Predicting stroke

Aleksandra Leszczyk

Abstract
A stroke is a severe medical condition that occurs when the part of your brain does not have access to the blood supply. Here, we build classification models that can diagnose a stroke.

Keywords
stroke – logistic classification — support vector machines — decision tree

Contents

1. Main objective

Stroke is one of the uppermost causes of death and disabilities in the world. Fast diagnosis and treatment are crucial for preventing its fatal consequences. In this analysis, we build different classification models that can detect a stroke. We aim at obtaining high sensitivity as the consequences of undiagnosed stroke have worse after-effects than false-positive classification.

Figure 1. Stroke symptoms usually start over seconds to minutes. Symptoms are not easy to recognize as they depend on the affected region of a brain.

2. Brief description of the data set and a summary of its attributes

The data set contains the information on patients that can be used for predicting a stroke, The data comes from Kaggle web service and it contains 5110 observations for 12 variables. The data set includes the identifications numbers of cases, genders, and age of patients, information if a person had hypertension or heart disease, marriage status, type of job, residence type, average glucose level, body-mass index, and smoking status.

The first rows of Table are presented in Table 1 while column names and types of variables are listed in Table 2. The first column contains the unique identification numbers, so it is not a valid variable. The column gender informs if a patient is a male, female, or another gender. The columns about hypertension and heart disease have a value of zero if a patient did not have a stroke, and one otherwise. There are only two options for a question about marriage, so yes means that a person is or was married. We distinguish only two types of residence here: rural and urban. We have five types of work,

Table 1. Head of a table.

id	gender	age
9046	Male	67
51676	Female	61
31112	Male	80
60182	Female	49
1665	Female	79
hypertension	heart disease	ever married
0	1	Yes
0	0	Yes
0	1	Yes
0	0	Yes
1	0	Yes
residence type	work type	avg glucose level
Urban	Private	228.69
Rural	Self-employed	202.21
Rural	Private	105.92
Urban	Private	171.23
Rural	Self-employed	174.12
bmi	smoking status	
36.6	formerly smoked	
NaN	never smoked	
32.5	never smoked	
34.4	smokes	
24.0	never smoked	

which include working for private subjects, being self-employed, working for the government, being a child, or never experienced working. The average glucose level is a continuous numerical variable of an unknown unit. The BMI stands for body-mass index, which is a measure of body fat based on height and weight. We investigate four types of smoking status that include people who smoke, people who do not smoke, people of unknown smoking status, and people who smoked in the past, but not now.

3. Data exploration, cleaning, and feature engineering

3.1 Distribution of a target variable

Data is not balanced. A stroke occurs in about 5% of cases.

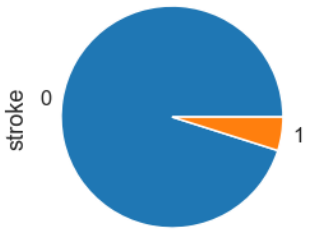
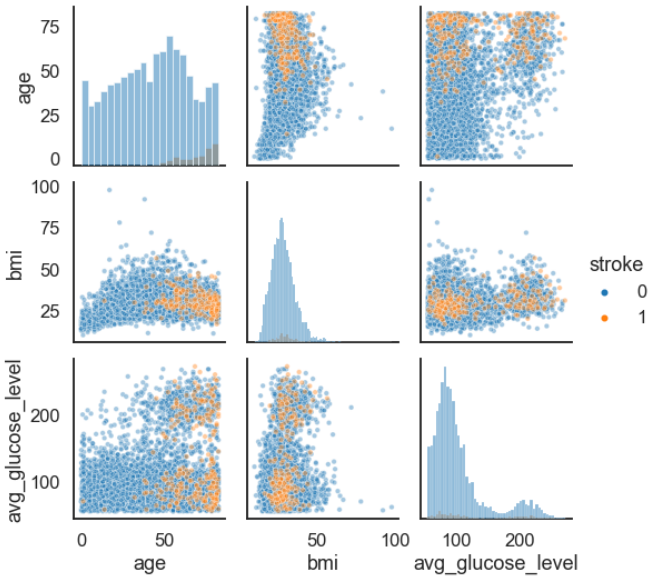


Table 2. Column names, number of non-null values, data types and variable types.

