UNIVERSITY OF PLYMOUTH

Stroke Risk Prediction System

L B Heshan Chandeepa Pathmakumara

BSc (Hons) Data Science

Undergraduate

University Of Plymouth

10899186@students.plymouth.ac.uk

Contents

1.Introduction

The World Stroke Organization [1] estimates that 13 million people worldwide suffer a stroke each year, with 5.5 million dying as a result. Since it is the primary cause of mortality and disability globally, it has a significant impact on all facets of life. A stroke has an impact on the sufferer as well as their family, job, and social surroundings. Furthermore, it can affect anyone, at any age, regardless of gender or physical condition, against popular perception [2].

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 stroke: ischemic and hemorrhagic. 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 [3], [4].

The following factors increase the risk of stroke: previous history of a similar stroke; presence of a transient stroke; myocardial infarction; other heart diseases, such as heart failure, atrial fibrillation; age (stroke can occur at any age, even in children); hypertension; carotid stenosis from atherosclerosis; smoking; high blood cholesterol; diabetes; obesity; sedentary lifestyle; alcohol consumption; blood clotting disorders; estrogen therapy; and use of euphoric substances, such as cocaine and amphetamines [5], [6].

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 [7], [8].

A CT scan, or computed tomography, is used to diagnose stroke patients right away. Magnetic resonance imaging (MRI) is effective when used to diagnose ischemic stroke patients. There are two types of strokes: severe and mild. The initial twenty-four hours are critical in most cases. The diagnosis will highlight the course of treatment, which consists primarily of medication and occasionally surgery. When a patient enters a coma, the intensive care unit must do mechanical breathing and intubation [10,11].

Depending on the severity of the stroke, the majority of patients continue to experience problems even after they recover, including memory, concentration, and attention issues, trouble speaking or understanding speech, emotional issues like depression, loss of balance or walking ability, loss of sensation on one side of the body, and trouble swallowing food [9], [10].

Following a stroke, recovery aids in regaining lost function. A suitable strategy is devised with the assistance of neurologists, kinesiotherapists, and speech therapists to ensure the patient's prompt psychological and social recovery [10], [11]. To reduce the risk of stroke, blood pressure should be checked frequently, physical activity should be sustained, weight should be maintained normally, alcohol and tobacco use should be stopped, and a healthy diet low in fat and salt should be followed [12].

Technologies of information and communication (ICTs), particularly artificial intelligence (AI) and machine learning (ML), are becoming increasingly important in the early diagnosis of a number of diseases, including diabetes, hypertension, cholesterol, COVID-19, sleep disorders, hepatitis C, CKD, and others. We will be especially concerned with the stroke in the context of this investigation. Machine learning models have been used in numerous research studies for this particular condition.

This study presents a way for creating efficient binary classification machine learning models for the incidence of stroke. The synthetic minority over-sampling technique (SMOTE) [13] method was used since class balancing is essential for the formulation of effective algorithms in stroke prediction. Next, a number of models are created, set up, and evaluated using 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 demonstrated a strong interest in creating instruments and strategies for tracking and forecasting a wide range of illnesses that significantly affect human health. The most recent studies that predict stroke risk using machine learning approaches will be discussed in this section.
First, to accurately diagnose a stroke, the authors in [14] used four machine learning algorithms: naive Bayes, J48, K-nearest neighbor, and random forest. While the accuracy of the J48, K-nearest neighbor, and random forest classifiers was 99.8%, that of the naive Bayes classifier was just 85.6%.

Furthermore, to categorize stroke risk levels, [15] 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, [16] applies the Kaggle dataset [17]. 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 [17]. 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	A unique identifier of each patient
Gender	"Male", "Female" or "Other"
Age	Age of the patients (1-82)
Hypertension	0 if the patient doesn't have hypertension, 1 if the patient has hypertension
Heart_disease	0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
Ever_married	"No" or "Yes"
Work_type	"Children", "Govt_job", "Never_worked", "Private" or "Self-employed"
Residence_type	"Rural" or "Urban"
Avg_glucose_level	Average glucose level in blood
BMI(Kg/m2)	Body mass index
Smoking_status	"Formerly smoked", "never smoked", "smokes" or "Unknown"
Stroke	1 if the patient had a stroke or 0 if not

