Cerebral Stroke Prediction

Team 11

2017170827 이병주

2018160339 차수지

2019240007 조여정

### Abstract

We used cerebral stroke data to predict the onset of stroke. One sample of this data was marked with a stroke of 0 or 1, along with 12 types of health-related data that may or may not be related to cerebral stroke. A total of 43,400 of these samples were used. During pretreatment, missing values ​​in bmi level and smoking status were treated and these data were removed because the stroke incidence rate was significantly lower in those under 35 years of age.

ROC auc score and F1 score were used for model evaluation. MLP, logistic regression, random forest classifier, XGB classifier, and Linear SVC were applied to the model. Among them, MLP had the highest model accuracy and ROC auc score, and the results of logistic regression, XGB classifier, and Linear SVC were not bad either. However, in the case of the random forest classifier, the accuracy and f1 score were high, but the ROC auc score was low, so it was judged not suitable.

Therefore, we considered it most appropriate for predicting the onset of MLP stroke, that is, for binary classification.

### 1. Introduction

A stroke, a cerebrovascular accident (CVA), happens when part of the brain loses its blood supply causing the part of the body that the blood-deprived brain cells control to stop working. There are 2 kinds of stroke: ischemic and hemorrhagic. Ischemic stroke happens when a major blood vessel in the brain is blocked, lacking blood flow into the brain. Hemorrhagic stroke occurs when a blood vessel in the brain bursts and spills bleeding into brain tissue. A stroke is a medical emergency because brain cells begin to die after a few minutes, and this leads to death or permanent disability. There are opportunities to treat ischemic strokes, but that treatment needs to be started in the first few hours after the signs of a stroke begin.

There are several risk factors for strokes. According to the Global Burden of Diseases, Injuries, and Risk Factors Study (GBD) 2019([1]), estimated 19 behavioral, environmental, and occupational, and metabolic stroke risk factors. In 2020, the World Health Organization (WHO) and the International Labor Organization (ILO) ([2]) systemically review and meta-analyze estimates of the effect of exposure to long working hours on stroke. Recent studies analyze the effects of risk factors on stroke (three outcomes: prevalence, incidence, and mortality). In this paper, we explore the effects of risk factors on stroke and build a classification model to classify whether the stroke occurs.

### 2. Background

#### 2.1 Binary Classification

Binary classification is the task of classifying the elements of a set into two groups on the basis of a classification rule. It is dichotomization applied to a practical situation. In many practical binary classification problems, the two groups are not symmetric, and rather than overall accuracy, the relative proportion of different types of errors is of interest. For example, in medical testing, detecting a disease when it is not present (a false positive) is considered differently from not detecting a disease when it is present (a false negative).

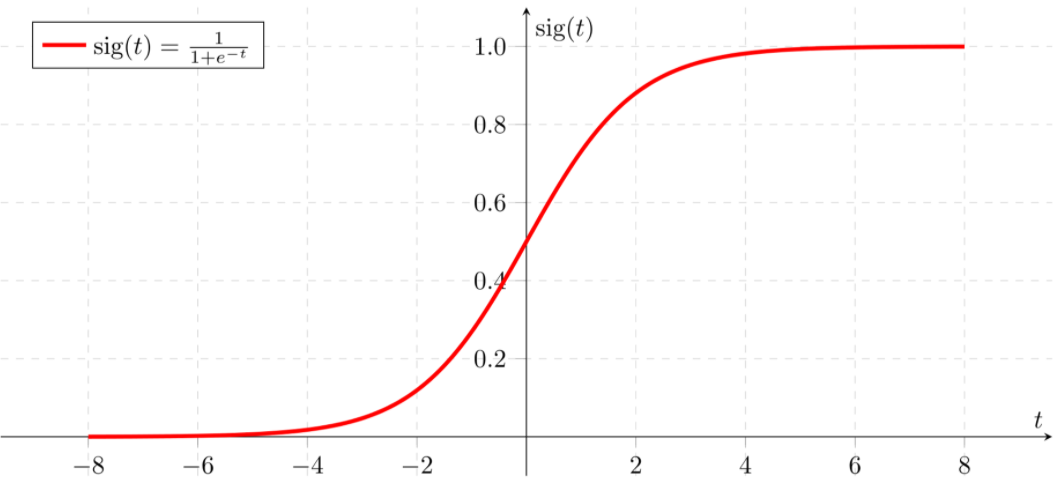
Statistical classification is a problem studied in machine learning. It is a type of supervised learning, a method of machine learning where the categories are predefined, and is used to categorize new probabilistic observations into said categories. When there are only two categories the problem is known as statistical binary classification. Some of the methods commonly used for binary classification are decision trees, random forests, bayesian networks, support vector machines, neural networks, logistic regression, probit model. Each classifier is best in only a select domain based upon the number of observations, the dimensionality of the feature vector, the noise in the data and many other factors.

#### 2.2 Logistic Regression

The logistic model (or logit model) is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick.

