Data Exploration and Stroke Prediction Using Logistic Regression on Patient Data

Burak C. Soyak
Department of AI Engineering
Bahcesehir University
2100841
burak.soyak@bahcesehir.edu.tr

Tarkan Özşen
Department of AI Engineering
Bahcesehir University
2102276
tarkan.ozsen@bahcesehir.edu.tr

Aleyna Benan Aydı
Department of AI Engineering
Bahcesehir University
2003977
aleynabenan.aydi@bahcesehir.edu.tr

Abstract

Stroke is one of the leading causes of death and disability worldwide, precise prediction models are needed to determine individuals who are at a high risk. A stroke is a sudden interruption of blood flow to the brain. This paper presents a thorough investigation of data exploration and stroke prediction using patient data and linear regression techniques. The main goal is to create a predictive model that calculates the probability of a stroke occurring using readily available patient data.

1 The Object

Explore a readily available patient data including metabolic risks (like high blood pressure, high body-mass index (BMI), or high cholesterol) and behavioral factors (such as smoking, poor diet, and low physical activity). Using the insight obtained by exploring, create a regression model to predict stroke probabilities on the given test data.

1.1 EDA(Exploratory Data Analysis)

Within the Exploratory Data Analysis, an overview of information processing is presented for the data exploration and visualization, in purpose of better comprehending the present data. This can be further explored as understanding the relationship between variables, identifying occurring patterns and trends. In the analysis, utilized visualization techniques are as follows: pie charts, bar plots, point plots, Kernel Density Estimate (KDE) plots, violin plots and a Spearman Correlation Matrix.

1.2 Prediction with Logistic Regression

Logistic regression is a statistical modeling technique used to analyze the relationship between a binary dependent variable (such as presence or absence of an event) and one or more independent variables. By applying a logistic function to the data, it calculates the likelihood that an event will occur based on the provided predictors, revealing the importance and influence of each variable. It is frequently used in a variety of fields, such as social sciences, economics, and healthcare, to forecast outcomes and make defensible decisions based on the probability estimates.

2 Exploratory Data Analysis

2.1 Visualization Techniques

The Exploratory Data Analysis consists of 6 different visualization techniques, each of them picked for the purpose of adapting a specific correlation within the data for higher human comprehension.

2.1.1 Pie Charts

Pie charts are applied as an initial in order to acknowledge and visualize the stroke rates within the training data, original data and combined data; as well as allowing the data to be compared. This can ensure the preparation for difference in results and a more accurate assessment of the algorithm results. As seen in Figure 1, the original data stroke rates are considerably higher than the training data.

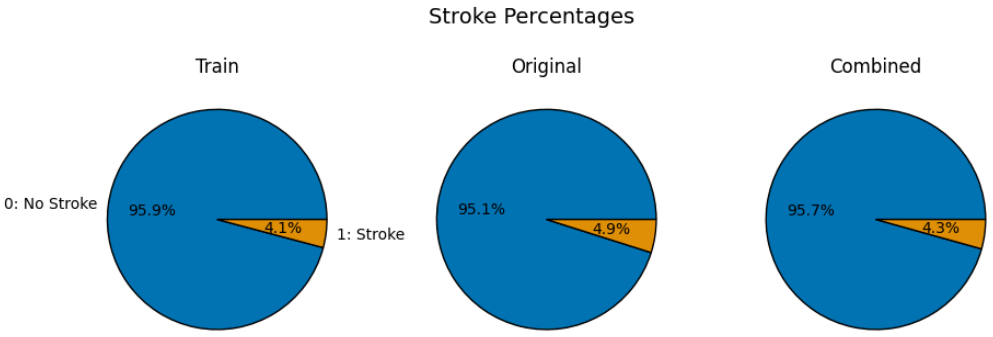

Figure 1: Pie Chart representing the Stroke percentages for the data.

2.1.2 Bar Plots & Point Plots

The bar plots below are used to assess the percentages of different variables. By examining the bar plots shown, insight regarding which variables may play a crucial role in stroke prediction can be gained. This is primarily useful in order to understand which characteristics are in the majority among patients and which of them are represented in problematically small values. When observed, point plots show the correlation between the before mentioned variables and stroke rates. These point plots going along with the bar plots result to easier comprehension of data analysis, specifically acknowledging how accurate the data is and how reliable it's ties to stroke rates are. In the bar plot

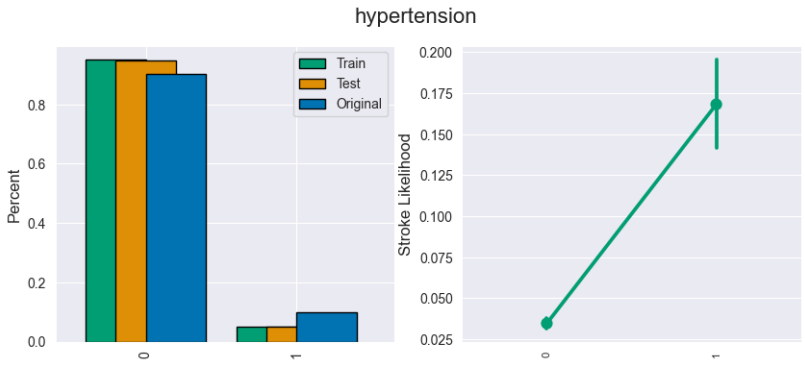

Figure 2: Hypertension in Data and vs. Stroke

of Figure 2, a big difference in numbers between patients with hypertension and no hypertension is observed. The point plot shows a much higher percentage of stroke for the patients with hypertension. These information can be analyzed to conclude that hypertension is a rare condition seen in less than 10% of the population; however, it increases the probability of a stroke by up to 4 times.

