Stroke Risk Factors

Kira Wolff

Introduction

After a stroke occurs, every minute counts (Darehed et al. 2020). 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 recognize 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.

Research Question

The goal of this project is to identify medical and/or lifestyle factors that are associated with stroke occurence by constructing a model that predicts stroke occurence and finding the important predictors.

Method

To answer the research question, data about people with and without a stroke in their medical history will be used. The goal is to build a model that classifies the people correctly as stroke patients and people without stroke history. This model could then be applied to other people whose basic demographic, medical and lifestyle data is known. If the model classifies them as stroke patients although they did not experience one, they should be treated as high risk.

Simple comparisons by group will be executed with t-tests or χ^2-tests for continuous and categorial variables, respectively.

As common procedure for binary classification problems, a logistic regression will be conducted. Considering potentially more complex connections between the variables, further modeling approaches will be used. Firstly, a regularization term in form of Elastic Net will be added which is especially sensible in prediction contexts as it reduces overfitting. Elastic Net was choosen instead of LASSO or Ridge regularization because it enables a sort of compromise between the other methods, as both parameters can be tuned. Secondly, to consider the case that the relationship of the variables is better represented by a more complex function, support vector machines (SVMs) with a linear kernel will be applied. Finally, a classification tree using the random forest algorithm will be applied. This might be especially useful to identify the most important/impactful variables, that is, predictors.

To train the models and tune the parameters, repeated 5-fold cross-validation will be used. Additionally, the final models' performance will be tested with a subset of the data that is not part of the training, consisting of 20% of the original dataset. The best model will be identified via accuracy as performance criterium. Sensitivity and Specifity will also be considered, as well as the general complexity of the model - in case of near identical performance, simpler models will be chosen.

Finally, the best model will be evaluated regarding the predictors. If possible, a p-value of 0.05 will be considered as a significant result.

Implementation

The analysis will be conducted using *R* version 4.0.3 (R Core Team 2020) in *RStudio* version 1.3.1093 (RStudio Team 2020). Data will be handled via *tidyverse* (Wickham et al. 2019), primarly with *dplyr* [dplyr2020], and plotted with *ggplot2* (Wickham 2016), *scales* (Wickham and Seidel 2020) and *ggpubr* (Kassambara 2020). The models will be implemented using *caret* (Kuhn 2020) and *randomForest* (Liaw and Wiener 2002). For resampling, the methods offered by *caret* will be used. Data will be preprocessed by centering and scaling before building the models.

Data

The data comes from the publicly accessible "Stroke Prediction" dataset which was published by the user fedesoriano on kaggle.net in January of 2021 (fedesoriano 2021). Unfortunately, no further information about the validity of the data is known. The dataset contains information about demographic, medical and lifestyle data from >5000 people.

The independent variable of interest, stroke occurence, is represented in the data as a binary variable, corresponding to having had a stroke in the own medical history or not. Unfortunately, the data is quite unbalanced with a ratio from 19.52:1 for the subjects not affected by a stroke.

Missing Data were handled via single imputation by predicting the missing values with a linear regression model.

A few variables show disputable values for which more context or explanation by the dataset's creator would have been useful. Since this information is not provided, I try to evaluate them as appropriate or inappropriate.

Among the age values, multiple cases show values smaller than one. Although this is not a common way to display age below one year, this is the most obvious explanation. It is not far-fetched since many other cases are underage as well. Although the age of people older than one year is just displayed as integer, the decimals might have been used for infants because an age of zero is even harder to interpret or rather incorrect. Additionally, several months make a large developmental difference for infants which could be considered here. Lastly, reducing the values to "<1" would have removed the possibility to calculate with age as a continuous variable.

Among the BMI values, outliers reach values up to 97. Although these values are extreme, they are not impossible and are thus likely valid.

Concerning the smoking status variable, information about 30% of the subjects in unknown which makes it difficult to evaluate the contribution of smoking to the research question. Nevertheless, the other extractable information from this variable can be useful.

