REVISTA E-PSI
REVISTA ELETRÓNICA DE PSICOLOGIA, EDUCAÇÃO E SAÚDE
ANO 4, VOLUME 1, 2014, pp.68-101.
https://www.revistaepsi.com
ISNN 2182-7591

Four Machine Learning Methods to Predict Academic Achievement of College Students: A Comparison Study

[*Quatro Métodos de Machine Learning para Predizer o Desempenho Acadêmico de Estudantes Universitários: Um Estudo Comparativo*]

HUDSON F. GOLINO[1], & CRISTIANO MAURO A. GOMES[2]

Abstract

The present study investigates the prediction of academic achievement (high vs. low) through four machine learning models (learning trees, bagging, Random Forest and Boosting) using several psychological and educational tests and scales in the following domains: intelligence, metacognition, basic educational background, learning approaches and basic cognitive processing. The sample was composed by 77 college students (55% woman) enrolled in the 2nd and 3rd year of a private Medical School from the state of Minas Gerais, Brazil. The sample was randomly split into training and testing set for cross validation. In the training set the prediction total accuracy ranged from of 65% (bagging model) to 92.50% (boosting model), while the sensitivity ranged from 57.90% (learning tree) to 90% (boosting model) and the specificity ranged from 66.70% (bagging model) to 95% (boosting model). The difference between the predictive performance of each model in training set and in the testing set varied from -2.60% to 23.10% in terms of the total accuracy, from -5.60% to 27.50% in the sensitivity index and from 0% to 20% in terms of specificity, for the bagging and the boosting models respectively. This result shows that these machine learning models can be used to achieve high accurate predictions of academic achievement, but the difference in the predictive performance from the training set to the test set indicates that some models are more stable than the others in terms of predictive performance (total accuracy, sensitivity and specificity). The advantages of the tree-based machine

[1] Faculdade Independente do Nordeste (BR). Universidade Federal de Minas Gerais (BR). E-mail: hfgolino@gmail.com.
[2] Universidade Federal de Minas Gerais (BR). E-mail: cristianogomes@ufmg.br.

ISNN 2182-7591

learning models in the prediction of academic achievement will be presented and discussed throughout the paper.

Keywords: Higher Education; Machine Learning; academic achievement; prediction.

Introduction

The usual methods employed to assess the relationship between psychological constructs and academic achievement are correlation coefficients, linear and logistic regression analysis, ANOVA, MANOVA, structural equation modelling, among other techniques. Correlation is not used in the prediction process, but provides information regarding the direction and strength of the relation between psychological and educational constructs with academic achievement. In spite of being useful, correlation is not an accurate technique to report if one variable is a good or a bad predictor of another variable. If two variables present a small or non-statistically significant correlation coefficient, it does not necessarily means that one can't be used to predict the other.

In spite of the high level of prediction accuracy, the artificial neural network models do not easily allows the identification of how the predictors are related in the explanation of the academic outcome. This is one of the main criticisms pointed by researchers against the application of Machine Learning methods in the prediction of academic achievement, as pointed by Edelsbrunner and Schneider (2013). However, their Machine Learning methods, as the *learning tree models*, can achieve a high level of prediction accuracy, but also provide more accessible ways to identify the relationship between the predictors of the academic achievement.

REVISTA E-PSI
REVISTA ELETRÓNICA DE PSICOLOGIA, EDUCAÇÃO E SAÚDE
ANO 4, VOLUME 1, 2014, pp.68-101.
https://www.revistaepsi.com
ISNN 2182-7591

Table 1 – Usual techniques for assessing the relationship between academic achievement and psychological/educational constructs and its basic assumptions.

Technique	Main Assumptions							
	Distribution	Relationship between variables	Homoscedasticity?	Sensible to outliers?	Independence?	Sensible to Collinearity	Demands a high sample-to-predictor ratio?	Sensible to missingness?
Correlation	Bivariate Normal	Linear	Yes	Yes	NA	NA	NA	Yes
Simple Linear Regression	Normal	Linear	Yes	Yes	Predictors are independent	NA	Yes	Yes
Multiple Regression	Normal	Linear	Yes	Yes	Predictors are independent/Errors are independent	Yes	Yes	Yes
ANOVA	Normal	Linear	Yes	Yes	Predictors are independent	Yes	Yes	Yes
MANOVA	Normal	Linear	Yes	Yes	Predictors are independent	Yes	Yes	Yes
Logistic Regression	True conditional probabilities are a logistic function of the independent variables	Independent variables are not linear combinations of each other	No	Yes	Predictors are independent	NA	Yes	Yes
Structural Equation Modelling	Normality of univariate distributions	Linear relation between every bivariate comparisons	Yes	Yes	NA	NA	Yes	Yes

The goal of the present paper is to introduce the basic ideas of four specific *learning tree's* models: single learning trees, bagging, Random Forest and Boosting. These techniques will be applied to predict academic achievement of college students (high achievement *vs.* low achievement) using the result of an intelligence test, a basic cognitive processing battery, a high school knowledge exam, two metacognitive scales and one learning approaches' scale. The tree algorithms do not make any assumption regarding normality, linearity of the relation between variables, homoscedasticity,

ISNN 2182-7591

collinearity or independency (Geurts, Irrthum, & Wehenkel, 2009). They also do not demand a high sample-to-predictor ratio and are more suitable to interaction effects than the classical techniques pointed before. These techniques can provide insightful evidences regarding the relationship of educational and psychological tests and scales in the prediction of academic achievement. They can also lead to improvements in the predictive accuracy of academic achievement, since they are known as the state-of-the-art methods in terms of prediction accuracy (Geurts et al., 2009; Flach, 2012).

Presenting New Approaches to Predict Academic Achievement

Machine learning is a relatively new science field composed by a broad class of computational and statistical methods used to extract a model from a system of observations or measurements (Geurts et al., 2009; Hastie, Tibshirani, & Friedman, 2009). The extraction of a model from the sole observations can be used to accomplish different kind of tasks for predictions, inferences, and knowledge discovery (Geurts et al., 2009; Flach, 2012).

Machine Learning techniques are divided in two main areas that accomplish different kinds of tasks: unsupervised and supervised learning. In the unsupervised learning field the goal is to discover, to detect or to learn relationships, structures, trends or patterns in data. There is a d-vector of observations or measurements of features, $\mathfrak{X} = \mathfrak{F}_1 \times \mathfrak{F}_2 \times \mathfrak{F}_3 \times \ldots \times \mathfrak{F}_d$, but no previously known outcome, or no associated response (Flach, 2012; James, Witten, Hastie, & Tibshirani, 2013). The features $\mathfrak{F}$ can be of any kind: nominal, ordinal, interval or ratio.

In the supervised learning field, by its turn, for each observation of the predictor (or independent variable) x_i, $i = 1, \ldots, n$, there is an associated response or outcome y_i. The vector x_i belongs to the feature space $\mathfrak{X}$, $x_i \in \mathfrak{X}$, and the vector y_i belongs to the output space $\mathfrak{Y}$, $y_i \in \mathfrak{Y}$. The task can be a regression or a classification. Regression is used when the outcome has an interval or ratio nature, and classification is used when the outcome variable has a categorical nature. When the task is of *classification* (e.g. classifying people into a high or low academic achievement group), the goal is to construct a labeling function (l) that maps the feature space into the output space

ISNN 2182-7591

composed by a small and finite set of classes $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k\}$, so that $l: \mathfrak{X} \rightarrow \mathcal{C}$. In this case the output space *is* the set of finite classes: $\mathfrak{F} \equiv \mathcal{C}$. In sum, in the classification problem a categorical outcome (e.g. high or low academic achievement), is predicted using a set of features (or predictors, independent variables). In the regression task, the value of an outcome in interval or ratio scale (for example the Rasch score of an intelligence test) is predicted using a set of features. The present paper will focus in the classification task.

