Revista
Portuguesa
de Educação

The Reasons why the Regression Tree Method is more suitable than General Linear Model to analyze complex educational datasets

ABSTRACT

Any quantitative method is shaped by certain rules or assumptions which constitute its own rationale. It is not by chance that these assumptions determine the conditions and constraints which permit the evidence to be constructed. In this article, we argue why the Regression Tree Method's rationale is more suitable than General Linear Model to analyze complex educational datasets. Furthermore, we apply the CART algorithm of Regression Tree Method and the Multiple Linear Regression in a model with 53 predictors, taking as outcome the students' scores in reading of the 2011's edition of the National Exam of Upper Secondary Education (ENEM; N = 3,670,089), which is a complex educational dataset. This empirical comparison illustrates how the Regression Tree Method is better suitable than General Linear Model for furnishing evidence about non-linear relationships, as well as, to deal with nominal variables with many categories and ordinal variables. We conclude that the Regression Tree Method constructs better evidence about the relationships between the predictors and the outcome in complex datasets.

Cristiano Mauro Assis Gomes[i]
Laboratory for Cognitive Architecture Mapping (LAICO), Federal University of Minas Gerais, Brazil

Gina C. Lemos[ii]
Research Centre on Education (CIEd), Institute of Education, University of Minho, Portugal

Enio G. Jelihovschi[iii]
Department of Exact and Technological Sciences, Universidade Estadual de Santa Cruz (UESC), Brazil

Keywords: Regression tree model; General linear model; National Exam of Upper Secondary Education (ENEM); Complex datasets

1. INTRODUCTION

Any quantitative method is shaped by certain rules or assumptions that constitute its own rationale (Gauer *et al.*, 2010; Golino & Gomes, 2014a, 2014b, 2016; Gomes, 2020; Gomes & Almeida, 2017; Gomes *et al.*, 2019; Gomes & Valentini, 2019; Pereira *et al.*, 2019). They provide the conditions as well the constraints which determine how the evidence can be constructed (Gomes *et al.*, 2017; Gomes & Gjikuria, 2017; Gomes *et al.*, 2013; Gomes & Jelihovschi, 2016). For example, factor analysis (Gomes, Linhares *et al.*, 2021; Matos *et al.*, 2019) and item response theory (Golino *et al.*, 2015; Golino &

Gomes, 2019; Gomes, 2013) assume that scientific constructs are latent variables that explain the variance of observable variables, which are, in general, respondents' performance in tasks (Gomes & Nascimento, 2021; Gomes et al., 2021a, 2021b, 2021c) or respondents' self-reports about certain statements, words, and so on (Fleith & Gomes, 2019; Gomes, Araujo *et al.*, 2020). When factor analysis or item response theory estimate latent variables in an individual, the time parameter is added, usually assuming that the previous response has an influence over the individual's response (Ferreira & Gomes, 2017; Gomes *et al.*, 2018; Rodrigues & Gomes, 2020). These rules constrain the evidence that can be constructed by the researcher. In other words, factor analysis and item response theory transform information from data into knowledge through a frame that bind latent and observable variables in a linear structure. In sum, the rationale of any quantitative method is both a possibility and a restriction required as the evidence is being constructed (Jelihovschi & Gomes, 2019; Pires & Gomes, 2017, 2018).

Educational large-scale assessments, such as the International Association for the Evaluation of Educational Achievement (IEA) (Härnqvist, 1975), the Programme for International Student Assessment (PISA) (OECD, 2019), and the National Exam of Upper Secondary Education (ENEM) (Brasil/INEP, 2015) are, essentially, complex datasets. Some properties that constitute these assessments as complex datasets are the following: (1) They have large amounts of information about students and their socioeconomic, psychological, familiar, and educational backgrounds; (2) not by chance, they involve many non-linear relationships among the variables; (3) besides, they have a large set of nominal variables with many categories. For example, the variable country in PISA has dozens of countries, and the variables states of Brazil in ENEM has dozens of states; (4) they have many ordinal variables, whose distances between their values, at least conceptually, is not the same.

The General Linear Model (i.e. Structural Equation Modeling, Multiple Regression, Multilevel Regression, ANOVA, correlations) is the mainstream approach in Education. However, its use in complex datasets is very questionable. The techniques from the General Linear Model can model nonlinear relationships among variables, and deal with nominal and categorical variables with many variables, but they do it through difficult ways.

The General Linear Model can only estimate non-linear relationships that are previously modeled by the analyst and inserted a priori in the model, so, in this way, many non-linear relations may be lost. For example, if the analyst suspects that there is a non-linear relationship between engagement and achievement, then he/she needs to introduce this non-linear relationship in the model, also defining the kind of non-linearity she believes is occurring between these variables, otherwise it will not work, that is, what is not modeled will not be identified and considered in the analysis. Since the General Linear Model is the mainstream in Education, being extensively applied in educational complex datasets (i.e. large-scale assessments), we may infer that many non-linear relationships between predictors and outcomes have been lost or not identified.

The General Linear Model is not suitable to deal with nominal variables with many categories. As an example, ENEM has a nominal variable with 27 categories, representing the States of Brazil. One common approach of the General Linear Model to deal with this variable is to transform the variable in 26 dummy variables, taking one category from the 27 categories as reference. For example, if the Category A (Minas Gerais State) is taken as the reference, all other categories (States of Brazil) will be compared only with this category (Minas Gerais State). Therefore, the slopes of the other dummy variables, representing the other categories, are estimated in comparison to the category A (Minas Gerais State). This is a real problem, because if the researcher wants to understand how the slope of the category B (Rio de Janeiro State) is related to the slope of the category C (São Paulo State), she will not be able to do that since all the slopes are comparable only to the slope of the category A (Minas Gerais State). This drawback of linear models narrows down substantially the information and interpretation from the data, compromising the constructed evidence from these models.

The General Linear Model is not appropriate for dealing with ordinal variables. It assumes, incorrectly, that ordinal variables are scales, that is, the distances between any two of their values are the same. This incorrect assumption is an issue. Educational large-scale datasets are plentiful of ordinal variables. Let's see an example. Suppose that an ordinal variable of motivation to study has three discrete values: 1- no motivation, 2- weak motivation, and 3- strong motivation. The General Linear Model assumes that the distance between the value 1 and value 2 is the same as the distance between the value 2 and value 3. The estimation of the intercept and the slopes of this variable are produced considering this assumption. However, conceptually this assumption is wrong, since ordinal variables are not scales, so there is no guarantee that the distances are the same.

By imposing several assumptions about normality, linearity, homoscedasticity, or independence of data for data analysis (Geurts, Irrthum & Wehenkel, 2009), the General Linear Model approach demands a great deal of effort to deal with complex educational datasets. On the other hand, the rationale of the Regression Tree Method is effective in dealing with complex datasets because this is a data-driven approach (Gomes & Jelihovschi, 2019). This method does not demand any assumption about the data, and this absence of assumptions makes it very suitable to deal with all types of variables, and all kind of linear and non-linear relationships among the variables (James et al., 2013).

The CART algorithm is, probably, the most famous and used technique of the Regression Tree Method. It has been originated in the machine learning field (James *et al.*, 2013). This algorithm was created by Breiman, Friedman, Olshen and Stone (1984) to implement the Classification and Regression Tree Method. Since the CART algorithm is not part of the mainstream quantitative approaches in Education and Psychology, we shortly explain the principles of this technique. [See Gomes and Almeida's (2017) as well Gomes and Jelihovschi's (2019) papers for more detailed information about this approach, especially its application to educational data.].

The CART algorithm contains some attractive properties. For instance, this algorithm does not make any assumptions about data, which makes it very useful for discovering linear or non-linear relationships among the variables (Geurts *et al.*, 2009). It also allows for the presence of many types of predictors in its models, without the need of preparing or transforming the variables. Another advantage is that its outputs generate results that are easy to read and to interpret, making them understandable and accessible to decision makers, managers, educators, as well as the general public (Gomes & Almeida, 2017).