2.1.3 KDE Plots & Violin Plots

Kernel Density Estimate plots allow for an easy comparison between datasets while observing the density of patient characteristics, resulting in a readable analysis of various information. Violin plots combine the original and training datasets and compare the densities of characteristics to the stroke values. Violin plots display the distribution of a continuous variable using a combination of a rotated kernel density plot on each side and a box plot in the middle. These two techniques combined allow for an assessment of the influence of variables and their weights across the board.

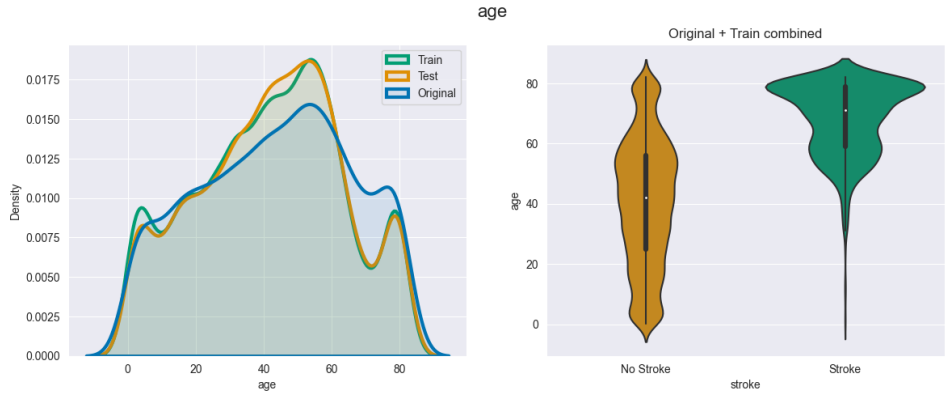

Figure 3: KDE of Age and Stroke vs Age Violin Plots.

Figure 3 displays the spread of patient ages with the KDE plot whilst visualizing the importance of age with the violin plot, pointing at a large increase in stroke rates as the patient age increases. Results worth noting are that all ages are fairly represented and accounted for, and that strokes are primarily a risk for people above the ages of 40, while 50% of patients who had a stroke were between the ages of 60 and 80.

2.1.4 Spearman Correlation Matrix

Spearman Correlation Matrix utilizes Spearman's rank correlation coefficient, measuring the strength and direction of the monotonic relationship. Hence, it is a technique that provides dense information to recognize the correlation between any and all variables, especially advantageous when the relationship is not linear. When applied to the datasets, SCM is highly useful in coming to general conclusions about trends seen within the number of patients.

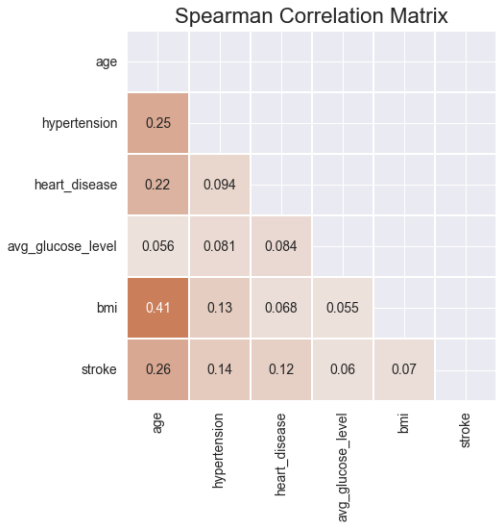

Figure 4: Correlation Matrix between Variables

When Figure 4 is analyzed, the following conclusions can be made. First of all, BMI and age seems to have the highest correlation of all data. Moreover, stroke rates appear to be most affected by age, followed by hypertension and heart disease, whereas glucose levels and BMI have lower amounts of correlation.

3 Prediction with Logistic Regression

Logistic regression can be extremely helpful in determining who is prone to have a stroke when it comes to datasets for stroke prediction. Using different input features like age, gender, hypertension, BMI values and smoking patterns, logistic regression models can be trained to predict the the risk that a patient will have a stroke.

3.1 Preprocessing

In the preprocessing stage, various steps are taken to prepare the data for the logistic regression model. Preprocessing techniques we used are explained here.

3.1.1 Dropping NA Values

There were only a handfull of NA values in the data and all of them are contained in the BMI column. Since we had a lot of present data, our decision of handling NA values was dropping them.

3.1.2 One-Hot Encoding the Categorical Values

To use the provided tools the Data needed to be reshaped and categorical values had to be turned into One-Hot encoded columns.

3.2 Training the Model

We Loop over the data five times Using sklearn.linear_model library's LogisticRegression function. While testing, we concluded that five loops were enough.

3.3 Predicting

After training the model enough over the training data, The values are predicted using the fit function and saved as a submission file.

The model we created using these methods got a score of 0.89 in the competition.

Github Page of the Code: `https://github.com/cancanasoyak/AIN2002`

References

Binary classification with a tabular stroke prediction dataset. Kaggle. (n.d.-a). https://www.kaggle.com/competitions/playground-series-s3e2/data

Fedesoriano. (2021, January 26). Stroke prediction dataset. Kaggle. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

Contributors:

Burak C. Soyak	Exploratory Analysis, Regression code and the Report(Typesetting).
Tarkan Özşen	The Report 1.1, 2.1, 2.1.1, 2.1.2, 2.1.3, 2.1.4, References
Benan Aydı	The Report 3, 3.1, 3.1.1, 3.1.2, 3.2, 3.3