International Journal of Mental Health and Addiction
https://doi.org/10.1007/s11469-024-01332-x

ORIGINAL ARTICLE

Shortening and Personalizing Psychodiagnostic Assessments with Decision Tree-Machine Learning Classifiers: An Application Example Based on the Patient Health Questionnaire-9

Daiana Colledani[1,2] · Egidio Robusto[1] · Pasquale Anselmi[1]

Accepted: 13 May 2024

Abstract

The development of psychological assessment tools that accurately and efficiently classify individuals as having or not a specific diagnosis is a major challenge for test developers and mental health professionals. This paper shows how machine learning (ML) provides a valuable framework to improve the accuracy and efficiency of psychodiagnostic classifications. The method is illustrated using an empirical example based on the Patient Health Questionnaire-9 (PHQ-9). The results show that, compared to traditional scorings of the PHQ-9, that based on decision tree (DT) algorithms is more advantageous in terms of accuracy and efficiency. In addition, the DT-based method facilitates the development of short test forms and improves the diagnostic performance of the test by integrating external information (e.g., demographic variables) into the scoring process. These findings suggest that DT-algorithms and ML applications such as feature selection represent a valuable method for supporting test developers and mental health professionals, and highlight the potential of ML for advancing the field of psychological assessment.

Keywords Machine learning classifiers · Psychological assessment · Patient Health Questionnaire-9 · Receiver operating characteristic curve · Decision tree

Preregistration statement The present study was not preregistered.

✉ Pasquale Anselmi
pasquale.anselmi@unipd.it

Daiana Colledani
daiana.colledani@uniroma1.it

Egidio Robusto
egidio.robusto@unipd.it

[1] Department of Philosophy, Sociology, Education and Applied Psychology, University of Padua, Via Venezia 14, Padua 35131, Italy

[2] Department of Psychology, Faculty of Medicine and Psychology, Sapienza University of Rome, Via Dei Marsi 78, 00185 Rome, Italy

Published online: 23 May 2024

Developing psychological assessment tools that accurately classify individuals as having or not having a specific diagnosis is one of the main challenges for test developers and mental health professionals. To obtain accurate assessments, the administration of a large number of items is often required. In these situations, respondents may feel annoyed and provide inaccurate answers with a negative impact on the quality of the assessment. Thus, developing assessment tools that consist only of the items that are most effective in classifying individuals is particularly useful. Taking into account the specificities of respondents may also contribute to classification accuracy. This work aims to show how machine learning (ML) can provide a valuable framework for developing psychodiagnostic tools that address all the aforementioned requirements.

ML has proved to be a precious resource in numerous fields, including healthcare, marketing, finance, and engineering (Dixon et al., 2020; Dzyabura & Yoganarasimhan, 2018; Reich, 1997; Sarker, 2021). Psychiatry and psychology, for instance, have recently started to use ML classifiers to improve and facilitate the diagnostic process (see, e.g., Dwyer et al., 2018; Paulus & Thompson, 2021; Yarkoni & Westfall, 2017) and to refine psychological assessments and psychometric instruments (Colledani et al., 2023; Gonzalez, 2021a, 2021b; Li, 2023; Stewart et al., 2016). In this regard, some studies have shown that decision tree (DT) algorithms can be used to develop effective testing procedures with good classification and diagnostic performance, and with an innovative approach that provides appealing clinical insight. In particular, they allow scoring procedures that emphasize the importance of each individual item (i.e., symptom) rather than relying solely on total test scores. It is worth noting that DT algorithms can be applied to tests that classify respondents according to diagnostic criteria, as well as to tests that estimate the respondent's trait levels or the severity of a disease (Gibbons et al., 2013; Yan et al., 2004). Moreover, DTs have been found to be an extremely valuable resource for the development of both adaptive (i.e., computerized adaptive testing, CAT) and static short versions of tests (Gonzalez, 2021b; Li, 2023; Stewart et al., 2016). In this regard, some studies have compared the performance of DT-based CAT with that of traditional CAT based on item response theory (IRT), the latter being a well-known framework that has been widely and successfully used in clinical and educational settings to develop adaptive tests (De Beurs et al., 2014; Eggen & Straetmans, 2000; Fliege et al., 2005; Moore et al., 2018; Tseng, 2016; Van der Linden & Glas, 2000). In general, IRT-based CAT has been shown to be effective in estimating the trait levels of examinees while ensuring considerable assessment efficiency (Brown & Weiss, 1977; Gibbons et al., 2008). However, recent research has shown that DT-based CAT can be even more efficient and accurate in classifying respondents than IRT-based CAT, and this makes it particularly valuable in diagnostic assessment (Delgado-Gomez et al., 2016, 2019; Lopez-Castroman et al., 2016; Michel et al., 2018; Ueno & Songmuang, 2010; Zheng et al., 2020). Furthermore, the DT approach offers several significant advantages over IRT. In fact, unlike DT methods, IRT relies on fundamental assumptions such as unidimensionality, local independence, and monotonicity, which may not hold for all real-world datasets. In this regard, some authors have argued that DT-based CAT may outperform IRT-based CAT, especially when the assumptions of IRT are not met (Gonzalez, 2021a; Michel et al., 2018; Riley et al., 2011; Ueno & Songmuang, 2010).

This work aims to demonstrate how the use of DT classifiers and ML-based methods in common psychological tests can contribute to the development of shorter, yet highly accurate assessment instruments. This work provides a novel contribution by illustrating how the flexibility of the method allows individual differences to be easily incorporated into the assessment process, resulting in an assessment that is personalized (and therefore more efficient and accurate) to a degree that would be difficult to achieve with traditional methods. This contribution highlights a relevant aspect in psychodiagnostic testing, which has not been previously explored, even if it can greatly improve the assessment process.

The next section describes the DT approach to testing and its advantageous features. The results of applying the method to real psychodiagnostic data are then presented. In particular, it is shown how ML-DT classifiers can be used to develop scoring algorithms that take into account not only the information derived from item responses, but also the information derived from individual variables (i.e., gender and age), thereby enabling highly personalized, accurate and efficient scoring procedures. It is also shown how the combined use of DT and feature selection can increase the efficiency of the assessment without compromising its accuracy. Implications for theory and practice and suggestions for future research directions conclude the argumentation.

Decision Trees in Psychological Testing

ML classifiers are algorithms that aim to predict the class (i.e., a discrete output variable) of a specific data point starting from a set of input variables (predictors). In the diagnostic field, they can be used to learn a classification function capable of predicting the clinical status of an individual (diagnosis vs. non-diagnosis) based on a set of relevant observed attributes, such as the results of specific clinical tests or other individual characteristics (e.g., age, gender, ethnicity). Typically, ML algorithms develop the classification function on one part of the dataset (the training dataset) and evaluate its predictive performance on a different part of the dataset (the test dataset; Mahesh, 2020; Yarkoni & Westfall, 2017). The predictors and the class variable are present in both datasets.

Although several ML classifier algorithms are available (e.g., *decision tree*, *random forest*, *support vector machine*, and *naive Bayes*), decision trees (DTs) are probably the most widely used in the clinical field. They are appreciated for their accuracy in the classification task and because, compared to other methods, they produce rules that are more understandable and capable of clearly indicating the role and importance of each variable in the prediction (Higa, 2018; Zhao & Zhang, 2008).

DT algorithms employ a top-down approach to create a set of "if-else" rules by using a recursive "divide-and-conquer" process (Breiman et al., 1984). In the first step, the algorithms select the attribute (i.e., the test item) to be placed at the root of the tree, which is then divided into branches. The root node includes the entire dataset and serves as the starting point for the classification algorithm. Typically, the attribute selected for the root is the one that leads to the purest node, that is, the attribute that divides the sample into subsets consisting of instances (i.e., individuals being evaluated) with the same classification (Witten et al., 2017). The selection process for the root node and subsequent nodes is based on the information gain, which measures the amount of information a variable provides relative to the class variable being predicted (Criminisi et al., 2012; Gupta et al., 2017).

After having defined the root node, DT algorithms develop the rules for growing the branches. A branch is a chain of nodes from the root to a leaf (i.e., the end of a branch, where the classification ends), with nodes being specific attribute variables. At each node, the instances are divided into branches (e.g., subsamples of individuals) based on the values assumed by the variable that constitutes the node itself (e.g., the score to the item that constitutes a node). This process continues recursively, with instances progressively distributed across branches according to rules aimed at generating subsets containing instances with the same classification relative to the class variable. For nominal attributes, the number of branches is equal to the number of possible attribute values. For numerical

variables, the algorithm searches for the value of the attribute (e.g., score ≤ 1 or > 1) that allows the instances to be divided into subsets that are as similar as possible relative to the class variable (Witten et al., 2017). Branch development proceeds recursively through the "divide-and-conquer" process until all instances have the same classification or further splitting does not improve classification accuracy.

