Stroke prediction

Alen Vlahovljak

December 21, 2021

Contents

Chapter 1

Introduction

1.1 Problem description

Stroke is a medical condition in which the blood arteries in the brain are ripped, causing damage to the brain. When the blood supply to the brain is interrupted, symptoms might develop. Stroke is the most common cause of death and disability (World Health Organization - WHO). Early recognition of the various warning signs of a stroke can help reduce the severity of the stroke.

As already said, a stroke occurs when the blood flow to various areas of the brain is reduced, resulting in the brain cells not receiving the nutrients and oxygen they require. A medical emergency is urgent when a stroke occurs. Early detection and appropriate is required to prevent further damage to the affected area of the brain.

With the growth in Med-Tech, it is achievable to predict a stroke by utilizing ML techniques. The algorithms included in Machine Learning are valuable as they let us perform accurate prediction and proper analysis.

The features are:

Feature	Explanation
id	Unique identifier
gender	"Male", "Female" or "Other"
age	age of the patient
hypertension	0 - the patient doesn't have hypertension, 1 - the patient has hypertension
heart_disease	0 - the patient doesn't have any heart diseases, 1 - the patient has a heart disease
ever_married	"No" or "Yes"
work_type	"children", "Got_job, "Never_worked", "Private" or "Self-employed"
Residence_type	"Rural" or "Urban"
avg_glucose_level	average glucose level in blood
bmi	body mass index
smoking_status	"formerly smoked", "never smoked", "smokes" or "Unknown" - info not available
stroke	1 - the patient had a stroke, 0 - not

1.2 Dataset analysis

The first step we have to complete before we can analyze data is to perform an operation called data cleansing. Data cleansing is the process of identifying and removing the errors in the data records. While collecting and combining data from various sources into a data warehouse, ensuring high data quality and consistency becomes a significant, often expensive and always challenging task [3].

The cleaning process in my example isn't performed in one go, but in iterations by the needs of Exploratory Data Analysis. To maintain the hornology, the process is described in one piece.

The first thing I always prefer to do when it comes to cleaning data is to handle missing values in the dataset. After examining all features, we have found out that the BMI feature has around 200 missing values. We have to take care of them either by dropping records or filling them with median values. We have 5110 records, so losing the data is not desirable, and that's the reason we'll take the second approach. We'll fill the null values with the median value.

After analyzing each feature separately on the chart, we noticed that the BMI values has outliers that have to be handled:

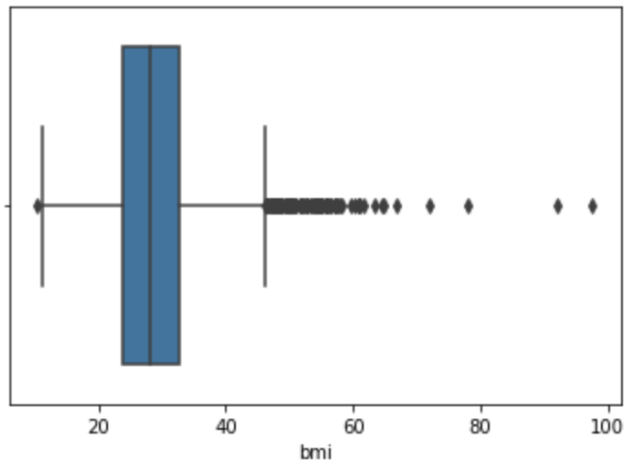

Figure 1.1: Distribution of the BMI within a population

In Figure 1.1, we can see that outliers occur for values greater than 45. After analyzing the BMI feature, we have found out that there are around 150 records that are greater than 45. To solve this issue, we'll replace outliers with the median value.

The next step is to prepare data type for a model because not all models can handle categorical data by applying integer encoding and one-hot encoding techniques to string-based records. The features which have the categorical type of data are *gender*, *ever_married*, *work_type*, *Residence_type* and *smoking_status*.