From among the classification methods of Machine Learning, the *tree based models* are supervised learning techniques of special interest for the education research field, since it is useful: *1)* to discover which variable, or combination of variables, better predicts a given outcome (e.g. high or low academic achievement); *2)* to identify the cutoff points for each variable that are maximally predictive of the outcome; and *3)* to study the interaction effects of the independent variables that lead to the purest prediction of the outcome.

A classification tree partitions the feature space into several R_m distinct mutually exclusive regions (non-overlapping). Each region is fitted with a specific model that performs the labeling function, designating one of the $\mathcal{C}_k$ classes to that particular space. The class is assigned to the R_m region of the feature space by identifying the majority class in that region. In order to arrive in a solution that best separates the entire feature space into more pure nodes (regions), recursive binary partitions is used. A node is considered pure when 100% of the cases are of the same class, for example, low academic achievement. A node with 90% of low achievement and 10% of high achievement students is more "pure" then a node with 50% of each. Recursive binary partitions work as follows. The feature space is split into two regions using a specific cutoff from the variable of the feature space (x_i) that leads to the most purity configuration. Then, each region of the tree is modeled accordingly to the majority class. Then one or two original nodes are split into more nodes, using some of the given predictor variables that provide the best fit possible. This splitting process continues until the feature space achieves the most purity configuration possible, with R_m regions or nodes classified with a distinct $\mathcal{C}_k$ class. Learning trees have two main basic tuning parameters (for more fine grained tuning parameters see Breiman, Friedman, Olshen &

REVISTA E-PSI
REVISTA ELETRÓNICA DE PSICOLOGIA, EDUCAÇÃO E SAÚDE
ANO 4, VOLUME 1, 2014, pp.68-101.
https://www.revistaepsi.com
ISNN 2182-7591

Stone, 1984): *1)* the number of features used in the prediction $n(\mathfrak{X})$, and *2)* the complexity of the tree, which is the number of possible terminal nodes $\alpha|T|$.

If more than one predictor is given, then the selection of each variable used to split the nodes will be given by the variable that splits the feature space into the most purity configuration. It is important to point that in a classification tree, the first split indicates the most important variable, or feature, in the prediction. Leek (2013) synthesizes how the tree algorithm works as follow: *1)* iteratively split variables into groups; *2)* split the data where it is maximally predictive; and *3)* maximize the amount of homogeneity in each group.

The quality of the predictions made using single learning trees can verified using the misclassification error rate and the residual mean deviance (Hastie et al., 2009). In order to calculate both indexes, we first need to compute the proportion of class $\mathcal{C}_k$in the node m. As pointed before, the class to be assigned to a particular region or node will be the one with the greater proportion in that node. Mathematically, the proportion of class $\mathcal{C}_k$ in a node m of the region R_m, with N_m people is:

$$\hat{p}_{m\mathcal{C}_k} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = \mathcal{C}_k)$$

The labeling function that will assign a $\mathcal{C}_k$ class to a node m is: $max_{\mathcal{C}_k} \hat{p}_{m\mathcal{C}_k}$. The misclassification error is simply the proportion of cases or observations that do not belong to the $\mathcal{C}_k$ class in the m region:

$$\frac{1}{N_m} \sum_{x_i \in R_m} I(y_i \neq \mathcal{C}_k) = 1 - \hat{p}_{m\mathcal{C}_k}$$

and the residual mean deviance is given by the following formula:

$$-2 \sum_m \sum_{\mathcal{C}_k} n_{m\mathcal{C}_k} \log \hat{p}_{m\mathcal{C}_k} \Big/ n - |T|$$

REVISTA E-PSI
REVISTA ELETRÓNICA DE PSICOLOGIA, EDUCAÇÃO E SAÚDE
ANO 4, VOLUME 1, 2014, pp.68-101.
https://www.revistaepsi.com
ISNN 2182-7591

where n_{mc_k} is the number of people (or cases/observations) from the $\mathcal{C}_k$ class in the m region, n is the size of the sample, and $|T|$ is the number of terminal nodes (James et al., 2013).

Deviance is preferable to misclassification error because is more sensitive to node purity. For example, let's suppose that two trees (A and B) have 800 observations each, of high and low achievement students (50% in each class). Tree A have two nodes, being A_1 with 300 high and 100 low achievement students, and A_2 with 100 high and 300 low achievement students. Tree B also have two nodes: B_1 with 200 high and 400 low, and B_2 with 200 high and zero low achievement students. The misclassification error rate for tree A and B are equal (.25). However, tree B produced more pure nodes, since node B_2 is entirely composed by high achievement people, thus it will present a smaller deviance than tree A. A pseudo R^2 for the tree model can also be calculated using the deviance:

$$\text{Pseudo } R^2 = 1 - \left(Deviance / Null\ Deviance\right).$$

Geurts, Irrthum and Wehenkel (2009) argue that learning trees are among the most popular algorithms of Machine Learning due to three main characteristics: interpretability, flexibility and ease of use. Interpretability means that the model constructed to map the feature space into the output space is easy to understand, since it is a roadmap of if-then rules. James, Witten, Hastie and Tibshirani (2013) points that the tree models are easier to explain to people than linear regression, since it mirrors more the human decision-making then other predictive models. Flexibility means that the tree techniques are applicable to a wide range of problems, handles different kind of variables (including nominal, ordinal, interval and ratio scales), are non-parametric techniques and does not make any assumption regarding normality, linearity or independency (Geurts et al., 2009). Furthermore, it is sensible to the impact of additional variables to the model, being especially relevant to the study of incremental validity. It also assesses which variable or combination of them, better predicts a given outcome, as well as calculates which cutoff values are maximally predictive of it.

ISNN 2182-7591

Finally, the ease of use means that the tree based techniques are computationally simple, yet powerful.

In spite of the qualities of the learning trees pointed above, the techniques suffer from two related limitations. The first one is known as the overfitting issue. Since the feature space is linked to the output space by recursive binary partitions, the tree models can learn *too much* from data, modeling it in such a way that may turn out a sample dependent model. Being sample dependent, in the sense that the partitioning is too suitable to the data set in hand, it will tend to behave poorly in new data sets. The second issue is exactly a consequence of the overfitting, and is known as the variance issue. The predictive error in a training set, a set of features and outputs used to grown a classification tree for the first time, may be very different from the predictive error in a new test set. In the presence of overfitting, the errors will present a large variance from the training set to the test set used. Additionally, the classification tree does not have the same predictive accuracy as other classical Machine Learning approaches (James et al., 2013). In order to prevent overfitting, the variance issue and also to increase the prediction accuracy of the classification trees, a strategy named *ensemble techniques* can be used.

Ensemble techniques are simply the junction of several trees to perform the classification task based on the prediction made by every single tree. There are three main ensemble techniques to classification trees: bagging, Random Forest and boosting. The first two techniques increases prediction accuracy and decreases variance between data sets as well as avoid overfitting. The boosting technique, by its turn, only increases accuracy but can lead to overfitting (James et al., 2013).

