Stroke Prediction Dataset Analysis

The following analysis refers to the Stroke Prediction Dataset, a public dataset that gathers clinical information from patients and whether or not they have suffered a stroke. Download made on the site https://www.kaggle.com/fedesoriano/stroke-prediction-dataset. The machine learning method - *Support Vector Machine (SVM)* made in the python programming language for *dataset* analysis was used.

The *dataset* contains just over 5000 observations, from clinical data such as whether the patient has hypertension to sociodemographic data such as age and whether he lives in the interior or city. The dataset gathers all these features and at the end records whether or not the patient has had a stroke.

The goal is to make a qualifying model with Support Vector Machine to predict if a person can have a stroke in the future.

Nós temos 11 features:

1) id: identification
2) gender: "Male", "Female" or "Other"
3) age: patient's age
4) hypertension: 0 if you don't have hypertension, 1 if you have hypertension
5) heart_disease: 0 if you have no heart disease, 1 if you have heart disease
6) ever_married: "No" or "Yes"
7) work_type: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
8) Residence_type: "Rural" or "Urban"
9) avg_glucose_level: blood glucose mean
10) bmi: body mass index
11) smoking_status: "formerly smoked", "never smoked", "smokes" or "Unknown"*

And the target:

1) stroke: 1 if you have a stroke or 0 if you have not had a stroke

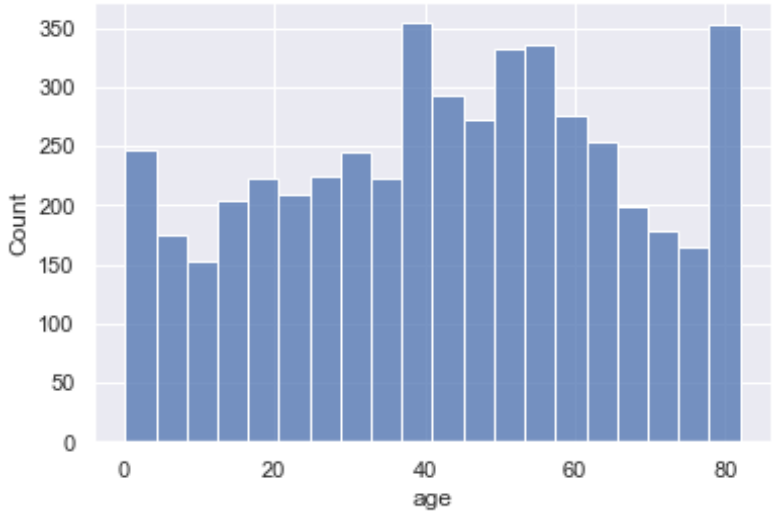
Table 1 - Dataset

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1
...	...	...	...	...	...	...	...	...	...	...	...	...
5105	18234	Female	80.0	1	0	Yes	Private	Urban	83.75	NaN	never smoked	0
5106	44873	Female	81.0	0	0	Yes	Self-employed	Urban	125.20	40.0	never smoked	0
5107	19723	Female	35.0	0	0	Yes	Self-employed	Rural	82.99	30.6	never smoked	0
5108	37544	Male	51.0	0	0	Yes	Private	Rural	166.29	25.6	formerly smoked	0
5109	44679	Female	44.0	0	0	Yes	Govt_job	Urban	85.28	26.2	Unknown	0

Source: The Author

In table 1, we see all the features and the target in the last column. You can also observe rows with "NaN" data that will be removed for analysis.

