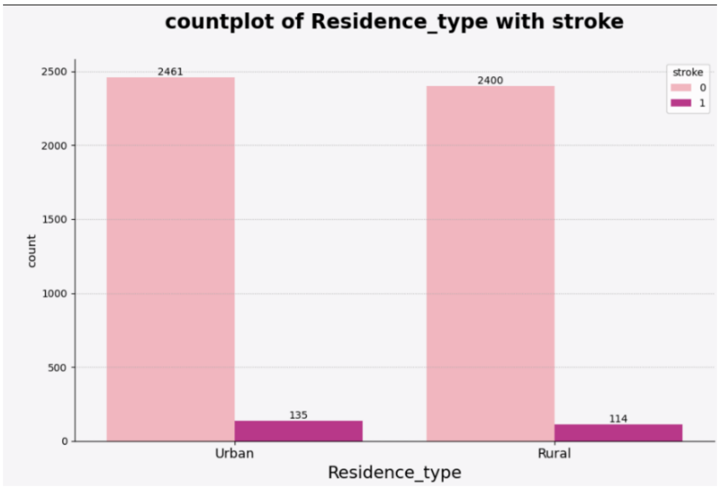


Figure 2-correlation between the type of residence.

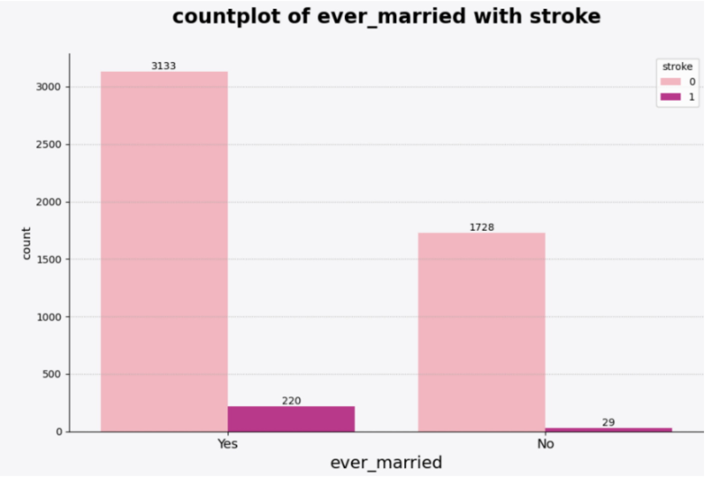


Figure 4- association between marital status and stroke.

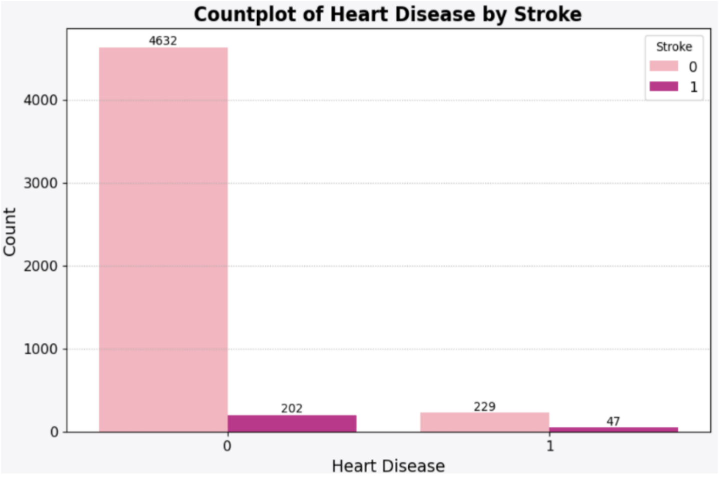


Figure 5- relationship between heart disease and strokes.

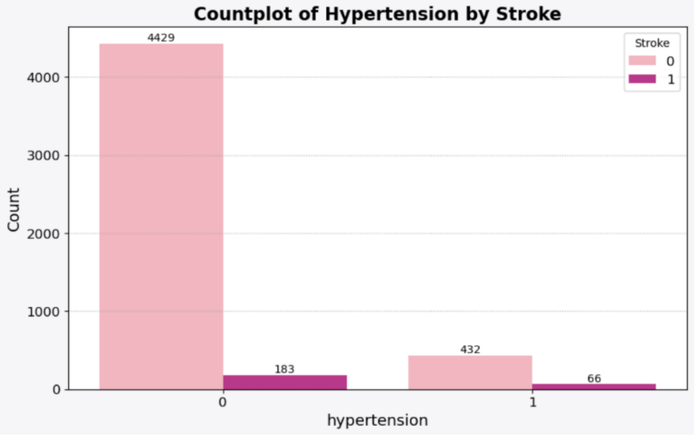


Figure 6- distribution of strokes by status of hypertension.

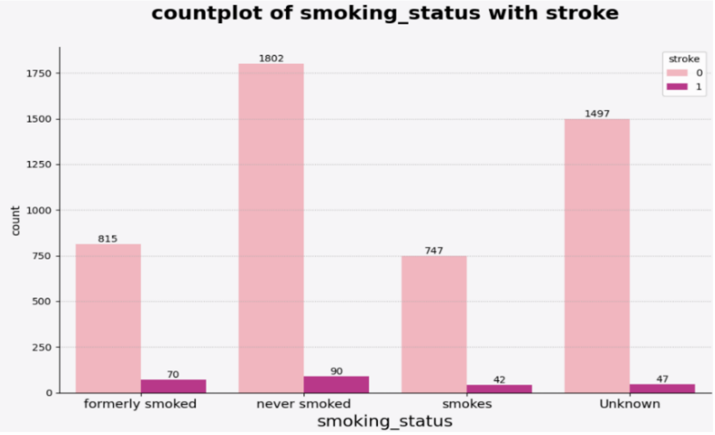
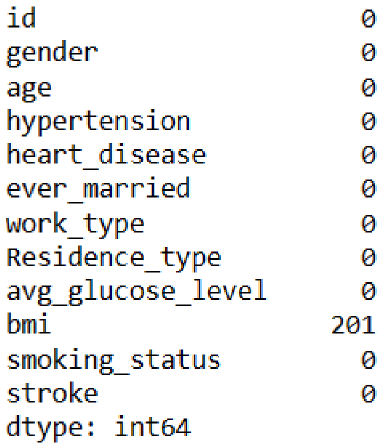


Figure 7-Smoking status and strokes

3.2. Data Preprocessing

The final prediction quality may degrade due to missing values and outliers in the raw data. Pretreatment processes including feature selection, data discretization, and redundant value removal are necessary to prepare data for mining and analysis (Fan *et al.*, 2021). First, we ascertained if the dataset had any null values. Out of 5110 data entries, 201 null values for the bmi are shown in Figure 8. We used the column mode of Scikit-learn's SimpleImputer to replace the null values with the missing ones. SimpleImputer, a univariate imputer from Scikit-learn, may replace missing values along each column with an explanatory statistic (such the mean, median, or most common).



```
id                     0
gender                 0
age                    0
hypertension           0
heart_disease          0
ever_married           0
work_type              0
Residence_type         0
avg_glucose_level      0
bmi                  201
smoking_status         0
stroke                 0
dtype: int64
```

Figure 8-Missing Values of the dataset.

Figure 9 shows the boxplots that were utilized to identify outliers in the 'bmi' and 'avg_glucose_level' columns. After that, an iterative process was employed to eliminate the outliers. Effective management of outliers is essential because they can seriously skew statistical analysis and machine learning models. Outliers are located and deleted iteratively until convergence is achieved in iterative techniques.

