Stroke prediction with techniques machine learning

Magirias Georgios
Department of Mechanical Engineering
University of Thessaly
Piraeus, Greece
magirias123@gmail.com

***Summary*— A stroke is defined as the acute neurological disorder of the brain's vessels caused when the blood supply to an area of the brain stops. It is vital to develop a model that can predict the risk of an upcoming stroke using risk factors such as advanced age, high blood pressure, heart disease and body mass index. In this work, the methods of Logistic Regression, Support Vector Machines (SVM), Random Forests and the Gradient Boosting Technique are applied to data of individuals, for which we know a multitude of characteristics, while the results of Random Forests are shown to show the best results by identifying 80% of people at risk.**

Keywords*—Logistic learning, stroke, logistic regression, machines of vectors support, random forests, boosting technique, recall, precision*

I. Introduction

A stroke occurs when the blood flow to various areas of the brain is disrupted or reduced and the cells in those areas do not receive the nutrients and oxygen they need, causing them to die. A stroke is a medical emergency that requires immediate care. According to the World Health Organization (WHO), stroke is the leading cause of death and disability worldwide [1].

Strokes are divided into ischemic and hemorrhagic. An ischemic stroke develops when a blood clot blocks a blood vessel in the brain. The clot can form in a blood vessel in the brain or be carried by another vessel in the system. About 8 out of 10 strokes are ischemic and are the most common type of stroke in older adults. of age. Hemorrhagic strokes occur when an artery in the brain ruptures. This causes bleeding inside the brain or near the surface of the brain. Hemorrhagic strokes are less frequent but more lethal than ischemic strokes [2]. Stroke can be prevented through a healthy and balanced lifestyle, away from bad habits such as smoking and drinking alcohol, while maintaining a healthy body mass index and controlling average blood glucose levels also play an important role.

There are several documented research papers in the literature that use machine learning models to predict stroke. Govindarajan [3], conducted a study and attempted to categorize individuals at risk of stroke using data from 507 individuals. In his analysis he used the ANN (Artificial neural network) and SGD (Stochastic Gradient Descent) algorithms and achieved an accuracy of 95%.

Amini [4] for his research collected data from 807 healthy and unhealthy subjects. Using 50 risk factors from all these individuals, he applied the c4.5 decision tree and k-nearest neighbors (k-nn) techniques and achieved accuracies of 95% and 94% respectively.

Cheng [5] published a report on the prognosis of ischemic stroke. He used 82 samples of patients who had experienced a stroke, while for the predictions he used neural networks and achieved an accuracy of 80%.

Cheon [6] conducted a study to predict the mortality of patients who had experienced a stroke. In his study he used 15099 patients to determine the occurrence of stroke, using a deep neural network.

Our contribution to this work is as follows:

- For stroke prediction, the performance of four different models that are common in classification problems (Logistic Regression, SVM, Random Forests, Gradient Boosting) is tested.
- The minority class is oversampled, that is, synthetic samples are created which

correspond to individuals who have experienced a stroke, as the data set is unbalanced.

The rest of the paper is organized as follows. Section 2 presents the data used, their pre-processing steps and finally describes the technical classifications applied. Then in section 3 the results from the application of the classification techniques are presented, as well as the scores of their quality assessment measures. Finally, in section 4, the conclusions drawn from the results are stated.

II. METHODOLOGY

This section is divided into three parts: The description of the data set used, the presentation of the classifiers and their quality assessment measures, and finally the description of the application process.

A. Data Description

The data used is available on the Kaggle website [7], which contains many kinds of data for different projects. The data contains 5110 samples corresponding to people who have either experienced a stroke or not. The following characteristics are known for each person:

age: Numerical characteristic that informs us about the age of the person.
gender: Categorical attribute referred to gender of the person.
hypertension: Binary attribute that takes the values 0 when the person is not hypertensive, otherwise the value 1.
work_type: Categorical attribute representing the type of work of each person.
residence_type: Categorical feature that us informs whether the person lives in an urban or rural environment.
heart_disease:0 if the person does not have heart disease, 1 if they do. Binary feature.
avg_glucose_level: Represents average glucose levels in the person's blood. Numeric attribute. **bmi:**The person's body mass index. Arithmetic characteristic.

ever_married: It tells us if the person is married or not. Categorical feature.
smoking_status: It represents the state of the person with based on smoking. Categorical feature.
strokes: This characteristic is also the goal for the predictions, it gives us the information about whether the person experienced (1) or not (0) a stroke.

