COGNITIVE, BEHAVIOURAL AND SOCIAL DATA

DETECT DISHONEST ANSWERS TO SURVEYS USING MACHINE LEARNING.

University of Padua

Department of Mathematics

Academic Year 2021/2022

Alberto Sinigaglia	mat. 2044712
Beatrice Sofia Bertipaglia	mat. 2054766
Chiara Colato	mat. 2062670
Sara Repetto	mat. 2055518

Abstract

Lying can be defined as the act of hiding the truth using a false statement with the intention to make someone else believe it. The evaluation of a statement to detect dishonesty is known as ***lie detection****.*
In this article we aim to address the ***reliability problem*** *and mitigate it by finding a reliable feature selection criterion, i.e. a model that always gives similar importance to features, which, at the same time, gives good performance in lie detection.*
We will show that not all feature selection techniques are good in terms of reliability, but among those that are, there is no benchmark method that outperforms all others in every dataset. There are, however, some good candidates who appear to be reliable across all datasets tested.
More in detail, we will test ***Model Agnostic techniques*** *and* ***Psychometric Inspired techniques*** *and we will compare their performances over the available datasets. For each of the proposed methods two metrics will be used for evaluation: firstly their* ***accuracy score*** *and secondly their "****robustness index****". The attention will be more focused on the second point, i.e. the best choice will be a compromise between the two measures, with a greater weight for the robustness of models.*
The conclusion reached will be that, comparing the average results obtained on all datasets, the best performance in terms of accuracy will be given by Model Agnostic techniques, but at the same time they will show worst results in terms of robustness. Psychometric Inspired techniques, instead, will show better values for the robustness index while they will reach smaller accuracy scores.
Following the goal of finding the best trade-off, we will finally conclude that we can obtain good results, on average, using Psychometric Inspired techniques.

1 Introduction

Lying can be defined as the act of hiding the truth using a false statement with the intention to make someone else believe it.
The evaluation of a statement to detect dishonesty is known as **lie detection**. There are different techniques used to examine the truth or lies of people, like *polygraph test* that analyzes heart and breathing rates, *psychological stress evaluation* that measures human stress by analyzing the voice. Other methods consist in observing micro facial expressions or brain waves.

But how can we tell if a person is saying the truth just by looking at the answers he gave to a survey, without knowing his heartbeat rate, his neural activity or his body movements while he was answering to the questions?

The solution to this problem is useful, for example, for a social security institution that has to decide whether to give economic support to a person who claims to be a victim of PTSD. Another example, which benefits from this solution, could be a judge of Family Court that has to evaluate a parent to determine if he is fit to look after his child. In these cases it is normal for humans to lie or exaggerate their condition in order to achieve their goal, therefore every compiled psychological test must be analyzed using lie detection methods.

1.1 Reliability problem

Besides wanting a model able to detect when a person is lying in a survey, we also want this model to be *reliable*, i.e. that produces similar results under consistent conditions. In other words this means that it can be generalized with good results on multiple surveys. Indeed the aim is that the final model can be used by anyone who wants to test the veracity of any psychological test.

1.2 Datasets description

The first step for the construction of a reliable predictor is collecting data. Thirteen different surveys were taken into consideration, in order to analyze the ability of the final model to be reliable. Participants were asked to respond twice to tests developed to identify different personality aspects or behavioral disorders, the first time answering honestly and the second time lying to appear worse (called Faking Bad) or better (called Faking Good), according to a given assignment. Each question is answered using a *Likert scale*, i.e. integers ranging, for example, from 1 (strong disagreement) to 5 (strong agreement). Below are shown the questionnaires analyzed with the relative specifications.

1. **(short) Dark Triad** [**1**]:
   (a) 241 participants answered twice to 27 questions formulated to identify machiavellism, narcissism and psychopathy disorders. Assignment for dishonest answers (Faking Good): *You are going through a job interview that you care a lot for, and you have to put yourself under the better light possible, avoiding showing undesirable personality traits*.
   (b) 432 participants answered twice to 27 questions formulated to identify machiavellianism, narcissism and psychopathy disorders. Assignment for dishonest answers (Faking Good): *You are a parent that is undergoing a cause for the custody of your children, therefore you have to show to be a good parent*.

2. **Prospective and Retrospective Memory** [**2**]:
702 participants answered twice to 16 questions formulated to identify specific memory difficulties. Assignment for dishonest answers (Faking Bad): *You have to fake a memory deficit in order to trick a social security institution into giving you economic support*.

3. **Post Traumatic Stress Disorder** [**3**]:
201 participants answered twice to 20 questions formulated to identify post traumatic stress disorder. Assignment for dishonest answers (Faking Bad): *You have to convince the National Centre for PTSD to suffer from post traumatic stress disorder in order to be given an economic support*.

4. **Negative Acts - Revised** [**4**]:
356 participants answered twice to 22 questions formulated to identify possible victims of mobbing. Assignment for dishonest answers (Faking Bad): *You have to pretend to be victim of mobbing, i.e. suffering physical and/or emotional abuse, in your workplace*.

5. **Depression and Generalized Anxiety Disorder** [**5**]:
559 participants answered twice to 16 questions formulated to identify possible victims of anxious-depressive syndrome. Assignment for dishonest answers (Faking Bad): *You have to trick a clinic to be suffering from anxiety and depression*.

6. **Personality Disorders** [**6**]:
412 participants answered twice to 220 questions formulated to identify personality traits including negative affect, detachment, antagonism, disinhibition, and psychoticism. Assignment for dishonest answers (Faking Bad): *You have to pretend to have a personality disorder in order to receive care from a clinic*.

7. **(short) Personality Disorders** [**7**]:
519 participants answered twice to 25 questions formulated to identify personality traits including negative affect, detachment, antagonism, disinhibition, and psychoticism. Assignment for dishonest answers (Faking Bad): *You have to pretend to have a personality disorder in order to receive care from a clinic*.

8. **Parental Reflective Functioning** [**8**]:
263 participants answered twice to 18 questions formulated to identify a parent's capacity to reflect upon his child's internal mental experience. Assignment for dishonest answers (Faking Good): *You are a parent that is going through a Family Court evaluation and being examined by a psychologist to determine if you are fit to look after your child*.

9. **Impact of Event Scale - Revised** [**9**]:
179 participants answered twice to 22 questions formulated to identify post traumatic stress disorder. Assignment for dishonest answers (Faking Bad): *You have to convince the National Centre for PTSD to suffer from post traumatic stress disorder in order to be given an economic support*.