Encoding is as it follows:

- gender (Male - 1, Female - 0)
- ever_married (Yes - 1, No - 0)
- work_type (Govt_job - 0, Never_worked - 1, Private - 2, Self-employed - 3, children - 4)
- Residence_type (Urban - 1, Rural - 0)

The last step is to drop columns that are not relevant to the analysis. In our case, there is just one column - *id*.

Chapter 2

Related work

2.1 Introduction

According to a GBD study [15] from 1990, cerebrovascular diseases is ranked the second leading cause of death after heart disease. The research was in the conduct of the World Health Organization on the burden of over 130 diseases. The research also included countries in which data were not available. In this case, the only expected approach to estimate the disease pattern was to perform extrapolations from populations in neighbouring countries.

Over 795,000 individuals in the United States experience the ill effects of a stroke [10]. Two-thirds of stroke deaths occurred in people living in developing countries, and 40% of the subjects were aged less than 70 years. Many surviving stroke patients will usually depend on other people's care. We can put the disease in control by studying critical data that will provide an entry point for public health initiatives to reduce the burden of stroke within a population.

2.2 Previous research

One of the biggest problems in data mining in medicine is that medical data is voluminous, heterogeneous and complex. The algorithm we utilize to extract meaningful data must be high precision as the diagnosis is an important task.

The most typical techniques we use to generate predictive models are the Naive Bayesian classifier and the decision tree. Kononenko [6] explains that the Bayesian classifier is one of the simplest yet fairly accurate predictive data mining methods. Unlike a heart attack, heart disease can be utilized with a decision support system [1]. This technique can extract hidden patterns and relationships among medical data using major risk factors.

Machine learning methods have become increasingly popular in the field of cerebrovascular diseases. One of

the most comprehensive explanations is in Eun-Jae Lee et al. [14]. study, where they describe that machine learning can be utilized as a next-gen decision support tool. They address several issues which the over-fitting problem is the biggest to be resolved before the deep learning techniques can be introduced as a standard clinical practice. Most of the logic is impossible to understand by humans, so that is the biggest problem hampering implementation in medicine.

Two methods worth mentioning is Robben's [12] and Pinto's [11]. Their method is developed for learning the tissue outcome from weakly supervised data for the ischaemic stroke. They use clinical meta-data as predictors to a convolutional neural network. Regardless, the performance of the deep learning-based is still too low for clinical usage.

Sung et al. [14] studied and establish a stroke severity index. They used different data mining methods (including linear regression) to create their predictive models. The most accurate result was by the k-nearest neighbour method (95% confidence interval).

2.3 Performance comparison

In this paper, two models to implement, one with high and one with lower precision. We'll go through previous studies and see the accuracy values for the most common models.

The first article by Nwosu et al. [9] introduce an example of utilizing the models on the Electronic Health record dataset. The methods they implemented were decision tree, random forest and multi-later perception, with a calculated accuracy: decision tree (74.31%), random forest (74.53%), and multi-layer perception (75.02%). The article provides a systematic analysis of risk factors for stroke prediction. The analysis shows that when the patient attributes are not highly correlated, the feature space for predictive modelling cannot be reduced significantly without a significant loss of information.

Another research [5] that compares three models was conducted by Kansadub et al. implemented three models utilizing Decision tree, Naive Bayes and Support Vector Machine. The result obtained maximum accuracy for all three models around 60%, where the decision tree is the most accurate and Naive Bayes is the best in AUC (ROC curves beyond binary classification). The dataset they have been used is the dataset from the Office of the National Economic and Social Development Board (NESDB) between 1994 and 2013.

Another engaging research, which ultimately has a practical implementation in the form of an application, is the research by Monirul Islam et al. [7]. They apply four models Logic Regression, Decision tree classifier and K-NN and Random forest, which has outstanding results with the accuracy of precision 96%, recall 96%, and F1-score 96%.

D.Shanthi et al. [13] tried to predict thromboembolic stroke disease using artificial neural networks. The method used for prediction was the backpropagation algorithm. This model was able to get an accuracy

of around 89%. The downside of Neural Networks, in general, is that they need more time to be trained. Generated results have been verified with the doctors and are found correct. This ANN model helps doctors to plan for better care and provide the patient with an early diagnosis.

