Exploratory Data Analysis

Patrick Brus

28.05.2021

Abstract

Stroke is a medical condition where too little blood flow to the brain causes cell death. Stroke is the second leading cause of death globally. If a stroke can be detected early, the chances of recovery are much better. The stroke prediction data set on Kaggle aims at predicting whether a patient is likely to get a stroke by using other input features. In this report the data set is analyzed and pre-processed such that machine learning can be executed directly afterwards.

1 Introduction

According to [3], stroke is the second leading cause of death globally with an increasing influence since the population ages. The Covid-19 pandemic is still not completely under control and is still affecting our daily routines. Stroke patients are also at higher risk for a severe course of COVID-19 disease [4]. It can be very helpful to recognize a stroke disease at an early stage and to initiate countermeasures afterwards. The Kaggle data set on [2] aims at predicting whether a patient is likely to get stroke based on certain input parameters. The underlying machine learning task is therefore a classification. Before applying machine learning, the data set should be explored and some feature engineering should be applied in order to understand the data and to get the best performance of the machine learning algorithm. This report summarizes the findings after performing some Exploratory Data Analysis (EDA). In addition, feature engineering and pre-processing is applied in order to prepare the data set for machine learning. In the end, a training set and a testing set is created. The testing set should only be applied after the machine learning model is trained and optimized. The aim of the hold-out testing set is to check the performance of the model in the real world.

2 Attribute Information

This chapter gives a quick overview of all available features and a short description of them. Table 1 shows all feature names and a short description of each feature.

3 Initial Plan for Data Exploration

This chapter describes the initial plan for performing EDA. Python is to be used as programming language, while pandas, seaborn and matplotlib are the libraries to be used for data exploration and visualizations. First, the data is to be loaded into a pandas data frame. As a next step, the pandas $info()$ function is to be used in order to get some insights about the data types of all features and if whether there are missing values or not. The pandas $describe()$ method is then to be used in order to get some statistics from the numerical features. Drop the features from the data frame, if there

Feature Name	Description
id	unique identifier
gender	Male, Female or Other
age	Age of patient
hypertension	0 if the patient doesn't have hypertension, 1 if the patient has hypertension
heart_disease	0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
ever_married	No or Yes
work_type	children, Govt_jov, 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
stroke	1 if the patient had a stroke or 0 if not

Table 1: An overview of all features available in the stroke prediction data set.

are only unique values of that feature, because that feature will not serve the machine learning algorithm as useful information. The target variable should also be checked for data imbalance in its distribution, since appropriate countermeasures would have to be taken. Afterwards, plot a histogram of all numerical features in order to get first insights in their distributions. If there are skewed histograms, then a log-scaling could be applied in order to repair the skewness. In addition, the pandas *value_counts*() function is to be used on all categorical features in order to check their categories and the number of samples for each category. If there are a lot of categories only containing one sample, then it could be useful to combine them to one category. As a next step, seaborn is to be used to plot histograms of categorical features and the *hue* parameter is to be used in order to already get some insights on the influence of some features on the target variable. Seaborn is then also to be used in order to create box-plots for all numerical features, separated by the target variable. As a last two steps, the seaborn *pairplot* and *heatmap* functions are to used for checking the distribution of all features pairwise and to check the correlation of the independent variables.

4 Key Findings of Exploratory Data Analysis

This section discusses some findings through the EDA process. Table 2 shows the mean, min and max values of the features age, body mass index and average glucose level. The oldest person in the data set is 82 years old, while the youngest person is less than a month old. The average age is at 43.2 years. That approximately equals the European average age, which is 43.7 years [1]. The average body mass index is at 28.89 and the average glucose level at 106.15. In total, 201 samples contain missing values for the body mas index feature. These values could be dropped, but dropping values is always a bad idea because maybe useful information could get lost. Therefore, a suitable data imputation has to be executed during the feature engineering and data cleaning part. The column "id" only contains unique values and can therefore be dropped later, because it