Results

This section will describe the sample in more detail, build the models and evaluate them.

Subjects

The data represents information about 5110 subjects. The sample has a large age range from <1 to 82 years with a mean of 43.23 years ($SD = 22.61$). It consists of 58.59% women, 41.39% men, and 0.02% identifying with other labels.

As Figure 1 shows, the majority of the sample has no history of hypertension, heart disease, stroke or elevated blood glucose level. Excluding the latter, 16.81% are affected by at least one of the aforementioned diagnoses

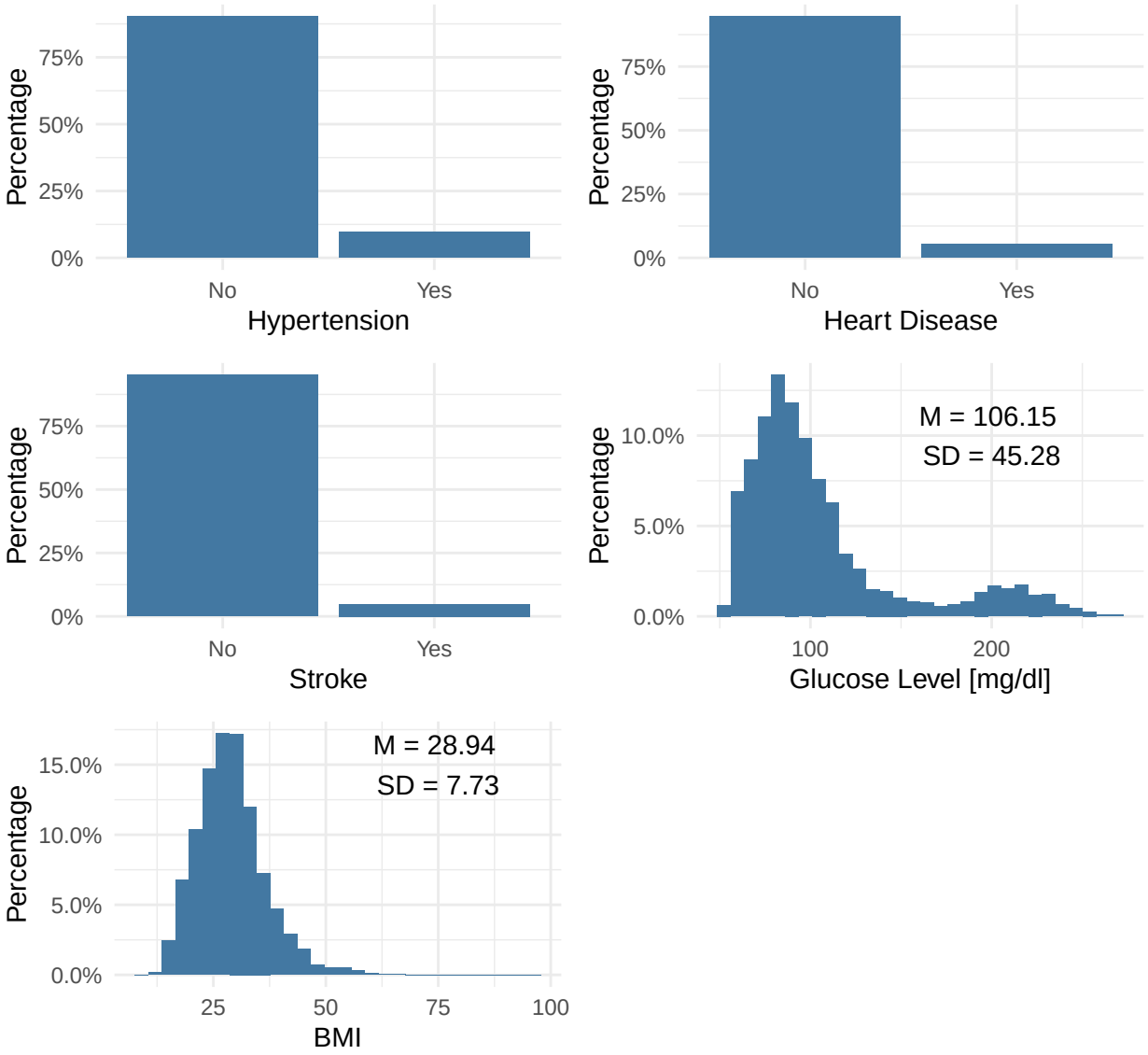

Figure 1: Distributions of Medical Data