10. **NEO Personality Inventory - Revised** [**10**]:
67174 participants answered to 120 questions formulated to assess an individual on five dimensions of personality, the so-called Big Five personality traits, that are openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism. True dishonest answers (Faking Good) are taken from a real job application setting where identity is disclosed, people are highly motivated to get the job they applied for. The other 10513 participants answered to the same 120 questions anonymously with the aim to know themselves, those are the honest responses.

11. **Dirty Dozen Dark Triad** [**11**]:
492 participants answered twice to 12 questions formulated to identify three socially undesirable personality traits that are narcissism, psychopathy, and machiavellianism. Assignment for dishonest answers (Faking Bad): *You have to convince the National Centre for PTSD to suffer from post traumatic stress disorder in order to be given an economic support*.

12. **International Adjustment Disorder** [**12**]:
225 participants answered twice to 9 questions formulated to identify adjustment disorder, a stress response syndrome defined as a maladaptive behavioural and/or emotional reaction that occurs when an individual is unable to properly cope or adjust to a stressful life event. Assignment for dishonest answers (Faking Bad): *You have to imagine being in an insurance setting and fake an adjustment disorder*.

13. **Big Five Inventory** [**13**]:

    (a) 221 participants answered twice to 10 questions formulated to assess an individual on five dimensions of personality. Assignment for dishonest answers (Faking Good): *You are going through a job interview for a salesperson position*.

    (b) 230 participants answered twice to 10 questions formulated to assess an individual on five dimensions of personality. Assignment for dishonest answers (Faking Good): *You are going through a job interview for a role in an humanitarian organization*.

(c) 243 participants answered twice to 10 questions formulated to assess an individual on five dimensions of personality. Assignment for dishonest answers (Faking Good): *You want to obtain child custody in the context of a litigation.*

1.3 Pipeline

For computational reasons we analyze feature selection techniques first only on a chosen dataset, then the best methods are selected and applied on all other datasets. The chosen dataset is number (1b), the short Dark Triad questionnaire with dishonest answers that assume a battle for child custody. We selected this dataset because, as it can be seen in Figure 1, it is one of the most complex in terms of features distribution, honest and dishonest answers are similar to each other. This choice makes it more difficult to choose a good reliable predictor, but, at the end, makes it easier to generalize the final model to other simpler datasets.

The chosen survey is available on the table below, each question has to be answered with one of the following integers: 1 (strongly disagree), 2 (disagree), 3 (neither agree nor disagree), 4 (agree), 5 (strongly agree).

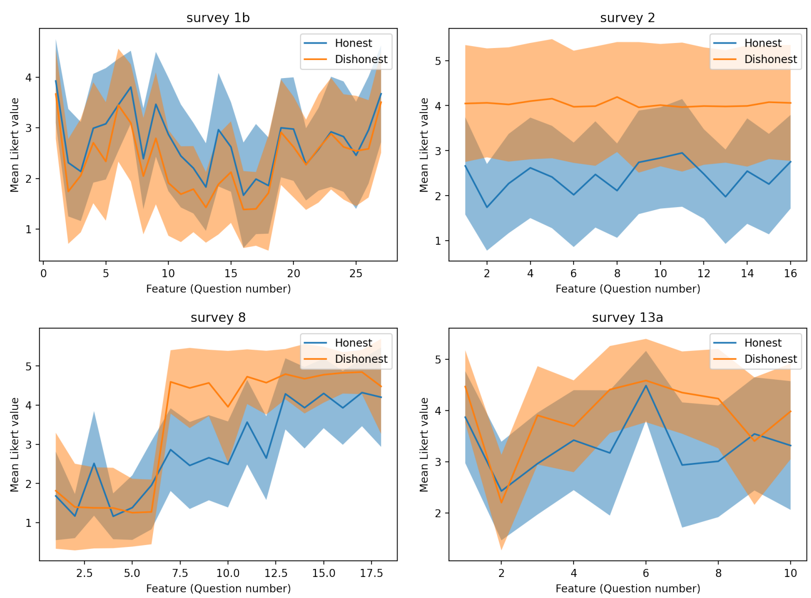

Figure 1: Features distribution of honest and dishonest answers at surveys (1b), (2), (8) and (13a).

(short) Dark Triad survey	
	Machiavellianism
1	It's not wise to tell your secrets.
2	I like to use clever manipulation to get my way.
3	Whatever it takes, you must get the important people on your side.
4	Avoid direct conflict with others because they may be useful in the future.
5	It's wise to keep track of information that you can use against people later.
6	You should wait for the right time to get back at people.
7	There are things you should hide from other people to preserve your reputation.
8	Make sure your plans benefit yourself, not others.
9	Most people can be manipulated.
	Narcissism
10	People see me as a natural leader.
11	I hate being the center of attention. (R)
12	Many group activities tend to be dull without me.
13	I know that I am special because everyone keeps telling me so.
14	I like to get acquainted with important people.
15	I feel embarrassed if someone compliments me. (R)
16	I have been compared to famous people.
17	I am an average person. (R)
18	I insist on getting the respect I deserve.
	Psychopathy
19	I like to get revenge on authorities.
20	I avoid dangerous situations. (R)
21	Payback needs to be quick and nasty.
22	People often say I'm out of control.
23	It's true that I can be mean to others.
24	People who mess with me always regret it.
25	I have never gotten into trouble with the law. (R)
26	I enjoy having sex with people I hardly know
27	I'll say anything to get what I want.

On the chosen dataset we apply the following techniques, for all of them we analyze their performances on making predictions and their ability to select important features:
2.2 **Model dependent feature selection**:
Random Forest, Logistic Regression, Gradient Boosting, Support Vector Machine with linear kernel, Voting and Staking Ensembles;
2.3 **Model independent feature selection with classifiers**:
Permutation Importance, Leave One Feature Out;
2.4 **Model independent feature selection without classifiers**:
PCA, Kernel PCA, Sparse PCA, Truncated SVD, Factor Analysis, Relief.

First, for each method, accuracies using all features and using only top 20% of features ordered by feature importance are compared, in order to see how much the performances change on the correspondent reduced dataset.
For each technique top 20% of features is compared with top 20% of all other methods, and its ability to give always the same order to top 20% of features is analyzed, defining a *robustness index* for better comparisons.