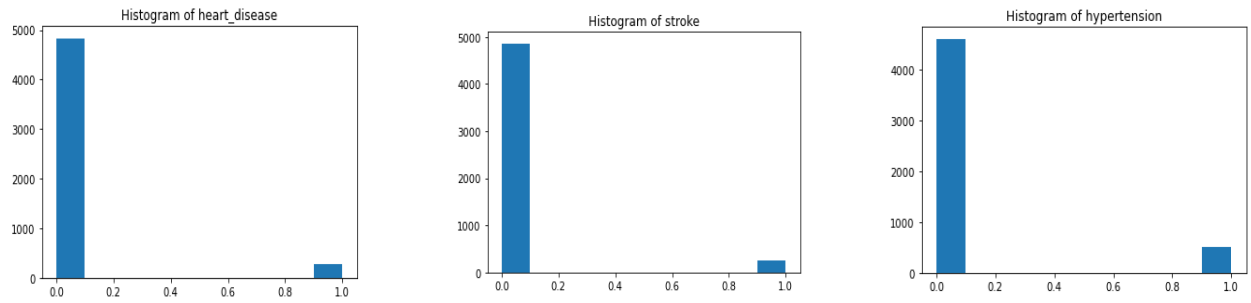

Figure 1: Histograms of heart disease (left), stroke (middle) and hypertension (right).

does not contain any valuable information for the machine learning algorithm. Figure 1 shows the histograms for the medical diseases in the data set. As one can see, the most people in the data set are healthy and only a small portion contains diseases. Figure 1 also contains the distribution of the target class "stroke", which is heavily imbalanced. In order to avoid the model being biased towards predicting "no stroke", a countermeasure has to be applied, like applying over-sampling of the minority class or under-sampling of the majority class. The best option is to apply different strategies and compare them to each other in order to find the best suited. In addition, the accuracy metric should not be used because of its weakness for imbalanced data. A better metric is the f1-score or even f2-score if more focus should be put on recall than on precision. Figure 2 shows the histograms of the numeric features body mass index, age and average glucose level. The distribution of the average glucose level is heavily skewed and can be corrected by applying a logarithmic scaling. This should enable the model to also perform well on samples having a larger average glucose level. Table 3 shows each different value of smoking status and the percentage of patients having strokes within each category. The groups formerly smoked and smokes are having the highest portion of stroke patients, but the percentages of the groups Unknown and never smoked are not that far away. Table 4 shows the same evaluation for the feature "ever_married", with 6.56% of the married people having stroke and only 1.65% of the never married people having stroke. Seems that being married can have some negative influence in regard to getting a stroke. Figure 3 shows the box plots for the average glucose level and the body mass index. The median of average glucose level of stroke patients is larger than the average of non stroke patients, so one could conclude that a larger average glucose level increases the risk of getting a stroke. Figure 4 shows the histogram of the age feature, with one histogram for non stroke patients and one for stroke patients. Both histograms are stacked together. As one can see, the higher the age the larger the risk of getting a stroke.

Feature	**Mean**	**Min**	**Max**
age	43.2	0.08	82
bmi	28.89	10.3	97.6
avg_glucose_level	106.15	55.12	271.74

Table 2: Some statistics of numerical features.

5 Data Cleaning and Feature Engineering

This chapter briefly describes the required actions for cleaning the data. The feature *id* is dropped from the pandas data frame, because it only contains unique values and is therefore not valuable

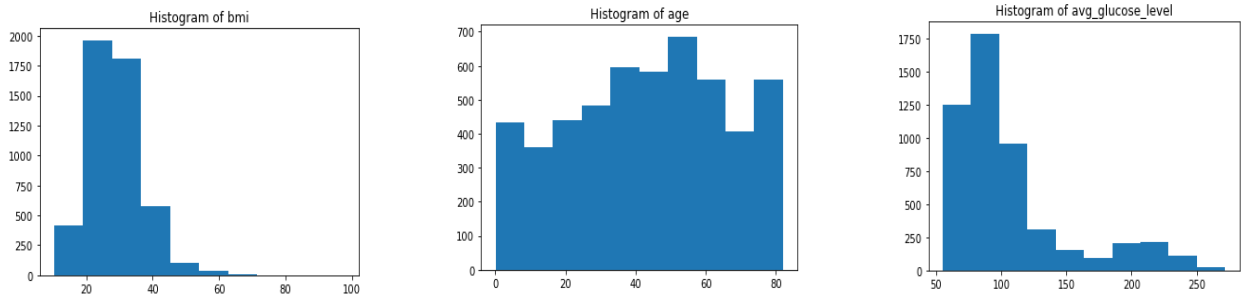

Figure 2: Histograms of body mass index (left), age (middle) and average glucose level (right).

