Master in Data Exploitation and Knowledge Discovery
Machine Learning

1st semester 2021

Practical work number 1

Participants:

López Malizia, Álvaro
Padula, Eliana Inés
Rossi, Fabiana Alejandra

University of Buenos Aires

College of Exact and Natural Sciences
Engineering Faculty

January 5, 2024

Abstract

The present document aims to explore in-depth the decision tree algorithm used in machine learning. To this end, we worked with a sample of patients with attributes that could be related to the occurrence of a stroke.

A pre-processing of the received sample was carried out by analyzing its missing data, the relationship of the attributes with the target variable, and defining their importance. The database was split into two subsets, one for development and another for testing the model to predict a stroke (80

During the development stage, different algorithms were tested to determine the best pruning and evaluate the performance of each model. This analysis was performed with the dataset with continuous numeric variables grouped dichotomously and ungrouped. Once the performance was validated in the training set, the best model generated was evaluated on the testing subset.

The ability to predict a stroke from the data was tested. The algorithm's calculation code and graphics can be found at: https://colab.research.google.com/drive/1XE_Si8Nh1tvQKC7SYtCG3c2jOpLBLL-N.

Introduction

Strokes are one of the leading causes of disability in adults and the elderly, potentially resulting in numerous socio-economic difficulties and even leading to death in the absence of treatment [1]. In this scenario, predicting which patient is more prone to developing a lesion is of particular clinical interest [2]. In this regard, numerous *machine learning* techniques have been applied to address this disease [1-3]. The objective of this practical work was to analyze the use of algorithms for generating decision trees. To achieve this, a sample from a population of individuals with different attributes that could be representative and indicative for predicting a stroke was utilized.

Data

The database comprises 10 attributes that may be significant for predicting a stroke. The following table summarizes their characteristics. Continuous numeric variables are highlighted in orange, indicating their treatment when grouped dichotomies.

variable	tipo de dato		tratamiento VARIABLES CONTINUAS	tratamiento VARIABLES DICOTOMIZADAS	niveles (cantidad)	código
id	*int64*	*numérico discreto*	*drop*	*drop*	*no es informativa*	
gender	object	categórico nominal	drop 'Other' (1) label encoder	drop 'Other' (1) label encoder	Female(2994) Male(2115)	0 1
age	float64	numérico contínuo	sin encoding cuando son contínuas	gini split label encoder	< 68 años (4253) >= 68 años (856)	0 1
hypertension	int64	numérico discreto	label encoder	label encoder	0(4611) 1(498)	0 1
heart_disease	int64	numérico discreto	label encoder	label encoder	0(4833) 1(276)	0 1
ever_married	object	categórico nominal	label encoder	label encoder	No(1756) Yes(3353)	0 1
work_type	object	categórico nominal	dummies (cat no ordinal no binaria)	dummies (cat no ordinal no binaria)	Private(2924) Self-employed(819) children(687) Govt_job(657) Never_worked(22)	no (0) yes (1)
Residence_type	object	categórico nominal	label encoder	label encoder	Rural(2513) Urban(2596)	0 1
avg_glucose_level	float64	numérico contínuo	sin encoding cuando son contínuas	gini split label encoder	>= 162.14 (659) < 162.14(4450)	0 1
bmi	float64	numérico contínuo	imputación c media de dev (agrupado por sexo), sin encoding	imputación c media de dev (agrupado por sexo) gini split + label	< 26.1 >= 26.1	0 1
smoking_status	object	categórico nominal	dummies (cat no ordinal no binaria)	dummies (cat no ordinal no binaria)	never smoked(1892) Unknown(1544) formerly smoked(884) smokes(789)	no (0) yes (1)
stroke	int64	numérico discreto	label encoder	label encoder	0(4861) 1(249)	0 1

Methodology

Data

A model was generated based on numeric variables to favor its simplicity. Using a database with characteristics of patients who experienced a stroke or not, an initial exploration of the variables was conducted. The presence of missing data and the correlation between different attributes were determined.

Data model The practical work was carried out using two preprocessing and analysis methodologies. On the one hand, continuous numerical variables were dichotomized (grouped into 2 groups) and trees were constructed based on this simplification. On the other hand, continuous numerical attributes were used for classifications.

Preprocessing The column corresponding to the attribute *id* was eliminated as it was not informative for our analysis, and the complete observations corresponding to the only individual whose gender had a value of 'other'.

Adequacy of continuous values The continuous attributes were discretized by performing a partition based on a threshold value that maximized the impurity reduction with the gini criterion.

Missing values

According to the analysis of missing values, 3.93% null data were found in the variable *bmi*. It was decided to impute them with the mean *bmi* of the development set, grouped by sex. This percentage is not substantial to alter the distribution of the variable.

