PREDICTING BRAIN STROKE IN THE YOUNGER TO MIDDLE-AGED ADULTS AFTER WORKING ON THE UNBALANCED DATASET WITH THE SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE (SMOTE)

Venkata Siva Rao Kamisetty and Sreevidhya P V
Venkataramana Pittala and Veena Upadhye and Pallavi Vaswani
Geetaharichandana Vemuri
Indiana University- Purdue University, Indianapolis, IN 46202, USA
vkamiset@iu.edu, sreepv@iu.edu, vuupadhy@iu.edu, pvaswan@iu.edu

13 December 2022

1 Abstract

The project's goal is to forecast stroke occurrence based on the various risk factors in the given dataset. Similarly, this initiative tries to determine if people under the age of 60 are at high risk of stroke based on various criteria. This is required to assist society in better understanding of the lifestyle and healthcare.

2 Keywords

Stroke, independent variables, python, data analysis, Synthetic Minority Oversampling Technique (SMOTE)

3 Introduction

A stroke, also known as a transient ischemic attack or cerebrovascular accident, occurs when blood supply to the brain is interrupted. This results in a shortage of oxygen or nutrients reaching the brain from the blood. Without adequate blood flow, oxygen, and nutrition, cells die, leading in bleeding of brain cells. Early symptoms might range from modest weakness to paralysis, long-term impairment, and even death in certain cases. The two kinds of strokes are as follows:

· Hemorrhagic stroke: This is a deadly form of stroke. Occurs when a blood vessel in your brain bursts, allowing blood to enter brain tissues.

· Ischemic stroke: This happens when a blood vessel in the brain becomes clogged. A clot might be caused by a fat deposit, plaque, or cholesterol.

According to the Centers for Disease Control and Prevention, stroke is one of the major causes of mortality in the United States. Every year, over 795,000 people have a stroke. Similarly, a stroke occurs every 40 seconds, and someone dies because of a stroke every 3.5 minutes. High blood pressure, smoking, cholesterol, and diabetes are all major risk factors for stroke. Therefore, early detention is critical. As the saying goes, "prevention is better than cure," and the odds of survival are higher when therapy begins early. A stroke, also known as a transient ischemic attack or cerebrovascular accident, occurs when blood supply to the brain is interrupted. This results in a shortage of oxygen or nutrients reaching the brain from the blood. Without adequate blood flow, oxygen, and nutrition, cells die, leading in bleeding of brain cells. Early symptoms might range from modest weakness to paralysis, long-term impairment, and even death in certain cases. The two kinds of strokes are as follows:

· Hemorrhagic stroke: This is a deadly form of stroke. Occurs when a blood vessel in your brain bursts, allowing blood to enter brain tissues.

· Ischemic stroke: This happens when a blood vessel in the brain becomes clogged. A clot might be caused by a fat deposit, plaque, or cholesterol.

According to the Centers for Disease Control and Prevention, stroke is one of the major causes of mortality in the United States. Every year, over 795,000 people have a stroke. Similarly, a stroke occurs every 40 seconds, and someone dies because of a stroke every 3.5 minutes. High blood pressure, smoking, cholesterol, and diabetes are all major risk factors for stroke. Therefore, early detention is critical. As the saying goes, "prevention is better than cure," and the odds of survival are higher when therapy begins early.

4 Aim

Our main goal is to predict the possibility of a cerebrovascular accident occurring in people under the age of 60 by analyzing variables such as gender, smoking status, hypertension, previous conditions, work status, and residence type and balancing the dataset using the Synthetic Minority Over-sampling Technique (SMOTE).

The project is based on two hypotheses, which is listed below:

- Null Hypothesis: The cause of stroke among people under the age of 60 is unrelated to the dataset's independent factors such as gender, hypertension, job status, residence type, and so on. Alternate Hypothesis: Gender, hypertension, job status, home type, and other factors in the dataset are associated with the etiology of stroke in adults under the age of 60.