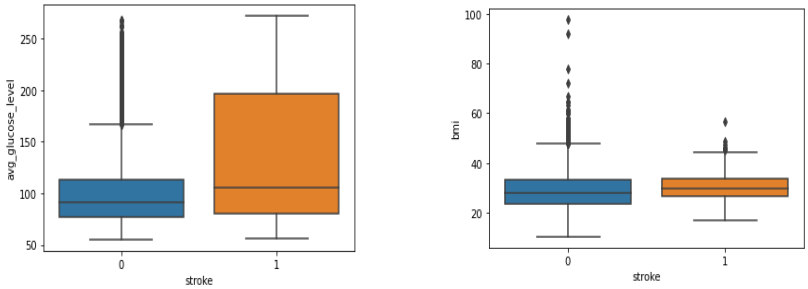

Figure 3: Box plots of average glucose level (left) and body mass index (right).

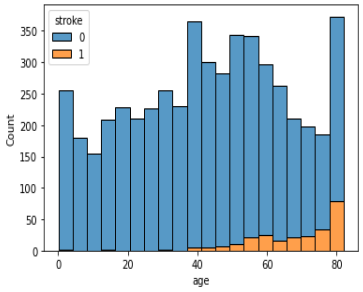

Figure 4: Histogram of age with hue parameter stroke.

Group	Percentage of Stroke Patients
Unknown	3.04%
Formerly smoked	7.92%
Never smoked	4.76%
Smokes	5.32%

Table 3: Different smoking status values and the percentage of stroke patients in each category.

Ever Married?	Percentage of Stroke Patients
Yes	1.65%
No	6.56%

Table 4: Different marriage status and the percentage of stroke patients within each group.

for the machine learning algorithm. The gender feature contains three possible outcomes, which are male, female and other. The category other only contains one sample. Therefore, this sample is dropped from the data set in order to avoid the machine learning model to overfit on this sample. As a next step, the gender feature is transformed to a numerical feature by mapping the category "male" to the numeric value one and the category "female" to the value zero. The same is applied for the feature "ever_married", where "Yes" is mapped to one and "No" is mapped to zero. Afterwards, all remaining categorical features are one-hot-encoded. The feature "avg_glucose_level" is scaled using log-scaling in order to repair the skewness in its distribution. The library scikit-learn is used in order to split the transformed data into a training set and a testing set. Afterwards, an iterative data imputer from scikit-learn is trained on the training set in order to impute the missing values for the feature "bmi". The data impute is only trained on the training set in order to avoid data leakage from the test set. As a last step, the data is normalized in order to ensure that all values are at the same scale. Again, the normalization values are only computed by using the training data and not the test data. This is again to avoid a data leakage problem form your test set into the normalization and therefore a risk of overly optimistic results on the hold-out test set.

6 Possible Hypothesis

This chapter describes three possible hypothesis, while the first one is tested using the library scipy. List of hypothesis:

1. Sample from feature age comes from normal distribution.
2. Feature gender does not influence stroke.
3. Feature smoking_status does not influence stroke.

The alpha for testing the first hypothesis is set to $\alpha = 0.05$. The returned p-value is less than α, which means that the null hypothesis can be rejected and the age distribution comes not from a normal distribution.

7 Data Quality

This chapter briefly summarizes the quality of the data set. The data set does only contain one feature with missing values. But these values can be easily imputed, so no data needs to get

dropped. The data set is heavily imbalanced, but there cannot be a data set which is perfectly balanced, because in the real world the amount of people having stroke is less than the amount of people not having stroke. The features are of good quality and no data cleaning on the feature values themselves is to be done. In total, there are 5110 samples which should be enough for training a precise machine learning model.

References

[1] eurostat. *Ageing Europe - statistics on population developments*. 2020. URL: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Ageing_Europe_-_statistics_on_population_developments#:~:text=In%202019%2C%20the%20median%20age,recorded%2C%20both%2037.7%20years). (visited on 05/28/2021).

[2] fedesoriano. *Stroke Prediction Dataset*. 2021. URL: https://www.kaggle.com/fedesoriano/stroke-prediction-dataset (visited on 05/28/2021).

[3] Mira Katan and Andreas Luft. "Global Burden of Stroke". In: *Seminars in Neurology* 38.2 (2018). ISSN: 10989021. DOI: 10.1055/s-0038-1649503.

[4] Hugh S. Markus and Michael Brainin. "COVID-19 and stroke—A global World Stroke Organization perspective". In: *International Journal of Stroke* 15.4 (2020). ISSN: 17474949. DOI: 10.1177/1747493020923472.