Categorical values: For non-ordinal categorical values, they were transformed into *dummies* variables, and each of these were incorporated as attributes to the dataset. For binary variables, they were assigned one or zero using *label encoder*.

Data separation: The data set was separated into Development and Held Out (or *test*). Since it is observed that in our data set the variable *target* is unbalanced, a separation strategy was carried out that considers this imbalance and proportionally distributes the *target* data between both sets. . In turn, the data set corresponding to the development category was subdivided into the sets of *train* and *validation*.

Algorithms

performance metrics were defined, the hyperparameters were evaluated, the model was simplified and the most important attributes were defined with the aim of the model efficiently and simply predicting a stroke:

Choice of metrics and *performance* Regarding the selection of metrics used, it was decided to give greater importance to *recall* than to *precision*. Given the objective of predicting a stroke, it is more important to detect all positive cases, even if there are false positives. We consider stroke prevention important. With this same logic, we consider a beta = 2 in the F-beta score, giving greater weight to *recall*.

Selection *a priori* of the hyperparameters The hyperparameters selected *a priori* to develop the tree model presented were the following: Gini impurity, 5 minimum values in each sheet, 15 levels deep at most and the model was required to balance the weights associated with the classes of the variables.

Evaluation of the defined hyperparameters The model with the defined hyperparameters *a priori* was evaluated in the development set through 50 different seeds and through cross-validation (*50k-fold cross validation*)

Tree pruning Through the maximum cost complexity algorithm provided by the sklearn library, a range of alpha hyperparameter values was extracted to evaluate a tree with less depth. Each provided alpha was evaluated using the *10k-fold* technique. The alpha whose tree best predicted the stroke with the least depth was selected, under the *performance* of F-beta in order not to overfit the model.

Attribute selection The recursive elimination technique was used to determine which variables were of most importance for the construction of the model. From the three main attributes, a new decision tree was built.

Results

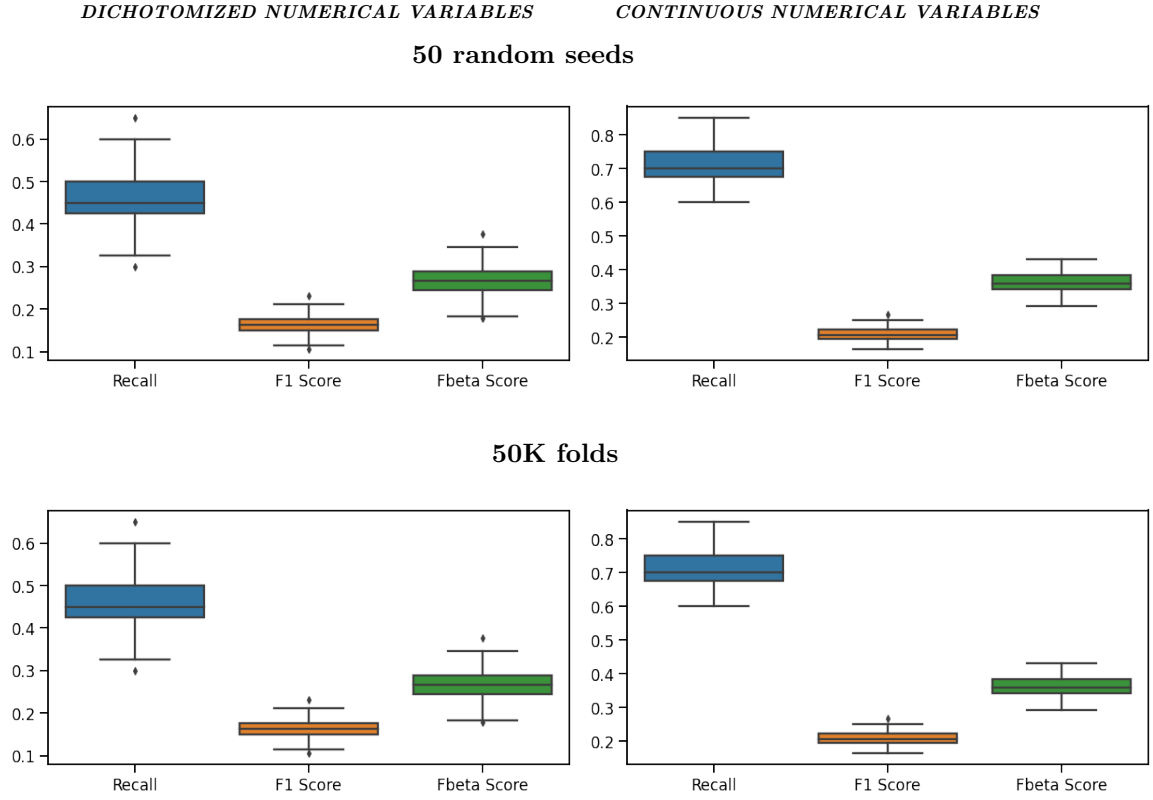
Correlations with *stroke* To evaluate the variables in their relationship with *stroke*, contingency tables represented in color maps were created, given that the database presented categorical and numerical variables . The relationships with the target variable are represented according to the number of observations in each class.