5 Purpose

The purpose of the research is to predict stroke occurrence based on parameters such as smoking status, age, BMI, hypertension, and heart disease. We believe that through defined research, we can forecast the occurrence of stroke in the stated age range, which will assist us to acquire an idea on the incidence of stroke, lowering our ability to conquer this sort of condition. Medically, all factors such as hypertension, smoking status, and heart disease have an impact on stroke attacks. This initiative will assist us in determining the cause of strokes and in adopting healthier living choices. Similarly, this effort will assist the larger population in taking care of their cardiovascular health system.

6 Methodology

Our project's main goal is to predict the likelihood that a cerebrovascular accident will occur in people under the age of 60 by looking at factors including gender, smoking status, hypertension, prior illnesses, employment status, and type of residence, utilizing the Synthetic Minority Oversampling Technique, as well as database technologies like Python jupyter notebook. Numerous steps make up our project's process, which will be covered in greater detail further in the report.
1) Data Collection
2) Data Extraction
3) Data Cleaning
4) Data Visualisation
5) Development of Models
6) SMOTE
7) Data Visualisation after Smote
8) Data Analysis after Smote The independent variables chosen are as follows:

1. Gender
2. Age
3. Hypertension
4. Heart disease
5. Ever married
6. Work type
7. Residence type
8. Average glucose
9. BMI
10. Smoking status

The dependant variable is:

1.Stroke

7 ORIGINAL TEAM MEMBERS AND RESPONSIBILITIES:

The interests and skill sets in our group are varied. The team members we formed for the project are listed below, along with the roles we decided on at its outset.

Name	Background	Roles
0.63	1.263	6.7

8 ACTUAL CONTRIBUTIONS FROM INDIVIDUAL TEAM MEMBERS:

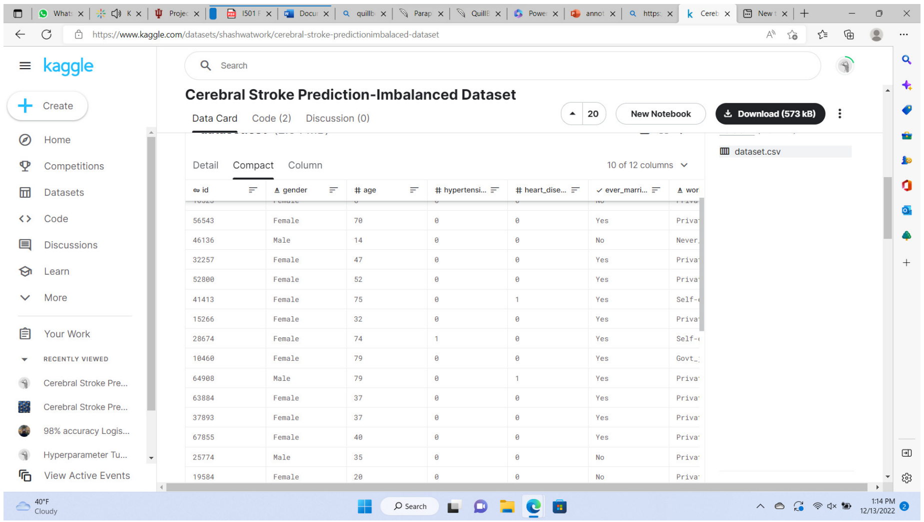
Name	Background	Roles
Venkata Siva Rao Kamisetty	Bachelor in Pharmacy	New to Python, SQL Project Proposal, Data transformat

9 PROJECT CHALLENGES:

There were numerous hurdles that we faced as a team while working on this project. The biggest issue was the time constraints. Everyone has distinct obligations. As a result, the time schedules in group meetings did not coincide. Secondly, because everyone came from a medical background and was new to informatics, it was tough to complete the coding and grasp what all models would work. It was challenging to grasp the codes and execute SMOTE in the data set because it was not part of the curriculum. However, our professor was quite helpful in this regard; he recommended numerous relevant websites from which we learned how to execute SMOTE. Overall, there were numerous minor problems that we overcame collectively.