The data is in table format, where each row corresponds to a sample, i.e. an individual. It should also be mentioned that out of the 5110 samples, only 249 correspond to people who have experienced a stroke. Below is a portion of the data and specifically 5 samples.

gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1

Table 1: Sample datasets for stroke prediction

B. Classifiers and Valuation Measures

As part of the work, the classification techniques described below were applied:

Accounting Regression

Logistic regression come true in problem cases where the dependent variable is dichotomous (binary). The hypothesis of the model for classifying new cases is derived from equation 1, and this method uses the sigmoid function to bound the outputs between 0 and 1. Essentially, the value obtained from the hypothesis can also be expressed as the probability the sample to belong to the positive class, that is to experience a stroke.

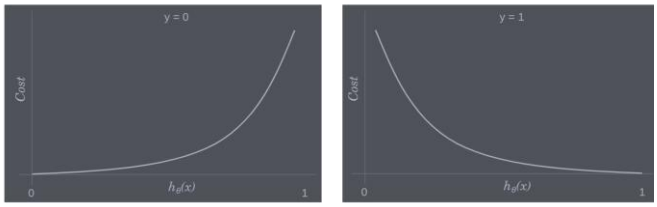
$$h(\ ) = \frac{1}{1 + \ -} = (\ = 1 | \ ; \ ), \text{ where } \quad 0 \le h(\ ) \le 1$$

Equation 1: Logistic function model hypothesis

Essentially what is being looked for are the parameter values that minimize the cost function used The cost function is shown below, where its properties are also presented.

$$(h(\ ),\ ) = - \ \log(h(\ )) - (1 - \ ) \log(1 - h(\ ))$$

Equation 2: Accounting function cost function



Picture 1: Logistic regression cost functions

As can be seen from the diagrams above, the cost function of the accounting function has the following properties.

When = 0:
= 0 if $h(\ ) = 0$ and → ∞ if $h(\ ) \to 1$

While when = 1:

= 0 if $h(\ ) = 1$ and $\rightarrow \infty$ if $h(\ ) \rightarrow 0$

Support Vector Machines (SVM)

SVMs [8] first map the training data into the space, which is as dimensional as the features used. Then a hyperplane is searched which will be able to separate the samples of the two different classes and at the same time will maximize the distance between them. This hyperplane, when only two features are used, as in figure 2, is a straight line, while in the case where 3 features are used, it is a plane. When the features are more, then it is a hyperplane of one dimension less than the number of features.

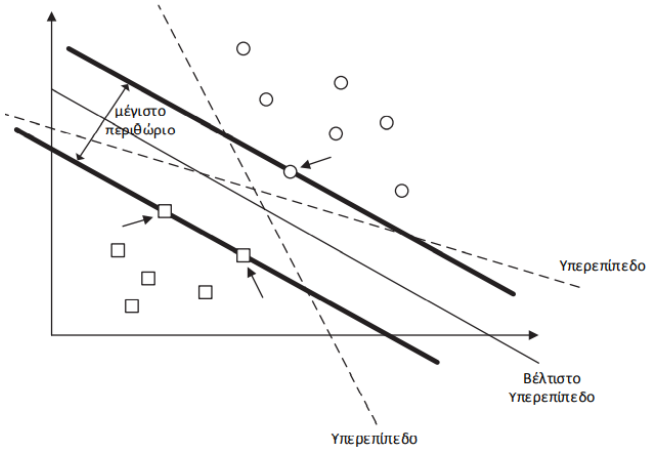


Figure 2: Super class separation layer

The resulting optimization problem for finding this hyperplane is:

$$, 2 \quad \frac{1}{-} \| \|_2 + \sum_{=1} 2$$

Equation 3: Objective functionSVM

Under the restrictions:

$$(\ + \ ) \geq 1 - \ , \quad = 1, \ldots ,$$
$$\geq 0 , \quad = 1, \ldots ,$$

After the optimums are calculated , , where vector perpendicular in the hyperplane and constant, then the new cases are sorted by the decision rule.