Do different methods that have similar accuracy order features by importance in a different order? Which techniques are the best in terms of accuracy and robustness index on the current 20%-reduced dataset?

At the end the best methods are applied on all other datasets.

Is there a technique that works better for all datasets? Do the reason why participants lie (i.e. datasets 1 and 13) affects the selection of most important features?

2 Methods

In this section the procedure and the methods used for dealing with our task are described. In particular we are interested into finding important features (corresponding to the most relevant proposed questions), i.e. the ones that show different behaviours, in terms of distributions, for the *honest* and *dishonest* classes. After the most important questions are identified, we can use the information contained in the corresponding answers to identify if a person is being honest or dishonest.

Important to mention is that most of the datasets have a dimensionality problem, the assumption $\#examples >> \#features$ does not always hold. In situations like those, encoding the features as one-hot would make this behaviour even worst. To overcome this problem, starting from now, we assume that "lying" is monotonous with respect to honest responses. This means that, given a honest response to a specific question, the dishonest version will be exaggerated on one of the two sides most of the times.

To be more clear, in Figure 2 are reported some of plots that show the difference we are looking for. Observing the distribution lines, in almost all the cases, we can see that the answers of honest and dishonest people follow quite a different path: more evident is the difference, more the corresponding feature is useful for reaching our goal.

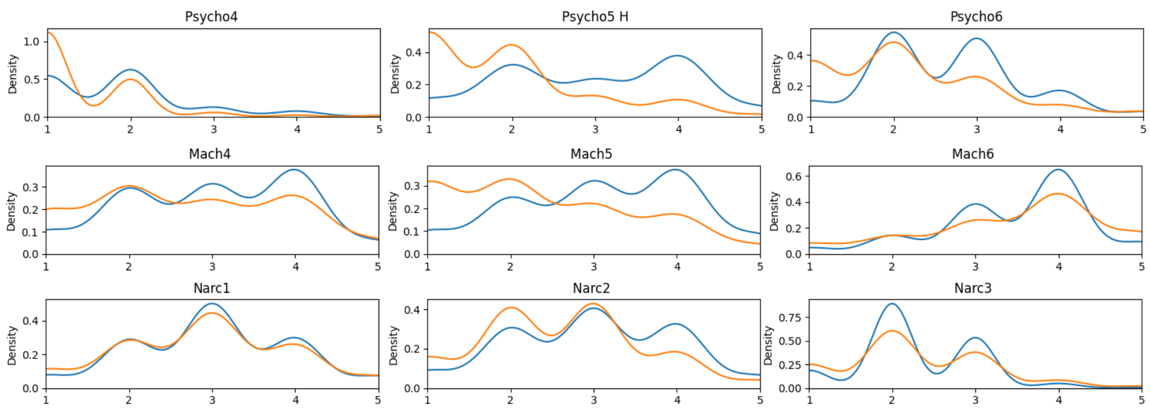

Figure 2: Distributions for the features *Psycho4, Psycho5 H, Psycho6, Mach4, Mach5, Mach6, Narc1, Narc2, Narc3*, honest class in blue and dishonest class in orange.

2.1 Correlation Check

As we saw the number of features potentially involved within the models could be significant, in these cases it is common to encounter collinearity and multi-collinearity problems, due also to the psychological nature of the questions themself.
To prevent the phenomenon from causing confusion and misunderstanding in the relations between features, the first step we perform is the analysis of the correlations between all features in each dataset. After a first exploratory analysis, we decide to remove the features correlated, using as threshold the value 0.5: for each feature we check its correlation with others, and if it overcomes the value chosen as threshold we delete it.
Then we are ready to start with our features selection procedure.

2.2 Model Dependent Feature Selection

In this section we analyze the only 4 classification models that, given our dataset in input, have available a built-in method that computes the importance of each feature in the classification.

1. **Random Forest**
   It constitutes a learning method for classification that operates by building multiple decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. This method can be used to rank the importance of variables in a natural way: during the fitting process, indeed, the out-of-bag error for each data point is recorded and averaged over the forest. To measure the importance of the $j-th$ feature after training, the values of the $j-th$ feature are permuted among the training data and the out-of-bag error is again computed on this perturbed data set. The importance score for the $j-th$ feature is computed by averaging the difference in out-of-bag error before and after the permutation over all trees. The score is then normalized. Features which produce large values for this score are ranked as more important than features which produce small values.

2. **Logistic Regression**
   In statistics, the logistic model describes the probability of a specific event taking place. In this case we are considering this tool as a binary classifier and, formally, in binary logistic regression there is a single binary dependent variable, coded by an indicator variable, where the two values are labeled "0" and "1". In our study, the label "1" correspond to a *honest* unit. As mentioned for the Random Forest, also here we look for a way to get the importance of the features used in the model. In this case, we can consider the absolute value of the parameters as the importance given from the model to the correspondent explanatory variable.

3. **Gradient Boosting**
   Gradient boosting is a machine learning technique that gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees; it usually outperforms random forest. A gradient-boosted trees model is built in a stage-wise fashion as in other boosting methods, but it generalizes the other methods by allowing optimization of an arbitrary differentiable loss function. Also the Gradient Boosting method has a built-in function that can give us the importance of each feature and provides a score that indicates how useful or valuable the features were in the construction of the boosted decision trees within the model. The more

an attribute is used to make key decisions with decision trees, the higher its relative importance.

4. **Support Vector Machine with Linear Kernel**
   It is a supervised learning model with associated learning algorithm that analyze data for classification analysis. Given the set of training examples, each marked as belonging to one of two categories, the SVM training algorithm builds a model that assigns new examples to one category or the other, maximising the width of the gap between the two categories. More formally, a support vector machine constructs a hyperplane or set of hyperplanes in a high or infinite-dimensional space, which can be used for classification tasks. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall. Also in this case, as in the Logistic Regression one, we can consider as features importance the coefficients assigned from the model to the independent variables.

In order to have a first idea of the importance given by the different models to each single question we are considering, we can look at the plot reported in Figure 3.