In psychodiagnostic testing based on DTs, a branch from the root to a leaf can be conceptualized as a sequence of items, which develops based on particular responses to them and allows for classifying individuals as having or not having a diagnosis. DTs constitute a useful means to create accurate scoring algorithms with an appealing diagnostic value (Colledani et al., 2023). In addition, DT-based assessments are usually more efficient than traditional ones. In fact, by following the sequence of items (nodes) suggested by the DT structure, it is possible to design adaptive testing procedures (DT-based CAT) capable of classifying individuals based on their responses to a reduced set of items. This is achieved by presenting individuals with only those items that belong to the branch in which they ultimately fall based on the responses they progressively provide (Delgado-Gomez et al., 2016, 2019; Lopez-Castroman et al., 2016; Michel et al., 2018; Ueno & Songmuang, 2010; Zheng et al., 2020).

A further interesting aspect of DT algorithms is the possibility of simultaneously and profitably handling variables of different nature. This feature is very useful in the field of testing where, in order to improve the assessment, it is often advisable to control for the effect of individual variables external to the test, such as gender, age, or specific psychiatric conditions (e.g., Colledani, 2018; Colledani et al., 2018, 2019, 2022; Hamilton, 1999; Vandenberg & Lance, 2000). ML classifiers allow for including these variables in the algorithm very easily. In particular, the variables are entered into the chain of nodes of specific branches if they contribute to increasing classification accuracy. For instance, the variable gender could be inserted at a specific position in a branch formed by a chain of items, leading to the development of different pathways to accurately classify males and females (i.e., the branch splits into two branches that suggest considering different items for males and females).

ML classifiers allow *feature selection*, an application that selects the most useful subset of items to be used for classification. In general, DT algorithms aim to choose the most promising attributes to place at each node. Although these algorithms are expected not to select useless attributes, such attributes can sometimes be inserted into the tree after the best ones have already been inserted. Intuitively, one could expect that having more attributes should always lead to a higher discriminating power. However, it has been shown that, in some cases, the presence of irrelevant or non-informative attributes in the dataset can be confusing for instance-based ML algorithms, such as DTs (Witten et al., 2017). By eliminating useless attributes (or items), feature selection helps to improve the performance of the learning algorithm (Karabulut et al., 2012) and, consequently, accuracy and efficiency. Furthermore, it leads to smaller and more easily interpretable trees (Novaković et al., 2011). Relevant attributes can be manually selected before the learning process starts. This feature selection procedure is called the *filter method* and can be very effective (Witten et al., 2017). However, there are also automatic and highly efficient methods (Kohavi & John, 1997), which are called *wrapper methods* because the selection process is embedded in the learning algorithm (Caruana & Freitag, 1994; John et al., 1994; Langley, & Sage, 1994). Wrapper methods systematically analyze all combinations (all pairs, triplets, etc.) of attributes in order to isolate the particular combination that maximizes prediction accuracy. Wrapper methods examine the effect of omitting or including different attributes in relation to the accuracy of the predictions that can be obtained with a specific algorithm.

If an attribute makes a significant difference to prediction accuracy, it is considered a high-quality attribute and is selected for inclusion (Witten et al., 2017).

In the field of psychological testing, feature selection can be used to create shorter versions of tests. Indeed, this approach can facilitate the identification of the subset of items that is most informative for classification purposes (as in DTs) or for predicting the scores that would be obtained by responding to the full set of items (as in regression or model trees). Abbreviated tests created using ML algorithms and feature selection procedures have been shown to be extremely effective at approximating the scores obtained on their full-length versions (Gonzalez, 2021b). For diagnostic purposes, however, accurate classifications are probably more critical than accurate estimates of test scores. DTs have been shown to be a promising approach for making accurate and efficient psychodiagnostic classifications (Colledani et al., 2023; Gibbons et al., 2013). Therefore, it is worth exploring whether DTs combined with feature selection procedures can allow the development of shortened tests that are accurate in respondent classification (rather than in score estimation) while achieving even greater efficiency. This aspect has not been explored in previous research and is considered in the present work.

Method

Participants

Data were obtained from the public repository FigShare (Doi et al., 2018; Ito et al., 2015). The dataset includes responses from 2,830 Japanese individuals (mean age = 42.44, *SD* = 10.39; 1,283 males) and is part of a larger study on psychopathology and emotion in the Japanese population. Participants were all at least 18 years of age and were recruited via a web-based survey based on a large panel of more than 1 million individuals, which included 389,265 individuals identified as "disease panelists" by annual self-report of current or past diagnosis of a disease. The sample of 2,830 individuals was randomly selected from the larger panel to ensure representation of different ages, genders, and geographical locations (this was done for both disease and non-disease panelists). Participants completed measures of mental health, including the Japanese version (Muramatsu et al., 2007) of the Patient Health Questionnaire-9 (PHQ-9; Kroenke et al., 2001). Furthermore, they provided some demographic information (e.g., age, gender, area of residence) and indicated whether they were currently or had previously been diagnosed and treated for any psychiatric disorders (i.e., Major Depressive Disorder – MDD; Social Anxiety Disorder – SAD; Panic Disorder – PD; Obsessive–Compulsive Disorder – OCD; and/or other disorders).

The dataset analyzed in this work is a subsample of 2,205 individuals (mean age = 42.65, *SD* = 10.51; 1,023 males) who were selected according to a single eligibility criterion, namely reporting being currently diagnosed with and treated for MDD in a medical setting or reporting not being currently diagnosed with and treated for psychiatric disorders in a medical setting. In the following, the former will be referred to as MDD individuals, the latter as nonclinical individuals. The sample included 1,042 MDD individuals (MDD $N = 406$; MDD-SAD $N = 95$; MDD-PD $N = 127$; MDD-OCD $N = 100$; MDD-PD-OCD $N = 52$; MDD-PD-SAD $N = 51$; MDD-SAD-OCD $N = 55$; MDD-PD-SAD-OCD $N = 156$) and 1,163 nonclinical individuals. The 625 people who reported being diagnosed with or treated for psychiatric disorders other than depression were not included in the analyzed sample to reduce noise (Uğuz, 2011).

Measures

The PHQ-9 (Kroenke et al., 2001) is one of the most used instruments for screening and diagnosing MDD (Costantini et al., 2021). It comprises nine items that cover the nine diagnostic criteria for major depression in the DSM-IV and asks respondents to evaluate the frequency with which they experienced the described symptoms over the last two weeks (4-point scale from 0 "not at all" to 3 "nearly every day"). The instrument is available in several languages and has been adapted to many different contexts (Gilbody et al., 2007). The literature indicates that it has good diagnostic accuracy, validity, and reliability (Costantini et al., 2021). A score ≥ 10 is usually considered as the optimal cut-off score (Kroenke et al., 2001; Manea et al., 2012; Spitzer et al., 1999) for MDD detection, with sensitivity from 0.37 to 0.98 and specificity from 0.42 to 0.99 (Costantini et al., 2021). For the Japanese version of the instrument, which is the one used in the present study, good validity and reliability were documented, as well as invariance across clinical and nonclinical populations (Doi et al., 2018). Cut-off scores ≥ 10 (sensitivity = 0.905, specificity = 0.766; Muramatsu et al., 2018) and ≥ 11 (sensitivity = 0.76, specificity = 0.81; Suzuki et al., 2015) have been reported to be optimal for diagnosing MDD in the Japanese population.

Analyses

Before running ML analyses, factor structure and gender and age invariance of the PHQ-9 were verified. Factor structure was tested through confirmatory factor analysis (CFA) using Mplus 7.4 (Muthén & Muthén, 2012), with the maximum likelihood means adjusted (MLM; Muthén & Muthén, 2012; see also Brown, 2006) estimator. Multiple-group analyses were run to test configural, metric, scalar, and strict invariance across gender and across three age groups (youngest participants: ≤ 37 years old; middle-aged participants: 38–49 years old; oldest participants: ≥ 50 years old), which were defined based on a tertile split (see Bianchi et al., 2022). To evaluate the goodness of fit of all models, several fit indices were inspected: χ^2, comparative fit index (CFI; Bentler, 1990), root mean square error of approximation (RMSEA; Browne & Cudeck, 1993), and standardized root mean square residual (SRMR; Bentler, 1995). A satisfactory fit is indicated by nonsignificant ($p \geq 0.05$) χ^2 values, CFI values greater than 0.90, and RMSEA and SRMR values smaller than 0.08. For testing the equivalence of nested models in measurement invariance, the tests of change in CFI, RMSEA, and SRMR (ΔCFI, ΔRMSEA, ΔSRMR) were considered. Invariance is supported by ΔCFI values $\leq |.01|$, paired with ΔRMSEA and ΔSRMR values $\leq |.015|$ (ΔSRMR values $\leq |.030|$ for metric invariance; Chen, 2007; Cheung & Rensvold, 2002).