Column name	Non-null	Data type	Variable type
id	5110	int64	
gender	5110	object	nominal
age	5110	float64	discrete numerical
hypertension	5110	int64	binary
heart disease	5110	int64	binary
ever married	5110	object	binary
work type	5110	object	nominal
residence type	5110	object	nominal
avg glucose level	5110	float64	continuous numerical
bmi	4909	float64	continuous numerical
smoking status	5110	object	nominal
stroke	5110	int64	binary

3.2 Examining the distributions of numerical data

To elucidate the distributions of numerical data, we used tools offered by descriptive statistics. We computed mean, median, quantiles, and ranges of age, average glucose level, and body-mass index of patients. The results are presented in Table 3. Also, we used histograms and scatter plots to visualize these distributions for people that had a stroke and for those who did not have a stroke.



We observe that the distributions of body-mass index and average glucose level are similar for these two groups of people, but the stroke victims are older than the average patient. Also, none of the investigated variables is characterized by a normal distribution.

3.3 Exploring the non-numerical features

To get a better understanding of non-numerical features, we determined the number of the unique values for non-

Figure 2. Bar plots illustrating distributions of nominal and binary variables.

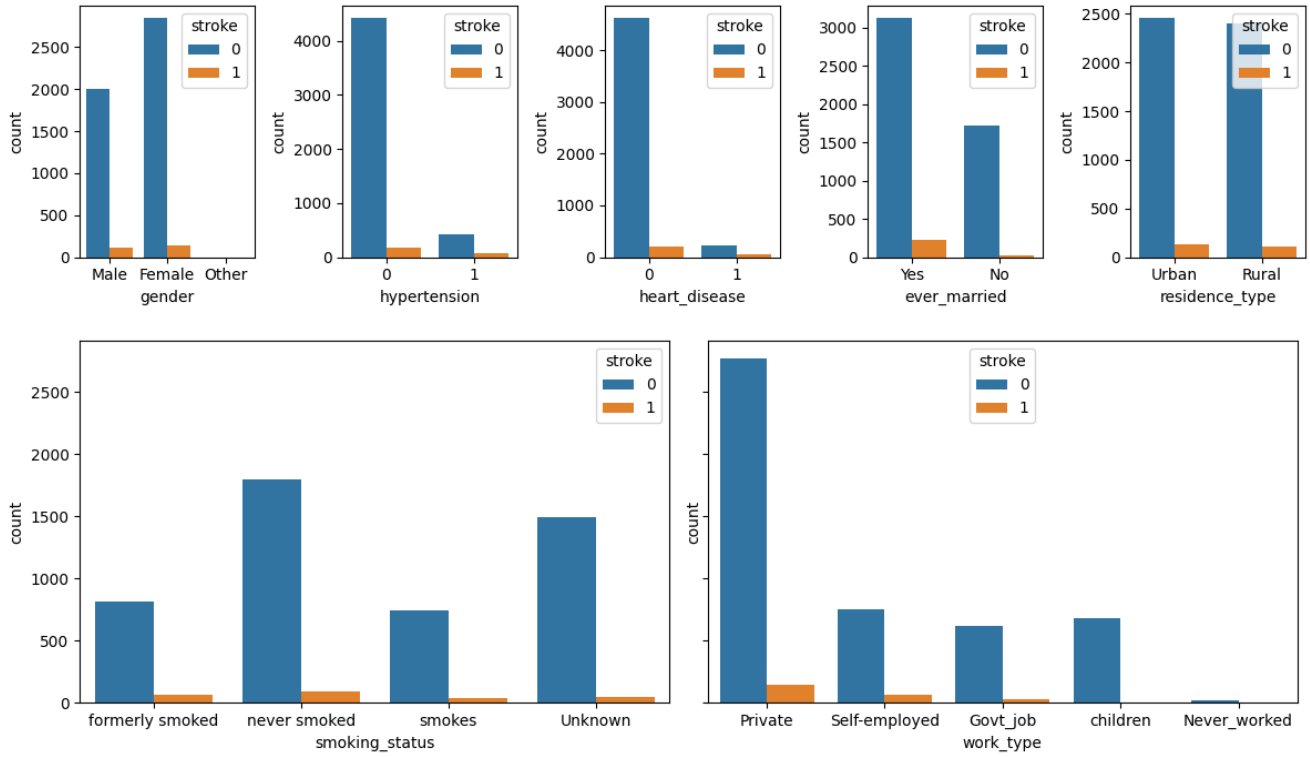


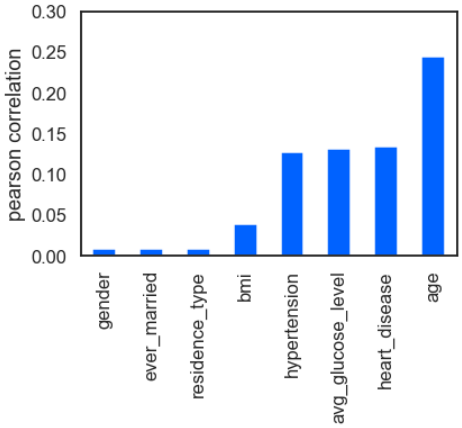
Table 3. Mean, median, quartiles, and skew of patient's age, average glucose level, and body-mass index.

	age	avg glucose level	bmi
mean	43.23	106.147	28.9
25%	25	77.245	23.5
median	45	91.885	28.1
75%	61	114.090	33.1
range	82	216.620	87.3
skew	-0.14	1.57	1.06

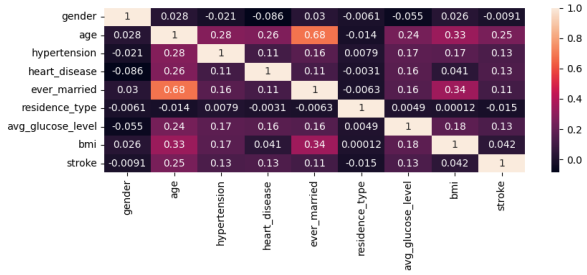
