Sleep quality analysis

Arturo Laflor

2017-05-10

Contents

Chapter 1

Prerequisites

Chapter 2

Introduction

Chapter 3

Data adquisition

Chapter 4

Data pre-process

Chapter 5

Feature selection

Chapter 6

Evaluation of Efficiency

Before testing the selected factors, models were trained using the 21 sleep hygiene factors to know the predictive efficiency that these models from various techniques of machine learning could achieve. The result was that both, the support vector machines (SVM) with linear kernel and logistic regression, were the two techniques with the best results. The SVM algorithm had an efficiency of 67% and the logistic regression reached an efficiency of 70%. With this background, the tests described below were made, taking into account only the four selected factors. If any of the techniques reaches an efficiency equal to or higher than the previous results, the selection of variables can be considered a successful process and these factors will be used for the prediction model of the study hereafter.

One of the steps in the development of the investigation project, includes the selection of a technique to train a predictive model on supervised automated learning. We did a review of the literature and we select three techniques under certain criterion based in the nature of the problem. The purpose is train the model with the available data and select the one given the best prediction. So, at the same time that the evaluation of efficience of the selected factors was performed, the selection of the technique that will be used for the final training was done. The three techniques that meet the inclusion criteria, were: artificial neural networks, vector supported machines And logistic regression with regularization. As in feature selection, a Shiny application was developed to process the data and compare the outcomes for these three algorithms, training a model with total of the records and only the four features selected in the feature selection process as was explained in Section 5.

The evaluation was performed by the cross validation technique using an iteration process of training, validation, analysis and refinement as the figure 6.1 shows. In this process a sixty percent of the data was used to train the model, when training conclude, the cross validation is performed through the prediction of the target variable in the cross validation set, containing a twenty percent of the main dataset. The analysis is done at that time and depending on the results, the parameters are adjusted to make a new iteration or reach the stop point. If the stop point was reached, the model is proved in the test set to obtain the final efficience of the model.

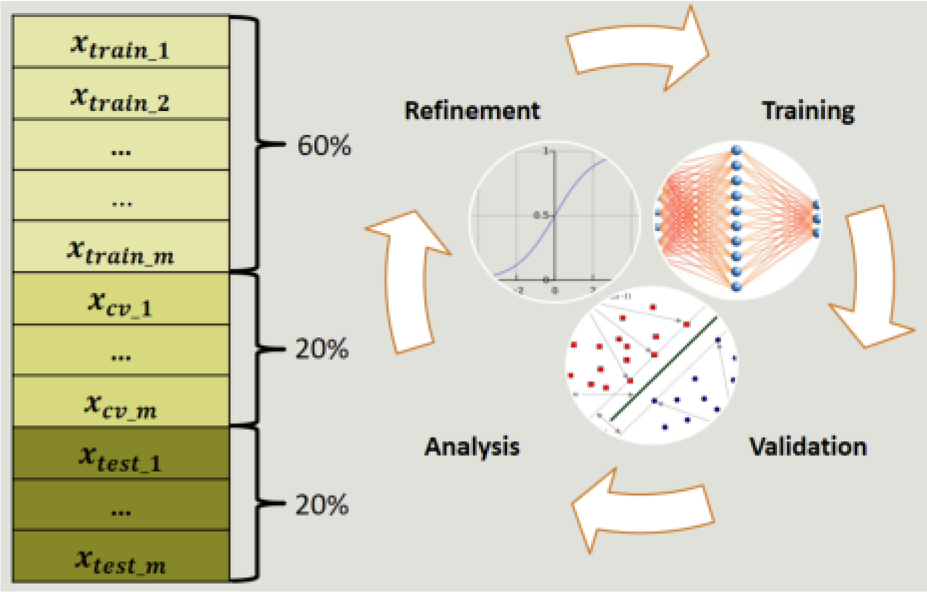


Figure 6.1: Cross Validation Process

6.1 Neural Networks Results

Two neural networks were trained and validated by cross validation process, both estructures with a hidden layer. The first neural network had three neurons in the hidden layer and the second four neurons. The Fig. 6.2 shows the structure of the neural network with four neurons in the input layer, one neuron for each factor selected in the feature selection process. The second layer is the hidden layer with four neurons and the last layer contains one neuron for the result (good sleep quality/bad sleep quality). Additionaly it is possible to observe the two activation neurons in the top of the figure.