Bagging (Breiman, 2001b) is the short hand for *bootstrap aggregating*, and is a general procedure for reducing the variance of classification trees (Hastie et al., 2009; Flach, 2012; James et al., 2013). The procedure generates B_j different bootstraps from the training set, growing a tree that assign a $\mathcal{C}_k$ class to the R_m regions of the feature space for every j. Lastly, the k class of m regions of each B tree is recorded and the majority vote is taken (Hastie et al., 2009; James et al., 2013). The majority vote is simply the most commonly occurring class over all B trees. As the bagged trees does not use the entire observations (only a bootstrapped subsample of it, usually 2/3), the remaining observations (known as *out-of-bag*, or OOB) is used to verify the accuracy of

ISNN 2182-7591

the prediction. The out-of-bag error can be computed as a «valid estimate of the test error for the bagged model, since the response for each observation is predicted using only the trees that were not fit using that observation» (James et al., 2013, p.323). Bagged trees have two main basic tuning parameters: *1)* the number of features used in the prediction, $n(\mathfrak{X})$, is set as the total number of predictors in the feature space, and *2)* the size *j* of the bootstrap set B, which is equal the number of trees to grow.

The second ensemble technique is the Random Forest (Breiman, 2001a). Random Forest differs from bagging since the first takes a random subsample n of the original data set N with replacement to growing the trees, as well as selects a subsample $\mathfrak{X}_m$ of the feature space $\mathfrak{X}$ at each node, so that the number of the selected features (variables) is smaller than the number of total elements of the feature space: $n(\mathfrak{X}_m) < n(\mathfrak{X})$. As points Breiman (2001a), the value of $n(\mathfrak{X}_m)$ is held constant during the entire procedure for growing the forest, and usually is set to $\sqrt{n(\mathfrak{X})}$. By randomly subsampling the original sample and the predictors, Random Forest improves the bagged tree method by decorrelating the trees (Hastie et al., 2009). Since it decorrelates the trees grown, it also decorrelate the errors made by each tree, yielding a more accurate prediction.

And why the decorrelation is important? James et al. (2013) create a scenario to make this characteristic clear. Let's follow their interesting argument. Imagine that we have a very strong predictor in our feature space, together with other moderately strong predictors. In the bagging procedure, the strong predictor will be in the top split of most of the trees, since it is the variable that better separates the $\mathcal{C}_k$ classes. By consequence, the bagged trees will be very similar to each other with the same variable in the top split, making the predictions highly correlated, and thus the errors also highly correlated. This will not lead to a decrease in the variance if compared to a single tree. The Random Forest procedure, on the other hand, forces each split to consider only a subset of the features, opening chances for the other features to do their job. The strong predictor will be left out of the bag in a number of situations, making the trees very different from each other. As a result, the resulting trees will present less variance in the classification error and in the OOB error, leading to a more reliable prediction. Random Forests have two main basic tuning parameters: *1)* the size of the subsample of features

ISNN 2182-7591

used in each split, $n(\mathfrak{X}_m)$, which is mandatory to be $n(\mathfrak{X}_m) < n(\mathfrak{X})$, being generally set as $\sqrt{n(\mathfrak{X})}$ and *2)* the size j of the set B, which is equal the number of trees to grow.

The last technique to be presented in the current paper is the boosting (Freund & Schapire, 1997). Boosting is a general adaptive method, and not a traditional ensemble technique, where each tree is constructed based on the previous tree in order to increase the prediction accuracy. The boosting method learns from the errors of previous trees, so unlikely bagging and Random Forest, it can lead to overfitting if the number of trees grown is too large. Boosting has three main basic tuning parameters: *1)* the size j of the set B, which is equal the number of trees to grow, *2)* the shrinkage parameter λ, which is the rate of learning from one tree to another, and *3)* the complexity of the tree, which is the number of possible terminal nodes $d = \alpha|T|$. James et al. (2013) point that λ is usually set to 0.01 or to 0.001, and that the smaller the value of λ, the highest needs to be the number of trees (B), in order to achieve good predictions.

The Machine Learning techniques presented in this paper can be helpful in discovering which psychological or educational test, or a combination of them, better predict academic achievement. The learning trees have also a number of advantages over the most traditional prediction models, since they doesn't make any assumptions regarding normality, linearity or independency of the variables, are non-parametric, handles different kind of predictors (nominal, ordinal, interval and ratio), are applicable to a wide range of problems, handles missing values and when combined with ensemble techniques provide the state-of-the-art results in terms of accuracy (Geurts et al., 2009).

The present paper introduced the basics ideas of the learning trees' techniques, in the first two sections above, and now they will be applied to predict the academic achievement of college students (high achievement vs. low achievement). Finally, the results of the four methods (single trees, bagging, Random Forest and boosting) will be compared with each other.

REVISTA E-PSI
REVISTA ELETRÓNICA DE PSICOLOGIA, EDUCAÇÃO E SAÚDE
ANO 4, VOLUME 1, 2014, pp.68-101.
https://www.revistaepsi.com
ISNN 2182-7591

Methods

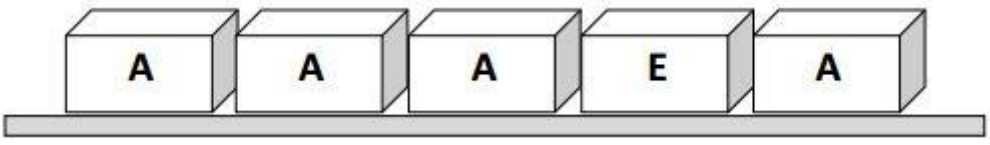
Participants

The sample is composed by 77 college students (55% woman) enrolled in the 2nd and 3rd year of a private Medical School from the state of Minas Gerais, Brasil. The sample was selected randomly, using the faculty's data set with the student's achievement recordings. From all the 2nd and 3rd year students we selected 50 random students with grades above 70% in the last semester, and 50 random students with grades equal to or below 70%. The random selection of students was made without replacement. The 100 random students selected to participate in the current study received a letter explaining the goals of the research, and informing the assessment schedule (days, time and faculty room). Those who agreed in being part of the study signed a inform consent, and confirmed they would be present in the schedule days to answer all the questionnaires and tests. From all the 100 students, only 77 appeared in the assessment days.

Instruments

The *Inductive Reasoning Developmental Test* (TDRI) was developed by Gomes and Golino (2009) and by Golino and Gomes (2012) to assess developmental stages of reasoning based on Common's Hierarchical Complexity Model (Commons & Richards, 1984; Commons, 2008; Commons & Pekker, 2008) and on Fischer's Dynamic Skill Theory (Fischer, 1980; Fischer & Yan, 2002). This is a pencil-and-paper test composed by 56 items, with a time limit of 100 minutes. Each item presents five letters or set of letters, being four with the same rule and one with a different rule. The task is to identify which letter or set of letters have the different rule.

Figure 1 – Example of TDRI's item 1 (from the first developmental stage assessed).



ISNN 2182-7591

Golino and Gomes (2012) evaluated the structural validity of the TDRI using responses from 1459 Brazilian people (52.5% women) aged between 5 to 86 years (*M*=15.75; *SD*=12.21). The results showed a good fit to the Rasch model (Infit: *M*=.96; *SD*=.17) with a high separation reliability for items (1.00) and a moderately high for people (.82). The item's difficulty distribution formed a seven cluster structure with gaps between them, presenting statistically significant differences in the 95% c.i. level (t-test). The CFA showed an adequate data fit for a model with seven first-order factors and one general factor [$\chi^2(61)$= 8832.594, *p*=.000; CFI=.96; RMSEA=.059]. The latent class analysis showed that the best model is the one with seven latent classes (AIC:263.380; BIC:303.887; Loglik:-111.690). The TDRI test has a self-appraisal scale attached to each one of the 56 items. In this scale, the participants are asked to appraise their achievement on the TDRI items, by reporting if he/she passed or failed the item. The scoring procedure of the TDRI self-appraisal scale works as follows. The participant receive a score of 1 in two situations: *1)* if the participant passed the *i*th item and reported that he/she passed the item, and *2)* if the participant failed the *i*th item and reported that he/she failed the item. On the other hand, the participant receives a score of 0 if his appraisal does not match his performance on the *i*th item: *1)* he/she passed the item, but reported that failed it, and *2)* he/she failed the item, but reported that passed it.

