STROKE RISK DETECTION USING SVM

Mendoza Magaña G. Author[1]
e-mail: gibrain@gmail.com

***Abstract* – Strokes are one of the leading causes of mortality in the world, detecting them before they even occur could save millions of lives, and prevent long-term disabilities. Support vector machine is an algorithm that comes handy with classification problems, it is also good when dealing with a big amount of data. On this paper, a SVM algorithm was tuned with the best hyperparameters using GridSearchCV, and a mean of scores was calculated using Cross Validation.**

***Keywords* – Average Glucose Level, BMI, Cross Validation, GridSearch, Hyperparameters, Support Vector Machine**

I. INTRODUCTION

A stroke, or bain attack, occurs when blood supply to the brain is blocked by a blocked artery or a broken vessel. Almost 87% of strokes are *ischemic strokes*, which occurs when blood clots or fatty deposits cause blockages in the blood vessels that feed the brain. Strokes can cause lasting brain damage, long-term disabilities, or even death. In the United States, every 3.5 minutes someone dies of a stroke[1][2][3].

Brain cells are very sensitive to a lack of oxygen, some of them would die within the first 5 minutes of lack of oxygen supply. Because of this, early action is important when strokes take place. Patients will have less disability after 3 months of a stroke if they are taken to the emergency room promptly; additionaly, most treatments for stroke only work if given within the first 3 hours after the incident took place[4][5].

According to the CDC, 4 in 5 strokes are preventable, which means that it is important to develop tools that can predict if someone might be at risk of suffering from a stroke, so that they can take the necessary actions to prevent one. In this paper, a simplify Data Set from McKinsey's Electronic Health Record was used in order to get more insights on the variables indicating a risk of suffering a stroke. A ML algorithm called Support Vector Machine was trained as well in order to predict which patients had a stroke[6].

II. Exploratory Data Analysis

The data set used consists of 5,110 observations of 11 different variables including patient ID, gender, age, hypertension, presence of previous heart diseases, marital status, BMI, level of glucose in blood, etc [7].

A. Age

The patients in the database range from 0 to 82 years old and are evenly represented. The mean age of this sample is of 43.23 years old. Figure 1 shows the distribution of patients by age, when it can be seen that most patients are aged between 23 and 61 years old.

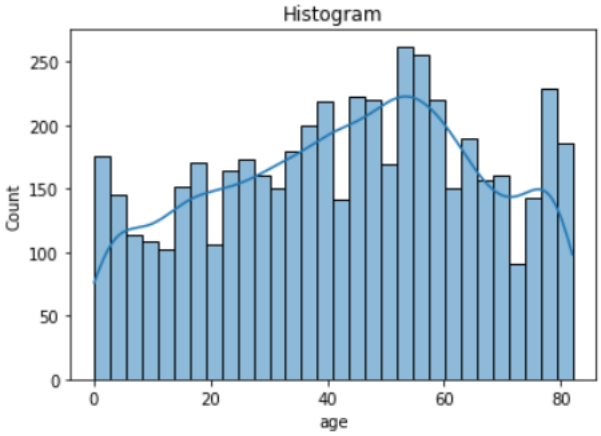

Fig. 1. Histogram of patients by age.

B. BMI

50% of the patients in the data set have a BMI between 23% and %35 body mass, the mean BMI of our data is %28.9. It can be seen on Figure 1b that there are many outliers as well (patients whose BMI is higher than %48. According to Princeton Health, a BMI between %18 and %25 is desirable; between %25 and %30, indicates overweight; between %30 and %39, indicates obesity; and %40+, indicates morbid obesity. Figure 2 shows the distribution[8]:

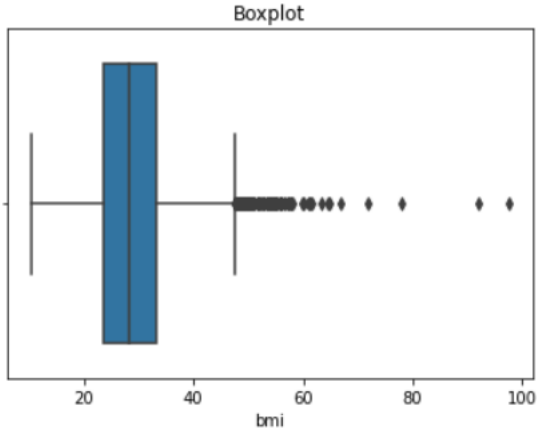

Fig. 2. Boxplot for BMI of patients in the data set.

