Project report

Topic: Stroke Issue Prediction

Group 8: Cấn Thị Mai Anh, Lê Thị Hương, Đỗ Minh Vương

1. Motivation

A stroke is a medical condition in which poor blood flow to the brain causes cell death. There are two main types of stroke: ischemic, due to lack of blood flow, and hemorrhagic, due to bleeding. Both cause parts of the brain to stop functioning properly (1).

In 2015, stroke was the second most frequent cause of death after coronary artery disease, accounting for 6.3 million deaths (11% of the total). About 3.0 million deaths resulted from an ischemic stroke while 3.3 million deaths resulted from hemorrhagic stroke. About half of people who have had a stroke live less than one year. Overall, two-thirds of strokes occurred in those over 65 years old (1).

So, by studying big data mining, we believe that it is possible to predict the likelihood that a person may have a stroke based on that person's medical record. We believe that this will give people some peace of mind in the face of this dangerous disease.

2. Progress

2.1. Loading and Cleaning data

The dataset file is ‘healthcare-dataset-stroke-data.csv’ located in Datasets directory. We need to load csv in pandas library by function: read_csv()

To clean data, we do:

Firstly, we drop some unnecessary data like: ‘id’, ‘gender’, ‘ever_married’, ‘work_type’, and ‘Residence_type’ because we found that these are information fields to be filled out in medical records. They have no impact on health problems in general and stroke in particular.

On the contrary, the remaining features all have a certain influence on stroke. They are all on the list of risk factors for stroke (2).

Secondly, we replace literal values that are not replaceable with more useful ones like: ‘never smoked’ to 0, ‘formerly smoked’ to 1, ‘smokes’ to 2, and ‘Unknows’ will be NaN.

Thirdly, we drop all rows containing the value NaN.

We also found that the results 1, 0 are appearing in order, that can interfere with the results so we swapped the rows in the dataset.

Fourthly, we split the data into X and y variables. Next, Normalize the dataset using the MinMaxScaler() function.

Finally, we create datasets for training and testing with a ratio of 7:3.

2.2. Visualizing and Analyzing data

In this dataset there are a total of 12 columns and 5110 rows. The columns include:

1) id: unique identifier
2) gender: "Male", "Female" or "Other"
3) age: age of the patient
4) hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
5) heart_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
6) ever_married: "No" or "Yes"
7) work_type: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
8) Residence_type: "Rural" or "Urban"
9) avg_glucose_level: average glucose level in blood
10) bmi: body mass index
11) smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"*
12) stroke: 1 if the patient had a stroke or 0 if not

*Note: "Unknown" in smoking_status means that the information is unavailable for this patient

After many changes, the analytical data set includes: 7 columns and 3426 rows.

To find null values we used isnull().sum(). As a result there are 1745 null values that we need to remove.

We also look for the shape and data types of each column.

2.3. Analyzing result

Precision: 0 (0.94), 1 (1.00), macro avg (0.97), weighted avg (0.95)

Recall: 0 (1.00), 1 (0.00), macro avg (0.50), weighted avg (0.94)

F1-score: 0 (0.97), 1 (0.00), accuracy (0.94), macro avg (0.49), weighted avg (0.91)

Support: 0 (969), 1 (59), accuracy (1028), macro avg (1028), weighted avg (1028)

3. Techniques

3.1. Weighted averaging

We used it for the best results for three tests.

3.2. Bagging

We used it because it is a popular and highly accurate method.

4. Findings

We found this to be rather unusual, but we could not find any other data sources for further study and evidence. Overall, using the methods that we consider the best and most relevant, here are our final results. With this model, we hope it can be useful for future reference and improvement.

5. Reference

(1) Wikipedia

(2) Overview of Stroke, 2020, Ji Y. Chong , MD, Weill Cornell Medical College