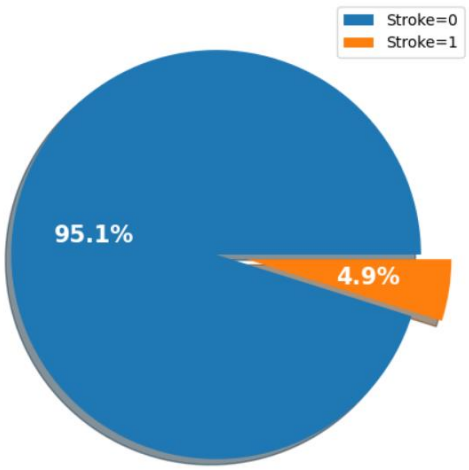

Figure 1-The visualizing count of classes (stroke and non-stroke) along with the percentage.

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.

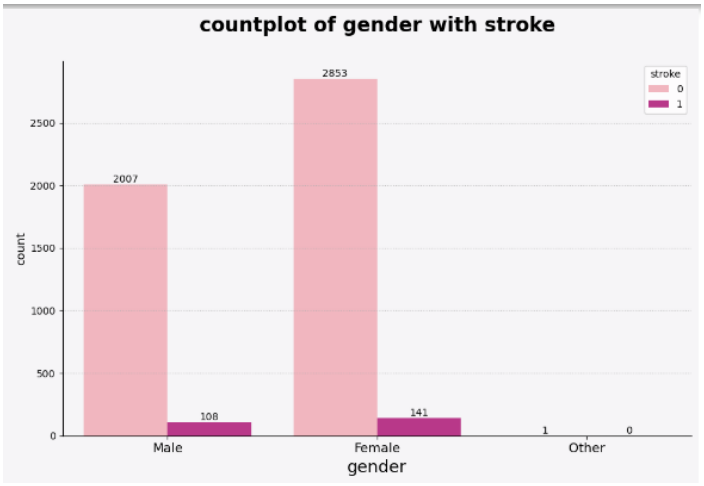

Figure 3- gender distribution of stroke.

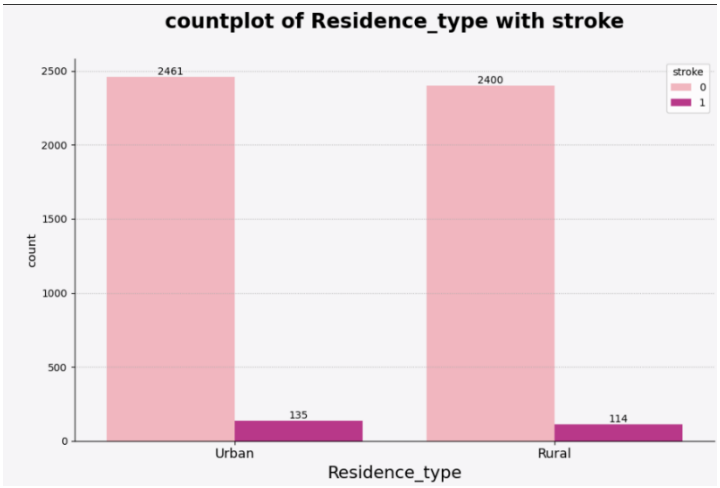

Figure 2-correlation between the type of residence.

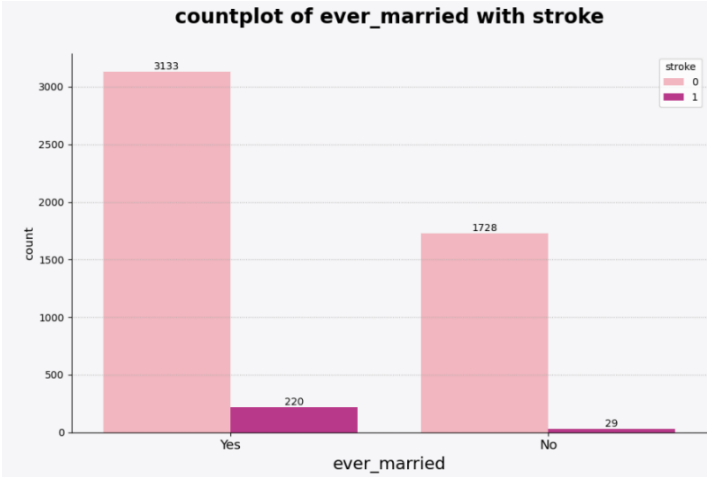

Figure 4- association between marital status and stroke.