ML analyses were implemented through the open-source software WEKA 3.8.5 (Waikato Environment for Knowledge Analysis, University of Waikato, New Zealand). The J48 classifier was used to build the DTs. This algorithm represents a Java extension of the better-known Quinlan C4.5 algorithm (Quinlan, 1993; Salzberg, 1994). Starting from a set of input data, the algorithm defines a DT that allows for classifying new data into the groups of a specific class variable. In the building phase, the algorithm develops the structure of the tree. The items that have to be placed at each node and the splitting rules for each of them (i.e., the scores that generate branches) are identified considering the information gain they provide (Lin, 2001; Prabhakar, et al., 2002; Sugumaran et al., 2007). In

the pruning phase, the nodes are progressively removed if their exclusion does not affect classification accuracy. A minimum leaf size of 10 was set. This prevents overfitting and allows for obtaining understandable DTs with high prediction accuracy and generalizability (Dekker et al., 2009; Song et al., 2011).

In this work, the J48 algorithm was used to build a DT that aims to classify individuals as having or not having an MDD diagnosis using the information provided by the responses to the nine items of the PHQ-9. After a first analysis in which only the nine items of the PHQ-9 were considered, a second analysis was run in which also two demographic variables (i.e., gender and age) were included. By incorporating them into the DT structure, chains of nodes (i.e., branches consisting of sequences of items) differentiated by gender and age are obtained. According to the literature, gender and age are associated with the onset and course of depression (Blazer et al., 1994; Brodaty et al., 2005; Patten et al., 2016). Consequently, incorporating them into the DT structure is expected to produce more accurate and efficient assessments. A third analysis was run in which feature selection based on a wrapper method was applied on the nine items of the PHQ-9. Wrapper methods are linked to a base classifier and aim to identify the subset of variables that maximizes discrimination between classes. In this work, the wrapper method was run using the J48 classifier and by adding one item at a time (forward selection). This method identifies the subset of items that maximizes classification accuracy among all possible subsets. A final analysis was run in which the wrapper method was applied to the nine items of the PHQ-9 plus the two demographic variables. Thus, a total of four models were run: a DT built on the nine items of the test (DT-I), a DT built on the nine items of the test plus two demographic variables (i.e., gender and age; DT-ID), a DT built on the nine items of the test to which feature selection was applied (i.e., DT-FS-I), and a DT built on the nine items of the test plus the two demographic variables to which feature selection was applied (i.e., DT-FS-ID).

A stratified tenfold cross-validation procedure was used to evaluate the performance of the four models. This procedure randomly divides the dataset into 10 subsets with equal size that are similar to the entire dataset in the class variable. The algorithm is tested on each of the 10 subsets and trained on the remaining 9. Finally, the learning algorithm is run a 11th and last time on the entire dataset to obtain the model that is printed out. The tenfold cross-validation procedure is strongly recommended to evaluate the algorithms because it has been found to provide generalizable results and accurate estimates (Cumming, 2008; Witten et al., 2017).

The performance of the four DTs (i.e., DT-I, DT-ID, DT-FS-I, and DT-FS-ID), as well as that of the two ROC-based cut-off scores for the Japanese population that are reported in the literature (i.e., ≥ 10 and ≥ 11; Muramatsu et al., 2018; Suzuki et al., 2015), was evaluated considering several indices and statistics. For all models, the following five classification assessment measures were computed and compared: accuracy (i.e., (true positive + true negative)/total cases), sensitivity (i.e., true positive/(true positive + false negative)), specificity (i.e., true negative/(true negative + false positive)), positive predictive value (i.e., true positive/(true positive + false positive)), and negative predictive value (i.e., true negative/(true negative + false negative)). Moreover, the accuracy of each model was compared with the no-information rate (NIR), which represents the proportion of the largest class (Hastie et al., 2009). A one-tailed binomial test was used to determine whether model predictions were more accurate than NIR (i.e., 0.527, which is the proportion of nonclinical individuals in the analyzed sample). Significant results ($p < 0.05$) indicate that model predictions are unlikely to be the result of chance. Cohen's Kappa coefficient (also known as Kappa score) was used to compare model predictions with actual classification

Table 1 Frequency (and Percentage) of Individuals Endorsing Each Item of the PHQ-9 and Mean Score (and *SD*) for the Individuals Diagnosed with and Treated for Major Depressive Disorder (MDD Individuals; $N = 1{,}042$), for the Individuals Diagnosed with and Treated for No Psychiatric Disorder (Nonclinical Individuals; $N = 1{,}163$), and for the Total Sample ($N = 2{,}205$)

Item	Item endorsement			Mean score		
	MDD individuals	Nonclinical individuals	Total sample	MDD individuals	Nonclinical individuals	Total sample
1	257 (24.66)	77 (6.62)	334 (15.15)	1.60 (1.02)	0.76 (0.89)	1.04 (1.04)
2	285 (27.35)	99 (8.51)	384 (17.42)	1.66 (1.01)	0.84 (0.92)	1.05 (1.05)
3	467 (44.82)	154 (13.24)	621 (28.16)	1.98 (1.08)	1.02 (1.03)	1.16 (1.16)
4	479 (45.97)	183 (15.74)	662 (30.02)	2.09 (.98)	1.22 (1.01)	1.09 (1.09)
5	348 (33.40)	121 (10.40)	469 (21.27)	1.71 (1.12)	0.86 (1.00)	1.14 (1.14)
6	385 (36.95)	156 (13.41)	541 (24.54)	1.79 (1.12)	0.90 (1.06)	1.17 (1.17)
7	243 (23.32)	54 (4.64)	297 (13.47)	1.39 (1.12)	0.53 (0.83)	1.07 (1.07)
8	164 (15.74)	35 (3.01)	199 (9.02)	1.07 (1.08)	0.37 (0.72)	0.97 (0.97)
9	227 (21.79)	61 (5.25)	288 (13.06)	1.24 (1.14)	0.46 (0.82)	1.05 (1.05)

Note. According to the DSM scoring algorithm (Kroenke et al., 2001), item endorsement is defined by a score ≥ 2. For each item, the frequency of respondents endorsing the item and the mean score were significantly larger in MDD individuals than in nonclinical individuals (for the frequency of endorsed items: z from 10.415 to 16.457, $ps < .001$; for the mean score: t from 17.58 to 21.285, $ps < .001$)

of instances. Kappa coefficient measures the agreement between predicted and actual classifications of a dataset, while correcting for agreement that occurs by chance (Witten et al., 2017). In ML, it can be used to evaluate how closely classifier categorizations matched the actual condition of instances. The higher the value of Kappa, the larger the agreement between the classifications and the better the performance of the classifier (values of 0 indicate that a classifier is useless since there is no agreement between predicted and actual classifications). According to the suggestions by Landis and Koch (1977), Kappa values between 0 and 0.20, 0.21 and 0.40, 0.41 and 0.60, 0.61 and 0.80, and over 0.81 indicate slight, fair, moderate, substantial, and almost perfect agreement, respectively. Finally, the McNemar test was used to compare the performance of the models. It is a χ^2 test (with 1 degree of freedom) that evaluates if two models have the same error rate. Significant results ($p < 0.05$) indicate that the two models perform differently. The McNemar test was used to compare DTs with and without demographic variables, and with and without feature selection, as well as to compare DT-I with the two ROC-based cut-off scores.

Finally, the percentage of correct classification was also calculated for the 625 individuals who reported being diagnosed with and treated for a psychiatric disorder other than MDD and who were initially excluded from the analyzed sample (to reduce noise). All of these 625 individuals had to be classified as not diagnosed because, although they reported some psychiatric disorder, it was not MDD.

Results

Descriptive statistics of the PHQ-9 items are reported in Table 1. For each item, the frequency of respondents endorsing the item and the mean score were significantly larger in MDD individuals than in nonclinical individuals (for the frequency of endorsed items: z

Table 2 Invariance of the PHQ-9 Across Gender and Age Groups

	χ^2	*df*	CFI	RMSEA	SRMR	ΔCFI	ΔRMSEA	ΔSRMR
CFA	368.55	27	0.970	0.079	0.027			
Gender invariance								
Configural model	381.306	50	0.971	0.078	0.027			
Metric model	409.263	58	0.969	0.074	0.030	0.002	0.004	-0.003
Scalar model	486.241	66	0.963	0.076	0.036	0.006	-0.002	-0.006
Strict model	506.288	75	0.962	0.072	0.037	0.001	0.004	-0.001
Age-group invariance								
Configural model	423.964	75	0.968	0.080	0.030			
Metric model	481.228	91	0.965	0.076	0.043	0.003	0.004	-0.013
Scalar model	563.787	107	0.959	0.076	0.046	0.006	0.000	-0.003
Strict model	693.217	125	0.949	0.079	0.053	0.010	-0.003	-0.007

Note. Three age groups (i.e., youngest participants: ≤ 37 years old; middle-aged participants: 38–49 years old; oldest participants: ≥ 50 years old). CFI = comparative fit index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual; ΔCFI = test of change in CFI; ΔRMSEA = test of change in RMSEA; ΔSRMR = test of change in SRMR; CFA = confirmatory factor analysis

from 10.415 to 16.457, *ps* < 0.001; for the mean score: *t* from 17.58 to 21.285, *ps* < 0.001). CFA analyses supported the one-factor structure of the PHQ-9 (but suggested correlating the residuals of Items 4 and 3, and Items 7 and 8) and its configural, metric, scalar, and strict invariance across gender and age groups (Table 2).