The *Metacognitive Control Test* (TCM) was developed by Golino and Gomes (2013) to assess the ability of people to control intuitive answers to logical-mathematical tasks. The test is based on Shane Frederick's Cognitive Reflection Test (Frederick, 2005), and is composed by 15 items. The structural validity of the test was assessed by Golino and Gomes (2013) using responses from 908 Brazilian people (54.8% women) aged between 9 to 86 years (*M*=27.70, *SD*=11.90). The results showed a good fit to the Rasch model (Infit: *M*=1.00; *SD*=.13) with a high separation reliability for items (.99) and a moderately high for people (.81). The TCM also has a self-appraisal scale attached to each one of its 15 items. The TCM self-appraisal scale is scored exactly as the TDRI self-appraisal scale: an incorrect appraisal receives a score of 0, and a correct appraisal receives a score of 1.

The *Brazilian Learning Approaches Scale* (EABAP) is a self-report questionnaire composed by 17 items, developed by Gomes and colleagues (Gomes, 2010; Gomes, Golino, Pinheiro, Miranda, & Soares, 2011). Nine items were elaborated to measure

ISNN 2182-7591

deep learning approaches, and eight items measure surface learning approaches. Each item has a statement that refers to a student's behavior while learning. The student considers how much of the behavior described is present in his life, using a Likert-like scale ranging from (1) not at all, to (5) entirely present. BLAS presents reliability, factorial structure validity, predictive validity and incremental validity as good marker of learning approaches. These psychometrical proprieties are described respectively in Gomes et al. (2011), Gomes (2010), and Gomes and Golino (2012). In the present study, the surface learning approach items scale were reverted in order to indicate the deep learning approach. So, the original scale from 1 (not at all) to 5 (entirely present), that related to surface learning behaviors, was turned into a 5 (not at all) to 1 (entirely present) scale of deep learning behaviors. By doing so, we were able to analyze all 17 items using the partial credit Rasch Model.

The *Cognitive Processing Battery* is a computerized battery developed by Demetriou, Mouyi and Spanoudis (2008) to investigate structural relations between different components of the cognitive processing system. The battery has six tests: *Processing Speed* (PS), *Discrimination* (DIS), *Perceptual Control* (PC), *Conceptual Control* (CC), *Short-Term Memory* (STM), and *Working Memory* (WM). Golino, Gomes and Demetriou (2012) translated and adapted the Cognitive Processing Battery to Brazilian Portuguese. They evaluated 392 Brazilian people (52.3% women) aged between 6 to 86 years (M= 17.03, SD= 15.25). The Cognitive Processing Battery tests presented a high reliability (Cronbach's Alpha), ranging from .91 for PC and .99 for the STM items. WM and STM items were analyzed using the dichotomous Rasch Model, and presented an adequate fit, each one showing an infit meansquare mean of .99 (WM's SD=.08; STM's SD=.10). In accordance with earlier studies, the structural equation modeling of the variables fitted a hierarchical, cascade organization of the constructs (CFI=.99; GFI=.97; RMSEA=.07), going from basic processing to complex processing: PS → DIS → PC → CC → STM → WM.

The *High School National Exam* (ENEM) is a 180 item educational examination created by Brazilian's Government to assess high school student's abilities on school subjects (see http://portal.inep.gov.br/). The ENEM result is now the main student's selection criteria to enter Brazilian Public universities. A 20 item version of the exam was created to assess the Medical School students' basic educational abilities.

REVISTA E-PSI
REVISTA ELETRÓNICA DE PSICOLOGIA, EDUCAÇÃO E SAÚDE
ANO 4, VOLUME 1, 2014, pp.68-101.
https://www.revistaepsi.com
ISNN 2182-7591

The student's ability estimates on the Inductive Reasoning Developmental Test (TDRI), on the Metacognitive Control Test (TCM), on the Brazilian Learning Approaches Scale (EABAP), and on the memory tests of the Cognitive Processing Battery, were computed using the original data set of each test, using the software Winsteps (Linacre, 2012). This procedure was followed in order to achieve reliable estimates, since only 77 medical students answered the tests. The mixture of the original data set with the Medical School students' answers didn't change the reliability or fit to the models used. A summary of the separation reliability and fit of the items, the separation reliability of the sample, the statistical model used, and the number of medical students that answered each test is provided in Table 2.

Table 2 – Fit, reliability, model used and sample size per test used.

Test	Item		Person	Infit: *M* (*SD*)	Model	Medical Students' *N* (%)
	Reliability	Infit: *M* (*SD*)	Reliability			
Inductive Reasoning Developmental Test (TDRI)	1.00	.96 (.17)	.82	1.00 (.97)	Dichotomous Rasch Model	59 (76.62)
TDRI's Self-Appraisal Scale	.83	1.01 (.16)	.62	.97 (.39)	Dichotomous Rasch Model	59 (76.62)
Metacognitive Control Test (MCT)	.99	1.00 (.13)	.81	.95 (.42)	Dichotomous Rasch Model	53 (68.83)
MCT's Self-Appraisal Scale	.96	1.00 (.16)	.72	.99 (.24)	Dichotomous Rasch Model	53 (68.83)
Brazilian Learning Approaches Scale (EABAP)	.99	1.01 (.11)	.80	1.03 (.58)	Partial Credit Rasch Model	59 (76.62)
ENEM	.90	.93 (.29)	.77	.96 (.33)	Dichotomous Rasch Model	40 (51.94)
Processing Speed	α=.96	NA	NA	NA	NA	46 (59.74)
Discrimination	α=.98	NA	NA	NA	NA	46 (59.74)
Perceptual Control	α=.91	NA	NA	NA	NA	46 (59.74)
Conceptual Control	α=.96	NA	NA	NA	NA	46 (59.74)
Short Term Memory	.99	.99 (.10)	.79	.98 (.25)	Dichotomous Rasch Model	46 (59.74)
Working Memory	.98	.99 (.07)	.81	.99 (.16)	Dichotomous Rasch Model	46 (59.74)

REVISTA E-PSI
REVISTA ELETRÓNICA DE PSICOLOGIA, EDUCAÇÃO E SAÚDE
ANO 4, VOLUME 1, 2014, pp.68-101.
https://www.revistaepsi.com
ISNN 2182-7591

Procedures

After estimating the student's ability in each test or extracting the mean response time (in the computerized tests: PS, DIS, PC and CC) the Shapiro-Wilk test of normality was conducted in order to discover which variables presented a normal distribution. Then, the correlations between the variables were computed using the heterogeneous correlation function (hector) of the *polycor* package (Fox, 2010) of the R statistical software. To verify if there was any statistically significant difference between the students' groups (high achievement vs. low achievement) the two-sample T test was conducted in the normally distributed variables and the Wilcoxon Sum-Rank test in the non-normal variables, both at the 0.05 significance level. In order to estimate the effect sizes of the differences the R's *compute.es* package (Del Re, 2013) was used. This package computes the effect sizes, along with their variances, confidence intervals, p-values and the *common language effect size* (CLES) indicator using the p-values of the significance testing. The CLES indicator expresses how much (in %) the score from one population is greater than the score of the other population if both are randomly selected (Del Re, 2013).