10 DATA EXTRACTION AND STORAGE:

The Cerebral Stroke Prediction-Imbalanced Dataset was extracted using Kaggle. There are 12 characteristics in the dataset, one of which has an uneven target column. It contains information about Cerebral Stroke in 43400 rows and 13 columns. The data is formatted in such a way that a main key, such as id, can be used to retrieve numerous values,



Cerebral Stroke Prediction-Imbalanced Dataset

Data Card Code (2) Discussion (0)

Detail Compact Column 10 of 12 columns

id	gender	age	hypertensi...	heart_dise...	ever_marri...	wor
56543	Female	70	0	0	Yes	Priva
46136	Male	14	0	0	No	Never_
32257	Female	47	0	0	Yes	Priva
52800	Female	52	0	0	Yes	Priva
41413	Female	75	0	1	Yes	Self-
15266	Female	32	0	0	Yes	Priva
28674	Female	74	1	0	Yes	Self-
10460	Female	79	0	0	Yes	Govt_
64908	Male	79	0	1	Yes	Priva
63884	Female	37	0	0	Yes	Priva
37893	Female	37	0	0	Yes	Priva
67855	Female	40	0	0	Yes	Priva
25774	Male	35	0	0	No	Priva
19584	Female	20	0	0	No	Priva

Figure 1: The dataset from Kaggle

and no column can be empty because the id column has a ”not null” condition. It supports a variety of data types, including integer for id, varchar for names, decimal for point values, and Boolean. It is a dataset record in which data can be evaluated based on gender, past health state, and remarks. There were 783 stroke patients and 42557 non-stroke cases in the sample.

11 DATA CLEANING :

Reviewing the chosen qualities was part of the data cleansing procedure. Due to the high number of null values in our data, this step was crucial. In order to detect erroneous values, we went through each row of information in Excel as part of the cleaning process.

The data cleaning process entails a number of steps, including the following components:
a)Filling all the blank values in the smoking status column as not known
b)The next step entails transforming the categorical data into a numeric form where the independent variables smoking status, gender, ever married, work type, and residence type are converted to numeric form.
c) In addition, we divided the age range in the column into three groups: those under the age of one to thirty, those between the ages of thirty to sixty, and those who are sixty years of age and over.
d)A column identifying the average glucose group has been added, where the average glucose levels are divided into three categories

e) A new column called BMI Group has been created, which groups BMI values into four categories: less than 18.5 as group 1, 18.5 to 24.9 as group 2, 25 to 29.9 as group 3, and all above 30 as group 4.
f) Removal of the ID, Age, Average Glucose Level, and BMI columns

12 DATA ANALYSIS:

Using Python, we implemented the following models: Linear Regression, LASSO (Least Absolute shrinkage and selection operator) and KNN classification for data analysis. The following steps were taken to analyze the data:

13 Correlation Among Attributes

14 DATA VISUALIZATION BEFORE SMOTE

Python is used to visualize our data,we imported many packages. The packages include Numpy for calculations, Matplotlib for visualization, Seaborn for sophisticated visualizations, and Pandas for processing. For showing the correlation, we used Heat map and Multivariable Bar chart. We created a correlation Heat map, comparing Stroke with Hypertension, Heart disease, ever married, Work type, Residence type, Smoking status, Age group. Average glucose group, BMI group.

15 Visualization code for Heat map:

16 Heat map:

According to the heat map, stroke risk is strongly connected with age group and adversely correlated with smoking status. Age group: 0.15 Smoking status: -0.28 Based on the correlated values from the Heat map, we created a Multivariable bar chart with stroke and age group and Stroke and Smoking status

17 Code for multivariable bar chart showing stroke and age group:

18 Multi variable Bar chart ‘Stroke and age group’:

1-stroke 0-non stroke The bar chart demonstrates the extreme imbalance in the data set, which has more non-stroke instances than stroke patients.

19 Code for multivariable bar chart showing stroke and smoking status:

20 Multi variable Bar chart ‘Stroke and smoking status’:

21 DATA ANALYSIS:

Steps to Test the Normality of the Data: The normality test shows us that the p value is ¡0.05, thus variables are not normally distributed. Since the data is not distributed, we performed Chi-square testing.
After performing Chi-square testing, we rejected the null hypothesis.

22 Model Development:

Using Python, we implemented the following models: Linear Regression, LASSO (Least Absolute shrinkage and selection operator) and KNN classification for data analysis. The following steps were taken to analyze the data:

Splitting the data set into training and testing data, divided at a ratio of 70 to 30.