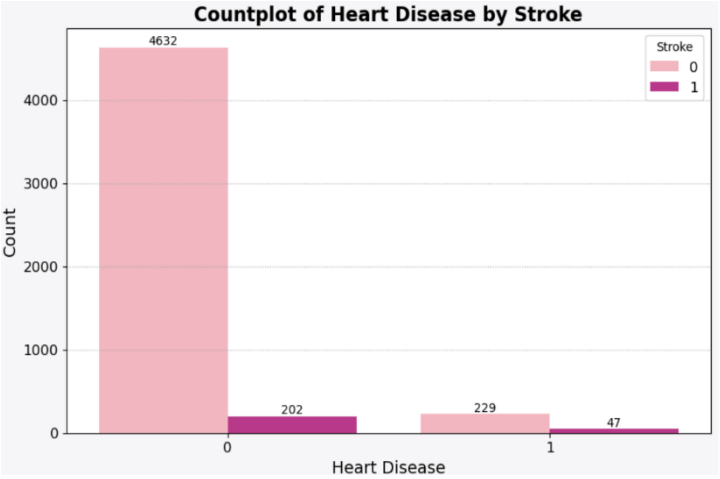

Figure 5- relationship between heart disease and strokes.

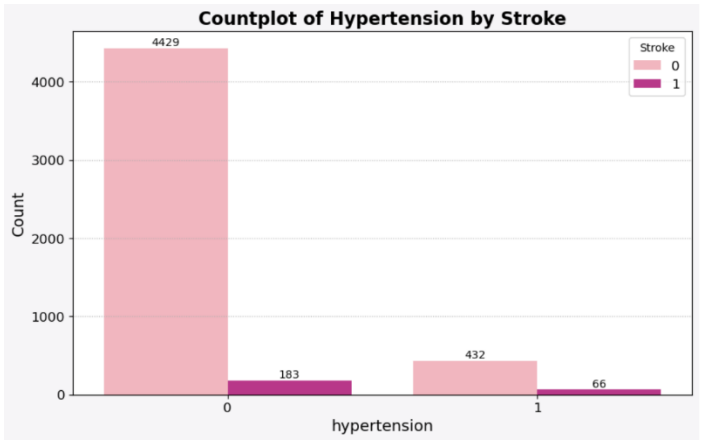

Figure 6- distribution of strokes by status of hypertension.

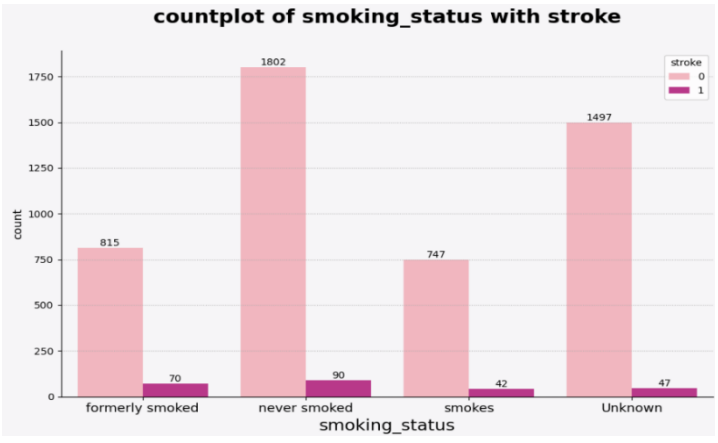

Figure 7-Smoking status and strokes

3.2. Data Preprocessing

Because of outliers and missing values, the final prediction quality may suffer from the quality of the raw data. In order to make data more suitable for mining and analysis, pretreatment steps such as feature selection, data discretization, and removal of redundant values are required [18]. Initially, we determined whether any null values existed in the dataset. Figure 8 displays the 201 null values for the bmi out of 5110 data entries. We utilized Scikit-learn's SimpleImputer to replace the null values with the missing values by utilizing the column mode. A univariate imputer called SimpleImputer from Scikit-learn may substitute missing values with an explanatory statistic (such the mean, median, or most common) along each column.

```
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.

Boxplots were used to find outliers in the 'bmi' and 'avg_glucose_level' columns as given in the Figure 9 . An iterative procedure was then used to remove the outliers. It's critical to manage outliers effectively since they have the potential to severely distort statistical analysis and machine learning models. In iterative approaches, outliers are found and eliminated iteratively until convergence is reached.

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.

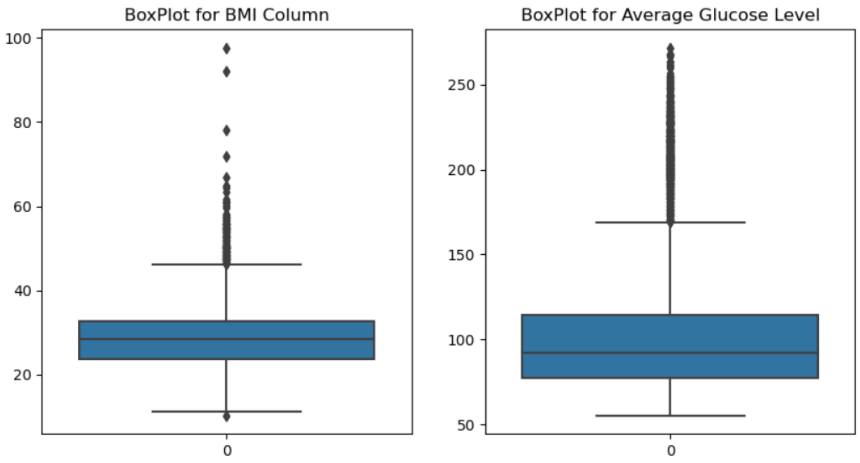