Chapter 3

Method(s)

Before model implementation, we have to do an Exploratory data analysis - a crucial step in any research analysis. The purpose of the exploratory analysis is to examine the data for distribution, outliers and anomalies to direct specific testing of your hypothesis. It also provides tools for hypothesis generation by visualizing and understanding the data, usually through graphical representation [4].

Firstly, we'll examine mostly important features via plot chart, and the features are *stroke*, *heart_disease*, *hypertension* and *avg_glucose_level*.

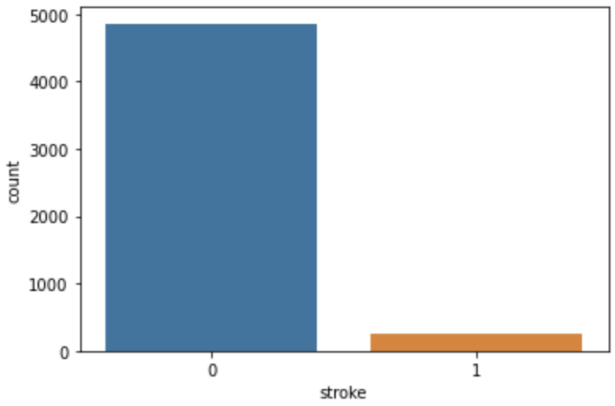

Figure 3.1: The plot chart of people who had a stroke in population

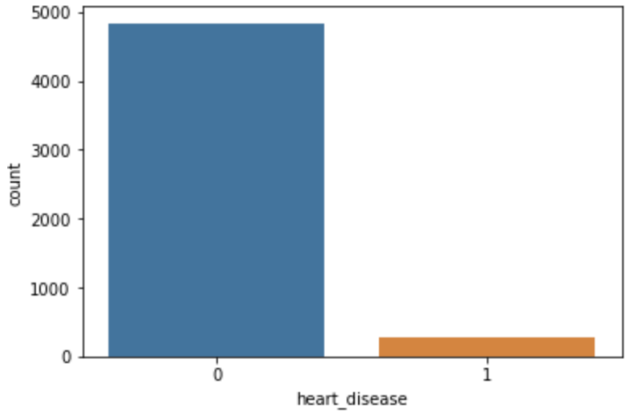

Figure 3.2: The plot chart of people who had a heart disease in population

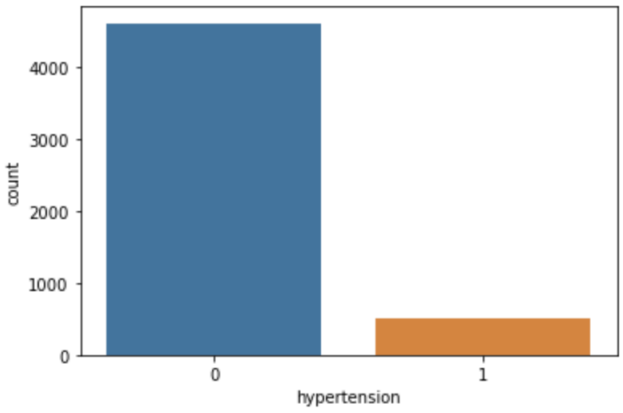

Figure 3.3: The plot chart of people who had a hypertension in population

We can see in the Figure 3.1 that data is highly imbalanced. Later, we are going to deal with this anomaly.

After some initial analysis, we'll investigate relations between independent variables and stroke (which is dependent in this case).

