Comparing the Predictive Power of the CART and CTREE algorithms

Cristiano Mauro Assis Gomes[1]
Universidade Federal de Minas Gerais, Belo Horizonte-MG, Brasil
Gina C. Lemos
Universidade do Minho, Braga, Minho, Portugal
Enio G. Jelihovschi
Universidade Estadual de Santa Cruz – UESC), Ilhéus-BA, Brasil

ABSTRACT

The CART algorithm has been extensively applied in predictive studies, however, researchers argue that CART produces variable selection bias. This bias is reflected in the preference of CART in selecting predictors with large numbers of cutpoints. Considering this problem, this article compares the CART algorithm to an unbiased algorithm (CTREE), in relation to their predictive power. Both algorithms were applied to the 2011 National Exam of High School Education, which includes many categorical predictors with a large number of categories, which could produce a variable selection bias. A CTREE tree and a CART tree were generated, both with 16 leaves, from a predictive model with 53 predictors and the students' writing essay achievement as the outcome. The CART algorithm yielded a tree with a better outcome prediction. This result suggests that for large data sets, called big data, the CART algorithm might give better results than the CTREE algorithm.

Keywords: algorithms; data mining; large-scale educational assessment; machine learning; National Exam of Upper Secondary Education.

RESUMO – Comparando o Poder Preditivo dos Algoritmos CART e CTREE

O algoritmo CART tem sido aplicado de forma extensiva em estudos preditivos. Porém, pesquisadores argumentam que o CART apresenta sério viés seletivo. Esse viés aparece na preferência do CART pelos preditores com grande número de categorias. Este artigo considera esse problema e compara os algoritmos CART e CTREE, este considerado não enviesado, tomando como resultado seu poder preditivo. Os algoritmos foram aplicados no Exame Nacional do Ensino Médio de 2011, no qual estão incluídos vários preditores nominais e ordinais com muitas categorias, o que pode produzir um viés seletivo. Foram geradas uma árvore do CTREE e outra do CART, ambas com 16 folhas, provenientes de um modelo com 53 variáveis preditoras e a nota da redação, como desfecho. A árvore do algoritmo CART apresentou uma melhor predição. Para grandes bancos de dados, possivelmente o algoritmo CART é mais indicado do que o algoritmo CTREE.

Palavras-chave: algoritmos; mineração de dados; avaliação educational em larga escala; aprendizagem de máquina; Exame Nacional do Ensino Médio.

RESUMEN – Comparando el Poder Predictivo de los Algoritmos CART y CTREE

El algoritmo CART es ampliamente utilizado en análisis predictivos. Sin embargo, los investigadores argumentan que el CART presenta un fuerte sesgo de selección. Este sesgo se refleja en el CART en la preferencia de seleccionar predictores con elevado número de categorías. Teniendo en cuenta este problema, el presente artículo compara el algoritmo CART y un algoritmo imparcial (CTREE) con relación a su poder predictivo. Ambos algoritmos se aplicaron en el Examen Nacional de la Enseñanza Secundaria de 2011, incluyendo predictores nominales y ordinales con diversas categorías, un escenario susceptible de producir el sesgo de selección de variables mencionado. Fueron generados un árbol CTREE y un árbol CART, ambos con 16 hojas, provenientes de un modelo predictivo con 53 variables y la nota del comentario de texto. El árbol del algoritmo CART presentó mejor predicción. Para grandes bases de datos el algoritmo CART puede proporcionar mejores resultados que el CTREE.

Palabras clave: algoritmos; minería de datos; evaluación educativa a gran escala; aprendizaje de máquina; Examen Nacional de la Enseñanza Secundaria.

The Classification and Regression Trees (CART) algorithm is a traditional, popular, and well-developed approach of the Regression Tree Method (Loh, 2014). Statistical antecedents of CART algorithm are of

[1] Endereço para correspondência: Departamento de Psicologia, Universidade Federal de Minas Gerais. Avenida Antônio Carlos, 6627, Sala 4010, Pampulha, 31270-901, Belo Horizonte, MG. Tel.: 55 (31) 3409-5027. *E-mail* cristianomaurogomes@gmail.com

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.

historical importance since they trace back to 1960s, when the Automatic Interaction Detection (AID) algorithm was created (Morgan & Sonquist, 1963). Nevertheless, it was only in the 1980s when researchers became interested in Regression Tree, due mainly from the technical improvements achieved by the creation of the CART algorithm (Breiman, Friedman, Olshen, & Stone, 1984). Although many different algorithms performing Regression Tree (Rusch & Zeileis, 2014) have been devised, the CART algorithm continues to be prominent (Loh, 2014).