numerical features that include gender, hypertension, heart disease, marriage status, residence type, work type, and smoking status. The results are presented in Figure 2 in form of bar diagrams. We plotted a count of specific cases for patients that were diagnosed with stroke and for those who did not have a stroke. We observe that we have more women than men included in the data set, but the number of cases of stroke does not differ so much for those two groups. It can suggest that men have a higher probability of having a stroke. Also, the data suggest that stroke probability for people with hypertension is higher.

Heart disease also appears to increase a stroke risk as the ratio of stroke cases is 13% in a group of people affected with heart disease against 4% not affected by this condition. We see that stroke is more prevalent among people who are married or who were married. Only the residence type seems to not affect the stroke frequency. The smoking status gives us quite surprising results indicating that most stroke cases appear for former smokers. We have a large group of people whose smoking status is unknown. The data shows that self-employed people suffer more frequently from stroke than people that are hired by private subjects or by the government. Stroke almost does not happen among children. We have also a very small sample of people who never worked.

3.4 Investigating correlations between features



We observe the highest correlation between age and stroke, which confirms our conclusions from Subsection 3.2 that age is a significant risk factor. On the other hand, age is also positively correlated with hypertension, heart disease, average glucose level, body-mass index, and marriage. In particular, the correlation between age and marriage is high. Thus, we need to consider removing this variable as it can provide misleading results.