Figure 9- Identified outliers using boxplots.

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.

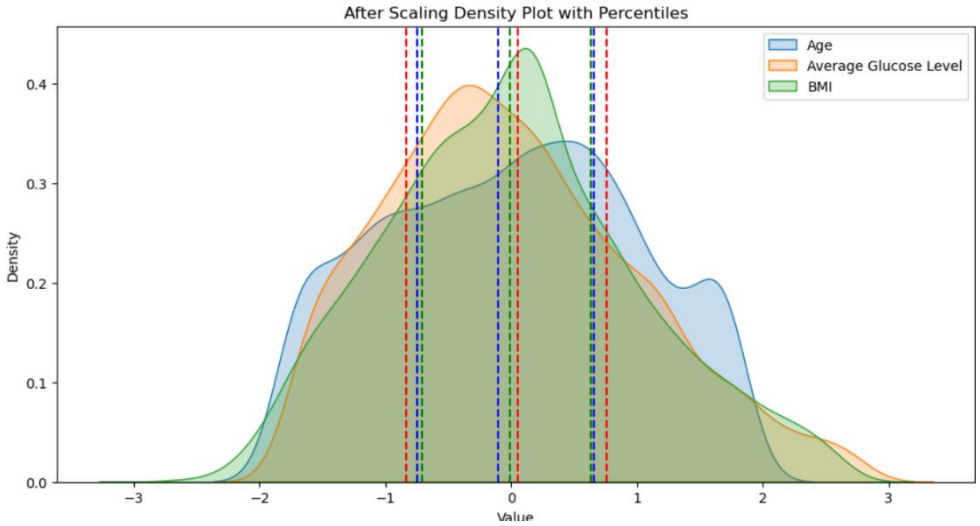

Figure 11-Distributed columns

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.

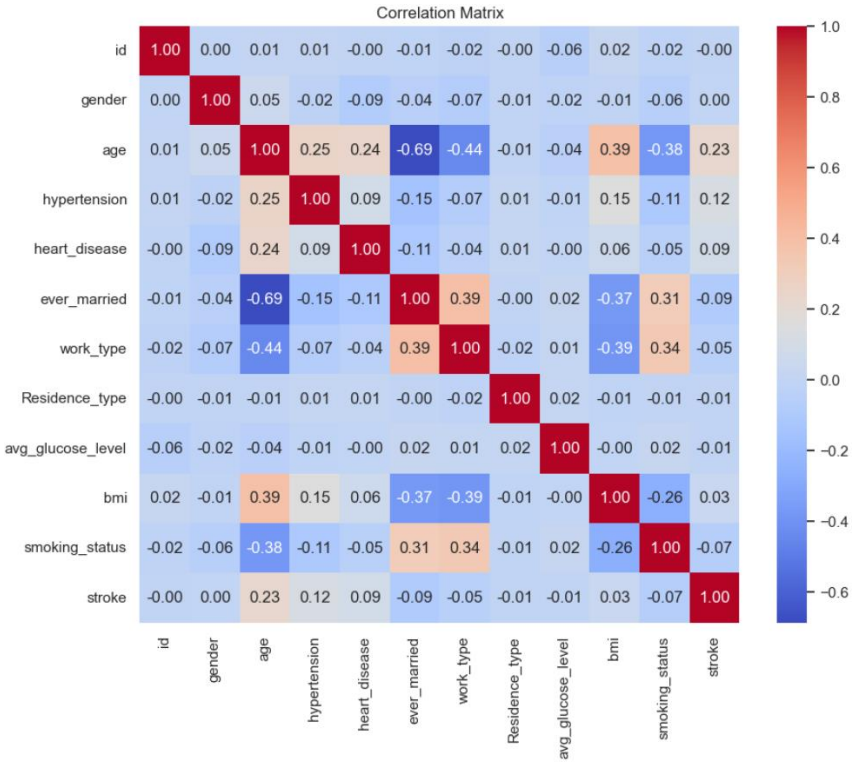

Figure 12-Heat map

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.

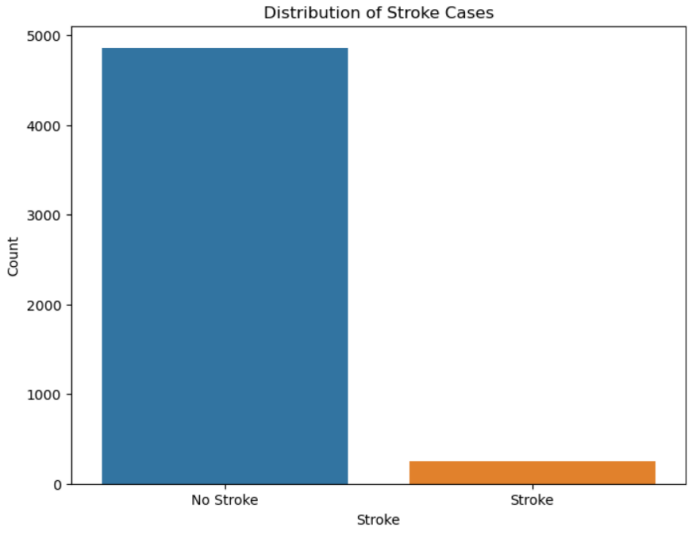

Figure 13-Strokes and non-strokes count.

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 [19]. 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 (true positive rate) or, otherwise, sensitivity, corresponds to the proportion of participants who had a stroke and were correctly considered as positive, with respect to all positive participants. Precision and recall are more suitable to identify the errors of a model when dealing with imbalanced data. Precision indicates how many of those who had a stroke actually belong to this class. Recall shows how many of those who had a stroke are correctly predicted. F-measure is the harmonic mean of the precision and recall and sums up the predictive performance of a model.

Figure 14 shows the confusion metrics structure.

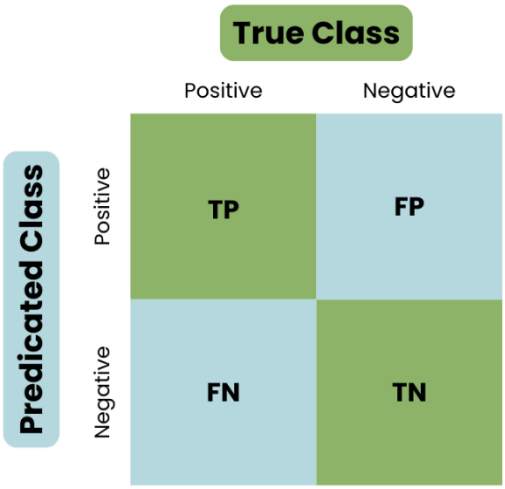

Figure 14-Confusion Metrics

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

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

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.

