Data Science and Economics

Universit`a degli Studi di Milano

Supervised learning

(Decision trees)

Student: Vojimir Ranitovic 963780

Email: vojimir.ranitovic@studenti.unimi.it

Module: Statistical Learning, Deep Learning and Artificial Intelligence

Abstract

In this project decision trees are used to classify patients according to key health indicators if they have stroke or not. It was found that dataset is fairly unbalanced, and combination of undersampling and oversampling was used to overcome this problem. In this situation accuracy measure should be taken with caution. Also, it was found that decision trees can easily become very complex – to overfit.

Introduction and dataset

According to World Health Organization and World Stroke Organization, stroke has already reached epidemic proportions and globally 1 in 4 adults over age of 25 will have a stroke in their lifetime and 12 million people globally will have their first stroke this year and 6 million will die as a result. These devastating results are shifting goals of many scientists to prevent it, also the programmers and data scientists who can help it to predict it and react on time. The goal of this project is to predict if patient would have a stroke or not according to 10 predictors: gender, age, if he/she has hypertension or hearth disease, marital status, residence and work type, BMI, smoking status and glucose level. The dataset is downloaded from Kaggle and has 4981 different patients and previously listed variables. First noticeable thing is that this dataset is highly unbalanced. As it is shown on Figure 1., there are 248 patient that had stroke and 4733 that have not.

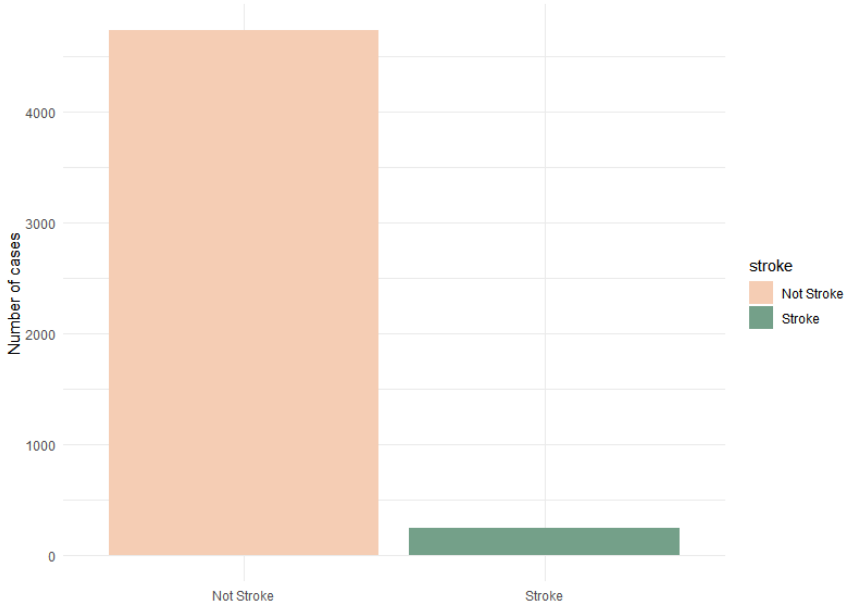

Figure 1: Number of cases for both classes – Unbalanced data

Figure 2: Raw counts of each feature and its categories

In Figure 2., we can see that there are more female patients, also there are lower numbers of hypertension and heart disease cases. Mostly, patients are/were married, and they mainly work at private firms. There is no big difference in numbers when they come from rural or urban places. Usually, patients have never smoked or its status is unknown. Figure 3., are set of plots displaying the within-feature category proportions of patients with stroke and without stroke. It can be seen that relatively there is slightly more man that has a stroke, and patients with hypertension and heart disease are proportionally more prone to stroke. If they have ever been married, strokes are more common, also if they are self-employed or former/current smokers. If we observe each plot individually, it is easy to be misled by the data. For example in the "ever married" plot it appears that strokes are far more common among those who are/were married than those who have not, but if we look at "work type" plot, there is a considerable amount of children, who presumably have never been married, and therefore cause the "ever married" plot be illusory.