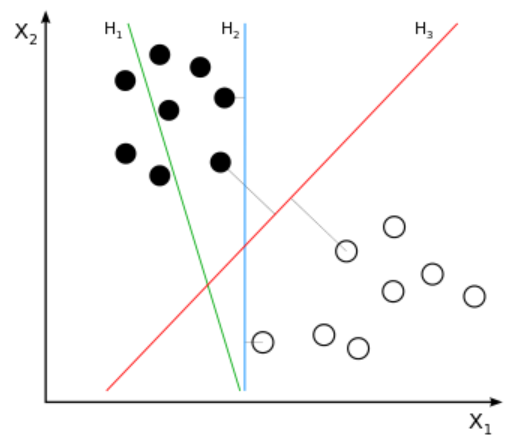
Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression is estimating the parameters of a logistic model (a form of binary regression). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labeled "0" and "1". In the logistic model, the log-odds (the logarithm of the odds) for the value labeled "1" is a linear combination of one or more independent variables ("predictors"); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling; the function that converts log-odds to probability is the logistic function, hence the name.

An explanation of logistic regression can begin with an explanation of the standard logistic function. The logistic function is a sigmoid function, which takes any real input, and outputs a value between zero and one. For the logit, this is interpreted as taking input log-odds and having output probability.



#### 2.3 SVM

Classifying data is a common task in machine learning. Suppose some given data points each belong to one of two classes, and the goal is to decide which class a new data point will be in. In the case of support-vector machines, a data point is viewed as a p dimensional vector (a list of p numbers), and we want to know whether we can separate such points with a (p-1)dimensional hyperplane. This is called a linear classifier. There are many hyperplanes that might classify the data. One reasonable choice as the best hyperplane is the one that represents the largest separation, or margin, between the two classes. So we choose the hyperplane so that the distance from it to the nearest data point on each side is maximized. If such a hyperplane exists, it is known as the maximum-margin hyperplane and the linear classifier it defines is known as a maximum-margin classifier; or equivalently, the perceptron of optimal stability.

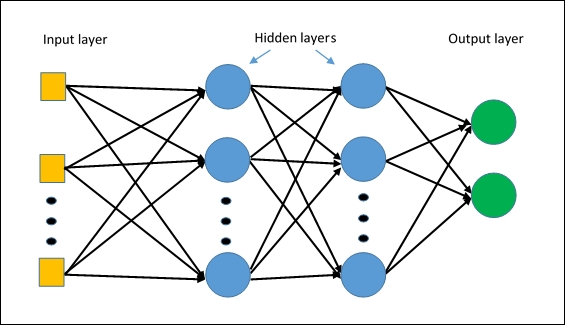


H1 does not separate the classes. H2 does, but only with a small margin. H3 separates them with the maximal margin.

#### 2.4 MLP

A multilayer perceptron (MLP) is a class of feedforward artificial neural network (ANN). The term MLP is used ambiguously, sometimes loosely to mean any feedforward ANN, sometimes strictly to refer to networks composed of multiple layers of perceptrons. Multilayer perceptrons are sometimes colloquially referred to as "vanilla" neural networks, especially when they have a single hidden layer.

An MLP consists of at least three layers of nodes: an input layer, a hidden layer and an output layer. Except for the input nodes, each node is a neuron that uses a nonlinear activation function. MLP utilizes a supervised learning technique called backpropagation for training. Its multiple layers and non-linear activation distinguish MLP from a linear perceptron.



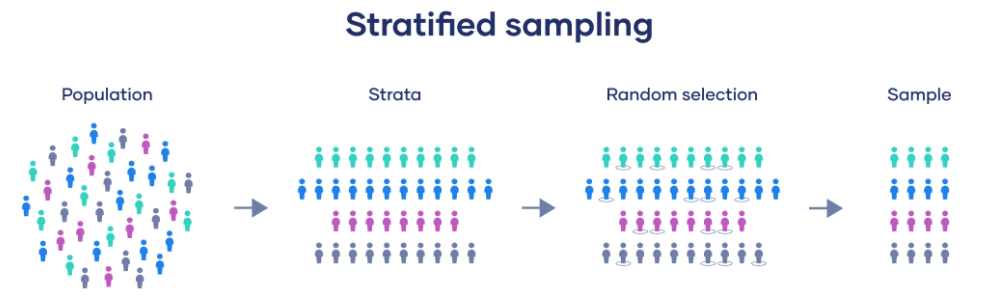
#### 2.5 XGB classification

XGBoost stands for eXtreme Gradient Boosting. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance that is dominative competitive machine learning. Although it is based on GBM, it is an algorithm that solves GBM's disadvantages such as slow execution time and overfitting regulations.