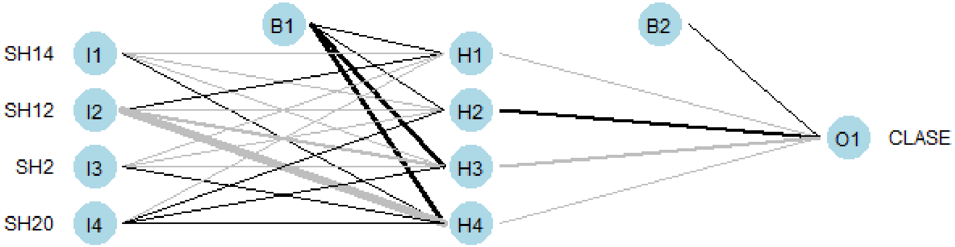
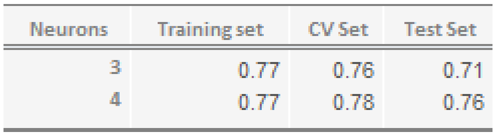


Figure 6.2: Structure of neural network with four neurons in the hidden layer

The results for the two neural networks and the appropriate comparison between them, are in the Fig. 6.3. The network with better efficiency of two networks is the network with four neurons. The table describes that in the three sets, the behavior was superior in terms of efficiency, while the plot represents the error per each set with three and four neurons. Clearly, the lines decrease in favor of the training and validation with four neurons, where the error of the prediction is smaller.



Neurons	Training set	CV Set	Test Set
3	0.77	0.76	0.71
4	0.77	0.78	0.76

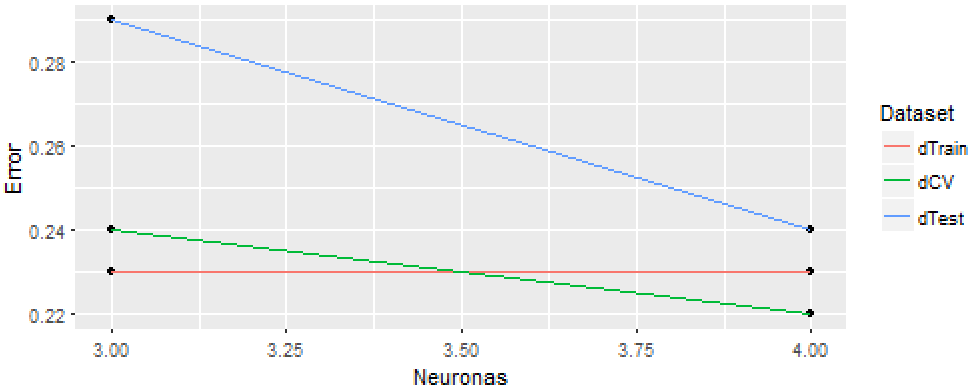


Figure 6.3: Comparison of the results for the two trained neural networks

The results of the neural network, satisfy the conditions sought, because although it is not greatly improved in efficiency when compared to what can be obtained by employing all the factors of sleep hygiene, we gain in the amount of factors that must be sensed to obtain input data. This fact has great relevance for the project because it greatly limits the design and infrastructure of the data acquisition module.