The sample was randomly split in two sets, training and testing. The training set is used to grow the trees, to verify the quality of the prediction in an exploratory fashion, and to adjust the tuning parameters. Each model created using the training set is applied in the testing set to verify how it performs on a new data set.

The single learning tree technique was applied in the training set having all the tests plus sex as predictors, using the package *tree* (Ripley, 2013) of the R software. The quality of the predictions made in the training set was verified using the misclassification error rate, the residual mean deviance and the Pseudo R^2. The prediction made in the cross-validation using the test set was assessed using the total accuracy, the sensitivity and the specificity. Total accuracy is the proportion of observations correctly classified:

$$Acc = \frac{1}{n|T_E|}\sum_{x \in T_E} I(y_i = C_k)$$

ISNN 2182-7591

where $n|T_E|$ is the number of observations in the testing set. The sensitivity is the rate of observations correctly classified in a target class, e.g. $\mathcal{C}_1 = low\ achievement$, over the number of observations that belong to that class:

$$Sens = \frac{\sum_{x \in T_E} I(y_i = \mathcal{C}_1)}{\sum_{x \in T_E} I(\mathcal{C}_1)}$$

Finally, specificity is the rate of correctly classified observations of the non-target class, e.g. $\mathcal{C}_2 = high\ achievement$, over the number of observations that belong to that class:

$$Spec = \frac{\sum_{x \in T_E} I(y_i = \mathcal{C}_2)}{\sum_{x \in T_E} I(\mathcal{C}_2)}$$

The bagging and the Random Forest technique were applied using the *randomForest* package (Liaw & Wiener, 2012). As the bagging technique is the aggregation trees using *n* random subsamples, the *randomForest* package can be used to create the bagging classification by setting the number of features (or predictors) equal the size of the feature set: $n(\mathfrak{X}_m) = n(\mathfrak{X})$. In order to verify the quality of the prediction both in the training (modeling phase) and in the testing set (cross-validation phase), the total accuracy, the sensitivity and specificity were used. Since the bagging and the random forest are black box techniques – i.e. there is only a prediction based on majority vote and no "typical tree" to look at the partitions – to determine which variable is important in the prediction two importance measures will be used: the mean decrease of accuracy and the mean decrease of the Gini index. The former indicates how much in average the accuracy decreases on the out-of-bag samples when a given variable is excluded from the model (James et al., 2013). The latter indicates «the total decrease in node impurity that results from splits over that variable, averaged over all trees» (James et al., 2013, p.335). The Gini Index can be calculated using the formula below:

REVISTA E-PSI
REVISTA ELETRÓNICA DE PSICOLOGIA, EDUCAÇÃO E SAÚDE
ANO 4, VOLUME 1, 2014, pp.68-101.
https://www.revistaepsi.com
ISNN 2182-7591

$$Gini = \sum_{k=1}^{K} \hat{p}_{mc_k}(1 - \hat{p}_{mc_k}).$$

Finally, in order to verify which model presented the best predictive performance (accuracy, sensitivity and specificity) the Marascuilo (1966) procedure was used. This procedure points if the difference between all pairs of proportions is statistically significant. Two kinds of comparisons were made: difference between sample sets and differences between models. In the Marascuilo procedure, a test value and a critical range is computed to all pairwise comparisons. If the test value exceeds the critical range the difference between the proportions is considered significant at .05 level. A more deep explanation of the procedure can be found at the NIST/Semantech website [http://www.itl.nist.gov/div898/handbook/prc/section4/prc474.htm]. The complete dataset used in the current study (Golino & Gomes, 2014) can be downloaded for free at http://dx.doi.org/10.6084/m9.figshare.973012.

Results

The only predictors that showed a normal distribution were the EABAP (W=.97, *p*=.47), the ENEM exam (W=.97, *p*=.47), processing speed (W=.95, *p*=.06) and perceptual control (W=.95, *p*=.10). All other variables presented a p-value smaller than .05. In terms of the difference between the high and the low achievement groups there was a statistically significant difference at the 95% level in the mean ENEM Rasch score ($\bar{x}_{High}$=1.13, σ^2=1.24, $\bar{x}_{Low}$=-1.08, σ^2_{Low}=2.68, *t*(39)=4.8162, *p*=.000), in the median Rasch score of the TDRI ($\tilde{x}_{High}$=1.45, σ^2= 2.23, $\tilde{x}_{Low}$ = .59, σ^2_{Low}=1.58, W=609, *p*=.008), in the median Rasch score of the TCM ($\tilde{x}_{High}$=1.03, σ^2=2.96, $\tilde{x}_{Low}$=-2.22, σ^2_{Low}=8.61, W=526, *p*=.001), in the median Rasch score of the TDRI's self-appraisal scale ($\tilde{x}_{High}$=2.00, σ^2=2.67, $\tilde{x}_{Low}$=1.35, σ^2_{Low}=1.63, W=646, *p*=.001), in the median Rasch score of the TCM's self-appraisal scale ($\tilde{x}_{High}$=1.90, σ^2=3.25, $\tilde{x}_{Low}$=-1.46, σ^2_{Low}=5.20, W=474, *p*=.000), and in the median discrimination time ($\tilde{x}_{High}$=440, σ^2=10.355, $\tilde{x}_{Low}$= 495, σ^2_{Low}=7208, W=133, *p*=.009).

ISNN 2182-7591

The effect sizes, its 95% confidence intervals, variance, significance and common language effect sizes are described in Table 3.

Table 3 – Effect Sizes, Confidence Intervals, Variance, Significance and Common Language Effect Sizes (CLES).

Test	Effect Size of the difference (d)	95% C.I. (d)	σ^2 (d)	*p* (d)	CLES
ENEM	1.46	0.73, 2.19	.13	.00	84.88%
Inductive Reasoning Developmental Test (TDRI)	0.64	0.11, 1.18	.07	.02	67.54%
Metacognitive Control Test (TCM)	0.87	0.29, 1.45	.08	.00	73.01%
TDRI' Self-Appraisal Scale	0.81	0.27, 1.36	.07	.00	71.73%
TCM' Self-Appraisal Scale	1.15	0.52, 1.78	.10	.00	79.21%
Discrimination	0.75	0.11, 1.38	.10	.02	70.19%