### 3. Methods

#### 3.1 Stratified Sampling

Stratified random sampling is a method of sampling that involves the division of a population into smaller sub-groups. In stratified random sampling, the strata are formed based on members' shared attributes or characteristics. The reason to use stratified sampling is, if measurements within strata have lower standard deviation, stratification gives smaller error in estimation. For example, if the values vary greatly, stratified sampling will ensure that estimates can be made with equal accuracy.



#### 3.2 One-Hot-Encoding

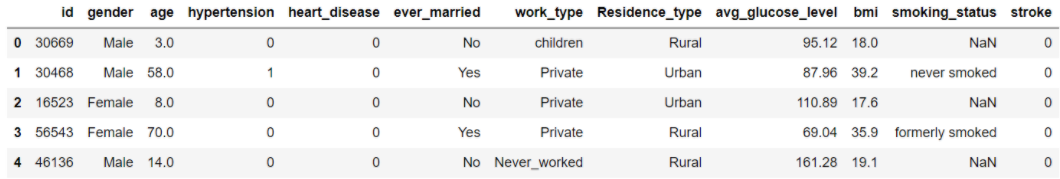
Machine learning algorithms cannot work directly with categorical data and they must be transformed into numeric values before training a model. Most common type of categorical encoding is One Hot Encoding (also known as dummy encoding) where each categorical level becomes a separate feature in the dataset containing binary values (1 or 0). This is ideal for features having categorical data. In other different scenarios other methods of encoding must be used.