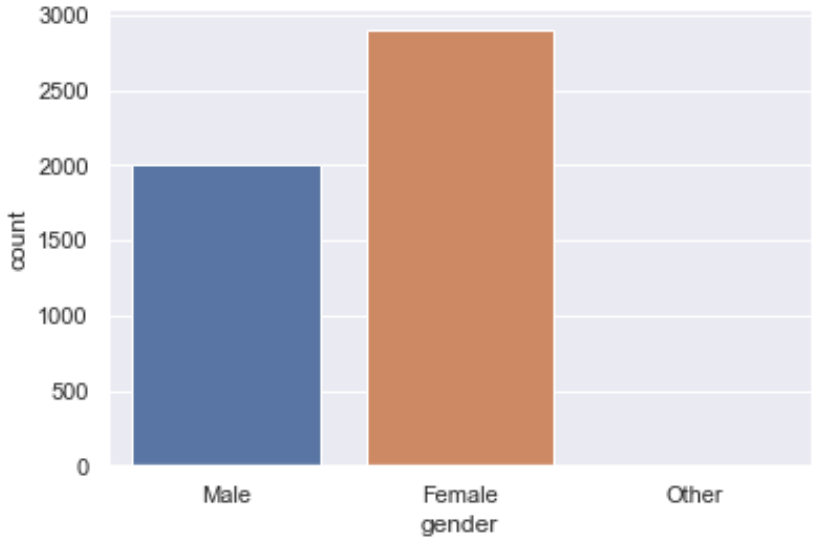
Figure 1 - Histogram 'age'



Source: The Author

Figure 1 shows the histogram of the age of the patients and reveals that most of them are between 40 and 60 years old. The mean age is 42.9 years.

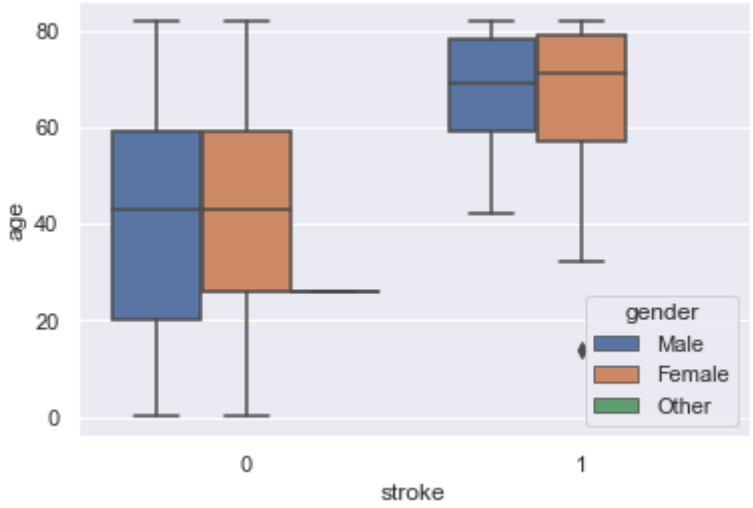
Figure 2 - Bar graph of the gender



Source: The Author

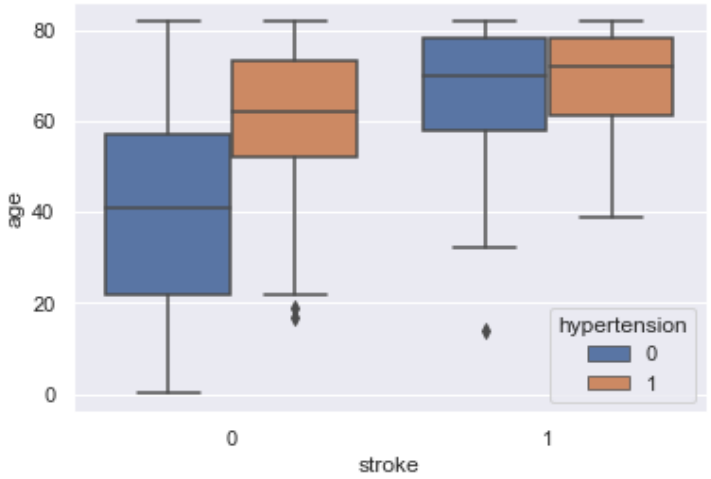
As seem in Figure 2, we have that women are the majority in this dataset.