Indicators of greater susceptibility According to the gini criterion, the variables that would provide the most information are: *age*, *avg glucose level*, *heart disease*, *hypertension*.

Balance of target variable The target variable (*stroke* or ACV) is unbalanced, given that only 4.87% of the data were positive. Consequently, the *stratify* parameter was used when creating the Held Out and development

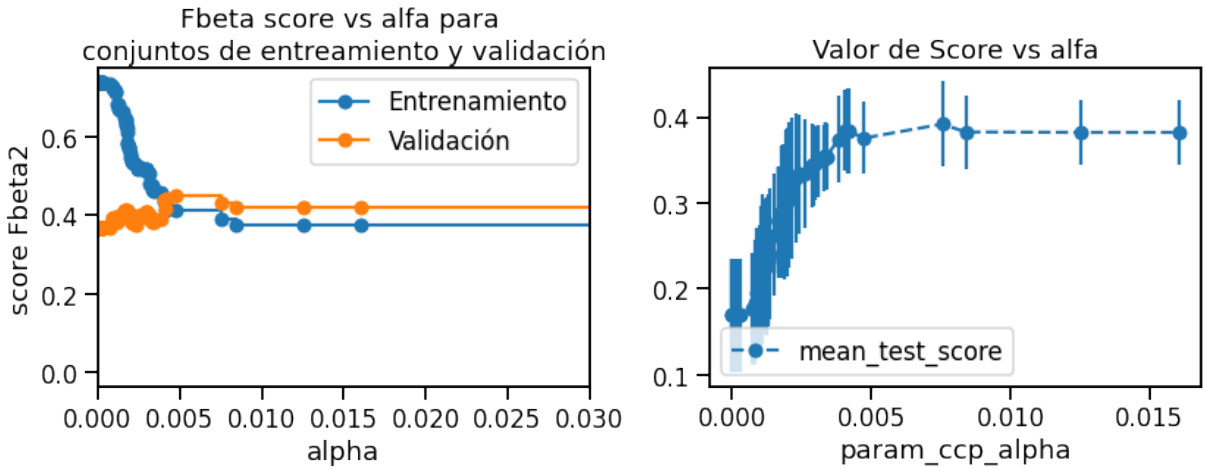
subsets, in order to guarantee the proportion of positive data in each case. On the other hand, the parameter *class weight* was used to compensate for the imbalance in the implementation of the models.

Model validation with hyperparameters *a priori*



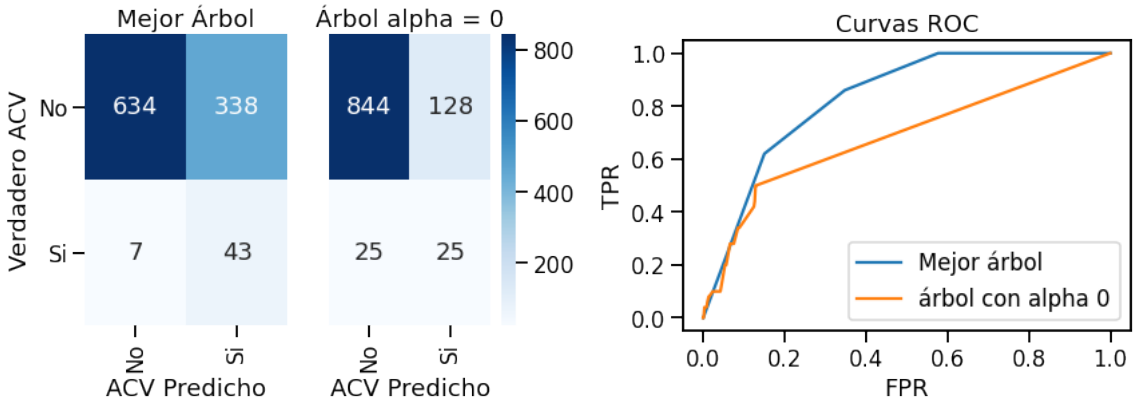
The hyperparameters defined *a priori* were evaluated with the data with "dichotomized" or continuous variables. Our results indicated that the *performance* metrics in the *50k fold* and 50-seed models for the variables in their binary form were lower than those of the models that used the continuous variables

Tree Pruning The results of re-training a tree iterating over different values of minimum complexity alphas using *10k folds* show that *performance* increases in the validation subset as the alpha increases.