C. Average glucose level

97% of the patients in the data set have an average glucose level ranging between 55.12mg/dl and 170mg/dl. Table 1 shows the levels of sugar in blood and what they indicate, according to the CDC[9].

TABLE I
Glucose levels in blood [mg/dl]

Fasting		
No.	**Condition**	**Glucose levels [mg/dl]**
1	Normal	< 100
2	Pre-diabetic	> 100 & < 125
3	Diabetic	> 126
2 hours after food		
No.	**Condition**	**Glucose levels [mg/dl]**
1	Normal	< 140
2	Pre-diabetic	> 140 & < 199
3	Diabetic	> 199

On Figure 3 below, we can see that 50% of the patients in our data set have a glucose level considered acceptable (depending on the type of glucose tolerance test performed on them). It is also notable that there are many patients above the normal levels of glucose on blood, even worse, many of these values are outliers (more than 170 mg/dl).

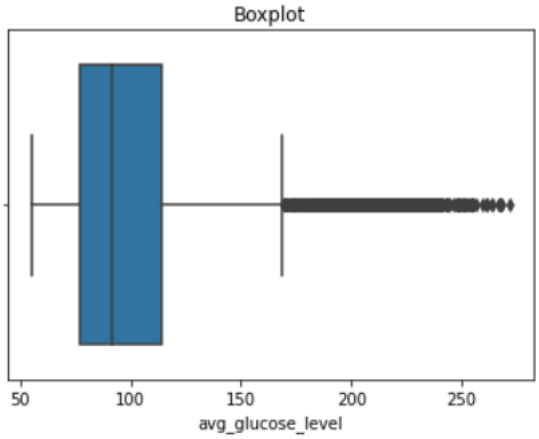

Fig. 3. Histogram of patients by age.

D. Gender

From the Patients that have not had a stroke, 2,994 patients(58.6%) are female and 2,115(41.4%) are male, only 4.9% of the women in the dataset had a stroke, while 5.4% of males did. The similarities in percentage is probably because the dataset has been previously cleaned.

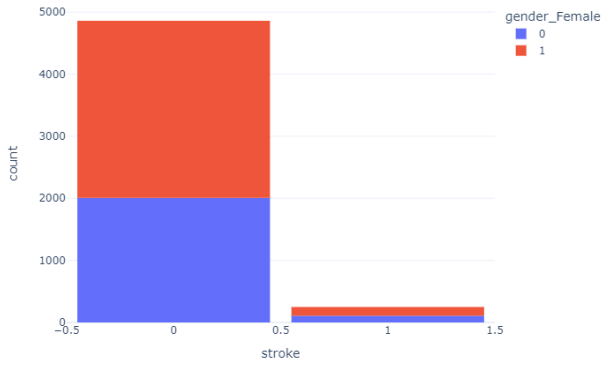

Fig. 4. Histogram of gender in dataset and stroke occurance.

E. Heart conditions

13.2% of the patients in the data that had hypertension, suffered from a stroke. This makes sense as the high blood pressure in the patients irrigation system might cause blood vessels in the brain to break. Additionally, 17% of the patients that have a heart disease reportedly had a stroke. The following figure shows the proportion of patients vs patients with a heart disease, and their stroke incidence.

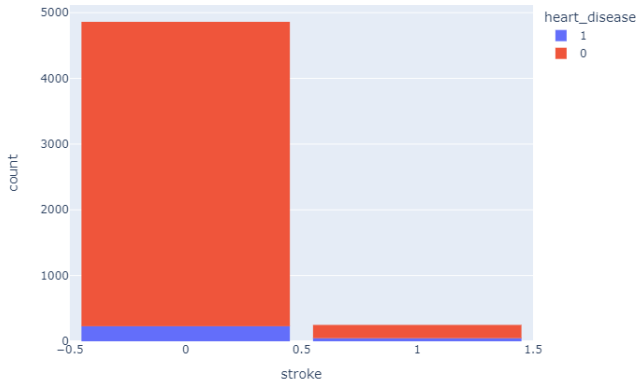

Fig. 5. Histogram of presence of heart diseases among patients in dataset and stroke occurance.