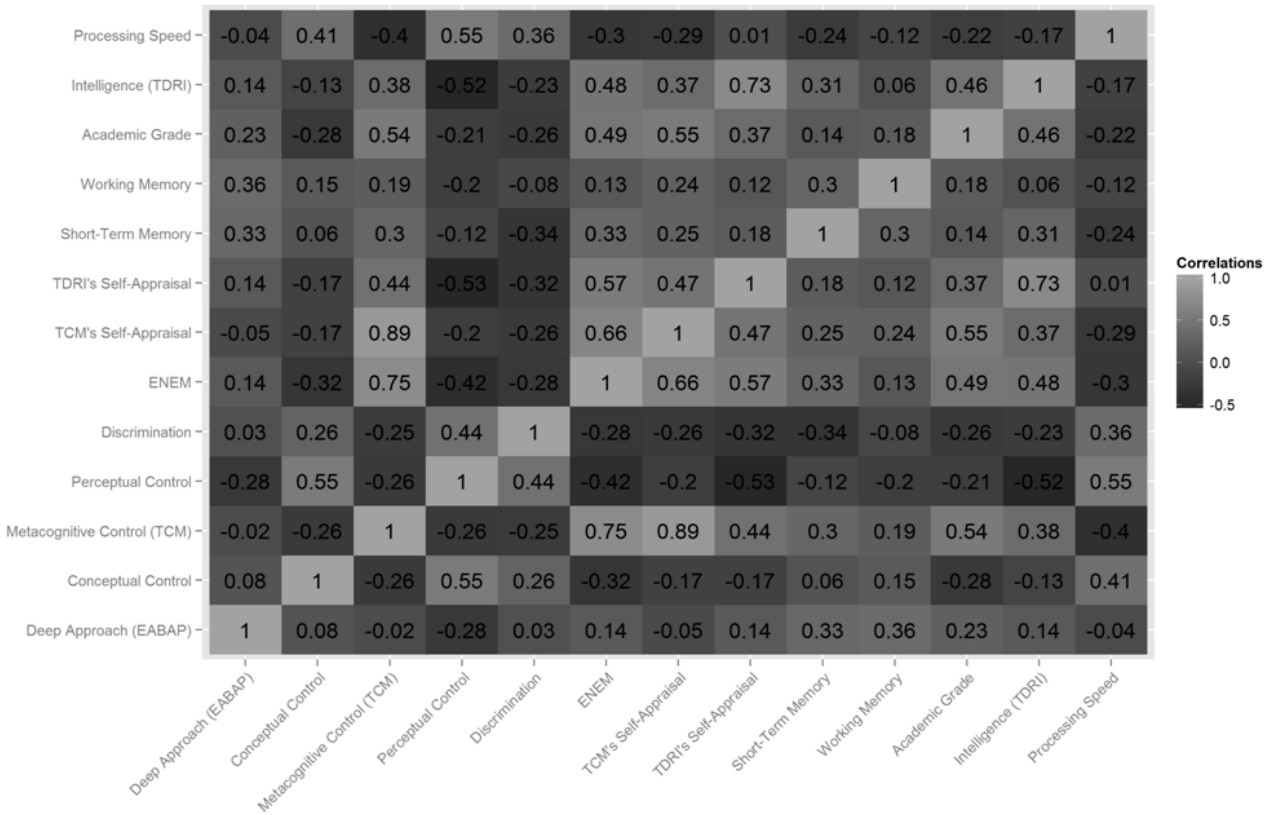
Considering the correlation matrix presented in Figure 2, the only variables with moderate correlations (greater than .30) with academic grade was the TCM (.54), the TDRI (.46), the ENEM exam (.49), the TCM Self-Appraisal Scale (.55) and the TDRI Self-Appraisal Scale (.37). The other variables presented only small correlations with the academic grade. So, considering the analysis of differences between groups, the size of the effects and the correlation pattern, it is possible to elect some variables as favorites for being predictive of the academic achievement. However, as the learning tree analysis showed, the picture is a little bit different than showed in Table 2 and Figure 2.

In spite of inputting all the tests plus sex as predictors in the single tree analysis, the *tree* package algorithm selected only three of them to construct the tree: the TCM, the EABAP (in the Figure 3, represented as DeepAp) and the TDRI' Self-Appraisal Scale (in the Figure 3, represented as SA_TDRI). These three predictors provided the best split possible in terms of misclassification error rate (.27), residual mean deviance (.50) and Pseudo-R^2 (.67) in the training set. The tree constructed has four terminal

REVISTA E-PSI
REVISTA ELETRÓNICA DE PSICOLOGIA, EDUCAÇÃO E SAÚDE
ANO 4, VOLUME 1, 2014, pp.68-101.
https://www.revistaepsi.com
ISNN 2182-7591

nodes (Figure 3). The TCM is the top split of the tree, being the most important predictor, i.e. the one who best separates the observations into two nodes. People with TCM' Rasch score lower than -1.29 are classified as being part of the low achievement class, with a probability of 52.50%.

Figure 2 – The Correlation Matrix.

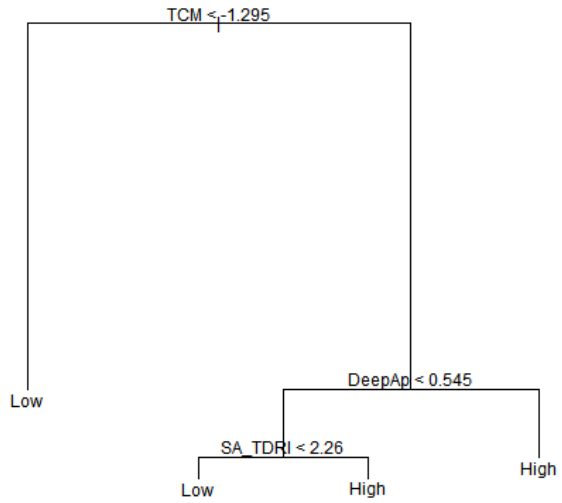


By its turn, people with TCM' Rasch score greater than -1.29 and with EABAP's Rasch score (DeepAp) greater than 0.54 are classified as being part of the high achievement class, with a probability of 60%. People are also classified as belonging to the high achievement class if they present a TCM' Rasch score greater than -1.29, an EABAP's Rasch Score (DeepAp) greater than 0.54, but a TDRI's Self-Appraisal Rasch Score greater than 2.26, with a probability of 80%. On the other hand, people are classified as belonging to the low achievement class with 60% probability if they have

REVISTA E-PSI
REVISTA ELETRÓNICA DE PSICOLOGIA, EDUCAÇÃO E SAÚDE
ANO 4, VOLUME 1, 2014, pp.68-101.
https://www.revistaepsi.com
ISNN 2182-7591

the same profile as the previous one but the TDRI's Self-Appraisal Rasch score being less than 2.26. The total accuracy of this tree is 72.50%, with a sensitivity of 57.89% and a specificity of 85.71%. The tree was applied in the testing set for cross-validation, and presented a total accuracy of 64.86%, a sensitivity of 43.75% and a specificity of 80.95%. There was a difference of 7.64% in the total accuracy, of 14.14% in the sensitivity and of 4.76% in the specificity from the training set to the test set.

Figure 3 – Single tree grown using the tree package.



The result of the bagging model with one thousand bootstrapped samples showed an out-of-bag error rate of .37, a total accuracy of 65%, a sensitivity of 63.16% and a specificity of 66.67%. Analyzing the mean decrease in the Gini index, the three most important variables for node purity were, in decreasing order of importance: Deep Approach (EABAP), TCM, and TDRI Self-Appraisal (Figure 4). The higher the decrease in the Gini index, the higher the node purity when the variable is used.

Figure 5 shows the high achievement prediction error (green line), out-of-bag error (red line) and low achievement prediction error (black line) per tree. The errors became more stable with more than 400 trees.

REVISTA E-PSI
REVISTA ELETRÓNICA DE PSICOLOGIA, EDUCAÇÃO E SAÚDE
ANO 4, VOLUME 1, 2014, pp.68-101.
https://www.revistaepsi.com
ISNN 2182-7591

Figure 4 – Mean decrease of the Gini index in the Bagging Model.