overall. According to the BMI data, all categories from underweight to obese are represented by the sample (though BMI calculations for minors have to be interpreted with reservations).

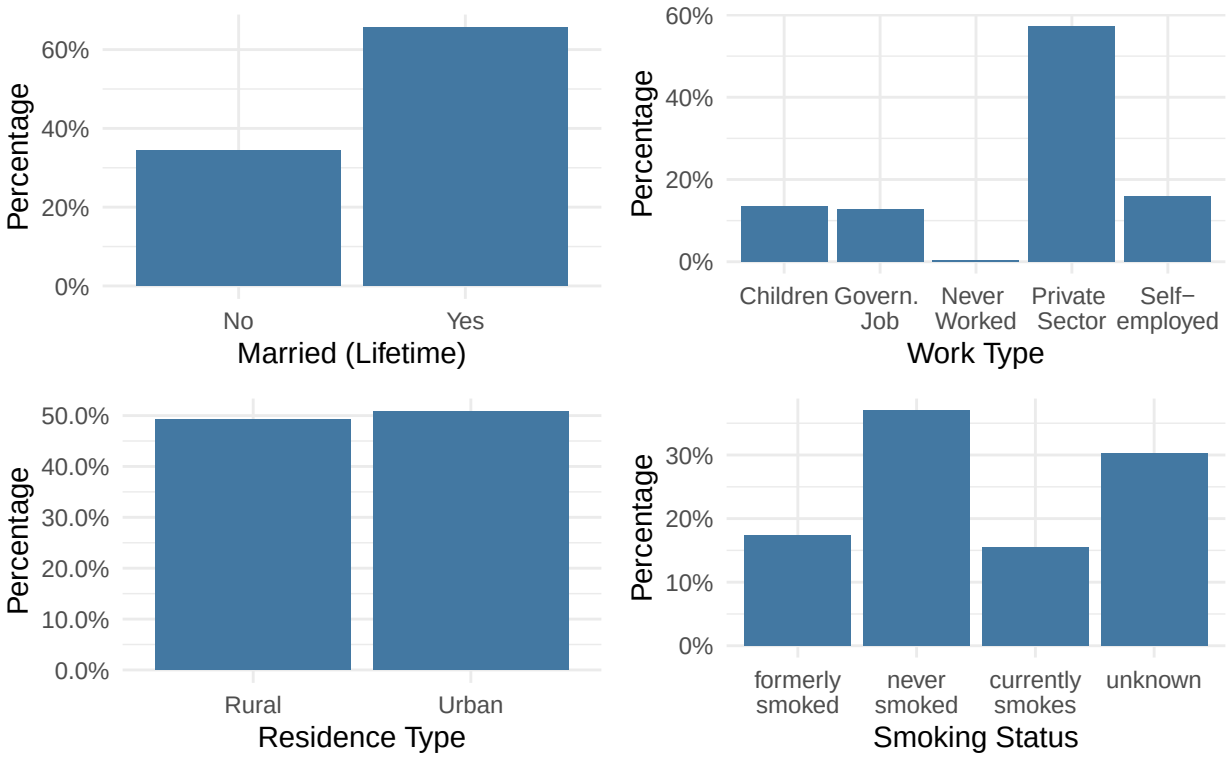

Figure 2: Distributions of Lifestyle Data

Concerning the lifestyle of the sample, the majority is married or has been at some point in their life. 86.13% were employed or self-employed at the time of data collection, while 13.44% were counted as children. Regarding their residence, the sample splitted almost fifty-fifty into rural and urban population. At last, considering only the subsample whose smoking status is known, 53.06% never smoked, while 24.82% smoked formerly and 53.06% currently still do. More information can be seen in Figure 2.

