Global Explanation of Tree-Ensembles Models Based on Item Response Theory

José Ribeiro[1,2*], Lucas Cardoso[2,3], Raíssa Silva[4,5], Vitor Cirilo[2,3], Níkolas Carneiro[2,3] and Ronnie Alves[2,3*]

[1*]Federal Institute of Education, Science and Technology of Pará - IFPA, Ananindeua, 67125-000, PA, Brazil.
[2]Federal University of Pará - UFPA, Belém, 66075-10, PA, Brazil.
[3]Vale Institute of Technology - ITV, Belém, 66055-090, PA, Brazil.
[4]Montpellier University, Montpellier, 34090, France.
[5]La Ligue Contre le Cancer, Montpellier, 34000, France.

*Corresponding author(s). E-mail(s): jose.ribeiro@ifpa.edu.br; ronnie.alves@itv.org;
Contributing authors: lucas.cardoso@pq.itv.org; r.lorenna@gmail.com; vitor.cirilo.santos@itv.org ; nikolas.carneiro@itv.org;

Abstract

Explainable Artificial Intelligence - XAI is aimed at studying and developing techniques to explain black box models, that is, models that provide limited self-explanation of their predictions. In recent years, XAI researchers have been formalizing proposals and developing new measures to explain how these models make specific predictions. In previous studies, evidence has been found on how model (dataset and algorithm) complexity affects global explanations generated by XAI measures *Ciu, Dalex, Eli5, Lofo, Shap* and *Skater*, suggesting that there is room for the development of a new XAI measure that builds on the complexity of the model. Thus, this research proposes a measure called Explainable based on Item Response Theory - *eXirt*, which is capable of explaining tree-ensemble models by using the properties of Item Response Theory (IRT). For this purpose, a benchmark was created using 40 different datasets and 2 different algorithms (Random Forest and Gradient Boosting), thus generating 6 different explainability ranks using known XAI measures along with 1 data purity rank and 1 rank of the measure *eXirt*, amounting to 8

global ranks for each model, i.e., 640 ranks altogether. The results show that *eXirt* displayed different ranks than those of the other measures, which demonstrates that the advocated methodology generates global explanations of tree-ensemble models that have not yet been explored, either for the more difficult models to explain or even the easier ones.

Keywords: Global Explanation, Item Response Theory, Explainable Artificial Intelligence, Black box

1 Introduction

Technology has been evolving day by day, and today artificial intelligence is already a reality in the daily life of society. There are many real-world problems that machine learning algorithms solve, making human daily life more automated and intelligent [1, 2].

Machine learning models based on tree-structured bagging and boosting algorithms are known to provide high performance and high generalization capabilities, and thus being widely used in intelligent systems embedded in real-world problems [3, 4].

Despite the popularity of these algorithms, a negative property they have is that they are not transparent[1], their predictions are not self-explanatory, thus being considered black box algorithms[2] and, therefore, less used in problems related to sensitive contexts, such as those of a social nature concerning health and safety [5–7].

With the increasing need for high-performance models — which implies low transparency[5] — in sensitive contexts, there is currently a growing need to develop measures or tools that can provide information about local explanations[3] and global explanations[4] as a means to make predictions more easily interpretable and also more reliable by humans [10].

Efforts have been made by the community that researches Explainable Artificial Intelligence - XAI in developing different measures to explain black box models. Many of these efforts are defined in XAI measures that use the well-trained machine learning model, its input data, and its outputs in order to explain the model. This type of measurement is called Analysis *Post-hoc* [5].

In this regard, measures such as *Ciu* [11], *Dalex* [12], *Eli5* [13], *Lofo* [14], *Shap* [15] e *Skater* [16] have emerged to promote the creation of model-agnostic explanations[5]. It should be noted that each of the tools mentioned above is capable of explaining models using different techniques and methodologies,

[1] Transparent Algorithms: Algorithms that generate explanations of how a particular output was produced. Such examples include Decision Tree, Logistic Regression, and K-Nearest Neighbors.

[2] Black box algorithms: machine learning algorithms that have classification or regression decisions hidden from the user.

[3] Local explanations: explanations in attribute relevance format generated around data instances[8]

[4] Global explanations: when it is possible to understand the logic of all instances of the model generating a global ranking of attribute relevance[8, 9]

[5] Model-Agnostic: means that it does not depend on type of model to be explained.

but one fact they have in common is that they all generate global relevance rankings of attributes related to the explanation of a model. And, therefore, are likely to have their results compared [17].

The terminologies Attribute Relevance Ranking and Attribute Importance Ranking are widely used as synonyms in the computing community, but have different definitions herein, as shown in [5]. Since attribute rankings are regarded as ordered structures whereby each attribute of the dataset used by the model appears in a position indicated by a score. The main difference being that, in relevance ranking, the calculation of the score is based on the model output, whereas to calculate the importance ranking of attributes, the correct label to be predicted is used [5, 18].

Thus, while attribute relevance ranking acts as an explanation of how attributes contribute to reaching a particular model output, attribute importance ranking acts as a performance measure related to how attributes contribute to reaching the correct prediction [5].

Global attribute relevance rankings — rather than local ones — are analyzed because they allow for general analyses of how a given model generalizes to a specific problem, along with analyses of how a given measure explains a specific model, without the need for a preliminary understanding of the context in which the model is embedded. [17].

Also, attribute importance rankings are analyzed in comparison with attribute relevance rankings in an attempt to emphasize to the computing community the existence of differences or even similarities between these two different types of rankings.

This way, the studies presented in [17] are continued by replicanting the experiments in benchmark format with the six measures of attribute relevance, an attribute importance measure is included for comparison; and finally, a new XAI measure is included to explain tree-ensemble models considering their complexity and using the Item Response Theory (IRT) perspective. This new measure was named Explainable based in Item Response Theory - *eXirt*.

Item Response Theory, used as the basis for developing the measure *eXirt*, is a very widespread theory, generally used in the process of evaluating candidates in selection processes. The theory uses the properties 'guessing', 'difficulty' and 'discrimination' to enable the evaluation of latent characteristics, which cannot be observed directly, of the responses of candidates in a selection process. This is intended to establish the relationship of hit probability to the candidate's ability [19].

Adding the new experiments to the methodology advocated in [17], by developing the measure *eXirt* and by adding attribute importance rankings to the comparisons made in the benchmark, seek to demonstrate that measure *eXirt* comes to fulfill a need in the XAI area, since this measure considers the complexity of the model to be explained. These additions also prompt new discussions about how model complexity affects model explainability and also how the importance and relevance of model attributes relate to each other.

Thus, the main contributions to the studies in Explainable Artificial Intelligence that this research generates are as follows:

- An innovative XAI measure, called *eXirt*, which is based on Item Response Theory, is still under-explored in computing for generating rankings of global explanations of tree-ensemble black box models;
- Investigations regarding the complexity of computational models as a function of their explanations, providing evidence of how these relationships are established, thus furthering the research [17];
- Comparative results on attribute relevance measure rankings and attribute importance rankings.

2 Related works

2.1 Explainable Artificial Intelligence

In recent years, there has been an increasing need to explain black box machine learning models in an agnostic manner. This includes machine learning models such as those based on ensemble algorithms [20–23].

Hence, studies can be found in the literature highlighting research on model-agnostic XAI measures that are proposed as general explanations for many different artificial intelligence-based systems. This context includes studies on XAI measures aimed at explaining simpler models — which use tabular data and tree-ensemble-based black box algorithms, for example — that are popularly used in solving various real-world problems and require explanations regarding their predictions [5, 22].

Based on a bibliographic survey conducted by the authors of the herein research, it was possible to find the main XAI measures included in the aforementioned context, that is, measures specifically aimed at generating global attribute relevance rankings in a model-agnostic manner that support tabular data. As a result, a total of six XAI measures were found to be properly validated and compatible with one another (at library and code execution dependencies level). These measures include: *CIU* [11], *Dalex* [24], *Eli5* [13], *Lofo* [14], *SHAP* [15] and *Skater* [16].

Due to incompatibilities between XAI measurement libraries and versions of machine learning model libraries and dependencies *scikit-lean,scikit-learn*, only the Random Forest and Gradient Boosting algorithms were used herein as models to be explained by the aforementioned measures. In other words, this problem is directly linked to the way these tools were programmed rather than the methodologies they advocate.

Note that the six measures presented herein generate relevance rankings based on the same previously trained machine learning models — with the same training and testing split —, manipulate their inputs and/or produce new intermediate models, if necessary (copies). Therefore, they are required to be compatible with each other so that a fair comparison of their final rankings of explanations can be made.

This research found other tools aimed at model explanation, including: *Alibe-ALE* [25], *Lime* [26], *Ethical ExplainableAI* [27], *IBM Explainable AI 360* [28], e *Interpreter ML* [29]. However, during the analysis, it was found that these tools did not meet the criteria established hereby, and for this reason they were not included in the tests.