To make sure the dataset is better suited for analysis and modeling, we put the iterative outlier elimination technique into practice. Our goal in eliminating outliers is to raise the data's quality, which will therefore raise the prediction models' accuracy.

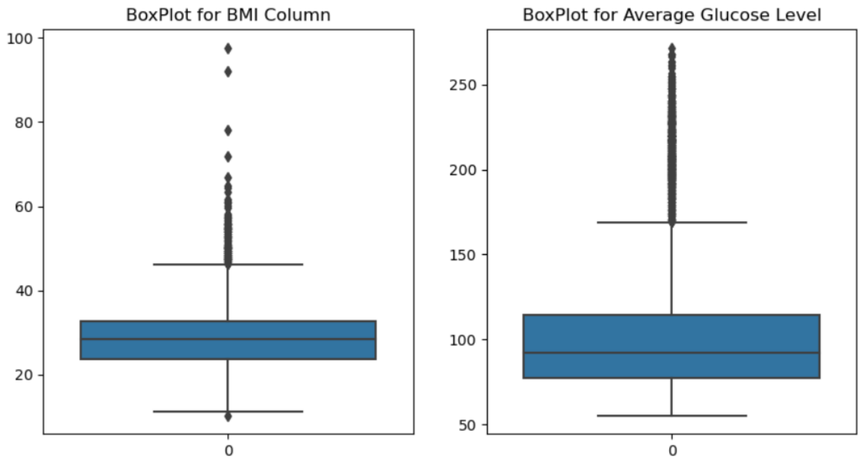
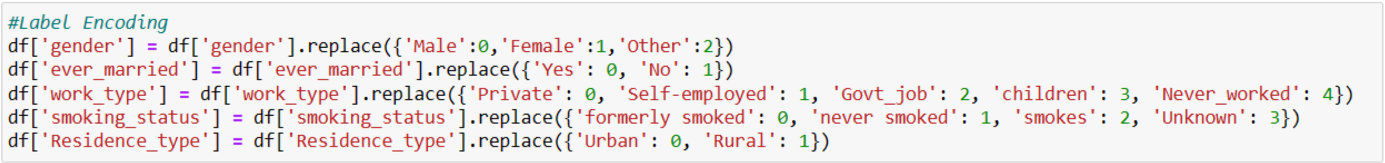


Figure 9- Identified outliers using boxplots.

3.3. Features Encoding

Only a few categorical factors (gender, ever married, work type, residence type, and smoking status) are present in the dataset. Since ML models require numerical characteristics as input, we used label-encoding to translate these properties to numerical values as given in the below Figure 10.



```
#Label Encoding
df['gender'] = df['gender'].replace({'Male':0,'Female':1,'Other':2})
df['ever_married'] = df['ever_married'].replace({'Yes': 0, 'No': 1})
df['work_type'] = df['work_type'].replace({'Private': 0, 'Self-employed': 1, 'Govt_job': 2, 'children': 3, 'Never_worked': 4})
df['smoking_status'] = df['smoking_status'].replace({'formerly smoked': 0, 'never smoked': 1, 'smokes': 2, 'Unknown': 3})
df['Residence_type'] = df['Residence_type'].replace({'Urban': 0, 'Rural': 1})
```

Figure 10-Converting categorical values into numerical values using Label Encoding.

3.3. Descriptive Analysis

We present summary statistics that include measures of position and central tendency in the section on descriptive analysis. These statistics provide a thorough overview of the dataset, highlighting important patterns, distributions, and data point dispersion. Our goal in conducting this inquiry is to identify fundamental patterns present in the dataset, which will serve as a basis for further investigation and analysis.

We examined the central tendency measures for age, body mass index (BMI), and average glucose level as the three main variables in the dataset. With an average BMI of 27.65 and an average glucose level of 89.18 mg/dL, the mean age was determined to be 40.80 years. These numbers provide information on the general health characteristics of the population being studied. The median age, BMI, and glucose level were also calculated and found to be 42.0 years, 87.09 mg/dL, and 27.6 kg/m², respectively. Because median values are unaffected by outliers, they are especially useful for skewed distributions. In addition, using the analysis of mode values for diverse categorical variables including smoking status, employment type, and hypertension, we were able to discern recurring patterns within the dataset. For example, the most common BMI was 28.7, the modal age was 45 years, and most individuals did not have a history of smoking, heart disease, or hypertension. In addition, the mode for job type showed that most participants worked in the private sector, while the mode for housing type showed that most of them lived in cities. These central tendency measures offer a thorough summary of the dataset, facilitating comprehension of the population under study's salient health-related and demographic features.

We looked at "measures of position," specifically percentiles, to analyze the distribution of the dataset's important variables and spot any possible outliers or extreme values. Insights into the relative positions of observations within a dataset are offered by percentiles, which facilitate the evaluation of central tendency and variability. We found that the age distribution was negatively skewed, with a median age of 42 and percentiles of 25 and 75 showing that, respectively, 25% and 75% of the population were under or equal to 22 and 58 years old. Likewise, there was negative skewness in the distribution of BMI and average glucose levels, with median values that were closer to the lower quartiles. In order to find predictors or risk variables in predictive modeling tasks and to direct future investigations or initiatives targeted at addressing health-related outcomes like stroke, it is imperative to comprehend these distributional features.

3.4. Normalization

We used the Min-max scaler technique to perform feature scaling on the numerical columns 'age', 'avg_glucose_level', and 'bmi' in order to guarantee uniformity and comparability in our research. By using this normalization procedure, the values of these attributes were changed to lie between 0 and 1, which helped to minimize any differences that may have arisen from different scales used for the variables. We normalized the features to have a mean of 0 and a standard deviation of 1 by using the equation z = (x — min) / (max — min), where z is the scaled output, x is the input data, min indicates the smallest value of the column, and max indicates the maximum value of the column. We next used percentiles to show the scaled distributions in order to better comprehend the characteristics of the dataset. The density plot of age after scaling is shown in Figure 11, where the 25th, 50th, and 75th percentiles are represented by percentile lines. Similarly, the same graph shows the density plot of the average BMI and glucose level, with percentile lines denoting the important percentiles for each variable. These graphics make it easier to understand the altered distributions and allow for a more thorough examination of the dataset.

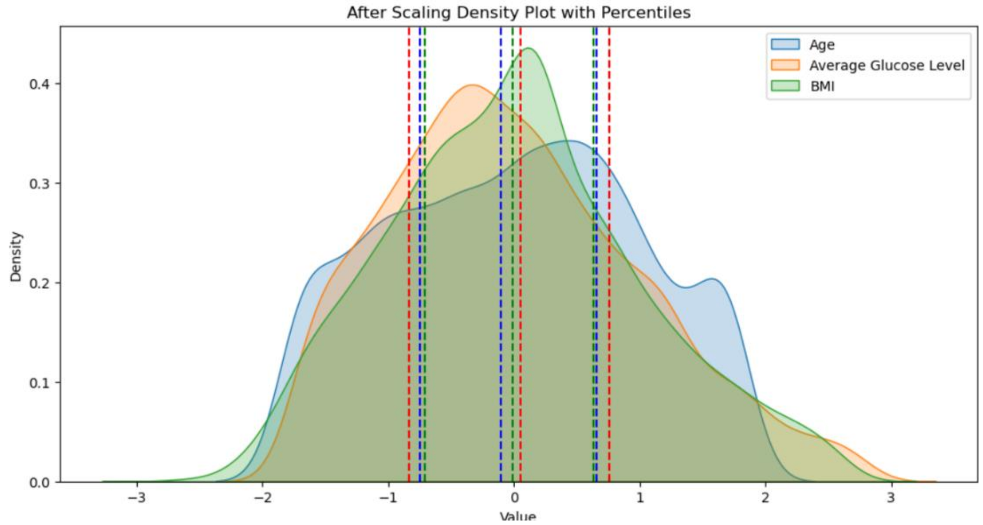


Figure 11-Distributed columns

3.5. Correlation Analysis

The heatmap's correlation values suggest that there exists a degree of association between all the features and the target variable (stroke), albeit at varying intensities. Age has the highest association (0.23) with stroke, but there are also moderate relationships seen with other variables, including hypertension, heart disease, work type, bmi, and smoking status. On the other hand, characteristics such as Residence_type and avg_glucose_level show less of a relationship with stroke.