In the case of regressions, the CART algorithm tries to reduce to the minimum possible the ordinary least squares error of the outcome prediction. To reduce this error, the CART algorithm divides the data into parts, and every data partition generates two new separated parts (Lantz, 2015). The CART algorithm' output looks like a tree. The original data is named root node. The parts created by the algorithm through the splits of the data are named nodes, and the nodes that are not broken are called terminal nodes or leaves (Zhang & Singer, 2010).

The process of data partition is complex. Every division of data is originated from the best split from a set of splits. For instance, for any partition of the data, the CART algorithm performs a specific split of that partition, for each value of every predictor. From these splits, the algorithm selects the best split, which is the split that reduces the most the outcome prediction error. The partition of the data is recursive and continues until it is no longer possible to decrease the outcome prediction error, or until another criterion is achieved, as for example, the minimum number of cases for each node.

As a result of this approach, the CART algorithm produces some data partitions that reduce the outcome prediction error only for the sample in question, but not for other samples, hindering the generalization of the results. The literature of machine learning recognizes this phenomenon and refers to it as overfit. To deal with overfit, the machine learning field recommends separating, randomly, the data in a training sample and a test sample, as well as to perform a cross-validation in the training sample (James *et al.*, 2013), and also to "prune" (name coined by the literature) some of the tree nodes created by the CART algorithm to help the generalization of the model or to facilitate the interpretability of the tree (Rokach & Maimon, 2015).

The literature of machine learning suggests two different ways to carry out the process of pruning the tree: one is called complexity cost criterion, while the other use the interpretability criterion (Rokach & Maimon, 2015). The complexity cost criterion identifies the number of tree nodes that generates the lowest outcome prediction error, pruning all the nodes that are beyond a given cut-off value. In turn, the interpretability criterion aims to maintain only a small amount of tree nodes, easy to interpret, as well as to generate substantial information, pruning all the other nodes (James *et al.*, 2013; Rokach & Maimon, 2015).

2. COMPARING REGRESSION TREE METHOD TO MULTIPLE REGRESSION

In this article, we compare the Regression Tree Method to Multiple Linear Regression applying both in a predictive model with 53 predictors and the languages domain as the outcome of reading achievement. We apply Multiple Linear Regression because this technique is a very representative technique of the General Linear Model and largely applied in Education. All variables of the model come from the 2011 National Exam of Upper Secondary Education (Exame Nacional do Ensino Médio [ENEM]). Currently, ENEM is the measure by which the quality of the Secondary Education is evaluated, and it is also the main national assessment measure to select students for the entrance in Brazilian public universities, through the Unified Selection System (SiSU), as well as in some universities abroad (MEC/INEP, 2013). Besides, if we consider the number of students who take the exam annually, the ENEM has a remarkable position in the world as an assessment for the entrance of students to Higher Education (Brasil/INEP, 2015). Since 2009, the ENEM has 180 multiple-choice items and an argumentative essay measuring four broad domains: natural sciences, mathematics, human sciences, and languages, and it is administered once a year, over two days, to millions of students. Each year, the ENEM microdata collect, register and store demographic, socioeconomic, educational, and motivational information about the students who take the exam.

The ENEM microdata are composed by many and diverse variables (i.e. quantitative, nominal variables with many categories and ordinal variables) which are freely available for download by the INEP at http://portal.inep.gov.br/web/guest/microdados and are supported by the Brazilian Ministry of Education (MEC/INEP, 2012). Some studies have applied the Regression Tree Method to predict academic achievement (i.e. Pazeto *et al.*, 2019, 2020). Some of them applied this method in educational complex datasets, as the ENEM (Gomes, Amantes *et al.*, 2020; Gomes & Jelihovschi, 2019; Gomes, Fleith *et al.*, 2020; Gomes, Lemos & Jelihovschi, 2020). However, concerning the ENEM, none of them have studied the suitable of the Regression Tree Method to construct educational evidence regarding the relationship between predictors and outcomes, in comparison to the General Linear Model. Our purpose, in this article, is to compare the Regression Tree Method and the Multiple Regression to empirically illustrate our argument that the Regression Tree Method is more adequate to construct evidence in complex educational datasets.

We believe the most important contribution of our article is to argue that the Regression Tree Method may be broadly applied in complex educational datasets. Additionally, we intend to show that the General Linear Model can be applied with important restrictions and cautious. We aim to challenge the almost exclusive use of the General Linear Model in complex educational datasets, inviting the researchers to use suitable alternatives.

3. METHOD

3.1. PARTICIPANTS

The ENEM's 2011 edition had 5,380,856 students enrolled in it (Brasil/INEP, 2015). We excluded the students who were not present on both days of the exam and did not answer the socioeconomic questionnaire of the 2011 ENEM's microdata, hence, our sample narrowed down to 3,670,089 students. This sample is predominantly female (59.51%), single (86.23%), and Caucasian (43.51%); most finished Secondary Education before 2011 (55.23%), completed Secondary Education through regular teaching (91.24%), attended public schools for Secondary Education (75.07%), attended schools in urban regions (97.58%), and possessed family monthly income equal or smaller than 2 minimum wages (74.63%). Moreover, this sample showed a high motivation to take the exam in order to pursue studies in Higher Education (90.60%), as well as to obtain a scholarship (82.81%).

3.2. VARIABLES OF THE ANALYSIS

Next, we briefly present the structure of our database and the variables of the analysis. The 2011 ENEM's microdata are comprised of seven parts or blocks of information: (1) Students' data; (2) students' school data; (3) municipality data, the place where the student took the exam; (4) multiple-choice exam data; (5) argumentative essay data; (6) basic education census data; (7) socioeconomic questionnaire (MEC/INEP, 2012).

Table 1
Predictors of the study, which were extracted from the 2011 ENEM's microdata

Type	n	Variables
Ordinal, 20 cat.	1	Q1. Number of people that live with the student.
Ordinal, 8 cat.	2	Q2. Student's father education.
Ordinal, 8 cat.	3	Q3. Student's mother education.
Ordinal, 11 cat.	4	Q4. Student's family monthly income.
Ordinal, 11 cat.	5	Q5. Student's own monthly income.
Nominal, 4 cat.	6	Q6. Student's home (own home or other options).
Nominal, 4 cat.	7	Q7. Student's home location (urban or other options).
Nominal, 2 cat.	8	Q8. Student's paid activity.
Nominal, 2 cat.	9	Q15. Courses attended by the student: Vocational course.
Nominal, 2 cat.	10	Q16. Courses attended by the student: Preparation for the Higher Education admission exam.
Nominal, 2 cat.	11	Q17. Courses attended by the student: Higher Education.
Nominal, 2 cat.	12	Q18. Courses attended by the student: Second language.
Nominal, 2 cat.	13	Q19. Courses attended by the student: Informatics.
Nominal, 2 cat.	14	Q20. Courses attended by the student: Preparation for public tender.
Ordinal, 6 cat.	15	Q24. Motivation to take the Exam: To test personal knowledge.
Ordinal, 6 cat.	16	Q25. Motivation to take the Exam: To pursue studies in Higher Education.
Ordinal, 6 cat.	17	Q26. Motivation to take the Exam: To obtain a Secondary Education certificate.
Ordinal, 6 cat.	18	Q27. Motivation to take the Exam: To obtain a scholarship.
Ordinal, 7 cat.	19	Q28. Years taken to complete Elementary Education.
Ordinal, 5 cat.	20	Q29. Hiatus in studies during Elementary Education.
Nominal, 9 cat.	21	Q30. Type of school attended in Elementary Education: Private, public or other options.
Ordinal, 5 cat.	22	Q32. Hiatus in studies during Secondary Education.
Nominal, 9 cat.	23	Q33. Type of school attended in Secondary Education: Private, public or other options.
Nominal, 27 cat.	24	I1. Location of the school attended at the time of the Exam: State located in Brazil.
Nominal, 4 cat.	25	I2. Administrative Unit of the School attended at the time of the Exam: Private or other options.
Nominal, 2 cat.	26	I3. School location attended at the time of the Exam: Urban or rural.
Nominal, 4 cat.	27	I4. Completion of Secondary Education: Finished, ongoing, or other options.
Nominal, 4 cat.	28	I5. Type of school institution in Secondary Education in which the student finished or would finish Secondary Education: Regular Teaching or other options.
Nominal, 4 cat.	29	I6. School functioning attended by the student when he/she performed the Exam: The school was still open and functioning at the time of the exam.
Nominal, 2 cat.	30	I7. Request for a Secondary Education Certification.
Nominal, 2 cat.	31	I8. Request to perform the Exam in Braille.
Nominal, 2 cat.	32	I9. Request to perform the Exam in larger letters.
Nominal, 2 cat.	33	I10. Request for reader assistance.
Nominal, 2 cat.	34	I11. Request for an easily accessible classroom.
Nominal, 2 cat.	35	I12. Transcript request.
Nominal, 2 cat.	36	I13. Libras request.
Nominal, 2 cat.	37	I14. Low vision indicator.
Nominal, 2 cat.	38	I15. Blindness indicator.
Nominal, 2 cat.	39	I16. Hearing impaired indicator.
Nominal, 2 cat.	40	I17. Physical disability indicator.
Nominal, 2 cat.	41	I18. Mental disability indicator.
Nominal, 2 cat.	42	I19. Attention deficit indicator.
Nominal, 2 cat.	43	I20. Dyslexia indicator.
Nominal, 2 cat.	44	I21. Indicator of pregnancy.
Nominal, 2 cat.	45	I22. Breast-feeding indicator.
Nominal, 2 cat.	46	I23. Lip reading indicator.
Nominal, 2 cat.	47	I24. Request for taking the Exam another day.
Nominal, 2 cat.	48	I25. Deafness indicator.
Nominal, 4 cat.	49	I26. Marital status.
Nominal, 6 cat.	50	I27. Declared color/race.
Numerical	51	I28. Age.
Nominal, 2 cat.	52	I29. Sex.
Nominal, 27 cat.	53	I30. Student's home location: State in Brazil.