Figure 3 - Boxplot de age and stroke with hue=gender



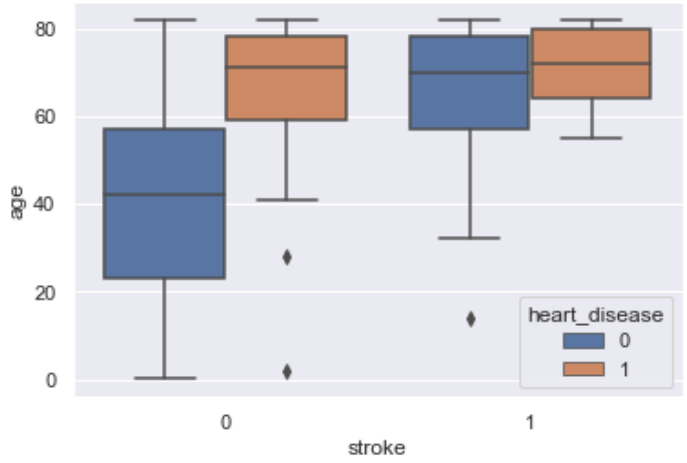
Source: The Author

Figure 4 - Boxplot de age and stroke with hue=hypertension



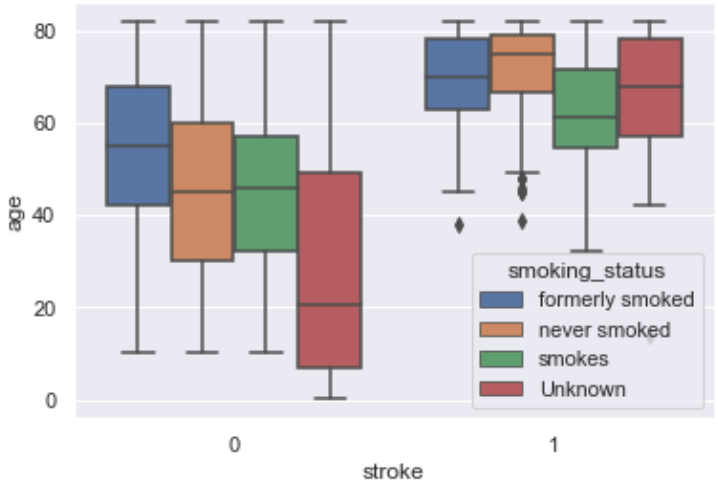
Source: The Author

Figure 5 - Boxplot of age and stroke with hue=heart_disease



Source: The Author

Figure 6 - Boxplot of age and stroke with hue=smoking_status

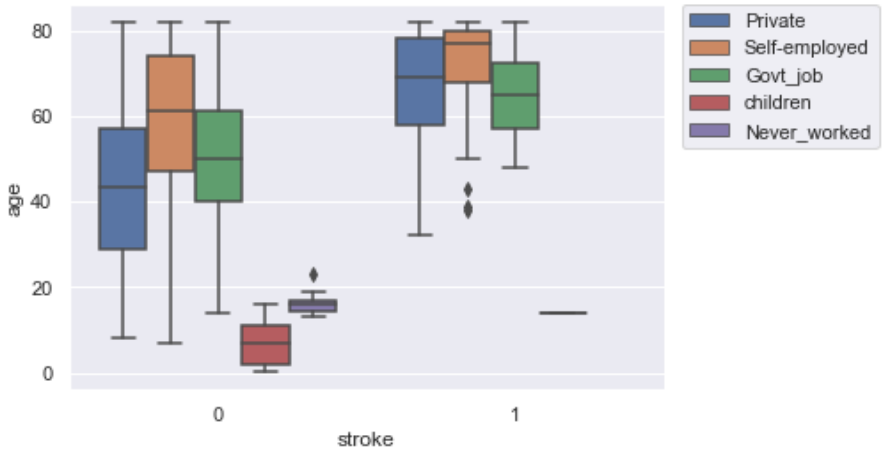


Source: The Author

To verify the age influences in cases of stroke differently in men and women, we have Figure 3. In it, it is shown that women begin to have strokes earlier than men, about 5 years earlier.