Correlation values displayed in the Figure 12, we may want to remove id from our analysis, as their associations with the target variable are not as strong. But it's important to recognize that domain knowledge should also be taken into consideration, and correlation is only one indicator of feature value. As such, you must carefully consider how removing these features may affect your model's predictive capability.

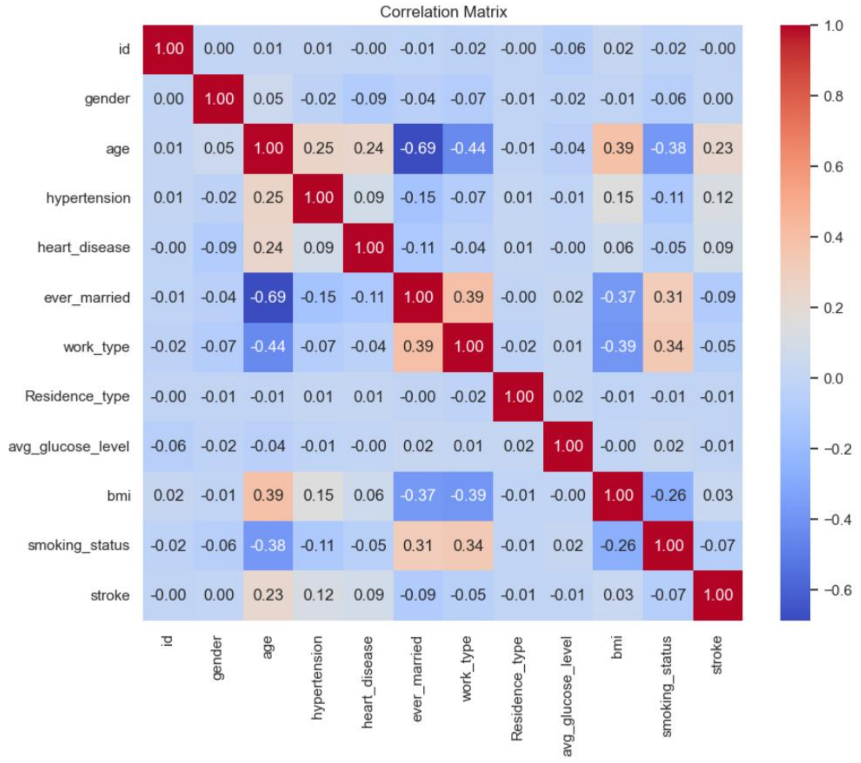


Figure 12-Heat map

3.6. Oversampling

Figure 13 illustrates the significantly unbalanced distribution of the dataset. Accurate prediction model training is hampered by imbalanced datasets. We used the Synthetic Minority Over-sampling Technique (SMOTE) from the imbalanced-learn module to perform oversampling in order to solve this problem. To balance the dataset, this approach creates synthetic samples of the minority class (stroke patients). Oversampling the minority class helps keep the model from being biased towards the majority class because the number of cases without strokes (class 0) is substantially higher than the number of examples with strokes (class 1). Thus, in order to address the problem of class imbalance, we applied SMOTE to the training data in order to generate synthetic instances of stroke cases.

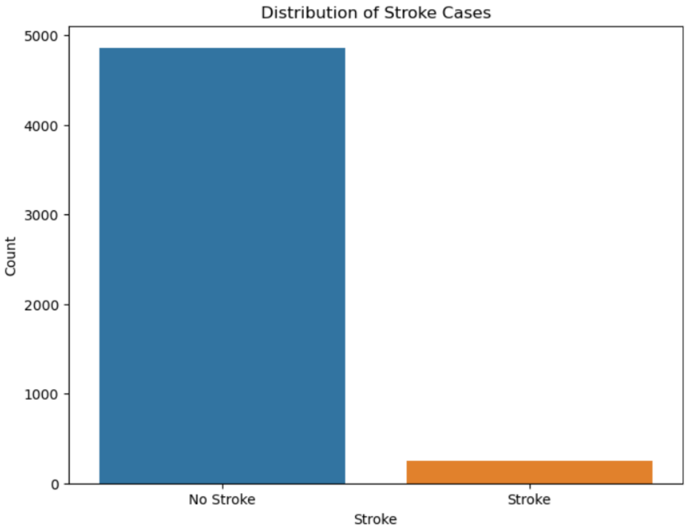


Figure 13-Strokes and non-strokes count.

3.6. Machine Learning Model

The model that will be applied in the stroke occurrence categorization framework is presented in this section. Using Scikit-Learn's train_test_split (80% for training and 20% for testing) technique, we divided the dataset, ensuring that the startify parameter was set to Yes.

3.6.1. Logistic Regression

Logistic regression (LR) is the model that will be included in the framework (Nusinovici *et al.*, 2020). This statistical classification technique was first created for binary tasks, but it has also been used to multi-class tasks. The output of the model is a binary variable, where $p = P(Y = 1)$ represents the likelihood that an instance belongs to the "Stroke" class, and $1 - p = P(Y = 0)$ represents the likelihood that an instance belongs to the "Non-Stroke" class.

3.7. Evaluation Metrics

As part of the ML model evaluation procedure, a number of performance measures were noted. We shall take into account the most widely utilized in the pertinent literature in the current analysis.

Recall, also known as true positive rate or sensitivity, is the percentage of participants who experienced a stroke and were correctly classified as positive, relative to all positive participants. When working with unbalanced data, precision and recall are more suited to pinpoint a model's faults. The precision shows the true number of stroke survivors in this class. Recall indicates the proportion of stroke patients who are accurately predicted. The F-measure, which summarizes a model's prediction ability, is the harmonic mean of precision and recall.

Figure 14 shows the confusion metrics structure.

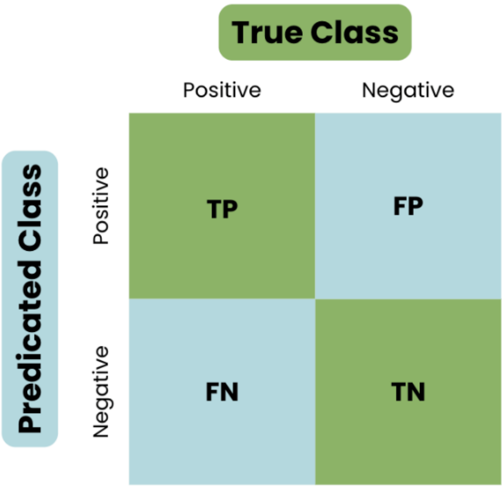


Figure 14-Confusion Metrics

Below equations shows method of calculating precision, recall, f1-score and figure 15 shows its values.