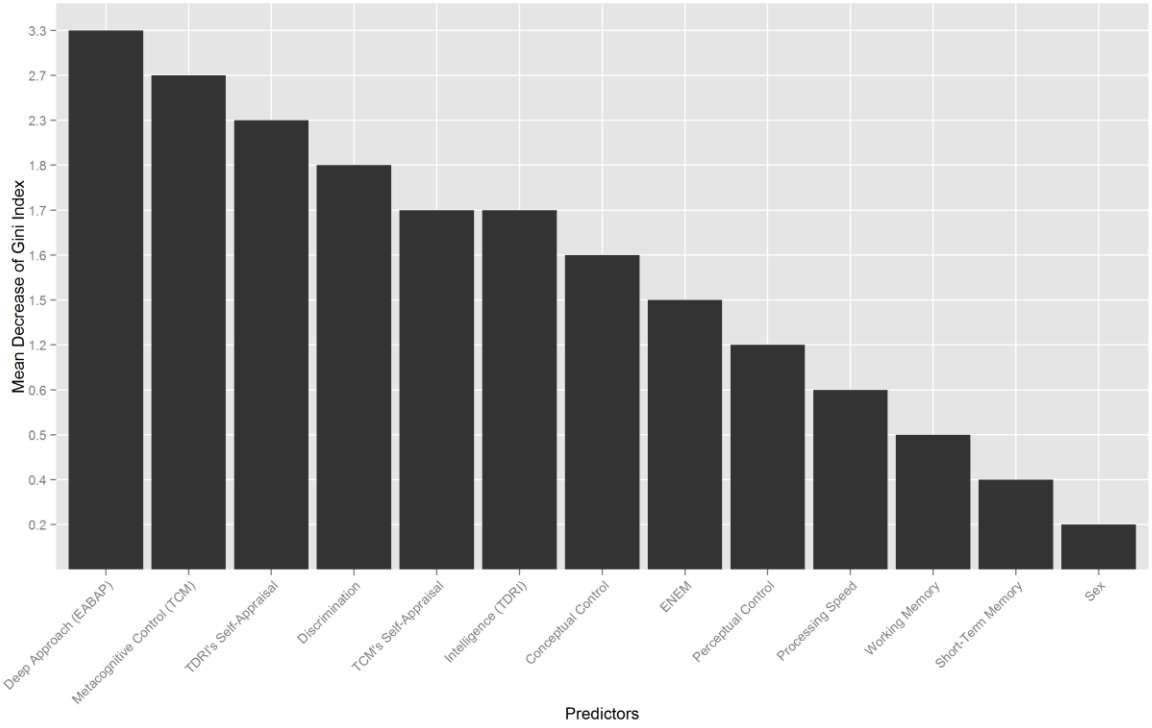
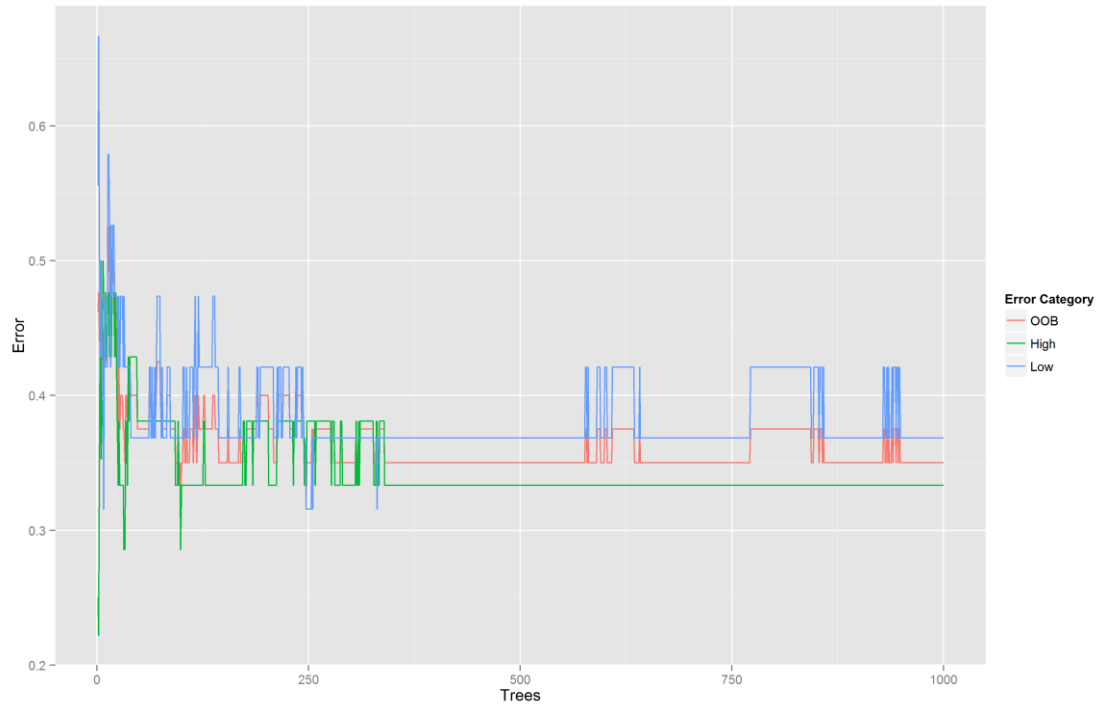


Figure 5 – Bagging's out-of-bag error (red), high achievement prediction error (green) and low achievement prediction error (blue).

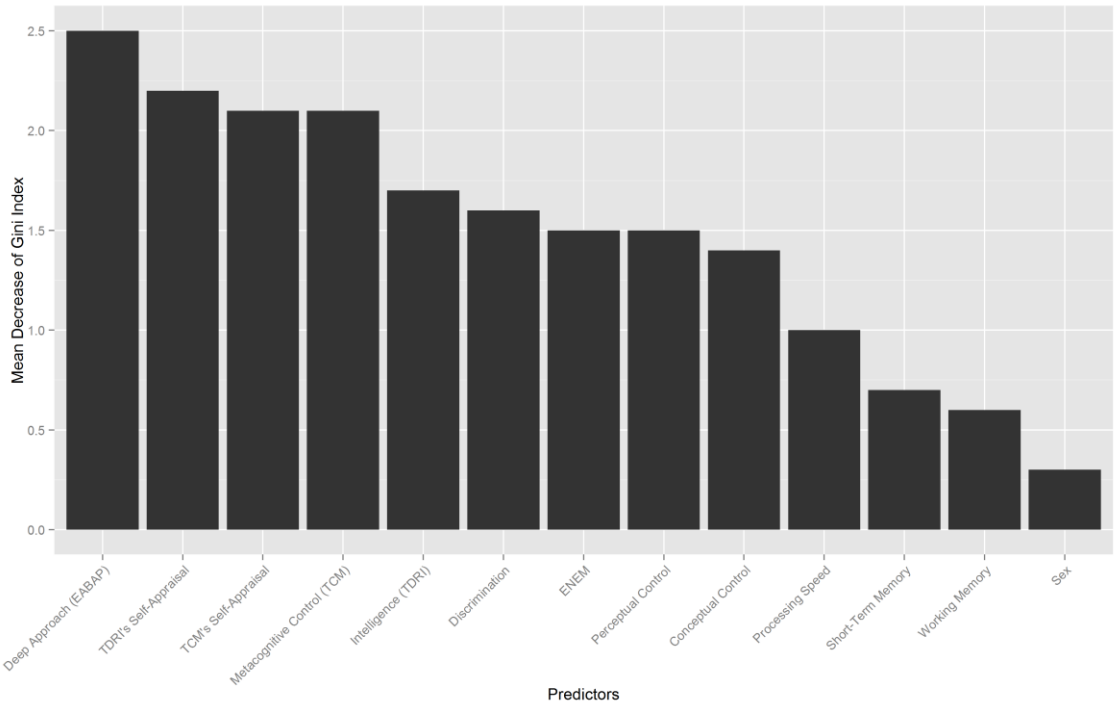


REVISTA E-PSI
REVISTA ELETRÓNICA DE PSICOLOGIA, EDUCAÇÃO E SAÚDE
ANO 4, VOLUME 1, 2014, pp.68-101.
https://www.revistaepsi.com
ISNN 2182-7591

The bagging model was applied in the testing set for cross-validation, and presented a total accuracy of 67.56%, a sensitivity of 68.75% and a specificity of 66.67%. There was a difference of 2.56% in the total accuracy and of 5.59% in the sensitivity. No difference in the specificity from the training set to the test set was found.

