UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE ENGENHARIA DE COMPUTAÇÃO

GABRIEL DE SOUZA SEIBEL

Classification of suicidality in a large occupational cohort: an analysis of machine learning algorithms applied to the ELSA-Brasil study

Work presented in partial fulfillment of the requirements for the degree of Bachelor in Computer Engineering

Advisor: Prof. Dr. Mariana Recamonde Mendoza

Porto Alegre
Dezembro 2020

"When you see a man casting pearls without getting even a pork chop in return—it is not against the swine that you feel indignation. It is against the man who valued his pearls so little that he was willing to fling them into the muck and let them become the occasion for a whole concert of grunting, transcribed by the court stenographer."

— Ayn Rand

ACKNOWLEDGMENT

A special acknowledgment is given to professors André Russowsky Brunoni, Ives Cavalcante Passos, and Mariana Recamonde Mendoza, for the valuable help and input on this work. As this study relates to both the medical and the informatics fields, their involvement and advising were crucial.

ABSTRACT

Suicide ideation is strongly correlated to suicide acts, a grave problem for society quantified in hundreds of thousands of deaths per year. Nevertheless, patterns regarding the emergence and presence of suicidal thoughts are not completely elucidated. Techniques from the Machine Learning field of study have shown great potential and success in tackling the problem of identifying or predicting suicidality in individuals, even though they still often face challenges in scenarios where the available data has a small percentage of occurrences of the class of interest. Thus, the main goal of this work is to train and evaluate models to identify instances as presenting *suicidality*, which we refer to as a combination of self-reported suicide ideation, *"taedium vitae"* (feeling that life is not worth living), and hopelessness. We also aim to analyse the factors involved in these models decision-making processes. Our proposed solution to this challenge is a general classifier-induction pipeline for dealing with datasets with class imbalance and feature abundance, which was validated using data obtained from the Brazilian Longitudinal Study of Adult Health (ELSA-Brasil), restricted to a subset only containing people with common mental disorders. Experiments using our approach to fit Elastic Nets, Neural Networks, Random Forests, and to combine them in a probability-averaging ensemble yielded classifiers with over 0.8 area under the receiver operating characteristic curve, F_2-Score from 0.6 to around 0.7, sensitivity up to 0.8, and specificity of ranging from 0.6 to 0.8 depending on the algorithm. The most important variables in the models' inference of suicidality were related to feelings of inferiority, sadness, disappearance of interests, unnecessary guilt and self-blaming, energy (disposition) levels, preclusion of activities including chores and leisure for bad feelings, income, anxiety, worrying, libido, irritability, obsession, and physical activities.

Keywords: Machine Learning. Classification. Suicide. Suicide Ideation. ELSA-Brasil.

Classificação de Suicidalidade em uma vasta coorte ocupacional: uma análise de algoritmos de aprendizado de máquina aplicados ao estudo ELSA-Brasil

RESUMO

Ideação suicida está fortemente correlacionada a atos de suicídio, um problema social grave quantificado em centenas de milhares de mortes por ano. Não obstante, padrões quanto ao surgimento e a presença desses pensamentos não estão completamente elucidados. Técnicas do campo de estudo de aprendizado de máquina já mostraram grande potencial e sucesso para enfrentar o problema de identificar ou prever suicidalidade em indivíduos, apesar de frequentemente encontrarem dificuldades em cenários em que os dados disponíveis têm uma porcentagem pequena de ocorrências da classe de interesse. Portanto, o objetivo deste trabalho é treinar e avaliar modelos para identificar instâncias apresentando *suicidalidade*, a que nos referimos como uma combinação de ideação suicida, *"taedium vitae"* (sentir que a vida não vale a pena ser vivida) e desesperança autorelatados. Nós também queremos analisar os fatores envolvidos no processo de decisão desses modelos. Nossa solução proposta para esse desafio é um fluxo geral de indução de classificadores para lidar com conjuntos de dados com desbalanço de classe e abundância de atributos, que foi validado usando dados obtidos do Estudo Longitudinal de Saúde do Adulto brasileiro (ELSA-Brasil), restritos a um subconjunto composto apenas por pessoas com transtornos mentais comuns. Experimentos usando nossa abordagem para ajustar modelos de Redes Elásticas, Redes Neurais e Florestas Aleatórias e então combiná-los em um modelo conjunto de média ponderada de probabilidades produziu classificadores com área sob a curva característica de operação do receptor maior que 0.8, Valor-F2 de 0.6 a por aproximadamente 0.7, sensibilidade de até 0.8 e especificidade variando de 0.6 a 0.8 dependendo do algoritmo. As variáveis mais importantes na inferência de suicidalidade pelos modelos são relacionadas a sentimentos de inferioridade, tristeza, desaparecimento de interesses, sensação de culpa mesmo que desnecessária, níveis de energia (disposição), incapacidade de realizar atividades incluindo responsabilidades e lazer por sentimentos ruins, renda, ansiedade, preocupação, libido, irritabilidade, obsessão e atividades físicas.

Palavras-chave: Aprendizado de Máquina, Classificação, Suicídio, Ideação Suicida, ELSA-Brasil.

LIST OF ABBREVIATIONS AND ACRONYMS

ELSA-Brasil	Longitudinal Study of Adult Health (Brasil)
CIS-R	Clinical Interview Schedule - Revised Version
MDD	Major Depressive disorder
MADD	Mixed Anxiety–Depressive Disorder
SLR	Systematic Literature Review
ML	Machine Learning
EN	Elastic Net
MLP	Multilayer Perceptron
ANN	Artificial Neural Network
RF	Random Forest
XGB	eXtreme Gradient Boosting
SVM	Support Vector Machine
FE	Feature Elimination
RFE	Recursive Feature Elimination
AUC	Area Under Curve
ROC	Receiver Operating Characteristic
AUCROC	Area under ROC curve
CV	Cross Validation
GS	Grid Search
NZV	Near-zero variance

LIST OF FIGURES

LIST OF TABLES

CONTENTS

1 INTRODUCTION

Suicide is a major concern worldwide, given its broad and severe impact: more than 800,000 people die from suicide each year - every 40 seconds, one person takes their own life (REID, 2010). Suicide globally represents the second highest cause of death among people aged between 15 to 29 years, and it is estimated that about 80% of the acts occur in developing countries (WHO, 2017). Therefore, it is evident the need for actions to better understand its patterns, especially on the population most affected by it, to support the creation of prevention methods.

Within the concept of suicidal behavior, there is the category of suicide ideation (SI; to consider committing suicide, or to think about it in general) (REID, 2010). Self-reported suicide ideation rates end up being underestimated depending on the interview method (SPIERS et al., 2014), which could present itself as a methodological limitation on the investigation of suicide causes and associated factors. Nevertheless, SI is a strong indicator of vulnerability to suicidal acts (BEBBINGTON et al., 2010), hence the prevention of the latter could be achieved by better understanding the profile of people affected by the former, ideally predicting its incidence. The term "suicidality" has often had a lack of clarity of definition, sometimes generalized up to the point of being conflated with self-injury, but a prevalent and surely constituent characteristic of this concept is *suicide intention* (CARBALLO et al., 2020). Although suicidality can also be identified in cases of suicide attempts or plans, in this work we specifically refer to it as a proxy to suicide intention, by considering self-reported feelings of hopelessness, or feelings that life is not worth living (also called *"taedium vitae"*), or direct suicide ideation.

In recent years, in a context of an exponential generation and availability of data regarding virtually any phenomena, the computer science field of machine learning, dedicated to deriving knowledge from information, has been steadily growing and flourishing in both academic and business fronts. In particular, in a scenario made possible by methodological advancements in machine learning combined with the availability of electronic medical records (EMRs) and socioeconomic and behavioral demographic data, automatic diagnosis or prediction of diseases has had considerable success in supporting physicians and health professionals in general, while also providing a better understanding of studied the phenomena (DARCY; LOUIE; ROBERTS, 2016). These approaches, however, still oftentimes face challenges in identifying adequate patterns when dealing with data that has fewer examples of one representative class than of the other, which is

a recurrent problem for diseases and medical conditions in general (BURKE; AMMERMAN; JACOBUCCI, 2019). This has been aptly the "class-imbalance problem".

In Brazil, a dataset with the potential of fruitful employment of machine learning techniques is the one produced by the Longitudinal Study of Adult Health (ELSA-Brasil) cohort study (SCHMIDT et al., 2015). The study evaluated social and biological factors related to, among other health topics, mental health. Moreover, it includes responses to a questionnaire called Clinical Interview Schedule-Revised (CIS-R), with information on the interviewees' self-reported thoughts surrounding suicide (NUNES et al., 2016; LEWIS et al., 1992) that compose our labeling of suicidality. Although the ELSA-Brasil project has interviewed over 15,000 adults in its first wave, in our study we restrict our analysis to the people presenting common mental disorders (CMD), around 4,000 individuals.

Therefore, we hypothesize that using state-of-the-art machine learning techniques over data from the ELSA-Brasil project, we can build models to identify patterns in suicidality, and based on certain structured characterizations of a person, correctly classify whether they present it. Thus, the goal of this work is to develop classifiers able to identify individuals presenting suicidality-associated patterns with high performance, despite data limitations due to the low sample size for this class of interest. In addition, we aim to provide useful knowledge regarding factors associated with suicidality that may be further explored by mental-health professionals in clinical and academic settings.

To pursue this goal, we train classification models using a combination of techniques for mitigating the class-imbalance problem, reducing the predictors set to keep only the most relevant for the task, and tuning model-specific hyperparameters. We adopted three algorithms, Elastic Net, Random Forest, and Multilayer Perceptron, comparing multiple performance metrics and motivating the aggregation of these models into a single probability-averaging ensemble.

The remainder of this work is organized as follows: Chapter 2 describes the basic concepts relevant to our methodology and related-work analysis; Chapter 3 reviews studies from the literature that have approached related problems; Chapter 4 motivates and introduces our solution to the problem in question; Chapter 5 explores the findings of an application of our solution; and, finally, Chapter 6 critically evaluates the impacts, the relevance and some improvement opportunities of the study as a whole.

2 THEORETICAL BACKGROUND

This section describes algorithms and techniques commonly employed in supervised machine learning, specifically in classification problems. We also discuss the *class imbalance problem*, mentioning and comparing mitigation techniques and elaborating on methods to evaluate results under class imbalance.

2.1 Classification Algorithms

Supervised learning consists of extracting knowledge by obtaining patterns from curated domain data. More specifically, through a process called "training" or "induction" of models, a function that relates independent input variables to a dependant output is approximated. This is made possible by the fact that in supervised learning the data points are examples of the true mapping from input to output variables. When the output of the models is categorical (that is, its values are always of a finite set of *classes*) this task is then named supervised learning classification, as opposed to the alternative of continuous values that is named regression. Thus, these models are useful for their ability to classify or predict new occurrences of a phenomenon, based on the values of the independent variables. Several algorithms are available for the creation of classification models, each with interesting and important peculiarities. In this section, we review the ones employed in the methodology of our study.

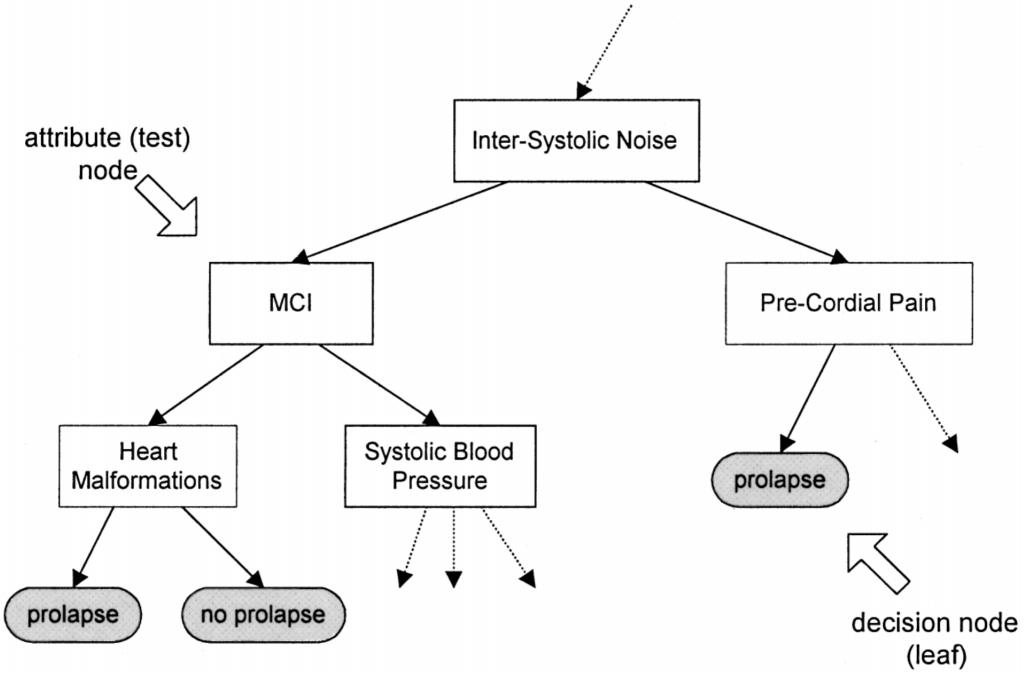
2.1.1 Decision Trees

Decision trees (DTs) are widely used in machine learning applications in general (KOTSIANTIS, 2013), and in the medical domain (BURKE; AMMERMAN; JACOBUCCI, 2019), for their potential for human understanding and classification effectiveness (PODGORELEC et al., 2002). Instances are classified in decision trees by following a path from the trees' root down to one of its leaves. At each node, as depicted in Figure 2.1, the instance has its value for some feature compared to some inducted rule of classification to define the decision path (KUBAT, 2017).

The problem of finding the shortest trees with the best rules and tests of variables is approached (though not optimally solved) with different *heuristics*, mainly distinguished

by their *node-splitting* criteria (e.g. *information gain*, *gain ratio*, and *Gini value*) (KOTSIANTIS, 2013). Growth in size and complexity in decision trees tends to lead to losses in interpretability generality, tending to overfit training data, which is the reason why techniques like *pruning* and *feature selection* are applied (KOTSIANTIS, 2013).

Figure 2.1 – A depiction of a decision tree



Source: Podgorelec et al. (2002)

2.1.2 Elastic Nets

The Elastic Net (EN) is a regularization and feature selection method for linear or logistic regression that essentially combines two other regression methods called ridge and lasso. It improves on lasso by generalizing it (ZOU; HASTIE, 2005). The technique introduces a *grouping effect* with which correlated variables stick together in regards to their relative contribution to the prediction, which can be estimated for the same usefulness as p-values for assessing their importance (BURKE et al., 2018). The nature of the algorithm begs the definition or selection of two hyperparameters: lambda (the regularization parameter) and alpha (the ridge-lasso-mixing parameter).

2.1.3 Artificial Neural Networks

Artificial Neural Networks (ANNs) are machine learning algorithms based on a graph structure of nodes (called *neurons*) interconnected by weighted links. Each neuron calculates an output that is a combination of its inputs (the network actual inputs or the outputs of other neurons) based on linear or non-linear functions. Arranged in layers, these structures are also called *multilayer perceptrons*, using *forward propagation* to classify instances. This process in the chain calculation of the neurons' activations up to the *output layer*, where the prediction is decided to be class represented by the neuron with the highest value. Based on a calculated error, the network updates its weights by the process called *backpropagation*, implementing error minimization by *gradient descent* (KUBAT, 2017). Since ANNs are a large family of algorithms, many hyperparameters can be explored depending on the exact variation employed, but arguably the most basic ones are the number of hidden layers and neurons per in each hidden layer of the network.

2.2 Ensemble Learning

Supervised learning algorithms can be grouped to improve prediction performance and robustness, constituting an *ensemble*. This can be achieved by several different approaches, each with varying degrees of complexity. The emerging trade-off between lesser interpretability by increased complexity and predictive power should always be considered, as its asymmetry is dependent on the domain, the dataset, and the underlying algorithms involved. That said, it has been shown that, in general, good results can be achieved with relatively simple combinations of predictors (CLEMEN, 1989). Commonly employed techniques of lower complexity include:

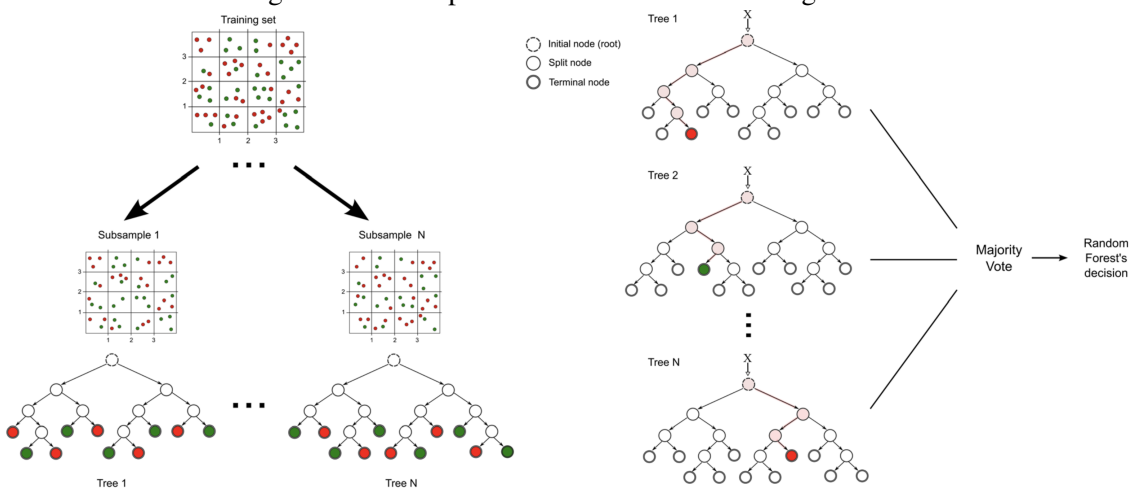
- majority voting: when the classification outcome is defined by the most predicted class among the ensemble's constituents;
- weighted averaging: when the class with the highest linearly-combined probability between constituents is chosen.

2.2.1 Random Forests

Random Forests are defined as ensembles of decision trees (grown considering random vectors) that classify instances by majority voting(BREIMAN, 2001). Random forests are also useful for generating ranks of features importance by assessing the automatically computed variable importance measures (VIMs) of the forest (BOULESTEIX et al., 2012).

This ensemble is trained using *bagging*, an acronym for *bootstrapping* and *aggregation* (majority voting). Bootstrapping works by composing a number of training subsets by sampling from the original training data with replacement, then inducing a tree classifier from each (KUBAT, 2017). This process is illustrated in the left half of Figure 2.2, while the right half shows how inference takes place. Bagging performance depends on how well the classifiers perform for different examples (the dependence between classifiers), and the individual performance of each (BREIMAN, 2001), coupled with the reliance that with sufficient randomness and forest size individuals' errors will be corrected by others (KUBAT, 2017).

Figure 2.2 – A depiction of the random forest algorithm



Source: Machado, Mendoza and Corbellini (2011)

Injecting randomness in the procedure through *random feature selection* in node splits can yield benefits in prediction performance including enhancement of generalization capacity and resilience to noise (variance), with overall results comparable to Adaboost (another tree-based ensemble algorithm, but based on boosting) (BREIMAN, 2001). Several variations on the classical random forests have been proposed, one of which being the Extremely Randomized Trees (GEURTS; ERNST; WEHENKEL, 2006).

This algorithm differentiates itself from the others by its approach to node-splitting, which (as the name suggests) strongly randomizes the choice of attribute and cut-point.

2.3 Feature Selection Wrappers

Classification models can be improved upon, in regards to their interpretability and inference capability, by being embedded in a feature selection wrapper algorithm. Different approaches to this optimization heuristic have been proposed, implemented, and tested. A common and intuitive approach is Feature Elimination (FE), which has been shown to improve performance and reduce model complexity (SVETNIK et al., 2004; GUYON et al., 2002). It consists of repeatedly eliminating irrelevant features to improve interpretability and performance upon model retraining. The variable importances can be assessed once, the first time the model is trained, or reassessed each time the model is retrained, configuring a Recursive Feature Elimination (RFE) (SVETNIK et al., 2004).

2.4 Class Imbalance

The *class imbalance problem* pervades machine learning research and applications in general (CHAWLA, 2009). A consequence of training from datasets skewed towards a majority class, it hinders the performance of classifiers by introducing biases in standard induction algorithms that work best with balanced distributions of classes (JAPKOWICZ, 2000). This situation consistently manifests itself in medical diagnosis contexts (VLUYMANS, 2019), where it is common to have many more examples of some healthy (generally called *negative*) class than the unhealthy (generally called *positive*) one. Moreover, the domains of application where these skewed distributions arise frequently have an inherently higher cost of mispredictions of the positive minority classes (e.g. labeling an ill instance as healthy) (KOTSIANTIS, 2013). Nevertheless, many works dealing with applications in mental health do not confer the appropriate attention to the evaluation and treatment of class imbalance (BURKE; AMMERMAN; JACOBUCCI, 2019). Vluymans (2019) separates techniques for the mitigation of class-imbalance effects into four groups: data level preprocessing (*external)*, algorithm-level (*internal*) approaches, and cost-sensitive and ensemble learning. The two latter are combinations or special applications of the two formers. While data-level preprocessing consists of modifying the

original dataset to have less class imbalance, algorithm-level approaches focus on slightly altering the mechanisms of standard learners to reduce their bias towards the majority class.