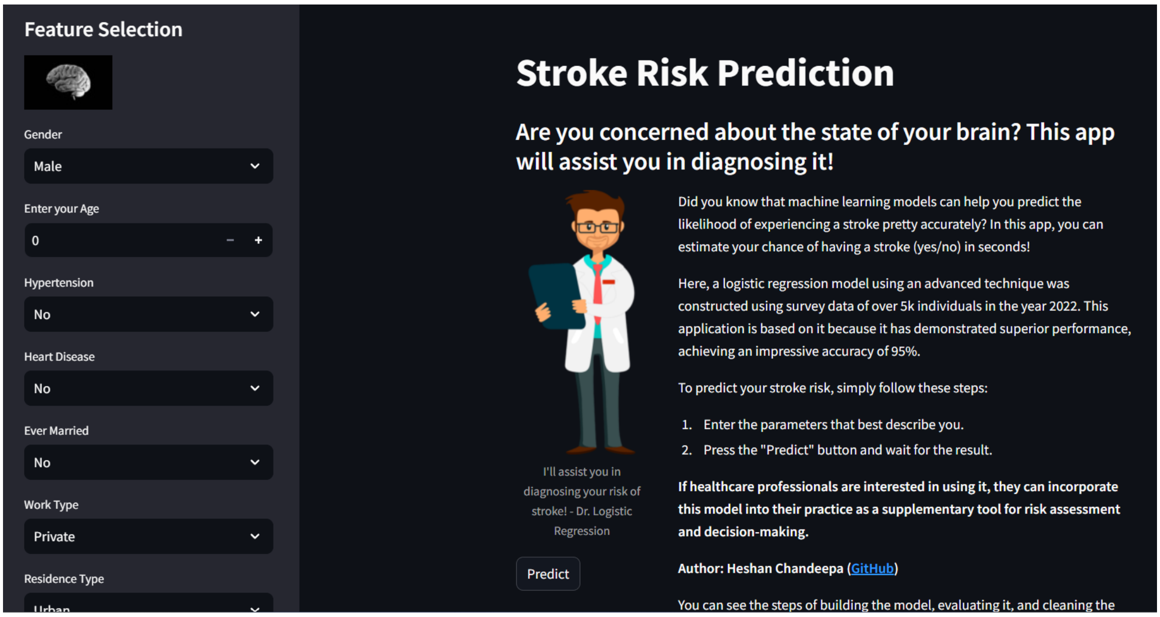
$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \qquad \text{Recall} = \frac{\text{TP}}{TP+FN,} \qquad \text{F-Measure} = 2.\frac{\text{Precision.Recall}}{\text{Precison+Recall,}} \qquad \text{Accuracy} = \frac{\text{TN+TP}}{\text{TN+TP+FN+FP}}$$

3.8. Web Application

The user interface of our web application, which may be accessed at https://stroke-risk-prediction-app.onrender.com/, is shown in Figure 16. A clear and user-friendly interface welcomes users and makes data entry for stroke risk prediction simple. Users can enter a variety of personal and health-related data into the interface's input fields, such as gender, age, marital status, occupation type, dwelling type, average glucose level, body mass index (BMI), and smoking status.

Our logistic regression model is used by the web application to calculate each user's risk score after they have input their pertinent data. The outcome gives the user important information about their risk profile by displaying the percentage chance of having a stroke within a given time range. With the help of this personalized risk assessment, individuals can make more informed decisions about their health and way of life and obtain a better understanding of their vulnerability to stroke.

Our web application is a useful tool for preventing strokes by utilizing sophisticated predictive analytics and integrating user-friendly features. Users may readily access and evaluate their risk projections through Figure 15 and the interactive features of the program, encouraging proactive health management and lowering the frequency of stroke-related problems.


Feature Selection

Gender: Male

Enter your Age: 0

Hypertension: No

Heart Disease: No

Ever Married: No

Work Type: Private

Residence Type: Urban

Stroke Risk Prediction

Are you concerned about the state of your brain? This app will assist you in diagnosing it!

Did you know that machine learning models can help you predict the likelihood of experiencing a stroke pretty accurately? In this app, you can estimate your chance of having a stroke (yes/no) in seconds!

Here, a logistic regression model using an advanced technique was constructed using survey data of over 5k individuals in the year 2022. This application is based on it because it has demonstrated superior performance, achieving an impressive accuracy of 95%.

To predict your stroke risk, simply follow these steps:

1. Enter the parameters that best describe you.
2. Press the "Predict" button and wait for the result.

If healthcare professionals are interested in using it, they can incorporate this model into their practice as a supplementary tool for risk assessment and decision-making.

Author: Heshan Chandeepa (GitHub)

You can see the steps of building the model, evaluating it, and cleaning the

I'll assist you in diagnosing your risk of stroke! - Dr. Logistic Regression

Figure 15-Interface of the web app

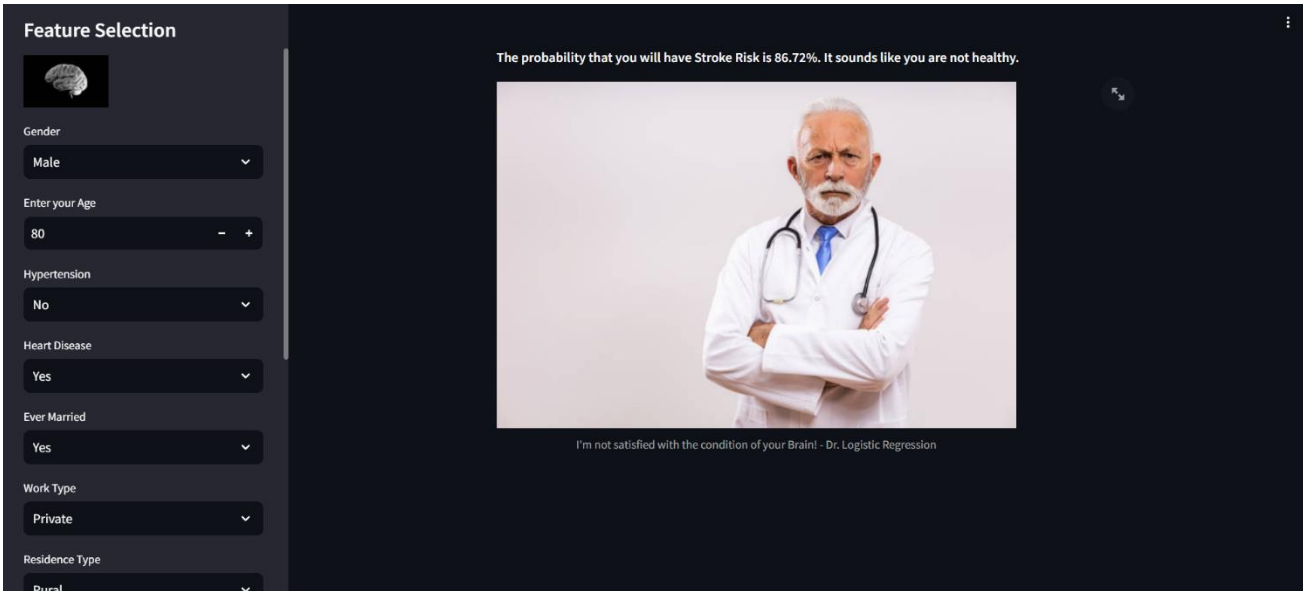
4.Results and Discussion

4.1. Experiments Setup

Apart from assessing the performance of the machine learning model, we also carried out tests to confirm the correctness and usefulness of our web application in practical situations. In order to guarantee consistency and dependability in our evaluations, we utilized data from the same dataset that was used to train and assess the model.