Figure 3: Within feature category proportions of stroke and no stroke

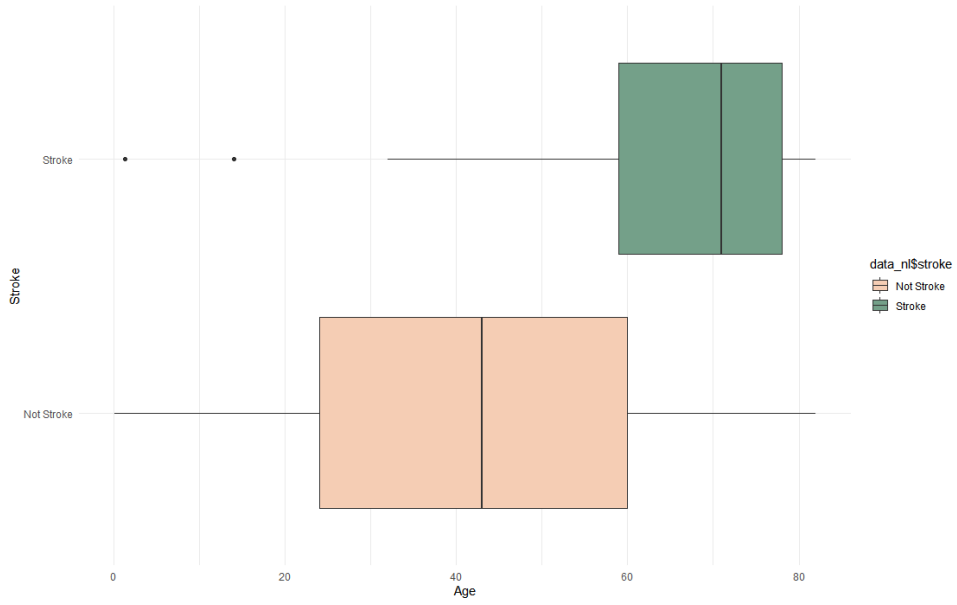

Figure 4: Age distributions

In Figure 4., we can see the age distribution of those who had a stroke is centralized at a much older age than the age distribution of those who did not have a stroke.

Model and Results

For predicting the classes "stroke" or "no stroke", this project will be used decision trees. Dataset is randomly divided into training and test sets in the proportion of 70% for train and 30% for test. As we saw at the beginning, this dataset is highly unbalanced, so it can mislead us that any model and algorithm that we use could have high accuracy. It is simple to predict one class all the time and have high accuracy like it could be in our case with predicting "no stroke" most of the time. There is some solution to this, like oversampling, undersampling, or a combination of these two. Oversampling methods duplicate or create new synthetic examples in the minority class, whereas undersampling methods delete or merge examples in the majority class. Both types of resampling can be effective when used in isolation, although can be more effective when both types are used together. For our model combination of methods was used. After splitting the dataset, there were 3313 "no stroke" and 174 "stroke" cases. Following the combination of oversampling minority class and undersampling majority one, there was 1730 "no stroke" and 1757 "stroke" class. After applying the model on training data and then on test data, with default CP (complexity parameter), we got:

Table 1: Statistics of training set – ***Default CP value***

Accuracy	0.8374
Sensitivity	0.9197
Specificity	0.7538

Table 2: Statistics of test set – ***Default CP value***

Accuracy	0.7276
Sensitivity	0.7973
Specificity	0.7239

After the cross-validated accuracy rate for the 50 different parameter values, we can see in Figure 5., that the best accuracy is accomplished when CP is 0.000. This will tend to overfit the model, the same as underfitting for higher CP values.

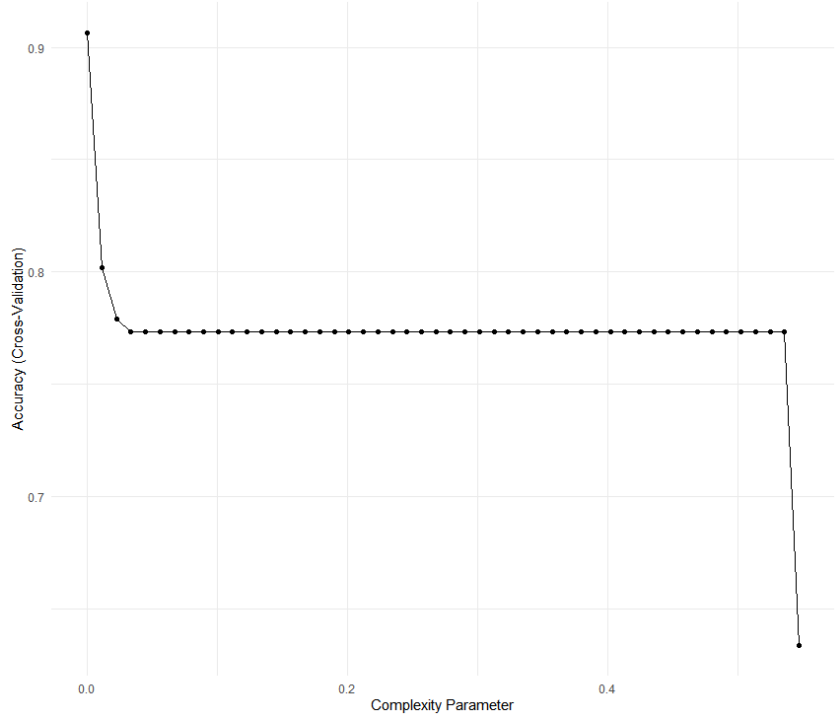

Figure 5: Different CP values and accuracy measure

*Table 3: Statistics of training set – **CP=0.00***

Accuracy	0.9475
Sensitivity	0.9875
Specificity	0.9069

*Table 4: Statistics of test set - **CP=0.00***

Accuracy	0.8213
Sensitivity	0.5675
Specificity	0.8345

If we set CP=0.54682081, our model will be too simple and will underfit with the test accuracy=0.0495. Although accuracy is important, it is not the main measure in our case. With an unbalanced test set, and the nature of the problem (this is the question of life and death), we should look more at measures such as sensitivity and specificity. Sensitivity is the ability of a test to correctly identify patients with a disease, and in our case, it is more important to look at. If we make more false negatives, it means that people with stroke are classified as they do not have one, which is a big problem. On the other side, if we have a false positive, it means that someone who does not have a stroke is considered as he has one, which could also be an inconvenience but not big as the first one. In our case the default value of CP will be set, it is CP=0.011 with the measures from Table 1. and Table 2. Below is a graphical representation of the last CP value.

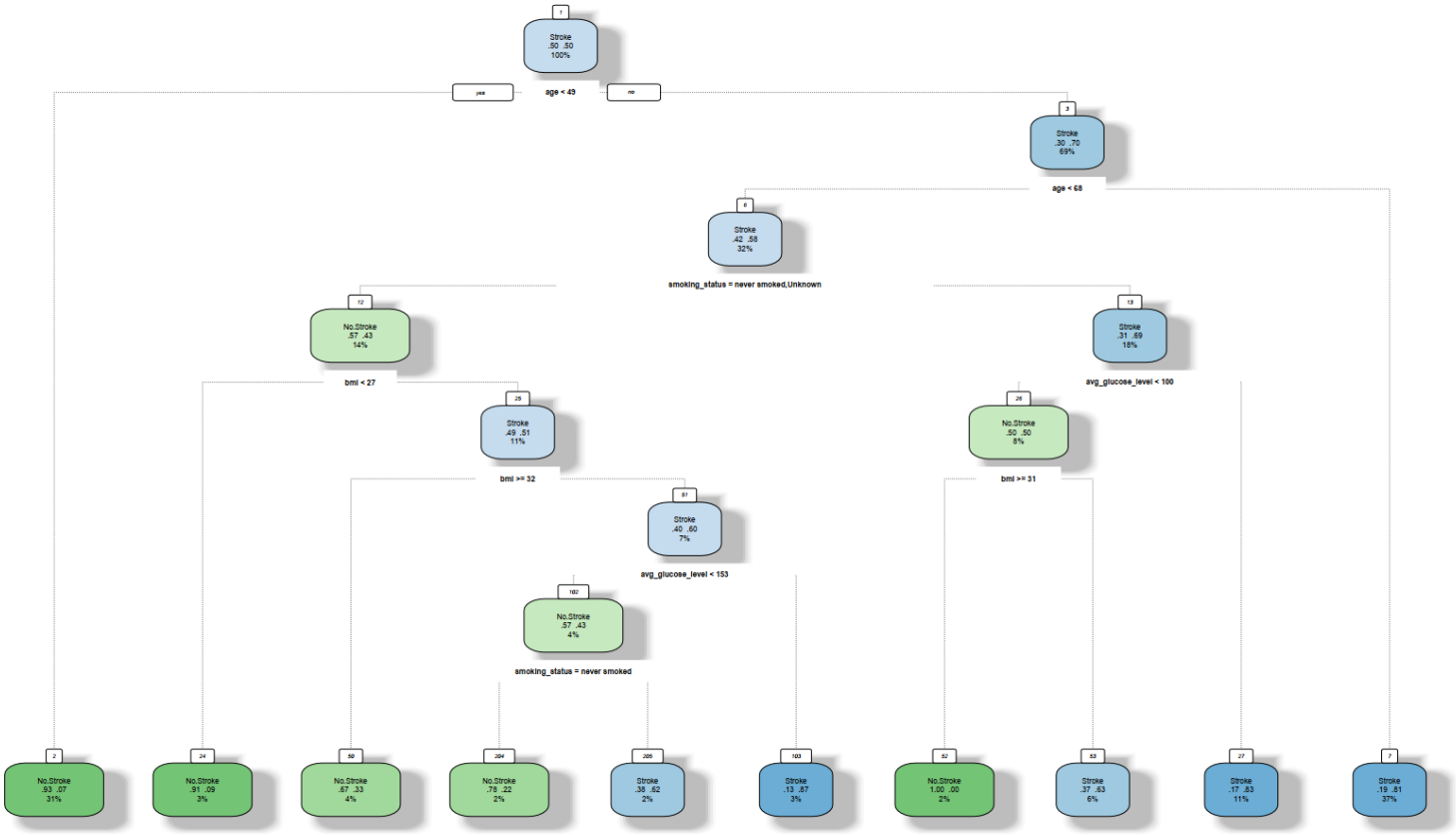

Figure 6: Graphical representation of decision tree - CP=0.011

Also, it is important to look at the confusion matrix, to see the raw numbers of true positives and true negatives.

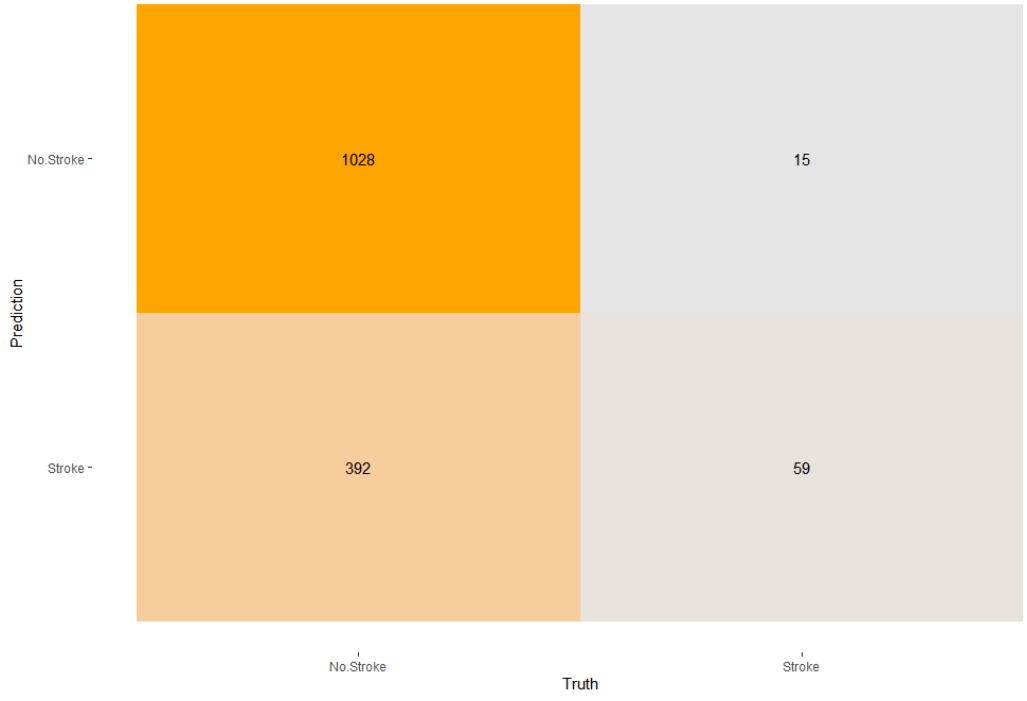

Figure 7: Confusion matrix - CP=0.011

In the confusion matrix, we can see that our model classified 15 people with "No stroke" even though they had a stroke (false negative), and 392 were classified as they had a stroke bud they did not (false positive). If interpretability is the goal, the complexity parameter can be lowered and get a tree like in Figure 8.

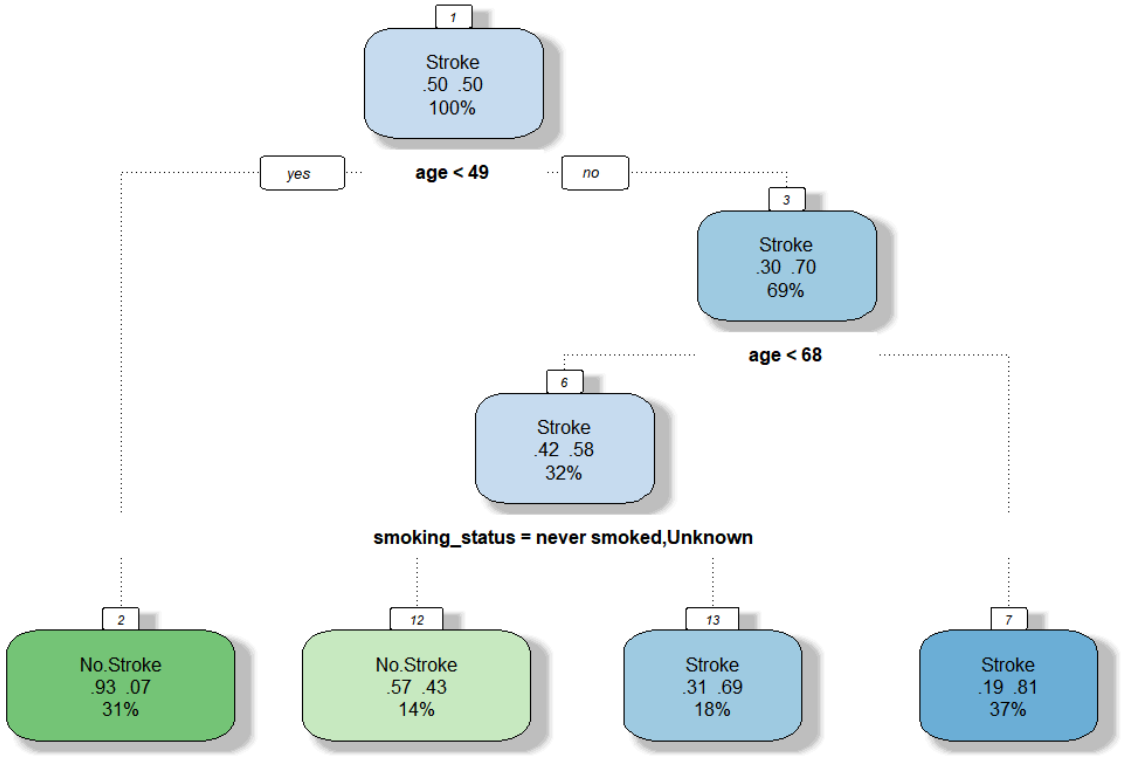

Figure 8: Simple tree - CP=0.2

It can be interpreted as: If the patient is less than 49 years old, there is a 93% probability that it is no stroke. On the other side: If the patient is older than 49 years but younger than 68 years and a smoker, there is a 69% probability that it is a stroke.

Conclusion

Even though decision trees are easy to interpret and to set up, it can easily overfit and become complex structure. Also, in the future there should be implemented other algorithms such as support vector machines or logistic regression, since decision trees are not so famous accuracy wise. Important thing in this unbalanced datasets is that accuracy is not that good measure to pursue, one should be focused more on sensitivity and specificity.