The outcome variable of our study comes from the fourth block of information and it refers to the students' scores in languages domain, which comprises reading. The exam has 180 items and 45 items measure the language domain. This is a broad domain composed of 9 language competences and 30 abilities that are defined by a theoretical reference matrix. The students' languages scores derive from a standardized scale with a mean of 500 points and standard deviation of 100 points, ranging from 0 to 1000 points (Brasil/INEP, 2015). This scale is produced by the Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), which is the Brazilian institute responsible for producing, storing, and making available to the public the ENEM' microdata. In turn, the 53 predictors of our study originate from the first, second, and the seventh block of the 2011 ENEM's microdata. All the predictors are outlined in Table 1. The reader can get more details about the ENEM and its microdata in MEC/INEP (2012).

3.3. DATA ANALYSIS

All statistical analyses were carried out in R language and environment for statistical computing (R Core Team, 2017). We used the functionalities of the rpart R package (Therneau & Atkinson, 2015) and the caret R package (Kuhn, 2017) to perform the CART algorithm. The following steps were applied, considering all the main recommendations of the literature (James *et al.*, 2013): [1] We divided randomly the data in two parts, a training sample (75% of cases) and a test sample (25% of cases), since the literature recommends as a suitable approach this ratio between the training and test sample size (James *et al.*, 2013); [2] We generated a non-pruned tree in the training sample, applying the CART algorithm to the predictive model with 53 predictors. This is the first step of the CART's output, that is, this algorithm generates a tree with a very large number of leaves in educational complex datasets (Gomes & Jelihovschi, 2019); [3] We applied the cross-validation 3-Fold to the predictive model in the training sample, since this number of folders is considered suitable in big data (James *et al.*, 2013); [4] We inspected the pruned tree suggested by the complexity cost criterion; [5] and generated the pruned tree through the interpretability criterion; [6] We checked the generalization of the results, comparing the R^2 index value obtained by the predictive model in the training sample to the R^2 index value obtained by the predictive model in the test sample; and, finally [7], obtained all possible information about the outcome variable, focusing on the reading and interpretation of the pruned tree.

In order to compare the Regression Tree Method with the General Linear Model, we apply the Multiple Regression approach to the same data. We used the same outcome, however we inserted as predictors for the model only those selected by the CART algorithm. Our purpose was to facilitate the comparison of the CART output with the Multiple Regression output. Since the CART algorithm used six predictors to produce the splits in the pruned tree (see results section), they were used as the predictors in the model of the Multiple Regression. They are: [1] Students' family monthly income; [2] type of schools attended by students in Primary Education; [3] students' motivation to perform the exam to obtain a Secondary Education certificate; [4] students' motivation to perform the exam to obtain a scholarship; [5] student's home location: State in Brazil; and [6] completion of Secondary Education.

Three predictors are ordinal variables: (1) Students' family monthly income; (2) students' motivation to perform the exam to obtain a Secondary Education certificate; (3) students' motivation to perform the exam to obtain a scholarship. The other three predictors are nominal variables. Since the General Linear Model assumes that the distance among the values of an ordinal variable is the same and its assumption needs to be tested, we treated the ordinal predictors of the model as dummy variables. Transforming them into dummy variables permits us to verify the distances among their values. Since dummy variables demand that one category be the reference and the other categories are compared to this reference, we defined a reference category for each predictor. The students' family monthly income had as reference the category "no income". The other 10 categories [(1) until 1 minimum wage; (2) 1 to 1.5; (3) 1.5 to 2; (4) 2 to 5; (5) 5 to 7; (6) 7 to 10; (7) 10 to 12; (8) 12 to 15; (9) 15 to 30; (10) above 30 minimum wages] are compared only to this reference category. The other two ordinal variables are the motivational variables and they have seven discrete numerical values ranging from 0 to 6. The reference category of these variables is the number 0, representing no motivation. Regarding the three nominal variables, they are usually transformed as dummy variables in the General Linear Model approach. For the nominal variable "type of schools attended by students in Primary Education", which has nine categories, the reference category is "only in public schools". The nominal variable "student's home location: State in Brazil" has 27 categories. The reference category is the Acre state of Brazil. The nominal variable "completion of Secondary Education" has four categories and the reference category is "I have concluded the Secondary Education".

We used the functionalities of the biglm R package (Lumley, 2020) to perform the Multiple Regression because this package is suitable to big data. The analysis followed three steps: (1) the model was trained in training sample; (2) the normality of the residuals was inspected through its kurtosis and skewness; (3) the model was applied to the test sample in order to verify the explained variance (R2) of the outcome; the caret R package (Kuhn, 2017) was used in this analysis.