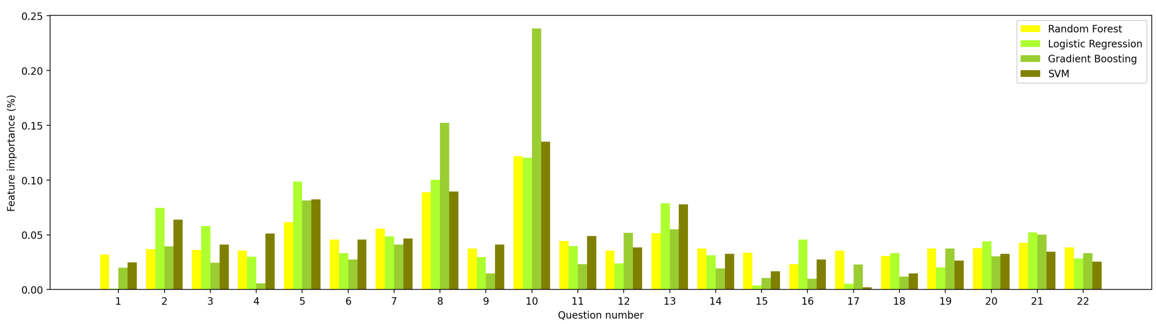

Figure 3: Features Importance assigned to each question described by Random Forest, Logistic Regression, Gradient Boosting and SVM models.

Before making conclusions, we retain useful to consider also the accuracy reached with the methods and in particular we compute it considering different percentages of the features considered as most relevant from each model as we can observe in Figure 4. In this way we are able to observe also how the performance of each used method changes. We reproduce the analysis considering, first, the top features (the ones considered as most important) selected by all the four models together, at the end we try to build a voting ensemble with the models as building blocks, in order to see if combining them can obtain better results, as we can see in Figure 5.

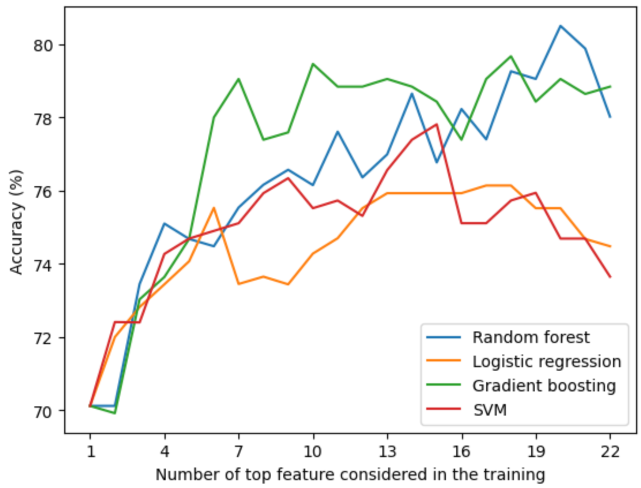

Figure 4: Percentage of reached accuracy considering only top X features ordered by importance according to the training phase, for each of the four proposed models.

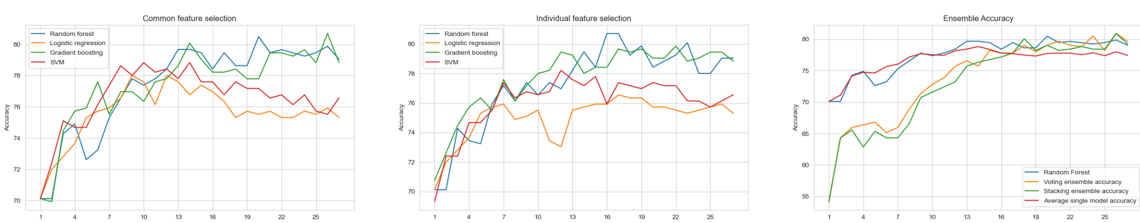

Figure 5: Percentage of reached accuracy considering only top X features ordered by importance: w.r.t. all the models together on the left, w.r.t. each single model at center, considering a voting ensemble on the right.

2.3 Model Independent Feature Selection with Classifiers

To perform feature selection, features importance are calculated and top 20% of features are selected, ordered by importance. Features importance was calculated using several methods. In model independent methods features importance is calculated as the average of features importance obtained using different classifiers. In addition to the previously described classifiers, five other classifiers were used: Naive Bayes classifier, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Adaptive Boosting and histogram-based Gradient Boosting classification Tree. The model agnostic techniques used in pairs with each of these classifiers are:

1. **Permutation Importance**
   The importance to each feature is given according to how much the accuracy of a model decreases when it is randomly shuffled. The random assignment of feature values to the observations obviously changes the relationship between the feature and the target. If the model relies heavily on that feature to make the prediction, this would cause a sharp reduction in the accuracy. We can observe some results and comparisons looking at the Figure 6.

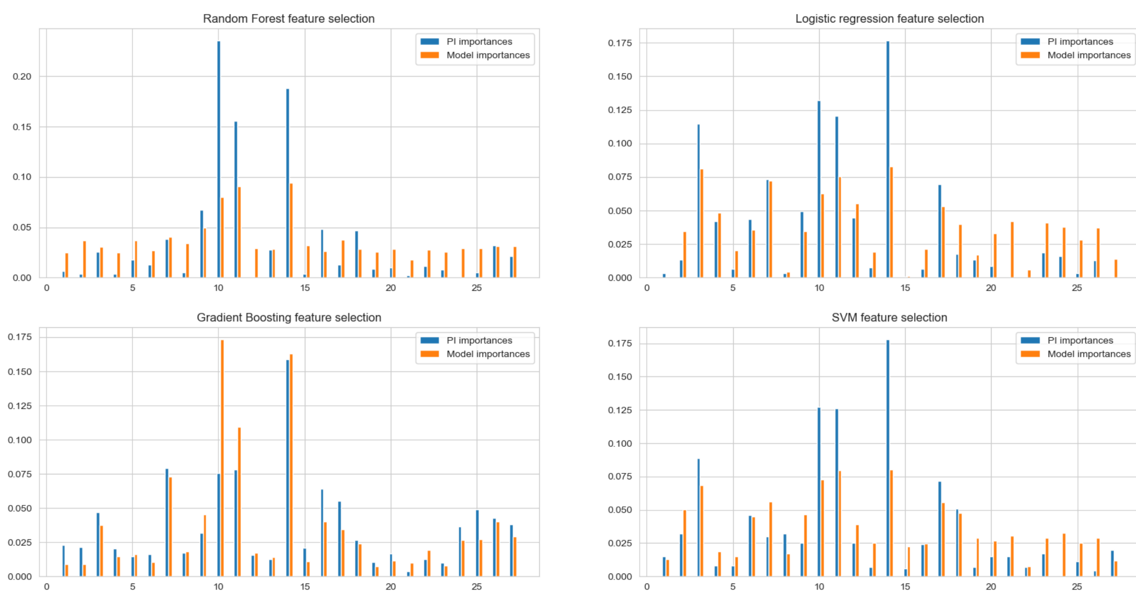

Figure 6: Comparison between the importance given to the features from the built-in methods of the models and the importance given by the Permutation Importance method.

