Project Outline

Kira Wolff

April 2021

Research Question

After a stroke occurs, every minute counts (**darehed2020?**). The sooner the affected person receives medical care, the higher the chances of a good recovery and little subsequent damages. Thus it is key to recocgnize a stroke occured. Besides knowledge about the visible symptoms of a stroke, knowledge about who is at high(er) risk to suffer a stroke can help channeling valuable attention resources on the these people.

Therefore, in my project I will work with the research question which medical or lifestyle factors are associated with a stroke occurence to draw conclusions about them as risk factors.

Data

The data with which I will investigate my research question is publically accessible via kaggle.com. The dataset is called “Stroke Prediction Dataset,” uploaded on the 26th of January 2021 by the user fedesoriano (https://www.kaggle.com/fedesoriano/stroke-prediction-dataset). Unfortunately, no further sources other than “confidential” for the data or information about the collection methodology or author are given. Thus, the validity of the data is uncertain and conclusions should be drawn with a grain of salt. Nevertheless, the data superficially seems fit to investigate the research question.

The dataset includes information about 5110 people. There is demographic data (age, gender), medical data (Hypertension, Heart Disease, Glucose Level, BMI, Stroke), and lifestyle data (Marriage, Work Type, Residence Type, Smoking Status). Except for age, BMI, and glucose level, all variables are categorial.

The sample has a large age range from children younger than a year to people older than 80 years. 59% of the participants are female, 41% are male. Most people are healthy, with only 16.8% suffering from having/having had hypertension, heart disease or a stroke.

Methods

The research question is about a classification problem with stroke occurence as a binary outcome.

I will use multiple methods and multiple combinations of covariates to identify the model with the highest accuracy.

The methods include logistic regression, regularized logistic regression, SVM, and random forest.

Logistic regression is useful as “standard procedure” and to establish a baseline with relatively simple methods. Since the goal is high accuracy of the prediction, a regularization term in form of Elastic Net might be useful. In case the relationship of the variables is better represented by a more complex function, SVMs will be applied. Finally, random forest as a faster alternative will be applied. This might be especially useful to identify the most important/impactful covariates.

To tune hyperparameters and estimate the test error, I will make use of k-fold-cross-validation.

Expectations

Regarding the modeling methods, I do not expect the more complex methods to necessarily achieve a higher accuracy. Depending on the relationship of the variables, elaborate methods might not be necessary. Nevertheless, it is interesting to compare the methods to get a feeling for possible gains in accuracy.

Regarding the research questions, I rate the medical data as most promising covariates. In the end, a stroke is a result of certain physical processes which are probably more associated with other physical processes than environmental factors like work or residence type. Although, some factors I identified as “lifestyle” are obviously also strongly associated with physical processes, e.g. smoking.

One potential problem I see with this dataset is the number of people who had a stroke. Although the sample is not small, only 249 people (roughly 5%) were affected by a stroke, and I am not sure if this suffices to achieve good accuracy, especially when this group is further splitted due to cross-validation. I think the modeling methods will technically still work, but the results may be unsatisfying. With this “small” subsample, it is likely that not all combinations of covariates will be covered with data. Thus, conclusions about the risk factors might only be applicable for the subgroup covered in the dataset.