Results on the Nine Items of the PHQ-9

Based on the sum scores on the nine items of the test, the two cut-off scores ≥ 10 and ≥ 11 classified 978 and 912 individuals as diagnosed, respectively. Conversely, the individuals classified as diagnosed by DT-I were 1,026 (Fig. 1).[1] These individuals were not classified according to their sum scores but according to the information provided by the single items progressively indicated in the tree structure. The DT algorithm placed Item 3 ("Trouble falling or staying asleep, or sleeping too much") at the root node. It was identified as the item that best differentiated the individuals into the two levels of the class variable (i.e., the item minimizing the misclassification rate in each branch). A score ≤ 1 or > 1 to Item 3 gave rise to two distinct branches, which contained 1,198 and 1,007 individuals, respectively. The branch originating from score ≤ 1 had Item 9 as the second node, while the branch originating from score > 1 had Item 1 as the second node (Fig. 1).

Table 3 shows the performance of the PHQ-9 scored using the two cut-off scores reported in the literature (≥ 10 and ≥ 11) and DT-I. DT-I was slightly superior to the cut-off score ≥ 10 in accuracy, sensitivity, PPV, and NPV, while the two models had the same

[1] In addition to the tenfold cross-validation method, all DTs were also trained and tested using both nine-fold and 11-fold cross-validation methods. The results were very close to those of the tenfold cross-validation method, suggesting that there are no relevant differences based on the size of the training and test datasets.

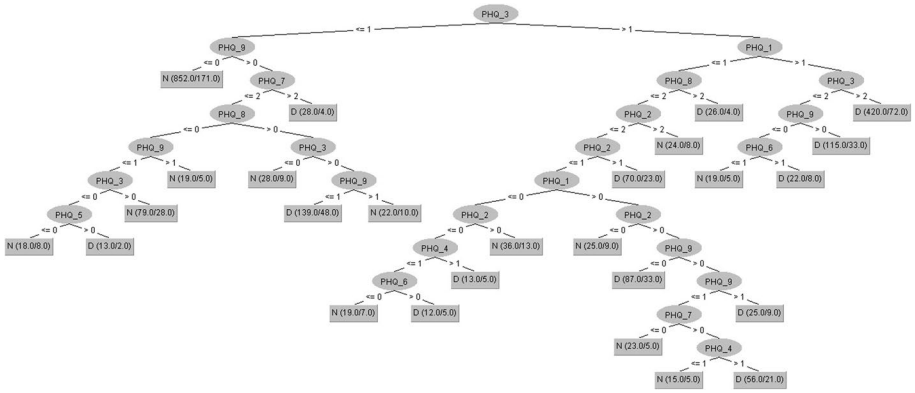

Fig. 1 Decision Tree Obtained Running the J48 Algorithm on the Nine Items of the PHQ-9 (DT-I) *Note*. The analyzed dataset included 2,205 individuals, 1,042 of whom diagnosed with and treated for major depressive disorder and 1,163 diagnosed with and treated for no psychiatric disorder. The rectangles at the end of each branch show how the DT classified the individuals falling in that branch (D = having major depressive disorder; N = not having the diagnosis), the number of individuals falling in that branch (first number in brackets), the number of incorrectly classified individuals (second number in the brackets)

Table 3 Performance of ROC-Based Cut-Off Scores and Decision Tree (DT) Algorithms

	Cut-off score ≥ 10	Cut-off score ≥ 11	DT-I	DT-ID	DT-FS-I	DT-FS-ID
Accuracy	.70	.70	.72	.73	.70	.73
Sensitivity	.69	.65	.70	.71	.72	.72
Specificity	.71	.74	.71	.75	.69	.75
PPV	.68	.70	.69	.72	.68	.72
NPV	.72	.71	.73	.74	.73	.75
Kappa	.41	.40	.41	.46	.41	.47

Note. DT-I = decision tree built on the nine items of the PHQ-9; DT-ID = decision tree built on the nine items plus two demographic variables (i.e., gender and age); DT-FS-I = decision tree built on the nine items to which feature selection was applied (Items 1, 2, 3, 7, 8, and 9 were retained); DT-FS-ID = decision tree built on the nine items plus the two demographic variables to which feature selection was applied (Items 3, 8, and 9, gender, and age were retained). PPV = positive predictive value; NPV = negative predictive value

specificity (Table 3). DT-I was superior to the cut-off score ≥ 11 in accuracy, sensitivity, and NPV, while it had lower specificity and PPV. The accuracy of the three models was significantly higher than the NIR (NIR = 0.53, $ps < 0.001$) and the Kappa coefficients were fair (from 0.40 to 0.41). The results of the McNemar tests showed that DT-I outperformed both the cut-off score ≥ 10 ($\chi^2 = 32.53$, $p < 0.001$) and the cut-off score ≥ 11 ($\chi^2 = 29.15$, $p < 0.001$).

Interestingly, DT-I was not only more accurate but also more efficient than the two ROC-based cut-off scores. In fact, to classify individuals, the cut-off scores required the administration of all nine items of the test whereas, following the DT-I scoring procedure in Fig. 1, individuals could be classified using fewer items. Specifically, if the PHQ-9 items were administered in a personalized way according to the process outlined by the DT structure, the respondents would be classified using the information deriving from 2 to 7 items (2.99 items on average).

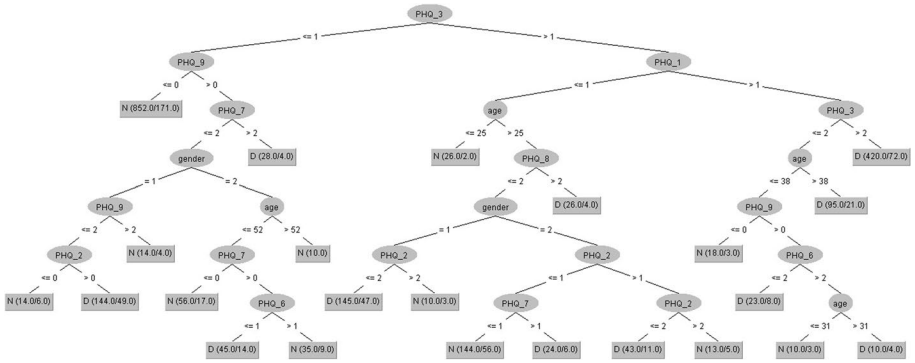

Fig. 2 Decision Tree Obtained Running the J48 Algorithm on the Nine Items of the PHQ-9 Plus the Two Demographic Variables Gender (1 = Male; 2 = Female) and Age (DT-ID) *Note*. The analyzed dataset included 2,205 individuals, 1,042 of whom diagnosed with and treated for major depressive disorder and 1,163 diagnosed with and treated for no psychiatric disorder. The rectangles at the end of each branch show how the DT classified the individuals falling in that branch (D = having major depressive disorder; N = not having the diagnosis), the number of individuals falling in that branch (first number in brackets), the number of incorrectly classified individuals (second number in the brackets)

In correctly classifying the 625 individuals who reported psychiatric disorders other than MDD as not diagnosed, DT-I was as accurate as the cut-off scores ≥ 10 and ≥ 11 (56.64%, 52.16%, and 57.76% of individuals correctly classified as not diagnosed by DT-I, cut-off score ≥ 10, and cut-off score ≥ 11, respectively; for the comparison between DT-I and cut-off score ≥ 10: $z = 1.59$, $p = 0.11$; for the comparison between DT-I and cut-off score ≥ 11: $z = -0.40$, $p = 0.69$), although it was more efficient than the cut-off score methods. For the results of the analysis conducted on the full sample of $N = 2{,}830$ individuals, see Table S1 in the Supplementary Materials.

Results on the Nine Items of the PHQ-9 Plus the Two Demographic Variables

When the two demographic variables (i.e., gender and age) were included in the analysis (DT-ID), a tree structure was obtained that embedded them into the chain of nodes, along with the PHQ-9 items. The resulting tree is displayed in Fig. 2. Following the scoring algorithm defined by the tree structure, the age of individuals with a score > 1 to Item 3 ("Trouble falling or staying asleep, or sleeping too much") and a score ≤ 1 to Item 1 ("Little interest or pleasure in doing things") is considered. For individuals aged 25 or younger, the assessment can end without considering additional items. For individuals older than 25 years, the algorithm suggests considering the score to Item 8. In other words, by considering the variable age, the algorithm defined two different sequences of items for younger or older participants. Overall, the variable age appeared four times in the tree and delineated different scoring algorithms for groups of individuals older or younger than 25, 31, 38, and 52 years. The variable gender, instead, appeared two times in the tree and defined specific sequences of items for classifying males and females.