3.5 Data cleaning

We performed the following actions:

1. Filling the missing BMI values with a mean value;
2. Removing the outlier case with "other" gender as we do not have enough data and the conclusions based on a single case are prone to overfitting;
3. Removing the outlier cases with patients that never worked.
4. Removing column "ever married" as marriage is highly correlated with age but not with stroke. Correlated features introduce errors into the model. Also, marriage is not considered a medical condition although it seems to be a symptom of problems with clear thinking in some cases.

We have 5064 observation after cleaning.

3.6 Encoding

We used two types of encoding

1. One-hot encoding for the "work type", and "smoking status" columns;
2. Binary encoding for the "gender", "hypertension", "heart disease", and "residence type" columns.

3.7 Scaling

Since numerical variables – "age", "avg glucose level", and "bmi"– vary in magnitude, units, and range, we use min-max scaling to standardize data.

3.8 Feature engineering

We considered including second-order polynomial features, but we discovered that they introduce over-fitting and decrease the general performance of tested models. Thus, we do not add nor transform features.

4. Preparing train and test data

Preparing training and testing sets of observations requires us to take data imbalance into account. First, we used stratified sampling to obtain two sets of data sets with the same distribution of a target variable – 95% versus 5%. Second, we used SMOTE oversampling in training data to increase the number of stroke cases in the train set.

5. Logistic regression

First, we prepared three logistic regression models: without regularization, with an L1 penalty, and with L2 penalty.

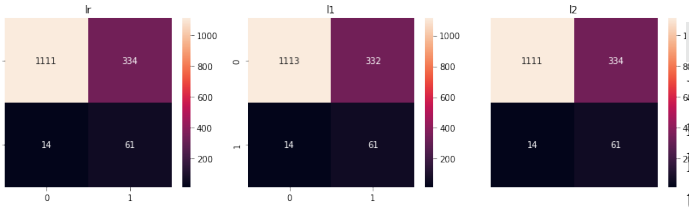
The prediction errors on the train set seem to balance different accuracy measures. Regularization does not change accuracy, which is expected in a data set with a low number of features.

train	no penalty	L1	L2
precision	0.818480	0.817461	0.817970
recall	0.814985	0.814095	0.814540
fscore	0.814476	0.813601	0.814039
accuracy	0.814985	0.814095	0.814540
auc	0.814985	0.814095	0.814540

We immediately see that we lose sensitivity while predicting the results in the test set.

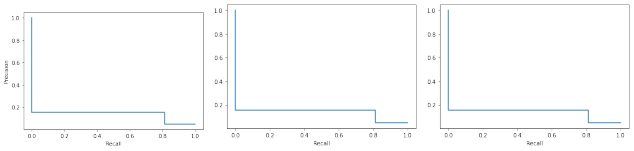
test	no penalty	L1	L2
precision	0.946447	0.946567	0.946537
recall	0.771053	0.773684	0.773026
fscore	0.834739	0.836526	0.836079
accuracy	0.771053	0.773684	0.773026
auc	0.791096	0.792480	0.792134

From the following confusion matrices, we see that we have a lot of type I errors, and we diagnose a lot of patients with stroke even if they do not have it. A culprit might be a SMOTE oversampling that increased the weights of cases with stroke.



The precision-recall curves (see Figure 3) are not impacted by highly different numbers of patients with and without the disease. Therefore, they might be the best estimate of the quality of the prediction model. Although precision, recall, f1 score, accuracy, and AUC are higher than 0.77, we see that the models are terrible at predicting stroke as distributions of people with and without stroke overlap.

Figure 3. Precision–recall curves for logistic regression models with no penalty, L1 penalty, and L2 penalty, respectively.