When reviewing the history of the Regression Tree Method and its algorithms, Loh (2014) argues that some well-studied and largely applied Regression Tree algorithms, e.g., CART, tend to produce the variable selection bias, which means "an artificial preference for variables offering more cutpoints - even if all variables are noise variables containing no information" (Strobl, 2014, p. 349-350). In other words, an algorithm that produces variable selection bias is one that prioritizes the predictors with large numbers of categories, even when they are not the best predictors.

According to Strobl (2014), the problem of variable selection bias has been ignored by many users of the Regression Tree Method. Biased algorithms such as the CART and C4.5 algorithms continue to be used, even though the use of more recent algorithms could mitigate this problem. One of these algorithms is the Conditional Inference Trees (CTREE), created by Hothorn, Hornik, and Zeileis (2006). The CTREE algorithm is considered unbiased because it selects the predictors through a "(...) global null hypothesis of independence between any of the m covariates and the response" (Hothorn et al., 2006, p. 2), followed by using statistical hypothesis testing and their p-values to inspect and choose the best predictors used in each split of the data and, in this way, build the tree. According to the authors: "If the global hypothesis can be rejected, we measure the association between Y and each of the covariates Xj, j = 1, . . . , m, by test statistics or P-values indicating the deviation from the partial hypotheses." (Hothorn et al., 2006, p. 3).

The current paper compares the predictive power of two algorithms: a biased one, the CART algorithm; and an unbiased one, the Conditional Inference Trees (CTREE). If the variable selection bias is significant, CART will generate more predictive noise and, consequently, a worse outcome prediction than CTREE.

Method

Dataset

We used the CART and the CTREE algorithms in the microdata from the 2011 National Exam of Upper Secondary Education (ENEM - *Exame Nacional do Ensino Médio*), which is a large-scale dataset. The ENEM is the national exam for students who complete Secondary Education in Brazil and the main assessment measure to select students for the entrance in Brazilian public universities. The ENEM's microdata is freely available to the general public by the INEP (*Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira*), and it can be downloaded at the site http://portal.inep.gov.br/web/guest/microdados (INEP, 2019). For this paper, the ENEM dataset was strategically selected, because it has many categorical (nominal and ordinal) variables (potential predictors) with many categories (cutpoints). This dataset has the potential of producing a significant variable selection bias (INEP, 2012). We chose the 2011 edition of the ENEM because its structural validity and reliability have already been evaluated (Gomes, Golino & Peres, 2016; 2018).

The sample is composed of students who completed the two days of the exam, answered the socioeconomic questionnaire, and wrote the argumentative essay. It consists of 3,670,089 students, mostly female (59.51%), single (86.23%) and Caucasian (43.51%), attending Secondary Education in public schools (75.07%), located in urban regions (97.58% of all schools) in Brazil, with a family monthly income equal or smaller than two minimum wages (74.63%).

Predictive Model

Table 1 and Table 2 show 53 variables of the 2011 ENEM's microdata (INEP, 2015), which make up the predictors of the current study. The outcome is the students' writing essay achievement score, a standardized scale produced by INEP with a mean of 500 points, a standard deviation of 100 points and a range from 0 to 1,000 points (INEP, 2015). Details of how INEP generates the students' writing essay achievement score are informed in INEP (2011).

To compare the predictive power of the CTREE and the CART algorithms, a CTREE' tree and a CART' tree were built, both with 16 leaves. We could have produced trees with either two, four, or eight leaves, instead of 16. However, this quantity of leaves allows for a good prediction performance, since the first 10 to 20 leaves from the Regression Tree algorithms are those that significantly explain the outcome variance. Any other leaf above that number usually will not explain any part of the outcome variance, and so it will not be relevant as a predictor. Moreover, the number of 16 leaves is easy to interpret and enough to get almost all information from the tree, as accounted by the outcome variance. A big tree with a large number of leaves, e.g., 2,000 leaves, will completely compromise the interpretation of the map of relations between the predictors and the outcome variable.