Hypertension, in this case, begins to appear from the age of 40 and does not show to be an aggravating factor when compared to the other group that had a stroke, but does not have hypertension, as seen in Figure 4. Figure 5 shows that the presence of heart diseases mainly affects older people and, in them, we have cases of stroke. In Figure 6, we have confirmation that smoking is related to stroke. Seen in green on the chart, those who smoke may have stroke from the age of 36 and those who do not smoke and who were smokers, approximately 43 years and 50 years, respectively.

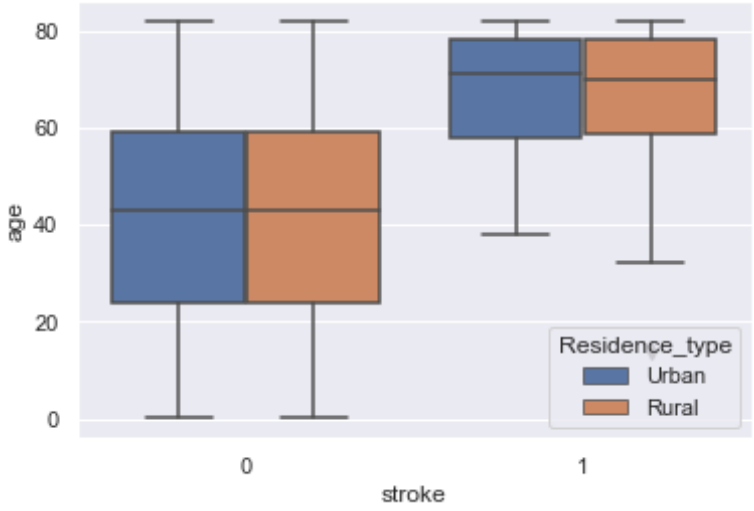
Figure 7 - Boxplot of age and stroke with hue=work_type



Source: The Author

The type of work, in private or self-employed companies, for example, are also relevant in future cases of stroke. In Figure 7, we have those cases of stroke begin to appear at the age of 30 for those who work in private companies, unlike those who are self-employed who begin to have cases of stroke with almost 50 years.

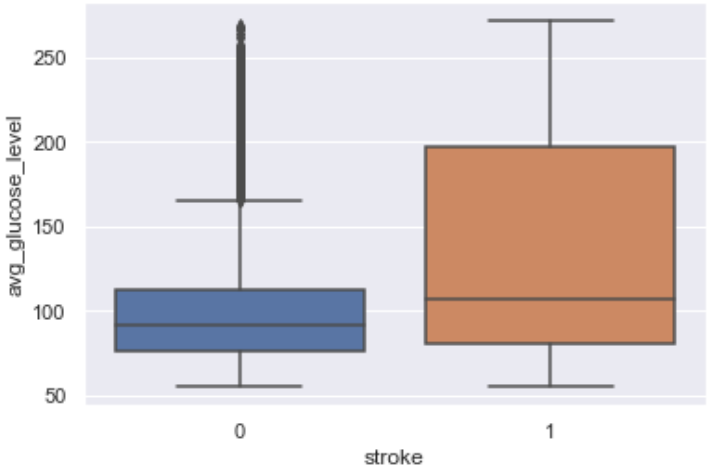
Figure 8 - Boxplot of age and stroke with hue=Residence_type



Source: The Author

Figure 8 shows that those who live in rural areas may have an earlier stroke, but the difference is little compared to those who live in urban areas.

Figure 9 - Boxplot of age and stroke with hue=avg_glucose_level



Source: The Author

Blood glucose levels contribute to having a stroke, as seen in Figure 9. From 126 mg/dL or more it is considered that the person has diabetes and almost more than half of the patients are diabetic. Therefore, we have that diabetes is a risk factor for stroke cases.