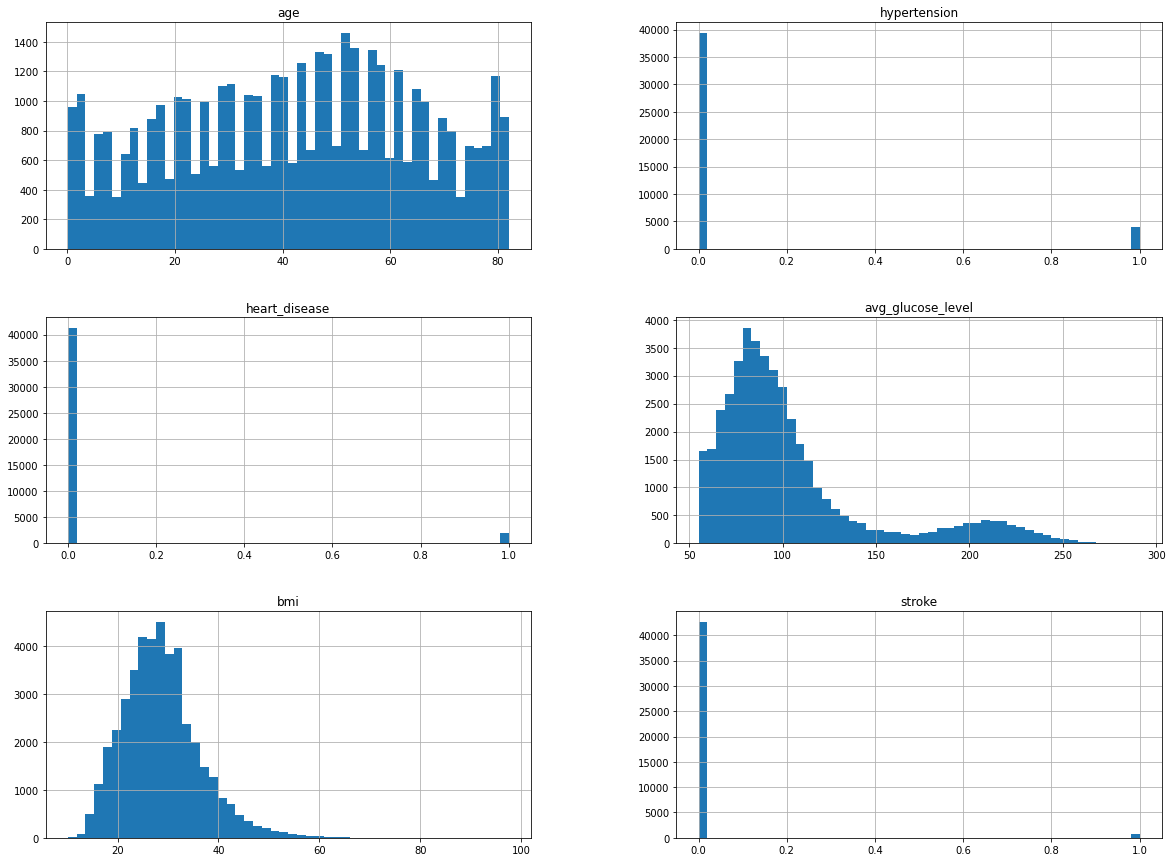
### 4. Experiments

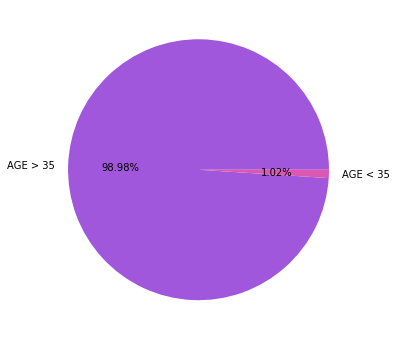
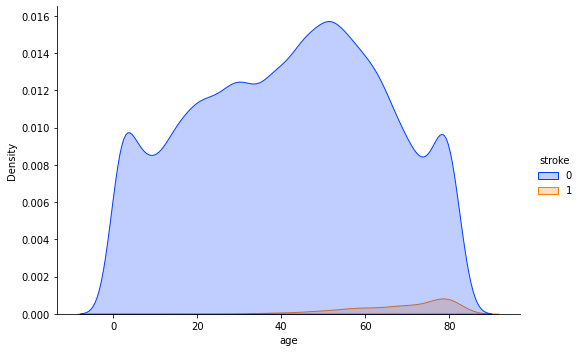
In our experiments, we endeavored to see which factor affects the occurrence of cerebral stroke and predict the probability to have cerebral stroke.

#### 4.1 Data Exploration

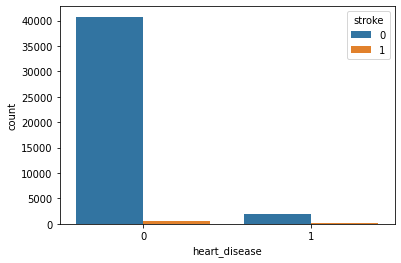
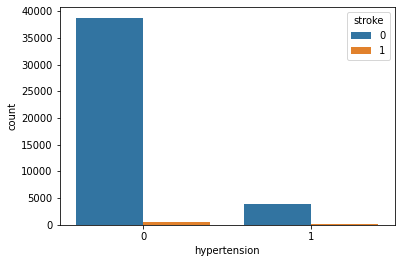
The cerebral stroke dataset consists of 43,400 people's information, which we used to predict cerebral stroke for our experiments. Each sample has 12 classes which are id, gender, age, hypertension, heart disease, ever married, work type, residence type, average glucose level, bmi, smoking status, and whether stroke or not.







The difference in stroke occurrence is large according to age. Looking at the two graphs above, most strokes occur in people over the age of 35. In addition, the incidence rate of stroke is significantly lower than that of non-stroke cases.



##### 4.1.1 Which factors will have the biggest impact?

In a sample that had a stroke, we calculated the proportion of patients with certain factors among all stroke patients. The probability that patients with hypertension had a stroke was only 10.70%, and patients with heart disease had only a 6.15% chance of having a stroke. Married people were roughly twice as likely to have a stroke than unmarried people, and those who were working were more than 90 percent higher than those who weren't working. In the case of residence, the ratio of urban and rural areas was the same, and the prevalence of stroke was twice as high for obese and smokers, respectively.

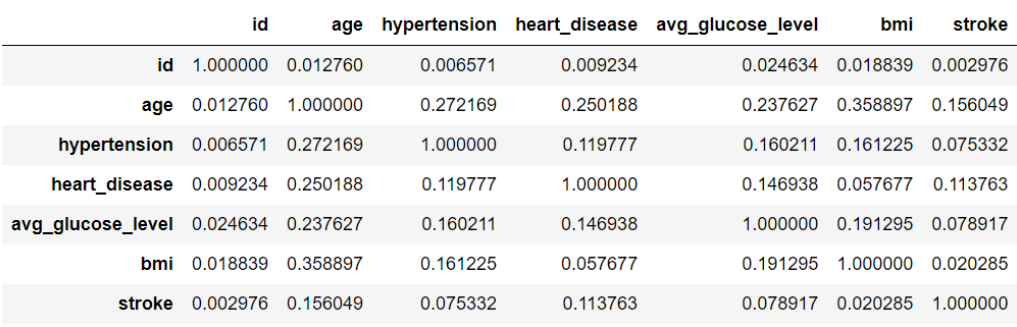
#### 4.2 Data Preprocessing

##### 4.2.1 Handling Missing Values

Most machine learning algorithms require numeric input values, and a value to be present for each row and column in a dataset. As such, missing values can cause problems for machine learning algorithms and it is common to identify missing values in a dataset and replace them with a numeric value. Therefore, missing values should be treated before training a machine learning model. When we checked missing values of our dataset, there were missing values in bmi and smoking status.

###### 4.2.1.1 Treating missing value of bmi

First, we checked bmi's correlation with other factors.