Table 1
Student's and School's Variables Data

Predictive variables	Type of variables	Categories
Age.	numerical (discrete)	0-100.
Sex.	nominal	0 – male. 1 – female.
Student's home location: state in Brazil.	nominal	27 states in brazil.
Completion of secondary education, or other options.	nominal	1 – already finished. 2 – 2011. 3 – after 2011. 4 – not enrolled.
School institution in secondary education in which the student finished or would finish secondary education.	nominal	1 – regular. 2 – adults. 3 – vocational. 4 – special needs.
Request for a secondary education certification.	nominal	0 – no. 1 – yes.
Request to perform the exam in braille.	nominal	0 – no. 1 – yes.
Request to perform the exam in larger letters.	nominal	0 – no. 1 – yes.
Request for reader assistance.	nominal	0 – no. 1 – yes.
Request for an easily accessible classroom.	nominal	0 – no. 1 – yes.
Ttranscript request.	nominal	0 – no. 1 – yes.
Libras request.	nominal	0 – no. 1 – yes.
Low vision indicator.	nominal	0 – no. 1 – yes.
Blindness indicator.	nominal	0 – no. 1 – yes.
Hearing-impaired indicator.	nominal	0 – no. 1 – yes.
Physical disability indicator.	nominal	0 – no. 1 – yes.
mental disability indicator.	nominal	0 – no. 1 – yes.
Attention deficit indicator.	nominal	0 – no. 1 – yes.
Dyslexia indicator.	nominal	0 – no. 1 – yes.
Indicator of pregnancy.	nominal	0 – no. 1 – yes.
Breast-feeding indicator.	nominal	0 – no. 1 – yes.
Lip reading indicator.	nominal	0 – no. 1 – yes.
Request to take the exam another day.	nominal	0 – no. 1 – yes.
Deafness indicator.	nominal	0 – no. 1 – yes.
Marital status.	nominal	0 – single. 1 – married. 2 – divorced. 3 – widow.
Declared color/race.	nominal	0 – not informed. 1 – white. 2 – black. 3 – brown. 4 – yellow. 5 – indigenous.
Student's school location: state in Brazil.	nominal	27 states in brazil.
Administrative unit of the school attended at the time of the exam.	nominal	1 – federal. 2 – state. 3 – municipal. 4 – private.
School location attended at the time of the exam.	nominal	1 – urban. 2 – rural.
School functioning attended by the student when he/she performed the exam.	nominal	1 – functioning. 2 – not functioning. 3 – closed. 4 – closed in previous years.

Table 2
Data from the Socioeconomic Questionnaire

Predictive variables	Type of variables	Categories
q01 – number of people that live with the student.	ordinal	1 = 1 person; 2 = 2 persons; (...) 20 = 20 or more persons.
q02 – student's father's education.	ordinal	a – no education. b – 1-4 degree. c – 5-8 degree. d – incomplete secondary education. e – secondary education. f – incomplete higher education. g – higher education. h – post-graduation.
q03 – student's mother's education.	ordinal	the same categories of q02.
q04 – student's family monthly income.	ordinal	a – no income. b – until 1 minimum wage. c – 1-1.5 minimum wage. d – 1.5-2 minimum wages. e – 2-5 minimum wages. f – 5-7 minimum wages. g – 7-10 minimum wages. h – 10-12 minimum wages. i – 12-15 minimum wages. j – 15-30 minimum wages. k – above 30 minimum wages.
q05 – student's monthly income.	ordinal	the same categories of q04.
q06 – student's home (own home or other options.	nominal	a – paid. b – financed. c – rented. d – ceded.
q07 – student's home location (urban or other options).	nominal	a – rural. b – urban. c – indigenous community. d – quilombola community.
q08 – student's paid activity.	nominal	a – yes. b – no.
q15 – courses attended by the student: vocational course.	nominal	a – yes. b – no.
q16 – courses attended by the student: preparation for the higher education admission exam.	nominal	a – yes. b – no.
q17 – courses attended by the student: higher education.	nominal	a – yes. b – no.
q18 – courses attended by the student: second language.	nominal	a – yes. b – no.
q19 – courses attended by the student: informatics.	nominal	a – yes. b – no.
q20 – courses attended by the student: preparation for public tender.	nominal	a – yes. b – no.
q24 – motivation to take the exam: to test personal knowledge.	ordinal	from 0 to 5.
q25 – motivation to take the exam: to pursue studies in higher education.	ordinal	from 0 to 5.
q26 – motivation to take the exam: to obtain a secondary education certificate.	ordinal	from 0 to 5.
q27 – motivation to take the exam: to obtain a scholarship.	ordinal	from 0 to 5.
q28 – years taken to complete elementary education.	ordinal	a – < 8 years. b – 8 years. c – 9 years. d – 10 years. e – 11 years. f – > 11 years. g – not finished.

Table 2 (continuation)
Data from the Socioeconomic Questionnaire

Predictive variables	Type of variables	Categories
q29 – hiatus in studies during elementary education.	ordinal	a – no. b – 1 year. c – 2 years. d – 3 years. e – 4 or more years.
q30 – type of school attended in elementary education.	nominal	a – only public. b – majority in public. c – only private. d – majority in private. e – only indigenous. f – majority indigenous. g – only quilombola community. h – majority quilombola community. i – not attended.
q32 – hiatus in studies during secondary education.	ordinal	the same categories of q29.
q33 – type of school attended in secondary education.	nominal	the same categories of q30.

Implementation of CART and CTREE algorithms

The CART algorithm was performed through the rpart R package (Therneau & Atkinson, 2015), version 4.1-13 (Therneau, Atkinson, & Ripley, 2018), while the CTREE algorithm was performed through the partykit R package, version 1.2-2 (Hothorn, Seibold, & Zeileis, 2018). The predictive power analysis of both algorithms was performed through the caret R package, version 6.0-80 (Kuhn, 2018). Statistical technical details of CART algorithm and CTREE algorithm are presented, respectively, in Breiman et al. (1984) and Hothorn and Zeileis (2015), as well as in Hothorn et al. (2006).

The sample of this study was randomly split into two parts, a learning sample (75% of the sample) and a test sample (25% of the sample). In both algorithms, the learning sample was used to build the predictive model, while the test sample was used to test the performance of each test as a predictive model. In the CART algorithm, we applied the cross-validation 3-Fold technique to the learning sample. This strategy was not applied to the CTREE algorithm since the pruning strategy is not necessary for this algorithm (Hothorn et al., 2006). To evaluate the CART tree and the CTREE tree predictive power, the outcome variance explained by these trees in the test sample was tested through R^2 index. In sum, all the stated technical steps in this study follow the essential recommendations of the Machine Learning field and the Regression Tree Method literature, regarding the use of the Regression Tree algorithms (Geurts, Irrthum, & Wehenkel, 2009; Gomes & Almeida, 2017; James, Witten, Hastie, & Tibshirani, 2013; Lantz, 2015; Rokach &Maimon, 2015).

The R code used to perform the CART tree of 16 leaves, as well as to test its predictive power in the test sample is the following:

```
library("rpart")
library("rpart.plot")

cfitpruningsplits <- rpart (`Writing essay Score` ~. - ID - `Mathematics Score` - `Human Sciences Score` - `Nature
Sciences Score` - `Languages Score`,
                  na.action = na.rpart,
                  data = learning,
                  method = 'anova',
                  control = rpart.control(xval = 3,
                                          cp = 1.18e-03,
                                          minsplit = 100))
library("caret")
predtestingpruningsplits <- predict(cfitpruningsplits,newdata=testing,type="vector")
accuracy <- R2(predtestingpruningsplits,testing$'Writing essay Score',formula="traditional")
accuracy
```

The R code used to perform the CTREE tree of 16 leaves, as well to investigate its predictive power in the test sample is on the next following:

```
library("partykit")
ctredacaokit <- ctree(`Writing essay Score` ~.,
              data = learning[ , !(names(learninging) %in% c('ID','Mathematics Score','Human Sciences
Score','Nature Sciences Score','Languages Score'))],
               control = ctree_control(teststat = "quad",
              testtype="Bonferroni",
            minsplit=100,
            minbucket=7,
            maxdepth=4))
library("caret")
predtesting <- predict(ctredacaokit,newdata=testing,type="response")
accuracy <- R2(predtesting,testing$'Writing essay Score',formula="traditional")
accuracy
```

Results and Discussion

Before reporting the results of the CART and the CTREE algorithms, we will briefly show some descriptive results, regarding the students' writing essay achievement in the learning sample and test sample. The learning sample scores have an average of 545.21 points and standard deviation of 146.45 points, they range from 40 points to 1000 points, and 95% of them range from 260 points to 840 points. Their distribution has the skew=0.08 and kurtosis=0.03, which may be considered to have a normal distribution. The test sample scores have average of 545.17 points and standard deviation of 146.48 points, they range from 40 to 1000 points, and 95% of them ranges from 260 to 840 points. They may also be considered to have a normal distribution with skew=0.09 and kurtosis=0.03.

The CART tree predicted 11.51% of the outcome variance in the test sample, while the CTREE tree predicted 3.31% of the outcome variance in the test sample, which shows a clear difference in predictive power. Figure 1 shows the CART tree and Figure 2 shows the CTREE tree. We must point out that in the CART tree the categories described in the predictor used to split a node go to the left of the reader after the split is done (see the term yes in Figure 1). For example, the first node has the split based on the variable "Type of school attended in Secondary Education," the listed categories used to split the node are: A – Only Public, B – Majority in Public, E – Only Indigenous, F – Majority Indigenous, G – Only Quilombola Community, H – Majority Quilombola Community, I – Not Attended., they go to the node (2). The remaining categories: C – Only Private, D – Majority in Private, go to the node (3). Node (3) has its split based on variable "Sex." The category 0 – male, which is listed, go to node (6) and the category 1 – female go to node (7).

Figures 1 and 2 display two completely different results and, undoubtedly, CART tree is much better than CTREE. Everyone familiar with the reality of Brazilian school knows about the large gap between public and private schools. Many studies point out that the students' achievement from private schools outperform the students' achievement from public schools in Brazil (i. e. Figueirêdo, Nogueira & Santana, 2014; Moraes & Beluzzo, 2014) as well in many other developing countries (Ashley et al., 2004; Rutkowski & Rutkowski, 2009). Thus, it makes sense that public versus private school should be an important predictor in our study and indeed CART chose it as the predictor for its first split, which is considered to be the most important split of a tree. Besides, the CART tree uses many sensible predictors, including age and sex. The CTREE algorithm uses only two predictors; age and sex, but they should not be so important, because male and female divide the population of the students in half at every level either economic or academic, on the other hand, most of the students taking the ENEM are between 17 to 21 years old, so that age must be a week predictor. CART algorithm only uses age as a predictor at the end of the tree and divides it in "above and below 30 years old".

To sum up, from the 53 predictors included in the predictive model of students' writing essay achievement, the CART algorithm selected nine predictors: the type of school attended by the students in Secondary Education, the number of years taken by students to complete Elementary Education, the students' request for a Secondary Education certification, the students' motivation to take the exam to obtain a Secondary Education certificate, the students' motivation to take the exam to obtain a scholarship, the students' family monthly income, and the course attended by the student in Higher Education, age and sex. The CTREE selected only two: age and sex. Also, the CART tree was more informative than the CTREE tree not only in the outcome prediction, but also in the map of relations between predictors and the outcome itself.

Figure 2 shows that the CTREE tree included a nominal predictor, sex, with only two categories and a discrete numerical predictor: age (see Table 1). CART tree selected a large scope of predictors. For example, the CART algorithm included the predictor type of school attended by the students in Secondary Education, a nominal variable with nine categories, as well as the predictor student's family monthly income, an ordinal variable with 11 categories (see Table 2). Usually, the first predictors used to split the nodes of the tree are considered the

most important ones. It is relevant to observe that the CTREE tree selected the variable age, a discrete numerical variable, to split the root node, while the CART tree selected the variable type of school attended by the students in Secondary Education, a nominal variable with 9 categories.

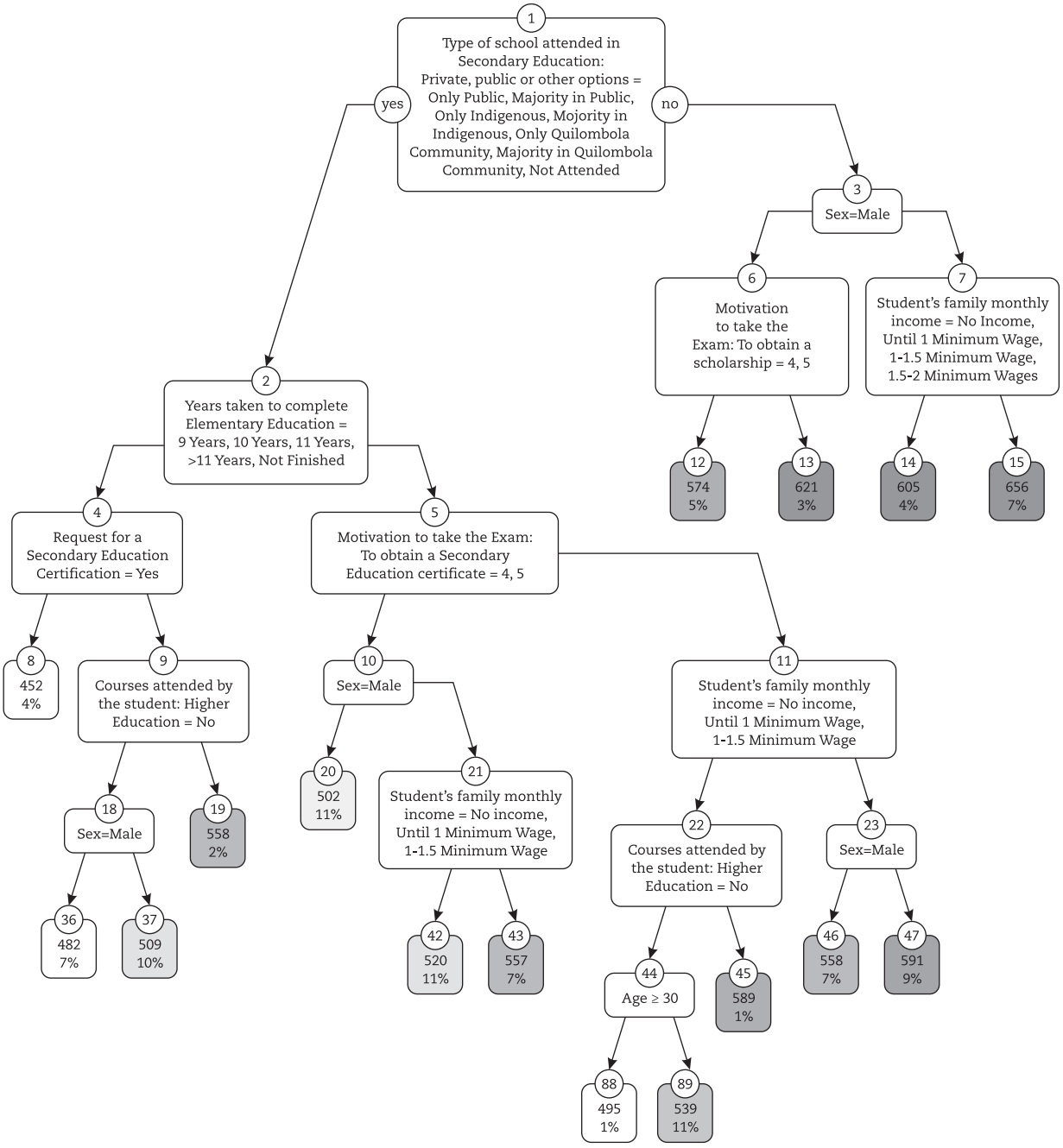

Figura 1. The CART tree of 16 leaves

The prediction of CART, which has a better prediction than CTREE, account for only 11.51% of the outcome variance, this is a low predictive value for the students' writing essay achievement. We believe that this low prediction was not caused by the algorithms themselves, since the machine learning literature defines the CART algorithm as a greedy technique to identify variables, linear and non-linear relationships, that are relevant for the prediction of the outcome variable. In our study, we used in the predictive model

all variables of the 2011 ENEM microdata that could serve as predictors of the outcome, excluding only the students' scores in the domains of mathematics, languages, human sciences, and natural sciences. However, the ENEM microdata collects basically socioeconomic and demographic information about the students and their schools, as well as their motivation to perform the exam. Many psychological variables which are predictors of the students' performance are not present in the ENEM microdata, as is the case of reasoning (Gomes, 2010a), intelligence (Gomes & Golino, 2012), students' learning approaches (Gomes, 2010b), and metacognition (Gomes, Golino & Menezes, 2014). So, we conclude that the ENEM's 2011 microdata are not very explanatory of the students' writing essay achievement.

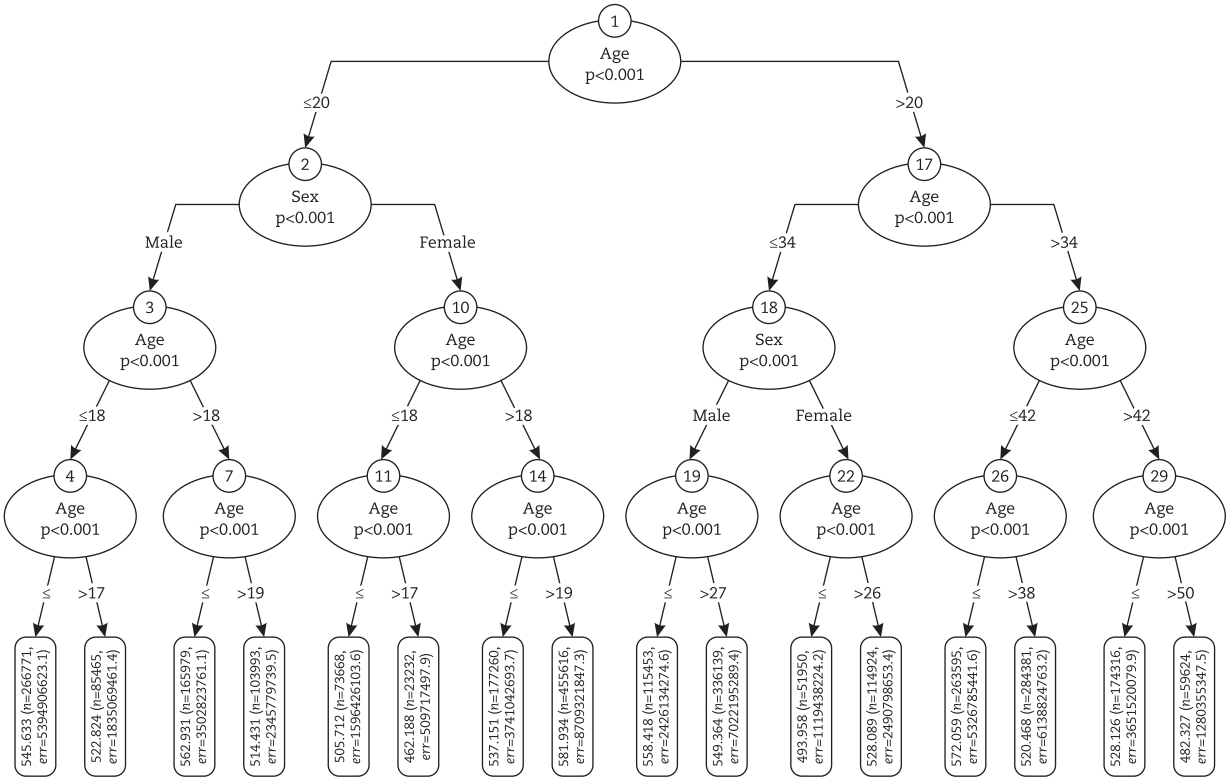

Figura 2. The CTREE tree of 16 leaves

Conclusion

As an additional evidence, we ran the CART and the CTREE algorithms without the limit of 16 leaves. CART algorithm led to the best predictive tree by using the cost complexity criterion, as recommended by the literature (Breiman et al., 1984), and found a pruned tree with 1,210 leaves, which explained 16.57% of the outcome variance in the test sample. In the case of CTREE, we did not impose any limit of leaves allowing the tree to grow to the maximum of its predictive power. To do so, we eliminated the R command "maxdepth=4" previously applied when we performed the tree with 16 leaves. This tree, with 39 leaves, explained only 3.37% of the outcome variance in the test sample, a very similar result obtained by the CTREE tree with 16 leaves, which explained 3.31% of the outcome variance in the test sample. The CTREE chose only two predictors when 16 leaves were used, therefore it is very plausible that it will choose only those two predictors whatever number of leaves its trees have, which will not improve its performance. Furthermore, this brings evidence that our strategy of forming trees with 16 leaves did not produce any artifact or any positive bias favoring the CART algorithm.

In conclusion, the present study has shown that the CART tree displayed a better outcome prediction in comparison to the CTREE tree. The most striking observation emerging from this study is that the CTREE tree only selected two predictors (age and sex), while the CART tree selected nine predictors to explain the outcome variance. It is also very interesting that the CTREE algorithm prioritized numerical or nominal variables with few categories, while the CART algorithm

was more eclectic, choosing numerical, ordinal, and nominal variables, with a diverse number of cutpoints.

It should be noted that, in seeking to avoid the variable selection bias, the CTREE algorithm may have worsen its prediction. Hothorn et al. (2006) give some examples of CTREE use and performance, with very good results, but the sample sizes of those examples were small. For small sample sizes CTREE indeed give good results, trees with a good prediction, using the right number of predictors. On the other hand, very large sample sizes, big data, belong to different realm, the difference between biased and unbiased prediction fades away. The permutation and likelihood ratio tests, which are designed for small to medium size samples may not work with big data. The splitting criteria of CART, which depends only on sum of squares, perform well with big data as well as small data, in which case CTREE might outperform CART nevertheless, as this paper shows, it does not work with big data. Further research might shed more light on the result suggested in this study.

While this study is not direct evidence that either CART or other old algorithms are better than more recent unbiased algorithms, it nonetheless offers some insight into the problem of variable selection bias, that is neither an easy question nor a question with a final answer.

References

Ashley, L. D., Mcloughlin, C., Aslam, M., Engel, J., Wales, J., Rawal, S., Batley, R., Kingdon, G., Nicolai, S., & Rose, P. (2014). *The role and impact of private schools in developing countries: A rigorous review of the evidence. Final report. Education Rigorous Literature Review.* Department for International Development. Retrieved from http://eppi.ioe.ac.uk/

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

Figueirêdo, E., Nogueira, L., & Santana, F. L. (2014). Igualdade de oportunidades: Analisando o papel das circunstâncias no desempenho do ENEM. *Revista Brasileira de Economia, 68*(3), 373-392. doi: 10.1590/S0034-71402014000300005

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. doi: 10.1039/b907946g

Gomes, C. M. A. (2010a). Avaliando a avaliação escolar: Notas escolares e inteligência fluida. *Psicologia em Estudo, 15*(4), 841-849. Retrieved from http://www.scielo.br/pdf/pe/v15n4/v15n4a19.pdf

Gomes, C. M. A. (2010b). Perfis de estudantes e a relação entre as abordagens de aprendizagem e rendimento escolar. *Psico-RS, 41*(4), 503-509. Retrieved from http://revistaseletronicas.pucrs.br/ojs/index.php/revistapsico/article/view/6336

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. Retrieved from http://pareonline.net/getvn.asp?v=22&n=10

Gomes, C. M. A., & Golino, H. F. (2012). O que a inteligência prediz: Diferenças individuais ou diferenças no desenvolvimento acadêmico? *Psicologia: Teoria e Prática, 14*(1), 126-139. Retrieved from http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1516-36872012000100010&lng=pt&tlng=pt

Gomes, C. M. A., Golino, H. F., & Menezes, I. G. (2014). Predicting school achievement rather than intelligence: Does metacognition matter? *Psychology, 5*(9), 1095-1110. doi: 1.4236/psych.2014.59122

Gomes, C. M. A., Golino, H. F., & Peres, A. J. S. (2016). Investigando a validade estrutural das competências do ENEM: Quatro domínios correlacionados ou um modelo bifatorial? *Boletim na Medida/INEP, 5*(10), 33-38. Retrieved from http://portal.inep.gov.br/informacao-da-publicacao/-/asset_publisher/6JYIsGMAMkW1/document/id/587206

Gomes, C. M. A., Golino, H. F., & Peres, A. J. S. (2018). Análise da fidedignidade composta dos escores do ENEM por meio da análise fatorial de itens. *European Journal of Education Studies, 5*(8), 331-344. doi: 10.5281/zenodo.2527904

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics, 15*(3), 651-674. doi: 10.1180/106186006X133933

Hothorn, T., Seibold, H., & Zeileis, A. (2018). *Partykit: A toolkit for recursive partytioning.* (Version 1.2-2) [Software]. Available from https://CRAN.R-project.org/package=partykit

Hothorn, T., & Zeileis, A. (2015). Partykit: A modular toolkit for recursive partytioning in R. *Journal of Machine Learning Research, 16*, 3905–3909.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira [INEP]. (2011). *Nota técnica: procedimento de cálculo das notas do ENEM.* Brasília: MEC/INEP. Retrieved from http://download.inep.gov.br/educacao_basica/enem/nota_tecnica/2011/nota_tecnica_procedimento_de_calculo_das_notas_enem_2.pdf

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira [INEP]. (2012). *Microdados do ENEM – 2011. Exame Nacional do Ensino Médio: Manual do Usuário.* Brasília: MEC/INEP. Retrieved from http://portal.inep.gov.br/web/guest/microdados

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira [INEP]. (2015). *Relatório pedagógico: Enem 2011-2012.* Brasília: INEP. Retrieved from http://www.publicacoes.inep.gov.br/portal/download/1401

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira [INEP]. (2019). *Microdados [Data file].* Retrieved from http://portal.inep.gov.br/web/guest/microdados

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

Kuhn, M. (2018). *Caret: Classification and regression learning.* (Version 6.0-80) [Software]. Available from https://CRAN.Rproject.org/package=caret

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

Loh, W.-Y. (2014). Fifty years of classification and regression trees. *International Statistical Review, 82*(3), 329-348. doi: 10.1111/insr.12016.
Moraes, A. G. E. de, & Belluzzo, W. (2014). O diferencial de desempenho escolar entre escolas públicas e privadas no Brasil. *Nova Economia, 24*(2), 409-430. doi: 10.1590/0103-6351/1564
Morgan, J. N., & Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association, 58*(302), 415–434. Retrieved from https://cs.nyu.edu/~roweis/csc2515.../morgan_sonquist63.pdf
Rokach, L., & Maimon, O. (2015). *Data mining with decision trees: Theory and applications.* Singapore: World Scientific Publishing.
Rusch, T., & Zeileis, A. (2014). Comments on fifty years of classification and regression trees. *International Statistical Review, 82*(3), 361-367. doi: 10.1111/insr.12062
Rutkowski, L., & Rutkowski, D. J. (2009). *Private and public education: A cross-national exploration with TIMSS 2003.* Paper presented at the Annual Conference of the American Educational Research Association.
Strobl, C. (2014). Comments on fifty years of classification and regression trees. *International Statistical Review, 82*(3), 349-352. doi: 10.1111/insr.12059
Therneau, T., & Atkinson, B. (2015). *An introduction to recursive partitioning using the rpart routines.* Retrieved from https://cran.rproject.org/web/packages/rpart/vignettes/longintro.pdf
Therneau, T., Atkinson, B., & Ripley. B. (2018). *The rpart package.* (Version 4.1-13) [Software]. Available from https://cran.r-project.org/package=rpart

recebido em fevereiro de 2019
aprovado em agosto de 2019

Sobre os autores

Cristiano Mauro Assis Gomes is Professor of Universidade Federal de Minas Gerais. Head of Laboratory for Cognitive Architecture Mapping (LAICO)/Federal University of Minas Gerais/Brazil. ORCID: http://orcid.org/0000-0003-3939-5807

Gina C. Lemos is Researcher of the Research Centre on Education (CIEd), Institute of Education, University of Minho, Portugal. ORCID: http://orcid.org/0000-0002-5975-2739

Enio G. Jelihovschi is Professor of Universidade Estadual de Santa Cruz (UESC). Campus Soane Nazaré de Andrade. ORCID: http://orcid.org/0000-0002-7286-1198.