As the first XAI measure compatible with the established criteria, the Contextual Importance and Utility measure stands out. - *CIU* is a XAI measure based on Decision Theory [30] that focuses on serving as a unified measure of model-agnostic explainability based on the implementation of two different scores, namely: Contextual Importance - CI and Contextual Utility - CU, which generally create equal attribute ranks, but with different values for each score, thus generating global explanations [11]. Only the CI rank is considered due to the fact that, even with different CI and CU values for each attribute, the positions of the attributes in the two ranks were always the same in the tests performed. — Using the term "importance" here is wrong, it should be "relevance" according to [5].

Another XAI measure is *Dalex*, a set of XAI tools based on the LOCO (leave-one covariate out) approach and it can generate explainabilities from this approach. In general, the measure receives the model and the data to be explained; it calculates the performance of the model, performs new training processes with new generated data sets, the iterative inversion of each attribute of the reported data is its main differentiator, since it can calculate how and which attributes are more relevant for the model through its performance [12, 31].

Following a similar strategy is the Leave One Feature Out - *Lofo*, but with a main difference regarding the inversion of the attribute iterativity, because in the Lofo measure the iterative step is the removal of the attribute to find its global relevance to the model based on the performance[14].

One of the most popular and widespread XAI measures today is the Explain Like I'm Five - *Eli5*, a tool that helps explore machine learning classifiers and explains their predictions by assigning weights to decisions, as well as exporting decision trees and presenting the relevance of the attributes of the model submitted to the tool [13].

One of the most used XAI measures today is the SHapley Additive exPlanations - *SHAP*, a unified measure of attribute relevance that explains the prediction of an instance X from the contribution of an attribute. The contribution of this attribute is calculated from the game theory of Shapley Value [32]. This measure calculates the explanation score of model attributes iteratively with each attribute and data instance by calculating shapley values [15, 33].

Finally, the *Skater* was found, which a set of tools capable of generating rankings of the relevance of model attributes, based on Information Theory [34], through measurements of entropy in changing predictions, through a perturbation of a certain attribute. The central idea is that the more a model is dependable upon an attribute, the greater the change in predictions [16].

Table 1 shows a general comparison between the main techniques selected and used herein.

Table 1 Main XAI measures surveyed

XAI meansure	Autor	Base algorithm	Explanation Technique	Global explanation (by rank)	Local explanation	API compatible
*eXirt**	-	Item Response Theory	Permutation of Feature	Yes	Yes	Yes
CIU	[11]	Decision Theory	Multiple Criteria Decision Making	Yes	No	Yes
Dalex	[24]	Leave-one covariate out	Permutation of Feature	Yes	Yes	Yes
Eli5	[13]	Assigning weights to decisions	Permutation of Feature and Mean Decrease Accuracy	Yes	Yes	Yes
Lofo	[14]	Leave One Feature Out	Permutation of Feature	Yes	No	Yes
SHAP for Tree	[15]	Game Theory	Permutation of Feature	Yes	Yes	Yes
Skater	[16]	Information Theory	Permutation of Feature	Yes	Yes	Yes

***Note: the eXirt, Explainable base in Item Response Theory, is the XAI measure defended by this article, it appears prematurely here only at the level of global comparison with the other measures.

Still in table 1, it can be seen that most existing XAI measures use the "Permutation of Feature" technique to perform the model explanation process. However, it should be emphasized at this point, that the *eXirt* differs from other measures by having a more robust model evaluation methodology than other techniques.

All the measures described above are capable of generating several types of model explanations — going beyond the generation of attribute relevance ranks. However, since the focus of this article is to compare this important basic structure of model explanation, the rank, it is restricted to comparing only this type of result generated by each measure.

2.2 Dataset Properties

After being processed through attribute engineering, a dataset is expected to exhibit characteristics that involve the nature of the problem in a contextual (problem properties) and technical (dataset properties) way that can be generalized by algorithms in order to obtain the solution for several computationally solvable problems [10].

It can be considered that although the contextual side of the problem helps significantly in the interpretation of the explanations to the models, the authors decided not to work with these characteristics of the analyzed data, since the datasets that were used involve knowledge from outside the area of computing, as they are data related to different real world problems.

On the other hand, there are dataset properties that are possible to be analyzed by means of techniques, as they are more easily computable, without having to take into consideration the problem context the data refers to. Properties such as dimensionality, number of numeric attributes, number of binary attributes, balance between classes and entropy are examples of properties that every tabular dataset has and which directly influence the generalization of the proposed model [35].

In a clustering process followed by a correspondence analysis, performed on properties from different datasets, information can be provided concerning

dataset clusters that show similarities among themselves, thus identifying the profile that each dataset cluster has in relation to the analyzed properties, as performed in [17].

2.3 Easy- and Difficult-to-Explain Models

Surveys performed in [17], show indications of the existence of computational models (dataset and algorithm) that are less complex to explain and others that are more complex — which, in other words, is the same as saying that there are models that are easier to explain and others that are more difficult, respectively. This statement was only possible due to an extensive comparative analysis performed between a considerable number of models — based on Random Forest and Gradiant Boosting algorithms — and explanations generated by the same 6 XAI measures mentioned herein.

In this study, it was identified that for two distinct clusters of datasets, analyzed by means of Multiple Correspondence Analysis - MCA [36], there were different values of correlations between the XAI measures analyzed, since for a first cluster it was found that the correlations were mostly insignificant, while for a second cluster it was found that the correlations found had considerable values, both positive and negative [17].

The line of logical reasoning that allows for assessing easy models to be explained and difficult models, according to the study [17], is as follows:

- If a model (dataset and algorithm) is considered easy to explain, then it is expected to have a small number of possible explanation ranks. Therefore, higher correlations between the various known measures of XAI are expected.
- However, if a model (dataset and algorithm) is considered difficult to explain, then it is expected to have a high number of possible explanation ranks. Thus, lower correlations between the various known measures of XAI are expected.

Reaching the aforementioned line of logical reasoning was only possible after verifying the results of 592 explainability ranks generated in [17], by observing the specificities of the different groups of datasets being used through the clustering process of their properties and also by a Multiple Correspondence Analysis [36] in each cluster.

2.4 Item Response Theory

Item Response Theory belongs to the field of Psychometrics and provides mathematical models for the estimation of latent traits[6], by proposing ways to represent: the relationship between the probability of an individual giving a specific answer to an item, or its latent trait, and the characteristics (parameters) of the items considering the analyzed knowledge area [19].

[6]latent traits: refers to a family of mathematical models that relate observable variables (test items, for example)[37].

According to [19], typically when evaluating the performance of individuals in a test, traditionally the total number of correct answers given by them is used. Even if this is the most natural way to evaluate individuals, one must consider that it has several limitations, such as: how to evaluate individuals who get the test questions right without having the necessary knowledge to get them right? — that is, individuals who "have guessed" certain answers and got them right — as well as, how to evaluate the level of difficulty posed by each question? And lastly, can a specific test, or even a specific question, distinguish individuals who have the knowledge needed to answer it from individuals who do not have such knowledge?

In this regard, the IRT aims to allow an evaluation of the latent characteristics of an individual that cannot be directly observed and aims to provide the relationship between the probability of an individual correctly answering an item and his/her latent traits. In other words, the individual's ability in the area of knowledge being assessed [38].

An interesting feature of the IRT is that it focuses on the items and not on the test as a whole. Thus, an individual's performance is assessed by taking into account his/her ability to get specific items on a test and not simply by considering how many items he/she got right. [19].

In general, IRT can be considered as a set of mathematical models that try to represent the probability of an individual getting an item right as a function of item parameters and respondent's ability[7]. Thus, the greater the individual's ability, the greater his chance to get answers right. [38].

The calculation of the item parameters has different implementations in the literature, such as "Rasch Dichotomous Model" [39] and "Birnbaum's Three Parameter Model" [40]. The implementation used herein is the most popular today, namely the logistic model $3PL$, shown in equation 1, which consists of a model capable of evaluating the respondents of a test from the estimated ability (θ_j), along with the probability of correct answer $P(U_{ij} = 1 \mid \theta_j)$ calculated as a function of an individual's ability j and the item parameters i.

$$P(U_{ij} = 1 \mid \theta_j) = c_i + (1 - c_i)\frac{1}{1 + e^{-a_i(\theta_j - b_i)}} \tag{1}$$

Seeking to better explain what each of the item parameters shown in the equation 1 is, the following definitions are given:

- a_i is the discrimination parameter and consists in how much a specific item i is able to differentiate between highly and poorly skilled respondents. It is understood that the higher its value, the more discriminative the item is. Ideally, a test should feature a gradual and positive discrimination;
- b_i is the difficulty parameter and represents how much a specific item i is hard to be responded correctly by respondents. So that when the ability of the respondent and the difficulty of the item are equal, the chances of getting it right are 50%;

[7] Respondent: an individual taking a test.

- c_i is the guessing parameter, representing the probability that a respondent gets a specific item i right randomly. It can also be understood as the probability that a respondent with low ability will get the item right. It is also the smallest possible chance that an item will be correct regardless of the estimated ability of the respondent.

Thus, once the item parameters are estimated and the hit probability is calculated using the equation 1, the item characteristic curve (ICC) can be obtained. The ICC defines the behavior of an item's hit probability curve according to the parameters describing the item (a_i, b_i e c_i) and the respondents' skill variance.

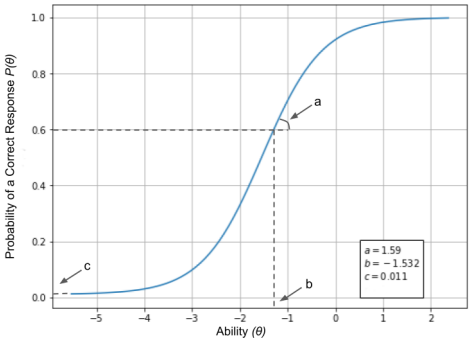

Fig. 1 Example of the representation of the parameter values of an item arranged on the Item Characteristic Curve - ICC. Source: The Authors.

As can be seen in figure 1, the hit probability on axis y is calculated by adding the values of the properties a_i, b_i e c_i found in an item and the variation of the skill θ. Given that each of these properties modifies the ICC in a way that makes it easier to see and understand how a particular item performs in the test.

Thus, the property a_i (discrimination) is responsible for the slope of the logistic curve; the property b_i (difficulty) plots the curve as a function of skill in the logistic function; and the property c_i (guessing) places the basis of the logistic function relative to the axis y.

The values found in the discrimination, difficulty and guessing properties are found for each instance and enable IRT to measure the ability θ of each respondent in the evaluation process, resulting in a rank of respondents arranged according to their skills. The higher the θ value the more skilled the individual.

As stated herein, the IRT emerges as a consolidated theory from the field of Psychometrics, which can be adapted to the field of Machine Learning - ML. Therefore, it is sufficient to consider that a test is a dataset, each test item is the instances of this dataset (independent variables are the questions and the dependent variable is the answer), each answer can be evaluated as right or wrong (in a dichotomous way, as advocated in the literature [19]), and each

respondent is a separate machine learning model (in this case, a large number of models is required).

That is, from this abstraction above, this important theory can be used to evaluate computational models, thus obtaining the evaluative benefits that IRT provides — which in general is related to the ability to analyze model complexity. Similar abstractions to this one have been made in machine learning research that uses this theory [41–44].

Item response theory currently has many applications in computing and its use goes beyond the abstractions presented above, such as [45–50]. It is worth mentioning that up to the time this article was written, the literature survey had not identified any research related to the use of IRT as an XAI measure to explain black box models based on attribute relevance ranking.

3 Materials and Methods

3.1 Benchmarck XAI

A benchmark was developed, figure 2, which performs all the comparative analyses required to assess the measure *eXirt* based on the other existing measures, because: in figure 2 (A) it uses tabular data already known and evaluated by the machine learning community; in figure 2 (B) it extracts properties from datasets; in the figure 2 (C) it groups the datasets according to their properties and performs multiple correspondence analyses; in the figure 2 (D)(E) it performs computational model building; in the figure 2 (F)(G) it generates ranks for all models, computes the correlations obtained from all pairs of ranks; and finally, in figure 2 (H)(I) it summarizes all the results in a boxplot. All these procedures are explained in detail in the following topics.

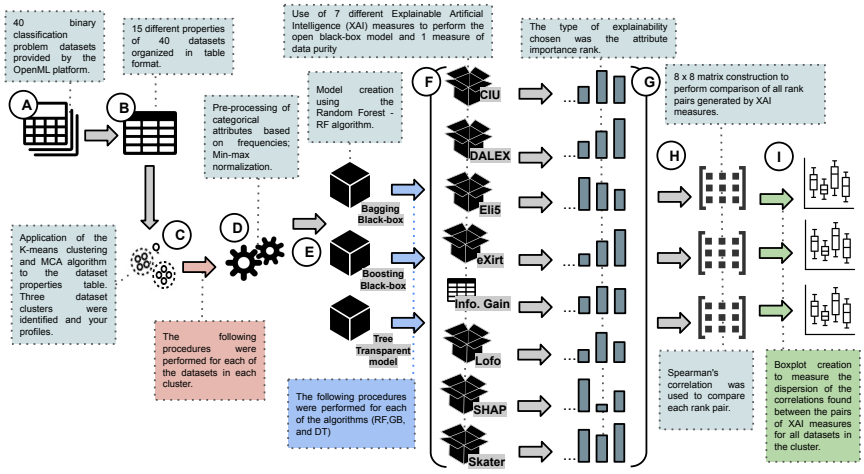

Fig. 2 Visual scheme of all steps and processes performed by the proposed benchmark.

3.1.1 Datasets and Preprocess

Different from what is presented in [17], where 41 datasets were used to run the benchmark, here 40 datasets were used, the reason for such reduction being related to dataset problems *"analcatdata lawsuit"* with measure *Shap*.

Despite the aforementioned elimination of the dataset in this new set of analyses, there were no significant impacts on the current results when compared to the results of the previous studies.

It is worth mentioning that the datasets concerned were selected from the *OpenML* [51] basis, while observing the fact that they refer to binary classification problems, without data loss and with a greater number of applications by the community, in order to ensure a standardization of the datasets used and also a better use of the data generated in the experiments performed here.

The datasets used were as follows: *australian, phishing websites, spec, satellite, banknote authentication, blood transfusion service center, churn, climate model simulation crashes, credit-g, delta ailerons, diabetes, eeg-eye-state, haberman, heart-statlog, ilpd, ionosphere, jEdit-4.0-4.2, kc1, kc2, kc3, kr-vs-kp, mc1, monks-problems-1, monks-problems-2, monks-problems-3, mozilla4, mw1, ozone-level-8hr, pc1, pc2, pc3, pc4, phoneme, prnn crabs, qsar-biodeg, sonar, spambase, steel-plates-fault, tic-tac-toe* and *wdbc*.

Increasing the number of datasets being analyzed was considered in order to identify new groups of datasets and also to further enrich benchmark execution results, but limitations of execution time and computational cost related to the internal processes of the execution environment being used impaired the execution of the benchmark.

3.1.2 Clustering

The process of analyzing different datasets based on their properties allows this research to compare the similarities and differences between them.

As noted above, this type of analysis is independent of the context in which the datasets are included, which facilitates the processing of the results of the herein study.

This research understands that, for example, by grouping a set of datasets based on values of their properties regarding class entropy and unbalances, relationships can be found between these dataset profiles and the explanations that can be extracted from models that use them.

Following the same idea above, by expanding the aforementioned analysis to a total of 15 different properties extracted from each of the 40 datasets, performing a clustering process becomes feasible and the identification of datasets having similar and different profiles is also possible.

The 15 dataset properties were used: *Number of Features, Number of Instances, Dimensionality, Percentage of Binary Features, Standard Deviation Nominal of Attribute Distinct Values, Mean Nominal Attribute Distinct Values, Class Entropy, Autocorrelation, Number of Numeric Features, Number of Symbolic Features, Number of Binary Features, Percentage of Symbolic*

Features, Percentage of Numeric Features, Majority Class Percentage, and *Minority Class Percentage*.

Thus, it used properties (provided by OpenML) extracted from each of the selected datasets and then used the k-means clustering algorithm [52] to identify dataset clusters based on their similarities.

Seeking to identify the optimal number of clusters to best separate the 40 datasets analyzed, the algorithm for interpretation and validation between data clusters was used, which is called Silhouettes [53], by varying the value K (clusters) between 2 and 10. In the end, $K = 3$ was found with average scores of *silhouette coefficient value* = 0.28, figure 3.

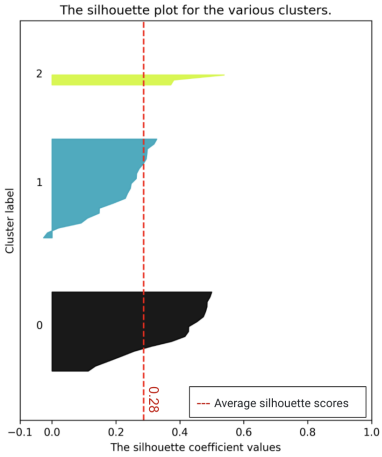

Fig. 3 Silhouette coefficients for clustering, using the Kmeans algorithm, for $K = 3$. Distance means (axis x) and label of clusters 0.1 and 2 (axis y).

Figure 3 shows the result of running the silluette algorithm, where it can be seen that for k=3 the distances between each cluster (0, 1 and 2) are above the average (red line), so this is an appropriate value of k.

The results that will be presented further on take into account only clusters 1 and 0, figure 3, due to their considerable number of datasets. Thus, analyses referring to the cluster 2 datasets were ignored, as that number of datasets is considered to be too low to perform any deeper analysis.

More information about the properties of all the datasets involved in this analysis can be found in the supplementary material item B.

3.1.3 Multiple Correspondence Analysis

In order to identify the relationships between the 15 properties of the datasets with each of the clusters identified in the clustering process, performing a Multiple Correspondence Analysis - MCA [36] on the properties datasets was decided, whereby the lines in this table are the observations or individuals (n) concerned — the datasets are here — and the columns are the different

categories of nominal variables (p) — here are the properties of each dataset. In this analysis, the label of the cluster each dataset belongs to was taken into consideration.

First, the binarization process was performed on the same table where the clustering was performed, replaced by h (equal to or above the average of the attribute values) and s (below the average of the attribute values). For more information about binarization of the dataset property table, see the supplementary material item C.

MCA provided important analyses in the herein study, since based on its results it was possible to identify the profile of datasets belonging to each cluster. Thus, once it is possible to identify which value ranges of specific properties each dataset is most related to, it is possible to identify the profile of each cluster, as recommended by the literature [36].

3.2 The *eXirt* Measure

The measure Explainable based on Item Response Theory - *eXirt* is one of the XAI measures performed in the developed benchmark. This measure is a new proposal to generate explanations for tree-ensemble models that is based on the generation of attribute relevance ranks by using item response theory.

Just like other XAI measures, *eXirt* only uses the training data, test data, the model itself together with its outputs, figure 4 (A). Initially, the test data and the model to be explained are passed on to the *eXirt*; 4 (B) the tool then creates the so-called "loop of models," shown in figure 4 (C) and (D), which is a basic iterative process relevant for the creation of the table with the respondents' answers used in the IRT run, figure 4 (E) and (F).

The "loop of models" process, figure 4 (D) is inspired on how the measures XAI Dalex [12] and Lofo [14] work, for in this process iterative variations are performed on different sets of attributes of the model, and at each iteration the answers of the classifier prediction are collected and, thus, it is possible to have a high number of responding candidates. It is noteworthy that each classifier is different from the other, as one presents different inputs and one or more attributes of their inputs.

Still referring to the "loop of models", 12 different ways of varying the attributes of the models have been implemented, figure 4 (C). An important detail is that variations can be combined, that is, they can occur in more than one attribute of the model in the same iteration. However, the standard use of 2 variations and combinations of up to 2 in parallel is indicated, so as to reduce the computational cost for datasets with high amounts of attributes.

The implemented variations are nothing more than different types of noise and interference that are inserted into the input values of the model, test set, such as: permutation of the indices, application of noise, modification of value to 0, application of normalization to scale different from that in the data, ordering of values and their indices in an ascending manner, ordering of values and their indices in a descending manner, inversion of index positions (from top to bottom), binning processing of values, multiplication of values by -1,

replacement of values by the mean, replacement of values by the standard deviation, and standardization of values.

The layout in figure 4 (D) shows several computational models only for ludic reasons, as only one computational model is used, varying its inputs iteratively.

In the results presented, only multiplication by -1 and the application of binning processes were used, as they have low computational costs.

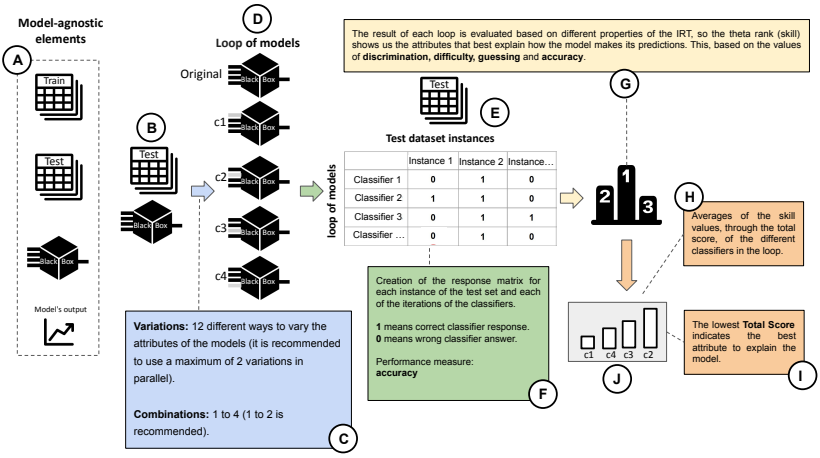

Fig. 4 Visual summary of all steps and processes performed by the XAI measure, called *eXirt*.

The creation of the response matrix, figure 4 (E) and (F), contains the answers of all the classifiers (respondents). The columns refer to the instances of the dataset that was passed on, while the rows refer to the different classifiers. Values equal to 0 (zero) are wrong answers of the prediction, while values 1 (one) are correct answers of the prediction, regardless of the number of classes that the problem may have. This matrix is used to calculate the values of the item parameters (discrimination, difficulty and guessing) for each of the instances, in figure 4 (G).

The implamentation of the IRT used was [47], called decodIRT, in a code developed exclusively for the purpose of this paper, as the code first receives the answer matrix, performs the calculations to generate the item parameter values — different algorithms can be used to calculate the IRT (such as: ternary, dichotomous, fibonacci, golden, brent, bounded or golden2) [54] — and, after this step, generates the rank of most skilled classifiers, figure 4 (G).

Among the new features of *decordIRT* is a new score calculation that involves the calculated ability of all respondents and their respective hits and misses, called Total Score. Total Score can be understood as an adaptation of the True-Score [55], whereby the score is calculated by summing up all the hit probabilities for the test items. However, in cases where respondents have

a very close ability, the True-Score result can be very similar or even equal, since only the hit chance is considered. To avoid equal score values and to give more robustness to the models' final score, the Total Score also considers the respondent's probability of error, given by: $1 - P(U_{ij} = 1|\theta_j)$. Thus, every time the model gets it right, the hit probability is added, and if the model gets it wrong, the error probability is subtracted. The calculation of the Total Score t_l is defined by the following equation 2, where i' corresponds to the set of items answered correctly, and i'' corresponds to the set of items answered incorrectly.

$$t_l = \sum_{i=1}^{i'} P(U_{ij} = 1|\theta_j) - \sum_{i=1}^{i''} 1 - P(U_{ij} = 1|\theta_j) \tag{2}$$

In this regard, a skilled model with high hit probability that ends up getting an item wrong will not have its score heavily discounted. However, for a low ability model with low hit probability, the error will result in a greater discount of the score value. For, it is understood that the final score value should consider both the estimated ability of the respondent and his/her own performance on the test.

The Total Score resulting from the execution of decodIRT is not yet the final rank of explainability of the model, because in this case it is necessary to calculate the average of the skills found for each attribute, figure 4 (H), involving the different variations and combinations of attributes used in the previous steps.

Ultimately, figure 4 (I) and (J), an explanation rank is generated where each attribute appears with a skill value. In this case, the lower the ability values, the more the attribute explains the analyzed model. Equation $T_{(f,r)}$ is presented, which represents the processes performed by *eXirt*, equation 3.

$$T_{(f,r)} = \sum_{j=1}^{(v*f)+(c^f)+1} e_j \sum_{i=1}^{r} p_{ji} + \sum_{l=1}^{f} t_l \tag{3}$$

Where v and c refer respectively to the number of variations and combinations used. The value of j represents the respondent's index of the loop of models. While f represents the total input attributes of the model, where $(v * f) + (c^f) + 1$ the total number of respondents with interference (variations and combinations) along with the respondent without interference (original model).

While e_j represents the process of building the response matrix, which is used in the following iterative process, where i represents the item index (respondent's answer) and r is the quantity of items to be considered in the calculation of the p_{ji}, where j is the analyzed respondent and i is the analyzed instance, presented earlier in equation 1 through the calculation of $P(\theta_j)_i$.

Finally, there is the iterative process for calculating the Total Score in t_l, where f is the quantity of input attributes of the model.

Based on the above, it can be seen that the execution of the measure *eXirt* depends directly on the number of respondents in the "loop of model" and also on the number of instances in the dataset. For this reason, we suggest using low values for variation and combination of attributes, as shown in figure 4 (C). While the IRT implementation used by *eXirt* has the same characteristics and limitations as seen in [56].

As an example of the ranking generated at the end of the eXirt application, the results collected for a random dataset chosen from the 40 datasets used in this study is shown below, figure 5.

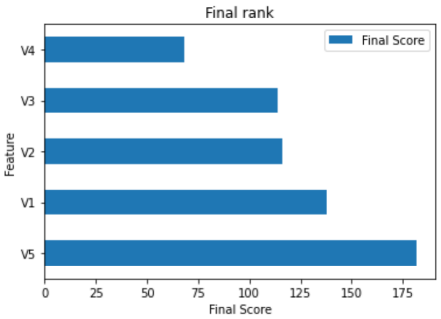

Fig. 5 Result of application of the *eXirt* in a Random Forest algorithm properly trained for the phoneme dataset.

In figure 5, the attributes that best explain the model are presented with the highest values in y and the lowest values in x.

3.3 Data Purity Measure

The Information Gain, also known as data purity measure, despite not being an attribute relevance measure, has an interesting place in the analyses and was one of the ranks included in the benchmark, since the main idea of including this attribute importance measure together with the other XAI measures is to analyze what correlations exist between the ranking of the attributes that explain (by using XAI measures) the model and a ranking of the attributes that provide a better performance [34].

This measure, information gain, has the capacity described above, since it generally represents how much is gained in "purity" by dividing a set according to an attribute, since its base calculation is based on entropy [57].

There are many other measures of attribute importance that could be included in the benchmark for the same purpose as described above, such as Out-of-Bag importance rank [58] or even the Gini Index [59], for example. However, only information gain was chosen to be used due to issues related to computational cost.

3.4 Ranks Correlations

For all 40 datasets (clustered into 3 different clusters), 2 machine learning models were created, based on the Random Forest and Gradient Boosting algorithms. In all, a total of 80 models were generated.

The overall explainability ranks were produced by the 8 measures used in each model, resulting in a total of 640 ranks.

In order to calculate the correlation between each pair of generated rankings, the rank correlation Spearman Rank [60] was selected. The reason for using this particular algorithm is that it measures the correlation between pairs of ranks considering the idea of ranks (positions) where different values (in this case, dataset attributes) may appear.

In this step, 2 rank correlation pair comparison matrices are generated for each dataset (one matrix for each algorithm). Figure 6 shows examples of correlation matrices created from the RF and GB models, along with the datasets *spambase* and *jEdirt_4.0_4.2*.

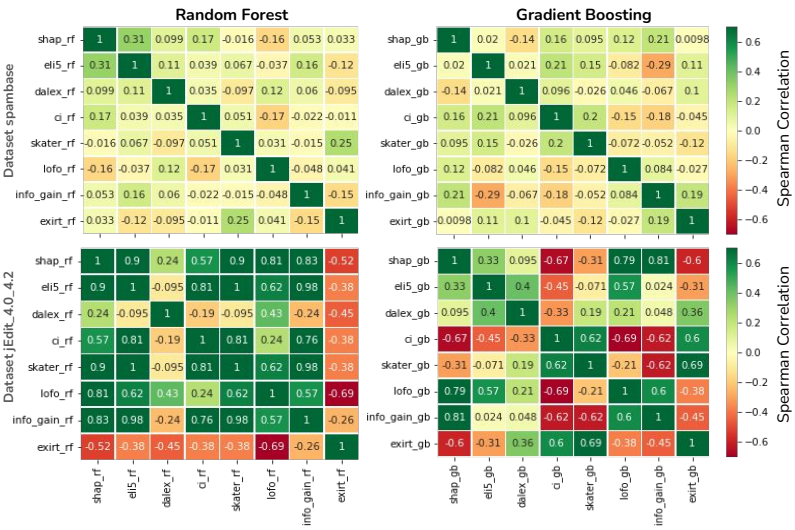

Fig. 6 Example of the results of all correlations calculated between the pairs of ranks generated for the RF and GB models, referring to the dataset phoneme.

By observing the correlations presented in figure 6, even in this stage of the analyses, it can be seen that there are greater disagreements — correlations closer to 0 (zero) — between the explanations created from models that use the dataset *spambase* and greater concordances — correlations closer to ±1 — between the explanations created from models using the dataset *jEdirt_4.0_4.2*. In both cases, there were minimal differences — but they do exist — in the results of the different algorithms being analyzed.

Seeking to expand on the findings by means of the analyses shown in the figure 6, more comprehensive results involving all the models created are presented below.

4 Results and Discussion

The collected results are divided into two parts (one for each cluster), and each of these parts is subdivided into two other parts (one for each algorithm tested). The purpose of having these analyses performed separately is to make it possible to check how the results of *eXirt* correlate with the results of other XAI measures and the data purity measure.

The first step to better understand the results is to understand the clustering and multiple correspondence analysis processes being used.

As previously stated, the clustering process found 3 clusters of datasets, but only clusters 0 and 1 were considered by this research, since they had considerable amounts of datasets:

- **Datasets of cluster 0:** *ionosphere, wdbc, credit-g, churn, Australian, eeg-eye-state, heart-statlog, ilpd, tic-tac-toe, jEdit-4.0-4.2, diabetes, prnn-crabs, monks-problems-1, monks-problems-3, monks-problems-2, delta-ailerons, mozilla4, phoneme, blood-transfusion-service-center, banknote-authentication, and haberman*;
- **Datasets of cluster 1:** *ozone-level-8hr, sonar, spambase, qsar-biodeg, kc3, mc1, pc3, mw1, pc4, Satellite, pc2, steel-plates-fault, kc2, pc1, kc1, and climate-model-simulation-crashes*;
- **Datasets of cluster 2:** *kr-vs-kp, PhishingWebsites, and SPECT.*

In order to consolidate the cluster profile analysis, a Multiple Correspondence Analysis - MCA was performed, as advocated in the literature [36], and the relationship between the datasets in each cluster and the value ranges of the 15 properties analyzed was verified . This provided a better understanding of the complexity of the datasets in both clusters.

To simplify this analysis, each range of property values was defined as above average (symbol h) and belo average (symbol s), figure 7.

Based on the inspection of the MCA result, shown in figure 7, one can notice the larger relations (smaller distances) of the datasets (small gray dots) belonging to the clusters (cluster 0 in blue and cluster 1 in red) with the different value ranges of the 15 analyzed properties (larger colored circles). Thus, it can be seen that each cluster is composed of:

- **Cluster 0:** datasets with above average property values for *ClassEntropy, PercentageOfSymbolocFeatures, PercentageOfBinaryFeatures, NumberOfBinaryFeatures, MinorityClassPercentage*, and *PercentageOfSymbolicFeatures*. And also below average values for *AutoCorrelation, NumberOfFeatures, PercentageOfNumericFeatures, NumberOfNumericFeatures, NumberOfInstances, Dimensionality*, and *MajorityClassPercentage*.

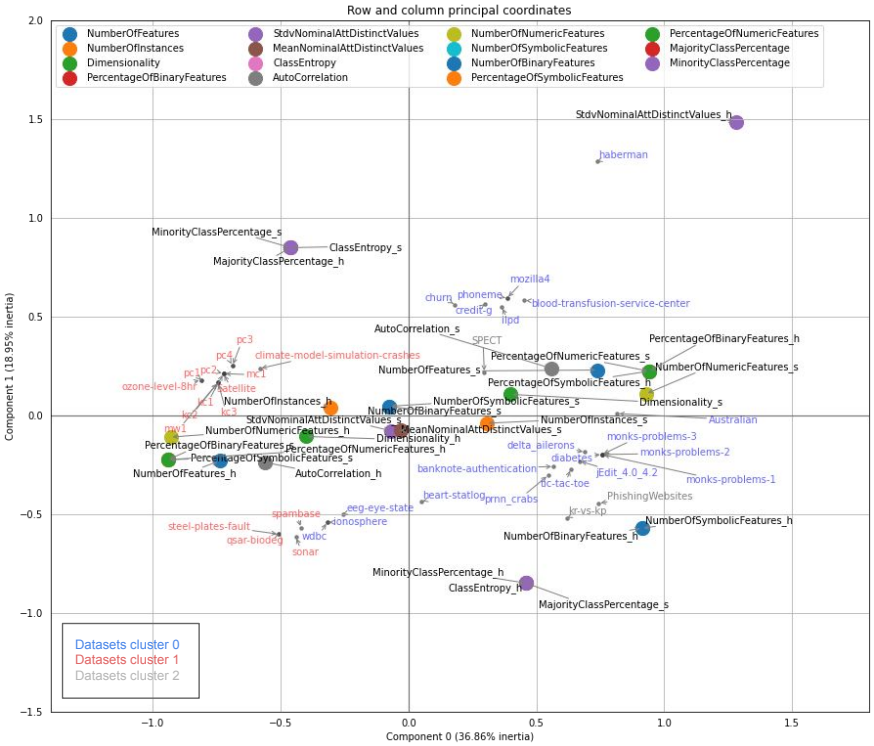

Fig. 7 Multiple Correspondence Analysis - MCA with rows (data sets) and columns (properties) the axis x and y are respectively the 0 and 1 components of the analysis.

- **Cluster 1:** datasets with above average property values for *MajorityClassPercentage, Dimensionality, NumberOfInstances, NumberOfFeatures, PercentageOfNumericFeatures, AutoCorrelation*, and *NumberOfNumericFeatures*. And also below average values for *ClassEntropy, PercentageOfSymbolicFeatures, MinorityClassPercentage*, and *PercentageOfBinaryFeatures*.

It is worth noting that some properties were not mentioned above because they appear at very similar distances for the two clusters.

Based on the relationships of the datasets belonging to each cluster with their value ranges of the 15 properties, it is assumed that the datasets belonging to cluster 0 are less complex and the datasets in cluster 1 are more complex.

It is understood that the explanation complexity of a model (dataset and algorithm) cannot only be evaluated by its dataset, because the capacity of the algorithm to generalize a given dataset must also be taken into account. In this sense, a benchmark was run for the two clusters of datasets, for the Random Forest and Gradient Boosting.

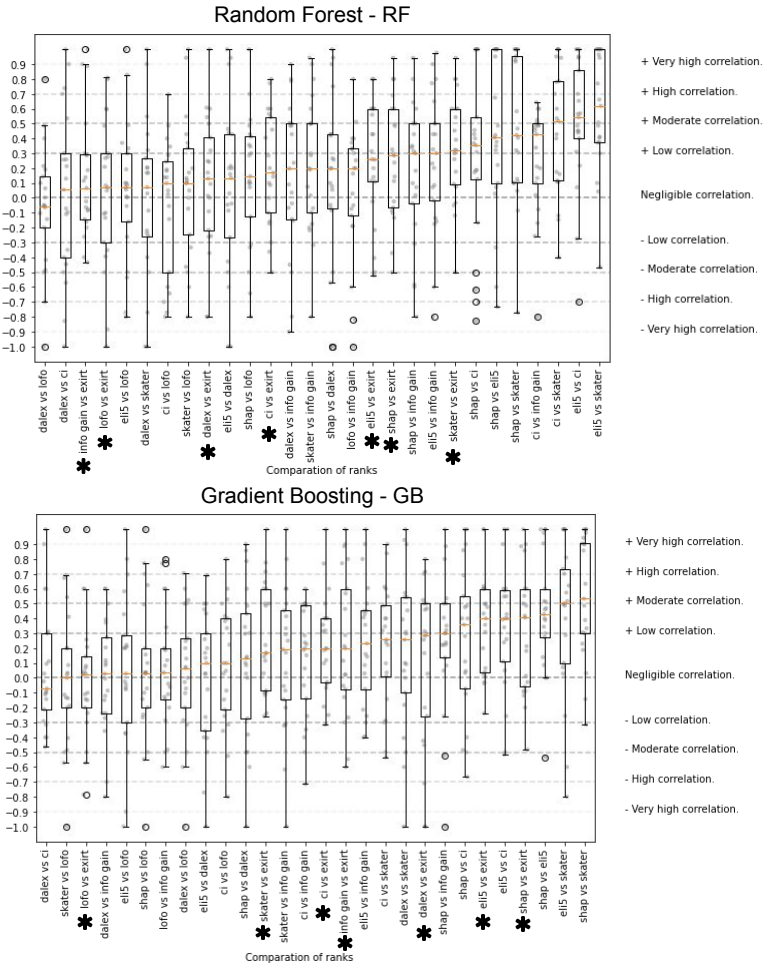

Fig. 8 Summary of correlations found between ranks of the models created from the less complex datasets. O '*' is a highlight for comparisons with the eXirt.

The benchmark run for the cluster 0 (less complex) datasets, figure 8, shows results with various correlation values, ranging from negligible correlations to high correlations — mostly positive for the two algorithms tested.

The results presented in figure 8 show evidence that XAI measures, applied to tree-ensemble-based models created from less complex data, tend to agree with each other and thus show higher correlations in comparisons of pairs of attribute relevance ranks, these models being considered the easiest to explain. These results are similar to those found in [17], even when adding the ranks from the *eXirt* and also the Information Gain.

Despite the high scatter and the presence of outliers in the boxplots presented at this stage of the analysis, it can be noticed in figure 8 highlight '*' tht the ranks generated by measure *eXirt* are considerably different from the

ranks of the other XAI measures and the Information Gain rank. For, at the positive or negative "high correlation" level, no central quartile boxplot was verified.

The result above indicates that even for models that are easier to explain — that show higher correlations between the results of XAI measures — *eXirt* is able to generate attribute relevance ranks different from those that already exist in the literature.

The result of the benchmark run for the cluster 1 datasets (more complex) is shown in figure 9.

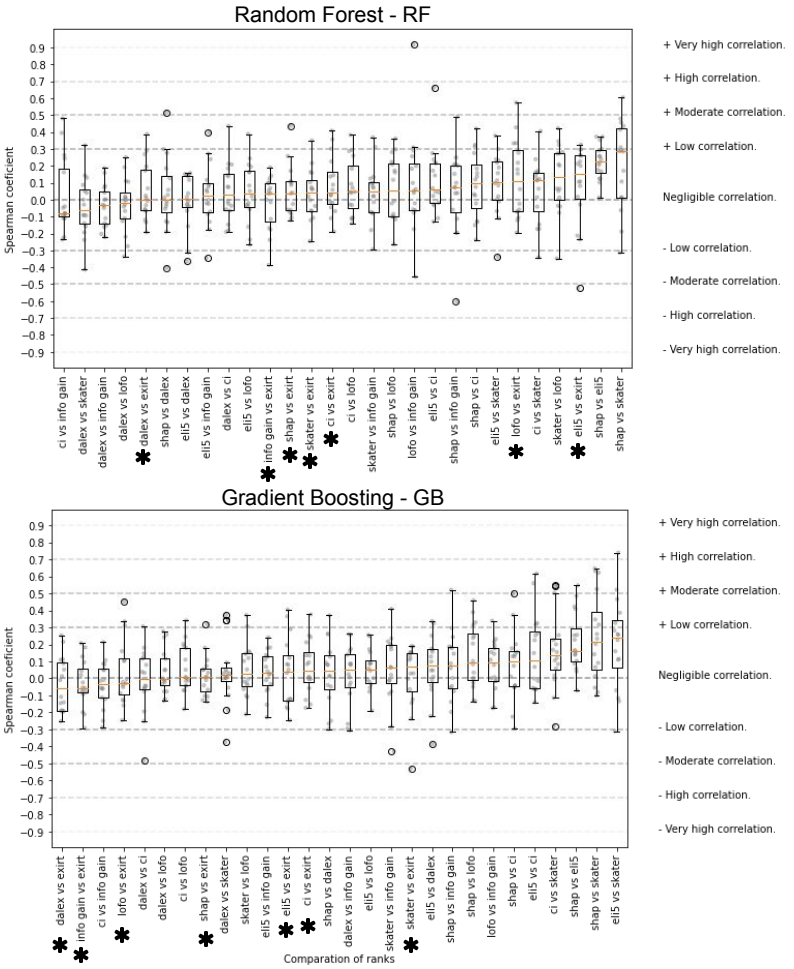

Fig. 9 Summary of correlations found in all runs performed - Cluster with more complex datasets.

The benchmark run for the datasets belonging to cluster 1 figure 9 show lower correlation values, with a fewer outliers and less scattering (compared to the previous experiment).

Based on these facts, figure 9 shows evidence that XAI measures, applied to tree-ensemble-based computer models created from more complex data, tend to disagree with each other and thus show lower correlations in comparisons of pairs of attribute relevance ranks — note the concentration of insignificant correlations found. These models are considered to be the most difficult to explain. These results are similar to those found in [17], even when adding the ranks from *eXirt* as well as the Information Gain.

Based on the results shown in figure 9 highlight '*', it can be said that the ranks gnerated by *eXirt* measures are considerably different from the ranks of the other XAI measures and the Information Gain rank, since said ranks were not found in any of the comparisons involving *eXirt* with resulta greater than the positive or negative "high correlation" level.

One way to observe how *eXirt* applies the IRT and how to relate the properties of discrimination, difficulty and guessing to the explanations of the easiest and most difficult models to explain is through detailed analysis of the item parameter values generated, figure 10.

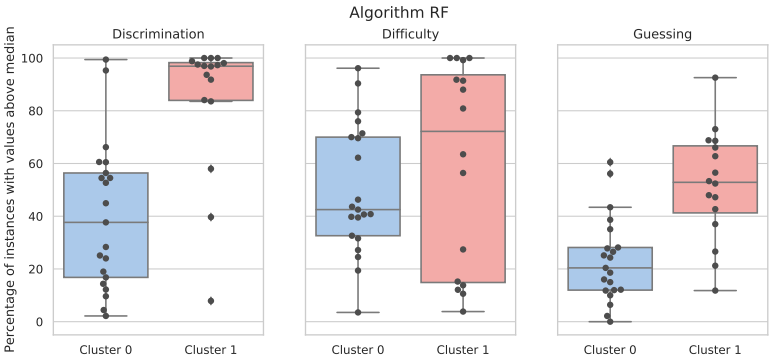

Fig. 10 Comparison of percentages of instances by item parameter values. Results only from the Random Forest algorithm.

In order to create figure 10, first the averages of item parameter values for each of the datasets and for each parameter were calculated — regardless of which cluster it belongs to — then, the mean of all averages calculated for each parameter, called thresholds, was calculated. Finally, the percentage of instances with item parameter values above or equal to these thresholds was calculated — for each of the clusters analyzed separately.

Therefore, it can be observed in figure 10 (Discrimination), by means of the perspective of the Random Forest algorithm, that the datasets belonging to the models in cluster 1 have a considerable number of very discriminative instances when compared to the datasets of the models in cluster 0. This means that the

datasets of the models in cluster 1 exhibit instances that discriminate between good and bad models in a more significantly manner, thus showing indications that models, which are difficult to explain, may show high discrimination of instances.

Following the same idea, figure 10 (Difficulty) contradicts expectations, as it is observed that most cluster 1 datasets had the lowest Difficulty percentages along with the highest spreads, when compared to cluster 0 datasets, thus showing evidence that difficult-to-explain models will not necessarily show the highest values of difficulty on the part of IRT.

Regarding figure 10 (Guessing), it can be seen that the datasets in cluster 1 have a high number of instances with considerable guess values compared to the values presented by the datasets in cluster 0. This means that the Random Forest algorithm used for generalizing the datasets hit a high number of random instances for cluster 1, thus showing indications that difficult-to-explain models may have high guessing values by IRT, which makes all logical sense, since explaining a model that hits predictions by chance should not be an easy task for any XAI measure.

It is noteworthy that the global explanations generated by *eXirt*, along with analyses of the item parameters generated for each model, allow the end user to assess their confidence in the explanation being generated. Since, as presented, the combination between the different properties of *IRT* shows indications and quantitative information regarding whether the analyzed tree-ensemble model is easy or difficult to explain.

All observations made based on figure 10 can be summarized in figure 11, since the latter represents the median of the mean values of the characteristic curves of the items in the models belonging to datasets 0 and 1. The ICC, despite usually being an instance-level analysis, can characterize and differentiate the models belonging to the two clusters, since, in y the average hit probability of the algorithms increases as the skill also increases, but in different proportions for the two clusters.

Therefore, by observing figure 11, where the left side of the figure shows greater median of guessing value for cluster 1 models (red line above the blue line), the central part of the figure shows higher median discrimination for

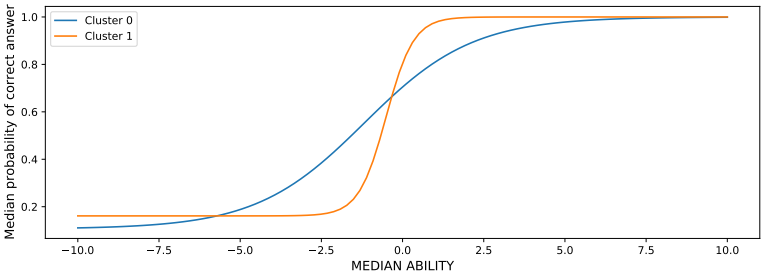

Fig. 11 Median of mean item parameter values for cluster 0 and cluster 1 models. Results for RF-based models.

models in cluster 1 (steeper slope of the curve in red) and the right side of the figure shows that the median of difficulty values of models belonging to cluster 0 is higher than those belonging to cluster 1, since for the ICC of cluster 0 to reach the maximum of "Median probability of corret answer" (axis y) a greater value of "Median Ability" (axis x) is required.

Figures 10 and 11 were selected for showing only the results for the Random Forest algorithm, as these are very similar to those faound by the Gradient Boosting algorithm.

Broadly speaking, the results presented by figures 8, 9, 10 and 11 show new evidence and insights into how explanations generated by XAI measures correlate to complex properties of the machine learning models being used.

Considering the evaluation processes carried out by the *eXirt* as an important differential of this measure, and it can be stated that the evaluation by means of the three IRT properties enables a comprehensive analysis of model complexity, since the results show evidence of the relations of difficult and easy models to be explained with different combinations of the values of the discrimination, difficulty and guessing properties of the IRT *eXirt*. Therefore, it can be stated that *eXirt* uses a different model evaluation method than the XAI measures based on "Permutation of Feature", such as *Dalex*, *Eli5*, *Lofo*, *Shap* and *Skater*.

It is not possible to say whether the ranks generated by *eXirt* are better or worse than the ranks generated by the other measures based on the context of each problem, since to making a statement of this nature would require the inclusion of human experts for each of the problems of the datasets analyzed, which is beyond the initial proposal of the work presented herein.

The results also show that the attribute relevance ranks are considerably different from the attribute importance ranks calculated by using Information Gain, both for easier models to explain and for the more difficult ones, figure 8 and figure 9.

It is clear that the explanations generated by the *eXirt* measure are different in the considerable majority of cases from the explanations obtained by the XAI measures found in the literature today, even in situations where the other XAI measures tend to show higher correlations among themselves. This shows that the use of Item Response Theory has enabled the *eXirt* to explain tree-ensemble models by creating ranks not yet explained by the current literature.

5 Final Considerations

In view of all the analyses performed, this research achieves its goal by presenting an innovative proposal for measuring XAI, called *eXirt*, which is capable of performing the explanation process of tree-ensemble machine learning models in a different way from the measures available in the literature. This research also provided analyses of the impacts of model complexity on their explanations by presenting results that complement previous research and enhance

evidence for the existence of easier and more difficult tree-ensemble-based models, as well as the identification of the properties that the data and algorithms of these models exhibit. However, more studies still need to be conducted, aiming at a better identification and separability between these two groups of models.

6 Future works

- Adapting the *eXirt* methodology to generate global and local explanations of models to be analyzed.
- Conduct an in-depth, robust investigation of the explanations generated by the *eXirt* measure in comparison with the explanations of the other XAI measures, intending to highlight the types of models and their properties that directly influence the explanations of the *eXirt* and making its explanations more different or more similar to the explanations of the other measures.
- Develop an interface for the *eXirt* seeking the interaction of the human with its explanations, thus enabling the creation of a collaborative explanation between man and machine.
- Expand the analyses with *eXirt* for other types of algorithms, which are not exclusively tree-ensemble by analyzing the potential of this measure to become Model Agnostic, since the whole methodology proposed by *eXirt* is compatible with the nature of this type of measure.

Funding

This research was not funded.

Conflicts of interest/Competing interests

It is declared that there are no conflicts of interest between the authors and their institutions belonging to any part of this research.

Ethics approval

Not applicable.

Consent to participate

Not applicable.

Consent for publication

All individuals and institutions involved in the research in question are in agreement with the publication of this article in the journal.

Availability of data, code and material (Supplementary information)

A - General repository for benchmark reproducibility:

- https://github.com/josesousaribeiro/eXirt-XAI-Benchmark;

B - Properties of normal value datasets:

- https://github.com/josesousaribeiro/eXirt-XAI-Benchmark/blob/main/df_dataset_properties.csv;

C - Properties of binary value datasets:

- https://github.com/josesousaribeiro/eXirt-XAI-Benchmark/blob/main/df_properties_binarized.csv;

D - Repository for reproducibility of the analysis of the item parameter value files of all generated models:

- https://github.com/josesousaribeiro/eXirt-XAI-Benchmark/blob/main/eXirt%20-%20Notebook%20-%20v0.1.ipynb

References

[1] Shalev-Shwartz, S., Ben-David, S.: Understanding Machine Learning: From Theory to Algorithms. Cambridge university press, ??? (2014)

[2] Ghahramani, Z.: Probabilistic machine learning and artificial intelligence. Nature **521**(7553), 452–459 (2015)

[3] Maclin, R., Opitz, D.: An empirical evaluation of bagging and boosting. AAAI/IAAI **1997**, 546–551 (1997)

[4] Haffar, R., Sánchez, D., Domingo-Ferrer, J.: Explaining predictions and attacks in federated learning via random forests. Applied Intelligence, 1–17 (2022)

[5] Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion **58**, 82–115 (2020). https://doi.org/10.1016/j.inffus.2019.12.012

[6] Zhang, X., Meng, F.: A large-scale group decision making method to select the ideal mobile health application for the hospital. Applied Intelligence, 1–21 (2022)

[7] Hernández-Pereira, E., Fontenla-Romero, O., Bolón-Canedo, V., Cancela-Barizo, B., Guijarro-Berdiñas, B., Alonso-Betanzos, A.: Machine learning techniques to predict different levels of hospital care of covid-19. Applied Intelligence **52**(6), 6413–6431 (2022)

[8] Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.-I.: From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence **2**(1), 56–67 (2020). https://doi.org/10.1038/s42256-019-0138-9. Number: 1 Publisher: Nature Publishing Group. Accessed 2021-05-24

[9] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM computing surveys (CSUR) **51**(5), 1–42 (2018)

[10] Gunning, D., Aha, D.: DARPA's Explainable Artificial Intelligence (XAI) Program. AI Magazine **40**(2), 44–58 (2019). https://doi.org/10.1609/aimag.v40i2.2850. Number: 2. Accessed 2021-05-23

[11] Främling, K.: Decision theory meets explainable ai. In: International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems, pp. 57–74 (2020). Springer

[12] Burzykowski, P.B.a.T.: Explanatory Model Analysis. https://ema.drwhy.ai/ Accessed 2021-05-14

[13] Korobov, M., Lopuhin, K.: Eli5. https://eli5.readthedocs.io/en/latest/index.html. Last accessed 21 Jan 2021.

[14] Çayır, U., Yenidoğan, I., Dağ, H.: Use case study: Data science application for microsoft malware prediction competition on kaggle. Proceedings Book, 98 (2019)

[15] Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.-I.: From local explanations to global understanding with explainable ai for trees. Nature Machine Intelligence **2**(1), 2522–5839 (2020)

[16] Skater. https://oracle.github.io/Skater/overview.html{#}skater. Last accessed 21 Jan 2021.

[17] Ribeiro, J., Silva, R., Cardoso, L., Alves, R.: Does dataset complexity matters for model explainers? In: 2021 IEEE International Conference on Big Data (Big Data), pp. 5257–5265 (2021). https://doi.org/10.1109/BigData52589.2021.9671630

[18] Molnar, C.: Interpretable Machine Learning. Lulu. com, ??? (2020)

[19] de Andrade, D.F., Tavares, H.R., da Cunha Valle, R.: Teoria da resposta ao item: conceitos e aplicações. ABE, Sao Paulo (2000)

[20] Qi, Z., Khorram, S., Li, F.: Visualizing deep networks by optimizing with integrated gradients. In: CVPR Workshops, vol. 2 (2019)

[21] Kindermans, P.-J., Schütt, K.T., Alber, M., Müller, K.-R., Erhan, D., Kim, B., Dähne, S.: Learning how to explain neural networks: Patternnet and patternattribution. In: International Conference on Learning Representations (2018). https://openreview.net/forum?id=Hkn7CBaTW

[22] Linardatos, P., Papastefanopoulos, V., Kotsiantis, S.: Explainable ai: A review of machine learning interpretability methods. Entropy **23**(1) (2021). https://doi.org/10.3390/e23010018

[23] Ethical ML. https://github.com/EthicalML/awesome-production-machine-learning#explaining-black-box-models/and-datasets. Last accessed 20 Apr 2021.

[24] Biecek, P.: Dalex: Explainers for complex predictive models in r. Journal of Machine Learning Research **19**(84), 1–5 (2018)

[25] Apley, D.W., Zhu, J.: Visualizing the effects of predictor variables in black box supervised learning models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **82**(4), 1059–1086 (2020)

[26] Ribeiro, M.T., Singh, S., Guestrin, C.: "why should I trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016, pp. 1135–1144 (2016)

[27] for Ethical AI, T.I., Learning, M.: Ethical XAI. Last accessed 22 Jul 2021. (2021). https://ethical.institute/xai.html Accessed 2021-05-14

[28] Arya, V., Bellamy, R.K., Chen, P.-Y., Dhurandhar, A., Hind, M., Hoffman, S.C., Houde, S., Liao, Q.V., Luss, R., Mojsilovic, A., *et al.*: Ai explainability 360: An extensible toolkit for understanding data and machine learning models. J. Mach. Learn. Res. **21**(130), 1–6 (2020)

[29] Nori, H., Jenkins, S., Koch, P., Caruana, R.: Interpretml: A unified framework for machine learning interpretability. arXiv preprint arXiv:1909.09223 (2019)

[30] Keeney, R.L., L, K.R., Howard, R.: Decisions with Multiple Objectives:

Preferences and Value Trade-Offs, Revised ed. edição edn. Cambridge University Press, Cambridge England ; New York, NY, USA (1993)

[31] Baniecki, H., Kretowicz, W., Piatyszek, P., Wisniewski, J., Biecek, P.: dalex: Responsible Machine Learning with Interactive Explainability and Fairness in Python. arXiv:2012.14406 (2020)

[32] Roth, A.E.: The Shapley Value: Essays in Honor of Lloyd S. Shapley. Cambridge University Press, ??? (1988)

[33] Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 4768–4777 (2017)

[34] Reza, F.M.: An Introduction to Information Theory. Courier Corporation, ??? (1994). Google-Books-ID: RtzpRAiX6OgC

[35] Oreski, D., Oreski, S., Klicek, B.: Effects of dataset characteristics on the performance of feature selection techniques. Applied Soft Computing **52**, 109–119 (2017)

[36] Abdi, H., Valentin, D.: Multiple correspondence analysis. Encyclopedia of measurement and statistics **2**(4), 651–657 (2007)

[37] Pasquali, L., Primi, R.: Fundamentos da teoria da resposta ao item: TRI. Avaliação Psicológica: Interamerican Journal of Psychological Assessment **2**, 99–110 (2003)

[38] Hambleton, R.K., Swaminathan, H., Rogers, H.J.: Fundamentals of Item Response Theory vol. 2. Sage, ??? (1991)

[39] Kreiner, S.: The rasch model for dichotomous items. Rasch models in health, 5–26 (2012)

[40] Birnbaum, A.L.: Some latent trait models and their use in inferring an examinee's ability. Statistical theories of mental test scores (1968)

[41] Prudêncio, R.B., Hernández-Orallo, J., Martınez-Usó, A.: Analysis of instance hardness in machine learning using item response theory. In: Second International Workshop on Learning over Multiple Contexts in ECML (2015)

[42] Martínez-Plumed, F., Prudêncio, R.B., Martínez-Usó, A., Hernández-Orallo, J.: Making sense of item response theory in machine learning. In: ECAI 2016, pp. 1140–1148. IOS Press, ??? (2016)

[43] Martínez-Plumed, F., Prudêncio, R.B., Martínez-Usó, A., Hernández-Orallo, J.: Item response theory in ai: Analysing machine learning

classifiers at the instance level. Artificial intelligence **271**, 18–42 (2019)

[44] Kline, A.S., Kline, T.J., Lee, J.: Item response theory as a feature selection and interpretation tool in the context of machine learning. Medical & Biological Engineering & Computing **59**(2), 471–482 (2021)

[45] Chang, T.-Y., Shiu, Y.-F.: Simultaneously construct irt-based parallel tests based on an adapted clonalg algorithm. Applied Intelligence **36**(4), 979–994 (2012)

[46] Gan, W., Sun, Y., Peng, X., Sun, Y.: Modeling learner's dynamic knowledge construction procedure and cognitive item difficulty for knowledge tracing. Applied Intelligence **50**(11), 3894–3912 (2020)

[47] Cardoso, L.F., Santos, V.C., Francês, R.S.K., Prudêncio, R.B., Alves, R.C.: Decoding machine learning benchmarks. In: Brazilian Conference on Intelligent Systems, pp. 412–425 (2020). Springer

[48] Martínez-Plumed, F., Prudêncio, R.B., Martínez-Usó, A., Hernández-Orallo, J.: Item response theory in ai: Analysing machine learning classifiers at the instance level. Artificial intelligence **271**, 18–42 (2019)

[49] Martínez-Plumed, F., Prudêncio, R.B., Martínez-Usó, A., Hernández-Orallo, J.: Making sense of item response theory in machine learning. In: ECAI 2016, pp. 1140–1148. IOS Press, ??? (2016)

[50] Bergner, Y., Droschler, S., Kortemeyer, G., Rayyan, S., Seaton, D., Pritchard, D.E.: Model-based collaborative filtering analysis of student response data: Machine-learning item response theory. International Educational Data Mining Society (2012)

[51] OpenML https://www.openml.org/search?q=qualities.NumberOfClasses%3A2%2520qualities.NumberOfMissingValues%3A0&type=data&sort=runs&order=desc. Last accessed 01 Mar 2021.

[52] KMeans. https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html. Last accessed 2 Mar 2021.

[53] Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics **20**, 53–65 (1987)

[54] Magis, D., Raîche, G.: Random generation of response patterns under computerized adaptive testing with the r package catr. Journal of Statistical Software **48**, 1–31 (2012)

[55] Lord, F.M., Wingersky, M.S.: Comparison of irt true-score and equipercentile observed-score" equatings". Applied Psychological Measurement **8**(4), 453–461 (1984)

[56] Martínez-Plumed, F., Prudêncio, R.B., Martínez-Usó, A., Hernández-Orallo, J.: Item response theory in ai: Analysing machine learning classifiers at the instance level. Artificial intelligence **271**, 18–42 (2019)

[57] Kent, J.T.: Information gain and a general measure of correlation. Biometrika **70**(1), 163–173 (1983)

[58] Breiman, L.: Out-of-bag estimation (1996)

[59] Lerman, R.I., Yitzhaki, S.: A note on the calculation and interpretation of the gini index. Economics Letters **15**(3-4), 363–368 (1984)

[60] Artusi, R., Verderio, P., Marubini, E.: Bravais-Pearson and Spearman Correlation Coefficients: Meaning, Test of Hypothesis and Confidence Interval. The International Journal of Biological Markers **17**(2), 148–151 (2002). Publisher: SAGE Publications Ltd STM. Accessed 2021-05-15