Sleep quality analysis

Arturo Laflor

2017-05-10

Contents

Chapter 1

Description of the work

This work describes the progress that has been made so far in estimating sleep quality, based on sleep hygiene factors. The characterization of the phenomenon is being carried out in two stages, the first, described in this work, has to do with the generation of a model that estimates the quality of sleep and that has been trained with subjective data from a questionnaire applied to 341 volunteers . The second stage is in process, this stage includes a logbook that records the factors of a person's sleep hygiene every day and the measurement of quality is not subjective, but is measured by an electronic device. In this way, in the end, an analysis of both characterizations will be made and a model will be constructed that takes into account the two perspectives to obtain a better approximation in their predictions.

Chapter 2

Introduction

The monitoring of sleep and the estimation of sleep quality has become relevant in recent years due to research that has positively correlated poor sleep quality with various problems ranging from mild to serious. Minor and short-term problems include lack of concentration, difficulty remembering and learning new things and irritability, while more serious problems talk about hypertension, type II diabetes, Alzheimer's and other chronic degenerative diseases.

The intention of this work is to construct a model that infer the quality of sleep of a person from the factors of the hygiene of the sleep. The literature discusses at least 21 factors of sleep hygiene that should be considered when studying the possible causes of a sleep disorder, physicians conduct interviews and apply questionnaires to their patients to investigate behaviors that favor or impair sleep quality, and from there they begin to generate a clinical file that allows them to diagnose or move to a second stage of the protocol that consists of laboratory studies such as polysomnography or actigraphy.

This work is aimed at the prevention of poor sleep quality and the possible prevention of a sleep disorder. It is a question of constructing a model that estimates the quality of sleep from the most relevant factors of sleep hygiene for a given population. The model will acquire data of the hygiene of the dream by means of sensors coupled to the surroundings of the user and will make an estimation of the quality of the dream, later the user will give a qualification to his quality of dream of such form that day to day the model is Adjustment to the lifestyle and sleep patterns of the person and the person make changes to that lifestyle, following model recommendations in such a way that their quality of sleep is favored.

The first part in the construction of the model consisted of collecting information through clinically proven questionnaires on the perception of sleep quality and sleep hygiene in a sample of people as Section 3 explains. With these data a study of the most relevant factors was made so that of the 21 original sleep hygiene factors that were taken from the literature, only the most relevant set was selected (see Section 5). This will make it possible to implement the model in a real system because a reduced number of factors implies few sensors in the context of the user, which translates into less design complexity, less intrusiveness and less infrastructure cost.

After selecting the most relevant factors for the estimation of sleep quality, we proceeded to test the model using the cross-validation technique. This was done with two purposes. The first one was to validate that the selection of factors was successful in verifying that with the subset of variables the model predicts with more efficiency the quality of sleep than when using the total factors. Second, this work was useful to obtain the first performance analysis of three supervised automated learning techniques to generate predictive models. The techniques tested were artificial neural networks (ANNs), logistic regression with regularization (LR) and vector supported machines (SVM). The three techniques had an efficient behavior as can be seen in the Section 6 and are candidates for any of them being chosen to be the basis of the model.

Chapter 3

Data adquisition

3.1 Questionnaire

In order to obtain data that contribute with evidence, regarding of relations between Sleep Hygiene Factors (FSH) and the Quality of Sleep (QS), we selected two questionnaires clinically used. The Sleep Hygiene Index (SHI) and the Pitsburgh Sleep Quality Index (PSQI). As the population where the new questionnaire would be applied is Spanish-speaking and original questionnaires are in English language, we proceed to do the process of translation. The valid process to obtain a reliable translation consist of following stages: A person `A` translate the questionnaires from English to Spanish, a person `B` getback the spanish translation to the English language, and, a person `C`, compare the questionnaire obtained by the translation of the `B` person agains the original questionnaire. The `C` person, writes comments regarding of those itemes that do not match in meaning, corrections are done, and the process iterate until reach a satisfactory result (see Fig. 3.1).

After this translation, the two questionnaires where joined in a single questionnaire, adding a section in order to obtain demographic, emoptional and health data of relevance to this study. The new questionnaire was comformed by three sections. The first section has six demographic items, one emotional and one health item, the second section is the PSQI questionnaire that consists of 20 items, and, the third section is the SHI whit a total of 21 items. In the end, the SHI survey was left with 21 items, unlike the original that has 13, this has been, only for data granularity reasons, however, the changes do not alter the SHI objective. For example, item six of the original questionnaire that asked about the use of tobacco, alcohol, or caffeine, became three items to ask separately about the use of these three substances.

In the end, the questionnaire consisted of 47 items, divided into three sections that are described later. The purpose of the questionnaire is to collect data to be analyzed using automated learning techniques

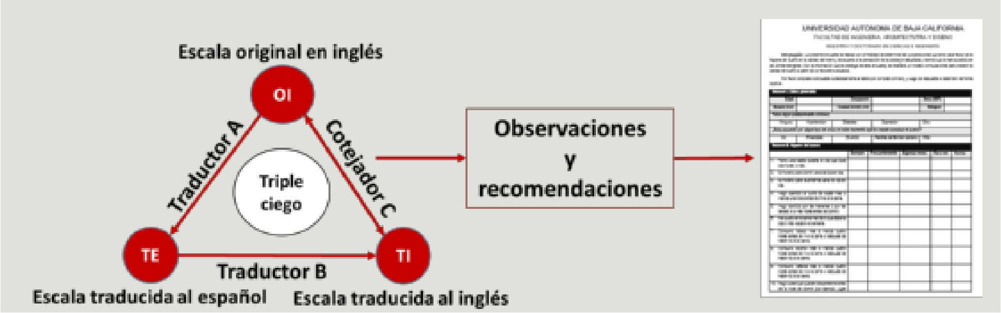

Escala original en inglés
OI
Traductor A
Cotejador C
Triple ciego
TE
Traductor B
TI
Escala traducida al español
Escala traducida al inglés
Observaciones y recomendaciones

Figure 3.1: Double translation

(specifically the techniques of feature selection) to determine those sleep hygiene factors that have a greater impact on the quality of sleep from the respondent's perception. One of the specific objectives of this research is to delimit the domain of input data to a subset of factors that explain an appropriate percentage of the variance of the phenomenon. The resulting factors will be used as the predictive variables in the first stage of training of the inference model to estimate the quality of sleep.

3.1.1 Demographic emotional and health data

3.1.1.1 Demographic

In this section we ask about six relevant data that allow to understand the context of respondent. The six variables are age, gender, ocupation, kind of work, religion and civil status. Of this six variables, **kind of work** provide relevant information to the study, since the literature says that phisical activity improve the sleep quality [cita] . The options for this question are: *intellectual*, *phisical*, *more intellectual than phisical* and, *more phisical than intellectual*. The other variables in this section are for exploration purposes regarding of sleep quality and sleep hygiene.

3.1.1.2 Emotional

This variable asks to the respondent if he/she are in a crisis time. The crisis can be financial, mourning, divorce, or other that can significantly alter the quality of sleep. This answer is a target population filter for this study. Data from people in this circumstances are noisy to the study and should not be part of the data that will be used for the analysis.

3.1.1.3 Health

Similar than *Emotional* variable, the health variable asks if the respondent suffers from a chronic degenerative disease such as diabetes, hypertension, depression or another that can directly or indirectly alter the quality of sleep. Data from people suffering some disease are removed before the analysis.

3.1.2 Quality of Sleep

The PSQI is considered the gold standard questionnaire to evaluate subjective sleep quality (Brick et al., 2010), and has been used to estimate the quality of sleep in clinical and nonclinical population (Mastin et al., 2006), and has been referred by numerous researchs in diverse sleep assessments (Bai et al., 2012). This questionnaire evaluates the quality of sleep using nineteen items grouped in seven components: subjective sleep quality, sleep duration, sleep latency, sleep disturbances, use of sleep medication, day time disfunction and sleep latency. The questionnaire provide a baremo to score each component and sumarize the final score resulting in a dicotomic varaible; *-good quality of sleep-* or *-poor quality of sleep-* (Buysse et al., 1989). For the purposes of this study, eighteen of the nineteen items was used in the second section, it fact does not affect the score results, since that latter item is not taken into account for the computation of the scale in the original questionnaire. On the other hand, the last item has to do with specific sleep disorders, for example, *-sleep apnea-*, while this study seeks to understand sleep habits in healthy people.

3.1.3 Sleep Hygiene Index

It is an instrument designed to measure the sleep hygiene behavior in a nonclinical population. Its theoretical basis is in the criteria that International Classification of Sleep Disorders (ICSD) uses to diagnose an inadequate sleep hygiene. The scale has thirteen items and has reported an internal validity of $\alpha = 0.71$, as well a high reliability in test-retest evaluation(Mastin et al., 2006).

For the purposes of this study and based on what the literature reports regarding sleep hygiene factors, the following adjustments were made to the instrument, without interfering with its essence:

- Following the structure and the meaning of the item four - *I exercise to the point of sweating within 1 h of going to bed*-, two items were added: - *I exercise to the point of sweating during the morning*- and -*I exercise to the point of sweating during the afternoon.*- . The main purpose was to know whether the exercise in the morning or in the afternoon is directly correlated with sleep quality.[cita]
- Item six of the original questionnaire that asked about the use of tobacco, alcohol, or caffeine, became three items to ask separately about the use of these three substances.
- Item 11 in the original questionnaire asks about an uncomfortable bedroom, due to four environment factors. In the questionnaire for this study, four items was generated from this one.
- Based on what the literature says about dinner type and schedule, and its negative impact on sleep quality, an item was added to the questionnaire (Posner and Gehrman, 2011, Stefano2014,Irish et al. (2015),Wentz and Wentz (2011)).

3.2 Validity and reliability

After the instrument was completed, it was validated by five experts that qualify each item on a escale of 1 to 5 for two metrics, clarity and pertinence. All items were qualified as clear and relevant. with the mean of 4.5 and 4.7 respectively. A sample of 30 people was randomly selected to perform the pilot test and obtain the internal validity of the instrument. Cronbach's alpha for the instrument after pilot test was $\alpha = 0.68$. This α value is acceptable and consistent with that reported by (Mastin et al., 2006) for the SHI scale, with this we proceeded to apply the questionnaire to a wider population to collect the dataset with which the analyzes were made for the selection of Variables that will be taken into account for the construction of the model.

3.3 Dataset

As a result of apply the questionnaire, a raw dataset ($m = 342, n = 47$) was obtained, this dataset, have missing data, some columns are no significant in terms of variance, there exist data in a wrong format to analyze, among other data quality issues. To obtain the dataset to the feature selection analysis, it was neccesary a pre-process of data, what included: The data quality analysis, the data quality plan to attend the issues and the implementation of the data quality plan. After this pre-process of data, the final dataset is conformed by one ID, 51 continuous and 10 categorical features grouped as shows the table 3.1.

Table 3.1: Columns distribution by its nature

Group	Categorical	Continuous	Total
ID	0	1	1
Demographics data	7	1	8
PSQI	14	4	18
SHI	21	0	21
Scale PSQI	8	1	9
Scale SHI	0	5	5
Total	50	12	62

The complete description of the dataset is in the tables 3.2, 3.3, 3.4, 3.5 and 3.6. At the end, the dataset contains 21 columns of main predictive variables (SHI), one column for continuous target variable (SQTT), and, one column for categorical target variable (SQCL). The other features in the dataset have diverse purposes as the last column of each table describes.

Table 3.2: Demographic features description

Grupo	Feature	Type	Values	Purpose
ID	EMAIL	Text	inf	identificator
Demographic	DD1	Continuous	[1-100]	Demographic data for statistical purposes only
	DD2	Categorical	Female, Male	
	DD3		Student, Employer, Teacher, Independent professional, Other	
	DD4		Intellectual, Physical, More intellectual than physical, More physical than intellectual	
	DD5		SDA, Catholic, Jehovah's withess, Evangelic, Other	
	DD6		Married, Single, Divorced, Free Union, Other	
	DD7		No, Hypertension, Diabetes, Depression, Other	Demographic information with filtering purposes
	DD8		No,Financial, Divorce process, Loss of Family, Other	Financial,

Table 3.3: PSQI features description

Group	Feature	Type	Values	Purpose
PSQI	SQ1	Continuous	A real numer. $0 \leq SQ1 \leq 12$	Provide information to PSQI Scale. A 0 value is the best for sleep quality and 3 is the worst
	SQ2		An integer number. $0 \leq SQ2 \leq 60$	
	SQ3		A real number. $0 \leq SQ3 \leq 12$	
	SQ4		A real number. $0 \leq SQ4 \leq 12$	
	SQ5a	Categorical	A level variable. $0 \leq SQ* \leq 3$.	
	SQ5b			
	SQ5c			
	SQ5d			
	SQ5e			
	SQ5f			
	SQ5g			
	SQ5h			
	SQ5i			
	SQ5j			
	SQ6			
	SQ7			
	SQ8			
	SQ9			

Table 3.4: PSQI Scale features description

Group	**Feature**	**Type**	**Value**	**Purpose**
PSQI Scale	SQDUR	Categorical	A level variable. $0 < SQ{*} \leq 3$.	Results of SQ duration. A 0 value is the best for sleep quality and 3 is the worst.
	SQDIS			Results of SQ disturbances. A 0 value is the best for sleep quality and 3 is the worst.
	SQLAT			Results of SQ latency. A 0 value is the best for sleep quality and 3 is the worst.
	SQDD			Results of SQ day dysfunction. A 0 value is the best for sleep quality and 3 is the worst.
	SQSE			Results of SQ sleep efficiency. A 0 value is the best for sleep quality and 3 is the worst.
	SQSQ			Results of SQ sleep quality general perception of the respondent. A 0 value is the best for sleep quality and 3 is the worst.
	SQMS			Results of SQ, needs meds. A 0 value is the best for sleep quality and 3 is the worst.
	SQTT	Continuous	An integer value. $0 < SQTT \leq 21$.	Total of PSQI
	SQCL	Categorical	A level value.	Good/ Poor

Table 3.5: SHI features description

Group	**Feature**	**Type**	**Values**	**Purpose**
SHI	SHI	Categorical	A level variable. $0 < SH{*} \leq 4$.	Predictive Features (Provide information for SHI scale). A 0 value is the best for sleep hygiene and 4 is the worst
	SH2			
	SH3			
	SH4			
	SH5			
	SH6			
	SH7			
	SH8			
	SH9			
	SH10			
	SH11			
	SH12			
	SH13			
	SH14			
	SH15			
	SH16			
	SH17			
	SH18			
	SH19			
	SH20			
	SH21			

Table 3.6: SHI scale feature description

Group	**Feature**	**Type**	**Value**	**Purpose**
SHI SCALE	SHSTR	Continuous	Sum of five SH features. $0 < SQ* \leq 20$.	Group of stress features
	SHDIS		Sum of five SH features. $0 < SQ* \leq 20$.	Group of disruptors features
	SHCH		Sum of six SH features. $0 < SQ* \leq 24$.	Group of circadian features
	SHDG		Sum of three SH features. $0 < SQ* \leq 12$.	Group of drugs features
	SHTT		Sum of SHSTR,SHDIS, SHCH and SHDG. $0 < SQ* \leq 76$.	Total of SHI

Chapter 4

Data pre-process

4.1 Estructuration and validation data process

Before to generate the quality report of the data, the data are loaded and passed for a process of validation and restructuration. This process includes the renaming of the columns and data validation for columns containing information of time and age. Additionally, the values for the responses of SHI and PSQI questionnaires was recodified from original responses (‘Nunca’,‘Casi nunca’,‘Algunas veces’,‘Frecuentemente’,‘Siempre’) to data that can be used to numerical and algorithmical analysis (‘0’,‘1’,‘2’,‘3’,‘4’).

The resulting dataset has 48 columns, distributed as the 4.1 shows:

Table 4.1: Distribution of features in the dataset by type

Type	Quantity	Columns in the dataset
CharID	1	1
Categorical	41	[3-9] and [14-48]
Continuous	6	2 and [10-13]

The quality analysis of data was based in the recommendations of (Kelleher et al., 2015). The resulting dataset will allow runs the algorithms to select the relevant features to generate the model. The analysis of the data quality includes the treadment of missing values, outliers and cardinality as well as correction of some possible bugs in the scripts that do the process of restructuration and validation of the dataset. The data quality analysis begins with a report of quality of continouos and categorical features. For continuous variables ten metrics were analyzed: quantity, missing values, cardinality, minimum value, first quartile, median, third quartile, maximum value, mean, and standard deviation. For categorical variables, nine metrics were analyzed: quantity, missing values, cardinality, mode, mode frequency, mode percent, second mode, second mode frequency and second mode percent. Before to do this report, a look at the complete dataset allowed to identify three records that contains no data for any feature. These records were deleted to avoid noisy information in the quality analysis.

4.2 Data quality report

The data quality report was performed using two scripts, one for continuous and one for categorical features, so, the features were grouped by type to do the analysis.

4.2.1 Continuous features

The dataset contains five continuous features, one for demographic data (DD1=AGE) and four features that measure the sleep duration (SQ1="Time to go to bed", SQ2="Latency of sleep", SQ3="Time to wake up", SQ4="Period of time between going to bed and waking up").

Table 4.2: Data quality report of continuous features

Feature	Count	Miss	Card	Min	Qrt1	Median	Qrt3	Max	Mean
DD1	338	6	49	16	27	35	44	66	35.92
SQ1	338	0	29	0	10	11	11.08	12.5	9.99
SQ2	338	1	15	1	5	15	20	60	14.99
SQ3	338	2	40	0.7	5.08	6	7	11	6.18
SQ4	338	3	107	-0.05	6.17	6.88	7.75	10.75	6.91

The table 4.2 shows some irregularities in continuous variables, as we can see, a negative value is the minimum value in the SQ4 variable; this variable represents the *Period of time between going to bed and waking up*, thus no negative value must be enter in this field, likewise, it appear a 0.0 value as the minimun value in the SQ1 variable, which is wrong because this is the time that respondents said go to bed, and, in this case, 0.00 is not a valid answer, in any case, an appropriate answer would be 12.00, referring to midnight. On the other hand, the variable SQ2, has a large standard deviation ($\sigma = 11.91$), since the variable represents the minutes that person takes to fall asleep in minutes ($\mu \simeq 15$).

No one of these features have a great quantity of missing values, DD1 is the variable with most of them, however the missing values represents only the 0.018% of the data, whis is not significant. If we assume that each record that contains a missing value is a different row, we have 12 records, which represents the 0.036%. As this percentage is small, and there are not in our hands, previous works describing the tendency of the data, these records with missing values could be deleted.

These continuos variables shows good cardinality, even though the ratio between the cardinality and number of records is not close to one. The nature of the data justifies this fact, because although the features are continuous and in theory they could take a large number of values, they take a small range of values, for instance, the variable SQ1 take an small range of values because the people answer commonly to this kind of question with onclock time, it means, people ask that they go to bed at $9:00$, $10:30$, $10:45$ or $11:00$, even when they, actually were to bed at $9:03$, or $10:33$ referring to the first to examples. The other variables have similar nature, thus the conclusion that cardinality is good for this variables, where the smallest cardinality was 15.

Additionaly to previous analysis, histograms and boxplot were generated to observe the behaviour of the data.

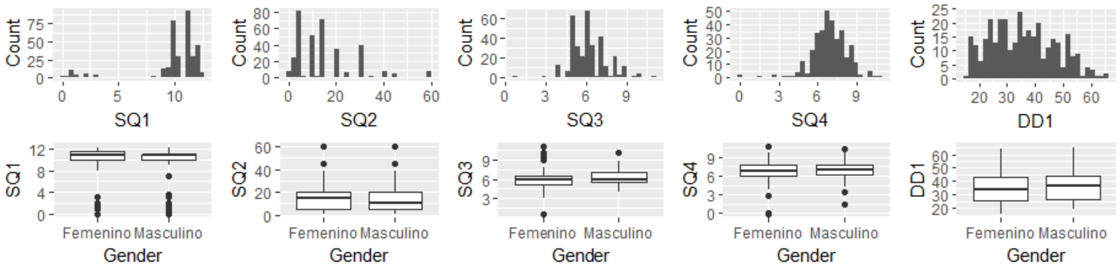


Figure 4.1: Histograms and boxplots of the continuous features

The plots in the Fig. 4.1 shows that the continuous features have outliers which should be analyzed to include/exclude from de dataset before of the training of the model to obtain better results. These variables do not intervene directly in the generation of the model, the model is generated from the SH features, however, these outliers could be indicators of some disorder of sleep in the respondent, thus the analysis must be performed.

The Table 4.3 summarizes the data quality issues in continuous features and the potential strategies to attend them.

Table 4.3: Data quality plan for continuous features

Feature	Data quality issue	Strategie
SQ1	Data contain 0.0, the correct value should be 12.00	Refine the process of convert data for this field
SQ4	Data contain negative values, it is wrong because the data represents the time to wake up	Refine the process validating the data, to correct the problem, or, eliminate the records with this issue.
SQ2	Standar deviation too large	Finding outliers visually and analytically. Excluding outliers from the modeling may improve the predictions. These analysis of outliers includes all continuous variables.
All	Missing values	The percentage of missing values is low, it allows to eliminate records with missing values. The decision of impute data is a few probable, since no reports are in our bibliography to know the tendency of the data.

4.3 Categorical Features for demographics data

The categorical features were divided in two groups, the demographic features are in the first group and the features containing all the information over the sleep hygiene are in other group.

Table 4.4: Data quality report of continuous features

Feat.	**Count**	**Miss**	**Card**	**Mode**	**MF**	**M%**	**Mode2**	**M2F**	**M2%**
DD2	338	0	2	Femenino	188	55.62%	Masculino	150	44.38%
DD3	338	1	5	Docente	143	42.31%	Empleado	70	20.71%
DD4	338	0	4	Más mental que físico	156	46.15%	Mental	143	42.31%
DD5	338	2	5	ASD	150	44.38%	Católica	129	38.17%
DD6	338	0	1	Unión Libre	338	100%	NA	NA	NA%
DD7	338	0	5	Ninguno	284	84.02%	Otro	26	7.69%
DD8	338	1	5	No	272	80.47%	Otra	34	10.06%

The table 4.4 shows only one irregularity in this set of variables, the variable DD6 has the hihest mode posible (100%), if data are right, all respondents are living in free union status which is very doubtful taken in account the population were the questionnaire was applied. There are less missing values than in the continuous features, so, it is possible to think in eliminate the records. The cardinality is not a problem in this set of features (except for the variable DD6 as commented above), since all posibilities have representation. In the two cases (features DD7 and DD8) where the mode capture a high percentage of the data, the information is good for this research. DD7 refers to people who suffer some chronic disease, the best answer to this work is “Any” because the intention is to work with healthy people, likewise in the DD8 variable that ask to the people if they are in some crisis that disrups their sleep, the best answer to this work is “No”, fortunatelly, this is the mode.

4.3.1 Categorical features for Slepp hygiene

Table 4.5: Report of quality of the SH features

Feature	**Count**	**Miss**	**Card**	**Mode**	**MF**	**M%**	**Mode2**	**M2F**	**M2%**
SH1	338	0	5	1	114	33.73%	0	111	32.84%
SH4	338	1	5	0	204	60.36%	1	67	19.82%
SH5	338	0	5	0	176	52.07%	2	64	18.93%
SH6	338	1	5	0	152	44.97%	1	75	22.19%
SH7	338	2	5	0	142	42.01%	1	107	31.66%
SH8	338	0	5	0	319	94.38%	4	7	2.07%
SH9	338	0	4	0	276	81.66%	1	34	10.06%
SH10	338	0	5	0	223	65.98%	1	57	16.86%
SH11	338	1	5	0	104	30.77%	2	78	23.08%
SH12	338	0	5	1	133	39.35%	2	114	33.73%
SH13	338	0	5	2	92	27.22%	1	76	22.49%
SH14	338	0	5	0	210	62.13%	1	61	18.05%
SH15	338	1	5	0	101	29.88%	1	91	26.92%
SH16	338	0	5	0	163	48.22%	1	84	24.85%
SH17	338	1	5	0	222	65.68%	1	82	24.26%
SH18	338	0	5	0	173	51.18%	1	82	24.26%
SH19	338	0	5	0	125	36.98%	1	81	23.96%
SH20	338	0	5	2	117	34.62%	1	90	26.63%
SH21	338	0	5	1	130	38.46%	2	96	28.40%

In Table 4.5, the cardinality shows that all possible value $[0-5]$ for each answer is represented in the data, except by the SH9 variable where one of the options was not selected as answer of the respondents. It is good for the quality of data, however, there are two variables with high mode. The 81.66 % of the respondents, answered *never (*0*)* to the question SH9 *"I use alcohol within 4 hours of going to bed or after going to bed."*, while the 94.38% answered *never (*0*)* to the question SH8 *"I use tobacco within 4 hours of going to bed or after going to bed."*, which means that there are few variability in the data in these two variables. It is possible to dispense with these data for the analysis, since they do not contribute much information to the studied phenomenon for this population. The other 19 categorical variables for SHI, have a cardinality of 5, and the higher mode is placed in SH10 *"I use cafeine within 4 hours of going to bed or after going to bed."* were a 66.98% of the respondents answered *never (*0*)* for this question. This means that the answers have a good range of variability to be analized, and to participate as candidate of be selected as feature to training the model.

This set of data has small number of missing values, however, in this case it is possible to impute data due to the features together represents a behavior of the person. Algorithms as KNN or a multiple logistic regression can performs data imputation to have a good approximation to the true data. The table 4.6 present the summary of the issues and potential strategies for the SH features.

Table 4.6: Potential strategies to attend SH features

Feature	Data quality issue	Strategie
SH8	The mode is very high (>94%)	Analyze the relevance of include this variable to the analysis due the few variability
SH9	The mode is high (>81%)	Analyze the relevance of include this variable to the analysis due the few variability
All	Missing values	The percentage of missing values is small, however, imputation will be performed for this missing values.

4.4 Following the quality plan to attend issues

The first step in order to attend the issues was the analysis of the code that validates the raw data, to avoid the suspicious of bugs that could be generate wrong data. After the code is validated, the errors in data can be adjudicate to human capture.

After the analysis of scripts, tree bugs were fixed. The first bug is related with the reason that the civil status have a mode equivalent to the number of records in the dataset. The bug was generated by omittining a condition in the evaluation of a missing value in the field in the same line of code that attempt to standardize the results so that all the sentences were in the same style of case. In the case of status civil feature, some answers are wroten as *'Unión libre'*, while other was wroten as *'Unión Libre'*.

The line with the bug:

```
ifelse(is.na(dataSet$DD6),dataSet$DD6<-NA,dataSet$DD6<-"Unión Libre")
```

The line after being fixed:

```
ifelse(is.na(dataSet$DD6),dataSet$DD6<-NA,dataSet[dataSet$DD6=='Unión libre',7]<-"Unión
Libre")
```

The second bug was identified in the script that calculates the time of sleep depending of the three variables, *'Time to go the bed'*, *'Time to wake up'*, *'Time to fall asleep'*, in this case, the condition for the calculation does not contemplate that a person could say that he went to bed and got up twelve hours apart. The problem was solved by modifying the conditional operator of $>$ to $\geq$.

Code with bug:

```
if(HD>HL){
  HD<-HD-12.00
}
SE<-abs(HL-HD)
SE<-SE-round(minutos/60,digits = 2)
```

Code after being fixed:

```
if(HD>=HL){
  HD<-HD-12.00
}
SE<-abs(HL-HD)
SE<-SE-round(minutos/60,digits = 2)
```

The third bug was corrected by adding two condition per values before ignored. Values in the range of $(-\infty, 0)$ must be taken a NA value, and values in the range of $[0, 1)$ must be transformed by adding 12.00.

The lines that were added are:

```
if(!is.na(s3)){
  if(as.numeric(s3)<0){
    s3<-NA
  }else if(as.numeric(s3)<1){
    s3=as.numeric(s3)+12
  }else{
    s3<-as.numeric(s3)
  }
}
```

The quality reports generated after apply this corrections, show a difference in the identified features with possible issues due to a wrong treatment.

Table 4.7: Quality report of continuous features after recoding the scripts

Feature	Count	Miss	Card	Min	Qrt1	Median	Qrt3	Max	Mean	Sdev
DD1	338	6	49	16	27	35	44	66	35.92	11.41
SQ1	338	0	27	1	10	11	11.5	12.75	10.17	2.46
SQ2	338	1	15	1	5	15	20	60	14.99	11.91
SQ3	338	2	40	3	5.21	6	7	12.7	6.21	1.33
SQ4	338	3	107	0.12	6.17	6.89	7.75	11.95	6.94	1.33

Table 4.8: Quality report of demographic features after recoding the scripts

Feature	Count	Miss	Card	Mode	MF	M%	Mode2	M2F	M2%
DD2	338	0	2	Femenino	188	55.62%	Masculino	150	44.38%
DD3	338	1	5	Docente	143	42.31%	Empleado	70	20.71%
DD4	338	0	4	Más mental que físico	156	46.15%	Mental	143	42.31%
DD5	338	2	5	ASD	150	44.38%	Católica	129	38.17%
DD6	338	0	5	Casada(o)	181	53.55%	Soltera(o)	119	35.21%
DD7	338	0	5	Ninguno	284	84.02%	Otro	26	7.69%
DD8	338	1	5	No	272	80.47%	Otra	34	10.06%

The minimum value in SQ1 is not zero as before, now it is one, which is reasonable; the variable SQ4 do not have a negative value as minimum value. The present also is very small (0.12) and was verified in the raw data, the conclusion is that it was a wrong user capture, the respondent said that he/she go to bed at 12:30 and wakes up at 12:42 every day. On the other hand, the civil status (DD6) has a congruent value for the sample of study, the 53.55% of the respondents said be married and 35.21% be single, in contrast with previous data where it was reported that 100% of the people had a ‘Free union’ civil status.

The second step was to work with the missing values in the continuous features and missing values in the categorical demographic features. Records containing missing values were eliminated as proposed in the data qulaity plan described in Table 4.3. The same apply to the categorical demographic features.

After applying the process to delete records with missing values in continuous and categorical demographics features the dataset reducts its dimentionality from 338 to 326, which represents a reduction of 0.036% of the original data.

In the third step of this stage the atypical values are identified to exclude them from the dataset that will serve as a source of the training of the model. The process excludes records that exceed three standard deviations in the variable SQ4 (‘time the person spent asleep’). This variable is of high impact on the quality of sleep and is not considered within the factors of sleep hygiene. Another indicator of a possible sleep disorder is the latency, latency is the time between a person go to bed and he/she fall asleep. The records containing outliers in this feature, also were eliminated.

After delete records with outliers n continuous variables the dataset reducts its dimentionality from 338 to 306, which represents a reduction of 0.095% of the original data.

The Fig. 4.2, shows the histograms and boxplots after eliminated records with outliers in SQ2 and SQ4, boxplots in these two variables make it clear that do not outliers were found for these features. DD1 has no outliers too, even though maintaining the original records.

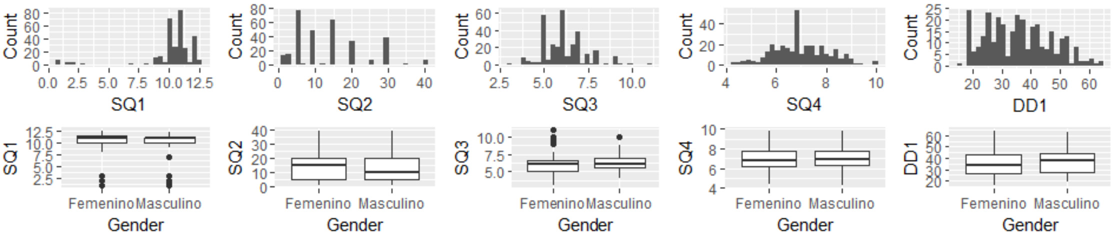


Figure 4.2: Histograms and boxplots of the continuous features after delete outliers

4.5 Imputation of missing values in SH and SQ features

The process of imputation was performed using *mice and VIM package* (van Buuren and Groothuis-Oudshoorn, 2011, Kowarik and Templ (2016)). the process begins with the analysis of missing values in the variables involved indirectly or directly in the generation of the model. The table 4.9 shows that the dataset has few missing values (Not all columns in the table are included for lack of space, however, all columns that were omitted do not have missing values). It has 297 complete records, a record with a missing value in the column SQ5a, one more with a missing value in the column SQ5c, an so on. The dataset contains eight records with a total of nine missing values in eight variables.

Table 4.9: Report of missing values

	SQ5b	SQ5d	...	SH20	SH21	SQ5a	SQ5c	SH4	SH6	SH11	SH15	SH17	SH7	
297	1	1	...	1	1	1	1	1	1	1	1	1	1	0
1	1	1	...	1	1	0	1	1	1	1	1	1	1	1
1	1	1	...	1	1	1	0	1	1	1	1	1	1	1
1	1	1	...	1	1	1	1	0	1	1	1	1	1	1
1	1	1	...	1	1	1	1	1	0	1	1	1	1	1
2	1	1	...	1	1	1	1	1	1	1	1	1	0	1
1	1	1	...	1	1	1	1	1	1	0	1	1	1	1
1	1	1	...	1	1	1	1	1	1	1	0	1	1	1
1	1	1	...	1	1	1	1	1	1	1	1	0	1	1
	0	0	...	0	0	1	1	1	1	1	1	1	2	9

A summary of the missing values its presented in the table 4.10

Table 4.10: Summary of missing values in SH features

Feature	Count of missing values
SQ5a	1
SQ5b	1
SH4	1
SH6	1
SH7	2
SH11	1
SH15	1
SH17	1
Total	7

The figure 4.3 shows in the left side, an histogram of the features with missing data depicting the influence of missing values in the dataset. The right side shows the pattern of missing values in the dataset, it concentrates all complete cases in the botton of the graph, which reach a 97.06% of the dataset. The remaining of the figure shows the features with missing data, placing in the right side the corresponding percentages per variable.

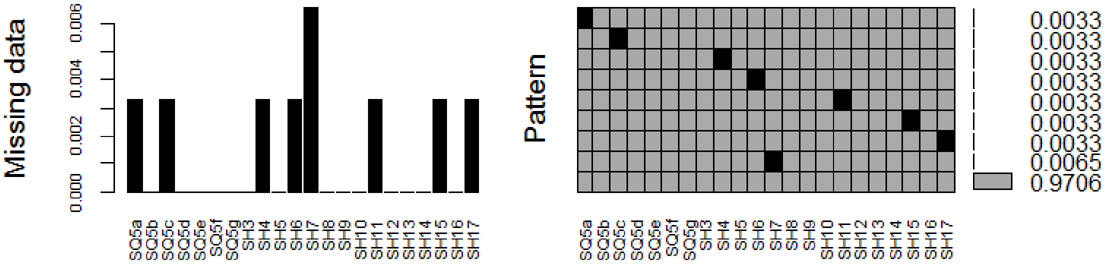


Figure 4.3: Pattern of missing values

The *mice* function was executed to impute data in the records containing missing values, the process was performed in a temporal dataset comformed only for those features with missing data. The *mice* function perform data imputation using *polytomous regression imputataion* for unordered categorical data with more of two leves, which is the case. The multinomial logistic regression was applied with 50 iterations and five datasets to obtain a table of results that allow to choose the best option to do the imputation.

Table 4.11: Five datasets with data imputation

Number of Row	DS1	DS2	DS3	DS4	DS5
ROW 225	2	1	0	1	3
ROW 72	3	1	0	2	2
ROW 12	0	0	0	0	3
ROW 233	2	1	1	4	2
ROW 17	0	1	2	1	0
ROW 146	0	1	2	0	2
ROW 144	3	0	0	3	1
ROW 183	0	4	0	3	4
ROW 201	0	2	0	0	0

The DS4 is the dataset most consistent with the proposals in the other datasets, it matches with at least one dataset of the remaining four, in seven of the rows. The DS4 was chosen to impute data in these variables.

Table 4.12: Report of missin gvalues after imputation

SQ5a	SQ5c	SH4	SH6	SH11	SH15	SH17	SH7	
1	1	1	1	1	1	1	1	0
0	0	0	0	0	0	0	0	0

The table 4.12, shows that no missing value are in the dataset after the imputation, now the new dataset of variables with data imputation will be merged with the others features of the original dataset in a new dataset to be saved and used in advanced to train the model.

4.6 Final results

The following tables show the report of quality in the dataset after apply deletion of missing values, exclusion of records containing outliers and imputation of missing values in variables that are closely related with the training of the model.

Table 4.13: Report of Quality of SH features after cleaning and imputation

Feature	Count	Miss	Card	Mode	ModeFrec	ModePerc	Mode2	Mode2Frec	Mode2Perc
SH1	306	0	5	1	109	35.62%	0	97	31.70%
SH2	306	0	5	1	109	35.62%	2	107	34.97%
SH3	306	0	5	1	147	48.04%	2	70	22.88%
SH4	306	0	5	0	186	60.78%	1	63	20.59%
SH5	306	0	5	0	159	51.96%	2	56	18.30%
SH6	306	0	5	0	135	44.12%	1	69	22.55%
SH7	306	0	5	0	128	41.83%	1	99	32.35%
SH8	306	0	5	0	292	95.42%	3	5	1.63%
SH9	306	0	4	0	254	83.01%	1	30	9.80%
SH10	306	0	5	0	201	65.69%	1	53	17.32%
SH11	306	0	5	0	95	31.05%	2	69	22.55%
SH12	306	0	5	1	122	39.87%	2	99	32.35%
SH13	306	0	5	2	85	27.78%	1	69	22.55%
SH14	306	0	5	0	188	61.44%	1	54	17.65%
SH15	306	0	5	0	94	30.72%	1	80	26.14%
SH16	306	0	5	0	150	49.02%	1	76	24.84%
SH17	306	0	5	0	200	65.36%	1	78	25.49%
SH18	306	0	5	0	158	51.63%	1	73	23.86%
SH19	306	0	5	0	112	36.60%	1	78	25.49%
SH20	306	0	5	2	107	34.97%	1	85	27.78%
SH21	306	0	5	1	122	39.87%	2	88	28.76%

With exception of high percentages in mode of the SH8 and SH9, all other features present a good behavior in the dataset. The issue of SH8 and SH9 will be attended in a future stage of the work, the process of feature selection will take care of this.

The last process to be carried out in this stage is the calculation of SHI and PSQI scores through the scales provided by the respective questionnaires.

The final dataset to be used in the process of feature selection is a 306 x 62 dataset containing 10 continuous features, 51 categorical features and one ID feature. To more detail of this dataset, see please the tables at the end of the previous section.

Chapter 5

Feature selection

This section describes the process that we perform to reduce the dimension of the sleep hygiene data set that contains the features to model the quality of sleep for respondents of the survey described in chapter 3. There exist two ways to addressed dimensionality reduction, feature extraction and feature selection. Feature extraction, consists in generate a new and small feature space. The application of a technique of feature extraction produce new features based in original ones. The new dataset is not understandable in terms of the original dataset, rather, it is an abstraction of this and its visualization have no practical meaning. On the other hand, feature selection as ilustrate the Fig. 5.1 choose a small subset of the relevant features from the original dataset according to certain relevance evaluation criterion, which usually leads to better learning performance, lower computational cost, and better model interpretability (Tang et al., 2014).

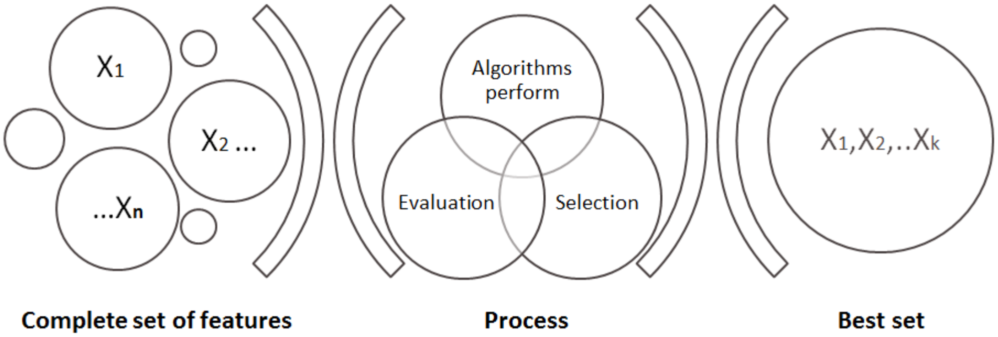


Figure 5.1: Feature selection Process

For the purposes of this study, the technique of selection of characteristics is the most appropriate. Our interest in reducing dimensionality is not related to the decrease in computational cost, rather, the purpose is to decrease the number of predictive variables due to the high cost of design and infrastructure that means capturing 21 different signals through sensors. If it is possible to characterize a high percentage of the phenomenon, through a reduced number of factors of sleep hygiene, the design of the system will be more feasible and less expensive.

The model accuracy for prediction of the sleep quality with the subset of features must be better than the training model using the total of sleep hygiene features.

5.1 Feature selection models

In 1996, (Liu and Motoda, 1998) proposes two models to achieve the reduction of features, that have been used as basis of diverse algorithms still in force. The filter model (see Figure 5.2) that uses as criterion of feature selection, some attributes concerning only to the data domain. Especifiacally in this model, Liu et. al. proposes that it is posible to analyze and make decisions over irrelevance or relevance of features based in measure information gain, dependence, distance and consistency.

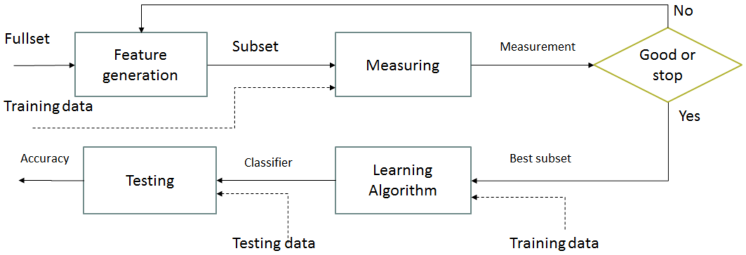


Figure 5.2: Filter model proposed by Liu et. al.

The second model showed in Fig. 5.3 proposed is the wrapper model that uses the accuracy of prediction as selection criterion, it means that this techniques are committed with a particular classifier in this stage of the learning process.

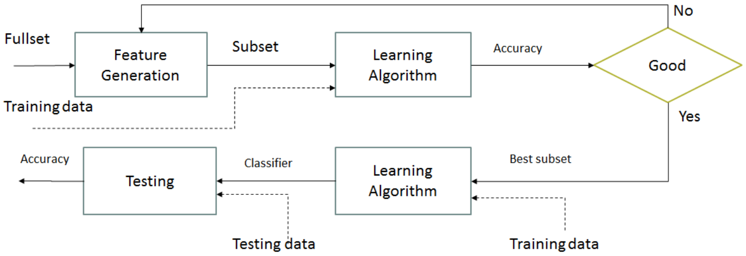


Figure 5.3: Wrapper model proposed by Liu et. al.

Both models have advanteges and disadvantages, techniques based in filter model, performs better than others based in wrapper model, however, researchers have no idea over prediction accuracy during the feature selection process. Some practitioners don't preffer to use these techniques because if accuracy prediction is not achieved in the proposed level, the first steep can be regarded as a waste of time. On the other hand, some researchers argued that select features based in determinated classifier, reduces the possibility to use other classifier to generate the prediction model, in this sense, the classifier to generate the final model should be choosed at the begining, and it is not convenient for all problems. In these order of thinks, (Kelleher et al., 2015) comment that wrapper models are more computationally expensive than filters models and that the argument of they are uncertain models respect to the accuracy, is not at all valid since filters model often generate models with good accuracy.

Additionaly, (Liu and Motoda, 1998) highlight *Search*, *Scheme* and *measure* as three important concepts that help to decide what technique is the most appropriate for an specific problem of dimentionality reduction by feature selection (see Fig. 5.4). Search refers to the activity of choose features in non deterministic, heuristic or complete form, Scheme must be determine if the search will be forward, backward or in random mode, and, measure has to do with tree ways to establish the threshold for stopping the feature search, the criterion used are accuracy, consistency, and, classic criterion involving distance, information gain and dependence.

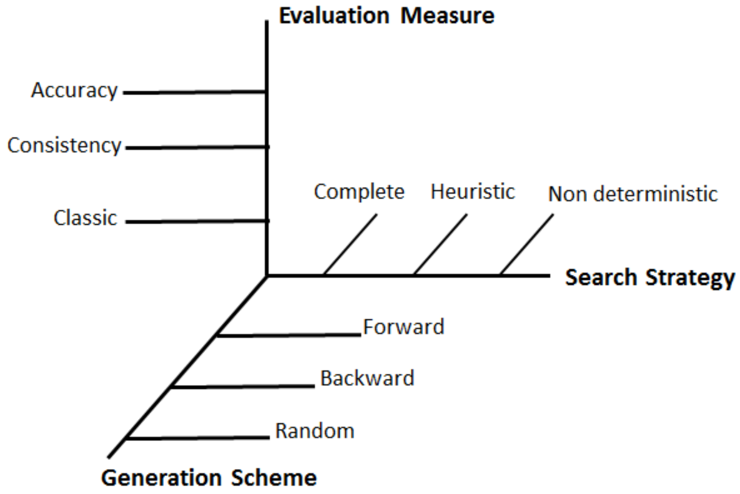


Figure 5.4: Main dimensions in feature selection, Liu et. al.

A third type of model has been proposed in last years, these models are called **embedded models**, since they allow practitoners select features while the prediction model is built. Embedded models have the advantage of filters model in terms of low computational cost, and take the advantage of wrapper model, because the prediction accuracy and classification model are involved in the process. (Tang et al., 2014) describe three type of embedded methods as we shows in the Table 5.1.

Table 5.1: Embedded methods as Tang et al. (2014) describes and quoted verbatim in his paper.

Method	Description	Cite
Pruning	Utilizing all features to train a model and then attemp to eliminate some features by setting the corresponding coefficients to 0, while maintaining model performance such as recursive feature elimination using support vector machines (SVM)	Guyon et al. (2002)
Build-in	Mechanism for feature selection as ID3 and C4.5	Quinlan (1986, 1993)
Regularization	Utilices objective functions that minimize fitting errors and in the mean time force the coefficients to be small or ti be exact zero.	Ma and Huang (2008)

These models are representative of the theoretical basis where a lot of algorithms for selection features in last twenty years have been fueled. Likewise four concepts are the most important and have been used for the generation of different feature selection algorithms in last two decades: distance, accuracy, inconsistency and information gain.

- Distance: The main goal to use distance, is to find similarity among instances in a dataset. The Equation proposed by Minkowski (see eq. (5.3)) is a generalization of the distances that are used in MLA. The most common distances are the particular cases where $p = 1$ called Manhatan distance (see Eq. (5.2)) and where $p = 2$, the well known Euclidian distance (see Eq. (5.1)). (All three equations were taken from (Kelleher et al., 2015)). The implication of use different values of p will be noted in the difference between two values of any feature in the final distance, it is directly proportional to the value of p. It means that large differences between two features in an instance, impact stronger in the final result when p grows.

$$Euclidean(A,B) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_1)^2 + \cdots + (a_n - b_n)^2} \tag{5.1}$$

$$Manhattan(a, b) = \sum_{i=1}^{m} abs(a[i] - b[i]) \tag{5.2}$$

$$Minkowski(a, b) = \left(\sum_{i=1}^{m} abs(a[i] - b[i])^{p} \right)^{\frac{1}{p}} \tag{5.3}$$

- Accuracy: Accuracy refers to the successes that a model had to predict each instance of a dataset, it is opposed to the miscalssification error as (Kelleher et al., 2015) defines in (5.4) and (5.5) equations. These two equation take relevance when accuracy is analized in the context of confusion matrix, a tool widely used to report the outcomes of the prediction thorugh a model. The confusion matrix together with the Receiver Operating Characteristics (ROC) curve, provides understanding and visualization of the specificity and sensibility, the most important metrics for evaluations of the models, especially in the health context.

$$misclassification\ rate = \frac{(FP + FN)}{(TP + TN + FP + FN)} \tag{5.4}$$

$$accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{5.5}$$

- Inconsistency: An inconsistency refers that two instances have the same value in all descriptive features, but they belong to a different class. We can compute two values to measure the inconsistency for a subset of features in a dataset. The first value, which is called the inconsistency count (IC), can be defined as $IC = nM - LCI$, where nM is the number of instances that coincide in all descriptive features, and, LCI is the largest class of the clases that are involved in this particular group of instances. The second value, is the inconsistency rate defined as $IR = \frac{\sum_{i=0}^{m} IC_i}{N}$, where m is the total of groups of matching instances in the dataset.
- Information Gain: Is a measure of the relevance that a predictive variable offers in relation to the target variable. To understand the concept of information gain, it is necessary to first understand the concept of information entropy as was raised by Shannon in 1948. In a dataset, an entropy value represents the heterogenity/homogenity of the target variable, in others words, if we have large probability of success to predict an outcome in the target feature, we have a set with small entropy and viceversa.

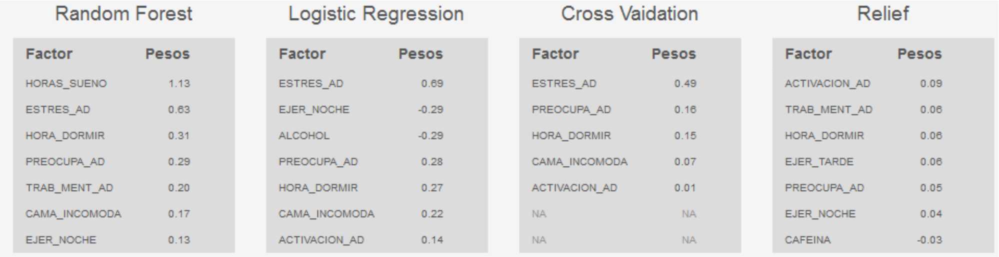
The process to calculate the information gain of one feature can be sumarized as follows:

1. Compute the total entropy.
2. Split the target feature in the levels of the predictive feature.
3. Compute the entropy of the target variable in each subset generate, and multiply the result by its weight. The weight is computed by dividing the number of instances in the subset, among the total number of instances in the dataset.
4. Subtrac form the total entropy, the entropy computed in the steps 2 y 3.
5. Sort the results in descending order to identify which are the best and the worst features in terms of provide information to characterize the phenomenon.

5.2 Feature selection process

Five methods for feature selection were selected to make the process of selection the relevant sleep hygiene factors. Each method works with the complete set of features and the total of the data, after the process, the features were ordered by relevance in descending order in each method. A merge process was performed to choose those features that were ranked in the first places in each method. This process ensure that features choosen are relevant features because the theory and math behind the methods are different in each one.

The Fig. 5.5 shows four of the six methods (for space reasons), and the corresponding features (Factor) and weights (Pesos) in descending order. The Fig. 5.6 shows the outcomes of selected factors by the merge process. The left side, is the table with features and the corresponding weights in the best algorithms, in these case Random Forest (RF), Logistic Regression (LR) and Logistic Regression with Cross Validation (LR_CV). The right side ilustrate in a line-graph, the comparisson of the data in the table of the left side. Both figures are screens capture of the application developed on Shiny R-Studio for this specific purpose.



Random Forest

Factor	Pesos
HORAS_SUENO	1.13
ESTRES_AD	0.63
HORA_DORMIR	0.31
PREOCUPA_AD	0.29
TRAB_MENT_AD	0.20
CAMA_INCOMODA	0.17
EJER_NOCHE	0.13

Logistic Regression

Factor	Pesos
ESTRES_AD	0.69
EJER_NOCHE	-0.29
ALCOHOL	-0.29
PREOCUPA_AD	0.28
HORA_DORMIR	0.27
CAMA_INCOMODA	0.22
ACTIVACION_AD	0.14

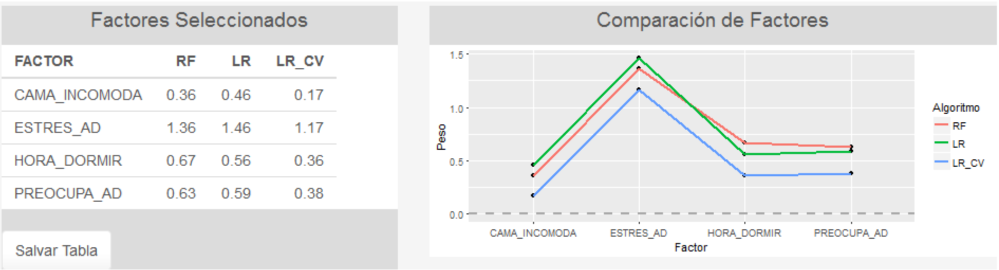
Cross Vaidation

Factor	Pesos
ESTRES_AD	0.49
PREOCUPA_AD	0.16
HORA_DORMIR	0.15
CAMA_INCOMODA	0.07
ACTIVACION_AD	0.01
NA	NA
NA	NA

Relief

Factor	Pesos
ACTIVACION_AD	0.09
TRAB_MENT_AD	0.06
HORA_DORMIR	0.06
EJER_TARDE	0.06
PREOCUPA_AD	0.05
EJER_NOCHE	0.04
CAFEINA	-0.03

Figure 5.5: Results of four methods after performs Feature Selection



Factores Seleccionados

FACTOR	RF	LR	LR_CV
CAMA_INCOMODA	0.36	0.46	0.17
ESTRES_AD	1.36	1.46	1.17
HORA_DORMIR	0.67	0.56	0.36
PREOCUPA_AD	0.63	0.59	0.38

Figure 5.6: Outcomes for selected factors after the merge process

From the original 21 hygiene factors that in theory are the predictive variables to characterize the sleep quality, the feature selection algorithms choose four features. It means that the remaining seventeen features, there are not relevant factors to characterize the phenomenon on this population. As additional information related with the studied phenomenon, it is possible to note that two features closely related with the state of mind, are present among features selected. Even, one of this two features, the stress before go to the bed (ESTRES_AD), is the most relevant feature of the four selected. If this selection provides the best model to characterize the phenomenon, a great challenge is perceived in the near future, due to how difficult it can be to measure a subjective variable, by means of an electronic device.

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.

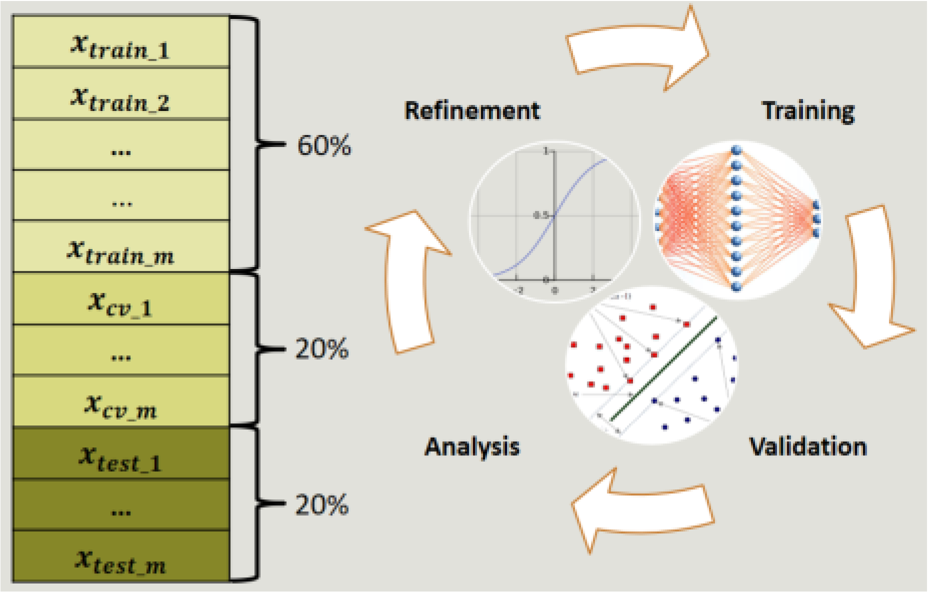


Figure 6.1: Cross Validation Process

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.

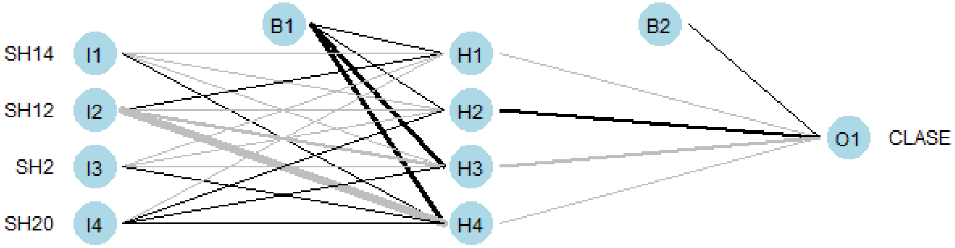
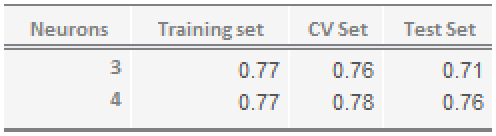


Figure 6.2: Structure of neural network with four neurons in the hidden layer

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

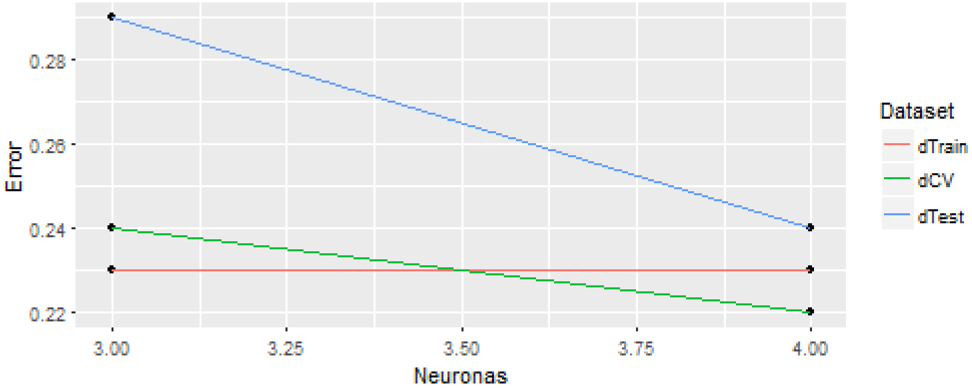


Figure 6.3: Comparison of the results for the two trained neural networks

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.

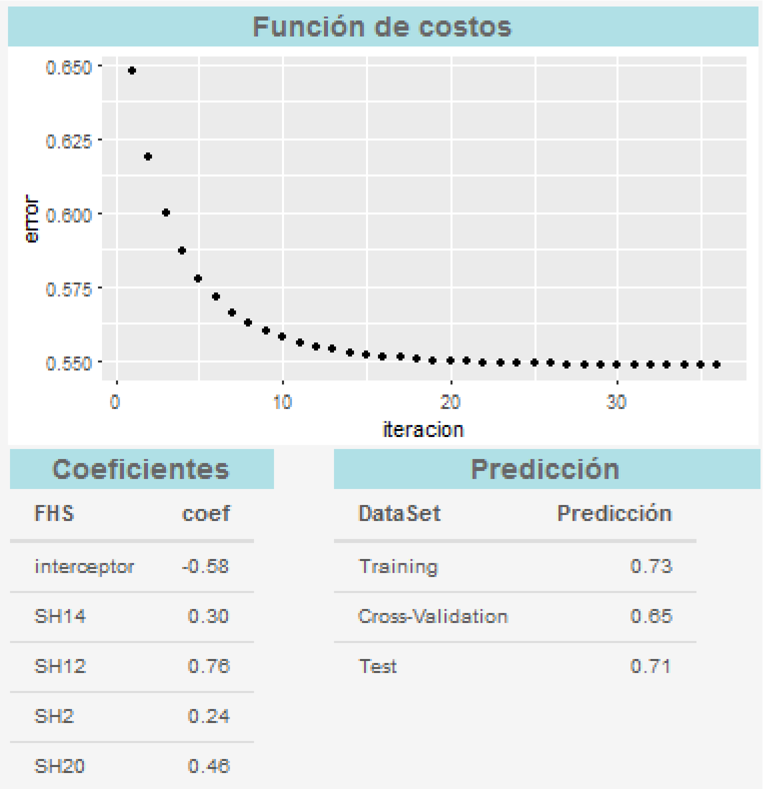
6.2 Logistic Regression Results

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.

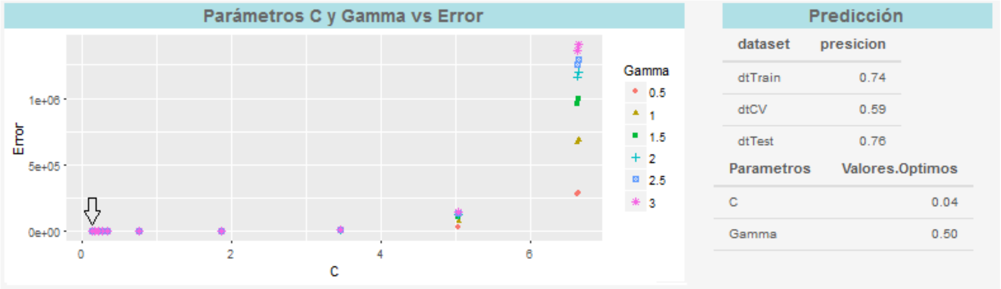
Table 6.1: Comparison of efficiency of LR with polynomials of degree one, two and three

	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

Final Words

To be continued.

Bibliography

Bai, Y., Xu, B., Ma, Y., Sun, G., and Zhao, Y. (2012). Will you have a good sleep tonight?: sleep quality prediction with mobile phone. In *Proceedings of the 7th International Conference on Body Area Networks*, pages 124–130. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

Brick, C. A., Seely, D. L., and Palermo, T. M. (2010). Association between sleep hygiene and sleep quality in medical students. *Behavioral Sleep Medicine*, 8(2):113–121. PMID: 20352547.

Buysse, D. J., Reynolds, C. F., Monk, T. H., Berman, S. R., and Kupfer, D. J. (1989). The pittsburgh sleep quality index: a new instrument for psychiatric practice and research. *Psychiatry research*, 28(2):193–213.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422.

Irish, L. A., Kline, C. E., Gunn, H. E., Buysse, D. J., and Hall, M. H. (2015). The role of sleep hygiene in promoting public health: A review of empirical evidence. *Sleep Medicine Reviews*, 22:23–36.

Kelleher, J. D., Namee, B. M., and D'Arcy, A. (2015). *Fundamentals of Machine Learning for Predictive Data Analytics: algorithms, worked examples, and case studies.* Number 1. The MIT Press, London.

Kowarik, A. and Templ, M. (2016). Imputation with the R package VIM. *Journal of Statistical Software*, 74(7):1–16.

Liu, H. and Motoda, H. (1998). *Feature selection for knowledge discovery and data mining.*

Ma, S. and Huang, J. (2008). Penalized feature selection and classification in bioinformatics. *Briefings in bioinformatics*, 9(5):392–403.

Mastin, D. F., Bryson, J., and Corwyn, R. (2006). Assessment of sleep hygiene using the sleep hygiene index. *Journal of behavioral medicine*, 29(3):223–227.

Posner, D. and Gehrman, P. R. (2011). Chapter 3 - sleep hygiene. In Perlis, M., Aloia, M., and Kuhn, B., editors, *Behavioral Treatments for Sleep Disorders*, Practical Resources for the Mental Health Professional, pages 31 – 43. Academic Press, San Diego.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1):81–106.

Quinlan, J. R. (1993). *C4.5: programs for machine learning.* Morgan Kaufmann.

Tang, J., Alelyani, S., and Liu, H. (2014). Feature Selection for Classification: A Review. *Data Classification: Algorithms and Applications*, pages 37–64.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67.

Wentz, D. and Wentz, M. (2011). *The Healthy Home: Simple Truths to Protect Your Family from Hidden Household Dangers.* Vanguard.