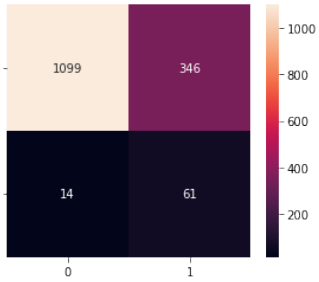


6. Support vector machines

We trained a linear support vector machine model. We have a similar situation as for the logistic regression model, where we lose sensitivity while predicting the results in the test set. The errors on the train and test set are presented as followed:

	train	test
precision	0.811411	0.946515
recall	0.806528	0.751316
f1	0.805767	0.821259
accuracy	0.806528	0.751316
auc	0.806528	0.787036

Again, we scan conclude from a confusion matrix that we have a lot of type I errors, and we diagnose a lot of patients with stroke even if they do not have it.



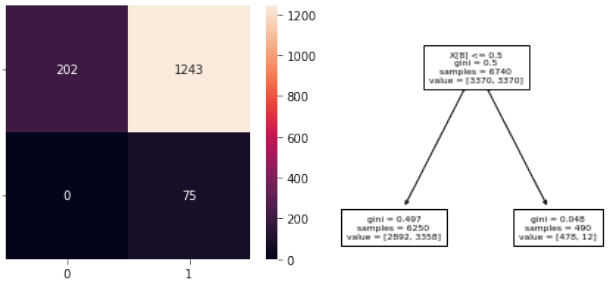
7. Decision tree

First, we trained a decision tree without tuning the parameters, but it resulted in strong over-fitting. The tree has a depth of 25 and 1351 nodes. Predictions for the train set are perfect, but the results for the test set are very poor.

no tuning	train	test
accuracy	1.0	0.855263
precision	1.0	0.162791
recall	1.0	0.466667
f1	1.0	0.241379

Thus, we need to look for better parametrization of the decision tree model. We found a tree that maximized a recall score that reached a value of 1. The tree has only 3 nodes and a depth of 1. However, the precision, accuracy, and f1 score of predictions are very low in a test set.

max recall	train	test
accuracy	0.569139	0.182237
precision	0.537280	0.056904
recall	0.996439	1.000000
f1	0.698129	0.107681



We checked if we can improve the general performance by maximizing an f1 score. The tree we obtained has a depth of 19 and 1207 nodes. The results are as followed:

max f1 score	train	test
accuracy	0.990059	0.853947
precision	0.986451	0.164384
recall	0.993769	0.480000
f1	0.990096	0.244898

Although almost perfect performance on the train set, we obtained low precision and f1 score in the test set.

8. Model selection

We trained three types of models used for stroke detection: logistic regression model, linear support vector machines model, and decision tree. The logistic regression model with the L1 penalty has the best performance and lowest number of undetected cases, and thus, we recommend this one for prediction purposes.

9. Summary key findings and insights

For all models, we had a large number of false-positive cases, which is a result of the strategy aimed at maximizing a recall and SMOTE oversampling, that increased weights of stroke-positive samples. In the case of the decision tree model, we encountered severe over-fitting problems, thus we do not recommend this model for stroke prediction. Although the accuracy, recall, f1 score, precision, and AUC for logistic regression models were higher than 0.77, we found out serious problems with the model while inspecting precision–recall curves. Turns out that the model cannot be treated as a reliable diagnosis tool which is not unexpected, as the data set is missing the key symptoms of a stroke. The reason might be selected features that do not include key stroke symptoms. We checked if adding polynomial features can increase the accuracy and precision, but we obtained only over-fitting. Thus, we suspect that this data set is created for learning purposes as it contains features that lead us to spurious correlations.

10. Next steps

We would advise removing the features that are not medical conditions nor hardly correlated with health, as marriage status, work type, or residence. The data set contains many features, but the characteristic symptoms of a stroke that are easy to detect are missing:

- numbness in the face, arm, or leg, especially on one side of the body
- confusion
- trouble speaking or understanding speech
- difficulty with seeing
- trouble walking, dizziness, lack of balance or coordination
- headache