After the exploratory analysis, a classification model was made to predict whether or not a person will have a stroke, based on all the features seen in table 1. The Support vector machine model was used because it has proven effective for this type of problem.

Table 2 - Data with dummies

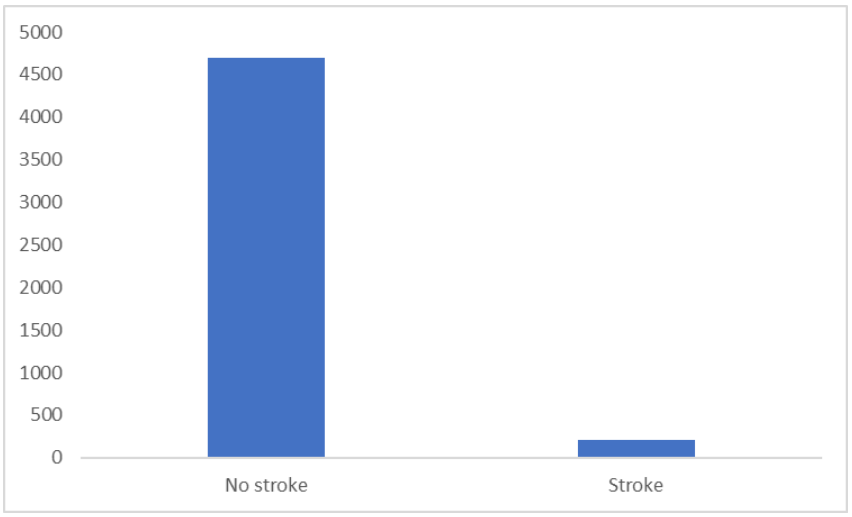
hypertension	heart_disease	avg_glucose_level	bmi	Male	Other	Yes	Never_worked	Private	Self-employed	children	Urban	formerly smoked	never smoked	smokes	stroke
0	1	228.69	36.6	1	0	1	0	1	0	0	1	1	0	0	1
0	1	105.92	32.5	1	0	1	0	1	0	0	0	0	1	0	1
0	0	171.23	34.4	0	0	1	0	1	0	0	1	0	0	1	1
1	0	174.12	24.0	0	0	1	0	0	1	0	0	0	1	0	1
0	0	186.21	29.0	1	0	1	0	1	0	0	1	1	0	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
0	0	103.08	18.6	0	0	0	0	0	0	1	0	0	0	0	0
0	0	125.20	40.0	0	0	1	0	0	1	0	1	0	1	0	0
0	0	82.99	30.6	0	0	1	0	0	1	0	0	0	1	0	0
0	0	166.29	25.6	1	0	1	0	1	0	0	0	1	0	0	0
0	0	85.28	26.2	0	0	1	0	0	0	0	1	0	0	0	0

Source: The Author

The first step was to replace categorical data, such as gender and work_type, with dummies, in order to have better results with the algorithm. The id column was then removed because they are not relevant to the analysis. In table 2 we see these changes.

The inputs and target were then separated and the dataset balancing was checked.