The accuracy of DT-ID was significantly higher than the NIR (NIR = 0.53, $p < 0.001$) and the Kappa coefficient was moderate (0.46). DT-ID outperformed DT-I in the five classification assessment measures (i.e., accuracy, sensitivity, specificity, PPV, and

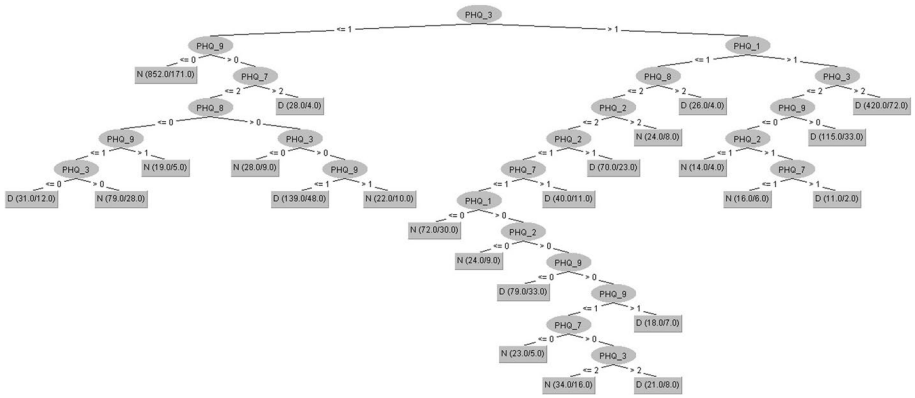

Fig. 3 Decision Tree Obtained Running the J48 Algorithm with Feature Selection on the Nine Items of the PHQ-9 (DT-FS-I; Items 1, 2, 3, 7, 8, and 9 were Retained) *Note.* The analyzed dataset included 2,205 individuals, 1,042 of whom diagnosed with and treated for major depressive disorder and 1,163 diagnosed with and treated for no psychiatric disorder. The rectangles at the end of each branch show how the DT classified the individuals falling in that branch (D = having major depressive disorder; N = not having the diagnosis), the number of individuals falling in that branch (first number in brackets), the number of incorrectly classified individuals (second number in the brackets)

NPV; Table 3). The McNemar test suggested that DT-ID performed better than DT-I even if the difference between the two models was not statistically significant ($\chi^2 = 3.14$, $p < 0.077$). DT-ID classified individuals using the information deriving from two to five PHQ-9 items (2.76 items on average) against the two to seven required by DT-I (2.99 items on average).

In correctly classifying the 625 individuals who reported psychiatric disorders other than MDD as not diagnosed, the DT-ID slightly outperformed DT-I, although the difference was not statistically significant (60.16%, and 56.64% of individuals correctly classified as not diagnosed by DT-ID and DT-I, respectively; $z = 1.26$, $p = 0.21$). For the results of the analysis conducted on the full sample of $N = 2{,}830$ individuals, see Table S1 in the Supplementary Materials.

Results Using Feature Selection on the Nine Items of the PHQ-9

In DT-FS-I, six PHQ-9 items (namely, Items 1, 2, 3, 7, 8, and 9; Fig. 3) were identified as the most effective for the classification of individuals. The accuracy of DT-FS-I was significantly higher than the NIR (NIR = 0.53, $p < 0.001$) and the Kappa coefficient was fair (0.41). These six selected items allowed for classifying individuals with an accuracy that was analogous to that of DT-I (Table 3). The results of the McNemar test showed that there was no difference in the performance of the two models ($\chi^2 = 0.521$, $p = 0.470$), although DT-FS-I was built on six PHQ-9 items out of nine.

In classifying the 625 individuals who reported psychiatric disorders other than MDD as not diagnosed, DT-FS-I was as accurate as DT-I (58.88% and 56.64% of individuals correctly classified as not diagnosed by DT-FS-I and DT-I, respectively; $z = 0.80$, $p = 0.42$), although it was more efficient. For the results of the analysis conducted on the full sample of $N = 2{,}830$ individuals, see Table S1 in the Supplementary Materials.

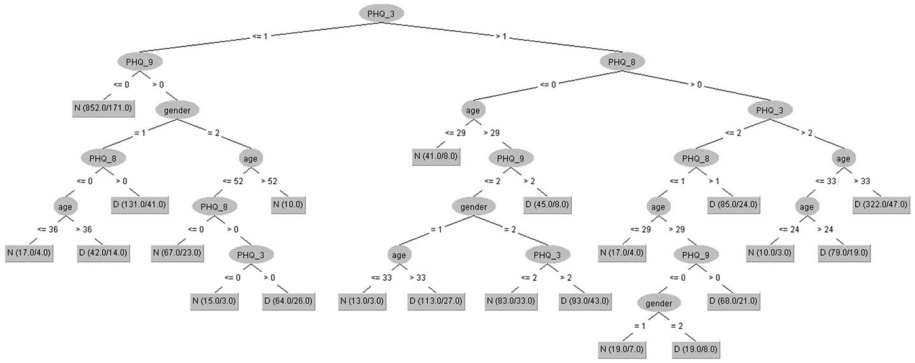

Fig. 4 Decision Tree Obtained Running the J48 Algorithm with Feature Selection on the Nine Items of the PHQ-9 Plus the Two Demographic Variables Gender (1 = Male; 2 = Female) and Age (DT-FS-ID; Items 3, 8, and 9, Gender, and Age were Retained) *Note*. The analyzed dataset included 2,205 individuals, 1,042 of whom diagnosed with and treated for major depressive disorder and 1,163 diagnosed with and treated for no psychiatric disorder. The rectangles at the end of each branch show how the DT classified the individuals falling in that branch (D = having major depressive disorder; N = not having the diagnosis), the number of individuals falling in that branch (first number in brackets), the number of incorrectly classified individuals (second number in the brackets)

Results Using Feature Selection on the Nine Items of the PHQ-9 Plus the Two Demographic Variables

Further improvements in terms of classification assessment measures (i.e., accuracy, sensitivity, specificity, PPV, and NPV) and efficiency were obtained when feature selection was applied to the nine items of the test plus the two demographic variables. Model DT-FS-ID, whose accuracy was again higher than the NIR (NIR = 0.53, $p < 0.001$), showed the largest classification assessment measures (Table 3) and the largest Kappa coefficient (0.47, moderate). This model retained only three PHQ-9 items (namely, Items 3, 8, and 9) and the two demographic variables gender and age (Fig. 4), which appeared in the tree two and seven times, respectively.

By allowing to classify individuals through the information deriving from two to five items (2.36 items on average), model DT-FS-ID turned out to be the most efficient. Moreover, its performance did not significantly differ from that of DT-ID ($\chi^2 = 1.13$, $p = 0.288$) and DT-FS-I ($\chi^2 = 1.38$, $p = 0.239$), even though DT-FS-ID was built on three PHQ-9 items out of nine.

In classifying the 625 individuals who reported psychiatric disorders other than MDD as not diagnosed, DT-FS-ID was more efficient than DT-FS-I and DT-ID but also less accurate (52.64%, 58.88%, and 60.16% of individuals correctly classified as not diagnosed by DT-FS-ID, DT-FS-I, and DT-ID, respectively; for the comparison between DT-FS-ID and DT-FS-I: $z = -2.22$, $p = 0.03$; for the comparison between DT-FS-ID and DT-ID: $z = -2.68$, $p = 0.007$). For the results of the analysis conducted on the full sample of $N = 2{,}830$ individuals, see Table S1 in the Supplementary Materials.

Discussion

This work showed that ML classifiers are a valuable tool for identifying the items that are most effective in categorizing the individuals as having or not having a certain condition. In particular, it illustrated how DT algorithms can increase the accuracy of existing tests and facilitate the development of shortened, yet accurate versions. It also showed how individual variables external to the test (e.g., demographic variables) can be profitably integrated into the assessment process.

The results of the analyses conducted on the responses to the PHQ-9 showed that scoring the nine items of the test according to the procedure outlined by the DT (i.e., DT-I) produces classifications that are more accurate and efficient than those produced by the two cut-off scores indicated in the literature. Moreover, it is worth noting that the DT procedure results in more informative assessments because it values qualitative differences in response patterns, which are usually overlooked by sum score methods. All individuals with the same sum score are classified in the same way by the ROC-based procedures whereas they can be classified in different ways by the DT procedure. For example, using the ROC-based procedures, all individuals with a test score of 8 are classified as not diagnosed, regardless of the specific pattern of responses that produced that score. Sometimes the classification is correct, sometimes it is not. Conversely, based on the DT, particular combinations of responses resulting in a sum score of 8 are classified as not diagnosed, others as diagnosed. In the analyzed dataset, an example is given by an individual who obtained a score of 8 by responding 2 ("More than half the days") to Item 6 ("Feeling bad about yourself — or that you are a failure or have let yourself or your family down"), 0 ("Not at all") to Items 5 ("Poor appetite or overeating") and 7 ("Trouble concentrating on things, such as reading the newspaper or watching television"), and 1 ("Several days") to the remaining six items. Since this response pattern resulted in a score lower than 10 and 11, the individual was classified as not diagnosed by the two cut-off score methods. This classification was incorrect. On the contrary, this individual was correctly categorized as diagnosed by the DT. The DT, in other words, was able to grasp qualitative differences in the response patterns that were not valued by the cut-off score methods.