In one such test scenario, we input specific demographic and health-related information into the web application to simulate a user's query. For instance, we provided the following details: gender=Male, age=80, hypertension=0 (No), heart disease=1 (Yes), ever married=Yes, work type=Private, residence type=Rural, average glucose level=105.92, BMI=32.5, and smoking status=never smoked. These inputs show an example of a common situation seen in clinical practice.

After the input data was submitted, the web application used the underlying machine learning model to interpret the data and produce a tailored stroke risk prediction. The user saw the results of the program, which showed an 89.38% probability of stroke risk. Furthermore, a contextual message was supplied that highlighted the possible health consequences linked to the determined risk level. It is depicted in Figure 16.



Feature Selection

Gender: Male

Enter your Age: 80

Hypertension: No

Heart Disease: Yes

Ever Married: Yes

Work Type: Private

Residence Type: Rural

The probability that you will have Stroke Risk is 86.72%. It sounds like you are not healthy.

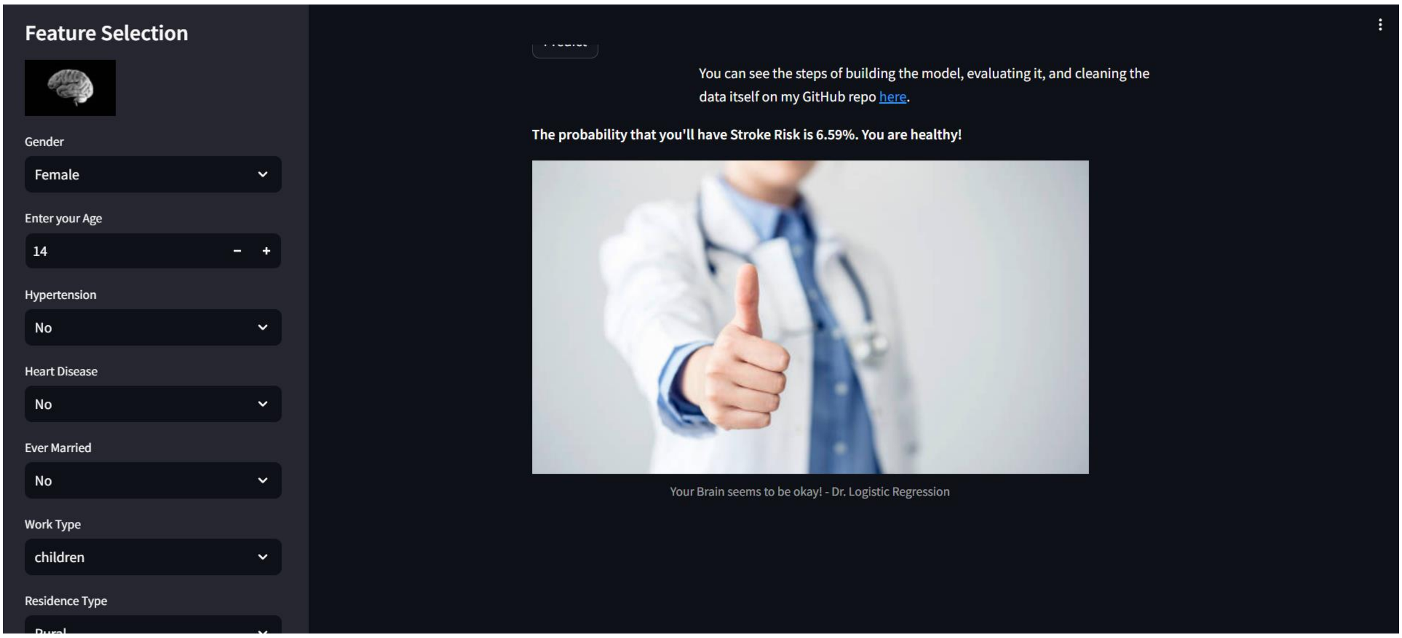
I'm not satisfied with the condition of your Brain! - Dr. Logistic Regression

Figure 16-High risk patient results.

By entering demographic and health-related data representative of a wide range of users, we sought to assess the accuracy and resilience of our online application in a different test scenario. In particular, we created a simulation of a situation in which the user's demographics were very different from those in the earlier test case.

For this test, we entered the following details into the web application: gender=female, age=14, hypertension=0 (No), heart disease=0 (No), ever married=no, work type=children, residence type=Rural, average glucose level=57.93, BMI=30.9, and smoking status=unknown. These inputs reflect a younger individual with no known medical conditions and an occupation categorized as "children," indicating a unique demographic profile.

After the input data was submitted, the web application used the underlying machine learning model to interpret the data and produce a tailored stroke risk prediction. The user saw the results of the program, which showed a low chance of 6.59% for stroke risk. In addition, the app gave a positive message indicating that the user's risk score indicates they are healthy. Figure 17 shows the results.



Feature Selection

Gender: Female

Enter your Age: 14

Hypertension: No

Heart Disease: No

Ever Married: No

Work Type: children

Residence Type: Rural

You can see the steps of building the model, evaluating it, and cleaning the data itself on my GitHub repo here.

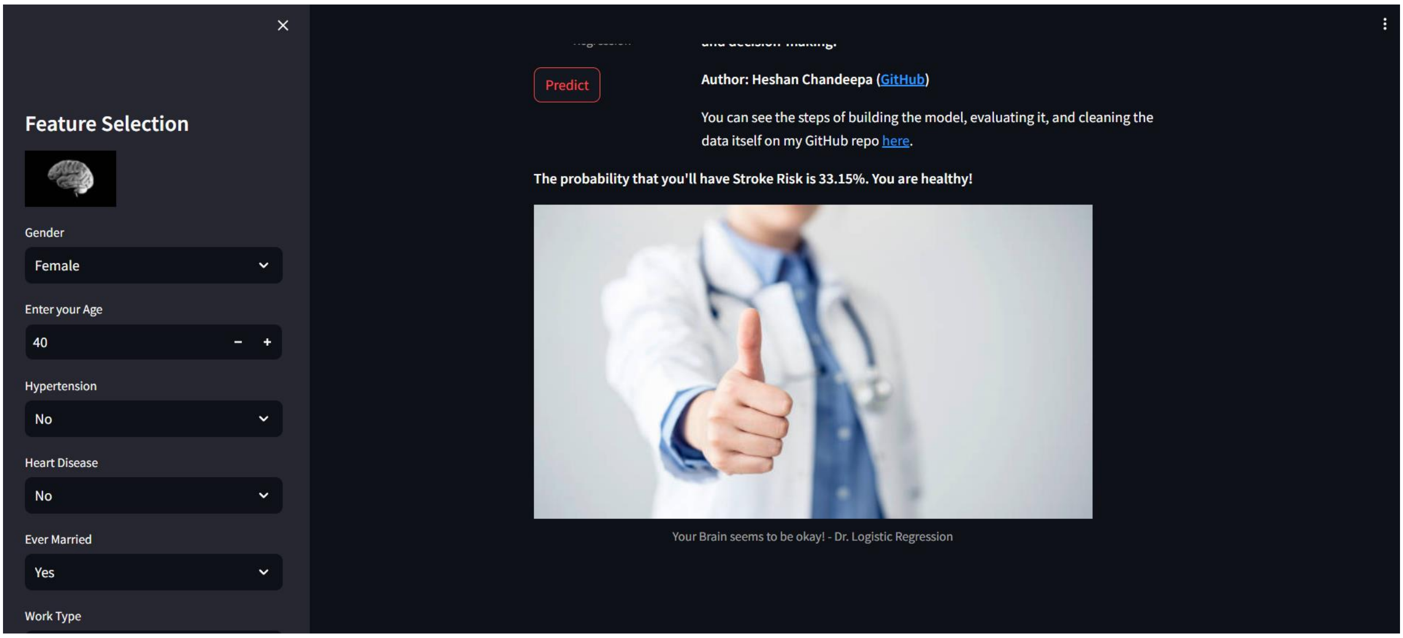
The probability that you'll have Stroke Risk is 6.59%. You are healthy!