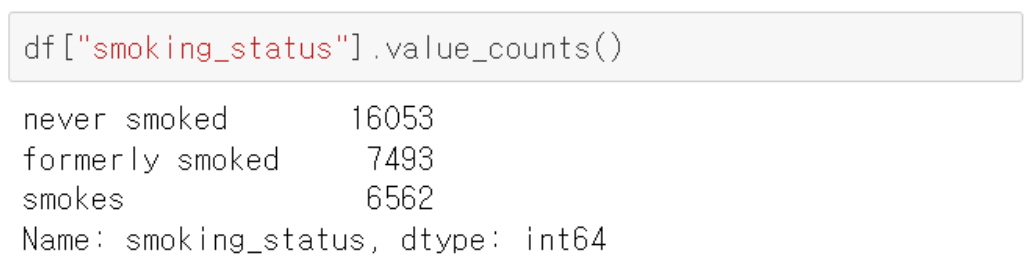


According to the result, we thought that bmi value has a high correlation with age. Therefore, from now on we are going to get the average value of bmi by age. Age values range from 0 to 82 years. So we split them into the 0's ~ 80's.

And then, we got average values of bmi by age. Finally, we applied these average values to the bmi missing values. After that, we could check null values of bmi are eliminated.

###### 4.2.1.2 Treating missing value of smoking status

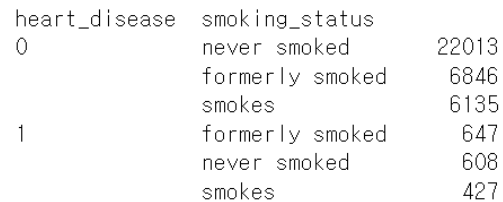
When we checked the number of people according to smoking status, those who have never smoked have the highest percentage.



Similar to the previous step which treated the missing values of mbi, we got the distribution of smoking by age. According to the results, we know that many children and teenagers are non-smokers. Therefore, we filled the missing value of smoking status with 'never smoked'.

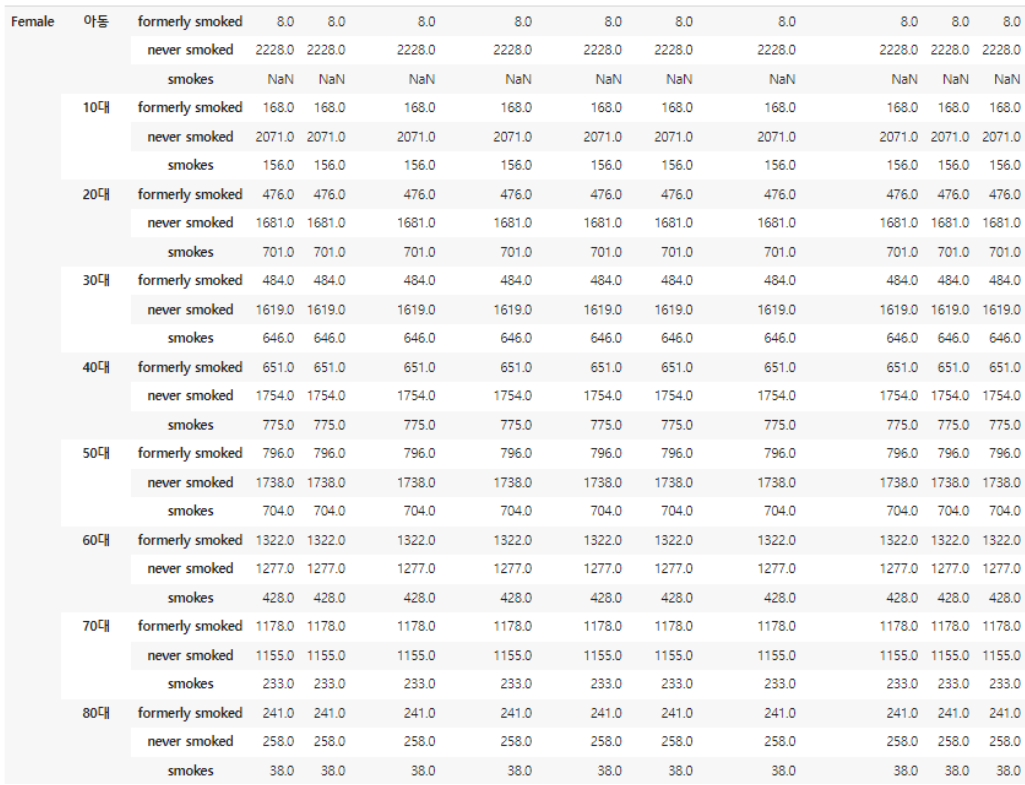
After that, the missing value of smoking status decreased from about 30% to about 15%.