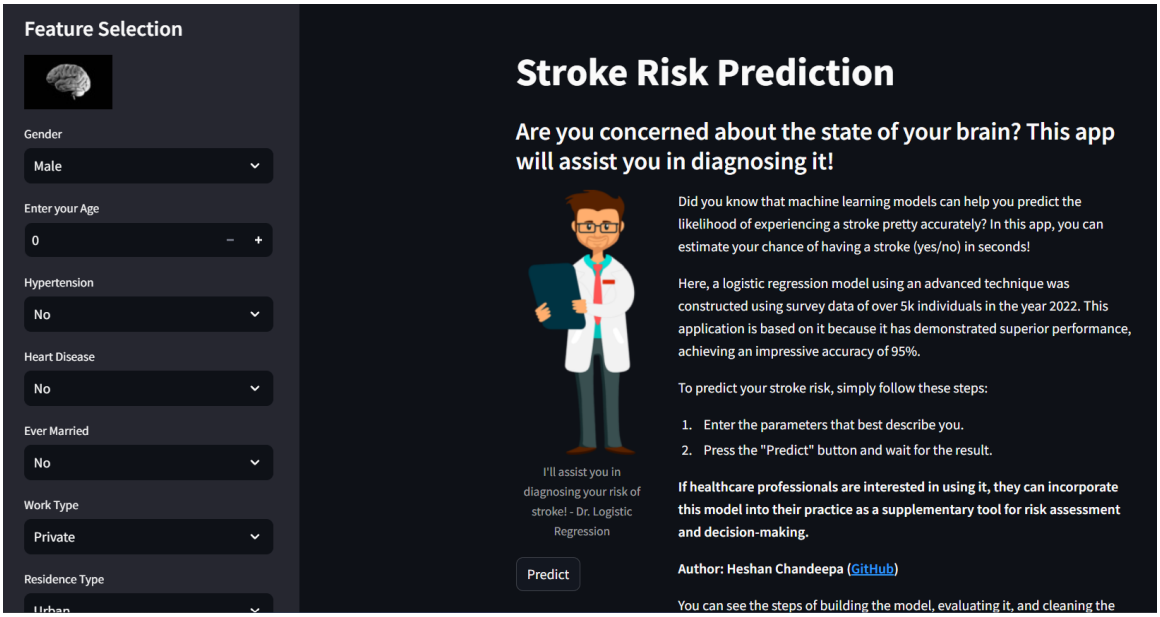

Figure 15-Interface of the web app

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.

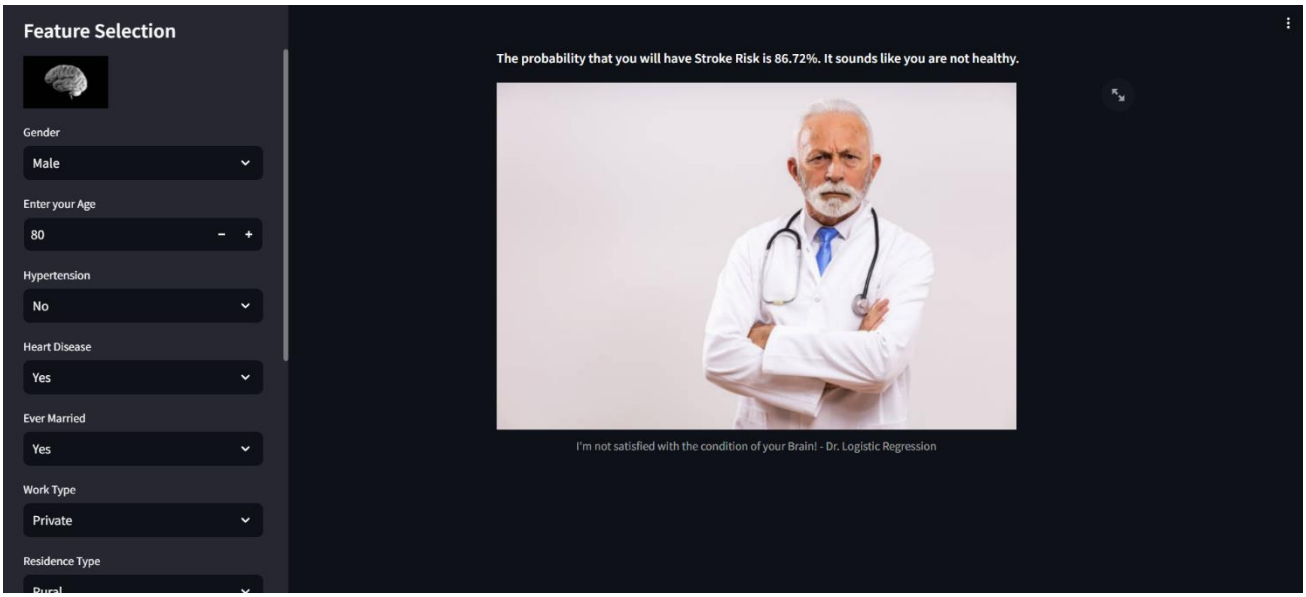

Figure 16-High risk patient results.

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.

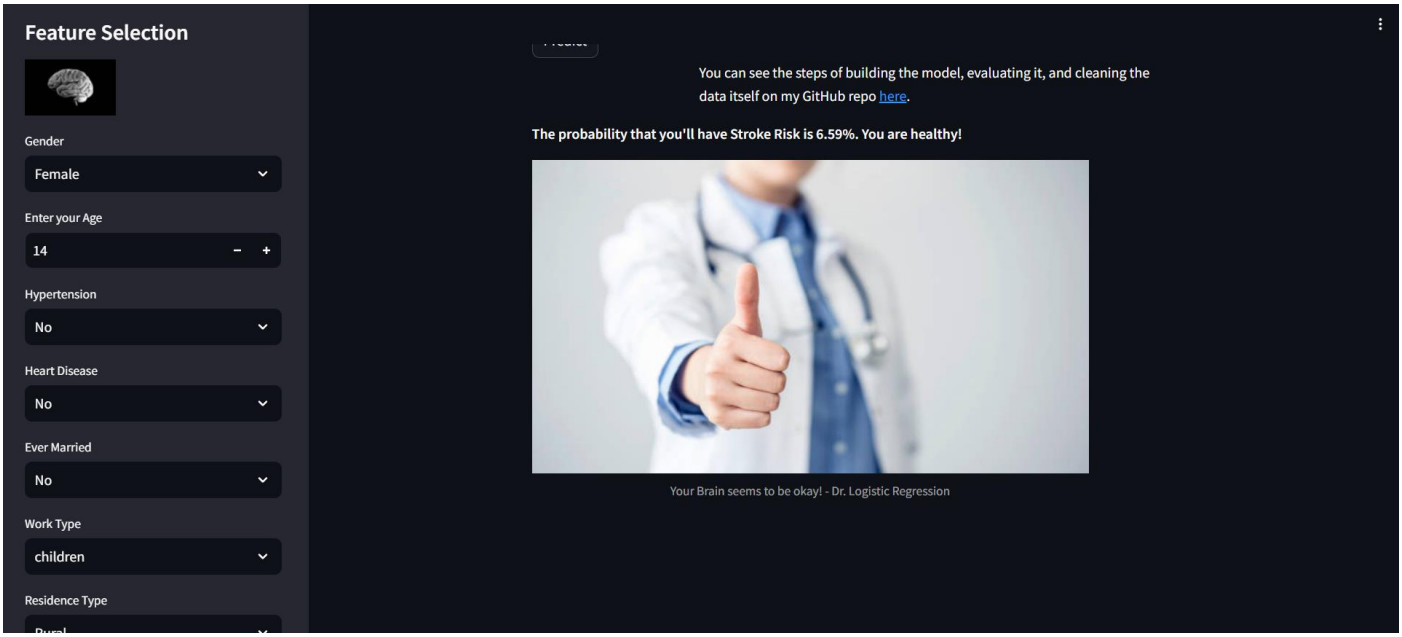

Figure 17-Low risk patient results.

These test findings support our belief that our web application is a trustworthy resource for managing your health and assessing your 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

[1] "Impact of Stroke," World Stroke Organization. Accessed: Apr. 05, 2024. [Online]. Available: https://www.world-stroke.org/world-stroke-day-campaign/about-stroke/impact-of-stroke

[2] T. Elloker and A. J. Rhoda, "The relationship between social support and participation in stroke: A systematic review," *Afr. J. Disabil.*, vol. 7, p. 357, 2018, doi: 10.4102/ajod.v7i0.357.

[3] M. Katan and A. Luft, "Global Burden of Stroke," *Semin. Neurol.*, vol. 38, no. 2, pp. 208–211, Apr. 2018, doi: 10.1055/s-0038-1649503.

[4] A. Bustamante *et al.*, "Blood Biomarkers to Differentiate Ischemic and Hemorrhagic Strokes," *Neurology*, vol. 96, no. 15, pp. e1928–e1939, Apr. 2021, doi: 10.1212/WNL.0000000000011742.

[5] "Prevalence and risk factors of stroke in the elderly in Northern China: data from the National Stroke Screening Survey - PubMed." Accessed: Apr. 05, 2024. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/30989368/