In general, the sample seems to be quite representative of the general population, without focusing on a specific subgroup. Thus, the results of the following analysis can potentially be transferred to the general population. Although one should keep in mind that some variable connections only appear in certain subgroups, so working with samples like this is not always beneficial.

Table 1: Sample Data by Stroke Occurence (Mean (SD) or Percentages)

	No Stroke	Stroke	p
n	4861	249	—
Gender (m/f)	58.7/41.3	56.6/43.4	0.79
Age	42 (22.3)	67.7 (12.7)	<0.001
Hypertension (n/y)	91.1/8.9	73.5/26.5	<0.001
Heart Disease (n/y)	95.3/4.7	81.1/18.9	<0.001
Glucose [mg/dl]	104.8 (43.8)	132.5 (61.9)	<0.001

	No Stroke	Stroke	p
BMI	28.9 (7.8)	30.3 (5.9)	<0.001
Ever Married (n/y)	35.5/64.5	11.6/88.4	<0.001
Residence (rural/urban)	49.4/50.6	45.8/54.2	0.298
Smoking Status (formerly/never/currently)	16.8/37.1/15.4	28.1/36.1/16.9	<0.001

A comparison of the variables by stroke, as depicted in Table 1, shows the potential of the variables as predictors: For almost all variables, a significant difference emerges. Especially age and the variables concerning health, i.e. hypertension, heart disease, and glucose show large differences.

Model Evaluation

The logistic regression provides a model with $Acc = 0.953$ which reaches $Acc = 0.943$ with the test data.

The hyperparameters of the logistic regression regularized with Elastic Net were tuned to $\alpha = 0.1$ and $\lambda = 0.008$. This provides a model with $Acc = 0.953$ which reaches $Acc = 0.943$ with the test data.

The hyperparameter of the SVM with Elastic Net was tuned to $C = 0.05$. This provides a model with $Acc = 0.953$ which reaches $Acc = 0.943$ with the test data.

The hyperparameter of the SVM with Elastic Net was tuned to $mtry = 2$. This provides a model with $Acc = 0.953$ which reaches $Acc = 0.943$ with the test data.

Table 2: Comparison of the Models

Model	Accuracy	Sensitivity	Specifity
GLM	0.943	1	0
Regularized Regression	0.943	1	0
SVM	0.943	1	0
Tree	0.943	1	0

The performance of the different models is summarized in Table 2.

Predictor Evaluation

Since the logistic regression reached highest accuracy while being the simplest model, it will be used to take a closer look at the predictors.

The logistic regression with stroke as outcome yields age ($b = 1.72$, $p < 0.001$) and glucose level ($b = 0.21$, $p = 0.001$) as significant predictors.

This means, that for each year older the odds to suffer a stroke are multiplied by 5.58, and for every increase of 1 mg/dl of average blood glucose, the odds to suffer a stroke are multiplied by 1.23. To be precise, these values apply to the "average" person from the sample with average values for all variables. The possibility that the odds change differently for people with values far from the mean cannot be eliminated with this analysis.

Although the other predictors are strictly speaking not significant, the p-value of two further predictors is close to 0.05. On the one hand, self-employment reduces the risk by 55.4% ($p = 0.065$), on the other hand urban residence increases the risk by 113.9% ($p = 0.089$).

Discussion

According to accuracy as performance criterium, there is no difference between logistic regression, regularized logistic regression, SVM and random forest. Due to the principle to keep modeling as simple as possible, as long as performance does not suffer, logistic regression is to be preferred in this case. The more complex models did not improve prediction performance.