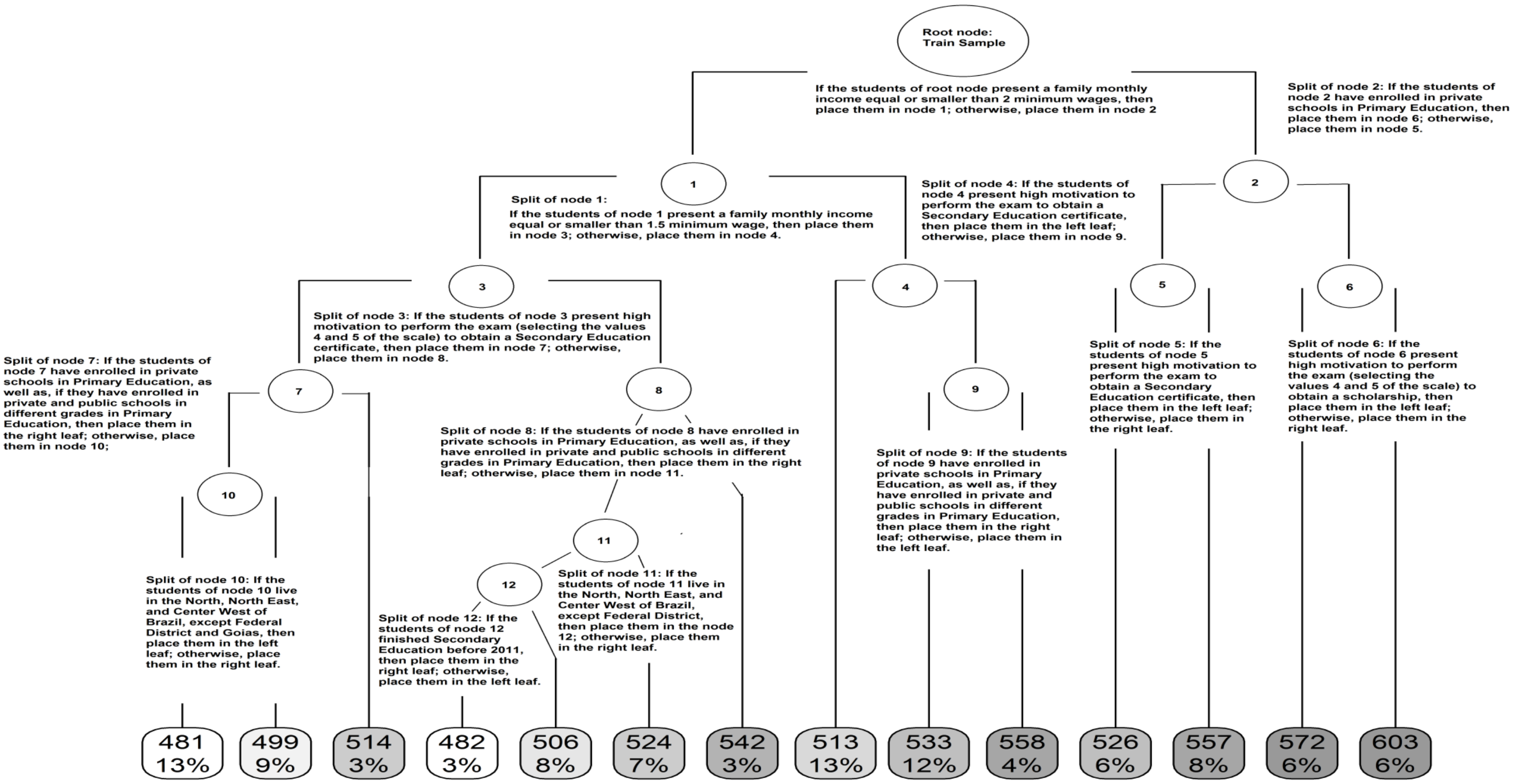
4. RESULTS AND DISCUSSION

The Regression Tree Method and the CART algorithm are not current approaches in Educational Sciences. So that, we will present and discuss the results together, focusing on the reading of the pruned tree, and showing how the structure of this tree can provide substantial information about the relationship between the predictors and outcome.

The CART algorithm generated a non-pruned tree with 32,131 leaves and 45,053 data splits. Many of these tree leaves only increase the outcome prediction error. The cost complexity criterion indicated that the first 2,074 leaves of this tree produced the lowest outcome prediction error (72.08%). Nonetheless, the cost complexity criterion suggested a pruned tree with still a very large number of leaves yet (2,074), leaving the tree with no interpretation and without any substantial information. Thus, we used the interpretability criterion to produce the final pruned tree. This remaining tree has only 13 splits and 14 leaves, and an outcome prediction error of 79.80%. Comparing this tree to the pruned tree indicated by the cost complexity criterion, there's a difference of only 6.72%, in terms of outcome prediction error favoring the cost complexity criterion. However, there is a difference of 2,060 leaves in favor of the final tree. Although this tree produced a relative worse outcome prediction, it allows the attainment of substantial information and interpretability. Therefore, our study resulted in this pruned tree with 14 leaves.

The final tree with 14 leaves explained 20.20% of the students' reading achievement variance in the training sample, as well as 20.14% of the students' reading achievement variance in the test sample. This small difference of 0.06% between the samples suggests a considerable generalization of the predictive model. The Figure 1 shows the pruned tree with 14 leaves. The root node is represented by the oval figure at the top of Figure 1. The nodes are represented by the numbered circles and the leaves (terminal nodes) are represented by the oval figures at the bottom of Figure 1. Every tree leaf in the Figure 1 has two numbers: The top number shows the reading average achievement of the students contained in the leaf, while the bottom number shows the percentage of the students in the training sample that belong to the leaf. For example, at the left corner of Figure 1 (taking the reader as reference), there is a leaf with the numbers 481 and 13%. These numbers indicate that this leaf is composed of students with an average of 481 points in reading achievement, which represent 13% of the training sample.

Figure 1
Pruned tree



If the students of root node present a family monthly income equal or smaller than 2 minimum wages, then place them in node 1; otherwise, place them in node 2

Split of node 2: If the students of node 2 have enrolled in private schools in Primary Education, then place them in node 6; otherwise, place them in node 5.

Split of node 1:
If the students of node 1 present a family monthly income equal or smaller than 1.5 minimum wage, then place them in node 3; otherwise, place them in node 4.

Split of node 4: If the students of node 4 present high motivation to perform the exam to obtain a Secondary Education certificate, then place them in the left leaf; otherwise, place them in node 9.

Split of node 3: If the students of node 3 present high motivation to perform the exam (selecting the values 4 and 5 of the scale) to obtain a Secondary Education certificate, then place them in node 7; otherwise, place them in node 8.

Split of node 7: If the students of node 7 have enrolled in private schools in Primary Education, as well as, if they have enrolled in private and public schools in different grades in Primary Education, then place them in the right leaf; otherwise, place them in node 10;

Split of node 8: If the students of node 8 have enrolled in private schools in Primary Education, as well as, if they have enrolled in private and public schools in different grades in Primary Education, then place them in the right leaf; otherwise, place them in node 11.

Split of node 9: If the students of node 9 have enrolled in private schools in Primary Education, as well as, if they have enrolled in private and public schools in different grades in Primary Education, then place them in the right leaf; otherwise, place them in the left leaf.

Split of node 5: If the students of node 5 present high motivation to perform the exam to obtain a Secondary Education certificate, then place them in the left leaf; otherwise, place them in the right leaf.

Split of node 6: If the students of node 6 present high motivation to perform the exam (selecting the values 4 and 5 of the scale) to obtain a scholarship, then place them in the left leaf; otherwise, place them in the right leaf.

Split of node 10: If the students of node 10 live in the North, North East, and Center West of Brazil, except Federal District and Goias, then place them in the left leaf; otherwise, place them in the right leaf.

Split of node 12: If the students of node 12 finished Secondary Education before 2011, then place them in the right leaf; otherwise, place them in the left leaf.

Split of node 11: If the students of node 11 live in the North, North East, and Center West of Brazil, except Federal District, then place them in the node 12; otherwise, place them in the right leaf.

Since every tree leaf is formed by a set of splits, which starts at the root node, we get information of each leaf by reading the tree through a top-down screening. For example, the leaf with 13% of the training sample and an average of 481 points in reading achievement, at the left corner of Figure 1, is the product of the splits of root node, node 1, node 3, node 7, and node

10. This leaf contains the students who: [1] have a family monthly income equal or smaller than 1.5 minimum wage (splits of root node and node 1), [2] show high motivation to perform the exam in order to obtain a Secondary Education certificate (split of node 3), [3] have not enrolled in private schools in Primary Education or attended private and public schools in different grades in Primary Education (split of node 7), and [4] live in the North, North East, and Center West of Brazil, except Federal District and Goiás State.

We now present the main results. Despite the large number of 53 predictors employed in the predictive model, we can observe in the Figure 1 that the CART algorithm used six predictors to produce the splits in the pruned tree. These variables are: [1] Students' family monthly income; [2] type of schools attended by students in Primary Education; [3] students' motivation to perform the exam to obtain a Secondary Education certificate; [4] students' motivation to perform the exam to obtain a scholarship; [5] student's home location: State in Brazil; and [6] completion of Secondary Education.