2. **Leave One Feature Out**

The importance to each feature is given according to how much the accuracy of a model decreases when that feature is removed from the features of the training set. We can observe the reached results in Figure 7.

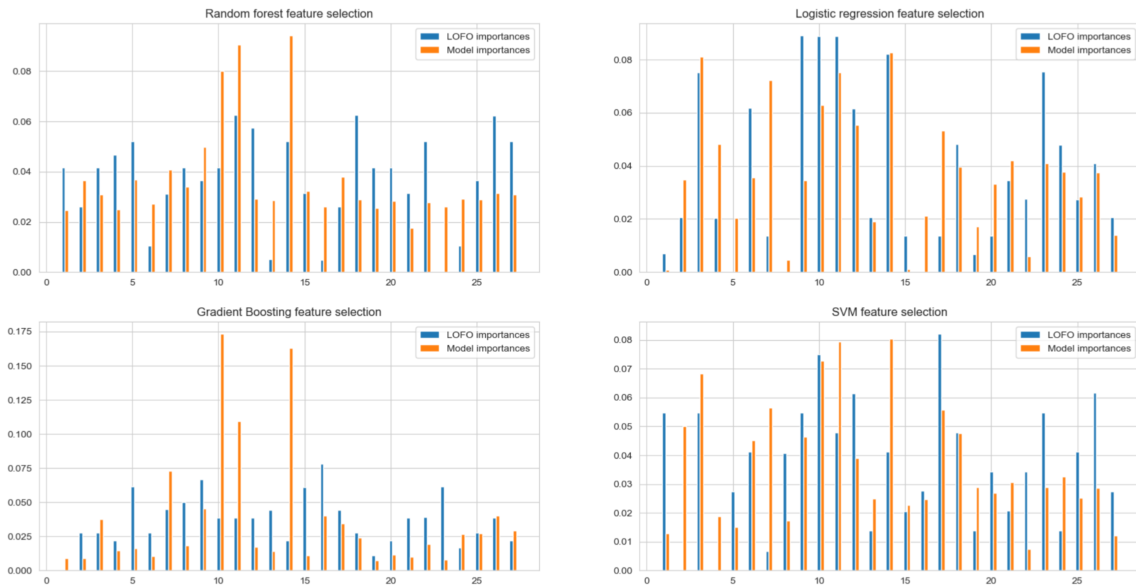

Figure 7: Comparison between the importance given to the features from the built-in methods of the models and the importance given by the Leave One Covariate Out (LOCO) method.

2.4 Model Independent Feature Selection without Classifiers

Psychometric inspired feature selection algorithms are now applied. In order to be able to compare in a correct way the performaces of each method, we follow this pipeline: as first we apply our techniques on the data and we look at the weights that are given to the original features. Once that we have available the information about which are the most relevant features for each method, we train our models on these subset of features, that are the ones "chosen" by the presented techniques.

1. **PCA**
   Principal component analysis (PCA) is a popular technique for analyzing datasets containing a high number of dimensions/features per observation and it can be useful for increasing the interpretability of data while preserving the maximum amount of information. Formally, PCA is a statistical technique for reducing the dimensionality

of a dataset, combining in a linear way the features and creating a new system of coordinates. In other words, each principal component is a linear combination of the original features. After the creation of the *principal components*, indeed, usually only the first few ones are considered and the rest is ignored. The new created components are basically new features with wich we can represent almost the same information contained in the original data that is been combined together, with the advantage of a less complexity.
In terms of features importance, we can easily compute the coefficients of the linear combinations in which the original features are involved: so we can understand the "weight" given to each variable inside each principal component. Knowing the most relevant new components, we can extract the information about the most relevant features.

2. **Kernel PCA**
Kernel PCA was developed in an effort to help with the classification of data whose decision boundaries are described by non-linear function. The idea is to project the problem into a higher dimension space in which the decision boundary becomes linear: in this way the usual PCA decomposition is again suitable. The function used to operate the projection is called *kernel* and there are of course different choices for it. Common ones are the Gaussian kernel or the polynomial kernel. Once we have the kernel, we follow the same procedure of the normal PCA.

3. **Sparse PCA**
As another variant, we try the Sparse Principal Component Analysis (sparse PCA) that is a specialised technique in particular in the analysis of multivariate data sets. It extends the classic method of PCA for the reduction of dimensionality of data by introducing sparsity structures to the input variables. Sparse PCA overcomes the issue given by the linear nature of the original variables combination by finding linear combinations that contain just a few input variables.

4. **Truncated Singular Values Decomposition**
It constitutes a tool of the linear algebra and it consists in a factorization of a data matrix into the form:
$$M = U\Sigma V^T$$
where U and V are unitary matrices and Σ is a diagonal one. The columns of U and the columns of V are called left-singular vectors and right-singular vectors of M, respectively and they form two sets of orthonormal bases. The diagonal entries are known as the singular values of M and they can be interpreted as the magnitude of

the semiaxes of an ellipsoid.
In many applications the number of the non-zero singular values is large and it may make the SVD impractical to compute. In such cases, the smallest singular values may need to be truncated to compute only fewer singular values. The truncated SVD is no longer an exact decomposition of the original matrix M, but rather provides the optimal low-rank matrix approximation for the original matrix. Truncated SVD is usually employed in latent semantic indexing.
The features importance are computed in the same way of the PCA.

5. **Factor Analysis**
   Factor Analysis is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables that are called factors. The original variables are considered as linear combinations of the potential latent factors.
   Factor analysis is commonly used in psychometrics, psychology, biology, and machine learning: it may help to deal with datasets where there are large numbers of observed variables that are thought to reflect a smaller number of underlying/latent variables. In our case we can interpret our latent factors as aspects of the personality that can be connected to some psicologic disorder.
   Particular of this tool is the possibility to use the loadings. They correspond to the correlations between the original features and the hidden dimensions and we can use them in order to understand the importance given from the factors to each original variables.

   In the plots presented in Figure 8, we show the different ways with which the methods assign different importance to the original features.