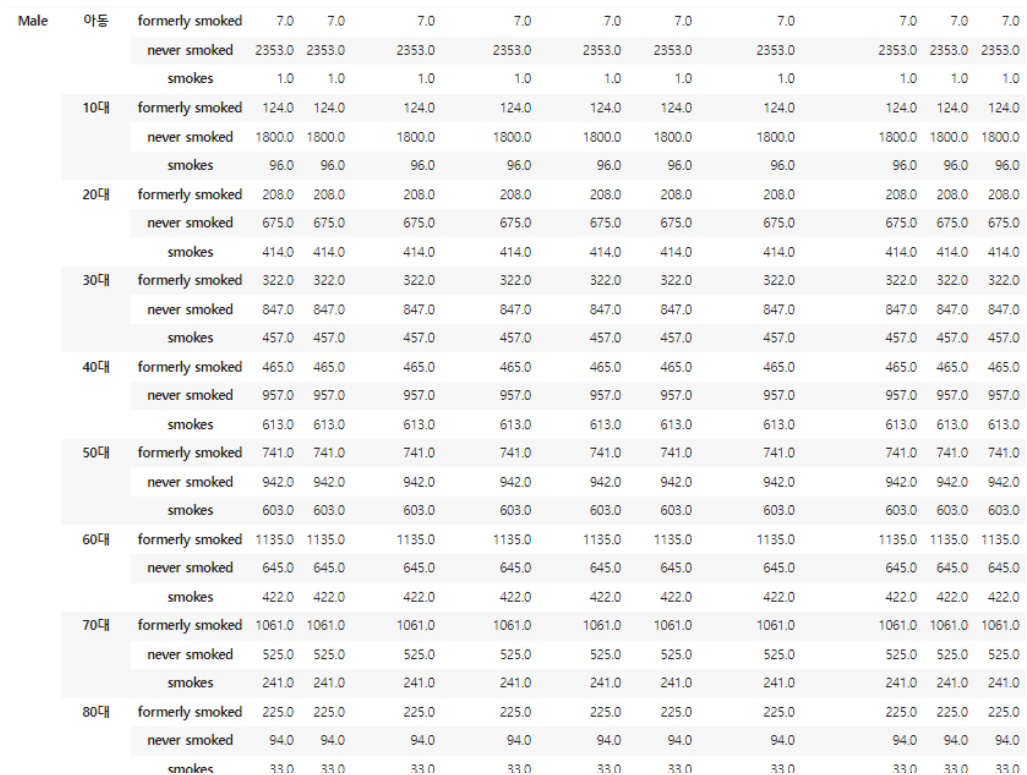
To decrease the missing value more, we applied a little information. First, many people with heart disease have smoked in the past. So we filled the smoking status's missing value of people who have heart disease into 'formerly smoked'.

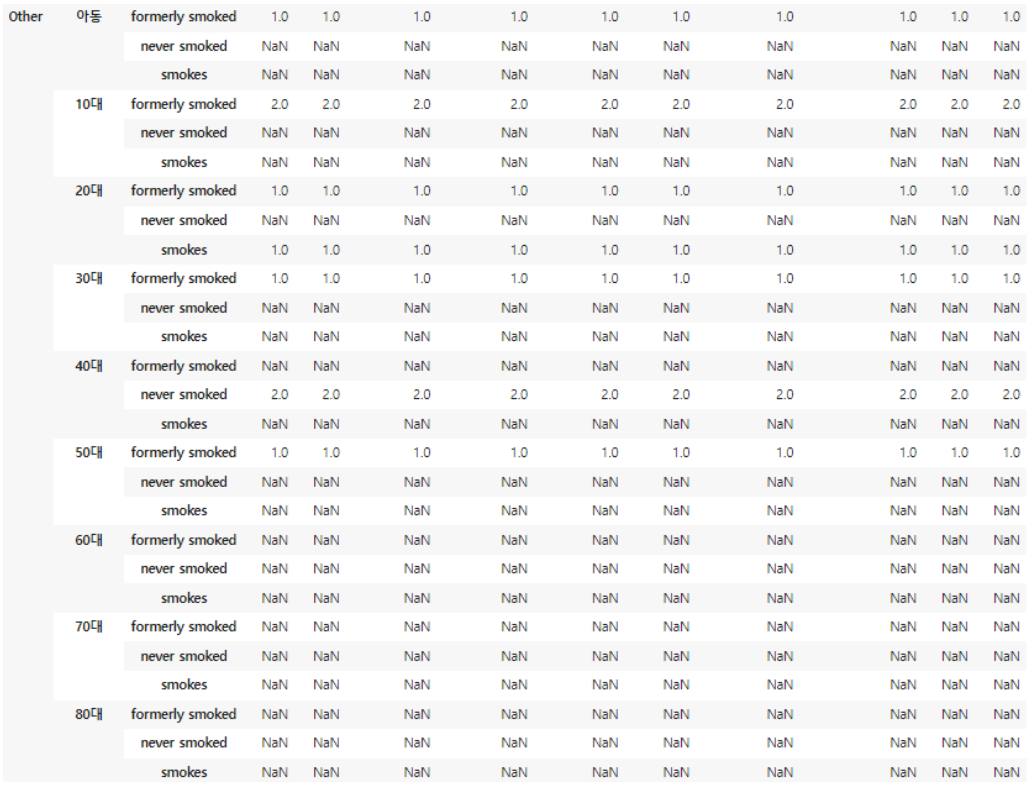


correlation between heart disease and smoking status

Second, we checked the smoking status by age and gender.