F. Smoking status

The proportion of patients that smoke or used to smoke but have never have never had a stroke and vice-versa is very high. Only 2% of patients in the dataset reported that hey had been or they still are smokers and had a stroke.

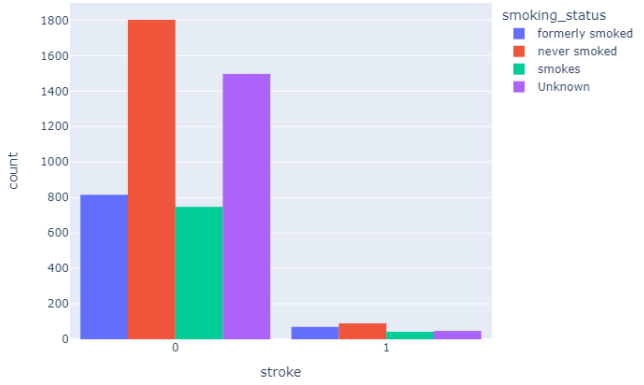

Fig. 6. Histogram of smoking status in dataset and stroke occurance.

III. Data cleaning

The data from McKinsey's EHR submitted to kaggle had previously been cleaned, which means that not much cleaning was needed. Among the data, 201 observations had no value for BMI, from which 40 belonged to patients that had a stroke. Since only 249 observations belong to patients that had a stroke, it was decided to fill the n.a values for BMI with the mean of the observations of this attribute.

Among the 5,109 observations, only 1 belonged to a

non-binary person, since this dataset does not have enough observations from non-binary people, it was decided to drop this value.

IV. Data Preprocessing

Categorical data was stored in the dataset as type object, so dummy variables were created in order to be able to process the data for modeling. Dummy variables were created among the observations of smoke status, gender, residence, and marital status. Redudant variables like "is_male" or "ever_married" were removed as their counterparts already describe gender and marital status.

After this process, the numbers that represented a category were converted to categorical variables, so that the algorithm does not treat them as a number from which one has more value than zero, but like labels where one means "not zero".

The EDA section in this paper already gives an introduction into which variables are significant for the prediction of strokes occurrence in patients. The following figure shows a pair plot of age, glucose levels, and BMI for patients that had and did not have a stroke.

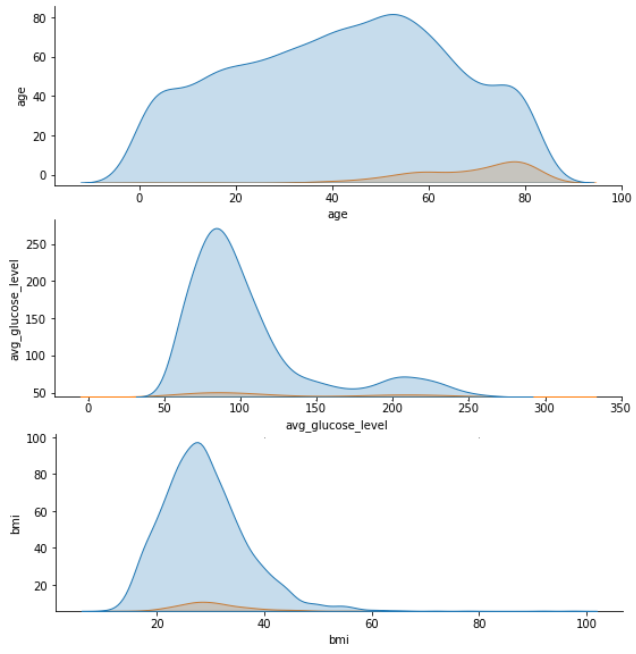

Fig. 7. Pairplot of age, glucose level and bmi for patiennts that had and had not have a storke.

It can be seen in the figure, that must of the patients that had a stroke are those among older patients, with higher glucose level, and a BMI higher than 22.

A heatmap was calculated to visualize the correlation between strokes and the other variables. Those with a correlation with our dependent variable higher than 10% were considered. The following table summarizes the variables that were considered and if they were chosen for training.