6. **Relief**
   Relief is an algorithm that takes a filter-method approach to feature selection that is very sensitive to feature interactions. Relief calculates a feature score that can be applied to rank and select top scoring features for feature selection.
   Relief feature scoring is based on the identification of feature value differences between nearest neighbor instance pairs. If a feature value difference is observed in a neighboring instance pair with the same class, the feature score decreases. Alternatively, if a feature value difference is observed in a neighboring instance pair with different class values, the feature score increases. What we obtain is shown in the Figure 9 .

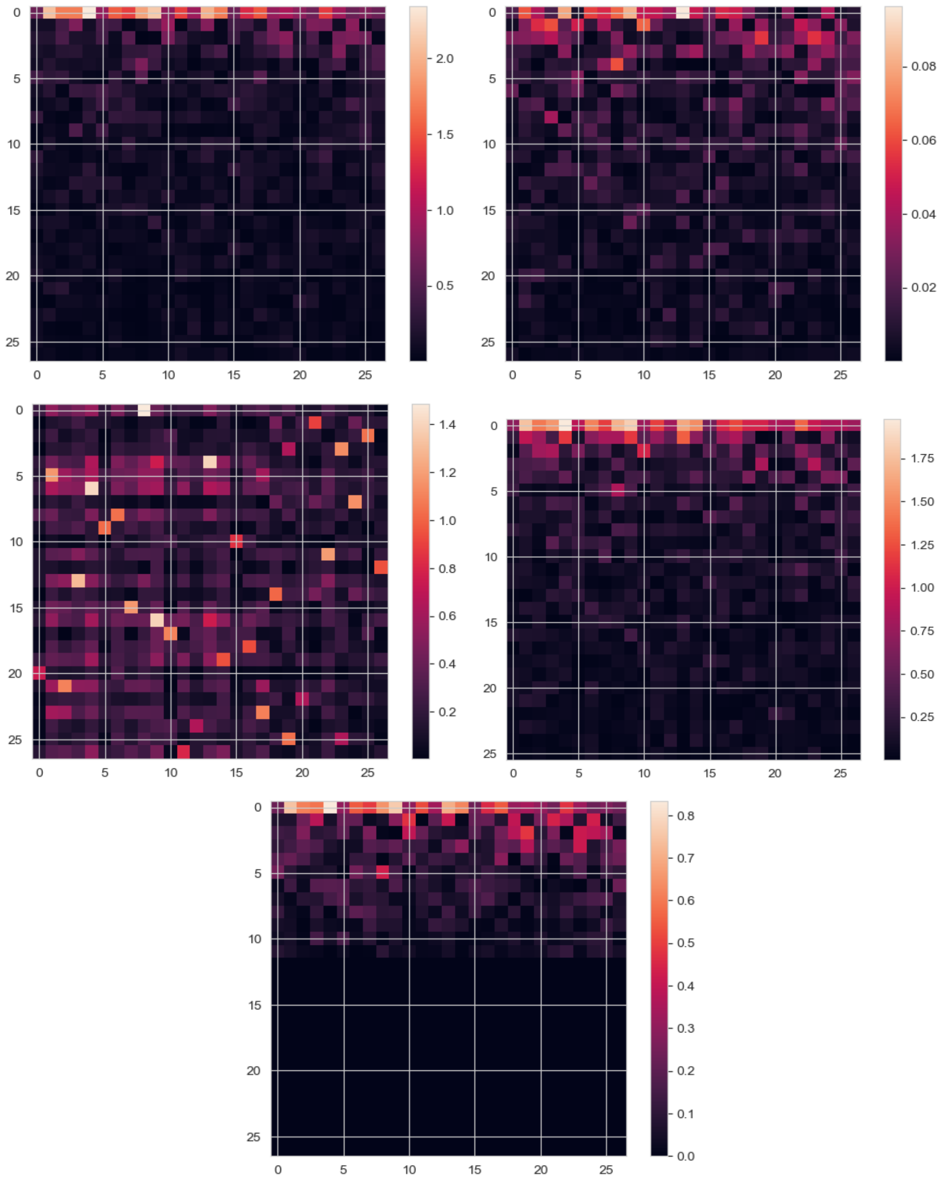

Figure 8: top-left: pca loadings, top-right: kernel-pca loadings, centre-left: sparse-pca loadings, centre-right: truncated-svd loadings, bottom: factor-analysis loadings.

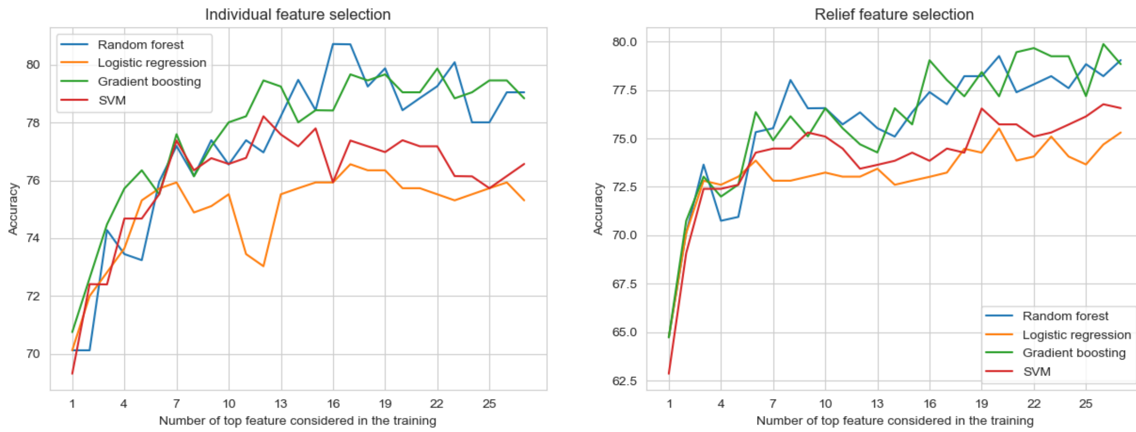

Figure 9: comparison between the importance given using simply the models and the importance given to the features introducing the Relief method.

3 Metrics

The previously described feature selection methods were compared using two different metrics: *accuracy*, to assess the quality of the predictions, and *robustness*, to assess the stability of the identified relevant features.
In particular our aim would be finding a method that works well in terms of both the metrics, but it's not really possible. So we will consider a "weighted trade-off" between this two indices where we choose to give more importance to the robustness part, in order to follow correctly the aim of the proposed project.