1. Age and stroke (the calculated correlation is 0.2453, which is a slightly positive correlation)

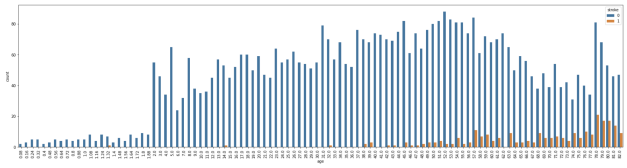

Figure 3.4: Age and the frequency of stroke

2. Heart disease and stroke (the calculated correlation is 0.1349, which is a slightly positive correlation)

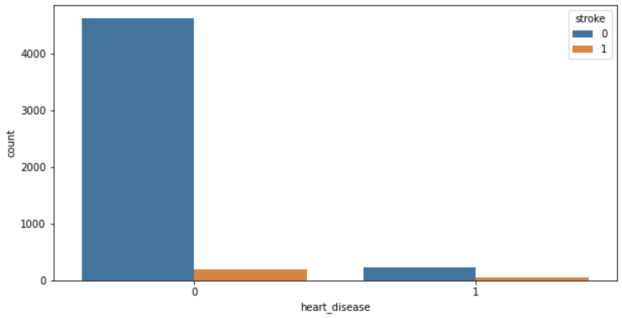

Figure 3.5: Heart disease and the frequency of stroke

3. Average glucose level in blood and stroke (the calculated correlation is 0.1319, which is a slightly positive correlation)

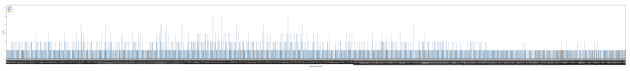

Figure 3.6: Average glucose level and the frequency of stroke

4. Hypertension and stroke (the calculated correlation is 0.1279, which is a slightly positive correlation)

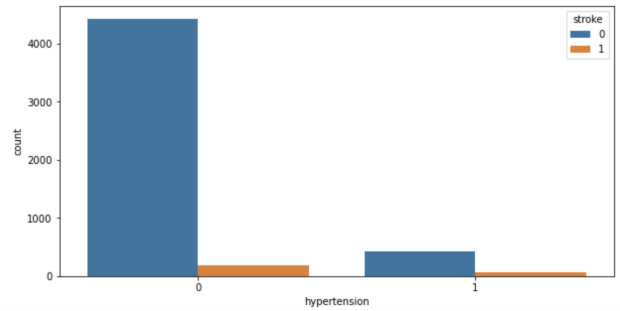

Figure 3.7: Hypertension and the frequency of stroke

5. Marital status and stroke (the calculated correlation is 0.1083, which is a slightly positive correlation)

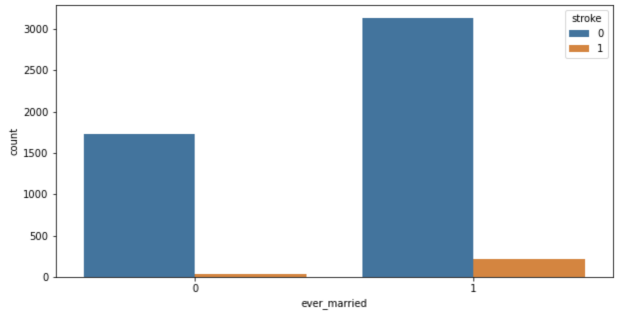

Figure 3.8: Marital status and the frequency of stroke

6. BMI and stroke (the calculated correlation is 0.0434, which is a very weak correlation)

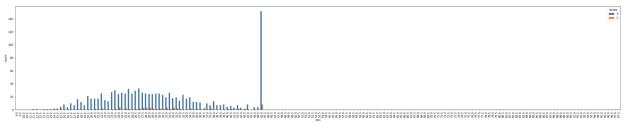

Figure 3.9: BMI and the frequency of stroke

7. Residence type and stroke (the calculated correlation is 0.0155, which is a very weak correlation)

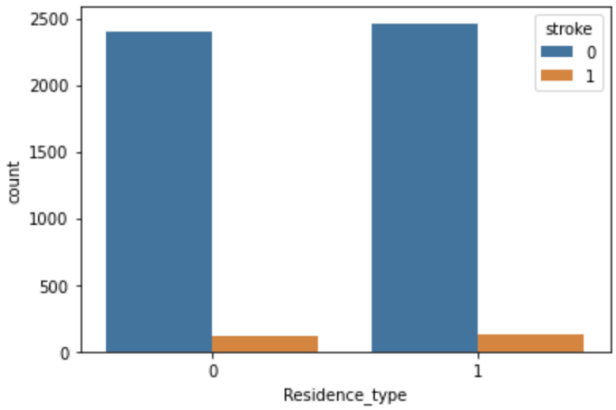

Figure 3.10: Residence type and the frequency of stroke

The last relationship we'll examine is between work type and stroke.