[6] A. K. Boehme, C. Esenwa, and M. S. V. Elkind, “Stroke Risk Factors, Genetics, and Prevention,” *Circ. Res.*, vol. 120, no. 3, pp. 472–495, Feb. 2017, doi: 10.1161/CIRCRESAHA.116.308398.

[7] I. Mosley, M. Nicol, G. Donnan, I. Patrick, and H. Dewey, “Stroke Symptoms and the Decision to Call for an Ambulance,” *Stroke*, vol. 38, no. 2, pp. 361–366, Feb. 2007, doi: 10.1161/01.STR.0000254528.17405.cc.

[8] “Response to symptoms of stroke in the UK: a systematic review | BMC Health Services Research | Full Text.” Accessed: Apr. 05, 2024. [Online]. Available: https://bmchealthservres.biomedcentral.com/articles/10.1186/1472-6963-10-157

[9] B. Delpont, C. Blanc, G. V. Osseby, M. Hervieu-Bègue, M. Giroud, and Y. Béjot, “Pain after stroke: A review,” *Rev. Neurol. (Paris)*, vol. 174, no. 10, pp. 671–674, Dec. 2018, doi: 10.1016/j.neurol.2017.11.011.

[10] crossref, “Chooser.” Accessed: Apr. 06, 2024. [Online]. Available: https://chooser.crossref.org/

[11] “Guidelines for Adult Stroke Rehabilitation and Recovery | Cerebrovascular Disease | JAMA | JAMA Network.” Accessed: Apr. 06, 2024. [Online]. Available: https://jamanetwork.com/journals/jama/article-abstract/2673525

[12] J. D. Pandian *et al.*, “Prevention of stroke: a global perspective,” *The Lancet*, vol. 392, no. 10154, pp. 1269–1278, Oct. 2018, doi: 10.1016/S0140-6736(18)31269-8.