According to the results, among men, formerly smoked is the most in their 60s, 70s, and 80s. Therefore we filled the smoking status's missing value of people who are 60s, 70s, and 80s men into 'formerly smoked'.

After these two steps, 10% of missing values are still left. Since we do not have correlation in data anymore, we decided to fill the remaining missing value into 'never smoked'.

##### 4.2.2 Splitting

The difference in stroke occurrence according to age was large, so we tried to discard samples under 35 years of age. And since the ratio of stroke values ​​differs greatly, layered extraction is performed. The ratio of the test set was set to 0.3. Then to proceed with model evaluation for model selection, layered extraction is performed once more with validation and train set.

##### 4.2.3 Pipeline

First, we seperated the categorical data and the numeric data.

To assemble several steps that can be cross-validated together while setting different parameters, we used a pipeline in scikit-learn.

We put a standard scaler inside the pipeline, and we didn't give parameter values. And then, we converted the data previously divided into numeric and categorical types. We used a standard scaler to convert numeric data, and used a one-hot-encoder to convert categorical data.

The data we have is markedly different in the number of the two classes. Therefore, we tried to solve the problem of unbalanced data and we used an asymmetric data processing program called SMOTE (synthetic minority oversampling technique). The operation method of SMOTE is an oversampling method that takes a sample of a class with a small amount of data and adds a random value to create a new sample and add it to the data.

### 5. Model Selections

For evaluation of each model, ROC auc score and f1 score were used.

ROC is a probability curve and AUC presents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. Higher the AUC, the better the model is at predicting 0 classes as 0 classes as 0 and 1 classes as 1.

The F1 score combines precision and recall. Precision and recall are two building blocks of the F1 score. The goal of the F1 score is to combine the precision and recall metrics into a single metric. At the same time, the F1 score has been designed to work well on imbalanced data.

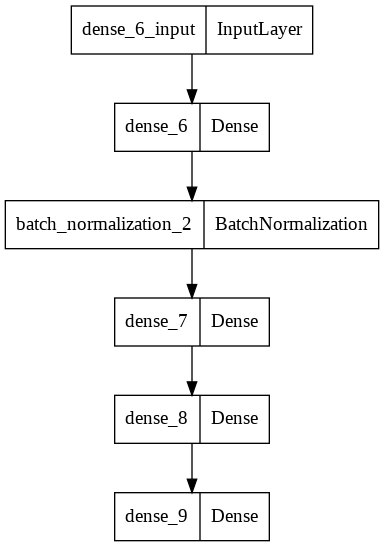
#### 5.1 MLP

##### 5.1.1 Training

We trained the ML on the 30,380 training data for 15 epochs and batch size for 300 using an Adam optimizer with a binary crossentropy loss function.

Initially, we created a model with 4 layers. The first layer using the relu function, the batch normalization layer, the second layer using the relu function, and the last layer using the sigmoid function. Then we checked accuracy and ROC auc score.

After that, we tried increasing the number of neurons used and adding one more layer using the relu function.



##### 5.1.2 Results

The final training accuracy was 95.58% and the final validation ROC auc score was 0.79. In general, if the AUC is 0.8 or higher, it is evaluated as a binary classifier with very good performance. Anything between 0.7 and 0.8 is a good binary classifier. If the AUC is 0.5-0.7, it is helpful to have a binary classifier. Since our MLP has an ROC auc score of 0.79, this is considered a pretty good result.

#### 5.2 Logistic Regression

##### 5.2.1 Training

We gave class weight 0.93 to training and 0.08 to testing.

##### 5.2.2 Results

The final training accuracy was 78.44%, final validation accuracy was 73.89%, f1 score was 0.80, and ROC auc score was 0.77. Although the ROC auc score is still high, the accuracy is considerably reduced compared to the previously applied MLP.

#### 5.3 Random Forest Classifier

Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.

##### 5.3.1 Training

We used previously preprocessed data (with samples under 35 years old) for training.

##### 5.3.2 Results

The final training accuracy was 100.00%, final validation accuracy was 92.58%, f1 score was 1.0, and ROC auc score was 0.57. The accuracy was very high in both the training data and the validation data, and the F1 score also reached its best value. However, the ROC auc score was 0.57, which was significantly lower than the previous two models. If the AUC is less than 0.5, then the binary classifier is useless. Therefore, we think that the random forest classifier is not appropriate.

#### 5.4 XGB Classifier

XGB is an algorithm that has recently been dominating applied machine learning and Kaggle competitions for structured or tabular data. XGB is an implementation of gradient boosted decision trees designed for speed and performance.

##### 5.4.1 Training

We used previously preprocessed data (with samples under 35 years old) for training.

##### 5.4.2 Results

The final training accuracy was 86.84%, final validation accuracy was 74.71%, f1 score was 0.86, and ROC auc score was 0.75.

#### 5.5 Linear SVC

The objective of a Linear SVC (Support Vector Classifier) is to fit to the data we provided, returning a “best fit” hyperplane that divides, or categorizes, our data. From there, after getting the hyperplane, we can then feed some features to our classifier to see what the predicted class is. This makes this specific algorithm rather suitable for our uses, though we can use this for many situations.