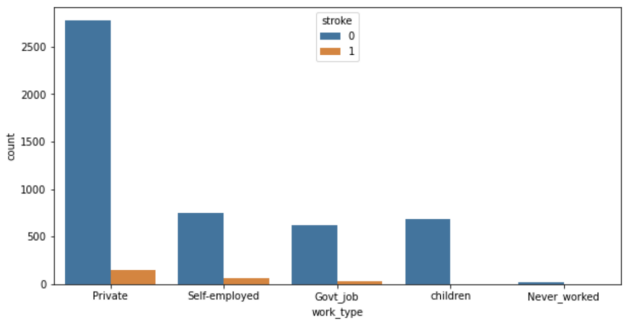

Figure 3.11: Work type and the frequency of stroke

As we can see from the figure, people whose work type is private have the highest risk of stroke. A reasonable explanation would be that these are people who often have to do multiple jobs at once, leading to increased stress. Also, working overtime means reduced physical activity, especially for people whose work is related to a sitting position.

Next, we have self-employed people. They have a slightly lower risk of stroke since this type of work involves fewer business activities.

People who work for someone are in the best position by looking at three statuses in which the person is in some kind type of employment type. That can be explained in such a way that their business activities are limited and that there is room for error, which is unacceptable for business in the case of self-employed or private individuals. Also, employees have less responsibility on their backs, and they work eight-hour shifts.

People who are not employed do not have a problem with work stress, so the risk of stroke is lower.

Finally, we have a group that includes children whose risk of stroke is minimal. Thus there is still a risk of stroke, which can be caused by various diseases of the cerebrovascular system. However, we cannot say this with certainty because we do not have their medical record for further investigation.

Chapter 4

Experimental results

Before the model can consume the data, there is one step we have to perform. We have to use MinMaxScaler from the scikit-learn library. We're going to apply MinMaxScaler for data normalization. For the fitting, we use the fit() function. It means that we use the training data for estimating the minimum and maximum observable values. The function transform() apply the scale to training data. It means you can use the normalized data to train your model. The default scale for the MinMaxScaler is to rescale variables into the range [0,1]. We could also apply StandardScaler, but according to experimental results, that is not efficient as the MinMaxScaler.

The difference between the MinMaxScaler and StandardScaler is that they use different scaling techniques. MinMaxScaler uses the normalization technique in which the values are shifted and rescaled, ending n the range between 0 and 1. On the other hand, StandardScaler uses scaling where the values are centred around the mean with a unit standard deviation. The mean of the variable becomes zero, and the resultant distribution has a unit standard deviation [2].

We need to keep in mind that there is also a rule on which technique is better under some circumstances. Normalization is more suitable to use when the data doesn't follow a Gaussian distribution (KNN or NN), while standardization is applied where the data follows a Gaussian distribution. Thus, this does not have to be necessarily true [8].

4.1 Building model

Before we apply data to the model, we'll split the dataset into two sets. One set is for testing purposes (80% of the dataset), the second one is for testing (20% of the dataset). The features we're using as independent parameters are all except the stroke parameter, which is dependent in our case.

We're going to build two types of models. One model is based on Logical regression, while the other is a random forest classifier. After we instantiate LogisticRegression and RandomForestClassifier method, we'll

use built-in fit (model is learning the relationship between x - digits and y - labels) and predict (predict the labels of new data - new images).

The next step is to use a confusion matrix and output the result (accuracy, precision and recall/sensitivity). After the first iteration the results are:

	Login Regression
accuracy:	0.9471624266144814
precision:	0.0
age	0.0
	Random Forest
accuracy:	0.9481409001956947
precision:	1.0
age	0.018518518518518517

It seems data is displaced since the precision and recall is equal to zero, which means we have to balance the data. We won't do a data balancing on our own, but with SMOTE function from the imblearn library.

