Project:

Prediction of Stroke

By Samd Guizani (Cohort A23 – SPOC)

Contents

[Problem statement 3](#_Toc171636097)

[Exploratory data analysis 4](#_Toc171636098)

[Exploring univariate categorical variables 4](#_Toc171636099)

[Exploring univariate continuous variables 6](#_Toc171636100)

[Exploring correlations 7](#_Toc171636101)

[Continuous variables - Pairwise correlations 7](#_Toc171636102)

[Continuous variables – Correlation vs. target 7](#_Toc171636103)

[Categorical variables vs. target correlations 8](#_Toc171636104)

[bmi – Correlations vs. categorical variables 9](#_Toc171636105)

[Features engineering 10](#_Toc171636106)

[Encoding categorical variables 10](#_Toc171636107)

[Modelling 11](#_Toc171636108)

[Conclusion 12](#_Toc171636109)

# Problem statement

The objective of the study is to propose a model that predicts whether a patient has stroke or not, based on descriptors (such as age, gender, existing disease, etc.)

The dataset contains 5111 observations with the following 12 variables:

1. **id**: unique patient identifier
2. **gender**: “Male”, “Female” or “Other”
3. **age**: age of the patient
4. **hypertension**: 0 (if the patient doesn’t have hypertension) or 1 (if the patient has hypertension)
5. **heart\_disease**: 0 (if the patient doesn’t have a heart disease) or 1 (if the patient has a heart disease)
6. **ever\_married**: “No” or “Yes”
7. **work\_type**: “children”, “Govt\_job”, “Never\_worked”, “Private” or “Self-employed”
8. **Residence\_type**: “Rural” or “Urban”
9. **avg\_glucose\_level**: average glucose level in the blood
10. **bmi**: body mass index
11. **smoking\_status**: “formerly smoked”, “never smoked”, “smokes” or “Unknown” (in this case the information for the patient is not available)
12. **stroke**: 1 (if the patient had a stroke) or 0 (if the patient didn’t have a stroke)

**stroke** is the target variable to be predicted. It is a binary value (0 or 1), hence the model to develop is a ***classification model***.

The rest of the variables, except id (used as an index), are the features or predictors, among which 3 are continuous (age, bmi and avg\_glucose\_level) and 8 are categorical variables (binary or multilevel).

# Exploratory data analysis

It is to be noted that bmi variable has 201 missing values and will be dealt with in section Features engineering.

## Exploring univariate categorical variables

|  |  |
| --- | --- |
| stroke variable shows a very high unbalance. Patients with stroke represent only 249 among the 5000+ observations. Unbalancing of target variable may be a problem to achieve good classification performance. |  |
| hypertension and heart disease variables are unbalanced. Patients with hypertension or heart disease are underrepresented. |  |
| Over-representation of “Private” work\_type is observed. Also, Never\_worked category is highly under-represented.  Exploring further “Never\_worked” patient category shows the age of these 22 patients is ranging between 13 and 23 and none had a stroke. Therefore, decision was made to group “Never\_worked” under “children” |  |
| For gender, a reasonable balance between “Male” and “Female” categories is observed. However, “Other” category is very rare. Indeed only 1 observation falls in this category, and it is a patient with no stroke). Hence, it was excluded from the dataset. |  |
| smoking\_status counts show that the status of a large proportion of patients is unknown. It could be detrimental to the model performance, since smoking is a risk factor for stroke. |  |
| Most of the patients have been married. |  |
| residence\_type is equally distributed between “Urban” and “Rural” categories. |  |

## Exploring univariate continuous variables

|  |  |
| --- | --- |
| age distribution is rather symmetric and close to uniform (although a higher representation of age range between 40 and 60 years can be seen on the histogram).  No obvious outlying values are observed. | A graph of a graph  Description automatically generated with medium confidence  A blue and white background  Description automatically generated with medium confidence |
| Avg\_glucose\_level shows a bimodal distribution.  No obvious outlying values are observed. |  |
| bmi has an approximate symmetric distribution.  However, 5 outliers are observed (bmi > 65, which is very uncommon). Further exploration of these 5 observations revealed the patients did not have stroke. So, decision was made to exclude them. |  |

## Exploring correlations

### Continuous variables - Pairwise correlations

Pairwise plots of the 3 continuous variables reveal a positive correlation between bmi, age and avg\_glucose\_level. These correlations can help with estimating the missing bmi values.

A collage of blue and white graphs

Description automatically generated with medium confidence

### Continuous variables – Correlation vs. target

Using box plots by stroke, it can be noticed that patients with stroke are older and have higher and (more dispersed) avg\_glucose\_level. These 2 features are likely to be interesting predictors in the classification model.

On the other hand, bmi correlation with stroke seems weaker. It is seen that higher bmi relates to stroke, but the difference in median or distribution between the 2 groups is small.

A diagram of a graph

Description automatically generatedA graph of a graph showing a stroke

Description automatically generated with medium confidenceA graph of a graph showing a diagram

Description automatically generated with medium confidence

### Categorical variables vs. target correlations

Based on contingency tables and bar charts, it can be confirmed a strong link of stroke vs.:

* hypertension and heart disease: both conditions yield a much higher risk of developing stroke.
* ever\_married: being or having been married is correlated with stroke, but the link could be coincidentally related to age range (young people, such as children, are not married and as presented previously, age is a factor increasing stroke risk)
* work\_type: self-employment increases risk of stroke
* smoking\_status: former smokers and smokers are at high risk of stroke.

These 5 variables are therefore considered interesting predictors to include in the model.

A graph with blue squares

Description automatically generatedA graph with blue squares

Description automatically generatedA graph with blue squares

Description automatically generated

A graph with blue rectangles and black text

Description automatically generatedA graph with blue rectangles and black text

Description automatically generated

Gender and residence\_type are not well correlated to developing stroke.

A graph with blue squares

Description automatically generatedA graph of a graph with a white line

Description automatically generated with medium confidence

### bmi – Correlations vs. categorical variables

A specific focus was put to find correlation between bmi and other variables. The purpose is to confirm relationships that could help estimating the missing bmi values. It was found slight dependency of bmi with hypertension, heart disease, ever\_married, work\_type and smoking\_status.

A diagram of a graph

Description automatically generated with medium confidenceA diagram of a heart disease

Description automatically generated A diagram of a couple of blue rectangular objects

Description automatically generated A diagram of a number of people

Description automatically generated with medium confidence A diagram of smoking status

Description automatically generated

# Features engineering

## Encoding categorical variables

gender (after discarding “Other” category), ever\_married and residence\_type are binary variables for which the levels are expressed in text values. Therefore, they were converted to Boolean variables.

work\_type and smoking\_status are multilevel categorical variables. Therefore, they have been converted into binary variables using one-hot encoding.

Imputing missing bmi values

# Modelling

# Conclusion