##### 5.5.1 Training

We used previously preprocessed data (with samples under 35 years old) for training.

##### 5.5.2 Results

The final training accuracy was 78.64%, final validation accuracy was 73.28%, f1 score was 0.80, and ROC auc score was 0.77.

### 6. Model Tuning and Testing

Since the data we have is unbalanced, we thought it would be appropriate to use the data oversampling with the SMOTE method. For this oversampling data, the best result was when XGBclassifier was applied. Therefore, in order to finally classify the oversampling data with the XGBclassifier, the validation set was also processed with the SMOTE technique.

As a result, the f1 score was 0.82, the accuracy was 79.48%, and the ROC auc score was 0.80. Although the accuracy was slightly lower than the results confirmed in the model selection earlier, the f1 score and the ROC auc score were able to obtain quite good values.

### 7. Discussion

We used cerebral stroke data to predict the onset of stroke. One sample of this data was marked with a stroke of 0 or 1, along with 12 types of health-related data that may or may not be related to cerebral stroke. A total of 43,400 of these samples were used.

Data exploration found that people who smoke, are obese, and who work are more likely to have a stroke. In the data preprocessing process, it was found that people under the age of 35 had a very low chance of developing a stroke, and this data was excluded in the future modeling process. A total of five models were used to predict the stroke onset probability for various factors (MLP, logistic regression, random forest classifier, XGBclassifier, and linear SVC). As a result, the rest of the models except for the random forest classifier were considered suitable for our binary classification model, and we considered it most appropriate for predicting the onset of MLP stroke.

We wanted to make a model that can calculate the probability of an outbreak by going further from predicting whether or not there is an outbreak with a more diverse and many datasets. In addition, we think that the current model can be improved by using Bayes networks, probit models, and other neural networks that can perform binary classification.

We first wanted to know which of the factors in each sample had the greatest influence on the occurrence of stroke, as well as predicting the occurrence of stroke. For now, working with data, we computed only one relationship for each factor. However, Previously, when calculating the probability of being a stroke patient for each factor, there were parts that did not match the above facts. This is thought to be due to the small amount of data. Therefore, it is thought that more data is needed. And since stroke is not caused by a single cause, but as a complication of several diseases, creating a model that can broadly understand the relationship between various factors will be more helpful in predicting stroke onset.

### 8. Conclusion

The brain is damaged when the blood vessels supplying blood to the brain are blocked or burst, and the resulting neurological abnormalities such as hemiparesis, speech disorders, and consciousness disorders are called strokes. Causes of the disease include high blood pressure, diabetes, smoking, heart disease, and obesity.

Various classification models were applied to predict the occurrence of cerebral hemorrhage. The dataset we used was a case of binary classification, and MLP, logistic regression, random forest classifier, XGBclassifier, and linear SVC were applied. As a result, pretty good results were obtained in most models except for the random forest classifier, and the accuracy was the highest when using MLP. Therefore, it was considered that MLP was suitable for binary classification. In addition, the XGBclassifier was evaluated with data oversampling by the SMOTE method. As a result, the accuracy was slightly lower than the results confirmed in the model selection earlier, but the f1 score and the ROC auc score were able to obtain quite good values.

In the case of the random forest classifier, the accuracy was high, but the ROC auc score was low, about 0.5. The reason for this seems to be that if a decision tree is used, the result or performance fluctuates widely. In order to improve this, it is thought that hyper parameter tuning to find an appropriate hyper parameter (ex. class\_weight) through trial and error is additionally needed.

If the probability of stroke can be predicted through the relationship between living characteristics or information on current diseases, it is expected that the probability of occurrence can be reduced by paying close attention to stroke.

### References

[1] Valery L Feigin, Benjamin A Stark, Catherine Owens Johnson at el. Global, regional, and national burden of stroke and its risk factors, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019, The Lancet Neurology, Volume 20, Issue 10, 2021, Pages 795-820

[2] Alexis Descatha, Grace Sembajwe, Frank Pega, Yuka Ujita, Michael Baer, Fabio Boccuni, Cristina Di Tecco, Clement Duret, Bradley A. Evanoff, Diana Gagliardi, Lode Godderis, Seong-Kyu Kang, Beon Joon Kim, Jian Li, Linda L. Magnusson Hanson, Alessandro Marinaccio, Anna Ozguler, Daniela Pachito, John Pell, Fernando Pico, Matteo Ronchetti, Yves Roquelaure, Reiner Rugulies, Martijn Schouteden, Johannes Siegrist, Akizumi Tsutsumi, Sergio Iavicoli, The effect of exposure to long working hours on stroke: A systematic review and meta-analysis from the WHO/ILO Joint Estimates of the Work-related Burden of Disease and Injury, Environment International, Volume 142, 2020, 105746