A key feature of DT algorithms that can be highly useful in psychological testing pertains to the possibility of using them to personalize the assessment. By their nature, DTs are remarkably suitable for developing adaptive tests, which present individuals with a subset of items that are appropriately chosen based on their responses to previous items (Colledani et al., 2023; Delgado-Gomez et al., 2016, 2019; Gonzalez, 2021a; Lopez-Castroman et al., 2016; Michel et al., 2018; Ueno & Songmuang, 2010; Zheng et al., 2020). Moreover, DTs allow for personalizing the assessment also by taking into account the information deriving from variables external to the test, which can be useful to improve assessment accuracy. In this work, for instance, two demographic variables were considered, namely gender and age. The DTs developed using these two variables (i.e., DT-ID and DT-FS-ID) outperformed those developed without them (i.e., DT-I and DT-FS-I) in accuracy. In the DTs that embedded the two demographic variables, branches were defined which contained different items and/or different splitting rules for individuals of specific gender and/or age groups. Achieving such a level of personalization would have been difficult using cut-off score methods because it would have required the definition of appropriate age groups and the identification of numerous clinical cut-off scores (i.e., cut-off scores specific to several gender and age groups). It should also be noted that the ML algorithm

automatically defined the age groups following a process aimed to maximize diagnostic accuracy. Outside an ML approach, the definition of the age groups would have had to be based on more or less arbitrary criteria. The personalization of the assessment is an important feature because it improves diagnostic performance even in bias-free instruments. In the PHQ-9, for example, full strict invariance across gender and age groups was observed. Nevertheless, the diagnostic performance of the test improved when gender and age were considered in the definition of the DT. Including external and relevant variables in the DT is quite easy in the ML approach, and the resulting advantages can be noticeable. This possibility has a crucial relevance considering that often individual characteristics such as gender, age, comorbidity, or ethnicity impact the assessment of psychopathological conditions (Achenbach, 2000; Hartung & Lefler, 2019). Such anamnestic data are normally available or collected but scarcely used to improve the psychodiagnostic performance even when their relevance is recognized by professionals and pointed out in the literature (Brown, 1986; Puente & Perez-Garcia, 2000).

Another application of ML-DT that can be useful in psychological testing is feature selection. It allows for identifying, within a set of variables, those that are most useful for classification purposes. In psychodiagnostic testing, feature selection could be extremely useful in selecting the subset of items that maximizes classification accuracy, and thus represents a valuable method for developing shortened, yet accurate psychodiagnostic tools. When applied to the PHQ-9, feature selection allowed for selecting six of the nine available items (DT-FS-I). The resulting shortened version performed analogously to the full-length scale, while saving one-third of the items. Interestingly, when feature selection was applied to the nine test items plus the two demographic variables (DT-FS-ID), the algorithm produced the most effective solution: It selected three PHQ-9 items and the two demographic variables and resulted in the most accurate and efficient DT.

Given the brevity of the PHQ-9, the development of a shortened version of it could not be a crucial objective. In the psychodiagnostic field, however, tests often contain many items (e.g., the Minnesota Multiphasic Personality Inventory-2 by Hathaway & McKinley, 1989; the Schedule for Nonadaptive and Adaptive Personality by Clark et al., 1993), and it is usually necessary to use more than one instrument. In such situations, feature selection could be an extremely useful method for developing accurate, personalized, and efficient assessment instruments. In addition, feature selection could also contribute to theoretical advances since it allows for identifying not only the attributes that are most relevant for the classification, but also those that are most relevant in the onset and course of psychopathological conditions. The 6-item version of the test obtained with feature selection and without taking into account demographic variables included Items 1 and 2 of the PHQ-9, which describe the core symptoms of depression pertaining to depressed mood and anhedonia, and Item 9 on suicidal ideation, which is a key feature for classification purposes according to the DSM scoring algorithm of the PHQ-9 (Kroenke et al., 2001). In contrast, the 3-item abbreviated version of the test obtained with feature selection and taking into account demographic variables, did not include Items 1 and 2, but did include Item 9.

Overall, the work has shown that DT-based methods allow the development of highly personalized tests that are equivalent or superior to traditional methods in terms of diagnostic classification accuracy, but with much greater efficiency. Interestingly, the good results observed with DT-based methods were obtained using a cross-validation procedure, which allows the results to be considered highly generalizable (James et al., 2013). Furthermore, even on data not used in the algorithm training phases, DT-based methods were as accurate as traditional methods, but much more efficient.

Implications for Theory and Practice

DT-based methods are valuable tools in the field of psychological testing. They facilitate the development of highly personalized adaptive testing procedures, which in turn allow for highly accurate and efficient diagnostic classifications. Compared to traditional cut-off-based methods, DT procedures give greater consideration to the contribution of each individual item in the classification process. In addition, they assign greater relevance to differences in test response patterns and, consequently, to differences in clinical profiles. By focusing on single items/symptoms, DTs allow for identifying not only the attributes that are most relevant for the classification, but also those that are most relevant in the onset and course of psychopathological conditions. This, in turn, is expected to promote theoretical advances.

DT-based procedures also allow for integrating variables external to the test (e.g., age, gender) into the assessment process. This feature appears to be very useful, as it can strongly and positively influence the accuracy of classification by exploiting information that is often available to clinicians but usually overlooked in the assessment. Besides the valuable features mentioned above, a practical advantage of DT-based methods is the high efficiency of the assessment procedures they generate. This aspect is of particular interest in large-scale assessments where multiple, often time-consuming and tedious tests have to be administered. In these situations, reducing the burden of respondents who have to answer lengthy questionnaires can improve attention and reliability of responses. In addition, improving the efficiency of clinical and diagnostic testing procedures is very useful for frail individuals who may have difficulty with long and complex tasks.

Future Research Directions

In this work, a feature selection procedure was used that identified the optimal set of items, without the number of items to be retained being predetermined. As a result, the length of the shortened tests could be any. In some cases, it may be useful to construct shortened tests containing a predetermined number of items. Feature selection procedures exist that are useful to this purpose (e.g., ranker methods; Witten et al., 2017). Future studies may be devoted to comparing the functioning of shortened tests obtained using different feature selection procedures, with and without a predetermined number of items. In addition, further studies are needed to test the effectiveness of feature selection and DT-based methods over traditional strategies in the development of static short forms of tests.

DT-based methods are a promising strategy for improving diagnostic testing procedures. However, the literature suggests that further studies are needed to understand their utility in assessing disease severity and levels of specific traits in respondents (Michel et al., 2018; Riley et al., 2011; Ueno & Songmuang, 2010).

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11469-024-01332-x.

Author contributions Daiana Colledani planned the study, conducted the literature search, analyzed and interpreted the data, and wrote the article. Pasquale Anselmi analyzed and interpreted the data, and wrote the article. Egidio Robusto interpreted the data and wrote the article.

Funding Open access funding provided by Università degli Studi di Padova within the CRUI-CARE Agreement. This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Data Availability The present study utilized a dataset that is publicly available on FigShare (Doi et al., 2018; Ito et al., 2015) and licensed under the Creative Commons Attribution 4.0 (CC BY 4.0).

Declarations

Ethic statement The present study utilized a publicly available dataset that is licensed under the Creative Commons Attribution 4.0 (CC BY 4.0). As such, no research ethics committee approval was required. However, it should be noted that the original study from which the dataset was obtained had received research ethics committee approval. Details regarding the data sources are documented in the Methods section of this paper and are supported by appropriate references.

Competing interests The authors declare no competing interests.

References

Achenbach, T. M. (2000). Assessment of psychopathology. In A. J. Sameroff, M. Lewis, & S. M. Miller (Eds.), *Handbook of developmental psychopathology* (pp. 41–56). Kluwer Academic Publishers. https://doi.org/10.1007/978-1-4615-4163-9_3

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Bentler, P. M. (1995). *EQS structural equations program manual* (Vol. 6). Encino, CA: Multivariate Software, Inc.

Bianchi, R., Verkuilen, J., Toker, S., Schonfeld, I. S., Gerber, M., Brähler, E., & Kroenke, K. (2022). Is the PHQ-9 a unidimensional measure of depression? A 58,272-participant study. *Psychological Assessment, 34*(6), 595. https://doi.org/10.1037/pas0001124

Blazer, D. G., Kessler, R. C., McGonagle, K. A., & Swartz, M. S. (1994). The prevalence and distribution of major depression in a national community sample: The National Comorbidity Survey. *The American Journal of Psychiatry, 151*(7), 979–986. https://doi.org/10.1176/ajp.151.7.979

Breiman, L., Friedman, J. H., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Pacific Grove: Wadsworth & Brooks.