Your Brain seems to be okay! - Dr. Logistic Regression

Figure 17-Low risk patient results.

Next, we entered the following details into the web application: gender=female, age=40, hypertension=0 (No), heart disease=0 (No), ever married=yes, work type=private, residence type=Rural, average glucose level=95.04, BMI=42.4, and smoking status=never smoked. These inputs reflect a younger individual with no known medical conditions and an occupation categorized as "Private", indicating a unique demographic profile.

After the input data was submitted, the web application used the underlying machine learning model to interpret the data and produce a tailored stroke risk prediction. The user saw the results of the program, which showed an 33.15% probability of stroke risk. Furthermore, a contextual message was supplied that highlighted the possible health consequences linked to the determined risk level. It is depicted in Figure 18.



Feature Selection

Gender: Female

Enter your Age: 40

Hypertension: No

Heart Disease: No

Ever Married: Yes

Work Type

Predict

Author: Heshan Chandeepa (GitHub)

You can see the steps of building the model, evaluating it, and cleaning the data itself on my GitHub repo here.

The probability that you'll have Stroke Risk is 33.15%. You are healthy!

Your Brain seems to be okay! - Dr. Logistic Regression

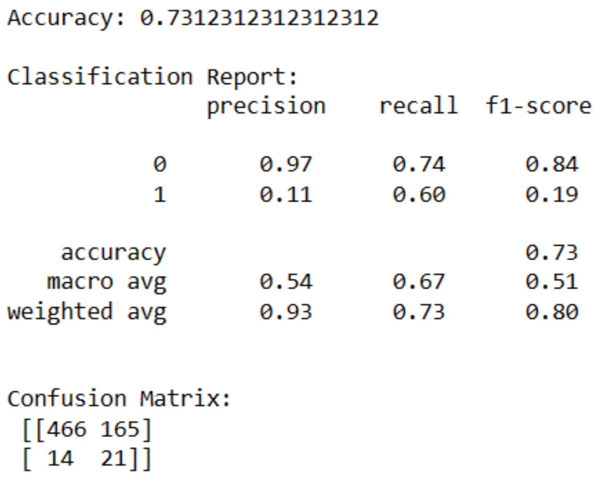
Figure 18 – normal aged person

These test findings support our belief that our web application is a trustworthy resource for managing our health and assessing our personal risk of stroke. Additionally, they stress the significance of utilizing cutting-edge machine learning methods to provide people with useful information for anticipatory healthcare decision-making. By putting our web application through such stringent testing and validation procedures, we hope to guarantee its usefulness and dependability in realistic situations, which will ultimately lead to better patient care and health outcomes.

4.1. Evaluation

We provide a thorough analysis of the machine learning model's predictive power for stroke risk in this section. First, we examine important performance measures like recall, precision, and F1-score, with a particular emphasis on the stroke class. Furthermore, we offer an analysis of the model's overall precision and efficacy in differentiating between positive (stroke) and negative (non-stroke) cases.

The machine learning model classified stroke risk with an overall accuracy of 73.12%, demonstrating its capacity to anticipate the outcome properly in most cases. We studied precision, recall, and F1-score parameters for the stroke class to assess its performance even more. It’s shown in the Figure 18.



```
Accuracy: 0.7312312312312312

Classification Report:
              precision    recall  f1-score

           0       0.97      0.74      0.84
           1       0.11      0.60      0.19

    accuracy                           0.73
   macro avg       0.54      0.67      0.51
weighted avg       0.93      0.73      0.80


Confusion Matrix:
 [[466 165]
 [ 14  21]]
```

Figure 18-Evaluation

The percentage of true positive predictions among all cases anticipated to be strokes is indicated by the model's precision of 0.11 for the stroke class (label 1). Recall, also known as sensitivity, is the percentage of genuine positive predictions among all real stroke cases; it is 0.60. The harmonic mean of these measures is 0.19, which represents the F1-score, which strikes a compromise between recall and precision.

A thorough analysis of the model's predictions is given by the confusion matrix, which highlights instances of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The model in this matrix detected 466 stroke cases (TP) properly but missed 14 stroke cases (FN). Furthermore, it properly identified 21 cases as non-stroke (TN) and wrongly classified 165 non-stroke instances as stroke (FP).

5.Conclusion

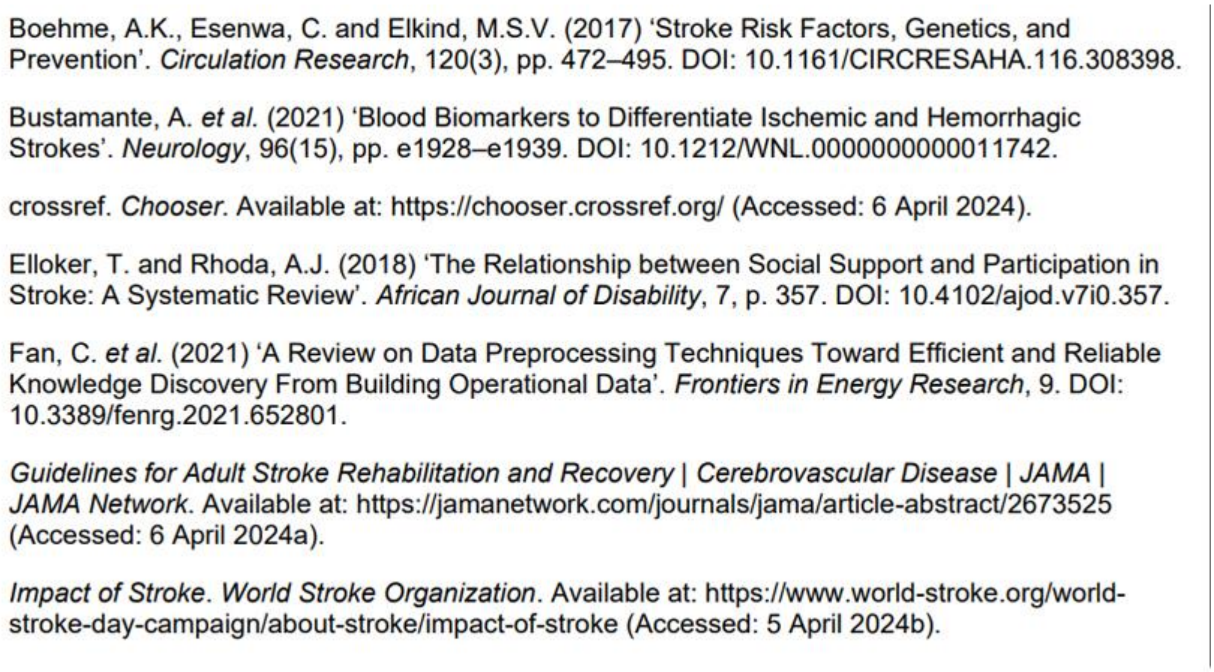
Finally, by applying a variety of demographic and health-related characteristics, our study concentrated on utilizing a logistic regression model to predict the risk of stroke. Our goal was to create a web application that would give people an easy-to-use way to determine their risk of stroke and take preventative action.