The most important variable used by the CART algorithm to discriminate students' reading achievement is students' family monthly income, as we can see in Figure 1, which is an indicator of socioeconomic status (OECD, 2019). This variable of the ENEM microdata is an ordinal variable with a large number of categories that represent different intervals of income. The pruned tree shows that students with a family monthly income higher than 2 minimum wages cluster around the four leaves at the right corner of Figure 1; in general, these leaves (526 points, 557 points, 572 points, 603 points) have the greatest averages in reading achievement. In turn, students with a family monthly income between 1.5 and 2 minimum wages compose the three leaves with average of 513 points, 533 points, and 558 points, representing an intermediate achievement in reading. Students who present a family monthly income equal or smaller than 1.5 minimum wage compose the seven leaves at the left corner of Figure 1; overall, these leaves have the lowest averages in reading achievement. These results show that a family monthly income higher than 2 minimum wages is related to the highest averages in reading achievement, while a family monthly income between 1.5 and 2 minimum wages is related to intermediate averages in reading achievement, and a family monthly income smaller than 1.5 minimum wage is related to the lowest averages in reading achievement. By looking into this result, we observe that only a few categories of this predictor are relevant to discriminate the variance of the outcome. This result is striking because it shows that the General Linear Model's assumption about the ordinal variables is incorrect for the analyzed data. It is not adequate to assume, as General Linear Model does, that all the values of ordinal variables have the same importance to explain the outcome variance. As the pruned tree shows, many categories of family income, particularly the higher intervals of income, have no significance. This outstanding result is an important parameter for the researcher's interpretation of the relationship between the students' income and reading achievement. Through this result the researcher can build evidence that only the lower income categories make any difference to differentiate the students' reading variance. Consequently, the researcher can conclude that if public politics act on income, changing the small income brackets of Brazilian families can provide better opportunities for students to increase their achievement.

In spite of the students' family monthly income be the main predictor, it is not, by itself, a powerful enough predictor of students' reading achievement (Figure 1). For example, the leaf with an average of 533 points, which comprises the students' family monthly income between 1.5 and 2 minimum wages, shows a higher average performance than the leaf with an average of 526 points, which is associated with students' family monthly income higher than 2 minimum wages. Thus, other variables in addition to students' family monthly income account for the shared variance with reading achievement. Along our pruned tree (Figure 1), it can be observed that one variable that discriminates students' reading achievement besides students' family monthly income is students' motivation to perform the exam to obtain a Secondary Education certificate. This finding suggests that if a student shows high motivation to perform the exam to obtain a Secondary Education certificate, then her reading achievement tends to be lower compared to their counterparts who have medium or low motivation to perform that exam. Moreover, students' motivation to perform the exam to obtain a scholarship is a relevant variable, namely for those students whose family monthly income is higher than 2 minimum wages (split of root node) and have enrolled in private schools in Primary Education (split of node 2). These results might seem incorrect, at least superficially, since they affirm that more motivation leads to lower reading achievement. However, these findings make sense. In Brazil, the students who intend to take ENEM to obtain a certificate are those who did not finish high school in the expected time. These are students who have failed and had to repeat one or more years of their education. They usually have learning difficulties, as well as, lower academic achievement. Therefore, more motivation to take the ENEM to obtain the certificate implying less academic reading achievement is a result that makes sense in the Brazilian context. The same applies to the result which shows that more motivation to perform the exam to obtain a scholarship implies less achievement in reading, notably in family monthly income higher than 2 minimum wages. In Brazil, the scholarships are provided by the Brazilian government only for those students which have lower income and intend to enroll in private universities. In comparison, private universities have high costs while public schools are free. In addition, private institutions are less prestigious than public universities, which centralize research in Brazil. Hence, students need to get a high score on ENEM to enroll in public universities, while a low score is enough to enroll in almost all private institutions. Thus, students with low family income and low ENEM scores enroll in private universities hoping to get a scholarship from the Brazilian government. Even if they receive a scholarship, their family income must be higher 2 minimum waves, otherwise they are not able to pay the high costs of private universities, because scholarships tend not to cover all the costs (Gomes & Jelihovschi, 2019).

Another very important aspect of the results regarding the predictors of motivation is the General Linear Model assumption involving the ordinal variables. As we can see in Figure 1, the values 4 and 5 are the substantial values to explain the variance of the outcome. The distances between 0 and 1, 1 and 2, 2 and 3 are not important, which supports the conclusion about family monthly income that only a few categories are indeed important. Both results show that the General Linear Model assumption about ordinal variables is very incorrect, at least for the analyzed data. This makes substantial difference when building evidence about the predictors and the outcome. Through the output of the pruned tree, the researcher can conclude that only the higher values of motivation to get a certificate or a scholarship make a difference in discriminating students' reading performance. She can claim that more or less motivation at the lower values of motivation makes no difference, claiming that the relation between motivation and reading achievement is non-linear and seems to work like the activation of the neurons, in that one needs to greatly increase motivation to influence variation in reading achievement.

It is worth mentioning that the output of the pruned tree shows how predictors are conditioned by other predictors to predict the outcome. As we said, the students' motivation to perform the exam to obtain a scholarship is conditioned by the variable “family monthly income” and the variable “type of schools attended by students in Primary Education”. The General Linear Model is not able to provide this kind of evidence about the relationships between the variables. The Multiple Regression approach, for example, assumes that the predictors are unrelated, that is, orthogonal, so the weight of each predictor is added to the weight of the other predictors to explain the variance of the outcome. In a very different approach, the Regression Tree Method assumes that if a node is a product of ancestral nodes that were created by using distinct predictors, then this node indicates that a relationship exists between these predictors.

Another relevant variable to discriminate students' reading achievement is the type of school attended by students in Primary Education. The highest averages in reading achievement are related to the students who have attended only private schools (split of node 2), or at least, have attended a mix of private and public schools in different grades in Primary Education (split of nodes 7, 8 and 9). Furthermore, students who finished Secondary Education before 2011 (right leaf of split of node 12, with 506 points) tend to have a better performance than the students who did not finished Secondary Education before 2011 (left leaf of split of node 12, with 482 points). In turn, students who live in the South, South East of Brazil, or in the Federal District, tend to have a better performance (right leaf of split of node 10, and right leaf of split of node 11, with 499 and 524 points, respectively) than the students who live in other regions (left leaf of split of node 10, and leaves from the left node (node 12) of split of node 11, with 481 and 482 or 506 points, respectively).

We may conclude then, that two main properties regarding the relationship between the predictors and the outcome variable of students' reading achievement are important. The first represent the fact that some predictors are conditioned by other predictors in terms of their own predictive roles. For example, the completion of Secondary Education only discriminates the reading achievement of the students who: [1] have a family monthly income equal or smaller than 1.5 minimum wage (split of node 1), [2] present medium or low motivation to perform the exam to obtain a Secondary Education certificate (split of node 3), [3] neither enrolled in private schools in Primary Education, nor attended in private and public schools in different grades in Primary Education (split of node 8), and [4] live in the North, North East, and Center West of Brazil, except Federal District (split of node 11). Furthermore, the completion of Secondary Education does not have a predictive role in other contexts. This variable is neither for students whose family has a monthly income higher than 2 minimum wages, nor of students whose family has a monthly income between 1.5 and 2 minimum wages. This is an evidence of non-linear relationship among a set of predictors and the outcome. Another example of non-linearity among variables is the fact that the variable student's home location does not have any predictive role for the students whose family has a monthly income higher than 2 minimum wages, nor the students whose family has a monthly income between 1.5 and 2 minimum wages. Another striking result refers to the second trend: there are categories of some variables that are related to either worse or better reading achievement under specific conditions. For example, in the case of type of school attended by students, for those students who present a family monthly income higher than 2 minimum wages, only private schools are related to a better performance. On the other hand, for students who have a family monthly income equal or smaller than 2 minimum wages, a mix of private and public schools attended by the students in different grades in Primary Education is also related to a better performance. In other words, as Figure 1 shows, depending on the socioeconomic context, the type of school attended by students is related to a better or to a worse reading achievement. Another example: Goiás State is related to a better reading achievement in the context of node 10 split (499 points vs 481 points), and to worse reading achievement in the context of node 11 split (482 or 506 points vs 524 points). This is a very interesting result that is very unlikely to be achieved by General Linear Model techniques.

The results of the Multiple Regression technique are shown in Table 2. The intercept is 527.64, indicating that if all predictors have the value of their reference category, then the score of the students who took the 2011 edition of ENEM will have 527.64 points in reading.

All the results on the predictors corroborate what we have previously argued. As stated, the reference category of the family monthly income variable is “no income”. It is notable that the multiple regression result supports that if the family’s monthly income is up to 1 minimum wage, compared to a family with no income, then these students will score 75.07 points higher than students living in families with no income. Note in Table 2 that the distances between the categories are very unequal. According the General Linear Model assumption, the distances should be equal and either increase or decrease. The result shows a very different pattern, that is, a non-linear shape. Some categories indicate a better reading achievement compared to the reference category, while other categories show lower reading achievement. As we said, even creating dummy variables for ordinal variables, as we did, this is not very appropriate, since the researcher is asked to compare all the categories only in against the reference category. Therefore, in our case, we can only compare the categories with respect to the reference category “no income”. We do not know anything else.

What is observed in the ordinal variable family monthly income also occurs in the nominal variable type of schools attended by Primary Education students. The reference category is students who enrolled only in public schools. The results show that the largest increase for this reference category is of 38.54 points in reading achievement, relative to the category of students enrolled in private schools only. The results also show that if students enrolled in any year in private school, then their reading achievement is better than that of the students who are only enrolled in public schools. On the other hand, if students enrolled in any other type of schools besides private and public schools, then these students have lower reading achievement. Note that all results are about the categories compared to students enrolled in public schools only. If the researcher wants to compare other interesting relationships between other categories, she is not able to do that.

Table 2

Results from the Multiple Regression Approach

parameters	coefficient	p-value	parameters	coefficient	p-value
(intercept)	527.64	0.0000			
family monthly income (no income)			student's home location (Acre state of Brazil)		
until 1 minimum wage	75.07	0.0000	AL	1.83	0.0007
1 to 1.5 minimum wage	-12.89	0.0000	AM	-5.85	0.0000
1.5 to 2 minimum wages	-3.52	0.0000	AP	7.21	0.0000
2 to 5 minimum wages	6.31	0.0000	BA	8.36	0.0000
5 to 7 minimum wages	-2.44	0.0000	CE	14.87	0.0000
7 to 10 minimum wages	-0.77	0.0009	DF	22.13	0.0000
10 to 12 minimum wages	-1.06	0.0000	ES	21.09	0.0000
12 to 15 minimum wages	0.41	0.0623	GO	16.17	0.0000
15 to 30 minimum wages	-0.00	0.9953	MA	4.26	0.0000
higher than 30 minimum wages	1.08	0.0000	MG	29.84	0.0000
type of schools attended by students in Primary Education (only in public schools)			MS	8.40	0.0000
main in public school	18.74	0.0000	MT	3.15	0.0000
only in private schools	38.54	0.0000	PA	8.51	0.0000
main in private schools	28.13	0.0000	PB	9.53	0.0000
only in indigenous schools	-20.93	0.0000	PE	11.41	0.0000
main in indigenous schools	-11.97	0.0050	PI	4.64	0.0000
only in quilombola schools	-11.18	0.0066	PR	23.16	0.0000
main in quilombola schools	-8.82	0.0320	RJ	27.30	0.0000
did not enroll in schools	-8.99	0.0004	RN	8.69	0.0000
students' motivation to perform the exam to obtain a Secondary Education certificate (distance between 0 and 1)			RO	7.55	0.0000
threshold 1 (distance between 1 and 2)	-16.94	0.0000	RR	4.90	0.0000
threshold 2 (distance between 2 and 3)	-1.51	0.0000	RS	25.87	0.0000
threshold 3 (distance between 3 and 4)	-3.50	0.0000	SC	27.70	0.0000
threshold 4 (distance between 4 and 5)	1.76	0.0000	SE	-2.21	0.0001
threshold 5 (distance between 5 and 6)	-0.02	0.8881	SP	28.91	0.0000
students' motivation to perform the exam to obtain a scholarship (distance between 0 and 1)			TO	-0.35	0.5694
threshold 1 (distance between 1 and 2)	-9.85	0.0000	completion of Secondary Education (concluded)		
threshold 2 (distance between 2 and 3)	-0.62	0.0011	the student is concluding the Secondary Education in the year of the exam	-12.37	0.0000
threshold 3 (distance between 3 and 4)	-7.92	0.0000	the student will conclude the Secondary Education after the year of the exam	-14.81	0.0000
threshold 4 (distance between 4 and 5)	5.55	0.0000	the student have never enrolled in Secondary Education	-18.92	0.0000
threshold 5 (distance between 5 and 6)	-0.90	0.0004			

Regarding students' motivation to take the exam to obtain a high school certificate, the results of the regression model show that the only relevant distance between the values is the one covering values 1 and 2. If the students move from value 1 to 2, their achievement in reading have a reduction of 16.94, compared to students moving from value 0 to 1. The other distances imply small reductions, and the distance between value 4 and 5 implies an increment in the achievement score. This indicates a non-linear relationship between this predictor and the outcome. A similar finding is observed in the motivation of students to take the exam to obtain a scholarship.

Regarding the nominal variable completion of Secondary Education, the results indicate that if the students do not complete high school before taking the exam, they will have a worse reading achievement, compared to students who have completed high school. As for the nominal variable of students' location, the results show that only those living in the states of Sergipe and Amazonas are related to a worse reading performance, compared to the Brazilian state of Acre ($p < .05$). In addition, students living in Minas Gerais will score 29.84 points higher in reading achievement than students living in Acre.

5. CONCLUSIONS

In this article, we presented the advantages of the Regression Tree Method, compared to the General Linear Model, for building evidence on complex educational datasets. Our comparison found many substantial non-linear relationships between the predictors and the outcome. In addition, we show why the logic of the General Linear Model is inadequate to handle non-linear relationships. Not dealing adequately with non-linearities makes the evidence fragile, since non-linear relations tend to change the results considerably, as well as, the scientific narrative that the researcher will hold about the data.

We also show that the General Linear Model is not the best idea when it comes to complex educational datasets. Unfortunately, its use is very dominant in Education and complex educational datasets. We hope that researchers will understand that the use of the General Linear Model should be applied to complex datasets only under special conditions. Taking our data as example, if a researcher argues that her interest primarily involves observing how the other categories of each ordinal and nominal predictors are related to a specific reference category of each predictor in which she has a theoretical interest, then the General Linear Model may be a reasonable approach. However, this goal is very restrictive and it is not the rule when the researcher intends to use a quantitative method. Typically, the researcher wants the method to be able to provide her with some substantial results about the relationships between various categories of nominal and ordinal variables, if these types of variables are present in the predictive model. In addition, the researcher wants the method to be able to provide a suitable approach to discovering complex non-linear relations, if they are relevant and represent important pattern in the data. For example, the pruned tree showed us that living in Southern, Southeastern of Brazil, or the Federal District, and having completed high school before ENEM 2011, showed only a predictive role for reading achievement in the context of the poorest families, who had a family monthly income of 1.5 minimum wages or less. This striking evidence is surprising and difficult to find by the General Linear Model. It substantially undermines scientific knowledge about how predictors are associated with outcomes and thus their effects on public policy and leads to inappropriate educational interventions. Imagine that you are an educational manager or an educational researcher and you discover that there is a non-linear relationship among type of school, family income and reading achievement. Also, suppose that you found that for the students who have a family monthly income higher than 2 minimum wages, only private schools predicted a better performance in reading. On the other hand, for the students who have a family monthly income equal or smaller than 2 minimum wages, a mix of private and public schools attended by the students in different grades in Primary Education predicted a better performance. Will this finding change your knowledge about reading and some factors that may affect this result producing substantial evidence? We found this result in our analysis only because we applied the Regression Tree Method. If we had used only Multiple Regression, we would not have found it.

Like many other studies, in spite of their originality and relevance, our study has some limitations that need to be considered. We did not simulate the data to be analyzed. We used an empirical data set to illustrate our arguments. Therefore, new studies could be interesting to corroborate or refute our arguments through simulation studies. However, although our data are not simulations, they do represent complex educational datasets everywhere, and not only in the Brazilian context. Our data is abundant in nominal and ordinal variables with many categories, just as our data seem to be abundant in non-linear relationships among variables. Although our goal was to use the data as an illustration to enrich our argument about the advantages of the Regression Tree Method, as compared to the General Linear Model, for analyzing complex educational datasets, we believe that our data, if not perfect as a simulation, was sufficient. Another possible limitation is that we only used Multiple Regression as a technique of the General Linear Model. We could apply a large number of techniques from this model to obtain more robust evidence. However, as we said, the Multiple Regression approach is the most widely used technique and represents well the logic of the General Linear Model. Therefore, we believe that using only Multiple Regression was not a problem and did not compromise our results and conclusions.

Our study has found evidence that highlights the relevance of investing in data-driven approaches, capable of discovering relevant non-linear relations that are hindered in the use of techniques from the General Linear Model. We hope see, in a short range of time, the Regression Tree Method pertaining to the techniques of the mainstream approaches in Education, especially in the context of the complex educational datasets.

ACKNOWLEDGEMENTS

Cristiano Mauro Assis Gomes: Productivity Fellowship, CNPq Brazil.
Gina C. Lemos: Postdoctoral Fellowship FCT (SFRH/BPD/93009/2013), Research Centre on Education (CIEd), FCT/MCTES-PT (projects UID/CED/1661/2013 and UID/CED/1661/2016), IE – University of Minho, Portugal.

REFERÊNCIAS

Brasil/Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira [INEP]. (2015). *Relatório pedagógico: Enem 2011-2012*. Inep.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Chapman & Hall/CRC.

Ferreira, M. G., & Gomes, C. M. A. (2017). Intraindividual analysis of the Zarit Burden Interview: a Brazilian case study. *Alzheimers & Dementia, 13*, P1163-P1164. https://doi.org/0.1016/j.jalz.2017.06.1710

Fleith, D. S., & Gomes, C. M. A. (2019). Students' assessment of teaching practices for creativity in graduate programs. *Avaliação Psicológica, 18*(3), 306-315. https://doi.org/10.15689/ap.2019.1803.15579.10

Gauer, G., Gomes, C. M. A., & Haase V. G. (2010). Neuropsicometria: Modelo clássico e análise de Rasch. In L. F. Alloy-Diniz (Org.), *Avaliação Neuropsicológica* (pp. 22-30). Artmed.

Geurts, P., Irrthum, A., & Wehenkel, L. (2009). Supervised learning with decision tree-based methods in computational and systems biology. *Molecular Biosystems, 5*(12), 1593-1605. https://doi.org/10.1039/b907946g

Golino, H. F., & Gomes, C. M. A. (2014a). Four Machine Learning methods to predict academic achievement of college students: a comparison study. *Revista E-Psi, 1*, 68-101. https://revistaepsi.com/artigo/2014-ano4-volume1-artigo4/

Golino, H. F., & Gomes, C. M. A. (2014b). Visualizing random forest's prediction results. *Psychology, 5*, 2084-2098. https://doi.org/10.4236/psych.2014.519211

Golino, H. F., & Gomes, C. M. A. (2016). Random forest as an imputation method for education and psychology research: its impact on item fit and difficulty of the Rasch model. *International Journal of Research & Method in Education, 39*(4), 401-421. https://doi.org/10.1080/1743727X.2016.1168798

Golino, H. F., & Gomes, C. M. A. (2019) *TDRI: Teste de Desenvolvimento do Raciocínio Indutivo*. Hogrefe.

Golino, H. F., Gomes, C. M. A., Amantes, A., & Coelho, G. (2015). *Psicometria contemporânea: compreendendo os Modelos Rasch* (1.ª ed). Casa do Psicólogo.

Gomes, C. M. A. (2013). A construção de uma medida em abordagens de aprendizagem. *Psico (PUCRS. Online), 44*(2), 193-203. http://revistaseletronicas.pucrs.br/ojs/index.php/revistapsico/article/view/11371

Gomes, C. M. A. (2020). Análises estatísticas para estudos de intervenção. In M. Mansur-Alves & J. B. Lopes-Silva (Orgs.), *Intervenção cognitiva: dos conceitos às práticas baseadas em evidências para diferentes aplicações* (pp. 93-107). T.Ser.

Gomes, C. M. A., & Almeida, L. S. (2017). Advocating the broad use of the decision tree method in Education. *Practical Assessment, Research & Evaluation, 22*(10), 1-10.

Gomes, C. M. A., Almeida, L. S., & Núñez, J. C. (2017). Rationale and applicability of exploratory structural equation modeling (ESEM) in psychoeducational contexts. *Psicothema, 29*(3), 396-401. https://doi.org/10.7334/psicothema2016.369

Gomes, C.M.A., Amantes, A., & Jelihovschi, E.G. (2020). Applying the regression tree method to predict students' science achievement. *Trends in Psychology, 28*, 99-117. https://doi.org/10.9788/s43076-019-00002-5

Gomes, C. M. A., Araujo, J., Nascimento, E., & Jelihovschi, E. (2018). Routine Psychological Testing of the Individual Is Not Valid. *Psychological Reports, 122*(4), 1576-1593. https://doi.org/10.1177/0033294118785636

Gomes, C. M. A., Araujo, J., & Jelihovschi, E. G. (2020). Approaches to learning in the non-academic context: construct validity of Learning Approaches Test in Video Game (LAT-Video Game). *International Journal of Development Research, 10*(11), 41842-41849. https://doi.org/10.37118/ijdr.20350.11.2020

Gomes, C. M. A., Fleith, D. S., Marinho-Araujo, C. M., & Rabelo, M. L. (2020). Predictors of students' mathematics achievement in secondary education. *Psicologia: Teoria e Pesquisa, 36*, e3638. https://doi.org/10.1590/0102.3772e3638

Gomes, C. M. A., & Gjikuria, J. (2017). Comparing the ESEM and CFA approaches to analyze the Big Five factors. *Avaliação Psicológica, 16*(3), 261-267. https://doi.org/10.15689/ap.2017.1603.12118

Gomes, C. M. A., Golino, H. F., & Costa, B. C. G. (2013). Dynamic system approach in psychology: proposition and application in the study of emotion, appraisal and cognitive achievement. *Problems of Psychology in the 21st Century, 6*, 15-28. http://www.journals.indexcopernicus.com/abstracted.php?level=5&icid=1059487

Gomes, C. M. A., & Jelihovschi, E. (2016). Proposing a new approach and a rigorous cut-off value for identifying precognition. *Measurement, 93*, 117-125. https://doi.org/10.1016/j.measurement.2016.06.066

Gomes, C. M. A., & Jelihovschi, E. (2019). Presenting the regression tree method and its application in a large-scale educational dataset. *International Journal of Research & Method in Education*, *43*(2), 201-221. https://doi.org/10.1080/1743727X.2019.1654992

Gomes, C. M. A., Lemos, G. C., & Jelihovschi, E. G. (2020). Comparing the predictive power of the CART and CTREE algorithms. *Avaliação Psicológica, 19*(1), 87-96. https://doi.org/10.15689/ap.2020.1901.17737.10

Gomes, C. M. A., Linhares, I. S., Jelihovschi, E. G., & Rodrigues, M. N. S. (2021). Introducing rationality and content validity of SLAT-Thinking. *International Journal of Development Research, 11*(1), 43264-43272, https://doi.org/10.37118/ijdr.20586.01.2021

Gomes, C. M. A., & Nascimento, D. F. (2021). Presenting SLAT-Thinking Second Version and its contente validity. *International Journal of Development Research, 11*(3), 45590-45596. https://doi.org/10.37118/ijdr.21368.03.2021

Gomes, C. M. A., Nascimento, D., & Araujo, J. (2021a). *Medindo a Inteligência Fluida: o Teste de Indução da Bateria de Fatores Cognitivos de Alta-Ordem (BAFACALO).* Research Gate. https://doi.org/10.13140/RG.2.2.17087.84641/3

Gomes, C. M. A., Nascimento, D., & Araujo, J. (2021b). *Projeto de Testes Gratuitos e Abertos do LAICO: Teste de Velocidade Perceptiva 3 da BAFACALO*. Research Gate. https://doi.org/10.13140/RG.2.2.36278.42563/2

Gomes, C. M. A., Nascimento, D., & Araujo, J. (2021c). *Teste de Velocidade Perceptiva 2 da Bateria de Fatores Cognitivos de Alta-Ordem (BAFACALO): Disponibilização Aberta e Gratuita aos Testes de Medida de Rapidez Cognitiva do LAICO*. Research Gate. https://doi.org/10.13140/RG.2.2.29567.53928/2

Gomes, C. M. A., Nascimento, E., & Peres, A. J. S. (2019). Investigating causal relations in personality by combining path analysis and Search algoritms. *Poster. 3rd World Conference on Personality*, World Association for Personality Psychology (WAPP), Hanoi, Vietnam.

Gomes, C. M. A., & Valentini, F. (2019). Time series in educational psychology: application in the study of cognitive achievement. *European Journal of Education Studies, 6*(8), 214-229. https://doi.org/10.5281/zenodo.3551953

Härnqvist, K. (1975). The international study of educational achievement. *Review of Research in Education, 3*, 85-109. http://rre.aera.net

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

Jelihovschi, E. G., & Gomes, C. M. A. (2019). Proposing an achievement simulation methodology to allow the estimation of individual in clinical testing context. *Revista Brasileira de Biometria, 37*(4), 1-10. https://doi.org/10.28951/rbb.v37i4.423

Kuhn, M. (2017). *caret: Classification and regression training.* https://CRAN.Rproject.org/package=caret

Lantz, B. (2015). *Machine learning with R*. Packt Publishing.

Lumley, T. (2020). *Bounded memory linear and generalized linear models* [Package 'biglm']. https://cran.r-project.org/web/packages/biglm/biglm.pdf

Matos, D. A. S., Brown, G. T. L., & Gomes, C. M. A. (2019). Bifactor invariance analysis of student conceptions of assessment inventory. *Psico-USF, 24*(4), 737-750. https://doi.org/10.1590/1413-82712019240411

Ministério da Educação [MEC]/Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira [INEP] (2012). *Microdados do ENEM – 2011. Exame Nacional do Ensino Médio: Manual do Usuário*. MEC/INEP.

Ministério da Educação [MEC]/Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira [INEP]. (2013). *Exame Nacional do Ensino Médio (Enem): Relatório pedagógico 2009-2010*. INEP/MEC.

OECD (2019). *PISA 2018 Assessment and Analytical Framework*. OECD Publishing. https://doi.org/10.1787/b25efab8-en

Pazeto, T. C. B., Dias, N. M., Gomes, C. M. A., & Seabra, A. G. (2019). Prediction of arithmetic competence: role of cognitive abilities, socioeconomic variables and the perception of the teacher in early childhood education. *Estudos de Psicologia, 24*(3), 225-236. https://doi.org/10.22491/1678-4669.20190024

Pazeto, T. C. B., Dias, N. M., Gomes, C. M. A., & Seabra, A. G. (2020). Prediction of reading and writing in elementary education through early childhood education. *Psicologia: Ciência e Profissão, 40*, e205497, 1-14. https://doi.org/10.1590/1982-3703003205497

Pereira, B. L. S., Golino, M. T. S., & Gomes, C. M. A. (2019). Investigando os efeitos do programa de enriquecimento instrumental básico em um estudo de caso único. *European Journal of Education Studies, 6*(7). https://doi.org/10.5281/zenodo.3477577

Pires, A. A. M., & Gomes, C. M. A. (2017). Three mistaken procedures in the elaboration of school exams: explicitness and discussion. *PONTE International Scientific Researches Journal, 73*(3), 1-14. https://doi.org/10.21506/j.ponte.2017.3.1

Pires, A. A. M., & Gomes, C. M. A. (2018). Proposing a method to create metacognitive school exams. *European Journal of Education Studies, 5*(8), 119-142. https://doi.org/10.5281/zenodo.2313538

R Core Team (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org

Rodrigues, M. N. S., & Gomes, C. M. A. (2020). Testing the hypothesis that the deep approach generates better academic performance. *International Journal of Development Research, 10*(12), 42925-42935. https://doi.org/10.37118/ijdr.20579.12.2020

Rokach, L., & Maimon, O. (2015). *Data mining with decision trees: theory and applications*. World Scientific Publishing.

Therneau, T. M., & Atkinson, E. J. (2015). *An introduction to recursive partitioning using the rpart routines*. https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf

Zhang, H., & Singer, B. H. (2010). *Recursive partitioning and applications*. Springer.

Informação dos autores:

i Laboratory for Cognitive Architecture Mapping (LAICO), Federal University of Minas Gerais, Brazil.
https://orcid.org/0000-0003-3939-5807

ii Research Centre on Education (CIEd), Institute of Education, University of Minho, Portugal.
https://orcid.org/0000-0002-5975-2739

iii Department of Exact and Technological Sciences, Universidade Estadual de Santa Cruz (UESC), Brazil.
https://orcid.org/0000-0002-7286-1198

Toda a correspondência relativa a este artigo deve ser enviada para:
Cristiano Mauro Assis Gomes
Departamento de Psicologia, Universidade Federal de Minas Gerais, Av. Antônio Carlos, 6627, Sala 4036, Pampulha, Belo Horizonte, MG, Brasil 31270-901
cristianomaurogomes@gmail.com

Recebido em 7 de junho de 2019
Aceite para publicação em 22 de outubro de 2021

A racionalidade do Método de Regressão em Árvore é mais apropriada do que o Modelo Linear Geral para analisar dados educacionais complexos

RESUMO

Qualquer método quantitativo é formatado por certas regras ou postulados que constituem a sua própria racionalidade. Não fortuitamente, esses postulados determinam as condições e constrangimentos segundo os quais as evidências podem ser construídas. Neste artigo, argumentamos por que a racionalidade do Método de Regressão em Árvore é mais apropriada do que a do Modelo Linear Geral para analisar dados educacionais complexos. Ademais, aplicamos o algoritmo CART do Método de Regressão em Árvore, assim como a Regressão Linear Múltipla, num modelo com 53 preditores, tomando como desfecho as pontuações dos estudantes em leitura da edição de 2011 do Exame Nacional do Ensino Médio (ENEM; N = 3.670.089), o qual é um dado educacional complexo. Esta comparação empírica ilustra como o Método de Regressão em Árvore é superior ao Modelo Linear Geral para fornecer evidências sobre relações não lineares, assim como para lidar com variáveis nominais com muitas categorias, e variáveis ordinais. Concluímos que o Método de Regressão em Árvore constrói melhores evidências sobre as relações entre os preditores e o desfecho em dados complexos.

Palavras-chave: Modelo de regressão em árvore; Modelo linear geral; Exame Nacional do Ensino Médio (ENEM); Dados complexos

La racionalidad del Método de Regresión de Árbol es más apropiada que el Modelo Linear General para analizar datos educacionales complejos

RESUMEN

Cualquier método cuantitativo está conformado por ciertas reglas o postulados que constituyen su propia racionalidad. No por casualidad, estos postulados determinan las condiciones y restricciones sobre las cuales se puede construir la evidencia En este artículo, argumentamos por qué la racionalidad del Método de Árbol de Regresión es más apropiada que el Modelo Lineal General para analizar datos educativos complejos. Además, se aplicó el algoritmo CART del Método de Regresión de Árbol, así como la Regresión Linear Múltiple, en un modelo con 53 predictores, tomando como variable de respuesta el desempeño de los estudiantes en lectura de la edición 2011 del Examen Nacional de Educación Secundaria (ENEM; N = 3.670.089), que es un dato educativo complejo. Esta comparación empírica ilustra cómo el Método de Árbol de Regresión es superior al Modelo Linear General al proporcionar evidencia de relaciones no lineales, así como al tratar con variables nominales con muchas categorías y variables ordinales. Llegamos a la conclusión de que el Método de Árbol de Regresión genera mejores evidencias sobre las relaciones entre los predictores y el variable de respuesta en datos complejos.

Palabras-clave: Método de regresión de árbol; Modelo linear general; Examen Nacional de la Enseñanza Media (ENEM); Datos complejos