Data-level methods can operate by oversampling the minority class, undersampling the majority class, or combining the two. Undersampling and oversampling can be done randomly or focused (KUBAT; MATWIN, 1997; LAURIKKALA, 2001), with reportedly mixed results of improvement of one over the other (JAPKOWICZ; STEPHEN, 2002). Oversampling methods work by copying or generating new instances. In the latter category, the *SMOTE* algorithm (CHAWLA et al., 2002) reportedly yields considerable performance gains from synthetically creating minority class examples. In effect, it enlarges decision regions that contain nearby minority class points.

Cost-sensitive learning consists of enhancing the importance of minority classes by altering the amount of cost to different types of errors, classes, or features (VLUYMANS, 2019). One example of this type of approach is MetaCost (DOMINGOS, 1999), which wraps general classifiers training with the introduction of a *cost matrix* that assigns different costs to false positive and false negative errors.

Lastly, in a broad context, ensembles are usually employed for improving accuracy, which by itself would not be of much help for the class imbalance problem. Thus, ensemble-based solutions for the class imbalance problem introduce specific adaptations, usually done by embedding data-level preprocessing methods or cost-sensitive learning in the induction of the ensemble (VLUYMANS, 2019).

2.5 Model Evaluation

To correctly evaluate the performance of a classification model, it is necessary an understanding of some key concepts, briefly explained in this section. Evaluation can be done with multiple techniques, assessing different metrics - each having peculiarities in relevance and interpretation depending on the specific context. Deciding the correct metrics to employ is crucial when dealing with class-imbalanced data because not every metric adequately describes classification quality reliably in this scenario.

2.5.1 Evaluation Metrics

Since all performance metrics for classifiers correlate to the model's predictions of the labels (or classes) of instances (also called examples), it's useful to employ some common naming of the possible classification cases. A binary classifier can either predict an instance's label correctly (inferring its true class) or incorrectly (inferring a false class). When a positive class is inferred correctly, it's called a *true positive*. Conversely, a misprediction of a positive class is a *false negative*. The terms *true negative* and *false positive* have analogous definitions. For convenience, throughout this paper, in the context of some set of predictions, we'll refer to the size of the subsets of *true positives*, *true negatives*, *false positives*, and *false negatives* as *TP*, *TN*, *FP*, and *FN*, respectively.

The first metrics that enable quantitative evaluation of a classifier are its accuracy and error rate. Both concepts are very intuitive: accuracy is the rate of correct labeling of a set of predictions. On the other hand, the error rate of this batch of inferences is the frequency of misclassifications. Accuracy (Acc) and error rate (E) are therefore correlated (by $Acc = 1 - E$), and can be defined as follows (KUBAT, 2017):

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.1}$$

$$E = \frac{FP + FN}{TP + TN + FP + FN} \tag{2.2}$$

However, error rate and accuracy are not good estimators for the performance of learners, especially in the presence of class imbalance. It's easy to see why by imagining the following scenario. Suppose some model learns from a dataset composed of 80% negative-class instances (healthy dogs) and 20% of positive-class instances (diseased dogs). If measured by accuracy, this learner could be deemed quite successful by always predicting negative labels (asserting that all dogs it analyses are fine). Yet, that predictor would be disastrous for identifying ill specimens, worse than useless - since animals bearing sickness would not be treated, after confidently being labeled healthy. Thus, accuracy can mask a learner's problematic performance profile and misjudge its usefulness, especially with imbalanced datasets and false-positive-critical applications - a frequent case in clinical medicine (VLUYMANS, 2019).

Alternatively, we could evaluate classification capacity by measuring *precision* (Pr) and *recall* (Re). These metrics improve on accuracy by focusing on a class of interest,

providing better insight from *TP*, *TN*, *FP*, and *FN* (KUBAT, 2017). Intuitively, precision is the probability that a learner is correct when inferring a positive class, whereas *recall* measures if true positives are identified accordingly. Depending on the context of the application, one would train its predictor with special attention to one of *precision* or *recall*. Precision may be more desired in applications where we do not want our model to yield many false positives. For example, in identifying criminals from pictures, we would not want to accuse innocents. On the other hand, in studies of clinical conditions, *recall* is generally the most important indicator, since we do not want to miss any positive instance by classifying it as a (false) negative. We can see that minimizing false positives yields higher precision and minimizing false negatives yields higher recall by analyzing the formulas for these quantities:

$$Pr = \frac{TP}{TP + FP} \tag{2.3}$$

$$Re = \frac{TP}{P} = \frac{TP}{TP + FN} \tag{2.4}$$

When a learner changes its profile to be more prone to label instances as positive in general, it will probably increase *recall* at the cost of reduced *precision*. Thus, a relevant performance metric is the F_β-Score, which combines precision and recall in a weighted harmonic mean. The β factor defines which of the two quantities will have more importance in the measure. The F_2-Score is particularly interesting in clinical applications where positive instances are fewer in number and at the same time more important to be noticed (i.e. FNs are more costly). This is because this metric is affected by changes in the class distribution (THARWAT, 2018), and has an intuitive meaning of its user valuing recall two times more than precision (C. J. van Rijsbergen, 1977). The F_2-Score can be reduced to the following equations:

$$F_2 = \frac{5 * TP}{5 * TP + 4 * FN + FP} \tag{2.5}$$

$$F_2 = \frac{5 * Pr * Re}{4 * Pr + Re} \tag{2.6}$$

Other important metrics for the medical domain are *sensitivity* and *specificity*. Although quite traditional to the field, these indicators are rarer in machine learning studies in general (KUBAT, 2017). *Sensitivity* is another name for *recall*, while *specificity* is

essentially *recall* for the negative class. More precisely:

$$Se = \frac{TP}{P} = \frac{TP}{TP + FN} \tag{2.7}$$

$$Sp = \frac{TN}{N} = \frac{TN}{TN + FP} \tag{2.8}$$

Just as there is often a trade-off between precision and recall, sensitivity and specificity can vary at odds with each other. This inference-behavior changes, reflecting the balance between sensitivity and specificity under different hyperparameters, are illustrated by *receiver operating characteristic* (ROC) curves (KUBAT, 2017). These graphs portray sensitivity (also called *true positive rate*) on the y-axis versus a proxy for specificity named *false positive rate* (also called *false alarm rate*) on the x-axis. The *false positive rate* is defined as (FAWCETT, 2006):

$$Fpr = 1 - Sp = \frac{FP}{N} = \frac{FP}{TN + FP} \tag{2.9}$$

ROC space has some notable points, each translating to a characteristic behavior in classification profile (FAWCETT, 2006):

- (0,0): to always infer the negative class
- (1,1): to always infer the positive class
- (0,1): to always infer the correct class

Based on the analysis of ROC curves, it is customary to consider a classifier's *area under the ROC curve* (*AUC* or *AUCROC*) - a scalar value (FAWCETT, 2006). A perfect classifier will have *AUC*=1, since its ROC curve will be the point (0,1). The area under the ROC curve can be interpreted as a model's capacity to distinguish instances of different classes as so - it may also be interpreted as the probability of ranking a positive random instance higher than a random negative one (FAWCETT, 2006). It is important to note, however, that the AUCROC is also subject to hiding the true performance of a model under class imbalance, although not as much as accuracy. With imbalanced classes, a high AUC could merely indicate a better capacity for identifying negative instances (BURKE; AMMERMAN; JACOBUCCI, 2019).

In conclusion, as each approach has its strengths and weaknesses, it is prudent to consider multiple metrics assessments to have a holistic evaluation of classification models.

2.5.2 Evaluation Methodologies

With the interest of inducing the best classifier in our reach, we are now ready to consider the methods for using performance metrics to achieve our goal of evaluating and choosing machine learning models. Several algorithms could be leveraged for that - hence we'll discuss baseline approaches, *random subsampling*, *N-fold Cross-Validation* (CV), and *stratification*. Nevertheless, The general idea is to perform a *model selection*, choosing the one with parameters that yield the best results, or to perform a final evaluation purely for estimation of performance on unseen data.

The ideal approach for model selection and evaluation would be to separate our dataset into three parts - for training, validation, and tests (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). The validation set would be used to assess the performance of models trained with the training set, while testing instances would be reserved for evaluating the chosen best classifier. There is also the approach to separate the data into only two sets (training and testing) (KUBAT, 2017). Although this could lead to test-error underestimation, this is done to check the generality of the learner (HASTIE; TIBSHIRANI; FRIEDMAN, 2009) - or else we could deem desirable a model overfit to the training data, with no guarantees of extending its prediction capability to new instances. Both methods are suitable for situations with data abundance. With smaller datasets, the distribution of instances within each subset is subject to be substantially different from the original dataset (KUBAT, 2017) - especially with class-imbalanced data.

As an alternative to these standard approaches, one could use *random subsampling* to mitigate the limitations in distribution representation fidelity of the subsets of the baseline methods. The difference in this procedure is that the training and testing separation is repeated multiple times, each composing different subsets randomly (KUBAT, 2017). This yields, each time, some performance metrics, which are averaged to finish the evaluation.

On the other hand, instead of randomly repeating the training-testing division, the dataset could be divided into K equally-sized parts (also called *folds*, typically 5 or 10). Then, for each fold, have a round of training, evaluations, and improvements, using the fold as a validation set and the others for training (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). One key (advantageous) difference that arises from this approach, called *K-fold cross-validation*, compared to *random subsampling*, is that each of the training sets (and test sets) is disjoint from the others (KUBAT, 2017). This concept can be extended to in-

corporate a repetition (then being loosely named N-times K-fold cross-validation), where new cutting-points to separate the folds are selected each time, to achieve assessments with more statistical robustness.

Although promising, CV is subject to problems when dealing with highly imbalanced datasets; the distribution of instances within folds could bear little resemblance to the complete dataset (KUBAT, 2017). *Stratified* approaches mitigate this by guaranteeing that all the folds have the same class ratio, respecting the original data's distribution.

3 RELATED WORK

To better understand the scenario of machine learning studies in medicine in recent years, and specifically in the field of mental health, a systematic literature review (SLR) was brought to analysis. Burke, Ammerman and Jacobucci (2019) focused its attention on suicide-related applications and employed a methodology oriented by inclusion and exclusion criteria, where papers from multiple sources were processed through a systematic pipeline.

The SLR targeted papers published until February 2018 in the PsycINFO, PsycARTICLES, ERIC, CINAHL and MEDLIN databases. From an initial set of 288 retrieved studies, derived by search terms of methodological (e.g. "machine learning", "data mining" and "big data") and domain ("suicide", "self-injury", "suicide ideation") imprint, only 35 papers met the inclusion criteria and went on to further analysis.

As for the conclusions of the review, we are interested mainly in its insights on the successes and difficulties of previously employed ML algorithms in problems with similar characteristics to ours - to predict suicide ideation. The study was able to demonstrate the potential of ML algorithms in the mental health field; it was shown that this branch of technology can greatly improve the prediction performance of mental disorders. Furthermore, the exploration and finding of predictor variables replicate established results but also find novel variables and identifies subgroups of interest. This SLR also emphasizes the importance of the interpretability of the models and their trade-off over performance. To that extent, this subset of the literature tends to favor simpler predictors like decision trees. As Figure 3.1 shows, over 60% of the analysed papers employed DTs, while ANNs (a generally less interpretable model, though a strong performer in a myriad of applications) appeared only in about 10% of the articles.

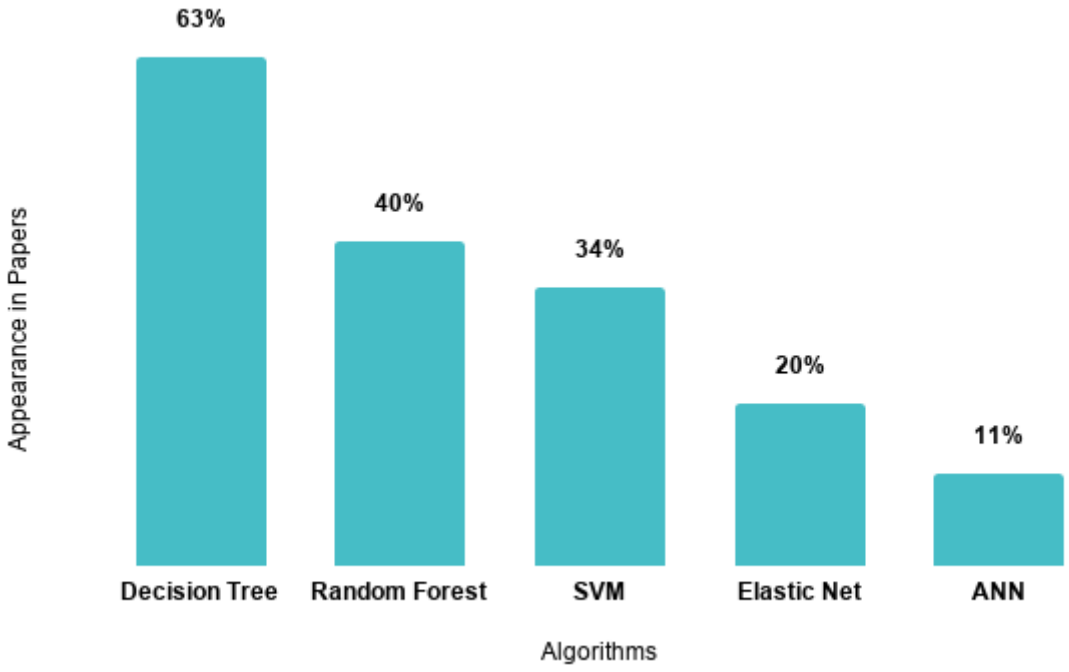
Burke, Ammerman and Jacobucci (2019) separate the study of suicide into distinct (though sometimes intersecting) categories: suicide death, suicide attempt, suicide planning, and suicide ideation. SI studies (a small subset of only 10 papers in this SLR) used both cross-sectional and longitudinal designs (with one to five years follow-ups), considering data from population samples distinct in age, locale, and mental health. The trained models incorporated, on average, 32 attributes (ranging from 3 to 62), and each study used one to four (M=1.6) ML methods. Almost all of the ten papers used structure data, safe from one that employed an NLP approach. As for the performance metrics, however, the inconsistency in strict and standardized reporting is evident. Some papers

do not show any performance estimation whatsoever, while others generally do not share the same indicators and scores (reporting exclusively AUC-ROC, accuracy, or sensitivity and specificity, etc.). Along with the small paper sample size, this makes the effort to derive statistics on their models' performance futile.

Nevertheless, the reported metrics are indicative of the potential of employing the algorithms from Figure 3.1 in the prediction of suicide ideation. As Burke, Ammerman and Jacobucci (2019) compiled, suicide ideation prediction studies reported AUC values ranging from 0.8 with DTs (HANDLEY et al., 2014) to 0.92 with DTs and RFs (GRADUS et al., 2017), and sensitivity and specificity reaching 0.88 and 0.94 (JUST et al., 2017) with decision trees, neural networks and Support Vector Machines (SVMs), although considering only 34 data points, half of which were of the positive class).

Besides the gaps and inconsistencies in reported performance, Burke's SLR brings to light the lack of consistency and depth in treatment and discussions of class imbalance in training data in the analysed studies. While many papers report AUC, which better describes the performance than just the accuracy, it still falls short in completely describing the actual performance of the predictor in the presence of class imbalance (BURKE; AMMERMAN; JACOBUCCI, 2019). And while some papers report precision and recall, few studies address the problem of unbalanced data in their sampling and training.

Figure 3.1 – Algorithms prevalence from Burke, Ammerman and Jacobucci (2019)



Source: Author

That said, even though the beneficial attribute selection (for data dimensionality reduction) and class-imbalance mitigation techniques are oftentimes lacking in classifications problems in the context of mental health (BURKE; AMMERMAN; JACOBUCCI, 2019), that is not always the case. An example of successful attribute selection is presented in (BARROS et al., 2017), which used a wrapper-based method in a Chilean mental-health patients dataset to reduce the feature set size from 343 to 22, reporting sensibility and specificity between 0.7 and 0.8, approximately. On the other hand, in the context of class-imbalance reduction, Schubach et al. (2017) employs both downsampling of the negative class and oversampling of the positive class (with the SMOTE algorithm) to balance data partitions and train a random forest on each, afterward combining them in what is described as an ensemble of ensembles. The models' performance estimation from Schubach et al. (2017) varied substantially depending on the evaluated dataset, achieving at best around 0.7 area under a precision and recall curve, but at worst barely over 0.4. That said, the calculated AUCROC was always around 0.98 (again showing a problem in relying only on this metric).

Since the afore-mentioned SLR did not review works after February 2018 (BURKE; AMMERMAN; JACOBUCCI, 2019), it is also worth mentioning some related works developed since then - again, searching for insights on their achievements and limitations, bringing to light their similarities to our study.

With a similar problem to the one our work aims to solve, future risk of suicidal thoughts was chosen as the outcome variable to the models of Roy et al. (2020), although the employed data was unstructured (text, from Tweeter). The study considered a vast amount of tweets per individual, but on the other hand, the suicide ideation cases in the dataset were less than 300. That said, the study achieved a high AUCROC value of 0.88 using a model that combines the outputs of several neural networks using a random forest.

Using structured data from Korean population samples, two studies (OH et al., 2020; JUNG et al., 2019) also tackled problems akin to ours, with similar approaches too. Both proposed to solve the problem of classification of suicide ideation (with our without suicide attempt), but with different datasets. The first focused adults (OH et al., 2020) and the second on adolescents (JUNG et al., 2019). With over 16,400 instances, the Korea National Health and Nutritional Examination Survey (KNHNES) data employed in Oh et al. (2020) is more than four times larger in this regard than our restricted (to common mental disorders) sample of the ELSA-Brasil dataset. On the other hand, although the Korean Young Risk Behavior Web-basedSurvey (KYRBWS) dataset from Jung et al. (2019)

included data from over 62,200 adolescents, Jung et al. (2019) employed a drastic down-sampling strategy in its preprocessing, achieving class balance to the cost of reducing the instances count to approximately 15,300 - just as the data in Oh et al. (2020), amounting to approximately 400% of our dataset size.

As far as assessed variables are concerned, a key difference between a fully automatic approach to variable selection (as we propose in Chapter 4) and the one from Oh et al. (2020) is that the study resorted to filtering based on the manual inspection and review of two health professionals, reducing the dimensionality from 800 variables to only 48. That said, Oh et al. (2020) also employed wrapper-based feature selection in their computational methodology, in a similar way to our efforts in that regard.

In regards to classification qualities, the best model from Jung et al. (2019), an artificial neural network, was estimated to have an AUCROC of almost 0.88, with sensibility of around 0.81 and specificity close to 0.77. The most successful algorithm used in Jung et al. (2019) (called *extreme gradient boosting*) yielded very similar results, with slightly lower AUCROC and sensitivity and higher specificity. Since these two studies also report the estimated positive predictive values (precision) of their classifiers, we can also calculate their F_2-Scores, which amount to 0.48 (OH et al., 2020) and a distinguishing value of 0.84 (JUNG et al., 2019).

Finally, in Brazil, some studies have already employed machine learning techniques over the ELSA-Brasil dataset. Both Brunoni et al. (2020) and Librenza-Garcia et al. (2020) treat as dependant variables the depression incidence or persistence - in other words, how interviewees' depression evolves in the four years between the two ELSA-Brasil waves. While Brunoni et al. (2020) focused on analyzing the risk factors of depression, Librenza-Garcia et al. (2020) performed a similar classification task to our own, with an elastic net reportedly having sensitivity of 0.67, specificity of 0.78, and AUCROC of 0.79. Since Librenza-Garcia et al. (2020) also reported the precision of its model, we can estimate its F_2-Score as 0.45.

A summary of the performance of some of the studies mentioned in this chapter is presented in Table 3.1, which has empty cells for the metrics that were not reported and could not be derived. The works included in the table are the ones that presented the most similarities to ours in scope and approach and are ordered by the F_2-Score and the AUCROC. Although the best performance is attributed to an application of the eXtreme gRadient Boosting (XGB) algorithm (JUNG et al., 2019), it is worth noting that this is not necessarily the most decisive factor to its success: the study employed the largest

dataset between the ones presented in the table by a substantial margin, and was able to circumvent the class-imbalance problem by discarding a huge chunk of the data and yet remain with a high number of instances.

Table 3.1 – Summary of performance estimates of related works

Paper	*Algorithm*	*F_2-Score*	*AUCROC*	*Sensitivity*	*Specificity*
A	XGB	0.84	0.86	0.79	0.79
B	ANNs + RF	0.71	0.88	0.80	0.79
C	ANN	0.48	0.88	0.81	0.77
D	EN	0.45	0.79	0.67	0.78
E	RFs		0.98		
F	RF		0.92		
G	SVM			0.77	0.79

Source: The Author
A: (JUNG et al., 2019);
B: (ROY et al., 2020);
C: (OH et al., 2020);
D: (LIBRENZA-GARCIA et al., 2020);
E: (SCHUBACH et al., 2017);
F: (GRADUS et al., 2017);
G: (BARROS et al., 2017).

4 METHODOLOGY

This chapter describes our methodology to approach the problem of prediction of suicidality using machine learning algorithms. First, we present the dataset used (the ELSA-Brasil study) and the steps adopted during its cleansing. Next, we describe our model-induction pipeline for a set of supervised learning algorithms of interest. Finally, we explain our strategies for model evaluation and for building an ensemble classifier.

Our proposal to approach the problem of prediction of suicidality can be summarized in three steps. First, the data (in our case, from the *ELSA-Brasil* study) is cleansed and minimally prepared. Afterward, it is used by a model-induction pipeline, for a set of supervised learning algorithms of interest. Finally, the models are evaluated and combined in an ensemble (which is also evaluated in the same manner).

4.1 Dataset

The Brazilian Longitudinal Study of Adult Health (ELSA-Brasil) dataset is a cohort of Brazilian adults (public universities' employees) in a longitudinal study with a follow-up of about 4 years (SCHMIDT et al., 2015; AQUINO et al., 2012). Up to now, two waves of interviews and examinations have been conducted: the first from 2008 to 2010, the second from 2012 to 2014 (OLIVERA et al., 2017). The baseline assessment included 15105 participants, with about 3000 assessed variables (although the availability of the attributes is restricted depending on the application). Given the scope and context of our application, after requesting data access with the interest of investigating suicidality, a version of the ELSA-Brasil dataset was made available to this study that includes 2288 "baseline" features.

Moreover, the ELSA-Brasil study assessed non-psychotic psychiatric morbidity using the Clinical Interview Schedule-Revised (CIS-R) (NUNES et al., 2016; LEWIS et al., 1992), which makes available information regarding depressive and/or suicidal thoughts on the cohort. The full version of the applied CIS-R questionnaire is made available in annex B. Although the full array of 176 questions is not explored in this section, the ones found to be of most importance are discussed in Chapter 5.

In our work, we analysed only the first wave of the study and focused on a specific population of the ELSA-Brasil dataset, corresponding to the individuals presenting any common mental disorders, indicated by the *mentalvar_A_TMC* feature. The reason for

restricting the dataset is that our goal is not to develop a model that identifies the presence of suicidality in the general population, as this would have limited utility, but rather to identify this condition for a population at risk. Individuals with CMD are probably under the assistance of mental health professionals, who could act upon the suggestion of machine learning classifiers that the patient is prone to suicidal ideation and, by proxy, suicide attempts. Besides, we also avoid broadening too much the training dataset to the point where class imbalance becomes too severe to handle, as the prevalence of suicidality is about 8.53% in this cohort. Although there is a clear trade-off between the ratio of instances in the positive class and the total number of instances in the training data, we apply this restriction since in our domain it is indispensable to make correct predictions for the positive class. As a result of excluding the instances without CMD, our number of data points is reduced from 15105 to 4039, incurring the loss of 169 (13.11%) positive class instances, but raising the probability of the class from 8.53% to 27.73%.

4.1.1 Input Variables - Features

Besides socio-demographic and economic attributes, the ELSA-Brasil study assessed the participants' health in different aspects and means, from electrocardiograms to retinal photographs and mental health questionnaires (SCHMIDT et al., 2015). The available data in ELSA-Brasil includes attributes previously described in the literature as associated with suicide ideation, including gender, age, marital status (NOCK et al., 2008), socioeconomic status (GUNNELL; PLATT; HAWTON, 2009; MELTZER et al., 2012b; MENEGHEL et al., 2004), physical activity, alcohol consumption, self and family education levels (SOUZA et al., 2010) and emotional difficulties (e.g. deaths of close relatives), social capital variables, pain conditions, chronic diseases, obesity and body mass index, the existence of physical disabilities (MELTZER et al., 2012a), and sexual orientation (SILENZIO et al., 2007).

As predictors of our models, we used all 2288 variables from the baseline dataset. Besides, we also included variables from the CIS-R questionnaire, which, after excluding the variables used as outcomes, consists of 173 variables. Therefore, the total number of predictors before data cleansing was 2461 features.

4.1.2 Outcome Variables - Labels

To represent the concept of suicidality, which is the outcome of our classifiers (i.e. a binary factor indicating the presence or absence of suicidality for a given instance), we assess and combine hopelessness, "taedium vitae", and suicidal ideation in logical disjunction, a boolean *OR*. These variables are, respectively, indicated by the responses to the CIS-R's H6, H8, and H9 questions:

- CIS-R H6 (hopelessness): *Have you felt hopeless at all during the past seven days, for instance about your future?*
- CIS-R H8 (taedium vitae): *In the past week have you felt that life isn't worth living?*
- CIS-R H9 (suicidal ideation): *In the past week, have you thought of killing yourself?*

Although the formulation of the questions implies a naturally binary "yes or no" response, both H8 and H9 have a third option, which is of presenting taedium vitae or suicidal ideation but not exactly in the last 7 days. To this study, the third-option responses are considered as positive ones (i.e., as if the person had responded to the questions with a "yes"). These variables had several missing values since the majority of the interviewees did not reach the point of being asked their respective questions - they responded negatively to prior ones that are more general, such that we can assume there is an implicit negative answer to the more specific ones which we are interested. Thus, absent values for CIS-R's H6, H8, and H9 variables were inferred as negative entries.

4.1.3 Data Cleansing

As a stride in the direction of improving data quality for the induction of the models, it is desirable to validate and clean (or *wrangle*) the data before usage. Kandel et al. (2011) describe data wrangling and analysis as an iterative process consisting of cleansing, merging, adapting, and evaluating the data.

Considering that the ELSA-Brasil study has several variables obtained by non-linear interviews (skipping questions depending on answers), some instances present missing values. Although our methodology includes a mechanism for inference of unavailable (NA) values (see Section 4.2), we considered important for data fidelity and for model quality to leave aside variables that have more NAs than a certain threshold - which we set as 10% in this study.

Our solution does not include any natural language processing techniques, thus all free-text variables in ELSA-Brasil were removed. We provide a list of textual variables removed in Table A.1.

Finally, and perhaps most importantly, given the chosen CIS-R questions (H6, H8, and H9) to be combined into our outcome label, we removed the predictors that indirectly make use of these variables - except for *mentalvar_A_ESCORETOTAL*, which is adapted to be the sum of the numerical value of answers in the CIS-R H section excluding H6, H8, and H9. These are, in the dataset, the variables shown in Table 4.1. This step is crucial to avoid data leakage in model training, where our input variables (i.e. predictors) would improperly contain information about the outcome to be predicted.

Table 4.1 – Attributes removed for introducing information leakage

Attribute	*Description*
mentalvar__TMAD	Mixed anxiety-depressive disorder (MADD)
mentalvar_a_DEP	Major depressive disorder (MDD)
mentalvar_A_DEPGRAVE	Severe MDD
mentalvar_A_DEPLEVCSINT	Mild MDD with somatic symptoms
mentalvar_A_DEPLEVSSINT	Mild MDD no somatic symptoms
mentalvar_A_DEPMODCSINT	Moderate MDD with somatic
mentalvar_A_DEPMODSSINT	Moderate MDD no somatic
mentalvar_A_ESCORETOTAL	CMD score (continuous)
mentalvar_A_SINTIDEIADEP	Depressive thoughts symptoms
mentalvar_A_TMC	Common Mental Disorder (CMD) (bin)
mentalvar_A_TMCGRAV	Common Mental Disorders (CMD) (3 levels)
mentalvar_MDD_trajectories	group(a_DEP b_DEP)
mentalvar_only_incident	Incident MDD
mentalvar_only_remitted	Remitted MDD

Source: The Author

Upon removal of these variables, it is desirable to supply the models with some sensible substitute to the data that was taken away, since the variables were of great importance for the clinical diagnosis of suicidality. Since interpretations of responses to the H section of CIS-R were summarized in the removed features, the idea is that the learners would instead interpret these relations by themselves given the necessary data. Thus, our cleansing needs steps to also impute missing values in CIS-R answers, which were frequent given the interview is non-linear, so they are not removed by our afore-mentioned 10%-maximum NA filter. For this task, we realized that the majority of the answers could be reliably inferred from the context, if the person responded negatively to a question X, then some question Y which only makes sense if X was answered positively can be

inferred to be negative too.

In conclusion, because of absent values, textual input, and information leakage, our cleansing procedure reduces the number of predictors to be employed in the model-fitting process, represented in Table 4.2, and does not change the number of instances in the data. The final high-level quantitative characteristics of our dataset are summarized in Table 4.3.

Table 4.2 – Number of variables in light of dataset cleansing process

Attribute Set	*Set Size*
Total (uncleansed)	2463 (100%)
Removed (information leakage)	13 (0.69%)
Removed (NA excess)	773 (31.38%)
Removed (free-text)	47 (1.91%)
Remaining (cleansed)	**1626 (66.02%)**

Source: The Author

Table 4.3 – Main quantitative characteristics of cleansed dataset

Dataset Characteristic	*Value*
#Instances	4039
#Attributes	1626
#Positives	1120 (27.73%)
#Negatives	2919 (72.27%)

Source: The Author

4.2 Pre-processing, Training, and Evaluation

The goals in the proposed approach to produce our learners are: to attain interpretability and prediction performance.

With this goal in mind, we define our methodology considering the particularities of our dataset that make the development of predictive models particularly challenging. The adopted dataset is highly skewed towards a majority class of non-suicidal instances, such that our procedures require the usage of techniques to mitigate class imbalance biases and to appropriately evaluate model performance in such a scenario. There is also the problem of having numerous features for each instance, which can make it harder for the learners to have both good predictive qualities and interpretability.

4.2.1 Pipeline and Pre-Processing

The ordered combination of the main steps of our approach is summarized in the three-staged pipeline diagram of Figure 4.1. The highest layer, of evaluation, encompasses the whole cleansed dataset and orchestrates the fitting of models (abstracted by lower layers) and their performance assessment for each supervised learning algorithm. The RFE procedure has its own layer (in blue), where the data is each training fold of the evaluation phase (also in blue), so the final test data is never used in induction. Finally, within the scope of the yellow data boxes, the RFE models training folds, the most basic supervised learning layer (also in yellow) induces classifiers with hyperparameter tuning.

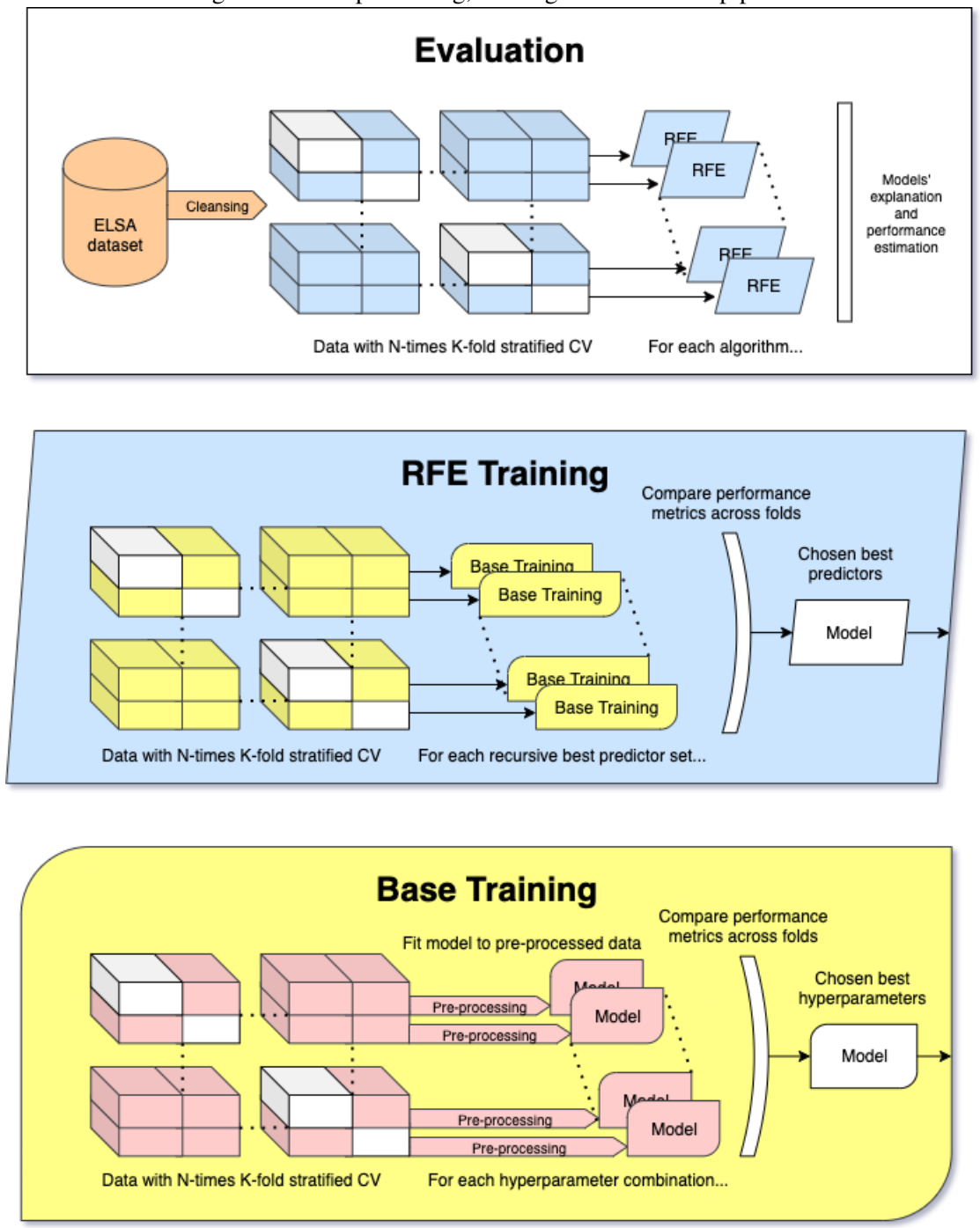
To avoid data leakage in the base learners induction, pre-processing is only applied in the lowest layer of our pipeline. If pre-processing were applied before, the training data in the lowest layer would have been processed in a manner that uses information from its corresponding test data. In this undesirable scenario, we could expect to have models with optimistic performance estimates that would perform (drastically) differently upon inferences on new data.

The preprocessing itself is proposed to be a sequential application of the following steps, each parameterized and described as general procedures:

- **downsampling**, removing negative instances until the class distribution is of a given ratio (e.g., 2N:1P, 3N:2P, etc - but not 1:1);
- **NA imputation**, inferring missing values using a function of the respective predictor values (e.g. the mean, or some classification or clustering ML technique, etc.);
- **near-zero variance (NZV) cut**, excluding useless predictors with a uniformity of values, defined by a parameterized quantitative criteria (e.g frequency distribution thresholds);
- **high-correlation filtering**, removing variables that are highly correlated to others (i.e. with correlation over a threshold) and that do not present distinct and useful information;
- **SMOTE**, synthetically creating positive instances derived from a fixed number of nearest neighbors (e.g. 3, 5, 7 etc) until the class distribution is of a given ratio (e.g. 1:1), (nearly) balancing the ratio of observations per class.

Note that the pre-processing steps of the pipeline do not necessarily promote per-

Figure 4.1 – Preprocessing, training and evaluation pipeline



Source: Author

fect class balance for subsequent model induction. This is controlled by adjusting the parameters related to the balance ratio, taking into account whether the supervised learning algorithm by itself can deal with the class imbalance to some extent.

4.2.2 Model Induction Approach

We also aim to reduce the number of attributes of the data used in training and predictions to have a more interpretable classifier, from which clinicians can obtain intuitive knowledge. In our work, we applied RFE to wrap "base learners" in a feature selection loop, where a model will have as the final predictor set the one that yielded the highest F_2-Score. Each iteration of feature elimination retrains the model and reassesses the variable importances rank so that only the best predictors are carried on to the next induction. Again, to avoid selection bias (AMBROISE; MCLACHLAN, 2002; REUNANEN, 2003) and for the other reasons presented in 4.2.3, the feature elimination procedure must be embedded in an N-times K-fold stratified cross-validation.

Finally, the essential and basic supervised learning procedure (which is wrapped in RFE) makes use of hyperparameter tuning by grid search (GS), where we select optimal models by performance comparisons (again, using the F_2-Score) of inductions done with several combinations of hyperparameters. Another cross-validation of the same sort as the ones just described must be applied here, for the same reasons.

4.2.3 Performance Assessment Approach

The F_2-Score is used as the optimization metric during model induction and as the main evaluation metric. In the latter case, other values are also calculated to estimate the generalization power of the classifier: the area under the ROC curve (AUCROC), the sensibility, and the specificity.

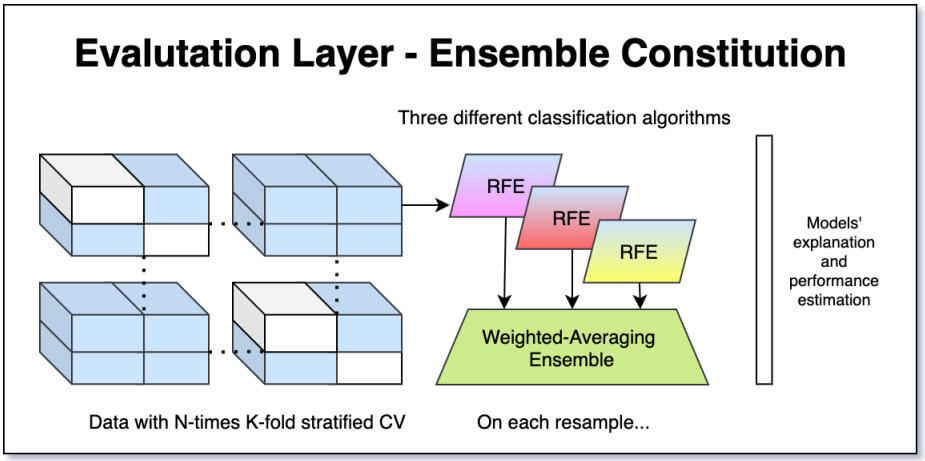
The F_2-Score is chosen as the most relevant measure not only for its intuitive meaning of valuing the positive class more than the negative one (which is essential for a suicide ideation classification, where the false negatives errors are the most costly) but also because it forces models not to sacrifice precision. This metric is often neglected in favor of specificity, thus in many cases being, without notice, quite low while the other is high. It shows possible deficiencies of having many errors in the positive predictions,

while specificity shows whether we have correctly found the true negatives within the negative instances. Thus, although the positive class requires the highest attention in our study, we must also assess both specificity and precision.

With a zeal for realistic and statistically-robust estimates of real-life performance, we employed a stratified N-times K-fold CV to separate our data, in our final evaluation mechanism.

4.2.4 Ensemble Composition

Figure 4.2 – Weighted-averaging ensemble constitution and evaluation



Source: Author

Lastly, our approach includes the composition of multiple models trained with different algorithms or pipeline parameters in an ensemble. This model is not trained by itself as a whole, its constituents are trained separately. For each CV resample of the pipeline's highest layer, the trained models simply have their inference outputs combined by weighted averaging of the classification probability, as illustrated in Figure 4.2. The idea is that the ensemble can compensate or mitigate each models' particular difficulties and consolidate their consensus. For that, it is desirable to make use of algorithms that provide this variability and diversity, although the nature of the pipeline already introduces a great mechanism of differentiation through RFE, such that models are not fit over the same predictors.

5 EXPERIMENTS AND RESULTS

In the following sections, our experiments are defined and have their results reported and analysed. We go over the minutia of the algorithmic details and the software and hardware resources employed in the conducted experiments, and afterward, we discuss how the proposed solution from Chapter 4 fared in this scenario.

5.1 Experiments Definition

The proposed methodology was implemented in the R language version 4.0.2 (2020-06-22), and the main libraries used were *caret*, *recipes*, and others from the *tidyverse* and *tidymodels* collections. Max Kuhn's *caret* (Classification And REgression Training) is a feature-rich machine learning framework (KUHN, 2008), which was crucial for the implementation of the model induction and evaluation pipeline, along with *recipes* and *rsample* that were leveraged for preprocessing, and resampling respectively.

On the data manipulation side, the main libraries used were *tibble* (for in-memory data model), *reader* (for data ingestion), *dplyr* and *tidyr* (for general tibble processing), *purrr* (for functional programming support), and *ggplot2* (for visualization), all part of the *tidyverse* collection.

As for the main supervised learning algorithms, we chose to run our experiments with three methods. Elastic net logistic regressions were selected so that we have linear models with good odds of reasonable interpretability. Additionally, as mentioned in Chapter 3, this algorithm was employed by Librenza-Garcia et al. (2020) over the dataset from the ELSA-Brasil study. The Elastic Net models were fit using the *glmnet* package implementation (FRIEDMAN; HASTIE; TIBSHIRANI, 2010).

Our second leaner of choice was the multilayer perceptron, as it is a generally well-performing and customizable technique and is among the most frequently used algorithms from the systematic literature review from Burke, Ammerman and Jacobucci (2019), as shown in Chapter 3. Its R implementation in this study was the *mlpML* from the package RSNNS (short for "an R port of the Stuttgart Neural Network Simulator") (BERGMEIR; BENÍTEZ, 2012), which although it is not the most efficient or the fastest available is sufficiently practical for our needs - we intended to keep the architectures simple.

Finally, random forests were selected to be trained too, as they are a quite popular

algorithm among medical applications of ML. Also, we are interested in RFs for their potentially higher complexity compared to the other two algorithms, for being an ensemble of several trees. Random forests, in our experiments, were induced using the *ranger* package, which is efficient and apt for high dimensionality data (WRIGHT; ZIEGLER, 2017).

Considering the afore-mentioned choices of algorithms, it would be reasonable to encode our attributes with one-hot encoding to avoid that the models infer ordinal relations in categorical data. That said, as our dataset is vast in features, we chose to not make use of this technique, mainly because it would multiply our number of variables which is already large, and manually analyzing which features to encode to reduce this effect would be costly.

As our pipeline requires lots of parameter definitions, we provide a list of the ones used in our experiments in Table 5.1. Arguably the most relevant one is the final ratio of class distributions, which was chosen to be 1:1 to balance the data. We decided to have the pipeline parameters fixed for every algorithm induction, as the opposite would be time-consuming.

To speed up the experiments, we used three different computers. Each one has an equal amount of RAM (16GB), but they differ in their processors and operational systems. The first one has a quad-core Intel Core i7-4770K CPU running in 7 threads over Arch Linux, and executed a pipeline run of *ranger*, taking about 7 days to finish. The second one runs Linux Mint Tara over an Intel Core i7-7700HQ quad-core CPU using 7 threads, which remarkably ran the procedures for *glmnet* in a single day. Finally, the third computer uses 14 threads in MacOS Catalina over an octa-core Intel Core i9-9880H and was used to train the *mlpML* multilayer perceptrons for a whole week.

5.2 Results Analysis

In the following subsections, we examine our learners' characteristics by focusing our attention on three different facets. First, Subsection 5.2.1 summarizes the performances obtained by our model and reviews the proposed weighted-averaging ensemble. Sequentially, Subsection 5.2.2 describes the trained models in terms of their tuned hyperparameters. Lastly, Subsection 5.2.3 presents the characterizations of the models in terms of their decision-making criteria, exploring the most determining attributes most for the suicidality classification.

Table 5.1 – Pipeline parameters used in experiments

Parameter	*Value*
Evaluation CV - K (folds)	10
Evaluation CV - N (times)	3
RFE Training CV - K (folds)	5
RFE Training CV - N (times)	2
Base Training CV - K (folds)	5
Base Training CV - N (times)	2
Downsampling - P-class ratio	33.3%
SMOTE - P-class ratio	50%
SMOTE - nearest neighbors	5
NZV Filter - Dominant value max prevalence	0.95
NZV Filter - Unique values min frequency	0.1
Correlation Filter - Threshold	0.9
Correlation Filter - Method	*Pearson*
NA Imputation - Method	Mean value
RFE Best Attribute set sizes	$(2^k)_{k=3}^{k=9}$
Elastic Net GS - Alphas values	0.1 , 0.325 , 0.550 , 0.775 , 1
Elastic Net GS - Lambda values	2e-4 , 9.2e-4 , 4.3e-3 , 2e-2 , 9.2e-2
Neural Network GS - Layer 1 #Units	1 , 2 , 3 , 4 , 5
Neural Network GS - Layer 2 #Units	0 , 1 , 2 , 3 , 4
Random Forest GS - Random-attributes set sizes	2, 17, 33, 48, 64
Random Forest GS - Node-splitting methods	*gini*, *extratrees*
Weighted-Averaging Ensemble - Weights	Equal (⅓, ⅓, ⅓)

5.2.1 Classification Performance Analysis

The most critical characteristic of our models is how competent they are at solving the task of classification of suicidality. To assess that information, we compare the elastic nets, multilayer perceptrons, random forests, and ensembles (having one of those for each final evaluation CV) with respect to their F_2-Score, AUCROC, sensitivity, and specificity. We display the mean and standard deviation measurements of these quantities in Table 5.2 (with entries ordered by F_2-Score), while Figures 5.1, 5.2, 5.3, 5.4 graphically compare the distribution of measurements of each metric over the cross-validation resamples.

Table 5.2 shows our ensemble approach was successful in increasing the overall performance robustness. It generally displays the best qualities of the other individual models in a single one. The averaging mechanism has the best F_2-Score and second-best sensitivity, paired with the ANNs, and the best AUCROC, together with the random forests. It also has the second-best specificity, although the score gap between it and the winner in that regard, the forests, is significant. In conclusion, the boxplots and the table

give us two crucial take-aways. On the one hand, our models are diverse and heterogeneous w.r.t. their error-type profiles. RFs tend to be way more restrained in predicting the positive class, as opposed to MLPs, while ENs (just as we saw when analyzing attributes) show an intermediate degree in that spectrum where the others are opposites. On the other hand, combining the three algorithms with an averaging-probability ensemble yields a performance profile considerably better than the mean of the performance of its constituents. As will be presented in subsection 5.2.3, the trained models also have complementary decision-making criteria, which further motivates and explains the success of the ensemble.

Table 5.2 – Final performance estimates mean and standard deviation

Algorithm	*F_2-Score*	*AUCROC*	*Sensitivity*	*Specificity*
Ensemble	0.690 ± 0.029	0.811 ± 0.018	0.780 ± 0.049	0.666 ± 0.047
Multilayer Perceptron	0.686 ± 0.042	0.764 ± 0.081	0.807 ± 0.093	0.591 ± 0.173
Elastic Net	0.659 ± 0.066	0.773 ± 0.042	0.747 ± 0.112	0.659 ± 0.095
Random Forest	0.608 ± 0.041	0.814 ± 0.019	0.630 ± 0.048	0.792 ± 0.026

Figure 5.1 – Comparison of measured F_2-Score for different algorithms

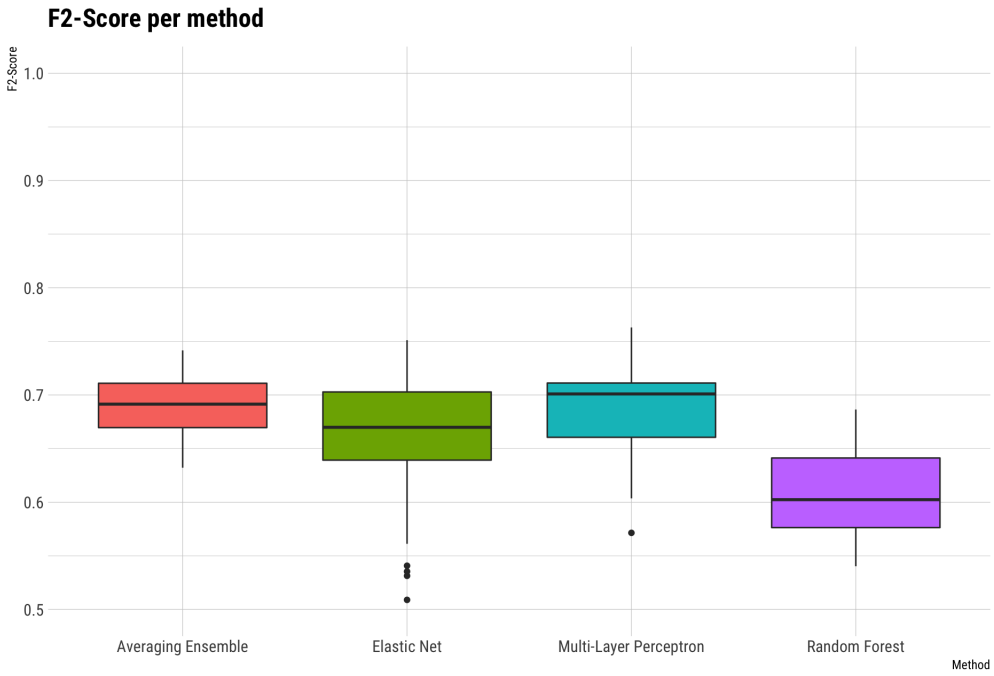


Figure 5.2 – Comparison of measured AUCROC for different algorithms

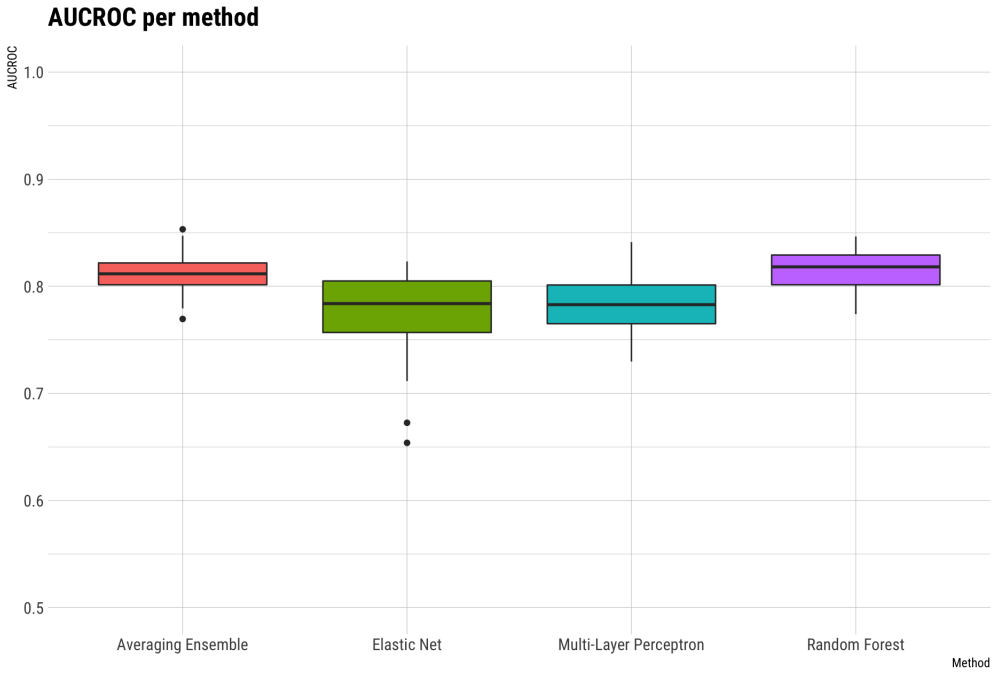


Figure 5.3 – Comparison of measured sensibility for different algorithms

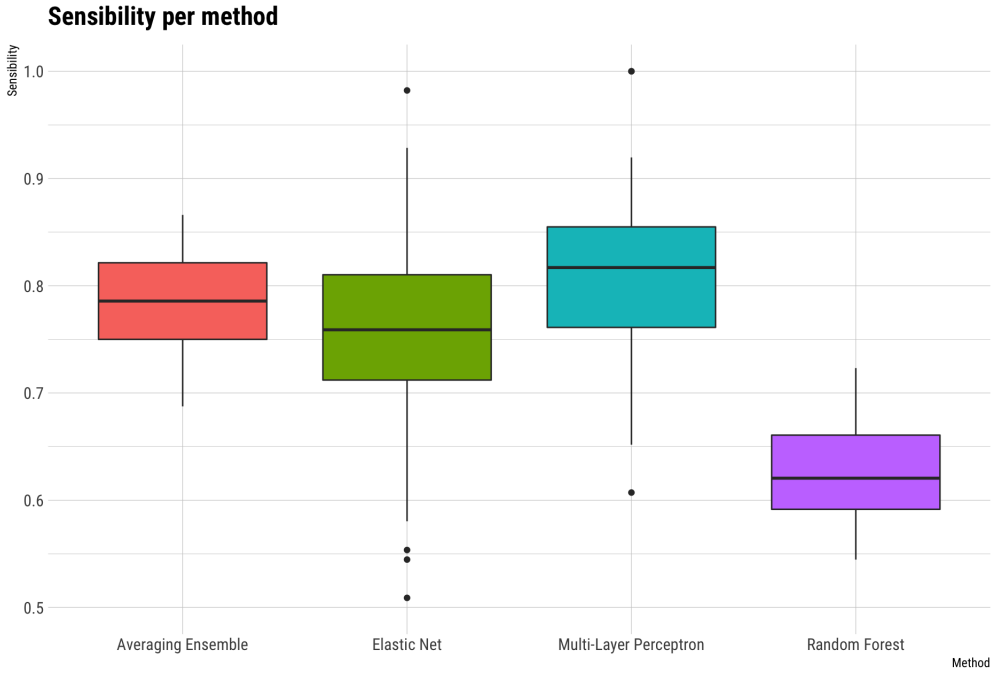
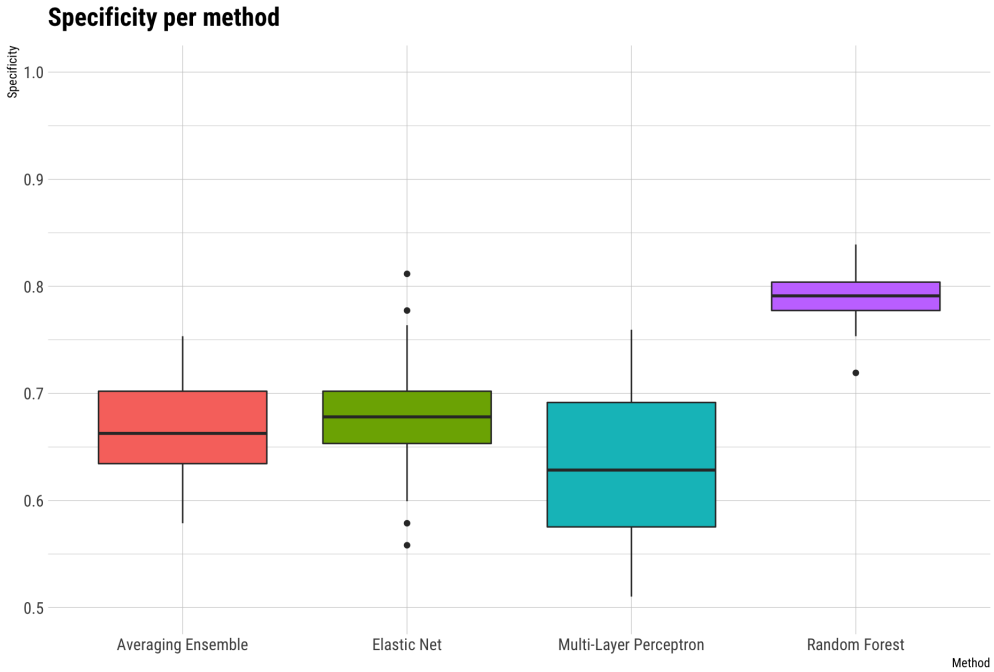


Figure 5.4 – Comparison of measured specificity for different algorithms



5.2.2 Hyperparameters Analysis

Table 5.3 – Hyperparameters most-frequently chosen in tuning

Parameter	*Value*
Elastic Net - Alphas	0.1
Elastic Net - Lambda	0.09
Neural Network - Layer 1 #Units	1
Neural Network - Layer 2 #Units	1
Random Forest - Node-splitting method	*extratrees*
Random Forest - #Random-attributes	2

For the analysis of the produced models internal characteristics, it is proper to first verify the results of the hyperparameters tuning process, as this plays an important role in models' performance. The most frequently chosen hyperparameters values of the final models are summarized in Table 5.3 and indicate that, in general, relatively simpler models tend to perform better in the training phase. This is not surprising, since the best sizes of variables subsets are low and in the hyperparameters tuning phase the available data has a relatively small number of instances, thus it is probably the case the more complex models could overfit and perform poorly.

Figures 5.5, 5.6, and 5.7 show plots of the F_2-Score during the training of the base-learners' model of a given resample, for multiple combinations of parameters of RFs, ANNs, and ENs (respectively). We see that, at least for this fold (and we keep the graphical analysis limited to this single one for the sake of simplicity), performance estimates seem to vary in an orderly manner throughout the hyperparameter values planes. More specifically, clear patterns are identifiable in every hyperparameters plot.

First, in the RF graph of Figure 5.5, where the *extratrees* (extremely randomized trees) and the *gini* tree-node-splitting methods are compared, we find that there is a big in the F_2-Score between the two, but for both we observe relatively flat lines for higher values of randomly selected predictors (*NRand*). For the lower values (i.e. from the range between 2 and 20), with the increasing number of predictors, F_2-Score increases in the *gini* line, but decreases for the extremely randomized trees splitting method. Most notably, the peak score is found in the left-most point of the *extratrees* line.

Figure 5.5 – Hyperparameter tuning for Random Forest in first CV resample

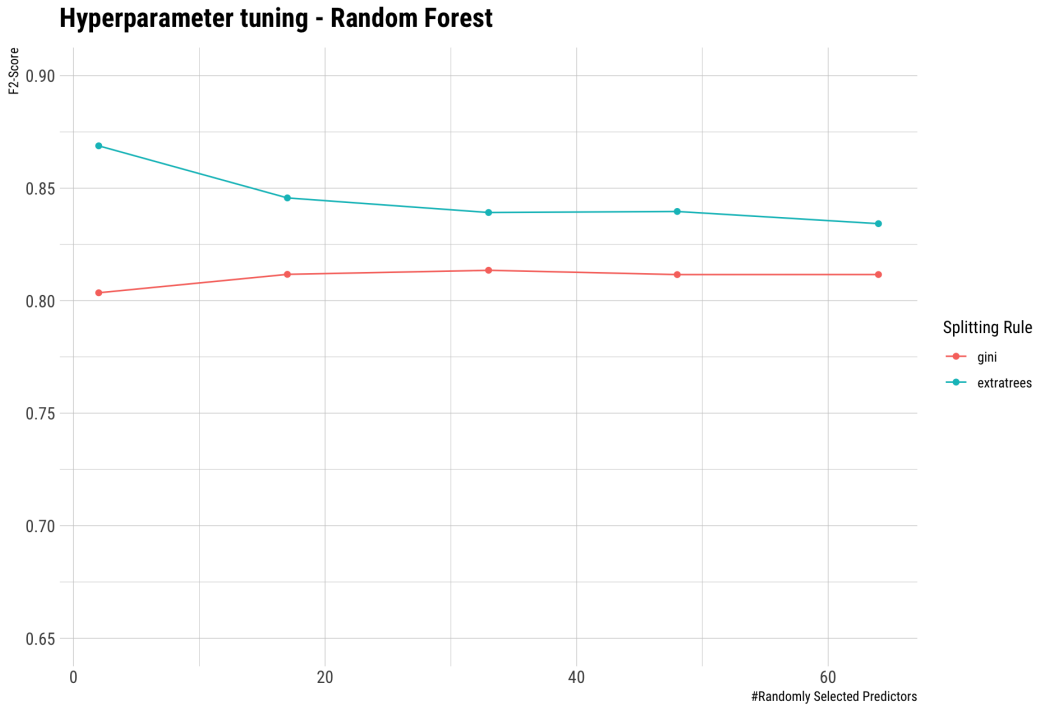


Figure 5.6 – Hyperparameter tuning for Multilayer Perceptron in first CV resample

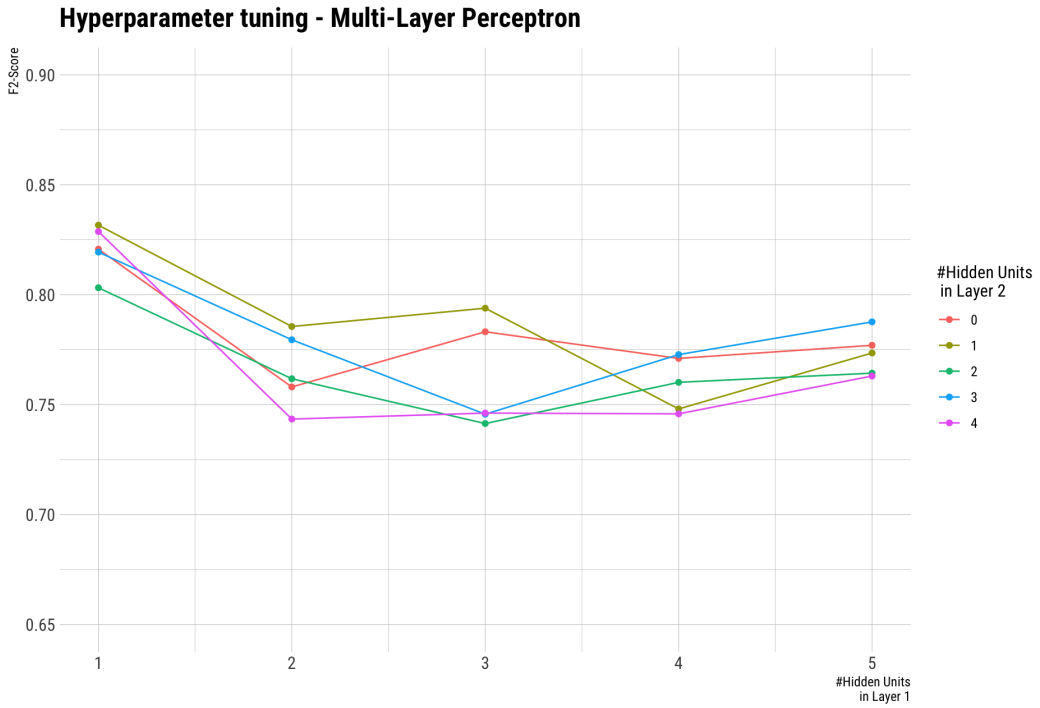
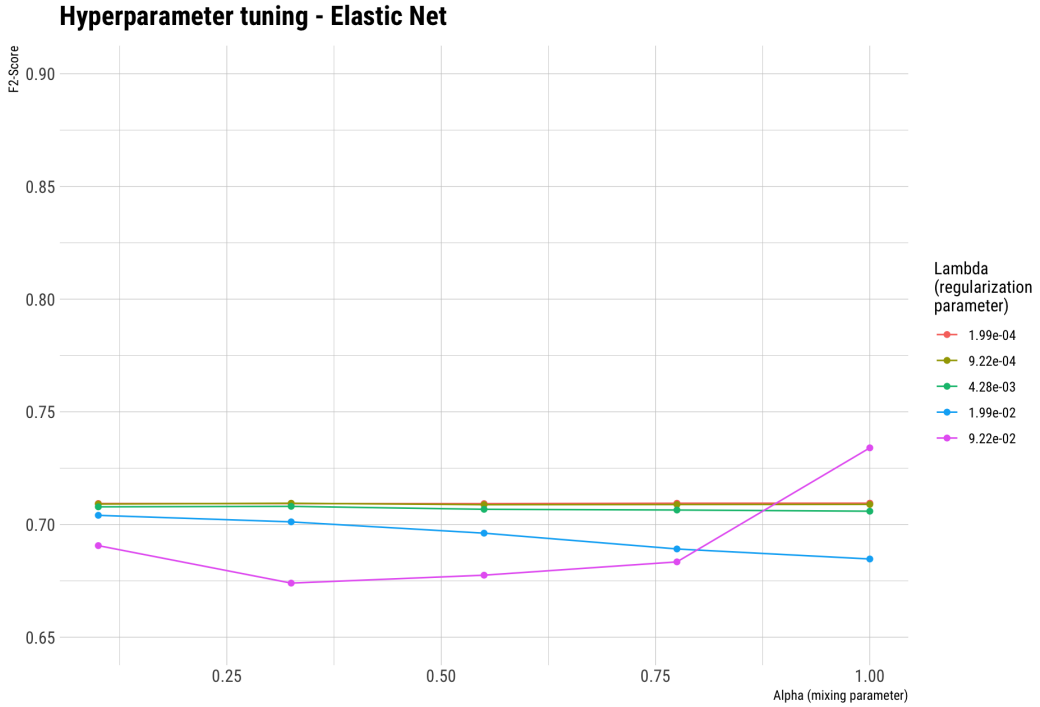


Figure 5.7 – Hyperparameter tuning for Elastic Net in first CV resample



As for Figure 5.6, every colored line represents a neural network with a given number of neurons in the second (and last) hidden layer. The commonalities between

these slightly different architectures is that performance is lower for higher numbers of hidden units of layer 1 (although it varies in a bowl-shaped curve). Increasing the number of neurons in layer 2 has a similar effect. In other words, simpler ANNs, in our scenario, seem to carry out a better classification.

Lastly, the plot of EN's hyperparameters of Figure 5.7 indicates a similar behavior to the one seen for the random forest, in the sense that each line generally presents just small variations of F_2-Score with respect to the variable from the horizontal axis (the lasso-ridge mixing parameter alpha in this case). Also in the same manner as the RF's grid search, the maximum performance of this particular elastic net was achieved in an "extreme" value, the highest *lambda* and highest *alpha* pair (as it is with the lowest RF's *NRand*). This is curious to notice, as an *alpha* value of 1 indicates a purely lasso regression.

5.2.3 Feature Selection Analysis

Besides hyperparameters tuning, another characteristic worth examining is the models' predictive capabilities across different predictor sets of the RFE loops. For each final RFE model produced, Figures 5.8, 5.9, and 5.10 show the F_2-Score variation depending on the number of variables considered in the base learners induction. To analyse the algorithms' performance according to distinct features set sizes, it is useful to start the analysis of the results by the right side of the graph, which represents the original features set size. This last mark in the horizontal axis indicates the data has as many variables as the cleansed dataset, which are then reduced in quantity down to about 600 by the base learners preprocessing. As the RFE retrains and reassesses the variables importances at each iteration, the difference between the number of predictors before and after preprocessing should drastically diminish. It is curious that both elastic nets (Figure 5.8) and ANNs (5.9) tend to perform better with fewer variables, showing a smoothly varying F_2-Score curve. On the other hand, random forests present a performance peak at the predictor set size of 64 (Figure 5.10), which indicates that after further removal of variables, too much valuable information is taken out from the model training process.

Figure 5.8 – Recursive feature elimination performance for Elastic Net

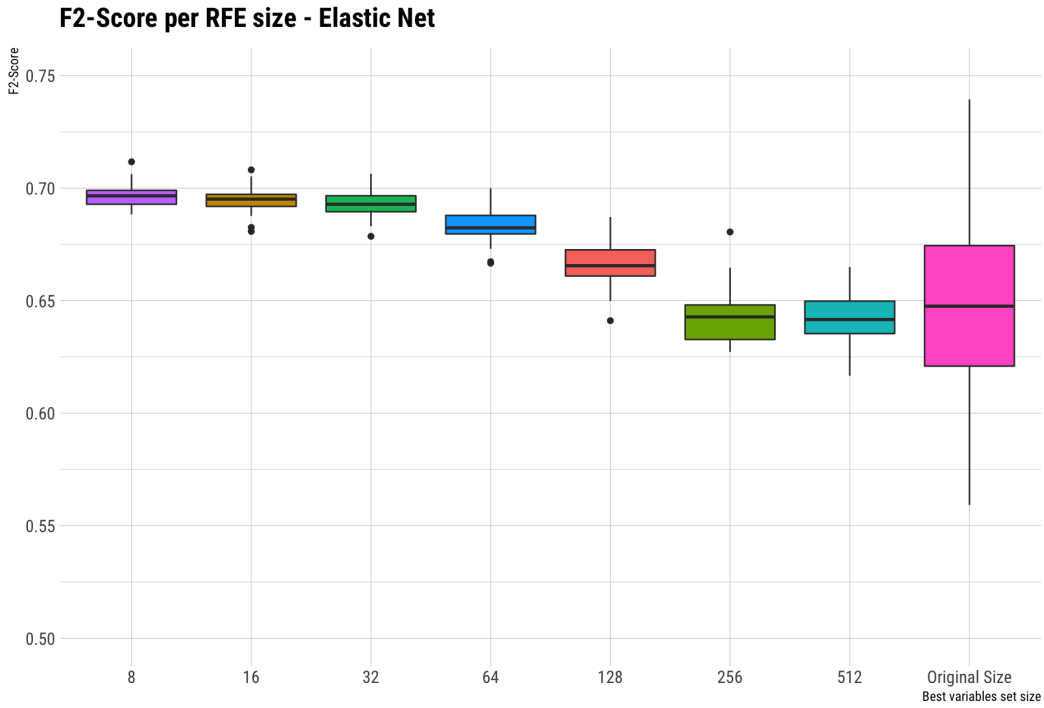


Figure 5.9 – Recursive feature elimination performance for Multilayer Perceptron

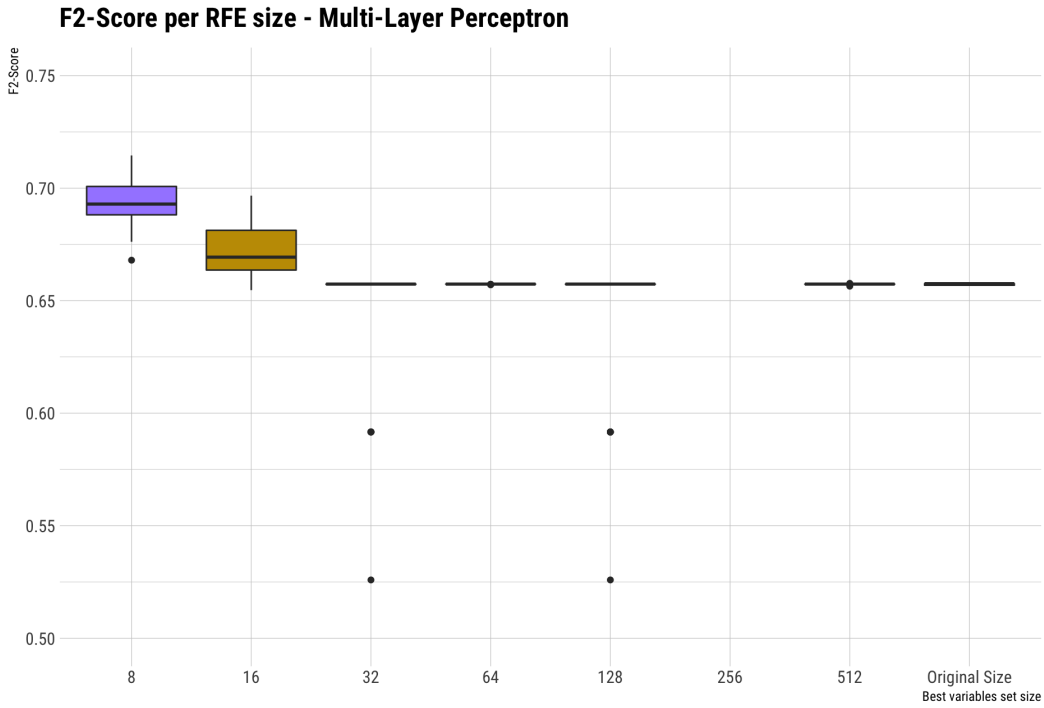
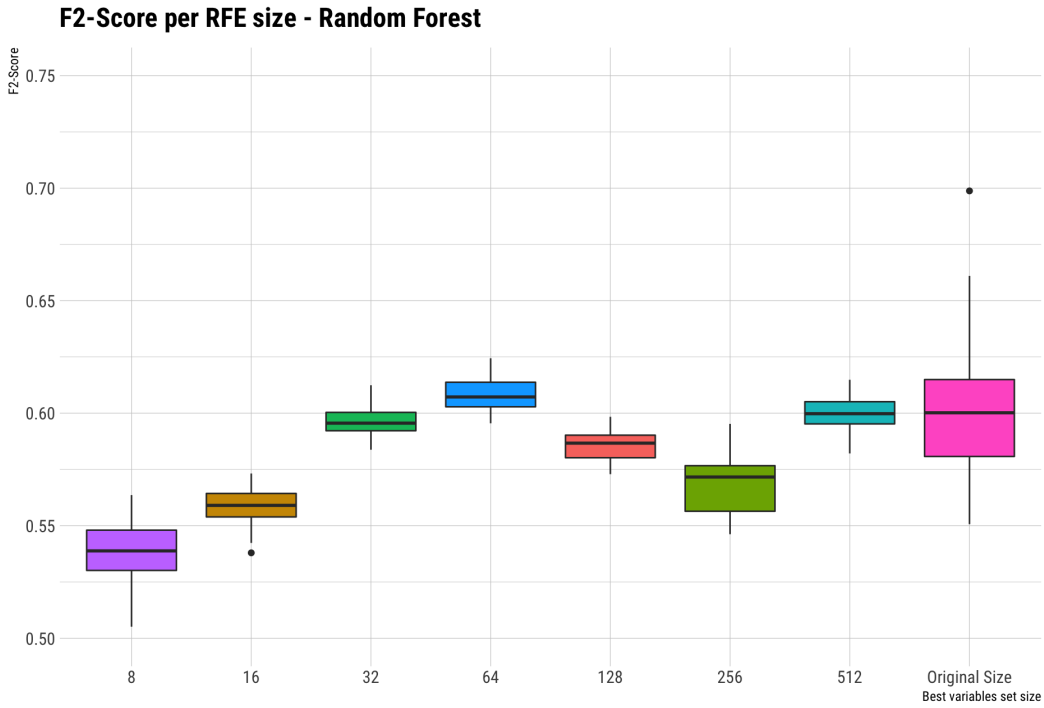


Figure 5.10 – Recursive feature elimination performance for Random Forest



To analyse the rankings of features importances for each algorithm, we produced "rank matrices" that can be plotted as heatmaps shown in Figures 5.11, 5.12, and 5.13. The idea of this visualization approach is the following: each predictor (in the vertical axis, ordered alphabetically for a naive semantical clustering) has its importance mapped to each of the trained models (with a 1:1 correspondence to a CV resample) by a ranking function F_{rank}. The complete list of meanings of the top-30 most relevant attributes found using each algorithm is available in Table 5.4. The value of each attribute-to-model cell is zero when the variable does not appear in the list of most important predictors of the model. Otherwise, it is defined as follows (where A_m is the set of best attributes of a model m, and $I(a, m)$ is the index of the attribute a in A_m):

$$F_{rank}(a,m) = \frac{\#A_m - I(a,m)}{\#A_m} \tag{5.1}$$

Our attribute-rank matrices then have higher values (indicated by brighter colors in the heatmaps) for the variables that are most important, but the magnitudes are relatively decreased given a fixed rank index for models that have larger best-variables sets. Moreover, the values of these matrices range from 0 to 1 after normalization according to the variables sets sizes. Thus, the intuitions to be derived from the heatmap plots of these matrices are that horizontal lines show how the feature relevances vary across the

collection of trained models, and clusters of horizontal lines with uniform colors indicate the importance of semantically similar attributes.

For each algorithm, we can average the F_{rank}s of each predictor across the multiple resamples to obtain generalized variable importances. More precisely, we calculate what we define as $G_{imp}(a)$ (the general importance of an attribute a), based on a set of models M, as:

$$G_{imp}(a) = \sum_{n=1}^{\#M} \frac{F_{rank}(a,m)}{\#M} \tag{5.2}$$

In Figures 5.11, 5.12, 5.13, we displayed only the top-30 G_{imp} features. We chose 30 because it matches the number of resamples in this plot, yielding a more pleasant plot aesthetic, but most importantly because it is a sufficient and efficient size for a set of best variables to be considered in our analysis, in the sense that we do not need to discuss many predictors but the ones we discuss are relevant and insightful (especially given the best attribute sizes indicated in Figures 5.8, 5.9 and 5.10).

The exception of this rule is Figure 5.12, where only 17 features are shown. We note that, for this algorithm, there are also relatively clearer horizontal and vertical patterns in the heatmap, and also low variance in its RFE plot (Figure 5.9). This indicates that these models are more concise, and also homogeneous between each other.

Figure 5.11 – Heatmap of attribute relevance for Elastic Net

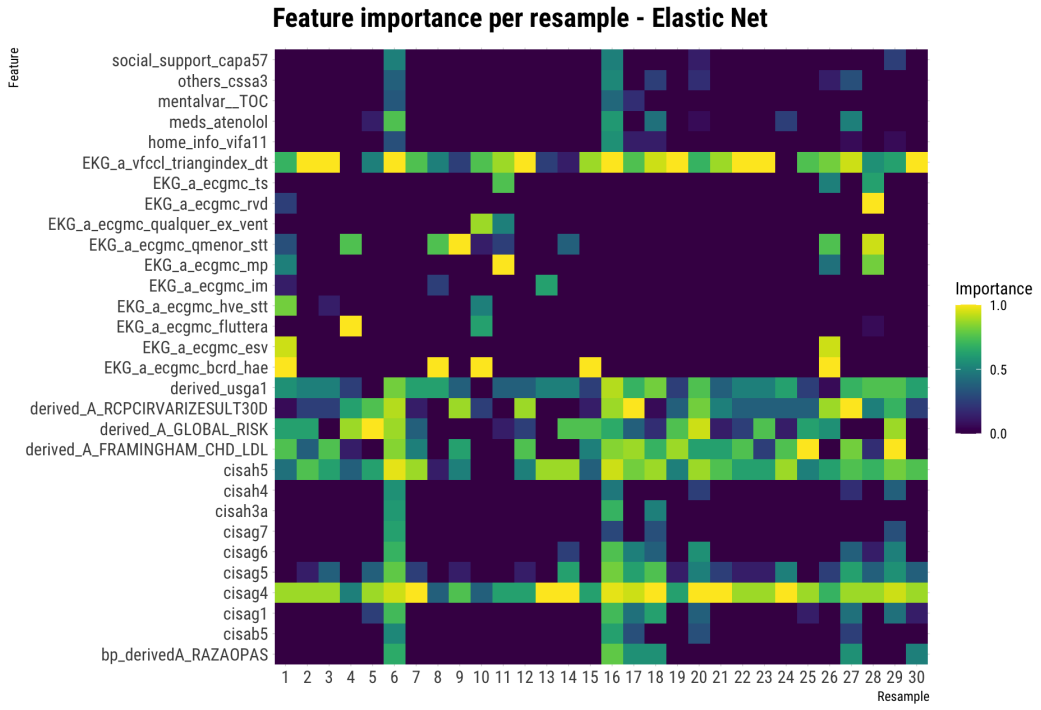


Table 5.4 – Variables ranked as most important for trained models in alphabetical order

Attribute	*Description*
bp_derivedA_RAZAOPAS	Ankle brachial index
cisab5	Tired for more than 3h in last 7d (B5)
cisae7	Argument fight in last 7d (E7)
cisag1	Sad/depressed in last 30d (G1)
cisag2	Enjoyed things as used to in last 30d (G2)
cisag4	Sad/depressed in last 7d (G4)
cisag5	Enjoyed things as used to in last 7d (G5)
cisag6	#days sad/incapable of enjoyment last 7d (G6)
cisag7	Sad/incapable of enjoyment 3h+ in last 7d (G7)
cisah1	Worst time of day for sadness last 7d (H1)
cisah2	Sexual desire in last 30d (H2)
cisah3a	When sad, got restless in last 7d (H3a)
cisah3b	When sad, did things more slowly in last 7d (H3b)
cisah3c	When sad got less talkative in last 7d (H3c)
cisah4	Guilty or blamed themself unnecessarily last 7d (H4)
cisah5	Felt not so good as other people in last 7d (H5)
cisai8	How unpleasant was the worrying in last 7d (I8)
cisaj8	How unpleasant was the general anxiety last 7d (J8)
cisan1	Unpleasant thoughts kept appearing last 30d (N1)
cisao1	Feelings precluded things last 7d (O1)
cisao1a	Feelings precluded more than once last 7d (O1a)
derived_A_FRAMINGHAM_CHD_LDL	Low LDL (Framingham criteria)
derived_A_GLOBAL_RISK	Global risk score (Framingham)
derived_A_RCPCIRVARIZESULT30D	Self-reported varicose vein surgery in last 30d
derived_A_VIFA30_PMCAT	Familial income
derived_cogA_FLUENCIA_LETRAF	Fluency score with words with the letter F
derived_nle	Any negative life event
derived_usga1	Carotid artery intima-media thickness
EKG_a_ecgmc_bcrd_hae	Complete right bundle block + left anterior hemiblock
EKG_a_ecgmc_esv	ventricular extrasystole
EKG_a_ecgmc_fluttera	Atrial Flutter
EKG_a_ecgmc_hve_stt	Left ventricular hypertrophy with ST-T alterations
EKG_a_ecgmc_im	Myocardial Infarction
EKG_a_ecgmc_mp	Artificial pacemaker
EKG_a_ecgmc_qmenor_stt	Lower Q with ST-T alterations
EKG_a_ecgmc_qualquer_ex_vent	Any extrasystole
EKG_a_ecgmc_rvd	High R waves right ventricle
EKG_a_ecgmc_ts	Sinus tachycardia
EKG_a_vfccl_triangindex_dt	Triangular Index VFCDT
home_info_vifa11	Children (Q11)
meds_atenolol	Usage of Atenolol
mentalvar__TAG	GAD - Generalized Anxiety Disorder
mentalvar__TOC	OCD - Obsessive Compulsive Disorder
mentalvar_A_SINTDEP	Depression symptoms
mentalvar_A_SINTMEMORIA	Concentration symptoms
metadata_ecga2	V4 value
negativelifeevents_evea12	Severe financial difficulties in last 12m (Q12)
neighborhood_viza06	Physical activities conditions in neighbourhood (Q6)
others_cssa3	Usage of aspirin or others for anti-coagulation
others_esca03	Work-related self-evaluation (Q3)
social_support_capa57	Someone takes care of children when away (Q57)

Figure 5.12 – Heatmap of attribute relevance for Multilayer Perceptron

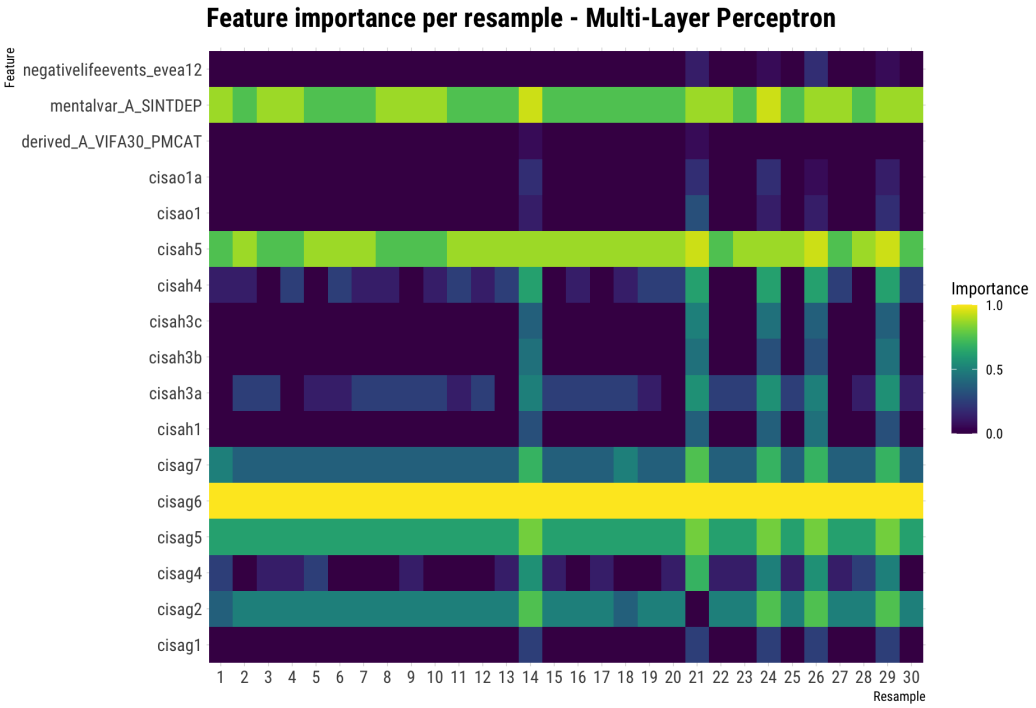


Figure 5.13 – Heatmap of attribute relevance for Random Forest

Another useful employment of our G_{imp} metric is to arrange the variables according to it, ordered from highest to lowest, so that we can understand at a glance what each algorithm tends to consider most when predicting the classes of instances. This is shown in Figures 5.14, 5.15, and 5.16.

In diametrical opposition to the MLPs' simplicity, RFs learned that many more predictors are relevant for correct classification. This is indicated by the already described performance peak at 64 RFE variables, by the F_{rank} heatmap (Figure 5.13) being overall brighter than the MLP's and EN's heatmaps, and by the top 30 attributes having small G_{imp} differences between each other (Figure 5.16).

On the other hand, based on the features heatmap and the rank plot of Figures 5.11 and 5.14, we can place the elastic nets in an "intermediate" spot between the MLPs and the RFs regarding simplicity and homogeneity. Figure 5.11 indicates patterns of some features being considered relevant across different resamples, but the patterns are not so clear and uniform as the MLP's, and the variables are not so numerous or few as the RF's or MLP's, respectively.

Figure 5.14 – Rank attribute relevance for Elastic Net

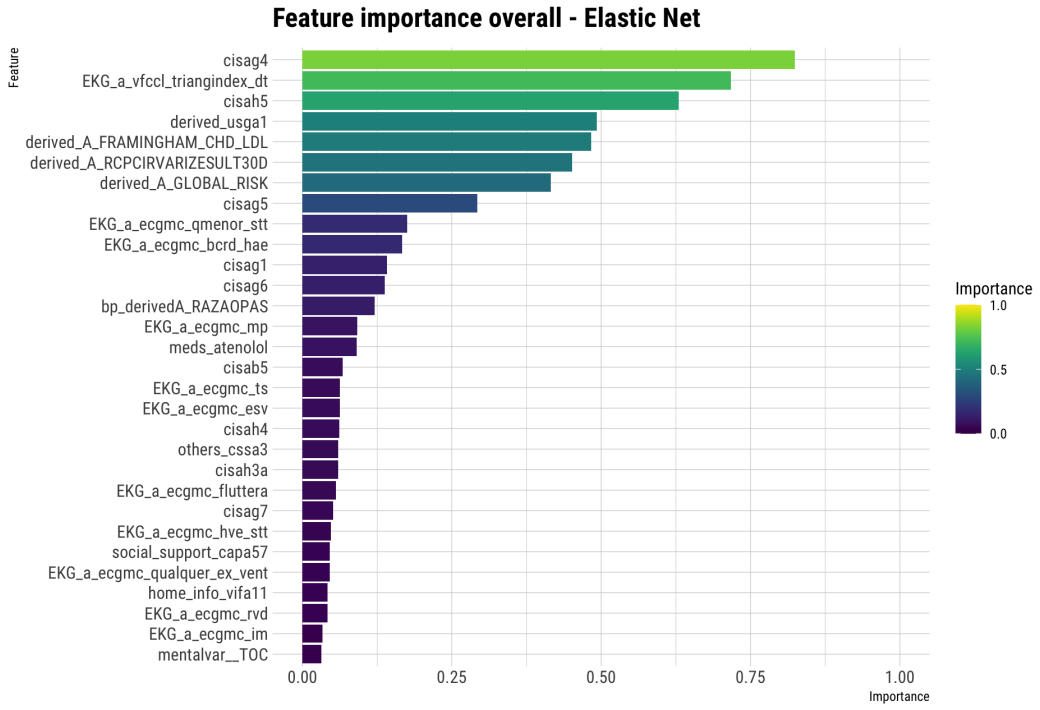


Figure 5.15 – Rank of attribute relevance for Multilayer Perceptron

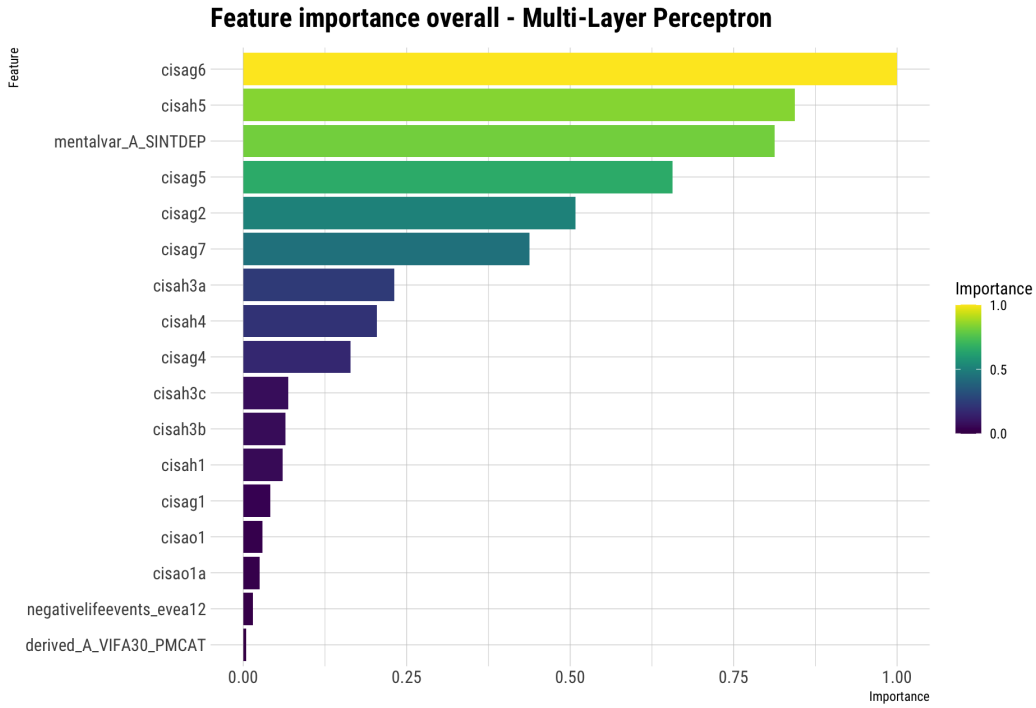
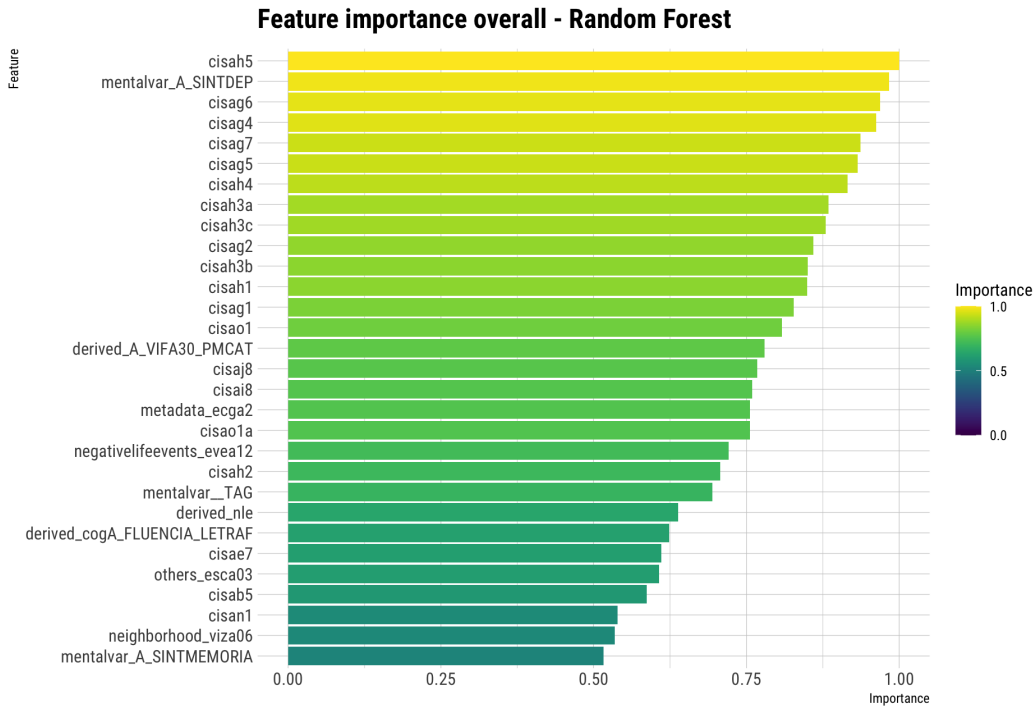


Figure 5.16 – Rank of attribute relevance for Random Forest



The combination of the three F_{rank} matrices (which, despite what is shown in Figures 5.11, 5.12, 5.13, and 5.17, includes *all* predictors in the dataset) by averaging the values of the cells in matching positions gives us a form of condensed and summarized knowledge on the criteria used by our classifiers and roughly estimates the attribute importances of our class-probability-averaging ensemble. We then calculate each predictor's G_{imp} to sort them by this quantity and again plot the top 30 in a heatmap and in a column chart as was done with the single classifiers.

Figure 5.17 – Heatmap of attribute relevance for Averaging Ensemble

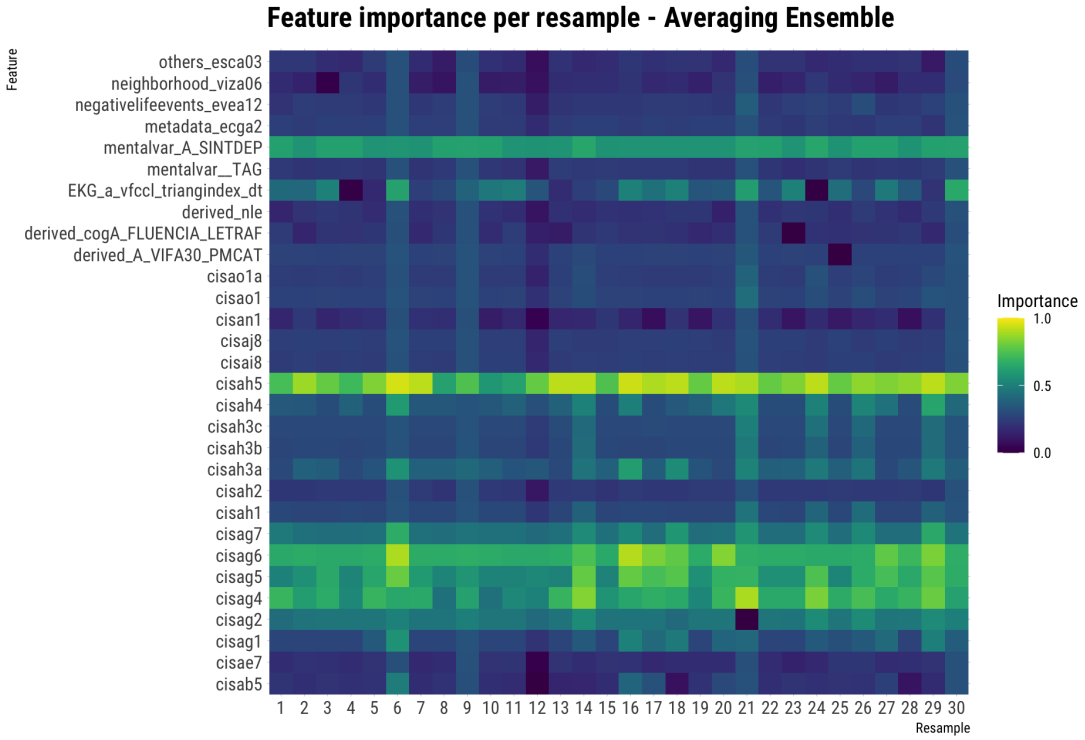
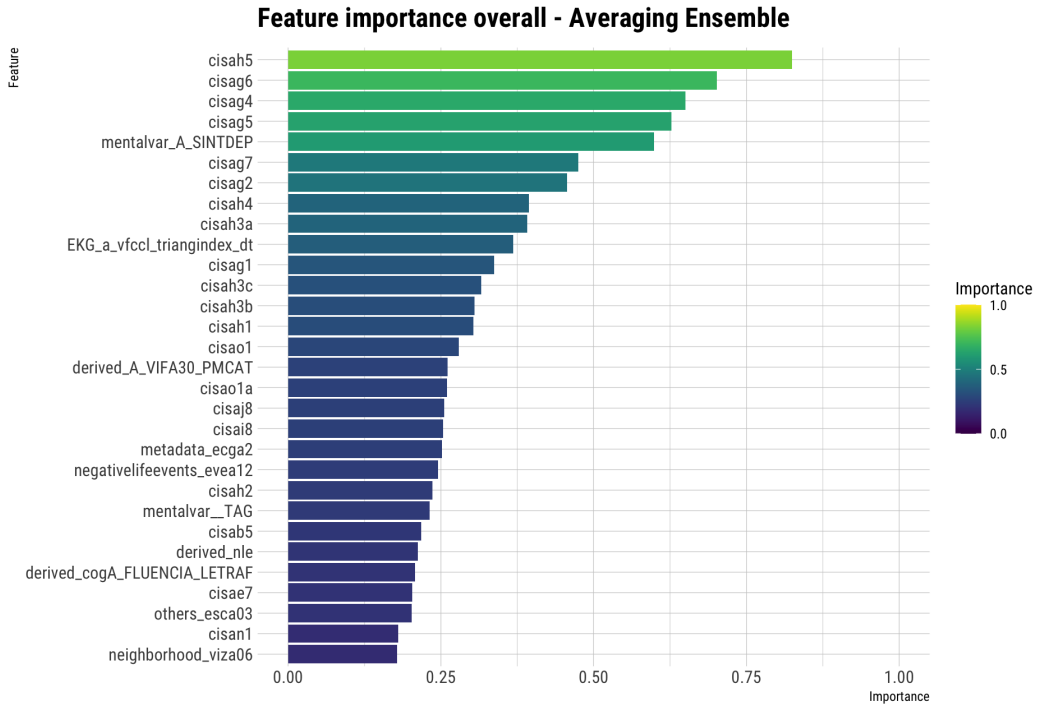


Figure 5.18 – Rank of attribute relevance for Averaging Ensemble



Regarding the chosen variables per se, Figure 5.18 gives us a clear top 5 indicated by the brightest colors and a rather abrupt G_{imp} jump from the fifth to sixth position, ordered by importance: *cisah5*, *cisag6*, *cisag4*, *cisag5*, and *mentalvar_A_SINTDEP*. All of these are responses to the CIS-R questionnaire, from section G (depression) and H (depressive ideas), safe for *mentalvar_A_SINTDEP*, but it relates to the same general theme. It is worth going over the exact definition of each of the respective questions here, for the analysis and understanding of the models. Thus, the top-5 variables are, ordered by relevance:

- CIS-H5: *During the past week, have you been feeling you are not as good as other people?*
- CIS-G6: *Since last (DAY OF WEEK) on how many days have you felt sad, miserable or depressed/unable to enjoy or take an interest in things?*
- CIS-G4: *In the past week have you had a spell of feeling sad, miserable, or depressed? Use informant's own words if possible*
- CIS-G5: *In the past week have you been able to enjoy or take an interest in things as much as usual? Use informant's own words if possible*

Question H5 stands-out from the others, for its meaning and its calculated G_{imp} - this indicates our models have an overall consensus that feeling inferior to other people plays a decisive role in increasing the chance of presenting suicidality. Both G4 and G6 reveal whether the interviewee had been feeling acute sadness, and the G5 and G6 pair provide information on people's loss of interest in things, such that they are not as enjoyable as before.

In short, our models consider as most relevant to their predictions degrees of:

1. **feelings of inferiority** (*cisah5*);
2. **sadness** (*cisag4, cisag6, cisag7*);
3. **disappearance of interests** (*cisag2, cisag7*);
4. **unnecessary guilt** (*cisah4*);
5. **energy (disposition)** (*cisah3a, cisah3b, cisah3c, cisab5*);
6. **preclusion** of activities including chores and leisure for bad feelings (*cisao1, cisao1a*);
7. **income** (*derived_A_VIFA30_PMCAT, negativelifeevents_evea12*);
8. **anxiety** (*cisaj8, mentalvar_A_TAG*);
9. **worrying** (*cisai8*);
10. **libido** (*cisah2*);
11. **irritability** (*cisae7*);
12. **obsessions** - or at least recurring bad thoughts (*cisan1*);
13. **physical activities** (*neighborhood_viza06)*.

Some outlier variables are found, which cannot be associated with another factor without being too speculative if no further studies on this particular variable are conducted, such as *derived_cogA_FLUENCIA_LETRAF*. Some predictors relate to physical characteristics and conditions (e.g. *EKG_a_vfccl_triangindex*) in the overall score, but they are found more prominently in the elastic nets' G_{imp} rank (Figure 5.14), like *derived_A_GLOBAL_RISK* that relate to physical activities, overweight, obesity and metabolic syndrome, and smoking (COKE, 2010).

6 CONCLUSION

In this chapter, we first reflect on our original goals and the contributions of the current work for the related literature. Next, we discuss the limitations of the proposed solutions and enhancement opportunities that could be explored in future works.

6.1 Contributions and Impact

Table 6.1 – Performance estimates of related works compared to ours, ordered by F_2-Score

Paper	*Algorithm*	*F_2-Score*	*AUCROC*	*Sensitivity*	*Specificity*
A	XGB	**0.84**	0.86	0.79	**0.79**
B	ANNs + RF	0.71	**0.88**	0.80	0.79
Ours	EN + ANN + RF	0.69	0.81	0.78	0.67
C	ANN	0.48	0.88	**0.81**	0.77
D	EN	0.45	0.79	0.67	0.78

Source: The Author
A: (JUNG et al., 2019);
B: (ROY et al., 2020);
C: (OH et al., 2020);
D: (LIBRENZA-GARCIA et al., 2020)

.

This study proposed and thoroughly tested a solution for identifying patterns of suicidality in a population of Brazilian adults based on data available in the first wave of the ELSA-Brasil study. Suicidality is defined as the presence of hopelessness, feelings that life is not worth living, and suicidal thoughts. We made use of special ML techniques chosen based on the challenges and goals of our mission, such that our approach to the problem was focused on minimizing false negative errors (as they are extremely detrimental in our domain) by mitigating the effect of the fewer number of examples labeled as the positive class and yielding high interpretability and insightful knowledge for physicians and clinicians from the produced classifiers. Thus, although our methodology could be applied to other datasets, our findings in this study elucidate patterns of suicidality in the specific population subset of Brazilian public servants with common mental disorders and do not necessarily directly apply to a more general population. Nevertheless, we believe the attainment of our goals carries high practical and academic value for dealing with the grave worldwide problem of suicide.

Although direct comparisons to the studies mentioned in Chapter 3 cannot yield

fair rankings of performance estimates, given all of them use different datasets (except for Librenza-Garcia et al. (2020)), it is important to contextualize the findings of this work among the related literature. That said, in relation to other works with similar objectives but with varying domain particularities, the estimated performance of our models was decently comparable, specially for the F_2-Score - our optimization metric. As shown in Table 6.1, we achieved very good results in terms of F_2-Score with our weighted-average ensemble model. With a mean value of 0.69, our estimated F_2-Score is close to the second-best work of the table (ROY et al., 2020). This promising performance indicator is considerably higher than the ones estimated in Oh et al. (2020) and (LIBRENZA-GARCIA et al., 2020), with similar data and objectives.

We note that most works do not report the F_2-Score, and focus on AUCROC, sensitivity, and specificity instead. The mean area under the ROC curve of our ensemble is, as the other work in the table using the ELSA-Brasil dataset (LIBRENZA-GARCIA et al., 2020), close to 0.8, while other works are closer to the 0.9 mark. Nonetheless, our results emphasize the importance of using metrics alternative to AUCROC as it does not take precision into account and, thus, can show very satisfactory results even when the models present low true positive rates. Even though in this work sensitivity and specificity did not achieve results as high as the best ones reported in the related literature, our models presented acceptable scores for these metrics.

Also, our analysis provides important insights about the most relevant features associated with suicidality, which after further investigation by specialists may contribute to new strategies to prevent suicidal ideation and attempts. Finally, we highlight that during the development of this work we were very careful regarding the methodological and reporting rigor of our research, especially concerning strategies to minimize data leakage and selection biases during model development, which although may result in lower performance provide us with higher confidence regarding the generalization power of our models.

6.2 Possible Improvements

The first limitation of this work is that it was mainly focused on the computational aspects of the discussed matters, thus it did not examine with all the possible depth the relations between and the explanations of the attributes found to be of most importance to the models. Also, although we went through what are the most relevant attributes to

predict suicidality, since our models are non-linear (safe for the elastic net), we are not able to discuss how variations in each attribute affect the output with the provided analysis material. Therefore, subsequent studies could analyse how the classification probability varies for each of them, or at least to the most important ones, using for instance Partial Dependence Plots. With that, stakeholders would have more accurate quantitative insights on the factors related to the class of interest. Similarly, the understanding of the models could be enriched by an analysis of the commonalities and patterns in the data that cause most classification errors (mainly the false negatives).

Considering the induction of the models to solve our problem, we could also explore different techniques. Cost-sensitive learning could be implemented and integrated into the procedure, possibly introducing more variation and richness to the constitution of the ensemble, but also a great model by itself. Other algorithm-level methods to mitigate the class-imbalance problem could be employed, using learners that naturally deal with this such as AdaBoost or eXtreme Gradient Boosting (XGB), or calibrating the probabilities predicted by our model (NICULESCU-MIZIL; CARUANA, 2005). These changes would not only provide richness and novelty to the body of literature of the suicidality prediction domain, but also preserve classification robustness in face of a smaller number of positive class instances, allowing future studies to also tackle other ELSA-Brasil dataset variations that are more unbalanced. For example, studies could consider the whole population instead of only the people presenting CMD, have as outcome label just the suicide ideation variable without combining it with others, or attempt to predict the incidence of suicidality and other labels on ELSA-Brasil's wave 2 based on input data of wave 1. In terms of data analysed, models could also be built by integrating ELSA-Brasil baseline features georeferenced data, as characteristics related to the geographic location of phenomena may play a role in its occurrence.

Finally, as a more practical application, one could develop user-facing applications to assist clinicians and/or patients and people with common mental disorders. The programs could use the knowledge uncovered by this study as a basis for a score associated with or some form of journaling done by the user, such that therapists, psychiatrists, and doctors in general can assess their patients' mental health with more tools and have automated support in their decisions.

REFERENCES

AMBROISE, C.; MCLACHLAN, G. J. Selection bias in gene extraction on the basis of microarray gene-expression data. **Proceedings of the National Academy of Sciences of the United States of America**, v. 99, n. 10, p. 6562–6566, 2002. ISSN 00278424.

AQUINO, E. M. et al. Brazilian Longitudinal Study of Adult health (ELSA-Brasil): Objectives and design. **American Journal of Epidemiology**, v. 175, n. 4, p. 315–324, 2012. ISSN 00029262.

BARROS, J. et al. Suicide detection in Chile: Proposing a predictive model for suicide risk in a clinical sample of patients with mood disorders. **Revista Brasileira de Psiquiatria**, v. 39, n. 1, p. 1–11, 2017. ISSN 15164446.

BEBBINGTON, P. E. et al. Suicidal ideation, self-harm and attempted suicide: Results from the British psychiatric morbidity survey 2000. **European Psychiatry**, v. 25, n. 7, p. 427–431, 2010. ISSN 09249338.

BERGMEIR, C.; BENÍTEZ, J. M. Neural networks in R using the Stuttgart neural network simulator: RSNNS. **Journal of Statistical Software**, v. 46, n. 7, 2012. ISSN 15487660.

BOULESTEIX, A. L. et al. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 2, n. 6, p. 493–507, 2012. ISSN 19424787.

BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5–32, 2001. ISSN 08856125.

BRUNONI, A. R. et al. Socio-demographic and psychiatric risk factors in incident and persistent depression: An analysis in the occupational cohort of ELSA-Brasil. **Journal of Affective Disorders**, Elsevier B.V., v. 263, p. 252–257, feb 2020. ISSN 15732517.

BURKE, T. A.; AMMERMAN, B. A.; JACOBUCCI, R. The use of machine learning in the study of suicidal and non-suicidal self-injurious thoughts and behaviors: A systematic review. **Journal of Affective Disorders**, v. 245, p. 869–884, feb 2019. ISSN 01650327. Available from Internet: <https://linkinghub.elsevier.com/retrieve/pii/S0165032718317506>.

BURKE, T. A. et al. Identifying the relative importance of non-suicidal self-injury features in classifying suicidal ideation, plans, and behavior using exploratory data mining. **Psychiatry Research**, v. 262, p. 175–183, 2018. ISSN 18727123.

C. J. van Rijsbergen. Information Retrieval . C. J. van Rijsbergen. **The Library Quarterly**, v. 47, n. 2, p. 198–199, 1977. ISSN 0024-2519.

CARBALLO, J. J. et al. **Psychosocial risk factors for suicidality in children and adolescents**. 2020. 759–776 p.

CHAWLA, N. V. Data Mining for Imbalanced Datasets: An Overview. In: **Data Mining and Knowledge Discovery Handbook**. [S.l.]: Springer US, 2009. p. 875–886.

CHAWLA, N. V. et al. SMOTE: Synthetic minority over-sampling technique. **Journal of Artificial Intelligence Research**, v. 16, n. 1, p. 321–357, 2002. ISSN 10769757.

CLEMEN, R. T. Combining forecasts: A review and annotated bibliography. **International Journal of Forecasting**, v. 5, n. 4, p. 559–583, 1989. ISSN 01692070.

COKE, L. A. Cardiac risk assessment of the older cardiovascular patient: the Framingham Global Risk Assessment Tools. **Medsurg nursing : official journal of the Academy of Medical-Surgical Nurses**, v. 19, n. 4, p. 253–254, 2010. ISSN 1092-0811.

DARCY, A. M.; LOUIE, A. K.; ROBERTS, L. W. **Machine learning and the profession of medicine**. [S.l.]: American Medical Association, 2016. 551–552 p.

DOMINGOS, P. MetaCost. In: **Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining**. [s.n.], 1999. v. 55, p. 155–164. ISBN 1581131437. ISSN 1550-4786. Available from Internet: <http://portal.acm.org/citation.cfm?id=312129.312220{&}type=ser>.

FAWCETT, T. An introduction to ROC analysis. **Pattern Recognition Letters**, v. 27, n. 8, p. 861–874, 2006. ISSN 01678655.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. Regularization paths for generalized linear models via coordinate descent. **Journal of Statistical Software**, v. 33, n. 1, p. 1–22, 2010. ISSN 15487660.

GEURTS, P.; ERNST, D.; WEHENKEL, L. Extremely randomized trees. **Machine Learning**, v. 63, n. 1, p. 3–42, 2006. ISSN 08856125.

GRADUS, J. L. et al. Gender Differences in Machine Learning Models of Trauma and Suicidal Ideation in Veterans of the Iraq and Afghanistan Wars. **Journal of Traumatic Stress**, v. 30, n. 4, p. 362–371, 2017. ISSN 15736598.

GUNNELL, D.; PLATT, S.; HAWTON, K. The economic crisis and suicide. **BMJ (Online)**, v. 338, n. 7709, p. 1456–1457, 2009. ISSN 17561833.

GUYON, I. et al. Gene selection for cancer classification using support vector machines. **Machine Learning**, v. 46, n. 1-3, p. 389–422, 2002. ISSN 08856125.

HANDLEY, T. E. et al. Predictors of suicidal ideation in older people: A decision tree analysis. **American Journal of Geriatric Psychiatry**, v. 22, n. 11, p. 1325–1335, 2014. ISSN 15457214.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. Elements of Statistical Learning 2nd ed. **Elements**, v. 27, n. 2, p. 745, 2009. ISSN 01727397. Available from Internet: <http://www-stat.stanford.edu/{~}tibs/book/preface.>

JAPKOWICZ, N. The Class Imbalance Problem: Significance and Strategies. **Proceedings of the 2000 International Conference on Artificial Intelligence**, p. 111—-117, 2000.

JAPKOWICZ, N.; STEPHEN, S. The class imbalance problem: A systematic study. **Intelligent Data Analysis**, v. 6, n. 5, p. 429–449, 2002. ISSN 1088467X.

JUNG, J. S. et al. Prediction models for high risk of suicide in Korean adolescents using machine learning techniques. **PLoS ONE**, Public Library of Science, v. 14, n. 6, jun 2019. ISSN 19326203.

JUST, M. A. et al. Machine learning of neural representations of suicide and emotion concepts identifies suicidal youth. **Nature Human Behaviour**, v. 1, n. 12, p. 911–919, 2017. ISSN 23973374.

KANDEL, S. et al. **Research directions in data wrangling: Visualizations and transformations for usable and credible data**. 2011. 271–288 p.

KOTSIANTIS, S. B. **Decision trees: A recent overview**. 2013. 261–283 p.

KUBAT, M. **An Introduction to Machine Learning**. [S.l.]: Springer International Publishing, 2017. 1–348 p. ISSN 18684408. ISBN 9783319639130.

KUBAT, M.; MATWIN, S. Addressing the Curse of Imbalanced Training Sets: One Sided Selection. **Icml**, 1997. ISSN 0717-6163.

KUHN, M. Building Predictive Models in R Using the caret Package. **Journal of Statistical Software**, v. 28, n. 5, 2008. ISSN 10738746. Available from Internet: <http://www.jstatsoft.org/>.

LAURIKKALA, J. Improving identification of difficult small classes by balancing class distribution. In: **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**. [S.l.]: Springer Verlag, 2001. v. 2101, p. 63–66. ISBN 3540422943. ISSN 16113349.

LEWIS, G. et al. Measuring psychiatric disorder in the community: A standardized assessment for use by lay interviewers. **Psychological Medicine**, v. 22, n. 2, p. 465–486, 1992. ISSN 14698978.

LIBRENZA-GARCIA, D. et al. Prediction of depression cases, incidence, and chronicity in a large occupational cohort using machine learning techniques: An analysis of the ELSA-Brasil study. **Psychological Medicine**, 2020. ISSN 14698978.

MACHADO, G.; MENDOZA, M. R.; CORBELLINI, L. G. What variables are important in predicting bovine viral diarrhea virus? A random forest approach. 2011.

MELTZER, H. et al. **Physical ill health, disability, dependence and depression: Results from the 2007 national survey of psychiatric morbidity among adults in England**. 2012. 102–110 p.

MELTZER, H. et al. The influence of disability on suicidal behaviour. **Alter**, v. 6, n. 1, p. 1–12, 2012. ISSN 18750672.

MENEGHEL, S. N. et al. Epidemiological aspects of suicide in Rio Grande do Sul, Brazil. **Revista de saude publica**, v. 38, n. 6, p. 804–10, 2004. ISSN 0034-8910. Available from Internet: <http://www.ncbi.nlm.nih.gov/pubmed/15608898>.

NICULESCU-MIZIL, A.; CARUANA, R. Predicting good probabilities with supervised learning. In: **ICML 2005 - Proceedings of the 22nd International Conference on Machine Learning**. [S.l.: s.n.], 2005. p. 625–632. ISBN 1595931805.

NOCK, M. K. et al. Cross-national prevalence and risk factors for suicidal ideation, plans and attempts. **British Journal of Psychiatry**, v. 192, n. 2, p. 98–105, feb 2008. ISSN 00071250.

NUNES, M. A. et al. Common mental disorders and sociodemographic characteristics: Baseline findings of the Brazilian Longitudinal Study of Adult Health (ELSA-Brasil). **Revista Brasileira de Psiquiatria**, v. 38, n. 2, p. 91–97, 2016. ISSN 15164446.

OH, B. et al. Prediction of suicidal ideation among korean adults using machine learning: A cross-sectional study. **Psychiatry Investigation**, v. 17, n. 4, p. 331–340, 2020. ISSN 19763026.

OLIVERA, A. R. et al. Comparação de algoritmos de aprendizagem de máquina para construir um modelo preditivo para detecção de diabetes não diagnosticada – ELSA-Brasil: Estudo de acurácia. **Sao Paulo Medical Journal**, FapUNIFESP (SciELO), v. 135, n. 3, p. 234–246, jun 2017. ISSN 15163180.

PODGORELEC, V. et al. **Decision trees: An overview and their use in medicine**. 2002. 445–463 p.

REID, W. H. **Preventing suicide: a global imperative**. 2010. 120–124 p.

REUNANEN, J. Overfitting in making comparisons between variable selection methods. **Journal of Machine Learning Research**, v. 3, p. 1371–1382, 2003. ISSN 15324435.

ROY, A. et al. A machine learning approach predicts future risk to suicidal ideation from social media data. **npj Digital Medicine**, v. 3, n. 1, 2020. ISSN 23986352.

SCHMIDT, M. I. et al. Cohort profile: Longitudinal study of adult health (ELSA-Brasil). **International Journal of Epidemiology**, v. 44, n. 1, p. 68–75, 2015. ISSN 14643685.

SCHUBACH, M. et al. Imbalance-Aware Machine Learning for Predicting Rare and Common Disease-Associated Non-Coding Variants. **Scientific Reports**, v. 7, n. 1, 2017. ISSN 20452322.

SILENZIO, V. M. et al. Sexual orientation and risk factors for suicidal ideation and suicide attempts among adolescents and young adults. **American Journal of Public Health**, v. 97, n. 11, p. 2017–2019, 2007. ISSN 00900036.

SOUZA, L. D. D. M. et al. Ideação suicida na adolescência: Prevalência e fatores associados. **Jornal Brasileiro de Psiquiatria**, v. 59, n. 4, p. 286–292, 2010. ISSN 00472085.

SPIERS, N. et al. **Trends in suicidal ideation in England: The National Psychiatric Morbidity Surveys of 2000 and 2007**. 2014. 175–183 p.

SVETNIK, V. et al. Application of Breiman's Random Forest to modeling structure-activity relationships of pharmaceutical molecules. **Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)**, v. 3077, p. 334–343, 2004. ISSN 16113349.

THARWAT, A. Classification assessment methods. **Applied Computing and Informatics**, 2018. ISSN 22108327.

VLUYMANS, S. Learning from imbalanced data. In: **Studies in Computational Intelligence**. [S.l.]: Springer Verlag, 2019. v. 807, p. 81–110.

WHO. **Suicide: Fact Sheet**. 2017. Available from Internet: <https://www.who.int/en/news-room/fact-sheets/detail/suicidehttp://www.who.int/mediacentre/factsheets/fs398/en/https://www.who.int/en/news-room/fact-sheets/detail/suicide{\%}0Ahttp://www.who.int/mediacentre/factsheets/fs398/>.

WRIGHT, M. N.; ZIEGLER, A. Ranger: A fast implementation of random forests for high dimensional data in C++ and R. **Journal of Statistical Software**, v. 77, n. 1, 2017. ISSN 15487660.

ZOU, H.; HASTIE, T. Regularization and variable selection via the elastic net. **Journal of the Royal Statistical Society. Series B: Statistical Methodology**, v. 67, n. 2, p. 301–320, 2005. ISSN 13697412.

ANNEX A — DATASET VARIABLES DESCRIPTIONS

Table A.1 – Variables removed during cleansing for being of free-text type

Attribute	*Description*
anamnesis_hmpa31h	Which other join issue (Q31)
anamnesis_hmpa37a	Which cancer (Q37)
anamnesis_hmpa38	Other health issues (Q38)
derived_A_CLASSEMEDANTHIPERT	SAH treatment - drug combinations
derived_A_RCPQOUTROPROBLULT12H	Any health issue last 12h
derived_A_STATUSVOP	VOP status
dietary_info_diea130i	Reason for diet (Q130)
dietary_info_diea132p	What supplement (Q132)
dietary_info_diea134a	What coffee (Q134)
discrimination_disa1al	Work discrimination: reason (Q1a)
discrimination_disa2al	Domestic discrimination: reason (Q2a)
discrimination_disa3al	Police discrimination: reason (Q2a)
discrimination_disa4al	Public discrimination: reason (Q4a)
discrimination_disa5al	School discrimination: reason (Q5a)
familial_hfda18d	Which cancer (father) (Q18)
familial_hfda18i	Which cancer (mother) (Q18)
familial_hfda18n	Which cancer (sibling 1)
familial_hfda18s	Which cancer (sibling 2)
familial_hfda18y	Which cancer (sibling 3)
home_info_vifa03a	Other conjugal situation (Q4)
home_info_vifa08a	Partner occupation (Q8)
home_info_vifa10a	Who is the other head of family (Q10)
job_related_hoca03	Main activities in work (Q3)
job_related_hoca14a	Main activities in work (Q14)
job_related_hoca23a	Occupation in first job (Q3)
job_related_hoca24a	Other type of first job (Q4)
job_related_hoca25a	Who is the other head of family (Q5)
job_related_hoca26	Other head of family occupation (Q6)
job_related_hoca27a	Other head of family type of work (Q7)
job_related_hoca32a	Which other "on duty" scheme (Q12)
metadata_centroa	Investigation center
metadata_ecga3	Time (mm:ss)
metadata_rcpadataapini	CPR date of application
negativelifeevents_evea05a	Hospitalized 1x: reason (Q5)
negativelifeevents_evea07a	Hospitalized +1x: reasons (Q7)
others_cssa11qou	What other interference in overload
others_cssa6	Start of ingestion of overload
others_cssa7	End of ingestion of overload
others_esca04	End time for interview
religion_hvsa14a	What other religion (Q14)
socio_economic_psea05a	Color or race (Q5)
socio_economic_psea08a	What other condition of the property (Q8)
women_mula17m	What other contraceptive (Q17)
women_mula20b	Missing medical data
women_mula22h	*<No description>*
women_mula30b	Missing medical data hormonal
women_mula5g	What reasons stopped menstruating (Q5)

ANNEX B — CIS-R QUESTIONNAIRE

CIS- R

A Somatic symptoms

A1 Have you had any sort of ache or pain in the past month?	
[1] Yes	
[2] No	**A2** During the past month have you been troubled by any sort of discomfort, for example, headache or indigestion? [1] Yes [2] No, **Go to section B**

A3 Was this ache or pain/discomfort brought on or made worse because you were feeling low, anxious or stressed? **If informant has more than one pain/discomfort, refer to ANY of them**
[1] Yes [2] No, **GO TO SECTION B**
A4 In the past seven days, including last (DAY OF WEEK), on how many days have you noticed the ache or pain/discomfort?
[1] 4 days or more [2] 1 to 3 days [3] None, **GO TO SECTION B**
A5 In total, did the ache or pain/discomfort last for more than 3 hours on any day in the past week/on that day?
[1] Yes [2] No
A6 In the past week, has the ache or pain/discomfort been RUNNING PROMPT
[1] muito desagradável [2] um pouco desagradável [3] ou No foi desagradável?
A7 Has the ache or pain/discomfort bothered you when you were doing something interesting in the past week?
[1] Yes [2] No/ has not done anything interesting
A8 How long have you been feeling this ache or pain/discomfort as you have just described? SHOW CARD
[1] Less than 2 weeks [2] 2 weeks but less than 6 months [3] 6 months but less than 1 year [4] 1 year but less than 2 years [5] 2 years or more

B Fatigue

B1 Have you noticed that you've been getting tired in the past month?	
[1] Yes	
[2] No	**B2** During the past month, have you felt you've been lacking in energy? [1] Yes [2] No, **GO TO SECTION C**

B3 Do you know why you have been feeling tired/lacking in energy?	
[2] No	
[1] Yes	**(a)** What is the **main** reason? Can you choose from this card? SHOW CARD [1] Problems with sleep [2] Medication [3] Physical illness [4] Working too hard (inc. housework, looking after baby) [5] Stress, worry or other psychological reason [6] Physical exercise, **GO TO SECTION C** [7] Other

B4 In the past seven days, including last (DAY OF WEEK) on how many days have you felt tired/lacking in energy?

[1] 4 days or more
[2] 1a 3 dias
[3] None**, GO TO SECTION C**

B5 Have you felt tired/lacking in energy for more than 3 hours in total on any day in the past week?
Exclude time spent sleeping

[1] Yes [2] No

B6 Have you felt so tired/lacking in energy that you've had to push yourself to get things done during the past week?

[1] Yes, on at least one occasion [2] No

B7 Have you felt tired/lacking in energy when doing things that you enjoy during the past week?

[1] Yes , at least once**, GO TO B9**
[2] No
[3] **Spontaneous**, Does not enjoy anything

B8 Have you in the past week felt tired/lacking in energy when doing things that you **used** to enjoy?

[1] Yes [2] No

B9 How long have you been feeling tired/lacking in energy in the way you have just described? SHOW CARD

[1] Less than 2 weeks
[2] 2 weeks but less than 6 months
[3] 6 months but less than 1 year
[4] 1 year but less than 2 years
[5] 2 years or more

C Concentration and forgetfulness

C1 In the past month, have you had any problems in concentrating on what you are doing?
[1] Yes, problems concentrating [2] No
C2 Have you noticed any problems with forgetting things in the past month?
[1] Yes [2] No

Se C1 e C2 = NO, PULE para a seção D
C4 Since last (DAY OF WEEK), on how many days have you noticed problems with your concentration/memory?
[1] 4 days or more [2] 1 to 3 days [3] None, **GO TO SECTION D**
SE C1 = YES In the past week could you concentrate on a TV programme, read a newspaper article or talk to someone without your mind wandering?
[2] Yes [1] No/ not always
SE C1 = YES **C6** In the past week, have these problems with your concentration actually **stopped** you from getting on with things you used to do or would like to do?
[1] Yes [2] No
SE C2 = YES (Earlier you said you have been forgetting things.) Have you forgotten anything important in the past seven days?
[1] Yes [2] No
C8 How long have you been having the problems with your concentration/memory as you have described? SHOW CARD
[1] Less than 2 weeks [2] 2 weeks but less than 6 months [3] 6 months but less than 1 year [4] 1 year but less than 2 years [5] 2 years or more

D Sleep problems

D1 In the past month, have you been having problems with trying to get to sleep or with getting back to sleep if you woke up or were woken up?	
[1] Yes	
[2] No	**D2** Has sleeping more than you usually do been a problem for you in the past month? [1] Yes [2] No, **GO TO SECTION E**

D3 On how many of the past seven nights did you have problems with your sleep?	
[1] 4 nights or more [2] 1 to 3 nights [3] None, **GO TO SECTION E**	
D4 Do you know why you are having problems with your sleep?	
[1] Yes	**(a)** Can you look at this card and tell me the **main** reason for these problems? SHOW CARD [1] Noise [2] Shift work/too busy to sleep [3] Illness/discomfort [4] Worry/thinking [5] Needing to go to the toilet [6] Having to do something (e.g. look after baby) [7] Tired [8] Medication [9] Other
[2] No	
Se D1 = YES Thinking about the night you had the least sleep in the past week, how long did you spend **trying** to get to sleep? (If you woke up or were woken up I want you to allow a quarter of an hour to get back to sleep). **Only include time spent trying to get to sleep.**	
[3] Less than 1/4 hr, **GO TO SECTION E** [1] At least 1/4 hr but less than 1 hr [2] At least 1 hr but less than 3 hrs	
[2] 3 hrs or more	**D6** In the past week, on how many nights did you spend 3 or more hours trying to get to sleep? [1] 4 nights or more [2] 1 to 3 nights [3] None
D7 Do you wake more than two hours earlier than you need to and then find you can't get back to sleep?	
[1] Yes, **GO TO D10** [2] No, **GO TO D10**	
Se D2 = YES Thinking about the night you slept the longest in the past week, how much longer did you sleep compared with how long you normally sleep for?	
[3] Less than 1/4 hr, **GO TO SECTION E**	

[1] At least 1/4 hr but less than 1 hr
[2] At least 1 hr but less than 3 hrs

[2] 3 hrs or more	**D9** In the past week, on how many nights did you sleep for more than 3 hours longer than you usually do? [1] 4 nights or more [2] 1 to 3 nights [3] None

D10 How long have you had these problems with your sleep as you have described?
SHOW CARD

[1] Less than 2 weeks
[2] 2 weeks but less than 6 months
[3] 6 months but less than 1 year
[4] 1 year but less than 2 years
[5] 2 years or more

E Irritability

E1 Many people become irritable or short tempered at times, though they may not show it. Have you felt irritable or short tempered with those around you in the past month?	
[1] Yes/no more than usual	
[2] No	**E2** During the past month did you get short tempered or angry over things which now seem trivial when you look back on them? [1] Yes [2] No, **GO TO SECTION F**

E3 Since last (DAY OF WEEK), on how many days have you felt irritable or short tempered/angry?		
[1] 4 days or more [2] 1 to 3 days [3] None, **GO TO SECTION G**		
E4 What sort of things made you irritable or short tempered/angry in the past week?		
R:		
E5 In total, have you felt irritable or short tempered/angry for more than one hour (on any day in the past week)?		
[1] Yes [2] No		
E6 During the past week, have you felt so irritable or short tempered/angry that you have wanted to shout at someone, even if you haven't actually shouted?		
[1] Yes [2] No		
E7 In the past seven days, have you had arguments, rows or quarrels or lost your temper with anyone?		
[1] Yes	**(a)** Did this happen once or more than once (in the past week)?	
	[1] Once	**E8** Do you think this was justified? [2] Yes, justified [1] No, not justified
	[0] More than once	**E9** Do you think this was justified on every occasion? [2] Yes [2] No, at least one was unjustified
[2] No		
E10 How long have you been feeling irritable or short tempered/angry as you have described ? SHOW CARD		
[1] Less than 2 weeks [2] 2 weeks but less than 6 months [3] 6 months but less than 1 year [4] 1 year but less than 2 years [5] 2 years or more		

F Worry about physical health

F1 Many people get concerned about their physical health. In the past month, have you been at all worried about your physical health? **Include women who are worried about their pregnancy**	
[1] Yes, worried	
[2] No/concerned	During the past month, did you find yourself worrying that you might have a serious physical illness? [1] Yes [2] No, **GO TO SECTION G**

F3 Thinking about the past seven days, including last (DAY OF WEEK), on how many days have you found yourself worrying about your physical health/that you might have a serious physical illness?
[1] 4 days or more [2] 1 to 3 days [3] None, **GO TO SECTION G**
F4 In your opinion, have you been worrying too much in view of your actual health?
[1] Yes [2] No
F5 In the past week, has this worrying been RUNNING PROMPT
[1] very unpleasant [2] a little unpleasant [3] or not unpleasant?
F6 In the past week, have you been able to take your mind off your health worries at least once, by doing something else?
[2] Yes [1] No, could not be distracted once
F7 How long have you been worrying about your physical health in the way you have described? SHOW CARD
[1] Less than 2 weeks [2] 2 weeks but less than 6 months [3] 6 months but less than 1 year [4] 1 year but less than 2 years [5] 2 years or more

G Depression

G1 Almost everyone becomes sad, miserable or depressed at times. Have you had a spell of feeling sad, miserable or depressed in the past month?	
[1] Yes	**G4** In the past **week** have you had a spell of feeling sad, miserable or depressed? **Use informant's own words if possible** [1] Yes [2] No
[2] No	
G2 During the past month, have you been able to enjoy or take an interest in things as much as you usually do?	
[1] Yes	
[2] No/no enjoyment or interest	**G5 -** In the past **week** have you been able to enjoy or take an interest in things as much as usual? **Use informant's own words if possible** [2] Yes [1] No/no enjoyment or interest

Se G1 = NO e G2 = YES, GO TO SECTION I

Se G4 = NO e G5 = YES, GO TO SECTION I

Since last (DAY OF WEEK) on how many days have you felt sad, miserable or depressed/unable to enjoy or take an interest in things?

[1] 4 days or more
[2] 1 to 3 days
[3] None

G7 Have you felt sad, miserable or depressed/unable to enjoy or take an interest in things for more than 3 hours in total (on any day in the past week)?

[1] Yes [2] No

G8 (a) What sorts of things made you feel sad, miserable or depressed/unable to enjoy or take an interest in things in the past week? Can you choose from this card?
What was the main thing? **Ring code in column (b)**
SHOW CARD

	(a) Code all that apply	(b) **Code one only**
Members of the family	[01]	[01]
Relationship with spouse/partner	[02]	[02]
Relationships with friends	[03]	[03]
Housing	[04]	[04]
Money/bills	[05]	[05]
Own physical health (inc. pregnancy)	[06]	[06]

Own mental health	[07]	[07]
Work or lack of work (inc. student)	[08]	[08]
Legal difficulties	[09]	[09]
Political issues/the news	[10]	[10]
Other	[11]	[11]
Don't know/no main thing	[99]	[99]

G9 In the past week when you felt sad, miserable or depressed/unable to enjoy or take an interest in things, did you ever become happier when something nice happened, or when you were in company?

[2] Yes, at least once
[1] No

G10 How long have you been feeling sad, miserable or depressed/unable to enjoy or take an interest in things as you have described? **Show card**

[1] Less than 2 weeks
[2] 2 weeks but less than 6 months
[3] 6 months but less than 1 year
[4] 1 year but less than 2 years
[5] 2 years or more

H Depressive Ideas

Informants who scored 1 or more at section G, Depression, GO TO SECTION I
I would now like to ask you about when you have been feeling sad, miserable or depressed/unable to enjoy or take an interest in things. In the past week, was this worse in the morning or in the evening, or did this make no difference? **Prompt as necessary**
[1] in the morning [2] in the evening [3] no difference/other
H2 Many people find that feeling sad, miserable or depressed/unable to enjoy or take an interest in things can affect their interest in sex. Over the past month, do you think your interest in sex has : RUNNING PROMPT
[1] increased [2] decreased [3] or has it stayed the same? [4] **Spontaneous** Not applicable
H3 When you have felt sad, miserable or depressed/unable to enjoy or take an interest in things in the past seven days, a…have you been so restless that you couldn't sit still?
[1] Yes [2] No
H3b – have you been doing things more slowly, for example, walking more slowly?
[1] Yes [2] No
H3c - have you been less talkative than normal?
[1] Yes [2] No
H4 Now, thinking about the past seven days have you on at least one occasion felt guilty or blamed yourself when things went wrong when it **hasn't** been your fault?
[1] Yes, at least once [2] No
H5 During the past week, have you been feeling you are not as good as other people?
[1] Yes [2] No
H6 Have you felt hopeless at all during the past seven days, for instance about your future?
[1] Yes [2] No
H7 – **Interviewer check** : Se H4 = No e H5 = No e H6 = No, **GO TO SECTION I** Se H4 = Yes ou H5 = Yes ou H6 = Yes
H8 – In the past week have you felt that life isn't worth living?
[1] Yes [2] **Spontaneous:** Yes, but not in the past week [3] No, **GO TO H10**
H9 In the past week, have you thought of killing yourself?
[1] Yes [2] **Spontaneous:** Yes, but not in the past week [3] No, **GO TO H10**

<table>
<tr><td colspan="2">(a) Have you talked to your doctor about these thoughts (of killing yourself)?
[1] Yes</td></tr>
<tr><td>[2] Spontaneous: No, but has talked to other people
[3] No</td><td>(b) (You have said that you are thinking about committing suicide.)
Since this is a very serious matter it is important that you talk to your doctor about these thoughts.</td></tr>
<tr><td colspan="2">READ
H10 (Thank you for answering those questions on how you have been feeling. I would now like to ask you a few questions about worrying.)</td></tr>
</table>

I Worry

I 1 (The next few questions are about worrying.) In the past month, did you find yourself worrying more than you needed to about things?	
[1] Yes, worrying	
[2] No/concerned	**I 2** Have you had any worries at all in the past month? [1] Yes [2] No**, GO TO SECTION J**

I 3 (a) Can you look at this card and tell me what sorts of things you worried about in the past month?
(b) What was the main thing you worried about?

	(a) Code all that apply	(b) **Code one only**
Members of the family	[01]	[01]
Relationship with spouse/partner	[02]	[02]
Relationships with friends	[03]	[03]
Housing	[04]	[04]
Money/bills	[05]	[05]
Own physical health (inc. pregnancy) **GO TO SECTION J**	[06]	[06]
Own mental health	[07]	[07]
Work or lack of work (inc. student)	[08]	[08]
Legal difficulties	[09]	[09]
Political issues/the news	[10]	[10]
Other	[11]	[11]
Don't know/no main thing	[99]	[99]

For the next few questions, I want you to think about the worries you have had **other** than those about your physical health.

I 6 On how many of the past seven days have you been worrying about things (other than your physical health)?

[1] 4 days or more
[2] 1 to 3 days
[3] None**, PULE P. SEÇÃO J**

I 7 In your opinion, have you been worrying too much in view of your circumstances?
Refer to worries other than those about physical health

[1] Yes [2] No

I 8 In the past week, has this worrying been: RUNNING PROMPT
Refer to worries other than those about physical health

[1] very unpleasant
[2] a little unpleasant
[3] or not unpleasant?

I 9 Have you worried for more than 3 hours in total on any one of the past seven days?
Refer to worries other than those about physical health

[1] Yes [2] No
I 10 How long have you been worrying about things in the way that you have described? SHOW CARD
[1] Less than 2 weeks [2] 2 weeks but less than 6 months [3] 6 months but less than 1 year [4] 1 year but less than 2 years [5] 2 years or more

J Anxiety

J1 Have you been feeling anxious or nervous in the past month?	
[1] Yes, anxious or nervous	
[2] No	**J2** In the past month, did you ever find your muscles felt tense or that you couldn't relax? [1] Yes [2] No
J3 Some people have phobias; they get nervous or uncomfortable about specific things or situations when there is no real danger. For instance they may get nervous when speaking or eating in front of strangers, when they are far from home or in crowded rooms, or they may have a fear of heights. Others become nervous at the sight of things like blood or spiders. In the past month have you felt anxious, nervous or tense about any specific things or situations when there was no real danger?	
[1] Yes [2] No	

J4 – **Interviewer check: Se J1 = YES ou J2 = YES e J3 = YES, go to J5** **Se J1 = YES ou J2 = YES e J3 =NO go to J6** **Se J1 = NO e J3 = NO, GO TO SECTION K**
J5 In the past month, when you felt anxious/nervous/tense, was this always brought on by the phobia about some **specific** situation or thing or did you sometimes feel **generally** anxious/nervous/tense?
[1] Always brought on by phobia, **GO TO SECTION K** [2] Sometimes felt generally anxious
J6 The next questions are concerned with **general** anxiety/nervousness/tension **only.** I will ask you about the anxiety which is brought on by the phobia about specific things or situations later. On how many of the past seven days have you felt **generally** anxious/nervous/tense?
[1] 4 days or more [2] 1 to 3 days [3] None, **GO TO SECTION K**
J8 In the past week, has your anxiety/nervousness/tension been: RUNNING PROMPT
[1] very unpleasant [2] a little unpleasant [3] or not unpleasant?
J9 In the past week, when you've been anxious/nervous/tense, have you had any of the symptoms shown on this card? SHOW CARD

[1] Yes	**(a)** Which of these symptoms did you have when you felt anxious/nervous/tense? **Code all that apply** [1] Heart racing or pounding [2] Hands sweating or shaking [3] Feeling dizzy tontura [4] Difficulty getting your breath [5] Butterflies in stomach [6] Dry mouth

	[7] Nausea or feeling as though you wanted to vomit
[2] No	
J10 Have you felt anxious/nervous/tense for more than 3 hours in total on any one of the past seven days?	
[1] Yes [2] No	
J11 How long have you had these feelings of general anxiety/nervousness/tension as you described? SHOW CARD	
[1] Less than 2 weeks [2] 2 weeks but less than 6 months [3] 6 months but less than 1 year [4] 1 year but less than 2 years [5] 2 years or more	

CIS-R - SEÇÃO K – FOBIAS

Se J3 =No

K2 Sometimes people avoid a specific situation or thing because the have a phobia about it. For instance, some people avoid eating in public or avoid going to busy places because it would make them feel nervous or anxious.

In the past month, have you avoided any situation or thing because it would have made you feel nervous or anxious, even though there was no real danger?

[1] Yes
[2] No, **GO TO SECTION L**

Se J3 = YES

K3(a) Can you look at this card and tell me which of the situations or things listed made you the **most** anxious/nervous/tense in the past month? SHOW CARD

K2 = Yes

(b) Can you look at this card and tell me, which of these situations or things did you avoid the most in the past month? SHOW CARD

[1] Crowds or public places, including travelling alone or being far from home
[2] Enclosed spaces
[3] Social situations, including eating or speaking in public, being watched or stared at
[4] The sight of blood or injury
[5] Any specific single cause including insects, spiders and heights
[6] Other **(specify)**:_____ _____

K4 - In the past seven days, how many times have you **felt** nervous or anxious about (SITUATION/THING)?

[1] 4 times or more
[2] 1 a 3 times
[3] None, **PULAR P. K6**

K5 In the past week, on those occasions when you felt anxious/nervous/tense did you have any of the symptoms on this card?
SHOW CARD

[1] Yes	**(a)** Which of these symptoms did you have when you felt anxious/nervous/tense? **Code all that apply** [1] Heart racing or pounding [2] Hands sweating or shaking [3] Feeling dizzy tontura [4] Difficulty getting your breath [5] Butterflies in stomach [6] Dry mouth [7] Nausea or feeling as though you wanted to vomit

[2] No

K6 In the past week, have you **avoided** any situation or thing because it would have made you feel

<table>
<tr><td colspan="2">anxious/nervous/tense even though there was no real danger?</td></tr>
<tr><td>[1] Yes</td><td>K7 How many times have your avoided such situations or things in the past seven days?
[1] 1 a 3 times
[2] 4 times or more
[3] None</td></tr>
<tr><td colspan="2">[2] No</td></tr>
<tr><td colspan="2">K8 How long have you been having these feelings about these situations/things as you have just described?
SHOW CARD</td></tr>
<tr><td colspan="2">[1] Less than 2 weeks
[2] 2 weeks but less than 6 months
[3] 6 months but less than 1 year
[4] 1 year but less than 2 years
[5] 2 years or more</td></tr>
</table>

L Panic

Informants who felt anxious in the past month

Thinking about the past month, did your anxiety or tension ever get so bad that you got in a panic, for instance make you feel that you might collapse or lose control unless you did something about it?
[1] Yes [2] No, **GO TO SECTION M**
L2 How often has this happened in the past week?
[1] Once [2] More than once [3] Not at all, **GO TO SECTION M**
L3 In the past week, have these feelings of panic been: RUNNING PROMPT
[2] a little uncomfortable or unpleasant [1] or have they been very unpleasant or unbearable?
L4 Did this panic/the worst of these panics last for longer than 10 minutes?
[1] Yes [2] No
L5 Are you relatively free of anxiety between these panics?
[1] Yes [2] No
L6 – Is this panic always brought on by (SITUATION/THING)? **Refer to situation/thing at K3.**
[1] Yes [2] No
L7 How long have you been having these feelings of panic as you have described? SHOW CARD
[1] Less than 2 weeks [2] 2 weeks but less than 6 months [3] 6 months but less than 1 year [4] 1 year but less than 2 years [5] 2 years or more

M Compulsions

M1 In the past month, did you find that you kept on doing things over and over again when you knew you had already done them, for instance checking things like taps or washing yourself when you had already done so?
[1] Yes [2] No, **GO TO SECTION N**
M2 On how many days in the past week did you find yourself doing things over again that you had already done?
[1] 4 days or more [2] 1 to 3 days [3] None , **GO TO SECTION N**
M3 Since last (DAY OF WEEK) what sorts of things have you done over and over again?
R:
M4 During the past week, have you tried to stop yourself repeating (BEHAVIOUR)/doing any of these things over again?
[1] Yes [2] No
M5 Has repeating (BEHAVIOUR)/doing any of these things over again made you upset or annoyed with yourself in the past week?
[1] Yes, upset or annoyed [2] No, not at all
M6 If more than one thing is repeated at M3 Thinking about the past week, which of the things you mentioned did you repeat the **most** times?
_____
M7 Since last (DAY OF WEEK), how many times did you repeat (BEHAVIOUR) when you had already done it? **Refer to BEHAVIOUR at M6, if applicable**
[1] 3 or more repeats [2] 2 repeats [3] 1 repeat
M8 How long have you been repeating (BEHAVIOUR)/any of the things you mentioned in the way which you have described? SHOW CARD
[1] Less than 2 weeks [2] 2 weeks but less than 6 months [3] 6 months but less than 1 year [4] 1 year but less than 2 years [5] 2 years or more

N Obsessions

N1 In the past month did you have any thoughts or ideas over and over again that you found unpleasant and would prefer not to think about, that still kept on coming into your mind?
[1] Yes [2] No, **GO TO SECTION O**
N2 Can I check, is this the **same** thought or idea over and over again or are you worrying about something in general?
[1] Same thought [2] Worrying in general, **GO TO SECTION O**
N3 What are these unpleasant thoughts or ideas that keep coming into your mind? **Do not probe Do not press for answer**
R:
N4 Since last (DAY OF WEEK), on how many days have you had these unpleasant thoughts?
[1] 4 days or more [2] 1 to 3 days [3] None , **GO TO SECTION O**
N5 During the past week, have you tried to stop yourself thinking any of these thoughts?
[1] Yes [2] No
N6 Have you become upset or annoyed with yourself when you have had these thoughts in the past week?
[1] Yes, upset or annoyed [2] No
N7 In the past week, was the longest episode of having such thoughts : RUNNING PROMPT
[1] a quarter of an hour or longer [2] or was it less than this?
N8 How long have you been having these thoughts in the way which you have just described? SHOW CARD
[1] Less than 2 weeks [2] 2 weeks but less than 6 months [3] 6 months but less than 1 year [4] 1 year but less than 2 years [5] 2 years or more

O Overall effects

Informants who scored 2 or more on any section, A to N.

Now I would like to ask you how all of these things that you have told me about have affected you overall. In the past week, has the way you have been feeling ever actually **stopped** you from getting on with things you used to do or would like to do?	
[1] Yes	**(a)** In the past week, has the way you have been feeling stopped you doing things once or more than once? [1] Once [2] More than once
[2] No	**(b)** Has the way you have been feeling made things more difficult even though you have got everything done? [1] Yes [2] No