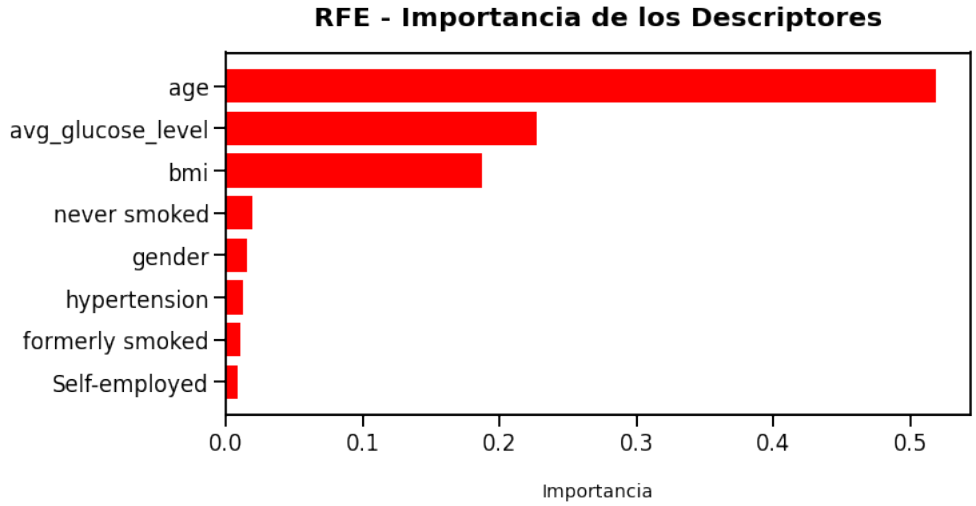
In this case, setting alpha = 0.0075 maximizes the Fbeta measure of the test.

Evaluation on test set - Pruned tree *versus* without pruning The pruned and unpruned tree were evaluated in the test data set. The results indicated that pruning improved the *performance* of the model in the *scoring* used. *score* values of 0.37 and 0.35 were obtained for the tree with and without pruning, respectively.



Regarding the confusion matrix, it indicated that the pruned model improves the number of positive cases as the *performance* objective has been defined. In other words, the ROC curve presents a higher ratio of positive cases.

Attribute selection through recursive elimination For the unpruned tree, the recursive elimination technique was used which indicated that only 7 of the 14 variables analyzed were significant for modeling.



It was found that age, average blood glucose level and *bmi* were the most important variables compared to those selected. The decision tree was retrained using these variables, with alpha=0 and iterating over different depth values. A higher F-beta2 value of 0.41 was obtained.

Conclusion

The comparison between the types of processing (continuous or "dichotomized continuous" variables) showed that transforming continuous numerical variables into binary produces a reduction in the predictive capacity of the generated trees. This result is attributable to the loss of information as a consequence of the grouping of the values. The selection of the most representative variables, as well as the pruning of the tree, impact the quality of the model to be developed. The prediction of having a stroke improves when only the most representative variables are used, and a tree with a lower depth is executed. These considerations will surely allow us to design more flexible models with greater capacity for making conjectures. It is concluded that it is important to validate our models in the selected hyperparameters and analyze which variables are representative of the analysis. Finally, given that the target variable was unbalanced (with few positive cases), despite the methodologies applied, the model would only reach an Fbeta *score* of 0.41.

Bibliography

[1] Yu, J.; Park, S.; Kwon, S.-H.; Ho, C.M.B.; Pyo, C.-S.; Lee, H. AI-Based Stroke Disease Prediction System Using Real-Time Electromyography Signals. Appl. Sci. 2020, 10, 6791. https://doi.org/10.3390/app10196791

[2] Truelsen T, Beggs S, Mathers CD (2006) The global burden of cerebrovascular disease. Geneva, Switzerland: WHO. https://www.who.int/healthinfo/statistics/bod_cerebrovasculardiseasestroke.pdf

[3] Amini L, Azarpazhouh R, Farzadfar MT, et al. Prediction and control of stroke by data mining. Int J Prev Med. 2013;4(Suppl 2):S245-S249.–

[4] Miroslav Kubat,An Introduction to Machine Learning. Second Edition 2007. Capítulo 6

[5] StatQuest with Josh Starmer,Decision Trees in Python from Start to Finish https://www.youtube.com/watch?v=q90UDEgYqeI&t=270s