23 Logistic Regression(Before SMOTE):

24 KNN classification(Before SMOTE)

25 Lasso Regression (Least Absolute Shrinkage and Selection Operation)

Inference: According to the above models, the coefficient of determination in KNN classification is a negative value, which portrays that there is no direct correlation between the variables and the occurrence of stroke (because of imbalanced data).

26 SMOTE:

Using SMOTE to balance the dataset:
The dataset is now balanced with After SMOTE : 0:29531,0:29531

27 DATA VISUALIZATION (AFTER SMOTE):

28 Code for Heat map:

29 Heat map:

Even after SMOTE, the correlation of Hypertension, work type, age group is positive and correlation of smoking status is negative.

30 Multi variable Bar chart after SMOTE:

Code for stroke, hypertension Barchart
Code for stroke, work type:
Code for stroke, age group:
Code for stroke, smoking status:

31 Logistic regression (After SMOTE):

32 Receiver Operating Characteristic Curve (ROC) after SMOTE

33 KNN Classification (After Smote):

34 LASSO Regression after SMOTE:

Inference: The Coefficient of determination has been reduced more negatively after applying SMOTE , in logistic regression, KNN classification, and LASSO.

35 Testing the Hypothesis:

The Coefficient of determination has been reduced more negatively after applying SMOTE , in logistic regression, KNN classification, and LASSO.

36 SUMMARY:

Both our regression and classification models demonstrated how the parameters we studied were related to stroke incidence. This means we can reject our null hypothesis: The factors investigated have a connection to stroke cases.

37 LIMITATIONS:

- The dataset is skewed, with many non-stroke patients (and fewer stroke instances.
- Medically, smoking status influences the occurrence of stroke, however heat maps show a negative correlation.
- An unbalanced data set has made the model more susceptible to most instances, resulting in bias.
- The model's accuracy reduced after applying SMOTE, as is anticipated, but the R2 value also decreased, which should not have happened.
- The coefficient of determination is more negative after using SMOTE, indicating that the model is still underperforming even after using SMOTE.

38 APPENDIX

39 REFERENCE:

Jang, D. E., Shin, J. H. (2019). Self-Care Performance of Middle-Aged Stroke Patients in Korea. Clin Nurs Res, 28(3), 263-279. https://doi.org/10.1177/1054773817740670

Liu, T., Fan, W., Wu, C. (2019). A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset. Artif Intell Med, 101, 101723. https://doi.org/10.1016/j.artmed.2019.101723

Virani, S. S., Alonso, A., Aparicio, H. J., Benjamin, E. J., Bittencourt, M. S., Callaway, C. W., Carson, A. P., Chamberlain, A. M., Cheng, S., Delling, F. N., Elkind, M. S. V., Evenson, K. R., Ferguson, J. F., Gupta, D. K., Khan, S. S., Kissela, B. M., Knutson, K. L., Lee, C. D., Lewis, T. T., . . . Tsao, C. W. (2021). Heart Disease and Stroke Statistics-2021 Update: A Report From the American Heart Association. Circulation, 143(8), e254-e743. https://doi.org/10.1161/cir.0000000000000950

Centers for Disease Control and Prevention. (2022, October 14). Stroke facts. Centers for Disease Control and Prevention. Retrieved December 13, 2022, from https://www.cdc.gov/stroke/facts.htm

YouTube. (2019, May 8). Smote (synthetic minority oversampling technique) for handling imbalanced datasets. YouTube. Retrieved December 13, 2022, from https://www.youtube.com/watch?v=U3X98xZ4$_{n}oab_{c}hannel = BhaveshBhatt$

YouTube. (2020, February 10). Handling imbalanced datasets smote technique. YouTube. Retrieved December 13, 2022, from https://www.youtube.com/watch?v=dkXB8HH$_{4}-k$

Tiwari, S. (2021, August 22). Cerebral stroke prediction-imbalanced dataset. Kaggle. Retrieved December 13, 2022, from https://www.kaggle.com/datasets/shashwatwork/cerebral-stroke-predictionimbalaced-dataset

U.S. Department of Health and Human Services. (n.d.). What is a stroke? National Heart Lung and Blood Institute. Retrieved December 13, 2022, from https://www.nhlbi.nih.gov/health/stroke