If $\cdot \ + \ \geq 0$ then it belongs to the class1, otherwise in class 2

Random Forests

Random Forests is an ensemble learning method [9] for classification, regression and other tasks, which works by building many decision trees during training. For classification tasks, the output of the random forest is the class selected from the majority of decision trees generated. Therefore in order to generate the random forests, the knowledge to construct the decision trees must be present.

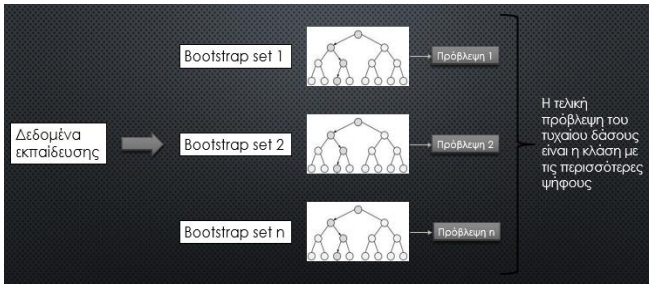


Figure 3: Random forest generation flow

As shown in figure 3, initially the bootstrap sets are created, which contain the same number of samples, but are allowed to have some samples more than once. Therefore, some samples may be duplicates, while other samples from the training data may not end up in any of the bootstrap sets. This way the trees are not related to each other.

Then, using each bootstrap set, we create a decision tree that aims to predict the dependent variable of the problem. Finally, when we need to classify a new instance for which we know its features, we give those features to each of the trees built, and the final prediction of the random forest is the class that receives the most votes from decision trees.

Gradient Boosting Classifier In machine learning, boosting techniques [10] are algorithms that aim to reduce bias as well as variance in supervised learning. The way these algorithms work is by creating a strong classifier, which is composed of a set of weak classifiers. Boosting is based on the question posed by Kearns and Valiant []: “Can a set of weak classifiers generate a single strong classifier?”. A weak classifier is defined as a classifier that can label examples little better than random guessing. In contrast, a robust classifier is a classifier that makes reasonably good predictions.

The iterative process of applying this particular algorithm is as follows: First, a tree is created

decision that predicts the dependent variable. Then the errors of the first tree are calculated, i.e. the difference between the predicted and observed value. After the errors are calculated, a new tree is created that uses the same samples, with the same characteristics, but now tries to predict the errors produced by the first tree. This process continues until the errors are minimized or until the set number of trees is built. Figure 4 shows the flow of this process.

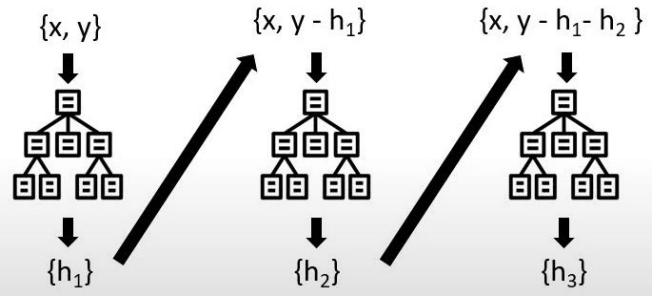


Figure 4: Process flow in amplification technique

The quality assessment measures of the classifiers used are calculated using the confusion matrix. This table gives us the following information: How many samples were correctly categorized as positive (TP), how many samples were correctly categorized as negative (TN), how many samples were incorrectly categorized as positive (FP), or incorrectly as negative (FN).

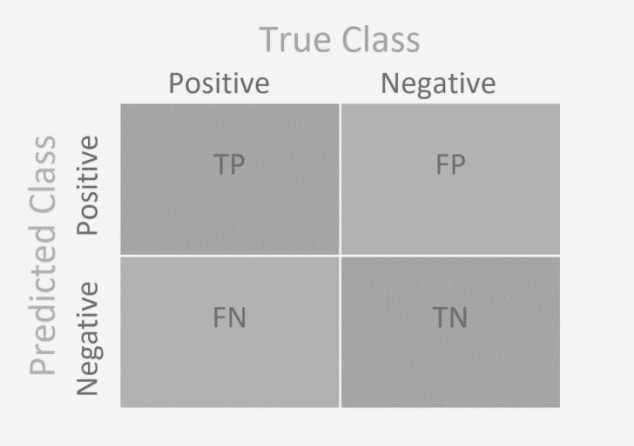


Figure 5: Table of confusion

Using the confusion table, the following are calculated:

- Precision, which is defined as the fraction of positive samples correctly predicted to the sum of positive samples correctly predicted and negative samples incorrectly predicted by the model. Accuracy is usually defined as a measure of efficiency when the objective is to limiting cases that are categorized as positive but should actually have been categorized as negative.

$$= \quad \frac{\quad}{+}$$

- Recall, which is defined as the fraction with positive samples correctly predicted as numerator and the sum of positive samples correctly predicted and positive samples incorrectly predicted as negative as denominator. Greater emphasis is placed on recall when the goal is to identify as many positive samples as possible. Recall can also be defined as the percentage of positive samples correctly predicted (TPR – True Positive Rate).

$$= \quad \frac{\quad}{+}$$

- The harmonic mean of precision and recall (F1 Score), which is used when both precision and recall of the model are equally important. The F1 Score is calculated by the formula below.

$$1 \quad = 2 * \quad \frac{+}{\quad}$$

- The percentage of negative samples incorrectly predicted as positive (FPR – False Positive Rate).

$$= \quad \frac{\quad}{+}$$

With the use of TPR and FPR, the Receiver Operating Characteristic (ROC) curve can then be designed, which is a useful tool used to analyze the performance of the classifier. This curve shows how the percentage of positive samples correctly predicted and the percentage of negative samples incorrectly predicted as positive changes as the classifier's decision threshold shifts. The decision threshold is the value of the decision function which, if exceeded, then the sample is categorized as positive, otherwise as negative. Through the ROC curve, the most appropriate decision threshold can be identified, depending on the goal of each project. The figure below shows the ROC curves for three classification systems, and it can be seen that as this curve moves towards the upper left corner it gets better. Thus by plotting the ROC for different models and calculating the area under the curve of each one, the comparison of the models can be made, as the larger the area they cover, the better the quality of the models. This area under the ROC curve is called AUC (Area Under Curve).

AUC is therefore defined as a measure of comparison of different models.

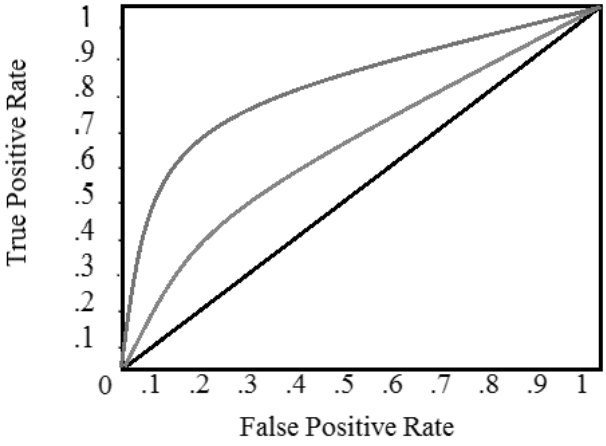


Figure 6: CurveROC

C. Application Process

The Python programming language, as well as the pandas, scikit-learn, numpy, matplotlib and seaborn libraries were used to carry out the process presented below.

1)**Input data:**First, the data presented earlier were loaded, containing 5110 samples of people who have either experienced a stroke or have not.

2)**Data Preprocessing:**For preprocessing, a check was first made for missing values and duplicate samples. Missing values were replaced with the mean value of the corresponding trait. Then the categorical features were transformed into numerical ones using LabelEncoder and the correlation matrix was calculated, which scores the features according to how useful they are for predictions.

3)**Data Separation:**70% of the samples were saved for training the models, while the remaining 30% were used for testing them. Also in this phase the numerical data were normalized, while only the training data was oversampled of the minority class. Oversampling created synthetic samples that correspond to people who have experienced a stroke, as the data set was unbalanced.

4)**Application of classification techniques:**The classifiers were trained with the training data and their scores were calculated using the test data.

5)**Model selection:**Finally, the model that showed the best results in predicting the stroke was selected.

III. ARESULTS

A. Correlation results

Pearson's correlation reveals the relationships between the dependent variable stroke and the characteristics. For each combination the table contains a correlation value, while the value range is between -1 and 1. The higher correlation values show the features that are more useful for learning the models.