TABLE II
Variables considered for training

Variable	Correlation	Status
Marital Status	0.10	Discarted
Age	0.21	Selected
Hypertension	-0.12	Selected
Heart disease	0.13	Selected
Average glucose level	0.35	Selected
Body Mass Index	0.19	Selected

Among the considered variables, only Marital Status was rejected as it is correlated with age, choosing it would be redundant. Among all the variables selected, Age and glucose level have the highest correlations with stroke occurances. There is a relatively small correlation between hypertension and stroke occurance, which surprisingly turned out to be negative.

After choosing the aforementioned variables, standard scaler was used in order to get an evenly distributed data with mean 0 and variance of 1, this way variables with higher orders of magnitude like glucose level would not weight more than bmi or age.

As mentioned before, the data for stroke prediction is heavily inbalanced, as only 4.9% of the records belong to a patient having had a stroke. This is not enough for training any algorithm, as it could just say "no risk of stroke" and be right 95.1% of the time. In order to counter this, SMOTE was used to create artificial "stroke = 1" observations to train the SVM model. After SMOTE was used, there were 3888 observations for stroke = 0, and 3888 for stroke = 1.

V. Modeling

SVM was trained with the final data, and the best parameters were selected with the help of GridSearchCV, a python library that iterates over multiple, given, parameters, and returns the ones that provided the best scores. The parameters to iterate from where the following:

- kernels = [’linear’, ’rbf’, ’poly’, ’sigmoid’]
- c = [0.001, 0.01, 0.1, 1, 10, 100]
- gammas = ([0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6]

After ca. 9 hours of computing, the parameters returned by the function where:

- kernels = ’rbf’
- c = 100
- gammas = 0.6

The metrics obtained by training the SVM model with these parameters are shown in the following figure:

```
Accuracy Score: 0.8498
SVC f1-score  : 0.8585
SVC precision : 0.9115
SVC recall    : 0.8114

              precision    recall  f1-score   support

           0       0.79      0.90      0.84       852
           1       0.91      0.81      0.86      1092

    accuracy                           0.85      1944
   macro avg       0.85      0.86      0.85      1944
weighted avg       0.86      0.85      0.85      1944
```

Fig. 8. Classification report for SVM with best hyperparameters.

In order to make sure that the model did not overfitted the training data, cross validation was used in order to iterate over different chunks of random training and testing data. After a 5 fold cross validation, 5 r-square were obtained with values ranging from 0.83 to 0.86. The mean of the scores was 0.843.

VI. Discussion

Since the aim of this project is to develop a model that can accurately predict if a patient is on the risk of having a stroke, the best metric to measure this model's performance is "recall" of strokes = 1. The following equation explains the meaning of recall:

$$Recall = \frac{TP}{TP+FN} \tag{1}$$

where:

TP - True Positive

FN - False Negative

The max recall score is 1 and the minimun is 0, providing the maximum amount of TP and lowest amount of FN (0) makes recall =1. A True Positive in this model implies a patient successfully recognized as in danger of having a stroke when this is actually true, while a False Negative implies that a patient was wrongly told that they were out of danger, when they actually were not.

A recall of 0.81 means that from all the real positives in the data, 81% were accurately detected. If this model was to be deployed, it could save many patients lives, it could also minimize the risk of misleading false negatives, by asking patients meeting the "risk criteria" (high bmi, high sugar levels, older age) to visit their doctors regularly.

Finally, it can be seen that the parameters chosen with the help of GridSearchCV were at the upper limit from the list of parameters provided (c=100 and gamma = 0.6), a second attempt was made with higher c and gamma values; however, after 14 hours of calculations, GridSearchCV has not been able to provide the best hyperparameters.

VII. Conclusions

Support Vector Machine is a great algorithm for classification, in this paper, the use of this algorithm after cleaning, processing and picking the best variables granted great results with an 81% accurracy detecting stroke cases. I further projects, with more computing time and power at hand, this algorithm can be improved further.

VIII. References

[1]https://www.cdc.gov/stroke/types_of_stroke.htm

[2]https://www.cdc.gov/stroke/facts.htm

[3]https://www.cdc.gov/stroke/about.htm

[4]https://medlineplus.gov/ency/article/001435.htm

[5]https://www.cdc.gov/stroke/women.htm

[6]https://www.princetonhcs.org/care-services/institute-for-surgical-care/the-center-for-bariatric-surgery-and-metabolic-medicine/resources/bmi-calculator

[7]https://www.cdc.gov/diabetes/basics/getting-tested.html

[8]https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset