Universidad de San Andrés

Departamento de Economía

Maestría en Economía

Understanding Psychological Distress: A Predictive Model using Machine Learning's Classification Trees

Andrés MARTIGNANO

38.996.295

Mentor: Walter SOSA ESCUDERO

Victoria, Buenos Aires

13 de noviembre, 2020

Tesis de Maestría en Economía de

Andrés MARTIGNANO

"Detrás del malestar psicológico: un modelo predictivo mediante árboles de clasificación"

Resumen

A partir de la pandemia de COVID-19, hubo un debate acerca de la implementación de restricciones, principalmente la cuarentena. Sin embargo, uno de los argumentos en contra giraba en torno a la psicología: la movilidad limitada y, para muchos, el desempleo temporal, podrían representar un gran desafío para el bienestar personal. Dada la naturaleza compleja de la salud mental individual, este trabajo busca comprender algunas de las características que subyacen al malestar psicológico, y cómo la presencia o ausencia de algunos indicadores puede ser clave para comprenderlo mejor. Utilizando la información de la encuesta NHIS, se ha generado un modelo predictivo a través de *Random Forest* en el entorno de Python. Los resultados muestran que las limitaciones físicas, las restricciones económicas y la calidad del sueño, entre otras, poseen una gran relevancia para comprender el malestar psicológico.

Palabras clave: Salud, psicología, árboles de clasificación, aprendizaje automático

"Understanding Psychological Distress: A Predictive Model using Machine Learning's Classification Trees"

Abstract

When the global COVID-19 pandemic saw its outburst, there was a debate whether it was best to impose restrictions, namely quarantine. However, one of the counter-arguments fell around psychology: limited mobility and for many, temporary unemployment, could present a major challenge for personal well-being. Given the complex nature of individual mental health, the purpose of this paper is to understand some of the features that lie behind psychological distress, and how the presence or absence of some indicators may be key to comprehend the latter better. Using NHIS information, a predictive model has been generated by the Random Forest algorithm in the Python environment. The results show that physical restrictions, financial limitations, and sleep quality, amongst others, have a major relevance in understanding psychological distress.

Keywords: Health, Psychology, Classification Trees, Machine Learning

Códigos JEL: I10, I31, J10

Contents

I. Introduction

As of the spread of COVID-19 during 2020, the world began to deal with a health crisis with few precedents, and in order to tackle it, in many countries the implementation of a mandatory and strict quarantine was resolved. Although at the moment of making this decision the measure had considered the expected benefits of containing and/or dealing with the virus, some voices arose against this regulation.

One of the arguments against has been about the impact on individual's psychological health: the absence of free mobility and leaving the home -among other factors- would affect negatively the mood of the people. On the other hand, in those cases where jobs were lost as a result of the cessation of production or commerce, financial problems would potentially emerge for these households, with probable emotional consequences too. Other concerns laid on the household status, whether people were accompanied during quarantine (by a lifetime partner, for instance), and also on forcing people to remain in their neighbourhoods, despite their possible negative externalities.

Due to the above, the need to understand to what extent these arguments have validity arises; and if so, to find out which determinants can induce psychological problems, given observable characteristics.

The objective of the present work is the generation of a predictive model that will shed light on the factors that can contribute to psychological distress. Using information from the NHIS survey, it will be sought to learn a model that allows predictions of a series of indicators, such as sadness, anxiety, or depression, based on the use of machine learning methodologies.

Although these topics have been covered by the psychology literature in depth, many of these works have resorted to descriptive methodologies and/or resources such as logistic regression. Instead, in this paper, the chosen instrument has been the classification tree methodology (or CART, for its acronym). One of the main advantages of this tool is that it allows a clear identification of variable interactions when analysing the determinants of the stated psychological disorders, since it will indicate whether there are attributes that weigh more than others, thus capturing the sources of non-linearities.

Unlike other classification methods, such as the previously mentioned logistic regression, classification trees also have the advantage of being visibly simpler to explain phenomena, since they classify from partitions of the data space, being able to model within each category, or node. What follows is that trees provide a more precise analysis, conditioned on characteristics previously defined by the algorithm itself, while conventional linear models such as logistic regression, ordinary least squares or discriminant analysis linear are not as adequate.

Because of these reasons, the election of the proposed non-linear method for the study of psychological distress is founded in the fact that the interactions amongst variables can generate a more accurate prediction, whereas considered on their own, they would not be able to provide an equally efficient explanation. As for other studied models, neither a linear design nor majority voting are capable of capturing this variable dynamics: while the former weighs every variable equally, the latter fails to provide a clear explanation of the phenomenon. To be more specific, it is not clear that the people undergoing psychological distress share common characteristics between them, as a number of reasons may apply to some but not for others.

The main findings of the Random Forest models suggest that the most relevant indicators behind the aforementioned outcomes are the amount -and to a lesser extent, quality- of sleep, a series of financial restrictions (including future consumption), limited mobility, and age. Other distinguishing regressors are the presence of migraines, labour conditions, family characteristics (e.g. marriage and/or parenthood), and the neighbourhood environment. As expected, the results show that Random Forest outperforms alternative models.

It is worth mentioning that the scope of this paper, as mentioned, is to find the most relevant features behind psychological distress. Therefore, no causal links shall be established.

The present paper is articulated as follows. First, in Part II., an insight on psychological distress will be presented, along with the literature review on both psychological findings and machine learning techniques used in the cited papers. Moreover, in Part III. the NHIS database shall be explored, as well as a statistical analysis for the purpose will be provided. Part IV. shall discuss the strategic instruments used, next to their respective machine learning background. Results will be illustrated on Part V., and concluding remarks shall be formulated in Part VI.. Additionally, the Appendix contains other descriptive features and relevant model findings.

II. Theoretical Framework

The literature on psychological distress is anything but scarce. Despite the wide research, a study of the papers that are related with the available features in the database, along with their respective implemented methodologies, can be carried out. In this Part, a brief introduction on psychological distress will be provided first, followed by a review of the most recent studies.

A) Behind Psychological Distress

In order to pursue an analysis on the matter, an insight of the phenomenon must be held, so as to grasp a better understanding of what is referred to as '*psychological distress*'. For this purpose, Drapeau et al., 2012 provides a distinguishing array of concepts, remarks and analysis on the subject.

Considering its clinical features, "psychological distress is largely defined as a state of emotional suffering characterised by symptoms of depression (e.g., lost interest; sadness; hopelessness) and anxiety (e.g., restlessness; feeling tense) (Mirowsky and Ross 2002). These symptoms may be tied in with somatic symptoms (e.g., insomnia; headaches; lack of energy)..."[1]

Moreover, "Psychological distress is usually described as a non-specific mental health problem (Dohrenwend and Dohrenwend 1982) . Yet, according to Wheaton (Wheaton 2007) , this lack of specificity should be qualified since psychological distress is clearly characterized by depression and anxiety symptoms." [2]

Compared to more significant disorders, the following can be said: "In effect, the scales used to assess psychological distress, depression disorders and general anxiety disorder have several items in common. Thus, although psychological distress and these psychiatric disorders are distinct phenomena, they are not entirely independent of each other (Payton 2009)." [3]

Psychological distress has been measured as the mean response to six items developed by Kessler et al., 2002 to assess symptoms of non-specific psychological distress. Respondents were asked to indicate how often in the past 30 days they felt around the following dimensions: (I)

[1] Drapeau et al., 2012
[2] *Ibidem*
[3] *Ibidem*

nervous, (II) restless or fidgety, (III) so sad nothing could cheer them up, (IV) hopeless, (V) everything was an effort, and (VI) worthless. Response categories ranged from 'never' {0} to 'very often' {4}.

B) Literature Review

A relationship between amount of sleep and health levels has been established in Sithey et al., 2017, using information from Bhutan's "Gross National Happiness Study". Having applied logistic regression, it was found that insufficient or excessive hours of sleep in the long run led to, among other impacts, a lower self-report of happiness. Furthermore, the association between both variables have been also analysed by Pearson, 2006, though this time by the opposite approach: higher self-reported levels of anxiety or depression lead to a lower quality of sleep. With data from the 2002 NHIS, this was concluded after using once again logistic regression.

For more on the relation between sleep and psychological distress, in Hill et al., 2009 the lack of sleep-quality is identified as a byproduct of neighbourhood unrest. Once the weighted ordinary least squares regression was performed, the authors conclude that neighbourhood disorder contributes to psychological distress by eroding protective resources.

Concerning physical characteristics, in the literature it has been widely shown that there is a the close relation between freedom of movement, particularly sports, and psychological conditions. Bragazzi et al., 2019 analyses systemic sclerosis patients, and find that they tend to feel more sadness than people without sclerosis.

Another relevant aspect involving physical traits has been found in the lack of sight: unlike blind from birth, people who lose progressively and/or partially their sight present higher symptoms of anxiety or depression. In De Leo et al., 1999, loss of sight is compared to loss of audition for patients that have committed suicide, and highlight that distress is produced merely on the fear of loss of sight.

Moreover, when it comes to budget concerns, Shapiro and Burchell, 2012 find that financial anxiety is a byproduct of financial mismanagement and lack of information and education, using survey and experimental data. Thus, highly indebted people tend to accumulate anxiety due to the financial whirlpool of constantly paying debts.

Also related to income but from another approach, Kessler and Neighbors, 1986 show that the relation between ethnics and socioeconomic status is not linear when understanding psychological distress, and that instead both regressors have an interactive dynamics. Having gathered information from a series of surveys in the US, the authors perform a linear regression with interactions, and find that there is effectively and asymmetry between low-income White and low-income Black populations. Similar findings were described in Kessler, 1979, where both linear and logistic regression were used to identify the linkage between stress and a series of indicators, such as social status, ethnics, gender and marital status.

A connection between financial constraints, unemployment and psychological distress has been revisited in Whelan, 1992. Parting from the General Health Questionnaire scoring criteria, the authors analyse for observations in Ireland a comparison between different indicators such as unemployment, lifestyle quality reduction, housing deprivation and income. The results of the regressions give sense to the notion that income affects psychological well-being indirectly via subjectively appraised financial strain.

Regarding the elder age, the dilemma of retirement is explored in Wels, 2018. Using Belgium's SHARE database, the author inquires over the effects of reduced labour in the late stage of life, comparing it with retirement. The results suggest that working partially reports more happiness than abandoning work due to retirement.

Mental health can also be traced back to gender issues. On the one hand, with the concern that women are twice as likely to be affected compared to men, Jothi et al., 2020 explore the Generalised Anxiety Disorder (GAD) in Malaysian women using a random forest technique, and conclude the more tired than usual a person is, and the higher the desires to commit suicide, and the less the interest in people and things, then the more significant the symptoms that s/he has developed GAD.

Additionally, numerous works have been carried out concerning the LGBTQ+ population. According to Gonzales et al., 2016, where a comparison of health and health risk factors was carried out between homosexual (both male and female) and bisexual adults, and heterosexual adults. Having implemented the logistic regression methodology using NHIS data, the findings related to psychological distress reflects that it is more present in non-heterosexual US adults.

Furthermore, there has been significant research between marriage and psychological distress. On the subject of divorcees, Hope et al., 1999 calculate distributions of Malaise Inventory using longitudinal data in the United Kingdom. The authors find that divorced people who never remarried presented increased distress, with both acute and longer-term components moderated by secondary factors such as childcare and declining socioeconomic status. As for widowers, in Umberson et al., 1992 a similar data strategy is pursued: using longitudinal data, the investigation seeks to explain the relation between depression and the current marital status by sex. The results of the linear regression suggest that widowers present higher levels of depression than married couples when affected by strains. Overall, Kessler and Essex, 1982 analyse the marital status and the effects on depression. Having implemented a linear regression, results similar to the aforementioned were obtained, explained by the coping mechanism of married couples, for both emotional and economic strains.

Although not identified clearly as a causal effect, Lawrence et al., 2011 present the high correlation between anxiety and smoking: patients with psychological disorder tend to smoke a lot, and find it harder to quit smoking.

Last but not least, a thorough analysis is provided by Hullam et al., 2019 on the subject of depression by contemplating a number of multicausal indicators, used to explore interactions and synergistic effects among the variables, which include social indicators, physical traits and daily activities, among other. Belonging to the neural network framework, they conclude that the environment operates by body weight, physical activity, parental depression and neurosis.

III. Data Analysis

In this Part, a data analysis shall be provided, as well as a brief comment on the database, and key relevant indicators.

A) The National Health Interview Survey (NHIS)

The NHIS ("National Health Interview Survey"[4]) is one of the major data collection programs of the NCHS ("National Center for Health Statistics"). Its main objective is to monitor the health of the United States population through the collection and analysis of data on a broad range of health topics.

As for the sampling design, the NHIS is carried out each year, and consists in a cross-sectional household interview survey. It uses geographically clustered sampling techniques to select the sample of dwelling units for the NHIS, across the country.

B) Database Selection

For the purpose of this analysis, the Sample Adult database has been selected, which collects information from population aged 18 or more. As for the years of information, the 2015, 2016, 2017 and 2018 databases were used.[5]

As it is well known, there is a series of challenges when dealing with latent variables on the subject of personal happiness and welfare. However, in the NHIS an attempt to capture these variables is performed, following the criteria presented in Part II., Section A).

Among the wide range of variables identified in the database, its variety can be summed in the following groups:

- **General features**: sex, age, weight, legal/civil status, ethnics & gender identity;
- **Lifestyle & consumption**: sport frequency, use of internet, hours of sleep, consumption of substances (alcohol, tobacco, drugs and other substances);
- **Socioeconomic indicators**: trust and reliance in neighbours, and budget related indicators such as income destined for health and standard of living;
- **Labour condition**: years in work, class of worker, earnings frequency, amount of jobs, among others;
- **Physical health state**: senses condition (such as sight and hearing), diseases (cardiorespiratory, cerebral, muscular, articulatory, immune), operations, duration of disease, duration of treatment, among others;
- **Psycho-emotional health state**: mental disorders, and psychological distress indicators

[4] https://www.cdc.gov/nchs/nhis/index.htm

[5] A couple of reasons are behind this decision. In the first place, to avoid relying on a single year, where a particular event might have influenced the overall survey report. However, and most importantly, given the highly unbalanced frequency of the desired outcomes, it was deemed convenient to add more observations.

(such as nervousness, anxiety, sadness, exhaustion, hopelessness); and

- **Health control**: medical control periodicity & health treatment.

C) Data Review

In this Section, a comment on the preparation process prior to the model elaboration will be delivered.

The first step implemented was the outer join of the 2015 - 2018 databases. The second measure taken was to remove the variables whose observations were a third (33%) at most 'missing values'. Thirdly, a group of variables underwent modifications & ordering, either because of their cardinal nature, or because they were assigned specific class values - the latter case driven by the criteria adopted for missing values[6], resolved by the survey designers.[7] Furthermore, the remaining missing values were all replaced by the mean.[8]

The next step was to generate new databases -copied from the original dataframe- that contained only one of the six proposed dependent variables for each of the newly created bases. Each of these databases would serve as the information set from where the corresponding models would be built upon. This decision was taken in order to reduce the effects of collinearity

[6]The question of missing values arises because a great number of features did not count with missing values given they were substituted by three additional classes: "Refused", "Not ascertained" or "Don't know". Thus, although the theory behind some questions was binomial, in practice the data was 'multinomial'.

This is a case of what is known as 'classification/label noise´, where observations are assigned a class different from the actual one. The higher the noise, the higher the probability of achieving a poorer accuracy, because the models are trained by these 'misclassified´ labels. This issue will be addressed once again in Part IV., Section A), Item 2-.

[7]A number of cases could be mentioned. For example, for variables related to physical activities, answers {996} were replaced to {-1} ('*MODNO*, '*VIGNO*' and '*STRNGNO*'), or from {0} to {6} ('*MODTP*', '*VIGTP*' and '*STRNGTP*'); in both cases, categories "'Never" and "Unable to do this type activity" were closed, allowing a more suitable scale. In the same spirit, for the responses "Not ascertained" or "Don't know" for questions related to financial worries and the neighbourhood -among others- whose range theoretic range of answers was {1; 2; 3; 4; 5}, their values were substituted by the average, hence replacing {8} and {9} for {2,5} in the following cases: '*ASINHELP*', '*ASINCNTO*', '*ASINTRU*', '*ASIRETR*', '*ASIMEDC*', '*ASISTLV*', '*ASICNHC*', '*ASICCOLL*', '*ASINBILL*', '*ASIHCST*' and '*ASICCMP*'. Analogously, this decision was also applied for dichotomous variables who also counted with the aforementioned categories.

Moreover, for the variables '*OCCUPN2*' and '*INDUSTRN2*' -both reflect the occupations and industrial sector of the observations-, replacement by the median lacked of sense because of their cardinal characteristic, so they were converted to l_i variables for each l class in $i = \{$'OCCUPN2', 'INDUSTRN2'$\}$. The original variables were later eliminated.

Last but not least, the demographic statistics that were inherently cardinal values -such as ethnics, marital status and gender- were used as a base for arranging new dichotomous variables. Hence, additional features were added responding to the 'Yes'/'No' logic for ethincs ('White - Non-white' (I), and 'Hispanic' - 'Non-Hispanic' (II)), martial status ('[Currently] Married' - '[Not Currently] Married') and gender ('Heterosexual' - 'Not-Heterosexual'). Also, the variable 'Minority' was created using sex, ethnics, and gender, resulting in $2^3 = 8$ classes.

[8]Considering continuous variables were scarce, it was preferred over the median or the mode.

among other dependent variables.

As for the relevant categories, the variables of interest were limited to the answers "All of the time" {1}, "Most of the time" {2}, "Some of the time" {3}, "A little of the time" {4} and "None of the time" {5}.[9]

Furthermore, in order to capture an index of psychological distress that would gather the information of all six relevant indicators, a simple average was calculated among them per observations, thus creating the new variable.Having rounded each average to the integer, each observation was assigned a class by replicating the former criterion.

Table 1 portrays the tentative outcome variables prior to modelling.

Table 1: Frequency per Dependent Variable

	Variable	**None/Little**	**%**	**Some/Most/Always**	**%**
I	*Sadness*	98.609	88,07%	13.354	11,93%
II	*Nervousness*	90.453	80,79%	21.510	19,21%
III	*Restlessness*	88.810	79,32%	23.153	20,68%
IV	*Hopelessness*	103.665	92,59%	8.298	7,41%
V	*Exhaustion*	92.373	82,50%	19.590	17,50%
VI	*Worthless*	105.152	93,92%	6.811	6,08%
VII	*Psychological Distress*	99.643	89,00%	12.320	11,00%

Because of their low frequency (under 10%) as stated in Table 1, variables *IV* and *VI* were not contemplated for modelling. As for the remaining variables, the unbalanced nature of the class frequency has been addressed by oversampling each training base[10] using the 'Borderline-SMOTE2' technique.[11] Additionally, this increase in the database has brought more computational costs, thus an undersampling on the "healthy" observations was performed on the training

[9]The observations that had one of the following options for the dependent variables were eliminated: "Refused", "Not ascertained" or "Don't know".

[10]Chawla, 2009 presents the problem of imbalanced datasets and its consequences on relevant indicators such as accuracy.

[11]Following Han et al., 2005, where and extension of the 'Synthetic minority oversampling technique' (SMOTE) is explored, an alternative to the oversampling issue suggested by the author is 'Borderline-SMOTE2'. This approach creates synthetic observations based on the nearest neighbours *within* the class that happen to be close enough to observations of the opposite class. Thus, oversampling is performed nearer the boundary with the opposite class rather than at a greater distance, allowing a significant reduction of misclassification at the border. For more literature on SMOTE, a detailed analysis is provided in Fernandez et al., 2018.

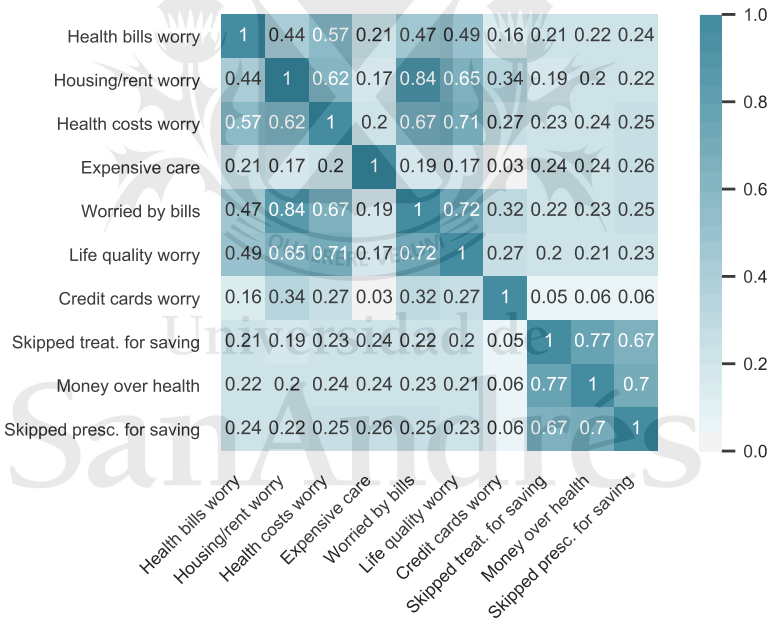
data.[12] The parameters selected per technique are addressed in Part IV., Section D).

D) Exploratory Data Analysis

Provided the structure and the identification of the main variables found in the database, a brief data description of some relevant indicators will be shown below.

As for variable correlations, a series of highly correlated predictors are presented in the following charts, each of which are grouped respective under common topica: Figure 1 depicts income issues related to health, while Figure 2 conveys physical limitations, and Figure 3 shows sleeping conditions.

Figure 1: Correlation Matrix for Several Indicators Related with Income and Health

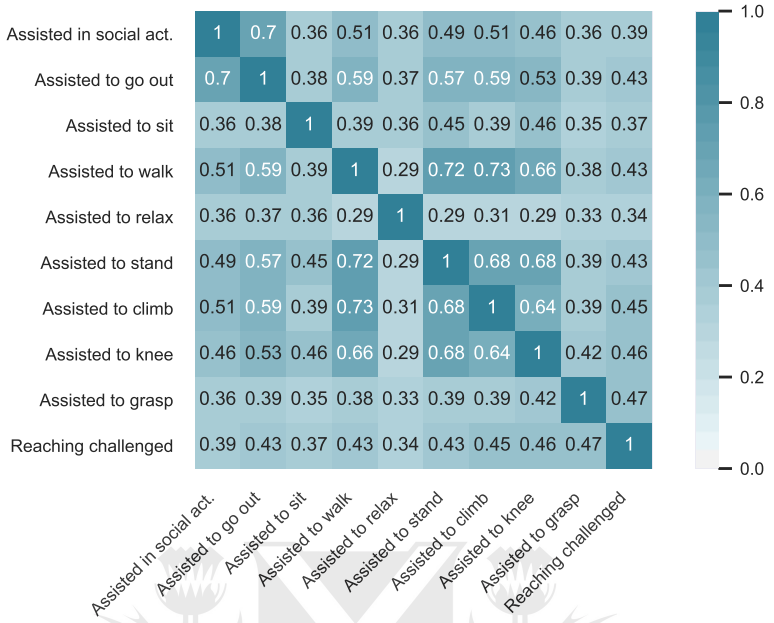


Source: NHIS

On the one hand, Figure 1 shows a fair to high correlation between "Health bills worry", "Housing/rent worry", "Health costs worry", "Worried by bills" and "Life quality worry" (I). It also presents an even more significant positive correlation between "Skipped treatment for saving","Saved money over health" and "Skipped prescription for saving" (II).

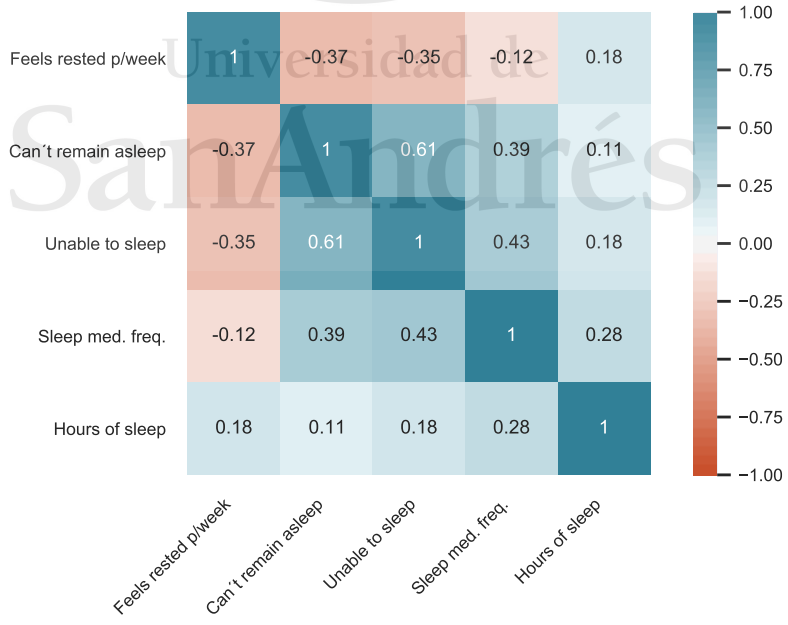
[12]This practice is pursued in, for instance, Rahman and Davis, 2013, where the SMOTE approach is accompanied by undersampling in the training data as well. However, the question of oversampling and undersampling is explored in Fernandez et al., 2018, where the trade-off between computational costs and loss of modelling information is stated.

Figure 2: Correlation Matrix for Several Indicators Related with Physical Limitations



Source: NHIS

Figure 3: Correlation Matrix for Several Indicators Related with Sleeping Conditions



Source: NHIS

On the other hand, Figure 2 portrays an overall high positive correlation between its variables, particularly for these two groups: "[Requires] assistance in social events" and "[Requires]

assistance to go out" (I); and "[Requires] assistance to walk", "[Requires] assistance to stand", "[Requires] assistance to climb" and "[Requires] assistance to knee" (II).

Lastly, Figure 3 depicts the relationship between a number of variables related to the sleeping conditions per observation. It strikes immediately the moderate inverse relation between the self-perception of personal rest, and the inability to fall or remain asleep. About these last features, they present a fairly positive correlation.

Given the aforementioned description, an index was created for each group of variables that presented elevated correlations: a simple average was calculated per subgroup.[13] This was carried out in order to reduce the effects of collinearity and to incorporate the most possible variance.[14]

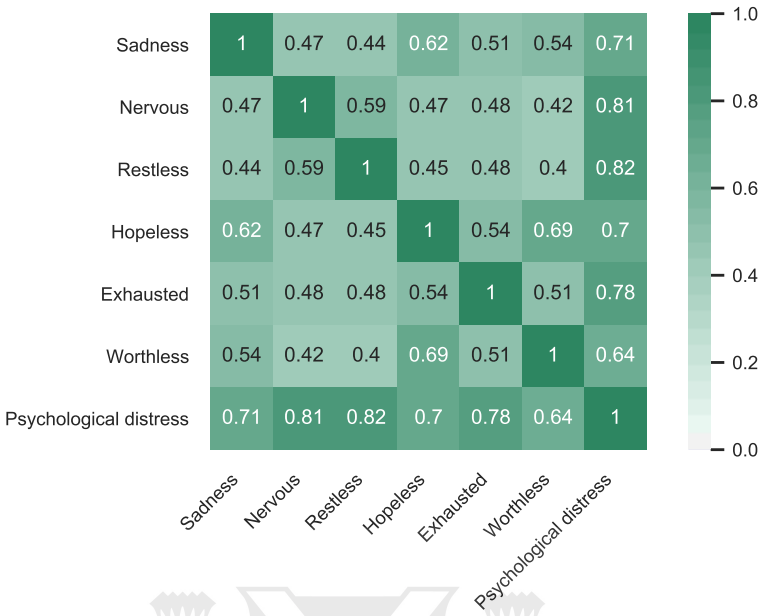
As for the possible outcomes, Figure 4 illustrates the six main indicators defined previously and their correlation.

[13]Thus, the following indexes were created:

- "*Need to save*": ("Skipped treatment for saving", "[Chose] money over medical treatment", and "Skipped prescriptions for saving");
- "*General cost worry*" ("Worried by Housing/rent", "Worried by health costs", "Worried by bills", "Worried by quality of life", and "Worried by health bills");
- "*Walk/stand/climb/knee*" ('Assisted to walk", "Assisted to stand", "Assisted to climb", and "Assisted to knee");
- "*Fall/remain asleep*" ("Cannot remain asleep" and "Unable to sleep");

[14]At first, the principal component analysis was deemed as the main tool to solve the reduction in dimensionality. However, provided the highly discrete classes per variable, its use resulted ineffective for this purpose.

Figure 4: Correlation Matrix for the Possible Outcomes



Source: NHIS

From Figure 4, it can be seen that the suggested variables present a significantly high relation between them. Although some features may not necessarily be present because of others, it seems quite remarkable the fact that people who perceive themselves as sad also tend to declare feeling anxious, depressed and/or fatigued. Given this context, the seventh variable associated with the constructed index gains significance as a measure that comprises the information of the precedent variables.

In order to provide descriptive statistics on mental health's common denominator, Figures 23 and 24 (found in the Appendix) depict the standardised means of psychological distress conditioned on a selection of demographic factors.

Figure 23 accounts for sex, ethnics and gender; it is observed that psychological distress affects broadly LGBTQ+ population, particularly in women; on the opposite end, heterosexual men seem to suffer less on average from mental health issues. As for ethnics, there does not seem to be a clear difference between white and non-white communities.

Meanwhile, on Figure 24 the focus is set on age ranges, each interval grouping 5 years (with the exception of the ends). It can be seen that psychological distress is highly present during the 20s, as well as in the transitioning from adulthood (45) to an advanced age (60). On the contrary, the

older the population gets, there is a significant average decrease in the self-perceived indicators.

IV. Empirical Strategy

In order to perform the classification of the formerly introduced variables, in this Part the adopted Machine Learning technique will be explained. Additionally, other possible predictive models shall be explored as alternatives, along with a concise description respectively. Finally, the main commands in Python as well as the main assumptions held will be presented.

A) Classification and Regression Trees

Given the nature of the problem in discussion, it is clear that there may be a number of determinants behind psychological distress; however, these might weigh more or less according to former presence of other relevant variables. Thus, it is fundamental to bear in mind the non-linear characteristics of these indicators before pursuing the analysis.

1- Classification Trees and the Bias-Variance Trade-off

In this case, the methodology to be implemented will rely on Classification and Regression Trees (or CARTs), a technique that allows classification using a binary criterion (i.e. "Yes" and "No", or more formally, {1;-1}, respectively). Particularly, classification trees are useful when encountering qualitative data (Hastie et al., 2009).

Unlike linear models, where the prediction *f(X)* is based on a linear combination of features, CARTs models follow the form expressed in 1:

$$f(X) = \sum_{m=1}^{M} c_m \cdot 1_{(X \in R_m)} \tag{1}$$

where R_1, ..., R_m represent a partition of feature space.

Under this circumstance, the prediction is built considering the probability of occurrence per class of training observations in the region to which each observation belongs. Thus, not only is it relevant to identify the prediction in a particular node, but also the the class proportions among the training observations that fall into that region.

Classification trees are often referred to as a 'decision tree', where for each level (or node), depending on a certain value a variable may adopt, the model shall determine whether the outcome belongs or not to a class, given a certain criterion. Once this process is iterated to a certain amount of nodes, a decision path shall be established, resembling the concept of a tree.[15]

Hence, what immediately follows is the trade-off between bias and variance in CARTs, that is: the point where the gain in prediction of adding another node in the tree is more than offset by the loss of prediction due to changes in the data. This could also be thought as learning a model that results too specific (or "overfit") for the available data, but will roughly be able to deliver an equally precise prediction for different data (*caeteris paribus*). This problem in CARTs is known as 'cost complexity pruning' (James et al., 2013), which calculates the cost of reducing a node.

In algebraic terms, the formulae is presented in the following expression (2):

$$\sum_{m=1}^{|T|} \sum_{i:x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha \cdot |T| \tag{2}$$

where α ($\forall \alpha \in \mathbb{R}^+$) is the hyperparameter that governs the trade-off between tree size and its goodness of fit to the data. In other words, the addition of a node bears a cost (α): the gain in accuracy is partially offset by the cost of fitting the model.[16]

However, classification trees may suffer from high variance - that is, the fact of performing partitions randomly in the training data, and later testing the modelled tree in the out of sample data for each half, may lead to significant differences in results. This latter point is of utter relevance in the context of a vast number of features; it would suffice to make a slight modification in data that the results could vary considerably, leaving the task of prediction difficult to fulfil given highly overfit models.

[15] A more detailed explanation is provided in Hastie et al., 2009, where the partition of data space is explored thoroughly. In short, the algorithm selects the variable with the best fit (i.e. that correlates the most with the outcome) and later parts the data space for a certain value conditioned on the mean of the outcome for that subarea. Thus the first step for classification has been carried on: beyond that threshold, the algorithm will consider the condition as a good predictor ("Yes"). Conversely, for values below it, as a bad predictor ("No"). It then performs this step repeatedly within each of the subareas, until it reaches the number of iterations predefined.

[16] In an extreme case, for $\alpha = 0$, the penalisation is null, so there would not be any cost in gaining very high precision; on the contrary, for a value of $\alpha \to \infty$, the penalisation is such that there would not be any gain in incorporating a single extra node.

Therefore, a way to sort the problem of high variance is to build models that minimise the error by instead considering the *average error* of a number of trees. For a sole database, 'different' bases may be created using bootstrap sampling, process described in Item 2- below.

2- Bootstrap Aggregation -'Bagging'- and Random Trees

When dealing with overfitting, and its inherent issue of high variance, Bootstrap aggregation offers a tentative solution: by generating B ($\forall B \in \mathbb{N}$) Bootstrap subsamples within the training data, the algorithm builds B trees per subsample, and then tests each model on the remaining deselected observations. It is worth mentioning that the Bootstrap samples are drawn repeatedly with replacement - that is, an observation that was already selected might be selected again in the same sample.

For the classification instance, the overall prediction of the B trees for each tested observation is defined by the 'majority vote' criterion. In this sense, bagging averages the variance of each tree by choosing the most commonly occurring class among the B predictions[17]. Equation 3 illustrates the bagging estimation.[18]

$$\hat{f}_{\text{bag}}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(x) \tag{3}$$

However, and despite its usefulness in tackling the overfitting problem, bagging will remain a strong classifier if and only if there are no significant variables that persistently emerge as relevant in the B trees; otherwise, the trees would end highly correlated, which still leaves space for high variance.

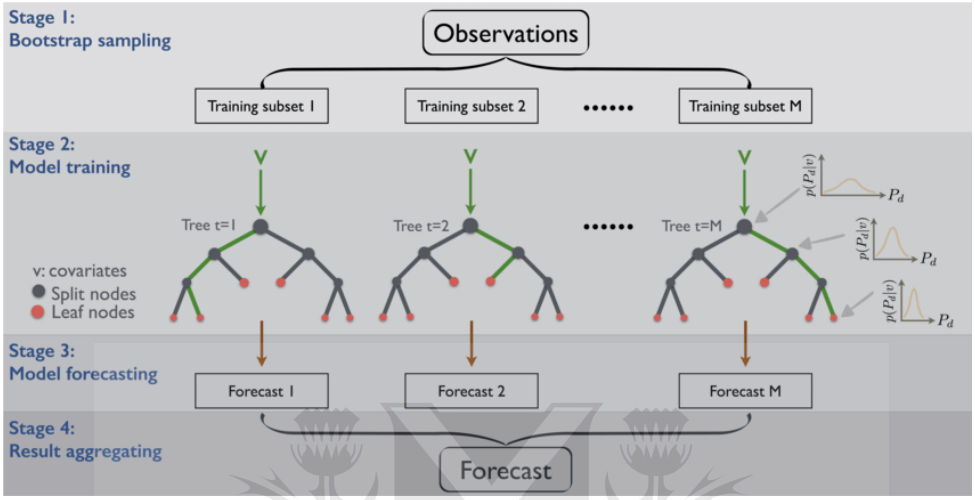
Therefore, a way to reduce this possibility is to alter the amount of available predictors when building each node per tree. More formally, given a total of p variables, only $m < p$ variables will be available to build each tree ($\forall m,\ p \in \mathbb{N}$). The fact that these m predictors are selected randomly reduces significantly the chance of trees being highly correlated. An illustration of the process can be found in HE et al., 2016, as shown in Figure 5.

Moreover, a distinctive feature of random forests -inherited from bagging- is the out-of-bag (OOB) samples, i.e. the samples that were not selected by Bootstrap in the training data to

[17] Given N observations, each of which has a variance of σ^2, then the variance of the mean will be σ^2/N.
[18] James et al., 2013

learn the tree: rather than minimising the Cross-validation error in a first instance, random forests may profit from the OOB samples for the purpose, thus allowing the estimators to be fitted in one sequence.

Figure 5: Representation of Random Forest Classifier



Source: HE et al., 2016

Alternative modelling strategies involving trees such as Boosting have been discouraged by a number of reasons, explained in Breiman, 2001, such as the higher computational cost, the unclear gain in prediction and yet the significant weakness in presence of noise[19], or sheerly explanatory preferences, as random forests are more intuitive.

B) Alternative Strategies

It would be unwise to limit the prediction analysis to a sole model, either because of the theoretical framework, or due to the nature of the data. Bearing this into consideration, additional models have been contemplated against the proposed model.

1- Logistic Regression

Another model worth contemplating is the logistic regression, whose classifier relies on the *odds-ratio* $\frac{p}{1-p}$, where $p = \frac{e^z}{1-e^z}$ (for $z = X\beta$). Essentially, the logistic regression assigns bivariate values ($\{0;1\}$) when classifying. Equation 4 presents the modelled regression:

[19] This point is explored in detail in Dietterich, 2001 and Frénay and Verleysen, 2014

$$\Pr(G = k \mid X = x) = \frac{\exp\left(\beta_{k0} + \beta_k^T x\right)}{1 + \sum_{\ell=1}^{K-1} \exp\left(\beta_{\ell 0} + \beta_\ell^T x\right)} \tag{4}$$

It is important to highlight the fact that Equation 4 contemplates every available variable to build the prediction. *A priori* this is useful to diminish the bias, thus increasing significantly the precision for the training data. However, as for the testing data, the model may result too specific to those regressors; hence any change in the data may imply a pronounced decrease in the precision due to the resulting high variance.

In order to compensate this trade-off, regularisation techniques result useful. At the moment of identifying the according type of penalisation -namely LASSO or Ridge-, a Solomonic criteria has been proposed, that contemplates both notions: *elastic-net* (Zou and Hastie, 2005). Among its most significant advantages, "The elastic-net selects variables like the [LASSO], and shrinks together the coefficients of correlated predictors like [R]idge." (James et al., 2013) Given the highly correlated regressors for the database stated in Part III. with an unseemly way to apply principal components analysis, this penalisation may be key to deal with this situation.

Equation 5 presents the problem and how elastic-net regularisation operates:

$$\tilde{\beta} = \underset{\beta}{argmin} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \left(\alpha\beta_j^2 + (1-\alpha)\,|\beta_j| \right) \right\} \tag{5}$$

As the second term of equation 5 shows, the algorithm shall favour one regularisation coefficient over the other conditioned on the value of the hyperparameter $\alpha \in [0, 1]$.[20] For the purpose of this analysis, the value of α has been selected by Cross-validation.

2- *K*-Nearest Neighbours (KNN)

Unlike the models presented in Section A) and Item 1-, this method is an non-parametric approach for prediction (James et al., 2013). It essentially determines whether an observation belongs to a certain class over another by contemplating the 'nearest' surrounding observations: having identified these 'neighbours', the algorithm assigns the class by majority vote. In other words, the observation will belong to the class that bears the highest frequency for the selected

[20] In Zhou et al., 2014, it is shown that this type of regularisation is quite similar and thus reducible to the support vector machine linear method.

'voters'.

K-Nearest Neighbours is especially effective when variable relations are hard to establish by linear or quadratic means, provided its non-parametric nature. The latter is also deemed as a distinct feature of this method, because no model assumptions are made, thus reducing the threat of selecting unsuitable models for a given database.

As for the bias-variance trade-off, the complexity shall be defined by the selected value for parameter k ($\forall k \in \mathbb{N}$) using Cross-validation; the lower the integer, the lower the bias, at the cost of a high variance. The converse holds analogously.[21]

Equation 6 presents the estimation of the conditional probability for class j as the fraction of points in $\mathcal{N}_0$ whose response values equal j:

$$\Pr\left(Y = j \mid X = x_0\right) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I\left(y_i = j\right) \tag{6}$$

C) Relevant Indicators

In the present Section, the key indicators to be considered in the analyses will be highlighted.

1- Mean Squared Error (MSE)

To begin with, the Mean Square Error (or MSE) is a vital indicator. It depicts the 'distance' between the predicted observation from the real one. For each i observation, MSE sums each squared error of prediction, hence determining the overall performance of the model by penalising high deviations and easing low deviations, and adding them up. Equation 7 reflects in algebraic terms this notion.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \hat{f}\left(x_i\right)\right)^2 = \frac{1}{n} \cdot RSS \tag{7}$$

From Equation 7 it is worth noticing that MSE is nothing else than the average of the Residual Sum of Squares (RSS).

[21] In an extreme case, if $k = 0$, the only relevant point of reference is the observation itself. On the other hand, for $k = N$, then there would be N nearest neighbours, i.e. the entire set.

Therefore, the rule of decision when defining whether a model is best than another is the minimisation of the MSE indicator; the model with the least MSE reflects that its loss of information is minimal.

2- Classification Error Indicators

Provided the binary splits that characterises the classification analysis of Random Forests, an alternative to the RSS approach can be found in the 'classification error rate'. The classification error rate is the fraction of the training observations in a given region that do not belong to the most commonly occurring class (James et al., 2013).

Equation 9 illustrates the classification error rate; it is important to bear in mind the role of proportion $\hat{p}_{mk}$[22] in the definition of this error.

$$E = 1 - \max_{k} \left(\hat{p}_{mk}\right) \text{[23]} \tag{9}$$

Although misclassification error may be the most straightforward method of identifying error, it fails to be the most effective. Due to this, two other measures of node impurity are the Gini index, and Cross-entropy -or deviance-, both conveyed below.

$$Gini = \sum_{k \neq k'} \hat{p}_{mk}\hat{p}_{mk'} = \sum_{k=1}^{K} \hat{p}_{mk}\left(1 - \hat{p}_{mk}\right) \tag{11}$$

$$Cross\ entropy = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk} \tag{12}$$

[22] where $\hat{p}_{mk}$ follows expression:

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I\left(y_i = k\right) \tag{8}$$

for a node m, representing a region R_m with N_m observations, and with k being the class.

[20] Equation 9 relies on the classification of observations in node m to class $k(m) = \underset{k}{argmax}\, \hat{p}_{mk}$. This is applied on the misclassification error, expressed below in Equation 10.

$$\frac{1}{N_m} \sum_{i \in R_m} I\left(y_i \neq k(m)\right) = 1 - \hat{p}_{mk(m)} \tag{10}$$

3- AUC, ROC Curve, and Other Relevant Indicators

The 'Area Under the Curve' (AUC) is an indicator that emerges from the ROC[24] curve. While the ROC curve portrays the graphic representation of goodness of an estimation from the higher or lower concavity of the curve -respectively-, the AUC is a value comprised between 0 and 1, both included. A prediction will be better over another as long as the AUC of the former model is higher than the latter. This is because the farther from an AUC of 0,5, the better the model will be; otherwise, a coin with a probability of 0,5 would result more suitable (i.e. the model would lack of value).

In terms of their relevance, ROC curves are useful for comparing different classifiers, since they take into account all possible thresholds (James et al., 2013). This is due to the fact that a ROC curve traces out two types of error as the threshold value varies for the posterior probability of default. The true positive rate is the 'sensitivity', i.e. the fraction of defaulters that are correctly identified, using a given threshold value. The false positive rate is the difference between the unit and the 'specificity', i.e. the fraction of non-defaulters that are classified incorrectly as defaulters, using that same threshold value. Both concepts shall be further illustrated in the list below.

For the four possible categories presented in the confusion matrix, each may me labelled as 'True positive' (TP) and 'False negative' (FN) for a certain prediction, and 'True negative' (TN) and 'False positive' (FP) for the other. Given these categories, the following calculi can be formulated.

Other Relevant Indicators

- **Precision**, or **Accuracy** = $\frac{TP+TN}{TP+FP+TN+FN}$

  This indicator measures the accuracy of the prediction; that is, how many true positives and true negatives were retrieved in the entire confusion matrix.

- **Sensitivity**, or **True Positive Rate**[25] = $\frac{TP}{TP+FN}$

  The true positive rate reflects the accurate prediction of True values over the whole prediction for positive values.

[24] 'Receiver Operating Characteristics'
[25] Also, 1 - type II error

- **False Positive Rate**[26] = 1 - *Specificity*[27] = $\frac{FP}{TN+FP}$

  The false negative rate reflects the prediction of false positive values over the whole prediction for negative values.

D) Python Parametrisation Specifications

Each of the four bases was split in 75% and 25% for the *training* data and the *testing* data, respectively. In other words, the variable selection process was performed with a sample representing 75% of the total available information, and its fit was tested in the remaining 25%.

Concerning the training data, as mentioned in Section C), oversampling and undersampling techniques were applied. The percentages to be used must bear, on the one hand, the higher representation of imbalanced values so as to construct a better predictive model, and on the other, the assertion that the balancing strategy implemented does not result in significant proportion alterations. With this under consideration, the respective percentages have been 0,30 for the former and 0,50 for the latter.[28]

As for the test base, given the nature of the data, for outcomes such as "Sadness", the chosen split might be somewhat challenging, as the individuals that have reported to feel at least some sadness represent approximately 12% of the total base, leaving it with eventual representation problems in the test data. Because of this, an identification of the participation of both classes per variable was carried out for both training and testing samples, as shown in the following Charts.

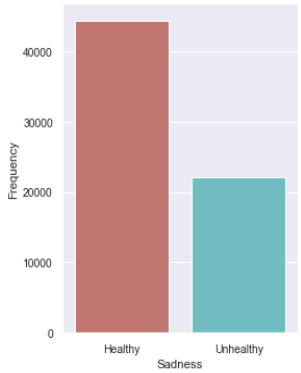
Figure 6: Outcome Decomposition in Training and Testing Samples for "Sadness"

[26] Also, type I error

[27] By construction, **Specificity** = $\frac{TN}{TN+FP}$

[28] This is to say that the training base saw its minority class ($Y[Y=1]$) incremented by 30% so as to be a third of the base, and later its majority class ($Y[Y=1]$) reduced until it reached being 50% of the minority class.

Training sample



Testing sample

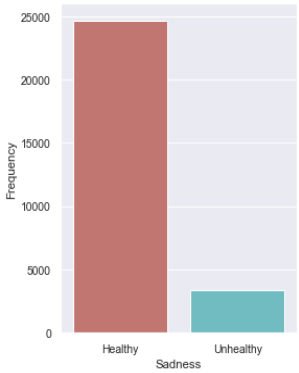
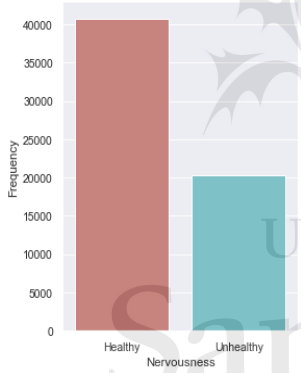


Figure 7: Outcome Decomposition in Training and Testing Samples for "Nervousness"

Training sample



Testing sample

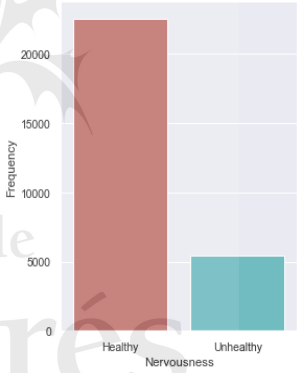
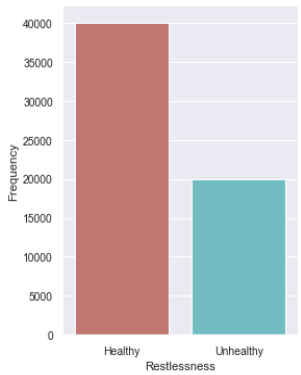


Figure 8: Outcome Decomposition in Training and Testing Samples for "Restelessness"

Training sample



Testing sample

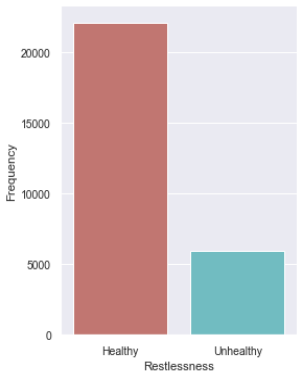


Figure 9: Outcome Decomposition in Training and Testing Samples for "Exhaustion"

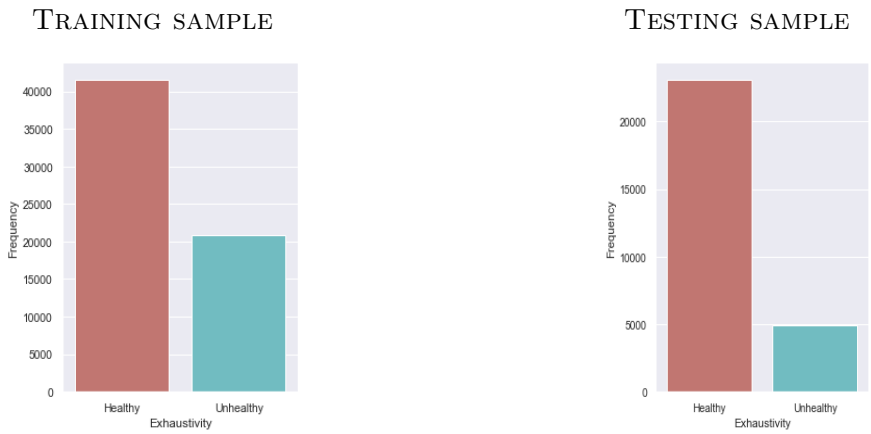
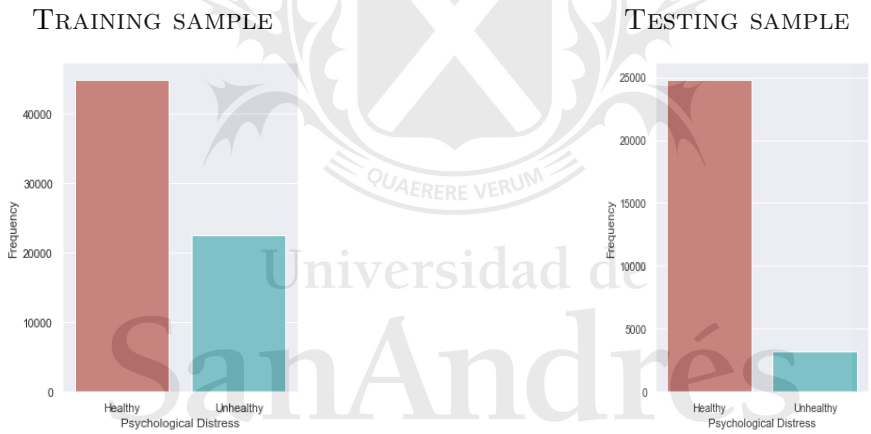


Figure 10: Outcome Decomposition in Training and Testing Samples for "Psychological Distress"



Thus, the investigator is inclined to think that the established percentages should not mean major threats for the findings.

Furthermore, the results of both oversampling and undersampling are shown in the Table 2 below.

Additionally, other required parameters that broadly affect the models in question is the K-fold for applying Cross-validation. This attribute was given the value of $K_{CV} = 5$. Additionally, and just as important, the seed required for the randomness was fixed at 42.

Moreover, the Python commands per model implemented were: (I) *RandomForestClassifier()* for Random Forests, (II) *KNeighborsClassifier()* for KNN with *GridSearchCV()* for Cross-

Table 2: Frequency per Dependent Variable, with Oversampling and Undesampling, in the Training Data

Variable	**None/Little**	**%**	**Some/Most/Always**	**%**
Sadness	44.370	66,67%	22.185	33,33%
Nervousness	40.732	66,67%	20.366	33,33%
Restlessness	40.034	66,67%	20.017	33,33%
Exhaustion	41.558	66,67%	20.779	33,33%
Psychological Distress	44.894	66,67%	22.447	33,33%

validation, and (III) *LogisticRegressionCV()* for Logistic regression, with 'elastic-net' regularisation. The parameters involved in the models presented in Sections A) and B) are defined as shown in Table 3:

Table 3: Parameter per Model

Model	**Parameter**	**Value**
Random Forest	N° of trees	5.000
	Criterion	'Gini'
	Minimum sample per split	250[29]
	Minimum sample per leaf	150
	Maximum depth i.e. nodes	5
	Maximum features (m)	$\sqrt{p}$[30]
	Maximum Bootstrap samples[31]	75%
	Class weight	'Balanced subsample'[32]
KNN	N° of neighbours (k)	{5; 7; 10; 15; 20; 30; 40; 50}[33]
	Type of distance	'Euclidean'
	Weight	'Distance'[34]
Logit LASSO	Elastic-net Penalty (L1/L2 ratio)[35]	{0; 0,1; 0,2; 0,3; 0,4; 0,5; 0,6; 0,7; 0,8; 0,9; 1}
	Tolerance	0,05
	Class weight	'Balanced'
	Maximum iterations	500

V. Results

In this Part, the results of the different models will be presented. Commencing with classification trees, the optimal trees for the selected outcomes shall be introduced. Next, a comparison will be made against the proposed alternative models (logit-LASSO and K-nearest neighbours)).

A) Classification using Random Forests

Because forests include a set of tree estimators, it would be unsuitable to present the entire forest. Instead, a tree per outcome will be shown below as a suggestion in Figures 11 - 15.

B) Model Comparison with Key Indicators

In this Section, the confusion matrices for the three models shall be presented: random trees, logit distribution, and K-nearest neighbours. Furthermore, relevant indicators such as MSE, AUC, ROC Curve, as well as accuracy, sensitivity and specificity.

[29] Approximately 1% of the class in the training base.

[30] Where p is the number of features in the database.

[31] The number of samples to draw to train each base estimator (i.e. decision tree) , as a percent of the training base.

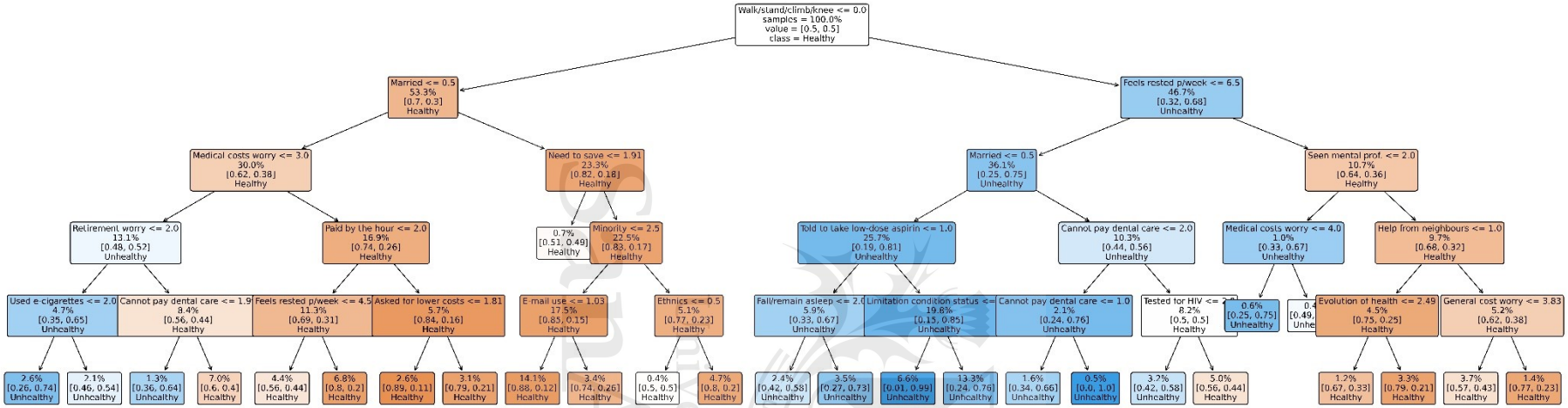
[32] Calculates the weight within each Bootstrap iteration.

[33] The scoring used was the negative MSE.

[34] As opposed to 'uniform', this mode favours neighbours that are closer and reduce the value of farther observations.

[35] The parameter was calculated by K-fold Cross-validation

FIGURE 11: CLASSIFICATION TREE FOR "SADNESS"



Source: NHIS

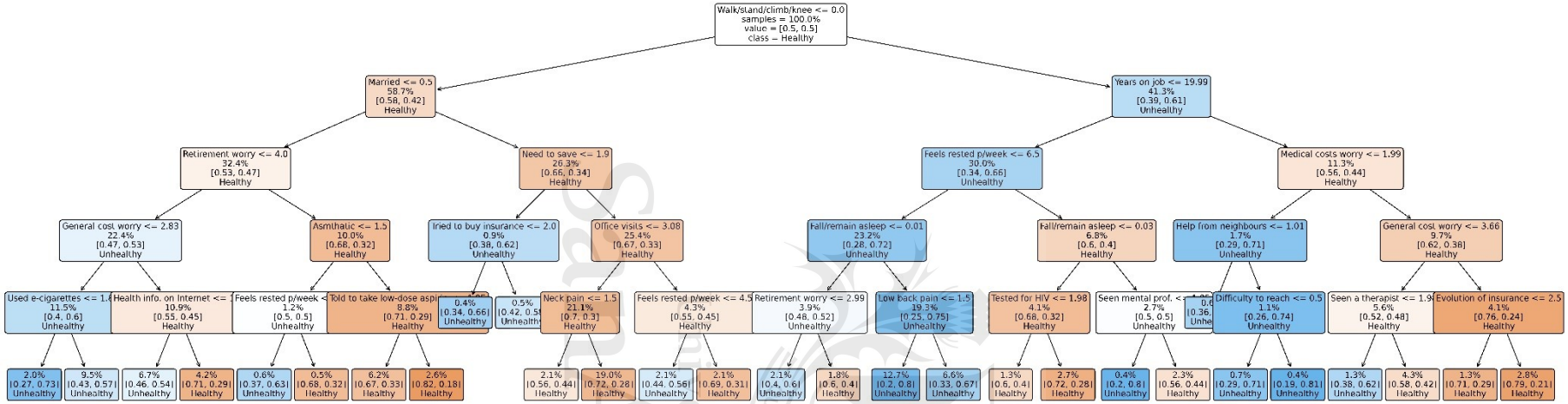
To begin with, Figure 11 presents the classification tree for outcome 'Sadness'. With its first partition defined by the ability to move legs -namely to walk, to stand, to climb or to knee-, the algorithm predicts that for all those individuals to have reported 'Not at all difficult' (0) present a minor probability to feel sadness.

For these, the next relevant feature is the fact of being married or not at the moment of the interview. According to the algorithm, being married offers a minor probability to report sadness, while the opposite occurs for divorcees, widowers, single, or living with the partner, among other categories. For both cases, the continuing nodes involve monetary aspects: worries about medical costs, and the need to postpone or refuse treatment in order to save money.

On the contrary, for individuals that had presented some physical constraint concerning the leg movement, what follows is the quality of the sleep, i.e. if they feel rested when they awake. For those who do feel rested, the next conditioning if whether they are married or not.

Other indicators that appear towards the end are the modality of work payment (that is if they are paid per hour), worries about the future, being white or not, and the help received from the neighbours.

FIGURE 12: CLASSIFICATION TREE FOR "NERVOUSNESS"

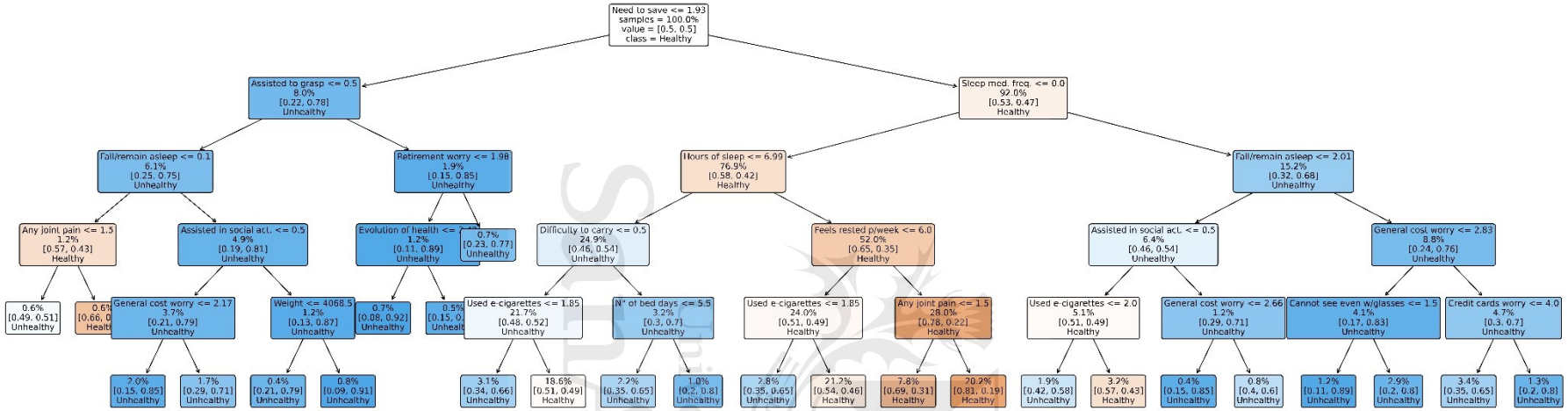


Source: NHIS

In second place, Figure 12 illustrates the classification tree for outcome 'Nervousness'. A number of common characteristics can be found with Figure 11, such as the initial node related with leg movement, or the fact of being married or not.

However, some differences emerge for example on the years on the job, where people who have job stability in the same place seem to feel less nervous than people with less permanent jobs. Furthermore, another factor fairly relevant is if the observations reported to suffer from asthma or not.

FIGURE 13: CLASSIFICATION TREE FOR "RESTLESSNESS"



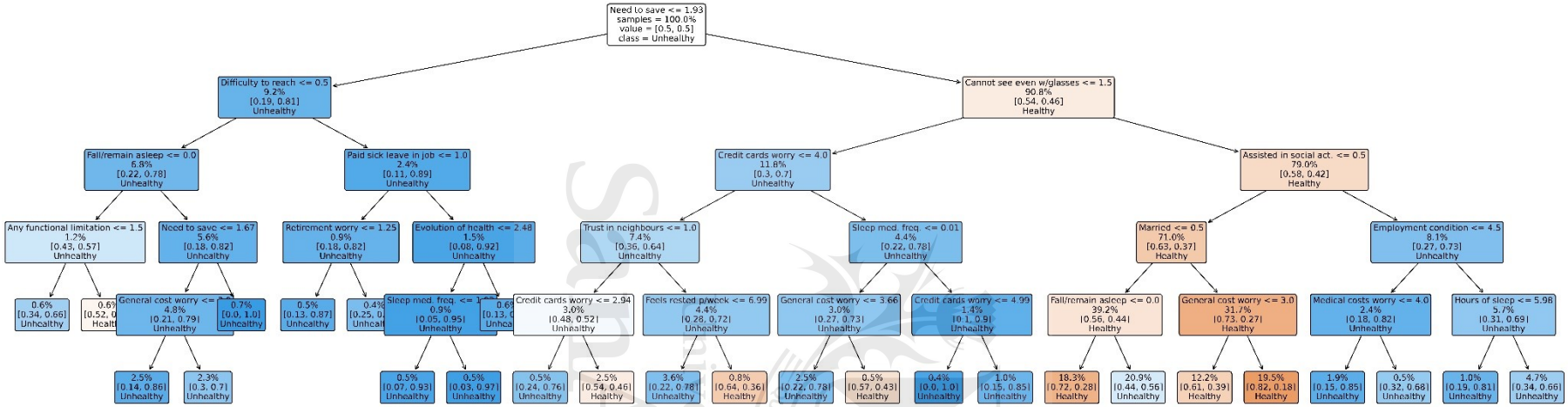
Source: NHIS

Morever, Figure 13 conveys the classification tree for outcome 'Restlessness', whose first partition is involved in economic reasons: the need to save money (over medicine), where individuals had to opt between following a medical prescription or treatment or even overall health, and saving that money.

For those who chose the latter over the former, the following condition involves physical mobility -the ability to grasp small objects without special equipment. Those that saved money and had this condition tended to report feeling restless more often. Meanwhile, individuals that did not face the budget trade-off, the next relevant feature concerns the frequency of consumption of medicine to sleep: for those that consumed at least once a week were assigned a higher probability.

Compared with the other outcomes, the share of sleep duration and quality seem to have a higher share than for people perceiving themselves as sad or nervous; other relevant indicators involve costs (both present and future) and pain in any joint of the body.

Figure 14: Classification Tree for "Exhaustion"

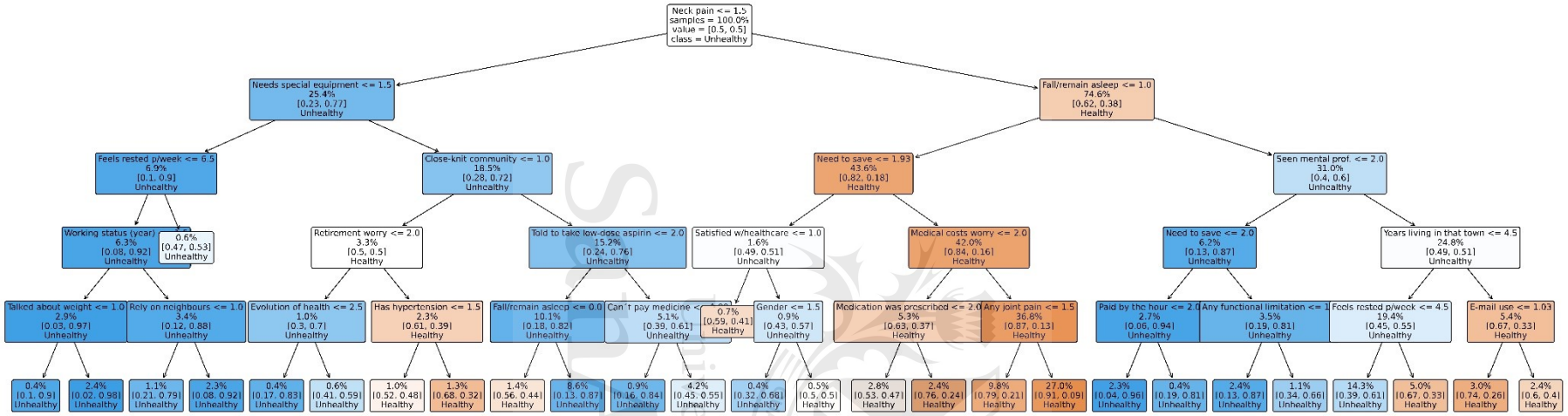


Source: NHIS

Additionally, Figure 14 depicts the classification tree for outcome 'Exhaustion'. Concerning the initial node, and as accounted for 'Restlessness' in Figure 13, it was defined by the need to save money over medicine. Moreover, and following the same spirit as the aforementioned outcome, the persons that chose money over health were also influenced by the presence or absence of the ability to reach over head without special equipment. Conversely, individuals without the economic dilemma were challenged by their sight, even when owning a set of glasses.

Regarding the differences with the other Figures presented formerly, there seems to be a significant concern about personal debt, manifested in the variable 'Credit card worry', which appears more than once. Once again, lack of sleep quality or quantity has a considerable role when understanding exhaustion.

Figure 15: Classification Tree for "Psychological Distress"



Source: NHIS

Finally, Figure 15 portrays the classification tree for outcome 'Psychological Distress'. Given this index was constructed based on the original six outcomes, it remains important to provide an deep insight of this outcome particularly. In the first place, the initial node concerns experiencing neck pain in last three months. For individuals to have reported feeling this, the next relevant feature is the need of special equipment because of health problems (I). Conversely, for those who did not suffer from neck pain, the following condition involves the ability to fall and/or remain asleep during the past week (II).

Branch I further assigns importance to the self-perceived rest per week for those who require special equipment, and to the community ties for those who do not. Meanwhile, branch II conditions the observations who did not experience trouble with their sleep to their need to save economic resources over health, while those who suffered from sleeping trouble are conditioned by the meeting of a mental professional.

It is interesting to highlight other features identified by the algorithm to understand psychological distress, such as the work status over the last week and year, concerns about future consumption over the elderly years, as well as weight issues, hypertension, and gender.

1- Analysis N°1: "Sadness"

Table 4: Confusion Matrices for "Sadness"

Logit Distribution[36]

	Y_0	Y_1
$\hat{Y}_0$	13.191	1.415
$\hat{Y}_1$	11.468	1.917

Random Forest

	Y_0	Y_1
$\hat{Y}_0$	20.658	1.295
$\hat{Y}_1$	4.001	2.037

K-Nearest Neighbours[37]

	Y_0	Y_1
$\hat{Y}_0$	21.864	2.813
$\hat{Y}_1$	2.795	519

Table 5: MSE & AUC Indicators per Model for "Sadness"

Model	**MSE**	**AUC**
Logit Distribution	46,03%	57,74%
Random Forest[38]	18,92%	80,91%
K-Nearest Neighbours	20,04%	54,85%

Figure 16: ROC Curves for "Sadness"

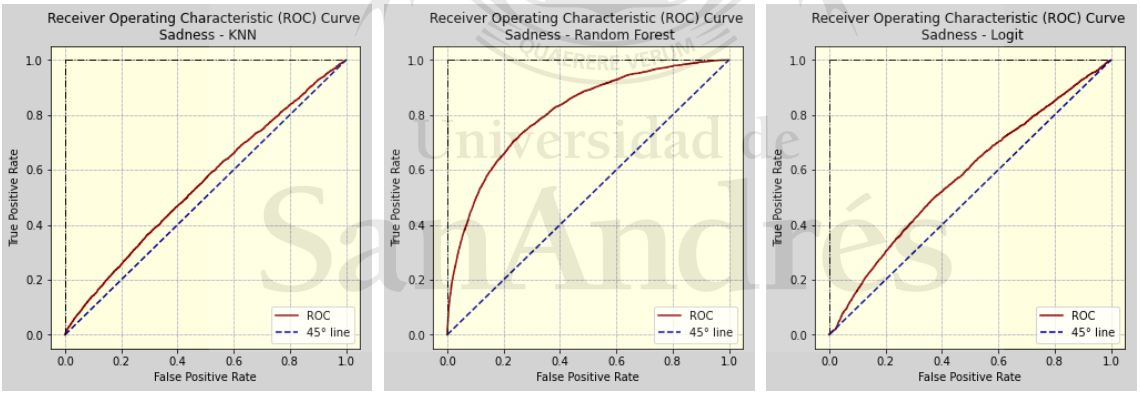


Table 6: Precision, Sensitivity & Specificity Indicators per Model for "Sadness"

Model	**Precision**	**Sensitivity**	**Specificity**
Logit Distribution	53,97%	90,31%	14,32%
Random Forest	81,07%	94,10%	33,74%
K-Nearest Neighbours	79,96%	88,60%	15,66%

[36] By CV, the L1/L2 ratio was determined at 0,0.

[37] By CV, the number of 'neighbours' was determined at 20.

[38] The OOB score was 85,19%.

2- Analysis N°2: "Nervousness"

Table 7: Confusion Matrices for "Nervousness"

Logit Distribution[39]

	Y_0	Y_1
$\hat{Y}_0$	17.396	3.735
$\hat{Y}_1$	5.169	1.691

Random Forest

	Y_0	Y_1
$\hat{Y}_0$	17.249	1.900
$\hat{Y}_1$	5.316	3.526

K-Nearest Neighbours[40]

	Y_0	Y_1
$\hat{Y}_0$	21.820	5.184
$\hat{Y}_1$	745	242

Table 8: MSE & AUC Indicators per Model for "Nervousness"

Model	**MSE**	**AUC**
Logit Distribution	31,81%	55,06%
Random Forest[41]	25,78%	78,66%
K-Nearest Neighbours	21,18%	52,80%

Figure 17: ROC Curves for "Nervousness"

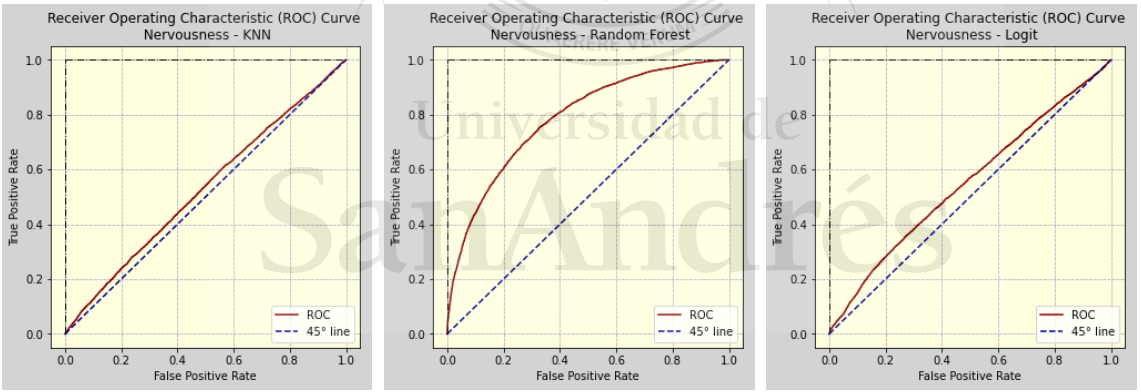


Table 9: Precision, Sensitivity & Specificity Indicators per Model for "Nervousness"

Model	**Precision**	**Sensitivity**	**Specificity**
Logit Distribution	68,19%	82,32%	24,65%
Random Forest	74,22%	90,08%	39,88%
K-Nearest Neighbours	78,82%	80,80%	24,52%

[39] By CV, the L1/L2 ratio was determined at 0,0.

[40] By CV, the number of 'neighbours' was determined at 40.

[41] The OOB score was 73,92%.

3- Analysis N°3: "Restlessness"

Table 10: Confusion Matrices for "Restlessness"

Logit Distribution[42]

	Y_0	Y_1
$\hat{Y}_0$	12.433	2.849
$\hat{Y}_1$	9.653	3.056

Random Forest

	Y_0	Y_1
$\hat{Y}_0$	17.013	1.970
$\hat{Y}_1$	5.073	3.935

K-Nearest Neighbours[43]

	Y_0	Y_1
$\hat{Y}_0$	21.651	5.720
$\hat{Y}_1$	435	185

Table 11: MSE & AUC Indicators per Model for "Restlessness"

Model	**MSE**	**AUC**
Logit Distribution	44,66%	55,48%
Random Forest[44]	25,16%	79,67%
K-Nearest Neighbours	22,00%	52,57%

Figure 18: ROC Curves for "Restlessness"

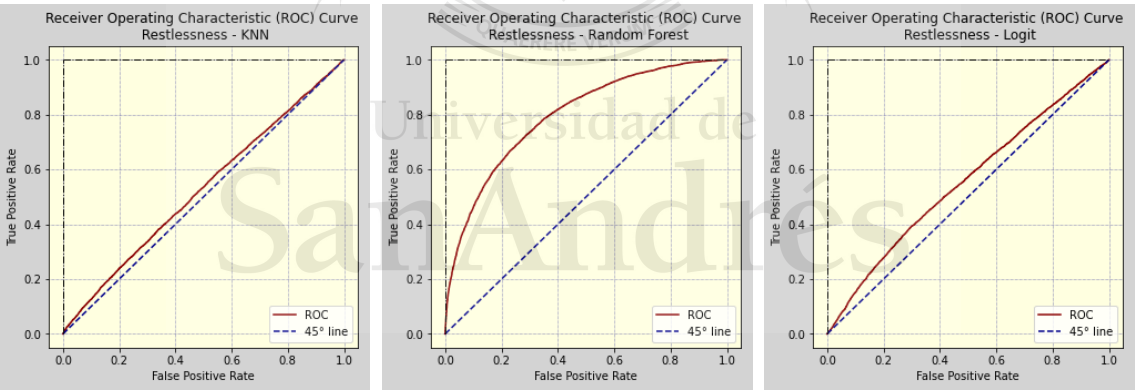


Table 12: Precision, Sensitivity & Specificity Indicators per Model for "Restlessness"

Model	**Precision**	**Sensitivity**	**Specificity**
Logit Distribution	55,33%	81,35%	24,04%
Random Forest	74,84%	89,62%	43,68%
K-Nearest Neighbours	78,01%	79,10%	29,83%

[42] By CV, the L1/L2 ratio was determined at 0,0.

[43] By CV, the number of 'neighbours' was determined at 50.

[44] The OOB score was 73,92%.

4- Analysis N°4: "Exhaustion"

Table 13: Confusion Matrices for "Exhaustion"

Logit Distribution[45]

	Y_0	Y_1
$\hat{Y}_0$	14.722	2.622
$\hat{Y}_1$	8.382	2.265

Random Forest

	Y_0	Y_1
$\hat{Y}_0$	18.500	1.694
$\hat{Y}_1$	4.604	3.193

K-Nearest Neighbours[46]

	Y_0	Y_1
$\hat{Y}_0$	21.969	4.474
$\hat{Y}_1$	1.135	413

Table 14: MSE & AUC Indicators per Model for "Exhaustion"

Model	**MSE**	**AUC**
Logit Distribution	39,31%	56,61%
Random Forest[47]	22,50%	80,63%
K-Nearest Neighbours	20,04%	54,61%

Figure 19: ROC Curves for "Exhaustion"

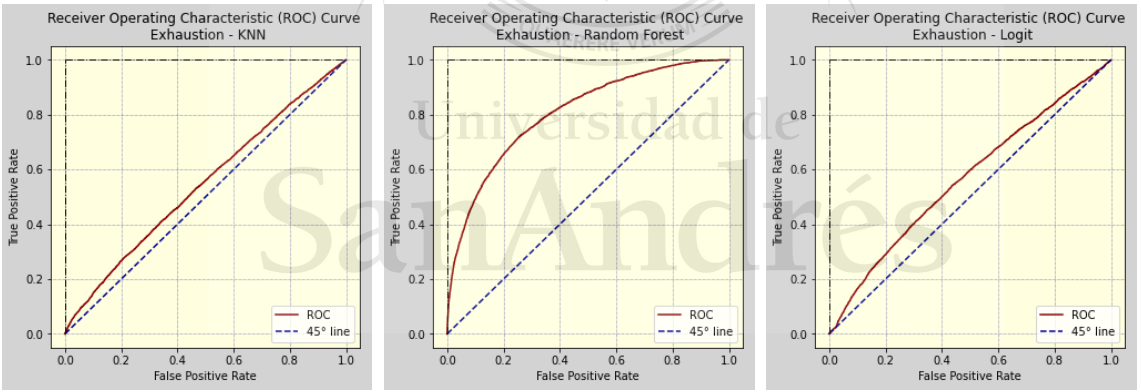


Table 15: Precision, Sensitivity & Specificity Indicators per Model for "Exhaustion"

Model	**Precision**	**Sensitivity**	**Specificity**
Logit Distribution	60,69%	84,88%	21,27%
Random Forest	77,50%	91,61%	40,95%
K-Nearest Neighbours	79,96%	83,08%	26,68%

[45] By CV, the L1/L2 ratio was determined at 0,0.

[46] By CV, the number of 'neighbours' was determined at 40.

[47] The OOB score was 77,23%.

5- Analysis N°5: "Psychological Distress"

Table 16: Confusion Matrices for "Psychological Distress"

Logit Distribution[48]

	Y_0	Y_1
$\hat{Y}_0$	16.978	1.694
$\hat{Y}_1$	7.836	1.483

Random Forest

	Y_0	Y_1
$\hat{Y}_0$	21.568	971
$\hat{Y}_1$	3.246	2.206

K-Nearest Neighbours[49]

	Y_0	Y_1
$\hat{Y}_0$	22.109	2.661
$\hat{Y}_1$	2.705	516

Table 17: MSE & AUC Indicators per Model for "Psychological Distress"

Model	**MSE**	**AUC**
Logit Distribution	34,00%	59,10%
Random Forest[50]	15,07%	87,14%
K-Nearest Neighbours	19,17%	54,59%

Figure 20: ROC Curves for "Psychological Distress"

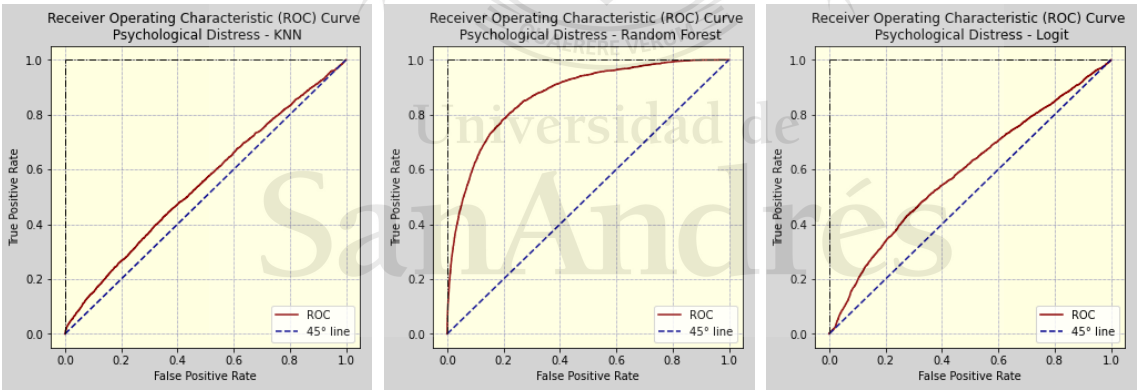


Table 18: Precision, Sensitivity & Specificity Indicators per Model for "Psychological Distress"

Model	**Precision**	**Sensitivity**	**Specificity**
Logit Distribution	65,95%	90,93%	15,91%
Random Forest	84,93%	95,69%	40,46%
K-Nearest Neighbours	80,83%	89,26%	16,02%

[48] By CV, the L1/L2 ratio was determined at 0,0.

[49] By CV, the number of 'neighbours' was determined at 20.

[50] The OOB score was 85,19%.

VI. Conclusion

In this last Part, a comment on the main findings will be formulated on Section A) concerning not only on the achieved results, but also on the overall performance of the model in relation to alternatives. Lastly, concluding remarks shall be delivered on Section B).

A) Main Findings

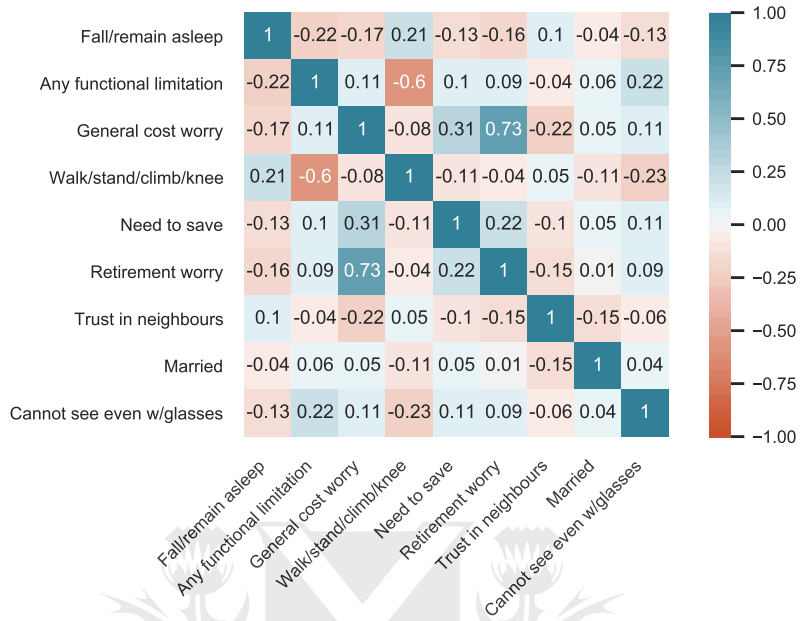
From Part V., a number of aspects can be discussed. To begin with, from Section A) it can be inferred that the most relevant variables (i.e. the first nodes of the tree) can be traced back to the main following:

- The ability to fall and/or remain asleep, and also related, the quality of the sleep (i.e. how rested a person feels);
- Difficulties in leg movement which involve walking, standing, climbing and/or kneeling;
- Any self-perceived functional limitation in movement;
- Worried by medical costs, or also general costs[51];
- Concerns raised by retirement;
- The marital status at the moment of the enquiry; and
- Computer use indicators, among other.

Provided the variables identified by the predictive model, it follows that a wide range of aspects are involved: from monetary to physical, including interactions with a significant other or the neighbourhood. In an attempt to establish the hierarchy of the features, Table 19 in the Appendix enumerates the variables, ordered by most important to the least, according to the Gini criterion. Although it may be true that all of these variables may be correlated in some particular instances, it roughly applies on a broader sense. A visualisation of the low correlations between indicators is presented in Figure 21.

[51] The latter worries are concerned on housing/rent, health costs, bills, quality of life, health bills, and medical costs.

Figure 21: Correlation Matrix for the Most Relevant Indicators According to the Gini Criterion



Source: NHIS

Because of this, the suggestion that non-linear model was the best theoretic approach seems validated: psychological distress is present in the aforementioned regressors, whereas these do not correlate significantly systematically between each other. In this sense, interactions between variables strike as vital in order to grasp a better comprehension of the phenomenon in question.

Moreover, concerning the alternative models, from B) it strikes at first that Random Forest outperforms the linear model of Logit-LASSO considering the MSE indicator. Furthermore, on average Random Forests predicts similarly to KNN; however, significant differences arise between both when contemplating the AUC and ROC indicators: while the former achieves 80,0% on average, the latter cannot surpass 60,0%. This represents a significant drawback for KNN, allowing to conclude with the original question of this work: the non-linear relation of the variables.

B) Concluding Remarks

The purpose of this work was to grasp a better understanding of psychological distress and its numerous realisations, which included sadness, nervousness, restlessness and exhaustion. Parting

from the prior that predictions vary significantly across different regressors, it was deemed as important to build a non-linear model that would cope with a wide range of variables; and the chosen model was random forests.

In the course of this work, it has been shown how differences in key variables condition the relevance of other features. In average for all the models, there seems to be a clear pattern around spatial mobility, financial restrictions, quality of sleep, and marital status. It is interesting to highlight the presence of these variables in the vast literature of psychological distress.

Rather than dealing with issues involving psychological or technical aspects, a major challenge faced by the work has fallen around the data and how it was structured. Highly unbalanced observations forced the grouping of the initial five classes per variable into a new pair, making the variable dichotomous at the investigator's discretion. Even though the inclusion of a number of databases moderated this issue to some extent, the resulting enlarged database had seldom brought some computational costs.[52] Moreover, it might be stated that additional indicators could be contemplated, such as family environment, or genetics.

Furthermore, it could be argued that a drawback of the study was the lacking of explanatory mechanisms; however, given that the scope was to build predictive -not causal- models, it leaves ground for future analysis around mechanisms that have not been already studied in the literature. As for the building of the indicator for 'psychological distress', other calculations might be pursued, such as Non-Linear Principal Component Analysis - NLPCA, or also CatPCA (Linting, 2012).[53] The same could also be applied to other highly correlated regressors. Last but not least, other non-linear methodologies can be explored in order to seek further relations among variables, for instance neural networks.

All in all, the work has shed light on the highly complex and often hard to measure (if not unobserved) questions around psychological distress. As a recommendation for policy, authorities may place their focus on some indicators more thoroughly than on others at the time of battling this issue. Elevated health costs, mobility limitations, labour market characteristics,

[52] For instance, the KNN command was highly sensitive to the number of neighbours - the higher the number, the more the time required. Another example can be found in the use of traditional oversampling techniques (where values even at 50% for both did not show significant reductions in time costs), or 'Bordeline-SMOTE1'.

[53] Principal Component Analysis (PCA) was first considered to build the proposed outcome; however, the resulting values were neither intuitive nor clear for the algorithm.

and unsettling neighbourhoods, just to mention a few, may be wanted to be paid more attention when dealing with psychological distress. This too should also be in mind for the COVID-19 pandemic, where lower income, less mobility and temporary job suspensions, accompanied by sleep alterations, should be brought onto the table when discussing both the length and depth of quarantine.

References

N. L. Bragazzi, A. Watad, A. Gizunterman, D. McGonagle, H. Mahagna, D. Comaneshter, H. Amital, A. D. Cohen, and D. Amital. The burden of depression in systemic sclerosis patients: a nationwide population-based study. *Journal of Affective Disorders*, 243:427–431, 2019. URL https://doi.org/10.1016/j.jad.2018.09.075.

L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. URL https://doi.org/10.1023/A:1010933404324.

N. V. Chawla. Data mining for imbalanced datasets: An overview. In O. Maimon and L. Rokach, editors, *Data Mining and Knowledge Discovery Handbook*, page 875–886. Springer, Boston, MA, 2009. ISBN 978-0-387-09822-7. URL https://doi.org/10.1007/978-0-387-09823-4_45.

D. De Leo, P. Hickey, G. Meneghel, and C. Cantor. Blindness, fear of sight loss, and suicide. *Psychosomatics*, 40(4):339–344, 1999. URL https://doi.org/10.1016/S0033-3182(99)71229-6.

T. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning*, 40:139–157, 2001. URL https://doi.org/10.1023/A:1007607513941.

A. Drapeau, A. Marchand, and D. Beaulieu-Prevost. Epidemiology of psychological distress. In P. L. LAbate, editor, *Mental Illnesses - Understanding, Prediction and Control*, chapter 5, pages 105–134. InTech, Oxford, 2012. URL https://doi.org/10.5772/30872.

A. Fernandez, S. Garcia, F. Herrera, and N. V. Chawla. Smote for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research*, 61:863–905, 2018. URL https://doi.org/10.1613/jair.1.11192.

B. Frénay and M. Verleysen. Classification in the presence of label noise: A survey. *Neural Networks and Learning Systems, IEEE Transactions on*, 25: 845–869, 05 2014. URL https://doi.org/10.1109/TNNLS.2013.2292894.

G. Gonzales, J. Przedworski, and C. Henning-Smith. Comparison of health and health risk factors between lesbian, gay, and bisexual adults and heterosexual adults in the united states. *JAMA Internal Medicine*, 176(9):1344–1351, 2016. URL https://doi.org/10.1001/jamainternmed.2016.3432.

H. Han, W. Wang, and B. Mao. Borderline-smote: A new over-sampling method in imbalanced data sets learning. *Lecture Notes in Computer Science Advances in Intelligent Computing*, 3644:878–887, 2005. URL https://doi.org/10.1007/11538059_91.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition*. Springer, New York, 2009. ISBN 978-0-387-84857-0.

X. HE, N. Chaney, M. Schleiss, and J. Sheffield. Spatial downscaling of precipitation using adaptable random forests. *Water Resources Research*, 52, 10 2016. URL 10.1002/2016WR019034.

T. Hill, A. Burdette, and L. Hale. Neighborhood disorder, sleep quality, and psychological distress: Testing a model of structural amplification. *Health and Place*, 15(4):1006–1013, 2009. URL https://doi.org/10.1016/j.healthplace.2009.04.001.

S. Hope, B. Rodgers, and C. Power. Marital status transitions and psychological distress: Longitudinal evidence from a national population sample. *Psychological Medicine*, 29(2):381–389, 1999. URL https://doi.org/10.1017/S0033291798008149.

G. Hullam, P. Antal, P. Petschner, X. Gonda, G. Bagdy, B. Deakin, and G. Juhasz. The ukb envirome of depression: from interactions to synergistic effects. *Sci-*

entific Reports, 9(1):9723, 2019. URL https://doi.org/10.1016/j.jad.2018.09.075.

G. James, D. Witten, T. Hastie, and R. Tibshirani. An Introduction to Statistical Learning: with Applications in R. Springer, New York, 2013. ISBN 978-0-387-84857-0.

N. Jothi, W. Husain, and N. A. Rashid. Predicting generalized anxiety disorder among women using shapley value. Journal of Infection and Public Health, 2020. URL https://doi.org/10.1016/j.jiph.2020.02.042.

R. Kessler. Stress, social status, and psychological distress. Journal of Health and Social Behavior, 20(3):259–272, 1979. URL https://doi.org/10.2307/2136450.

R. Kessler and M. Essex. Marital status and depression: The importance of coping resources. Social Forces, 61(2):484–507, 1982. URL https://doi.org/10.2307/2578238.

R. Kessler and H. Neighbors. A new perspective on the relationships among race, social class, and psychological distress. Journal of Health and Social Behavior, 27(2):107–115, 1986. URL https://doi.org/10.2307/2136310.

R. C. Kessler, G. Andrews, L. J. Colpe, E. Hiripi, D. K. Mroczek, S. L. Normand, E. E. Walters, and A. M. Zaslavsky. Short screening scales to monitor population prevalences and trends in non-specific psychological distress. Psychological Medicine, 32(6):959–976, 2002. URL https://doi.org/10.1017/s0033291702006074.

D. Lawrence, F. Mitrou, and S. Zubrick. Non-specific psychological distress, smoking status and smoking cessation: United states national health interview survey 2005. BMC Public Health, 11(256), 2011. URL https://doi.org/10.1186/1471-2458-11-256.

A. Linting, Mariëlle Van Der Kooij. Nonlinear principal components analysis with catpca: A tutorial. Journal of Personality Assessment, 94:12–25, 2012. URL https://doi.org/10.1080/00223891.2011.627965.

N. J. Pearson. Insomnia, trouble sleeping, and complementary and alternative medicine. *Archives of Internal Medicine*, 166(16):1775, 2006. URL https://doi.org/10.1001/archinte.166.16.1775.

M. M. Rahman and D. N. Davis. Addressing the class imbalance problem in medical datasets. *International Journal of Machine Learning and Computing*, 3(2):224–228, 2013. URL https://doi.org/10.7763/ijmlc.2013.v3.307.

G. K. Shapiro and B. J. Burchell. Measuring financial anxiety. *Journal of Neuroscience, Psychology, and Economics*, 5(2):92–103, 2012. URL https://doi.org/10.1037/a0027647.

G. Sithey, L. M. Wen, P. Kelly, and M. Li. Association between sleep duration and self-reported health status: Findings from the bhutan's gross national happiness study. *Journal of Clinical Sleep Medicine*, 13(01):33–38, 2017. URL https://dx.doi.org/10.5664/jcsm.6382.

D. Umberson, C. Wortman, and R. Kessler. Widowhood and depression: Explaining long-term gender differences in vulnerability. *Journal of Health and Social Behavior*, 33(1):10–24, 1992. URL https://doi.org/10.2307/2136854.

J. Wels. Assessing the impact of partial early retirement on self-perceived health, depression level and quality of life in belgium: A longitudinal perspective using the survey of health, ageing and retirement in europe (share). *Ageing and Society*, 40(3):512–536, 2018. URL https://doi.org/10.1017/s0144686x18001149.

C. Whelan. The role of income, life[U+2010]style deprivation and financial strain in mediating the impact of unemployment on psychological distress: Evidence from the republic of ireland. *Journal of Occupational and Organizational Psychology*, 65:331–344, 1992. URL https://doi.org/10.1111/j.2044-8325.1992.tb00509.x.

Q. Zhou, W. Chen, S. Song, J. Gardner, and K. Weinberger. A reduction of the

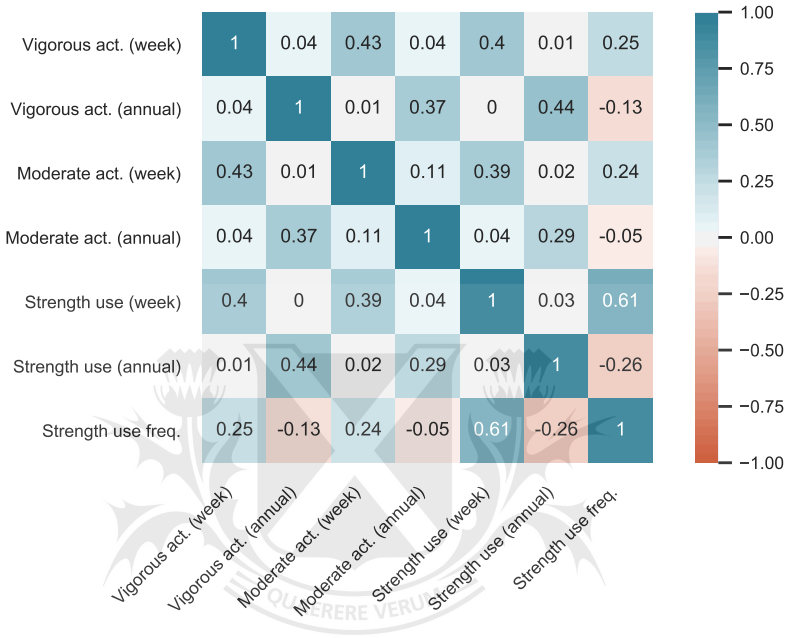
elastic net to support vector machines with an application to GPU computing. *ArXiv e-prints*, 2014.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005. URL https://doi.org/10.1111/j.1467-9868.2005.00503.x.

Appendix

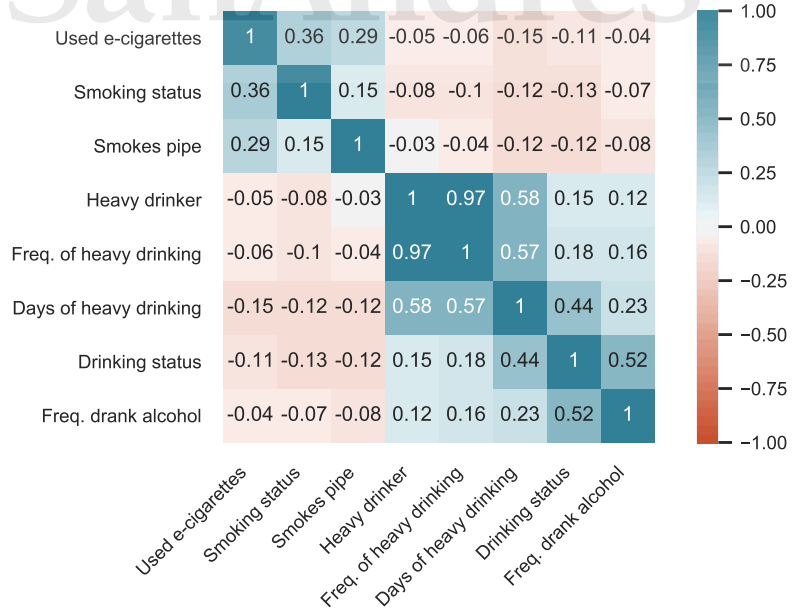
Other Correlations

Correlation Matrix for Several Indicators Related with Physical Activity



Source: NHIS

Correlation Matrix for Several Indicators Related with Substance Consumption such as Cigarettes and Alcohol



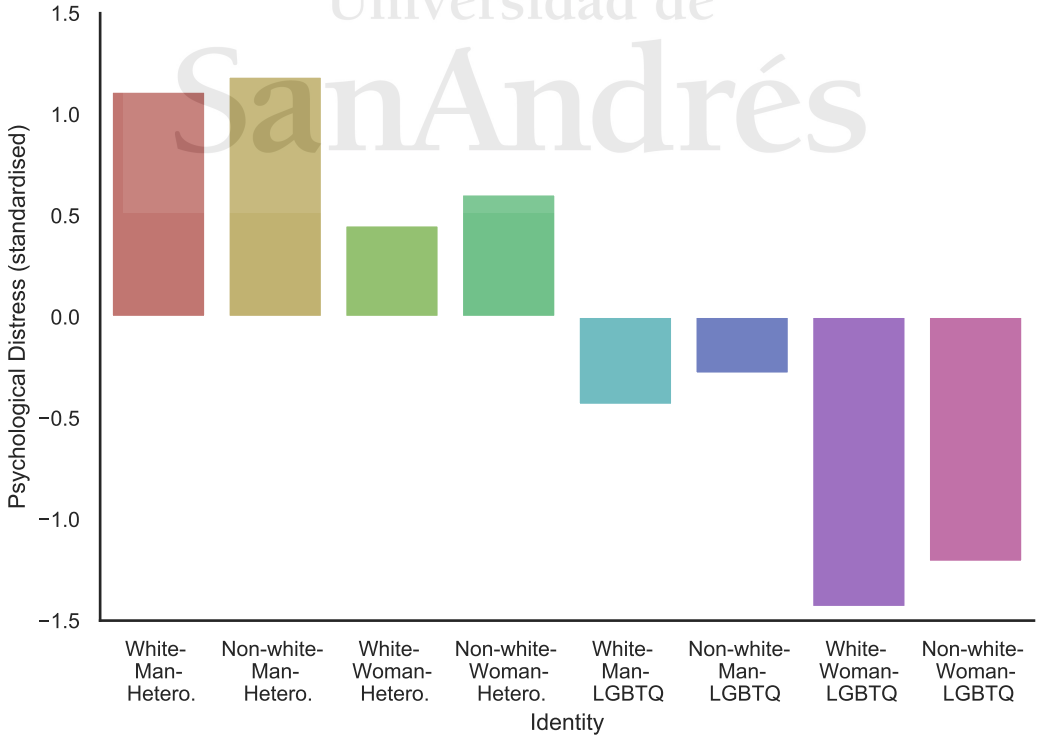
Source: NHIS

Figure 22: Correlation Matrix for Several Indicators Related with Labour Conditions

Source: NHIS

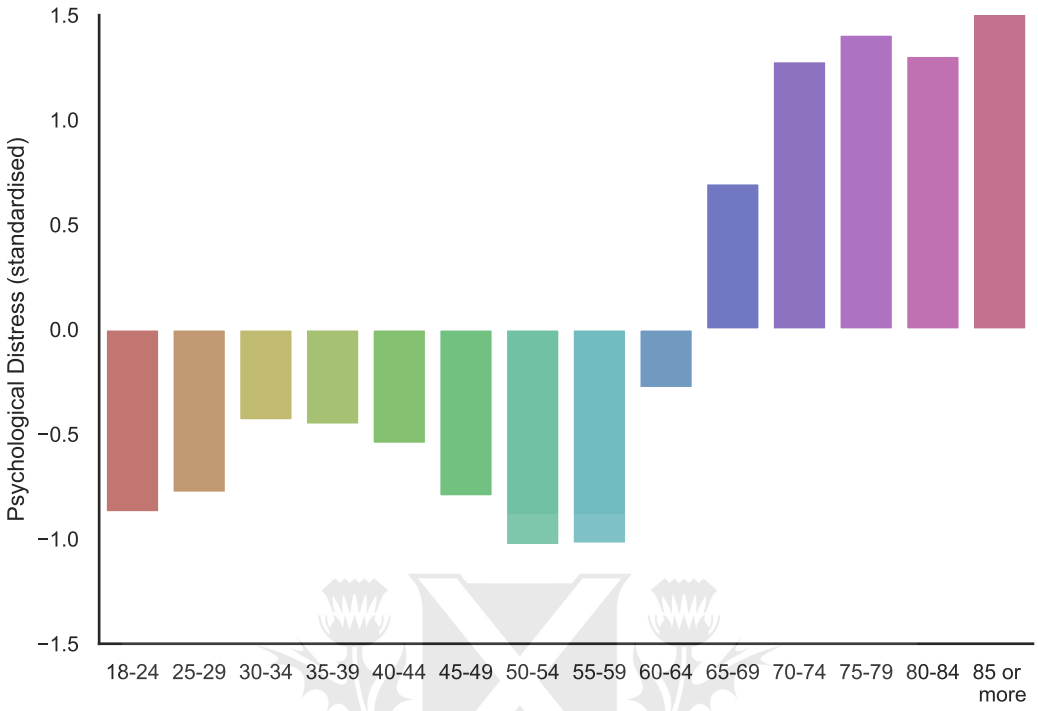
Psychological Distress and Demographics

Figure 23: Psychological Distress on Sex, Ethnics and Gender



Source: NHIS

Figure 24: Psychological Distress on Age



Source: NHIS

Random Forest Feature Relevance

Table 19: Random Forest Feature Relevance using Gini criterion

#	Feature	Gini Importance
1	Cannot fall/remain asleep	0,10143
2	Feels rested (week)	0,07887
3	Presents any functional limitation	0,06062
4	Worried by general cost	0,05959
5	Has seen mental professional	0,05629
6	Low back pain	0,05032
7	Can walk/stand/climb/knee without help	0,04734
8	Has migraines	0,04643
9	Worried about retirement	0,04585
10	Has neck pain	0,02923
11	Worried about medical costs	0,02915
12	Cannot afford dental care	0,02537
13	Joint pain/aching/stiffness	0,02523
14	Assisted in social activities	0,02240
15	Satisfaction with healthcare	0,01747
16	Evolution of health	0,01741
17	Cannot afford medicine	0,01656
18	Need to save money	0,01425
19	Trust in neighbours	0,01390
20	Is currently married	0,01294
21	Takes sleep medication	0,01265
22	Assisted to go out	0,01258
23	N° of bed days	0,01246
24	Cannot afford glasses	0,01244
25	Paid sick leave in job	0,01125
26	Medication was prescribed	0,00988

#	Feature	Gini Importance
27	Cannot see even w/glasses	0,00966
28	Paid by the hour	0,00917
29	Tested for HIV	0,00872
30	Smoking status	0,00806
31	Limitation condition status	0,00693
32	Hours of sleep	0,00639
33	Rely on neighbours	0,00577
34	Work Status	0,00573
35	Assisted to push	0,00549
36	Has used e-cigarettes	0,00500
37	Cannot afford mental care	0,00487
38	Has arthritis	0,00460
39	Office visits	0,00445
40	Age	0,00434
41	Assisted to sit	0,00429
42	Had to ask for lower costs	0,00408
43	Worried by credit card payments	0,00375
44	Help from neighbours	0,00306
45	Told to take low-dose aspirin	0,00298
46	Years living in that town	0,00298
47	Difficulty to carry	0,00257
48	Needs special equipment	0,00219
49	Computer use frequency	0,00215
50	Employment condition	0,00206
51	Cannot afford specialist	0,00177
52	Seen a NP/ PA/midwife	0,00172
53	Seen eye doctor	0,00162
54	Pneumonia shot	0,00156
55	Asthmatic	0,00138

#	Feature	Gini Importance
56	Jaw/front of ear pain	0,00136
57	E-mail use	0,00133
58	Marital Status	0,00132
59	Is supervisor	0,00129
60	Seen medical specialist	0,00123
61	Appointment delayed	0,00104
62	Close-knit community	0,00102
63	Was talked about own weight	0,00101
64	Years working on job	0,00089
65	Has searched health information on Internet	0,00087
66	Is the current job the longest?	0,00081
67	Finds healthcare expensive	0,00078
68	Worried for children's college costs	0,00078
69	Internet use freq. (year)	0,00073
70	Internet use	0,00071
71	Ever smoked cigar	0,00058
72	Where to go when sick	0,00056
73	Evert tested hepatitis	0,00056
74	Stomach problems	0,00054
75	Evolution of insurance since LY	0,00052
76	Smokes pipe	0,00050
77	Working status (year)	0,00046
78	Internet use frequency (units)	0,00046
79	Had diabetes	0,00035
80	Difficulty to reach	0,00030
81	Assisted to relax	0,00030
82	Vigorous activity (annual)	0,00029
83	Had 12 or more drinks (year)	0,00029
84	Has chronic bronchitis	0,00025

#	Feature	Gini Importance
85	Parent of child	0,00022
86	Seen a therapist	0,00021
87	Had an ulcer	0,00020
88	Tried to buy insurance	0,00020
89	Ethnics	0,00019
90	Assisted to grasp	0,00015
91	Drinking status	0,00014
92	Alternatives for saving	0,00013
93	Emailed physician	0,00011
94	Hispanic	0,00011
95	Not open when spare time	0,00010
96	Moderate activity (annual)	0,00010
97	N° of employees	0,00010
98	Times in ER	0,00009
99	Had to wait for doctor	0,00008
100	Region	0,00007
101	Type of worker	0,00007
102	Freq. drank alcohol (daily)	0,00006
103	Couldn't get appointment on the phone	0,00005
104	Time since saw dentist	0,00005
105	Vigorous activity (week)	0,00004
106	Strength use (annual)	0,00004
107	Freq. drank alcohol	0,00004
108	Opt. sleep	0,00004
109	Sex	0,00003
110	Minority	0,00003
111	Medical appointment on Internet	0,00002
112	Moderate act. (week)	0,00002
113	Strength use (week)	0,00002

#	Feature	Gini Importance
114	Strength use frequency	0,00002
115	Weight	0,00001
116	Hearing without aid	0,00001
117	Weight without shoes	0,00001
118	Body Mass Index	0,00001
119	Gender	0,00001
120	Employed in... Manufacturing (IND_05)	0,00001
121	Employed in... Education Services(IND_15)	0,00001
122	Employed in... Health Care and Social Assistance (IND_16)	0,00001
123	Works as... Chief executives, or managers (OCC_01)	0,00001

Logit Coefficients

Table 20: Logit Coefficients after 'Elastic-Net' Regularisation, in Descending Order by Absolute Value, for Psychological Distress

#	Variable	Coefficient
1	N of bed days	0,00122
2	Freq. drank alcohol	-0,00081
3	Age	-0,00030
4	Years on job	-0,00028
5	Weight without shoes	0,00027
6	Feels rested p/week	-0,00021
7	Fall/remain asleep	0,00019
8	Sleep med, freq,	0,00010
9	Moderate act, (week)	-0,00009
10	Office visits	0,00009
11	Assisted in social act,	0,00006
12	Body Mass Index	0,00006
13	General cost worry	-0,00006
14	Hours of sleep	-0,00006
15	Medical costs worry	-0,00006
16	Retirement worry	-0,00006
17	Vigorous act. (week)	-0,00006
18	Walk/stand/climb/knee	0,00006
19	Assisted to go out	0,00005
20	Assisted to push	0,00005
21	Any functional limitation	-0,00004
22	Assisted to sit	0,00004
23	Computer use freq,	-0,00004
24	Difficulty to carry	0,00004
25	Height	-0,00004

#	Variable	Coefficient
26	Low back pain	-0,00004
27	Smoking status	-0,00004
28	Strength use freq,	-0,00004
29	Years living in that town	-0,00004
30	Any joint pain	-0,00003
31	Close-knit community	0,00003
32	Drinking status	-0,00003
33	Employment condition	0,00003
34	Freq, drank alcohol (daily)	-0,00003
35	Help from neighbours	0,00003
36	Marital Status	0,00003
37	Migrane	-0,00003
38	Neck pain	-0,00003
39	Rely on neighbours	0,00003
40	Strength use (week)	0,00003
41	Times in ER	0,00003
42	Trust in neighbours	0,00003
43	Vigorous act. (annual)	0,00003
44	Asmthatic	-0,00002
45	Assisted to grasp	0,00002
46	Assisted to relax	0,00002
47	Cannot pay dental care	-0,00002
48	Cannot see even w/glasses	-0,00002
49	College for child worry	-0,00002
50	Difficulty to reach	0,00002
51	Has arthritis	-0,00002
52	Health info, on Internet	-0,00002
53	Married	-0,00002
54	Moderate act. (annual)	0,00002

#	Variable	Coefficient
55	Needs special equipment	-0,00002
56	Satisfied w/healthcare	0,00002
57	Seen mental prof.	-0,00002
58	Strength use (annual)	0,00002
59	Time since saw dentist	0,00002
60	Alternatives for saving	-0,00001
61	Asked for lower costs	-0,00001
62	Cannot afford glasses	-0,00001
63	Cannot pay medicine	-0,00001
64	Cannot pay mental care	-0,00001
65	Credit cards worry	-0,00001
66	Current job, the longest?	0,00001
67	Delayed appointment	-0,00001
68	E-mail use	0,00001
69	Ethnics	-0,00001
70	Evert tested hepatitis	-0,00001
71	Evolution of health	-0,00001
72	Evolution of insurance	-0,00001
73	Expensive care	-0,00001
74	Had an ulcera	-0,00001
75	Had to wait for doctor	-0,00001
76	Has chronic bronchitis	-0,00001
77	Has had cancer	-0,00001
78	Hearing without aid	0,00001
79	Hispanic	-0,00001
80	Internet use freq. (units)	-0,00001
81	Jaw/front of ear pain	-0,00001
82	Limitation condition status	0,00001
83	Medical appointment on Internet	-0,00001

#	Variable	Coefficient
84	Medication was prescribed	-0,00001
85	N° of employees	0,00001
86	Need to save	-0,00001
87	Opt. sleep	0,00001
88	Paid by the hour	-0,00001
89	Paid sick leave in job	0,00001
90	Parent of child	-0,00001
91	Seen a NP/ PA/midwife	-0,00001
92	Seen a therapist	-0,00001
93	Seen medical specialist	-0,00001
94	Stomach problems	-0,00001
95	Talked about weight	-0,00001
96	Tested for HIV	-0,00001
97	Type of worker	-0,00001
98	Unaffordable specialist	-0,00001
99	Used e-cigarettes	-0,00001
100	Weight	-0,00001
101	Work Status	0,00001
102	Working status (year)	0,00001

List of Tables

List of Figures