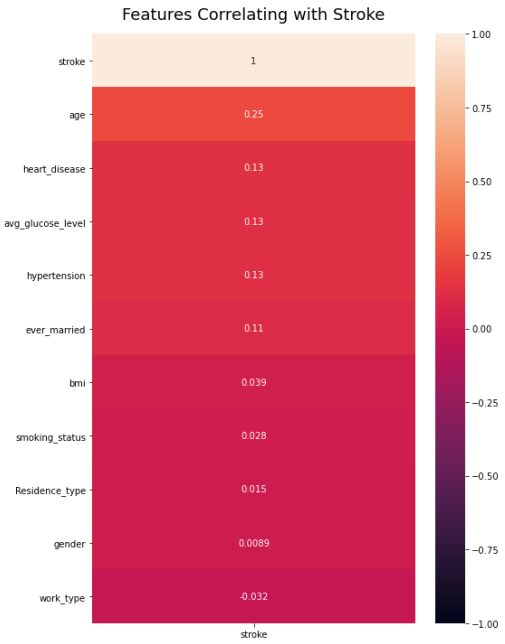


Figure 7: Correlations between dependent variable and characteristics

B. Results of experiments

In this phase the predictions made by the models when given the control data will be discussed. Three computational experiments were performed. In the first case no oversampling was performed and also the 6 most important features were used, based on the correlation values. In the second case, oversampling was performed and the 6 most important features were used again. Finally, in the third experiment, all the

characteristics of the data set and oversampling was also performed.

Experiment 1: 6 features, no oversampling

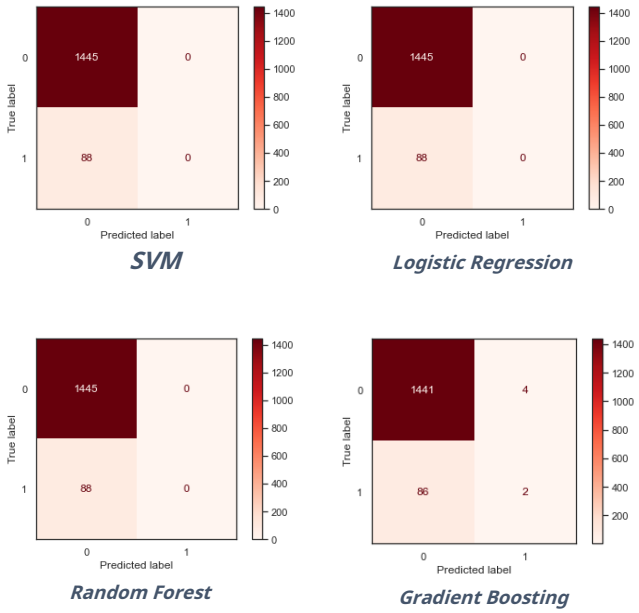


Figure 8: Experiment confounding tables1

Μοντέλα	Recall	Precision	F1-Score
Logistic R.	0	0	0
SVM	0	0	0
Random F.	0	0	0
Gradient B.	0.33	0.02	0.04

In the case where the 6 most important features are used, but at the same time no oversampling is performed, it seems that the models fail to correctly categorize the positive samples, which correspond to people who have experienced a stroke. Nevertheless, their accuracy is high, as they categorize all samples as negative and the majority of samples are indeed negative.

As can be seen in the confusion tables, the control data contains 1445 negative class samples and only 88 positive class samples. Thus, the models can achieve 95% accuracy simply by categorizing all samples as negative. But in this particular work we are mainly interested in identifying the people who are at risk.

Experiment 2: 6 features, with oversampling

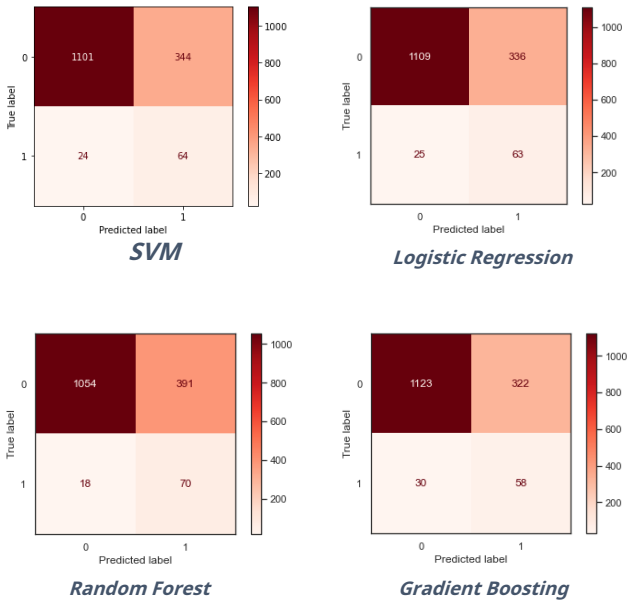
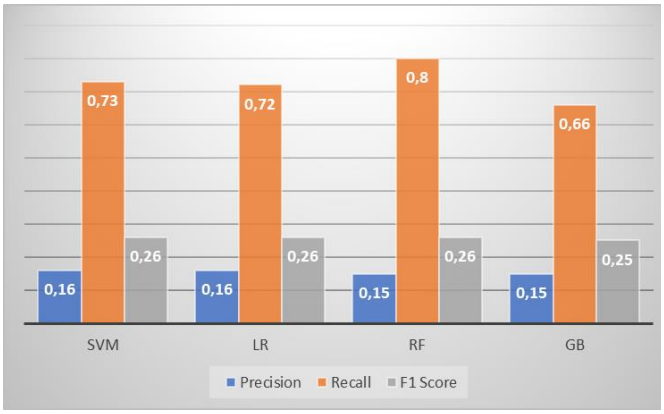


Figure 9: Experiment confounding tables2



In this experiment the results are clearly improved. The models identify several samples that correspond to people who have experienced a stroke, but also misclassify as positive a large number of samples that correspond to people who are not at risk. The best results are presented by the random forest, where it detects 80% of the positive samples (0.8 recall), while only 15% of the samples it categorizes as positive are actually positive (0.15 precision). Finally, the harmonic mean, i.e. the F1 score, reaches the value of 0.26.

Experiment 3: all features, with oversampling

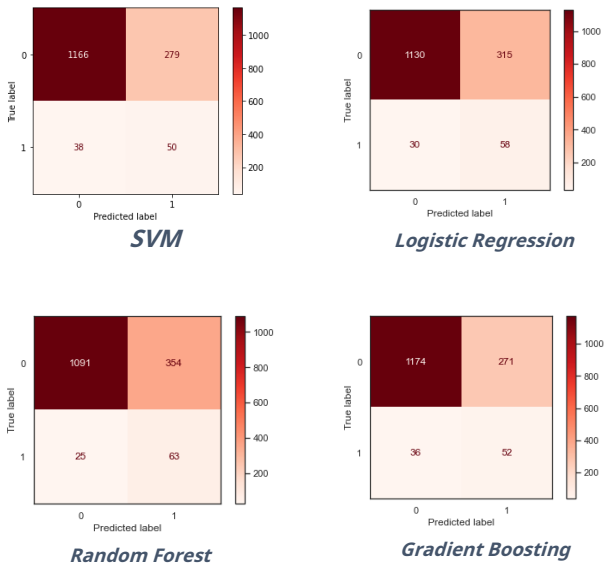
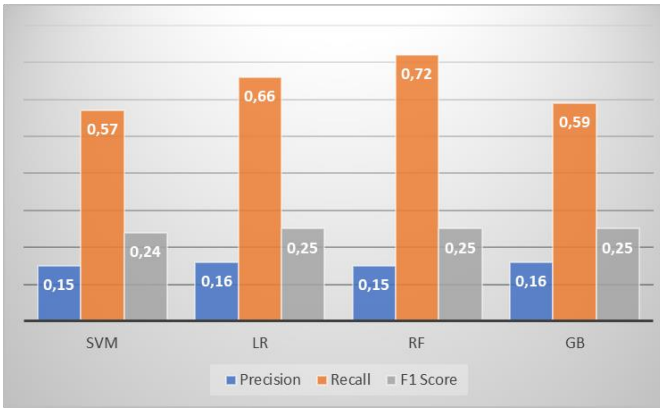


Figure 10: Experiment confounding tables3



In the third and last experiment, where oversampling is performed and all the features of the data set are used, it seems that again the models detect a sufficient number of positive samples, but a reduction is observed compared to experiment 2. Also in this case the precision is low for all classifiers, i.e. many negative samples are misclassified as positive. And in this case the random forest seems to show the best results.

Comparing the results of this work, with the results from the works mentioned in the literature, it can be seen that most of the time, importance is given to the accuracy of the model, while in reality we should be mainly interested in the recall of the models . High accuracy is not important when the models fail to detect them people at risk, which we care about the most.

For the above reason, the model that is more important for the specific purpose is the random forest that uses the 6 most important features, while oversampling has also been carried out. This particular classifier is chosen because it identifies 80% of people who have experienced a stroke. Below is presented the ROC curve of the specific classifier, while the area it covers is also calculated. The AUC score, when it gets values higher than 0.8, the model is considered quite good.

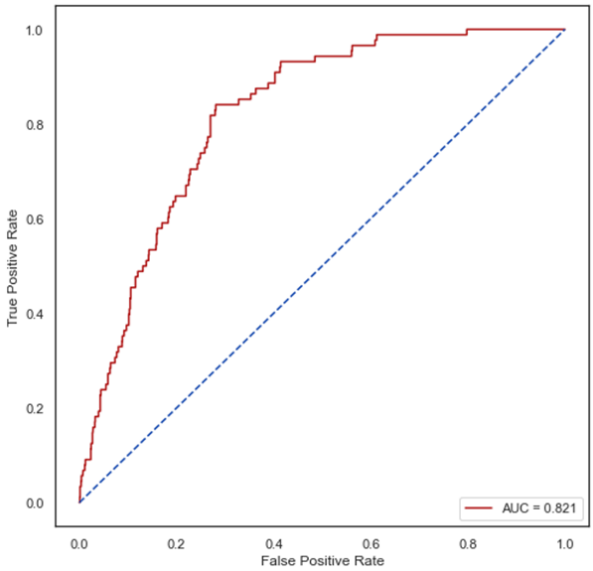


Figure 11: CurveROC, Random Forest

IV.CONCLUSIONS

From the study of the stroke data and the results of the predictions given by the classifiers, it seems that oversampling of unbalanced data is necessary, while using all the features of the data set rather than the most important ones can introduce noise on forecasts.

The best method based on the goal of this project was the random forest which was able to identify 80% of individuals at risk and misclassify many samples corresponding to individuals not at risk. The specific model can be used by doctors, so that they form an initial opinion about each patient and then act with the appropriate moves, while one cannot rely only on his prediction.

BBIBLIOGRAPHY

[1] World Health Organization Official website, http://www.emro.who.int/health-topics/stroke-cerebrovascularaccident/index.html , 2021.

[2] Thomas Truelsen, Stephen Begg, Collin Mathers, “The global burden of cerebrovascular disease,” 2006.

[3] P. Govindarajan, RK Soundarapandian, AH Gandomi, R. Patan, P. Jayaraman, and R. Manikandan, “Classification of stroke disease using machine learning algorithms,” Neural Computing and Applications, vol. 32, no. 3, pp. 817–828, Feb. 2020.

[4] L. Amini, R. Azarpazhouh, MT Farzadfar, SA Mousavi, F. Jazaieri, F. Khorvash, R. Norouzi, and N. Toghianfar, “Prediction and control of stroke by data mining,” International Journal of Preventive Medicine, vol. 4, no. Suppl 2, pp. S245–249, May 2013.

[5] C.-A. Cheng, Y.-C. Lin, and H.-W. Chiu, “Prediction of the prognosis of ischemic stroke patients after intravenous thrombolysis using artificial neural networks,” Studies in Health Technology and Informatics, vol. 202, pp. 115–118, 2014.

[6] S. Cheon, J. Kim, and J. Lim, “The Use of Deep Learning to Predict Stroke Patient Mortality,” International Journal of Environmental Research and Public Health, vol. 16, no. 11, 2019.

[7] Dataset named 'Stroke Prediction Dataset' from Kaggle: https://www.kaggle.com/fedesoriano/stroke-prediction-dataset .

[8] Vapnik, V., 1995. Support-vector networks. Machine learning, 20(3), pp. 273–297.

[9] L. Breiman, JH Freidmen, RA Olsen and CJ Stone, Classification and Regression Trees, Wadsworth, 1984.

[10] Friedman, JH (2001). Greedy function approximation: a gradient boosting machine. Annals of Statistics, 1189–1232.