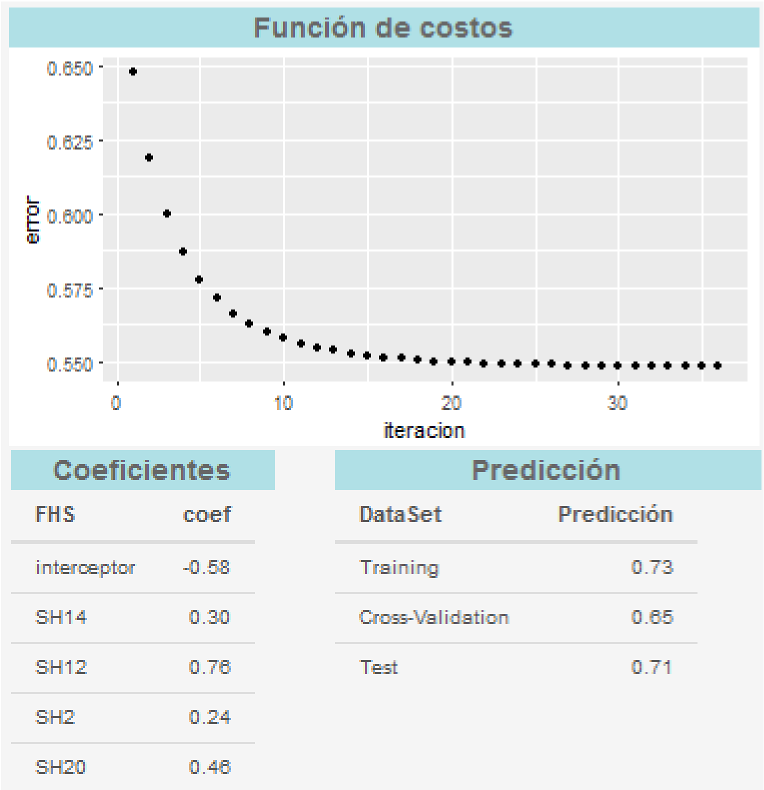
6.2 Logistic Regression Results

We train the model through logistic regression (LR) with regularization parameter and polynomials of degree one, two and three, in order to look for the optimal point between over fit and bias. The regularization parameter based on the norm l_2 takes the form of the equation (6.1), where λ took values from 0.1 to 0.6 with intervals of 0.03 to choose the optimal value.

$$reg = \frac{\lambda}{2m} \sum_{j=2}^{n} \theta_j^2 \tag{6.1}$$

The stop condition for the adjustment of the parameters of the regression is of the order of one hundred thousandths, that is to say, while the previous and the current cost function did not have a difference of 0.00003 between both, the regression continued to iterate.

The results of LR's are shown by the application in the format of the Fig. 6.4. This figure shows the original results for the LR with the polynomial of degree one, we obtained five coeficients including the intercept coeficient, the right table have the data of prediction, 70% of efficiency for the training set, 76% for the cross validation set and 69% in the test set. The plot in the top of figure, shows the behavior of the cost funtion through the iterations in the compute and refinement of the parameters.



Coeficientes

FHS	coef
interceptor	-0.58
SH14	0.30
SH12	0.76
SH2	0.24
SH20	0.46

Predicción

DataSet	Predicción
Training	0.73
Cross-Validation	0.65
Test	0.71

Figure 6.4: Results of the LR and polynomial grade one

For polynomials of degree two and three we have a similar figure, the difference is the number of coeficients that in the case of the polynomial of degree two are 16 and, in the polynomial of grade three are 35, including in both cases *dummy factors*. In the case of polynomial of degree two the cost function iterated 150 times and the predictions were 72% of efficiency for the training set, 78% for the cross validation set and 69% for the test set. The LR with the polynomial of degree three, did 446 iterations, having a precision in the prediction of 75% for the training set, 70% for the cross validation set, falling to 62% for the test set.

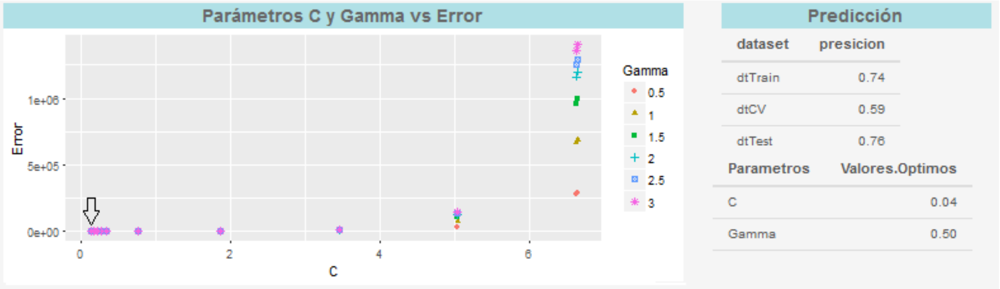
Table 6.1: Comparison of efficiency of LR with polynomials of degree one, two and three

	degree 1	degree 2	degree 3
training set	70 %	72 %	75 %
cross validation set	76 %	78 %	70 %
test set	69 %	69 %	62 %

Comparing the three results in the table 6.1, we conclude that the polynomial of degree one is the best choice for this study, because, is the algorithm that consumes the lower resources of the processor and memory and have similar predictions than the other two models of degree two and three. Results, also are satisfactory if they are compared with the results using the 21 input data.

6.3 Support Vector Machine Results

As in the previous algorithms, for support vector machines algorithm, a cross valitation test was performed. In this case, were used four kernels, two lineal kernels with polynomials of degree one and two, one radial kernel and one sigmoide kernel. The Fig. 6.5 shows the results as they are presnted in the Shinny application, we can see in the left panel, the plot showing diferents values of C and Gamma parameters and how is the behavior of the error depending of these two parameters. In the right side we observe that the best values for C is 0.04 and the best value for Gamma is 0.5 to reach the best prediction for this kernel, 76% of prediction in the test set.



Predicción

dataset	presicion
dtTrain	0.74
dtCV	0.59
dtTest	0.76

Parametros	Valores.Optimos
C	0.04
Gamma	0.50

Figure 6.5: Results of SVM with sigmoide kernel

The Table 6.2 show a comparative framework of results of the four kernels that were tested. We can observe that the sigmoid and radial are the best evaluated with a slight advantage of 3 percentage points of the sigmoid over the radial. The linear kernel is not a bad choice if one thinks in terms of simplicity to program it and the little memory and processor that consumes.

Table 6.2: Results of officiency in prediction with SVM algorithm

	Lineal degree one	Lineal degree two	Radial	Sigmoide
Training dataset	72 %	69 %	80 %	74 %
Cross Validation dataset	70 %	78 %	69 %	59 %
Test dataset	71 %	67 %	73 %	76 %
Parameter C	0.04	0.14	0.40	0.04
Parameter Gamma	0.5	0.5	0.5	0.5

The Table 6.2, also presents the best C and Gamma parameters that the cross validation process selected for these algorithm with different kernels, the parameter Gamma maintains its value in each one of the two kernels that is required (radial and sigmoide), 0.5 is the best value among the six values tested. On the other hand, the parameter C, shows that small values are more appropriate than big values. For C parameter, the best value is 0.40 for Radial kernel and 0.04 for sigmoide kernel. C was chosen in a range of 0.01 to 1000. It means that in both cases the algorithm selected a wide margin classifier.

6.4 Comparing the results of the three algorithms and their variants

All models that were trained with four factors exceed the precision of the prediction to the models that were trained with the 21 factors considered in the applied survey (see Table 6.3). The mean of the prediction in the models trained by the 21 factors is $\mu = 63.33$ with a standard deviation of $\sigma = 2.45$, less than in the models trained with the four factors selected by the algorithms described in Section 5 is $\mu = 70.33$ and standard deviation of $\sigma = 2.64$.

After obtaining these results, we decided to use only the four factors selected to train the model for the estimation of sleep quality. The next step is to choose the algorithm to be used for model generation. The metrics that will be used to choose the algorithm will be, precision of prediction, computational cost, implementation complexity in a mobile device, and the flexibility of scaling in The time required. In the table 6.3) we can see that the ANN of four neurons and the SVM with radial kernel are the best algorithms

Table 6.3: Results of all algorithms tested

Algorithm	Variant	Features	Precision	Time (sec)	Features	Precision	Time (sec)
ANN	3 Neurons	21	64%	0.14	4	71%	16.40
	4 Neurons	21	64%	0.15	4	76%	24.32
LR	Linear, dg 1	21	66%	0.25	4	71%	0.59
	Linear, dg 2	21	68%	0.30	4	70%	0.78
	Linear, dg 3	21	60%	0.70	4	64%	4.44
SVM	Linear, dg 1	21	62%	132.53	4	71%	11.90
	Linear, dg 2	21	62%	147.21	4	69%	12.46
	Radial	21	62%	20.47	4	73%	8.11
	Sigmoid	21	62%	15.94	4	68%	7.48

in prediction, however, the execution time are also of the highest. In terms of implementation, the simplest is the LR and we can see that the LR-trained model is below the SVM only two percentage points, and five percentage points below the ANN Of four neurons in the hidden layer. This allows us to have a preliminary idea of what should be done, however it is necessary to do more tests to arrive at more solid conclusions.

Chapter 7

Applications

Chapter 8

Placeholder

Chapter 9

Final Words

Bibliography