A result is a decreased accuracy, but the model isn't biased, which means it will perform better on unseen data.

Chapter 5

Conclusion

As we could see in the initial analysis, the parameters that affect the probability of stroke are the expected ones. However, I'm sure that we would not be able to rank them accurately from the one who influences the most to the one whose influence is negligible.

After building the model, we got maybe a little discouraging result:

	Logistic Regression	**Result(%)**
accuracy:	0.7701446280991735	77.01%
precision:	0.7671092951991828	76.71%
recall/sensitivity	0.7758264462809917	77.58%
	Random Forest	**Result(%)**
accuracy:	0.8047520661157025	80.48%
precision:	0.9238505747126436	92.39%
recall/sensitivity	0.6642561983471075	66.43%

Ultimately, the results are satisfactory, but I guess it would have been better if the data were balanced.

Bibliography

[1] Syed Amin, Kavita Agarwal, and Rizwan Beg. Genetic neural network based data mining in prediction of heart disease using risk factors. pages 1227–1231, 04 2013.

[2] David Cournapeau. scikit-learn library, June 2007.

[3] Ratnadeep Deshmukh and Vaishali Wangikar. Data cleaning: Current approaches and issues. 01 2011.

[4] N. Heckert, James Filliben, C Croarkin, B Hembree, William Guthrie, P Tobias, and J Prinz. Handbook 151: Nist/sematech e-handbook of statistical methods, 2002-11-01 00:11:00 2002.

[5] Teerapat Kansadub, Sotarat Thammaboosadee, Supaporn Kiattisin, and Chutima Jalayondeja. Stroke risk prediction model based on demographic data. pages 1–3, 11 2015.

[6] Igor Kononenko. Inductive and bayesian learning in medical diagnosis. *Applied Artificial Intelligence*, 7:317–337, 1993.

[7] Md. Rokunojjaman Jahid Hasan Rony Al Amin Susmita Kar Md. Monirul Islam, Sharmin Akter. Stroke prediction analysis using machine learning classifiers and feature technique. 2021.

[8] K. Muralidharan. A note on transformation, standardization and normalization. 02 2010.

[9] Chidozie Nwosu, Soumyabrata Dev, Peru Bhardwaj, Bharadwaj Veeravalli, and Deepu John. Predicting stroke from electronic health records. volume 2019, pages 5704–5707, 07 2019.

[10] U.S. Department of Health Human Services. Statistics of stroke.

[11] Adriano Pinto, Richard Mckinley, Victor Alves, Roland Wiest, Carlos A Silva, and Mauricio Reyes. Stroke lesion outcome prediction based on MRI imaging combined with clinical information. *Frontiers in neurology*, 9:1060–1060, December 2018.

[12] David Robben, Anna M.M. Boers, Henk A. Marquering, Lucianne L.C.M. Langezaal, Yvo B.W.E.M. Roos, Robert J. van Oostenbrugge, Wim H. van Zwam, Diederik W.J. Dippel, Charles B.L.M. Majoie, Aad van der Lugt, and et al. Prediction of final infarct volume from native ct perfusion and treatment parameters using deep learning. *Medical Image Analysis*, 59:101589, Jan 2020.

[13] Dumpala Shanthi, Gadadhar Sahoo, and Dr.N.Saravanan. Designing an artificial neural network model for the prediction of thrombo-embolic stroke. *International Journal of Biometric and Bioinformatics*, 3, 03 2009.

[14] Sheng-Feng Sung, Cheng-Yang Hsieh, Yea-Huei Kao Yang, Huey-Juan Lin, Chih-Hung Chen, and Yu-Wei Chen. Developing a stroke severity index based on administrative data was feasible using data mining techniques. *Journal of Clinical Epidemiology*, 68, 01 2015.

[15] Thomas Truelsen and Stephen Begg. The global burden of cerebrovascular disease. *World Health Organization*, 01 2006.