On the first glance, the accuracy of 0.953 seems to be a good result for the goal to build a model to predict stroke occurence. Unfortunately, when considering specifity as an additional performance criterium, the prediction does not work well. Considering the data, the reason for this probably lies in the ratio from subjects with to subjects without stroke which is 1:19.52. This ratio leads to an unbalanced dataset, and accuracy can be a misleading performance criterium: Because of the high portion of healthy subjects, high accuracy is easy to acquire by always predicting "no stroke." There are solutions for unbalanced datasets like Over- or Under-Sampling which unfortunately go beyond the scope of this project work.

Although the created models will thus not work well to predict future stroke cases, the predictors can still be interpreted. The logistic regression presented age as a factor which multiplies the odds to suffer a stroke by 5.58 for each passing year. This is not unexpected, as the prevalence of many diseases increases with age and stroke is among the most common causes for people 65 years of age and older (Sahyoun et al. 2001). As second predictor, the average glucose level emerged by multiplying the odds of a stroke by 1.23. The glucose level is tightly-knit to the diagnosis of type 2 diabetes which is known to increase stroke risk as well (Sander, Sander, and Poppert 2008).

In addition to the significant predictors, the other variables are also interesting to answer which factors are rather not relevant to assess the stroke risk. Here, demographic and medical data seem to be more important than lifestyle data, though the results for residency type and self-employment are interesting. It is possible that some of the other, not-significant variables correlate with age and thus could not contribute more than age already did.

Limitations

The limitation of this work lies primarly in the unbalanced dataset, which impeded the model construction with the known methods.

Additionally, the validity of the data is not certain, though no variables or values stood out in an all too negative way. Also referring to metadata, it would have been useful to know how the data where collected, especially in which city, region, country or culture. With this lack of information, generalization of the results is limited.

Future Research

Future research should try to work around the aforementioned limitations, i.e. use methods to transform the data into a more usable, balanced dataset. Another possibility would be to use different data that is more balanced to begin with.

Although a heterogenous dataset that represents the general population well can be important for classification/prediction research, it can also be of use to work with a specific subset to identify risk factors of this subset. It would even be useful to identify which risk factors apply to everyone and which are more relevant for specific subgroups.

Conclusion

Unfortunately, the original goal of this project work to build a well-working stroke prediction model could not be achieved due to the unbalanced data. Nevertheless, age and glucose level could be identified as important predictors.

References

Darehed, David, Mathias Blom, Eva-Lotta Glader, Johan Niklasson, Bo Norrving, and Marie Eriksson. 2020. "In-Hospital Delays in Stroke Thrombolysis: Every Minute Counts." *Stroke* 51 (8): 2536–39.

fedesoriano. 2021. "Stroke Prediction Dataset." https://www.kaggle.com/fedesoriano/stroke-prediction-dataset.

Kassambara, Alboukadel. 2020. *Ggpubr: 'Ggplot2' Based Publication Ready Plots.* https://CRAN.R-project.org/package=ggpubr.

Kuhn, Max. 2020. *Caret: Classification and Regression Training.* https://CRAN.R-project.org/package=caret.

Liaw, Andy, and Matthew Wiener. 2002. "Classification and Regression by randomForest." *R News* 2 (3): 18–22. https://CRAN.R-project.org/doc/Rnews/.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

RStudio Team. 2020. *RStudio: Integrated Development Environment for r.* Boston, MA: RStudio, PBC. http://www.rstudio.com/.

Sahyoun, Nadine R, Harold Lentzner, Donna Hoyert, and Kristen N Robinson. 2001. "Trends in Causes of Death Among the Elderly." *Aging Trends* 1 (1): 1–10.

Sander, Dirk, Kerstin Sander, and Holger Poppert. 2008. "Stroke in Type 2 Diabetes." *The British Journal of Diabetes & Vascular Disease* 8 (5): 222–29.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, and Dana Seidel. 2020. *Scales: Scale Functions for Visualization.* https://CRAN.R-project.org/package=scales.