Brodaty, H., Cullen, B., Thompson, C., Mitchell, P., Parker, G., Wilhelm, K., ... & Malhi, G. (2005). Age and gender in the phenomenology of depression. *The American Journal of Geriatric Psychiatry*, *13*(7), 589−596. https://doi.org/10.1097/00019442-200507000-00007

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. Guilford Press.

Brown, J. M., & Weiss, D. J. (1977). *An adaptive testing strategy for achievement test batteries (Research Report 77–6)*. Minn, University of Minnesota, Computerized Adaptive Testing Laboratory.

Brown, L. S. (1986). Gender-role analysis: A neglected component of psychological assessment. *Psychotherapy: Theory, Research, Practice, Training, 23*(2), 243–248.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Sage.

Caruana, R., & Freitag, D. (1994). Greedy attribute selection. In *Proceedings of the Eleventh International Conference on International Conference on Machine Learning* (pp. 28−36). Morgan Kaufmann. https://doi.org/10.1016/B978-1-55860-335-6.50012-X

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Clark, L. A., McEwen, J. L., Collard, L. M., & Hickok, L. G. (1993). Symptoms and traits of personality disorder: Two new methods for their assessment. *Psychological Assessment, 5*(1), 81–91. https://doi.org/10.1037/1040-3590.5.1.81

Colledani, D. (2018). Psychometric properties and gender invariance for the Dickman Impulsivity Inventory. *TPM-Testing, Psychometrics, Methodology in Applied Psychology, 25*(1), 49–61. https://doi.org/10.4473/TPM25.1.3

Colledani, D., Anselmi, P., & Robusto, E. (2018). Using item response theory for the development of a new short form of the Eysenck Personality Questionnaire-Revised. *Frontiers in Psychology, 9*, 1834. https://doi.org/10.3389/fpsyg.2018.01834

Colledani, D., Anselmi, P., & Robusto, E. (2019). Using multidimensional item response theory to develop an abbreviated form of the Italian version of Eysenck's IVE questionnaire. *Personality and Individual Differences, 142*, 45–52. https://doi.org/10.1016/j.paid.2019.01.032

Colledani, D., Meneghini, A. M., Mikulincer, M., & Shaver, P. R. (2022). The Caregiving System Scale: Factor structure, gender invariance, and the contribution of attachment orientations. *European Journal of Psychological Assessment, 38*(5), 385–396. https://doi.org/10.1027/1015-5759/a000673

Colledani, D., Anselmi, P., & Robusto, E. (2023). Machine learning-decision tree classifiers in psychiatric assessment: An application to the diagnosis of major depressive disorder. *Psychiatry Research*, *322*, 115127. https://doi.org/10.1016/j.psychres.2023.115127

Costantini, L., Pasquarella, C., Odone, A., Colucci, M. E., Costanza, A., Serafini, G., ... & Amerio, A. (2021). Screening for depression in primary care with Patient Health Questionnaire-9 (PHQ-9): A systematic review. *Journal of Affective Disorders*, *279*, 473−483. https://doi.org/10.1016/j.jad.2020.09.131

Criminisi, A., Shotton, J., & Konukoglu, E. (2012). Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends® in Computer Graphics and Vision, 7*(2–3), 81–227. https://doi.org/10.1561/0600000035

Cumming, G. (2008). Replication and p intervals: P values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science, 3*(4), 286–300. https://doi.org/10.1111/j.1745-6924.2008.00079.x

De Beurs, D. P., de Vries, A. L., de Groot, M. H., de Keijser, J., & Kerkhof, A. J. (2014). Applying computer adaptive testing to optimize online assessment of suicidal behavior: A simulation study. *Journal of Medical Internet Research, 16*(9), e207. https://doi.org/10.2196/jmir.3511

Dekker, G. W., Pechenizkiy, M. & Vleeshouwers, J. M. (2009, July 1−3). *Predicting students drop out: A case study*. EDM'09 - Educational Data Mining 2009: 2nd International Conference on Educational Data Mining, Cordoba, Spain.

Delgado-Gomez, D., Baca-Garcia, E., Aguado, D., Courtet, P., & Lopez-Castroman, J. (2016). Computerized adaptive test vs. decision trees: Development of a support decision system to identify suicidal behavior. *Journal of Affective Disorders, 206*, 204–209. https://doi.org/10.1016/j.jad.2016.07.032

Delgado-Gomez, D., Laria, J. C., & Ruiz-Hernandez, D. (2019). Computerized adaptive test and decision trees: A unifying approach. *Expert Systems with Applications, 117*, 358–366. https://doi.org/10.1016/j.eswa.2018.09.052

Dixon, M. F., Halperin, I., & Bilokon, P. (2020). *Machine learning in Finance* (Vol. 1170). New York, NY: Springer International Publishing.

Doi, S., Ito, M., Takebayashi, Y., Muramatsu, K., & Horikoshi, M. (2018). Factorial validity and invariance of the Patient Health Questionnaire (PHQ)-9 among clinical and non-clinical populations. *PLoS ONE, 13*(7), e0199235. https://doi.org/10.1371/journal.pone.0199235

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology, 14*, 91–118. https://doi.org/10.1146/annurev-clinpsy-032816-045037

Dzyabura, D., & Yoganarasimhan, H. (2018). Machine learning and marketing. In N. Mizik & D. M. Hanssens (Eds.), *Handbook of marketing analytics* (pp. 255–279). Edward Elgar Publishing.

Eggen, T. J. H. M., & Straetmans, G. J. J. M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement, 60*(5), 713–734. https://doi.org/10.1177/0013164002197086

Fliege, H., Becker, J., Walter, O. B., Bjorner, J. B., Klapp, B. F., & Rose, M. (2005). Development of a computer-adaptive test for depression (D-CAT). *Quality of Life Research, 14*, 2277–2291. https://doi.org/10.1007/s11136-005-6651-9

Gibbons, R. D., Weiss, D. J., Kupfer, D. J., Frank, E., Fagiolini, A., Grochocinski, V. J., ... & Immekus, J. C. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services*, *59*(4), 361−368. https://doi.org/10.1176/ps.2008.59.4.361

Gibbons, R. D., Hooker, G., Finkelman, M. D., Weiss, D. J., Pilkonis, P. A., Frank, E., ... & Kupfer, D. J. (2013). The computerized adaptive diagnostic test for major depressive disorder (CAD-MDD): a screening tool for depression. *The Journal of Clinical Psychiatry*, *74*(7), 3579. https://doi.org/10.4088/JCP.12m08338

Gilbody, S., Richards, D., Brealey, S., & Hewitt, C. (2007). Screening for depression in medical settings with the Patient Health Questionnaire (PHQ): A diagnostic meta-analysis. *Journal of General Internal Medicine, 22*(11), 1596–1602. https://doi.org/10.1007/s11606-007-0333-y

Gonzalez, O. (2021a). Psychometric and machine learning approaches for diagnostic assessment and tests of individual classification. *Psychological Methods, 26*(2), 236–254. https://doi.org/10.1037/met0000317

Gonzalez, O. (2021b). Psychometric and machine learning approaches to reduce the length of scales. *Multivariate Behavioral Research, 56*(6), 903–919. https://doi.org/10.1080/00273171.2020.1781585

Gupta, B., Rawat, A., Jain, A., Arora, A., & Dhami, N. (2017). Analysis of various decision tree algorithms for classification in data mining. *International Journal of Computer Applications, 163*(8), 15–19. https://doi.org/10.5120/ijca2017913660

Hamilton, L. S. (1999). Detecting gender-based differential item functioning on a constructed-response science test. *Applied Measurement in Education, 12*(3), 211–235. https://doi.org/10.1207/S15324818AME1203_1

Hartung, C. M., & Lefler, E. K. (2019). Sex and gender in psychopathology: DSM–5 and beyond. *Psychological Bulletin, 145*(4), 390–409. https://doi.org/10.1037/bul0000183

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Vol. 2). Springer.

Hathaway, S. R., & McKinley, J. C. (1989). *MMPI-2: Minnesota Multiphasic Personality Inventory-2: Manual for administration and scoring*. University of Minnesota Press.