1. **Accuracy**
   Accuracy is a widely used metric to assess the quality of a prediction. It is calculated by dividing the number of correct predictions by the total number of predictions.

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + FP)}$$

   TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative.

2. **Robustness**
   To assess the stability of the identified relevant features, the importance of the selected features computed by using different models were compared.
   First, the selected features were sorted by order of importance for each model. Next,

the orders assigned by each model were compared one to each other. For each comparison, the difference between the two orders was calculated and the robustness was calculated as the average of these differences. Therefore, the lower the better.

4 Conclusions

Which techniques are the best in terms of accuracy and robustness index on the 20%-reduced datasets? Is there a technique that works better for all datasets?

On the following table are reported averages of the obtained results on all datasets (the complete table is at the end of the report 4.2). The averages are done separately on all complete datasets ("correlated") and on all datasets in which correlated features were removed ("uncorrelated"). The robustness index is denoted as "ROB." and the accuracy score as "ACC.". Green values are the best ones and red values are the worst ones.

Average of results on all datasets								
	P I	LOFO	PCA	K PCA	S PCA	T SVD	F A	Relief
correlated								
ROB.	2.08	2.02	2.09	1.86	1.58	1.61	1.26	1.23
ACC.	82.95	83.90	81.77	81.87	81.75	82.43	81.16	82.43
uncorrelated								
ROB.	1.58	1.79	1.17	1.24	1.35	1.51	1.88	1.26
ACC.	73.66	75.95	72.01	72.99	72.55	71.35	71.88	72.87

It can be seen that, on both "correlated" and "uncorrelated" datasets, model agnostic techniques Permutation Importance (PI) and Leave One Feature Out (LOFO) give the best performances in terms of accuracy. On the other hand, if we work on a "correlated" dataset the best robustness index is given by Factor Analysis (FA) and Relief, and if the dataset is "uncorrelated" then, for the same robustness purpose, is best to use PCA and Kernel PCA. Summing up, model agnostic techniques with classifiers give the best accuracy, and psychometric inspired features selection techniques are best in terms of features importance assignments.

Looking at the same table, keeping in mind that we are looking for a tradeoff between robustness and accuracy that give some more importance to the first one, we can state that the best reliable predictor for a "correlated" dataset is given by classifiers trained on top

20% of features ordered by importance by Relief algorithm, followed by Factor Analysis. If we are considering a "uncorrelated" dataset, instead, the best reliable predictor is given by classifiers trained on top 20% of features ordered by importance by Kernel PCA, followed by Relief algorithm.

4.1 Interpretation of important features

Do the reason why participants lie affects the selection of most important features?

The results obtained with the dataset (1b) under investigation were compared with the results obtained on the same survey but where people lie for a job interview (1a), to assess whether the reason for lying causes selected features to be different. Indeed the answer is yes, different features are selected.

In the dataset (1b) features selected reflected aspects psychologically related to parenting and child custody.
All methods select question 13, suggesting that ambition may be an important element in child custody consideration. This is probably also because ambition is generally associated with greater economic security that protects children more.
Furthermore, question 9 is selected by most of the models, suggesting that people who have to lie for child custody tend to be more likely to declare themselves as good leaders. Being a good leader is associated with the ability to lead and motivate people, as well as the ability to listen to people. All of which are fundamental aspects of raising children.
Question 10 is also selected by most of the models. Therefore, people who battle for custody of their children often lied about wanting to be the centre of attention. Probably being less likely to be the centre of attention fosters a listening skill that is fundamental to being a good parent.

In the other dataset (1a) features selected reflected aspects need to be good workers.
Most models select questions 18, 24, 23. This seems to be in line with the motive of lying. Indeed, respect towards the authorities (question 18) is a fundamental aspect in the employment context. On the other hand, question 24 investigates the criminal record, another fundamental element in the context of employment. Finally, question 23 investigates tolerance and respect for others, necessary for working in a team.

From this comparison it is therefore possible to state that the models used do indeed select the most critical questions regarding the purpose of lying and are therefore effective methods.

4.2 Results on all datasets

	P I	LOFO	PCA	K PCA	S PCA	T SVD	F A	Relief
a. (short) Dark Triad								
correlated								
ROB.	4.58	4.58	3.07	4.31	3.47	5.33	3.33	1.42
ACC.	64.96	74.47	58.68	60.3	59.26	59.72	58.82	64.41
uncorrelated								
ROB.	4.58	4.58	3.07	4.31	3.16	4.89	4.31	3.24
ACC.	64.96	75.07	58.68	60.3	58.39	60.14	59.58	63.52
b. (short) Dark Triad								
correlated								
ROB.	3.51	3.6	3.02	1.47	2.93	3.38	2.27	0.93
ACC.	74.72	74.47	71.89	72.72	69.96	69.76	71.68	72.3
uncorrelated								
ROB.	2.44	3.6	2.98	1.91	3.64	2.8	3.64	2.36
ACC.	73.53	76.64	70.91	70.39	69.61	70.59	69.61	73.66
Prospective and Retrospective Memory								
correlated								
ROB.	0	0.93	0.93	0.93	0.93	0.4	0.4	0
ACC.	88.66	88.56	88.78	88	87.95	87.53	87.55	88.39
uncorrelated								
ROB.	2.44	3.6	2.98	1.91	3.64	2.8	3.64	2.36
ACC.	68.42	69.03	62.39	70.12	62.39	70.43	62.39	64.26
Post Traumatic Stress Disorder								
correlated								
ROB.	2.18	2.27	3.02	1.33	1.82	2.04	1.07	1.11
ACC.	82.23	82.06	79.4	78.41	78.55	78.67	80.23	79.61
uncorrelated								
ROB.	4.58	4.58	3.07	4.31	3.16	4.89	4.31	3.24
ACC.	61.72	70.99	54.66	60.04	59.2	60.77	54.66	56.91
Negative Acts - Revised								
correlated								
ROB.	3.47	3.24	1.82	2.49	1.82	2.04	2	3.56
ACC.	88.52	89.43	85.63	82.36	85.63	88.95	81.85	87.97
uncorrelated								
All features are correlated, we can't remove all of them.								