According to our research, the logistic regression model predicted the risk of stroke with a performance accuracy of 73.12%. Even while this is a significant step in the right direction toward early stroke diagnosis, there is still opportunity for improvement, especially in terms of resolving the imbalance in our dataset. To improve the model's performance and accuracy, more balanced datasets will be gathered for future implementations.

Our study's findings highlight the potential of machine learning to help predict strokes early and lessen their serious effects. Even though logistic regression produced encouraging findings, our model's prediction power can still be further increased by integrating deep learning techniques in future research projects.

Furthermore, we see a difficult but intriguing path in which we use brain CT scan imaging data to assess how well deep learning models forecast the occurrence of strokes. By adopting these developments, we hope to enhance patient outcomes in clinical practice and promote measures for preventing strokes.

References



Boehme, A.K., Esenwa, C. and Elkind, M.S.V. (2017) ‘Stroke Risk Factors, Genetics, and Prevention’. *Circulation Research*, 120(3), pp. 472–495. DOI: 10.1161/CIRCRESAHA.116.308398.

Bustamante, A. *et al.* (2021) ‘Blood Biomarkers to Differentiate Ischemic and Hemorrhagic Strokes’. *Neurology*, 96(15), pp. e1928–e1939. DOI: 10.1212/WNL.0000000000011742.

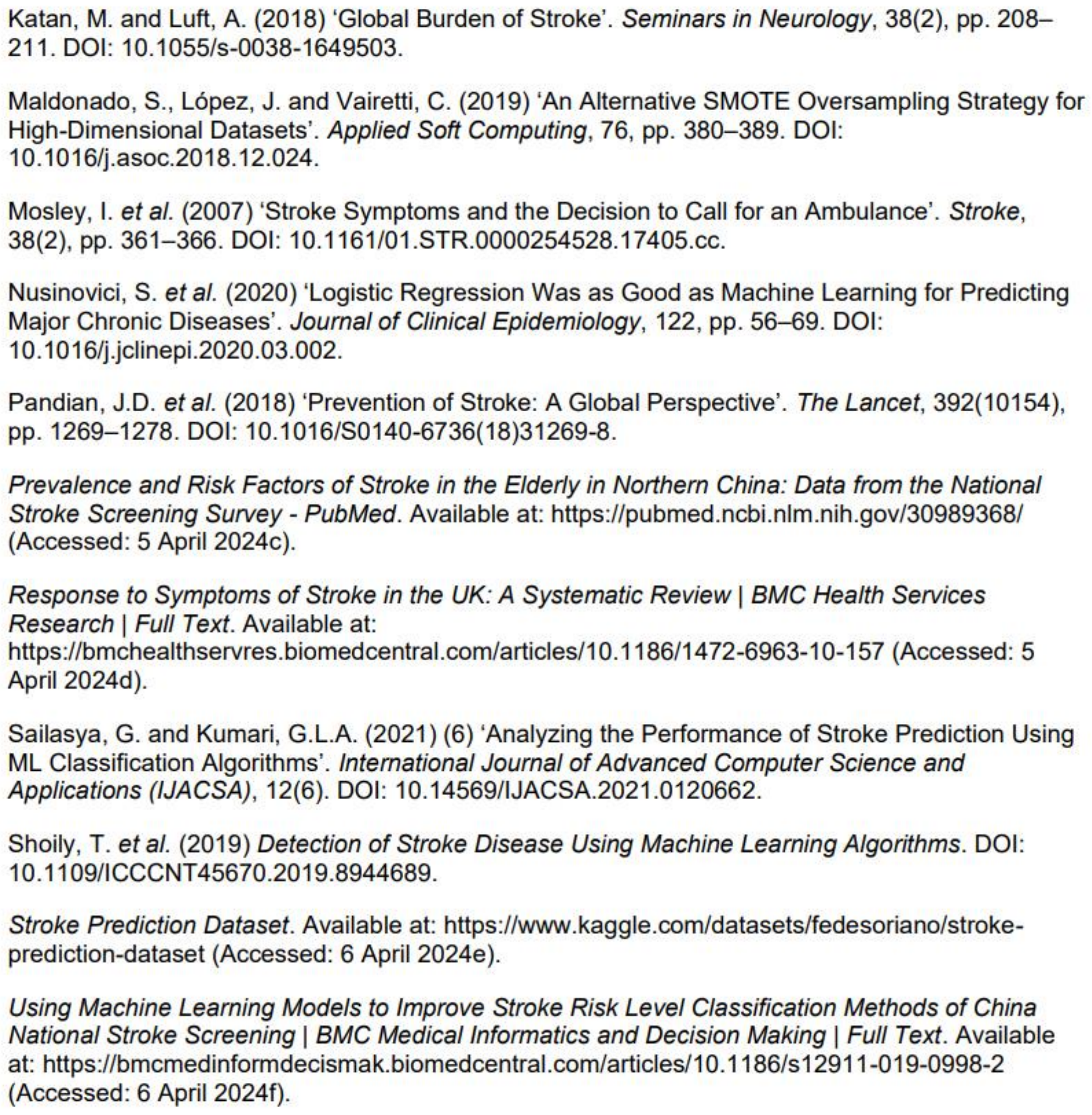
crossref. *Chooser*. Available at: https://chooser.crossref.org/ (Accessed: 6 April 2024).

Elloker, T. and Rhoda, A.J. (2018) ‘The Relationship between Social Support and Participation in Stroke: A Systematic Review’. *African Journal of Disability*, 7, p. 357. DOI: 10.4102/ajod.v7i0.357.

Fan, C. *et al.* (2021) ‘A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data’. *Frontiers in Energy Research*, 9. DOI: 10.3389/fenrg.2021.652801.

Guidelines for Adult Stroke Rehabilitation and Recovery | Cerebrovascular Disease | JAMA | JAMA Network. Available at: https://jamanetwork.com/journals/jama/article-abstract/2673525 (Accessed: 6 April 2024a).

Impact of Stroke. *World Stroke Organization*. Available at: https://www.world-stroke.org/world-stroke-day-campaign/about-stroke/impact-of-stroke (Accessed: 5 April 2024b).



Katan, M. and Luft, A. (2018) ‘Global Burden of Stroke’. *Seminars in Neurology*, 38(2), pp. 208–211. DOI: 10.1055/s-0038-1649503.

Maldonado, S., López, J. and Vairetti, C. (2019) ‘An Alternative SMOTE Oversampling Strategy for High-Dimensional Datasets’. *Applied Soft Computing*, 76, pp. 380–389. DOI: 10.1016/j.asoc.2018.12.024.