[13] S. Maldonado, J. López, and C. Vairetti, “An alternative SMOTE oversampling strategy for high-dimensional datasets,” *Appl. Soft Comput.*, vol. 76, pp. 380–389, Mar. 2019, doi: 10.1016/j.asoc.2018.12.024.

[14] T. Shoily, T. Islam, S. Jannat, S. Tanna, T. Alif, and R. Ema, *Detection of Stroke Disease using Machine Learning Algorithms*. 2019, p. 6. doi: 10.1109/ICCCNT45670.2019.8944689.

[15] “Using machine learning models to improve stroke risk level classification methods of China national stroke screening | BMC Medical Informatics and Decision Making | Full Text.” Accessed: Apr. 06, 2024. [Online]. Available: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-0998-2

[16] G. Sailasya and G. L. A. Kumari, “Analyzing the Performance of Stroke Prediction using ML Classification Algorithms,” *Int. J. Adv. Comput. Sci. Appl. IJACSA*, vol. 12, no. 6, Art. no. 6, 30 2021, doi: 10.14569/IJACSA.2021.0120662.

[17] “Stroke Prediction Dataset.” Accessed: Apr. 06, 2024. [Online]. Available: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

[18] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, “A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data,” *Front. Energy Res.*, vol. 9, Mar. 2021, doi: 10.3389/fenrg.2021.652801.

[19] S. Nusinovici *et al.*, “Logistic regression was as good as machine learning for predicting major chronic diseases,” *J. Clin. Epidemiol.*, vol. 122, pp. 56–69, Jun. 2020, doi: 10.1016/j.jclinepi.2020.03.002.