	P I	LOFO	PCA	K PCA	S PCA	T SVD	F A	Relief
Depression and Generalized Anxiety Disorder								
correlated								
ROB.	0.4	0.4	0	0	0	0	0	0
ACC.	97.7	97.7	97.79	97.83	97.79	97.83	92.73	97.83
uncorrelated								
All features are correlated, we can’t remove all of them.								
(short) Personality Disorders								
correlated								
ROB.	6.8	6.76	4.49	6.98	4.36	2.71	2.13	3.64
ACC.	95.36	95.18	94.99	95.42	94.6	94.24	94.22	95.28
uncorrelated								
ROB.	0	0	0	0	0	0	0	0
ACC.	86.99	86.99	86.78	86.78	86.78	86.78	86.78	86.78
Parental Reflective Functioning								
correlated								
ROB.	3.47	3.24	1.82	2.49	1.82	2.04	2	3.56
ACC.	89.04	90.42	89.24	89.32	86.44	89.13	89.32	88.81
uncorrelated								
ROB.	2.09	2.09	0.4	0.4	0.4	0.4	2.89	1.16
ACC.	78.97	79	77.78	77.66	77.77	77.84	76.77	79.24
Impact of Event Scale - Revised								
correlated								
ROB.	3.24	1.42	2.36	2.58	2.36	1.51	1.11	1.82
ACC.	91.8	92.56	91.4	91.19	91.4	92.53	89.56	90.7
uncorrelated								
All features are correlated, we can’t remove all of them.								
Dirty Dozen Dark Triad								
correlated								
ROB.	0.71	0.71	0	1.07	0	0.71	1.07	0.71
ACC.	73.07	73.07	69.38	66.99	69.38	72.04	66.99	68.94
uncorrelated								
ROB.	0.4	0.4	0	0.4	0	0.4	1.07	0.71
ACC.	71.36	71.36	71.56	71.83	71.56	71.83	68.59	68.94

	P I	LOFO	PCA	K PCA	S PCA	T SVD	F A	Relief
International Adjustment Disorder								
correlated								
ROB.	0.4	0.71	1.07	0.93	1.07	0.93	0.71	0
ACC.	78.49	80.78	82.78	82.78	82.78	82.78	82.44	81
uncorrelated								
ROB.	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4
ACC.	76.84	76.84	79.78	79.78	79.78	79.78	79.78	79.78
a. Big Five Inventory								
correlated								
ROB.	0.4	0	1.11	1.11	1.11	1.11	1.11	0.4
ACC.	80.44	79.19	80.1	80.1	80.1	80.1	80.1	80.49
uncorrelated								
ROB.	0.4	0.4	0	0	0.4	0	0.4	0.4
ACC.	77.65	79.08	78	78	80.49	73.2	80.49	80.49
b. Big Five Inventory								
correlated								
ROB.	0	0	0	0	0	0	0	0
ACC.	78.91	78.91	79.95	79.95	79.95	79.95	79.95	79.95
uncorrelated								
ROB.	0	0	0	0	0	0	0	0
ACC.	74.59	74.59	74.84	71.25	75.33	71.25	75.33	71.25
c. Big Five Inventory								
correlated								
ROB.	0	0.4	0.4	0.4	0.4	0.4	0.4	0
ACC.	77.41	77.77	80.76	80.76	80.76	80.76	80.76	78.4
uncorrelated								
ROB.	0	0	0	0	0	0	0	0
ACC.	75.2	75.82	76.75	76.75	76.75	62.28	76.75	76.75

References

[1] Jones, Daniel N., & Delroy L. Paulhus. *Introducing the short dark triad (SD3) a brief measure of dark personality traits*. Assessment 21.1 (2014): 28-41.

[2] Crawford, J., Smith, G., Maylor, E., Della Sala, S., & Logie, R. (2003). *The Prospective and Retrospective Memory Questionnaire (PRMQ): Normative data and latent structure in a large non-clinical sample*. Memory, 11(3), 261-275.

[3] Blevins, C. A., Weathers, F. W., Davis, M. T., Witte, T. K., & Domino, J. L. (2015). *The posttraumatic stress disorder checklist for DSM-5 (PCL-5): Development and initial psychometric evaluation*. Journal of traumatic stress, 28(6), 489-498.

[4] Notelaers, Guy, & Ståle Einarsen. *The world turns at 33 and 45: Defining simple cutoff scores for the Negative Acts Questionnaire–Revised in a representative sample*. European Journal of Work and Organizational Psychology 22.6 (2013): 670-682.

[5] Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). *A brief measure for assessing generalized anxiety disorder: the GAD-7*. Archives of internal medicine, 166(10), 1092-1097.

[6] Fossati, A., Krueger, R. F., Markon, K. E., Borroni, S., & Maffei, C. (2013). *Reliability and validity of the Personality Inventory for DSM-5 (PID-5) predicting DSM-IV personality disorders and psychopathy in community-dwelling Italian adults*. Assessment, 20(6), 689-708.

[7] Fossati, A., Krueger, R. F., Markon, K. E., Borroni, S., & Maffei, C. (2013). *Reliability and validity of the Personality Inventory for DSM-5 (PID-5) predicting DSM-IV personality disorders and psychopathy in community-dwelling Italian adults*. Assessment, 20(6), 689-708.

[8] Luyten, P., Mayes, L. C., Nijssens, L., & Fonagy, P. (2017). *The parental reflective functioning questionnaire: Development and preliminary validation*. PloS one, 12(5), e0176218.

[9] Weiss, D. S. (2007). *The impact of event scale: revised. In Cross-cultural assessment of psychological trauma and PTSD* (pp. 219-238). Springer, Boston, MA.

[10] Costa Jr, P. T., & McCrae, R. R. (1997). *Stability and change in personality assessment: the revised NEO Personality Inventory in the year 2000*. Journal of personality assessment, 68(1), 86-94.

[11] Jonason, P. K., & Webster, G. D. (2010). *The dirty dozen: a concise measure of the dark triad*. Psychological assessment, 22(2), 420.

[12] Shevlin, M., Hyland, P., Ben-Ezra, M., Karatzias, T., Cloitre, M., Vallières, F., & Maercker, A. (2020). *Measuring ICD-11 adjustment disorder: The development and initial validation of the International Adjustment Disorder Questionnaire*. Acta Psychiatrica Scandinavica, 141(3), 265-274.

[13] Guido, G., Peluso, A. M., Capestro, M., & Miglietta, M. (2015). *An Italian version of the 10-item Big Five Inventory: An application to hedonic and utilitarian shopping values*. Personality and Individual Differences, 76, 135-140.