Mosley, I. *et al.* (2007) ‘Stroke Symptoms and the Decision to Call for an Ambulance’. *Stroke*, 38(2), pp. 361–366. DOI: 10.1161/01.STR.0000254528.17405.cc.

Nusinovici, S. *et al.* (2020) ‘Logistic Regression Was as Good as Machine Learning for Predicting Major Chronic Diseases’. *Journal of Clinical Epidemiology*, 122, pp. 56–69. DOI: 10.1016/j.jclinepi.2020.03.002.

Pandian, J.D. *et al.* (2018) ‘Prevention of Stroke: A Global Perspective’. *The Lancet*, 392(10154), pp. 1269–1278. DOI: 10.1016/S0140-6736(18)31269-8.

Prevalence and Risk Factors of Stroke in the Elderly in Northern China: Data from the National Stroke Screening Survey - PubMed. Available at: https://pubmed.ncbi.nlm.nih.gov/30989368/ (Accessed: 5 April 2024c).

Response to Symptoms of Stroke in the UK: A Systematic Review | BMC Health Services Research | Full Text. Available at: https://bmchealthservres.biomedcentral.com/articles/10.1186/1472-6963-10-157 (Accessed: 5 April 2024d).

Sailasya, G. and Kumari, G.L.A. (2021) (6) ‘Analyzing the Performance of Stroke Prediction Using ML Classification Algorithms’. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 12(6). DOI: 10.14569/IJACSA.2021.0120662.

Shoily, T. *et al.* (2019) *Detection of Stroke Disease Using Machine Learning Algorithms*. DOI: 10.1109/ICCCNT45670.2019.8944689.

Stroke Prediction Dataset. Available at: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset (Accessed: 6 April 2024e).

Using Machine Learning Models to Improve Stroke Risk Level Classification Methods of China National Stroke Screening | BMC Medical Informatics and Decision Making | Full Text. Available at: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-0998-2 (Accessed: 6 April 2024f).

GitHub Link of the Project

https://github.com/HChandeepa/Stroke_Prediction_System-Machine_Learning_Approach

Contribution

Lanka Pathmakumara – 10899186

I was crucial in leading our study project from the beginning to the end as the group leader. Developing the project concept and determining the problem description was my main contribution. With my training in data science, I was able to identify how important it was to use machine learning techniques to anticipate the risk of stroke and lessen its negative effects.

Finding and selecting the stroke dataset used for model training and assessment was one of my major contributions. I searched widely for pertinent datasets that included a range of demographic and health-related characteristics. Furthermore, I conducted correlation analysis on the dataset to identify meaningful relationships between various features and stroke occurrence. This analysis guided the feature selection process and informed the development of our machine learning model.

In addition, I helped with the dataset's normalization, which is an important step in assuring consistency throughout various attributes and standardizing the data. In order to improve machine learning algorithm performance and lessen the impact of outliers, normalization entails scaling numerical features to a consistent range. I contributed to maximizing the efficacy and efficiency of our machine learning model by standardizing the dataset.Making use of my experience with machine learning methods, I led the creation of the logistic regression model that predicts the risk of stroke. Furthermore, I led the development of the web application, ensuring its functionality and user-friendliness to facilitate easy access to our predictive model. In addition, I helped with a lot of the paperwork during the project, making sure that our procedures, conclusions, and insights were all properly recorded and arranged. My documentation efforts made it easier for the team to collaborate and gave us useful reference information for upcoming projects.

My responsibilities as the group leader included developing project concepts, obtaining datasets, doing correlation analyses,scaling, creating models, implementing web applications, and providing documentation assistance. I contributed to the accomplishments and significance of our research in the area of stroke prediction and prevention by taking the lead and being actively involved.

Ponnahennadige Dias – 10899285

I was instrumental in getting the dataset ready for analysis and prepping it for our research project. Making sure the dataset was clear, well-structured, and prepared for usage in our machine learning model was the main focus of my duties.

I carefully cleaned the dataset as part of the data pretreatment step, addressing missing values, outliers, and inconsistencies to guarantee data integrity and quality. I contributed to the removal of any potential biases and errors that might have affected our model's performance by carrying out exhaustive data cleaning procedures.

I also helped with feature encoding, which involved converting category variables into numerical representations that could be used as input by our machine learning algorithms. During this procedure, categorical features were encoded using methods like label encoding to help our model understand and make use of the data.

Throughout the project, I also contributed significantly to the documentation process by making sure that our data preprocessing methods were replicable and thoroughly documented.

As a whole, my efforts as a data pretreatment specialist on the project helped to establish the foundation for our machine learning model. I contributed to making sure that the predictions made by our model were accurate and dependable by carefully cleaning the data and encoding its features.

Gabbalage Dilshan - 10899287

I was a key contributor to the data visualization portion of our research project, using a variety of methods to clearly convey the insights from the dataset. I offered insightful explanations of the underlying patterns and trends in the data using visual aids including charts, graphs, and plots.

I produced aesthetically pleasing and educational visualizations that facilitated the study and comprehension of the dataset using programs like Matplotlib, Seaborn, and Plotly. These visual aids aided in determining correlations between various factors, spotting anomalies, and emphasizing significant patterns associated with the prediction of strokes.

I also made sure that our visualizations had the appropriate annotations and documentation, which helped with the documentation process. I contributed to making sure that members could easily access and comprehend the research findings by offering concise explanations and interpretations of the visuals.

Overall, the clarity and interpretability of our research findings were greatly improved by my efforts as a data visualization specialist. I contributed to the discovery of important discoveries and the facilitation of well-informed decision-making in the context of stroke prediction and prevention by using aesthetically appealing representations of the dataset.

Bathala Wicramasinhe - 10899497

My main responsibility for our study project was to perform a descriptive analysis on the dataset. In order to provide important insights into the structure and patterns of the data, this involved examining and summarizing its main features and distributions. I contributed to laying the groundwork for later phases of data preparation and model development by carrying out descriptive analysis.

Throughout the project, I offered comprehensive documentation help in addition to descriptive analysis. I made sure that all of our methods, processes, and conclusions were well recorded in order to promote openness and repeatability in our studies. My documentation efforts were vital to the team's ability to communicate our research method and findings effectively and to preserve unity and clarity.

Overall, my contributions as a member of the research team encompassed descriptive analysis and documentation support. Through my efforts, I helped enhance the quality and reliability of our research outcomes, contributing to the success and impact of our project in the field of stroke prediction and prevention.

Balasuriya Balasuriya - 10899180

As a member of the project team, my main duty was to assess our machine learning model's performance and offer insightful analysis of its efficacy. I carefully examined the predictive power of the model, evaluating its recall, accuracy, precision, and other pertinent parameters. I made sure our model was strong and trustworthy in estimating the risk of stroke by putting it through rigorous testing and validation processes.

My main duty in our research endeavor was to conduct a thorough performance evaluation of our machine learning model. This required examining the model's accuracy in predicting stroke risk and performing in-depth evaluations of its predictive capabilities.

I used a range of evaluation criteria, including accuracy, precision, recall, and F1-score, to complete this work.

In addition, I made a significant contribution to documentation efforts, making sure that our evaluation processes and conclusions were transparently and thoroughly documented. I made sure that our model evaluations were well documented, which helped to ensure accountability and reproducibility throughout the study process.

Overall, my contributions as an evaluator and documenter were essential in ensuring the credibility and effectiveness of our research outcomes. Through meticulous evaluation and documentation efforts, I helped lay the groundwork for impactful advancements in stroke prediction and prevention.