Higa, A. (2018). Diagnosis of breast cancer using decision tree and artificial neural network algorithms. *International Journal of Computer Applications Technology and Research, 1*(7), 23–27. https://doi.org/10.7753/ijcatr0701.1004

Ito, M., Bentley, K. H., Oe, Y., Nakajima, S., Fujisato, H., Kato, N., Miyamae, M., Kanie, A., Horikoshi, M., & Barlow, D. H. (2015). Assessing depression related severity and functional impairment(warning) the Overall Depression Severity and Impairment Scale (ODSIS). *PLoS ONE, 10*(4), e0122969. https://doi.org/10.1371/journal.pone.0122969

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

John, G. H., Kohavi, R., & Pfleger, K. (1994). Irrelevant features and the subset selection problem. In *Proceedings of the Eleventh International Conference on International Conference on Machine Learning* (pp. 121−129). Morgan Kaufmann. https://doi.org/10.1016/B978-1-55860-335-6.50023-4

Karabulut, E. M., Özel, S. A., & Ibrikci, T. (2012). A comparative study on the effect of feature selection on classification accuracy. *Procedia Technology, 1*, 323–327. https://doi.org/10.1016/j.protcy.2012.02.068

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence, 97*(1–2), 273–324. https://doi.org/10.1016/S0004-3702(97)00043-X

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine, 16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 363−374. https://doi.org/10.2307/2529786

Langley, P., & Sage, S. (1994). Induction of selective Bayesian classifiers. In *Uncertainty Proceedings 1994* (pp. 399−406). Morgan Kaufmann. https://doi.org/10.48550/arXiv.1302.6828

Li, P. (2023). The Application of Decision Tree Algorithm in Psychological Assessment Data. *The International Conference on Cyber Security Intelligence and Analytics* (pp. 185–194). Springer Nature Switzerland: Cham.

Lin, J. (2001). Feature extraction of machine sound using wavelet and its application in fault diagnosis. *NDT and E International, 34*(1), 25–30. https://doi.org/10.1016/S0963-8695(00)00025-6

Van der Linden, W. J., & Glas, C. A. (Eds.) (2000). *Computerized adaptive testing: Theory and practice*. Springer Science & Business Media.

Lopez-Castroman, J., Delgado-Gomez, D., Courtet, P., & Baca-Garcia, E. (2016). Optimizing the assessment of suicide attempters with a decision tree. *European Psychiatry, 33*(S1), S602–S603. https://doi.org/10.1016/j.eurpsy.2016.01.2251

Mahesh, B. (2020). Machine learning algorithms -a review. *International Journal of Science and Research (IJSR), 9*, 381–386. https://doi.org/10.21275/ART20203995

Manea, L., Gilbody, S., & McMillan, D. (2012). Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): a meta-analysis. *CMAJ: anadian Medical Association Journal, 184*(3), E191–E196. https://doi.org/10.1503/cmaj.110829

Michel, P., Baumstarck, K., Loundou, A., Ghattas, B., Auquier, P., & Boyer, L. (2018). Computerized adaptive testing with decision regression trees: An alternative to item response theory for quality of life measurement in multiple sclerosis. *Patient Preference and Adherence, 12*, 1043–1053. https://doi.org/10.2147/PPA.S162206

Moore, T. M., Calkins, M. E., Reise, S. P., Gur, R. C., & Gur, R. E. (2018). Development and public release of a computerized adaptive (CAT) version of the Schizotypal Personality Questionnaire. *Psychiatry Research, 263*, 250–256. https://doi.org/10.1016/j.psychres.2018.02.022

Muramatsu, K., Kamijima, K., Yoshida, M., Otsubo, T., Miyaoka, H., Muramatsu, Y., & Gejyo, F. (2007). The patient health questionnaire, Japanese version: Validity according to the mini-international neuropsychiatric interview–plus. *Psychological Reports, 101*(3), 952–960. https://doi.org/10.2466/pr0.101.3.952-960

Muramatsu, K., Miyaoka, H., Kamijima, K., Muramatsu, Y., Tanaka, Y., Hosaka, M., ... & Shimizu, E. (2018). Performance of the Japanese version of the Patient Health Questionnaire-9 (J-PHQ-9) for depression in primary care. *General Hospital Psychiatry*, *52*, 64−69. https://doi.org/10.1016/j.genhosppsych.2018.03.007

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Muthén & Muthén.

Novaković, J., Strbac, P., & Bulatović, D. (2011). Toward optimal feature selection using ranking methods and classification algorithms. *Yugoslav Journal of Operations Research, 21*(1), 119–135. https://doi.org/10.2298/YJOR1101119N

Patten, S. B., Williams, J. V., Lavorato, D. H., Wang, J. L., Bulloch, A. G., & Sajobi, T. (2016). The association between major depression prevalence and sex becomes weaker with age. *Social Psychiatry and Psychiatric Epidemiology, 51*, 203–210. https://doi.org/10.1007/s00127-015-1166-3

Paulus, M. P., & Thompson, W. K. (2021). Computational approaches and machine learning for individual-level treatment predictions. *Psychopharmacology (berl), 238*(5), 1231–1239. https://doi.org/10.1007/s00213-019-05282-4

Prabhakar, S., Mohanty, A. R., & Sekhar, A. S. (2002). Application of discrete wavelet transform for detection of ball bearing race faults. *Tribology International, 35*(12), 793–800. https://doi.org/10.1016/S0301-679X(02)00063-4

Puente, A. E., & Perez-Garcia, M. (2000). Psychological assessment of ethnic minorities. In G. Goldstein & M. Hersen (Eds.), *Handbook of psychological assessment* (pp. 527−551). Amsterdam: Pergamon Press.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

Reich, Y. (1997). Machine learning techniques for civil engineering problems. *Computer-Aided Civil and Infrastructure Engineering, 12*(4), 295–310. https://doi.org/10.1111/0885-9507.00065

Riley, B. B., Funk, R., Dennis, M. L., Lennox, R. D., & Finkelman, M. (2011, October 3–5). *The use of decision trees for adaptive item selection and score estimation.* Annual conference of the international association for computerized adaptive testing. Pacific Grove, CA.

Salzberg, S. L. (1994). C4.5: Programs for machine learning by J. Ross Quinlan. Morgan Kaufmann publishers Inc, 1993. *Machine Learning, 16*, 235–240.

Sarker, I. H. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science, 2*(3), 160. https://doi.org/10.1007/s42979-021-00592-x

Song, E., Huang, D., Ma, G., & Hung, C. C. (2011). Semi-supervised multi-class Adaboost by exploiting unlabeled data. *Expert Systems with Applications, 38*(6), 6720–6726. https://doi.org/10.1016/j.eswa.2010.11.062

Spitzer, R. L., Kroenke, K., Williams, J. B., Patient Health Questionnaire Primary Care Study Group, & Patient Health Questionnaire Primary Care Study Group. (1999). Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. *JAMA, 282*(18), 1737–1744. https://doi.org/10.1001/jama.282.18.1737

Stewart, R. W., Tuerk, P. W., Metzger, I. W., Davidson, T. M., & Young, J. (2016). A decision-tree approach to the assessment of posttraumatic stress disorder: Engineering empirically rigorous and ecologically valid assessment measures. *Psychological Services, 13*(1), 1. https://doi.org/10.1037/ser0000069

Sugumaran, V., Muralidharan, V., & Ramachandran, K. I. (2007). Feature selection using Decision Tree and classification through Proximal Support Vector Machine for fault diagnostics of roller bearing.

Mechanical Systems and Signal Processing, 21(2), 930–942. https://doi.org/10.1016/j.ymssp.2006.05.004

Suzuki, K., Kumei, S., Ohhira, M., Nozu, T., & Okumura, T. (2015). Screening for major depressive disorder with the Patient Health Questionnaire (PHQ-9 and PHQ-2) in an outpatient clinic staffed by primary care physicians in Japan: A case control study. *PLoS ONE, 10*(3), e0119147. https://doi.org/10.1371/journal.pone.0119147

Tseng, W. T. (2016). Measuring English vocabulary size via computerized adaptive testing. *Computers & Education, 97*, 69–85. https://doi.org/10.1016/j.compedu.2016.02.018

Ueno, M., & Songmuang, P. (2010). Computerized adaptive testing based on decision tree. Proceedings of the *10th IEEE International conference on advanced learning technologies, ICALT 2010* (pp. 191–193). https://doi.org/10.1109/ICALT.2010.58

Uğuz, H. (2011). A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm. *Knowledge-Based Systems, 24*(7), 1024–1032. https://doi.org/10.1016/j.knosys.2011.04.014

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4–70. https://doi.org/10.1177/109442810031002

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2017). *Data Mining: Practical Machine Learning Tools and Techniques (4th Edition)*. Morgan Kaufmann.

Yan, D., Lewis, C., & Stocking, M. (2004). Adaptive Testing with Regression Trees in the Presence of Multidimensionality. *Journal of Educational and Behavioral Statistics, 29*(3), 293–316. http://www.jstor.org/stable/3701355.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science, 12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Zhao, Y., & Zhang, Y. (2008). Comparison of decision tree methods for finding active objects. *Advances in Space Research, 41*(12), 1955–1959. https://doi.org/10.1016/j.asr.2007.07.020

Zheng, Y., Cheon, H., & Katz, C. M. (2020). Using machine learning methods to develop a short tree-based adaptive classification test: Case study with a high-dimensional item pool and imbalanced data. *Applied Psychological Measurement, 44*(7–8), 499–514. https://doi.org/10.1177/0146621620931