Figure 10 - No stroke and stroke cases

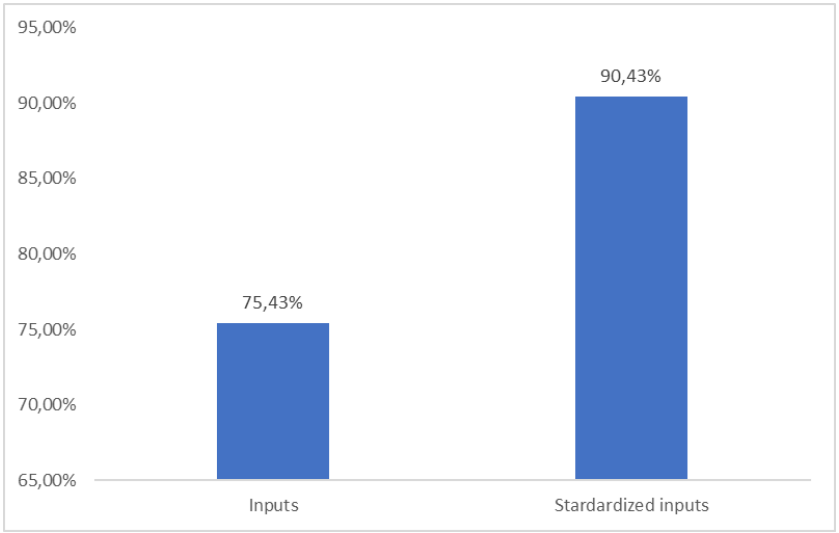


Source: The Author

From Figure 10, we see that the dataset is unbalanced and therefore can affect the performance of the model. Then, this dataset was balanced with two techniques, SMOTE and RandomUnderSampler with pipeline to get the best result.

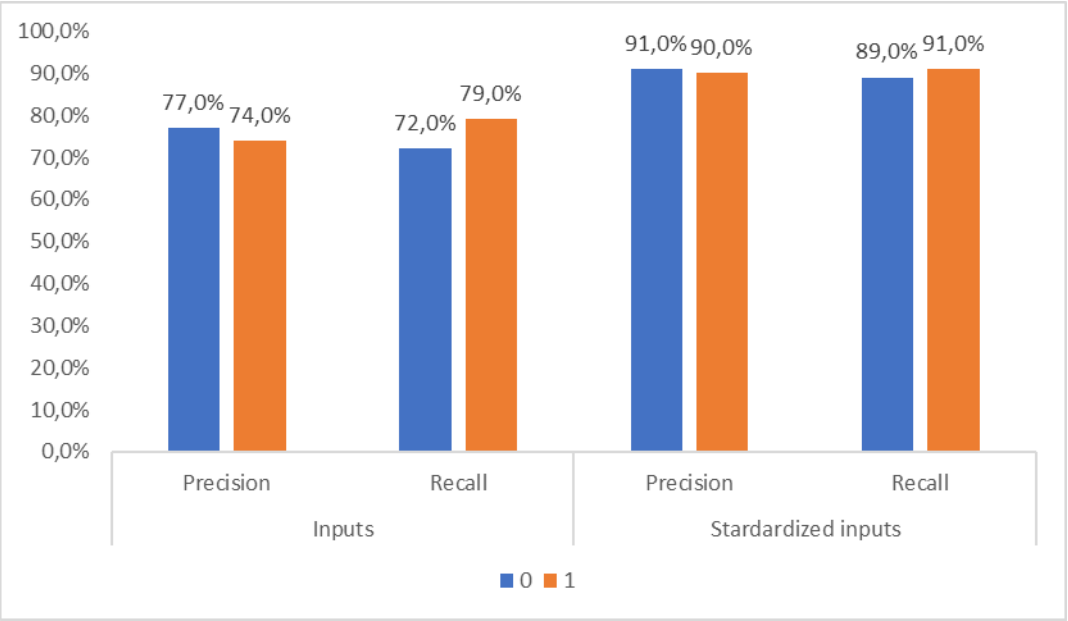
Two models were made, one with unchanged inputs and the other with normalized inputs to check for improvement in results.

Figure 11 - Accuracy Score



Source: The Author

Figure 12 - Precision and Recall



Source: The Author

As seen in Figure 11, with the normalization of inputs, we had an increase of 15 percentage points in accuracy score and also a good increase in the metrics of Precison and Recall, which are very important in a case of classification of cases of stroke, because the presence of false-negatives (recall) affects people's lives.

All the code can be found on my github:

https://github.com/williamausenka/ML_estudos_de_caso/tree/main/Stroke%20Prediction%20-%20Support%20Vector%20Machine%20(SVM)