The result of the Random Forest model with one thousand trees showed an out-of-bag error rate of .32, a total accuracy of 67.50%, a sensitivity of 63.16% and a specificity of 71.43%. The mean decrease in the Gini index showed a similar result of the bagging model. The four most important variables for node purity were, in decreasing order of importance: Deep Approach (EABAP), TDRI Self-Appraisal, TCM Self-Appraisal and TCM (Figure 6).

Figure 6 – Mean decrease of the Gini index in the Random Forest Model.

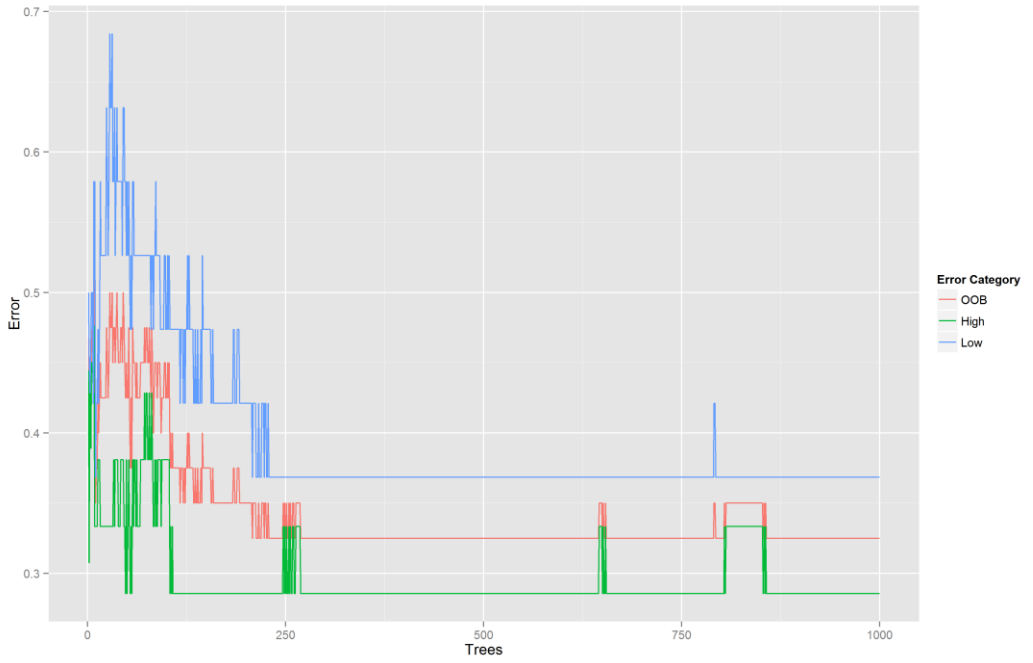


The Random Forest model was applied in the testing set for cross-validation, and presented a total accuracy of 72.97%, a sensitivity of 56.25% and a specificity of 81.71%. There was a difference of 5.47% in the total accuracy, of 6.91% in the sensitivity, and of 10.28% in the specificity.

REVISTA E-PSI
REVISTA ELETRÓNICA DE PSICOLOGIA, EDUCAÇÃO E SAÚDE
ANO 4, VOLUME 1, 2014, pp.68-101.
https://www.revistaepsi.com
ISNN 2182-7591

Figure 7 shows the high achievement prediction error (green line), out-of-bag error (red line) and low achievement prediction error (black line) per tree. The errors became more stable with approximately more than 250 trees.

Figure 7 – Random Forest's out-of-bag error (red), high achievement prediction error (green) and low achievement prediction error (blue).



The result of the boosting model with ten trees, shrinkage parameter of 0.001, tree complexity of two, and setting the minimum number of split to one, resulted in a total accuracy of 92.50%, a sensitivity of 90% and a specificity of 95%. Analyzing the mean decrease in the Gini index, the three most important variables for node purity were, in decreasing order of importance: Deep Approach (EABAP), TCM and TCM Self-Appraisal (Figure 8).

The boosting model was applied in the testing set for cross-validation, and presented a total accuracy of 69.44%, a sensitivity of 62.50% and a specificity of 75%. There was a difference of 22.06% in the total accuracy, of 27.50% in the sensitivity, and of 20% in the specificity. Figure 9 shows the variability of the error by iterations in the training and testing set.

REVISTA E-PSI
REVISTA ELETRÓNICA DE PSICOLOGIA, EDUCAÇÃO E SAÚDE
ANO 4, VOLUME 1, 2014, pp.68-101.
https://www.revistaepsi.com
ISNN 2182-7591

Figure 8 – Mean decrease of the Gini index in the Boosting Model.

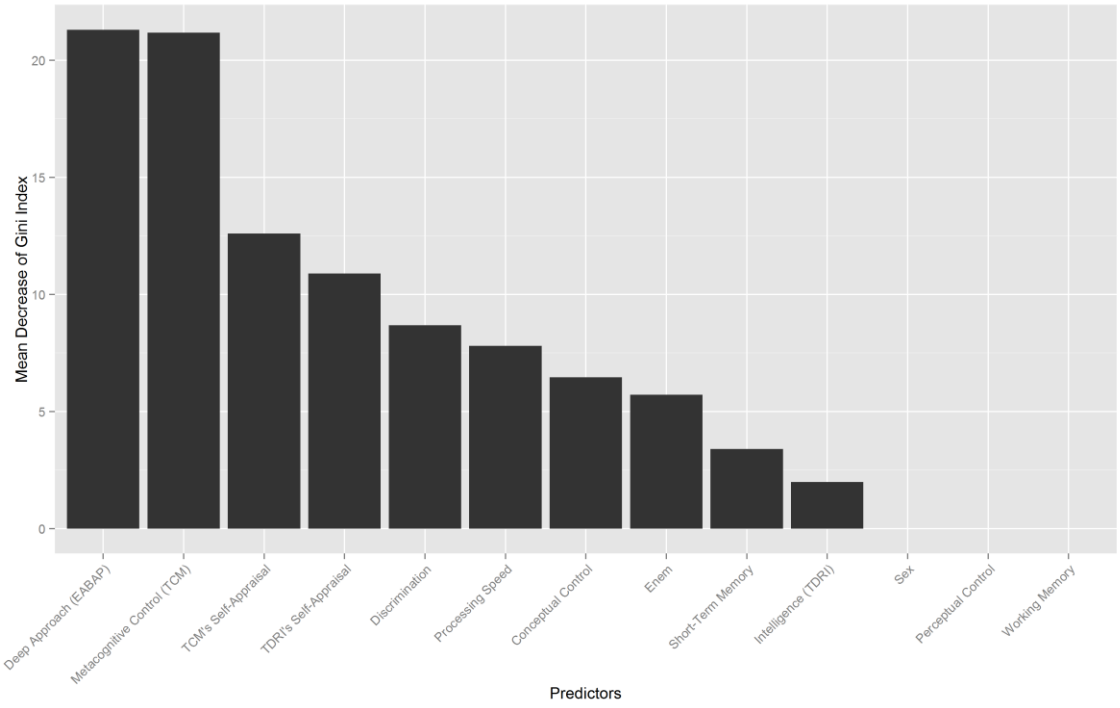
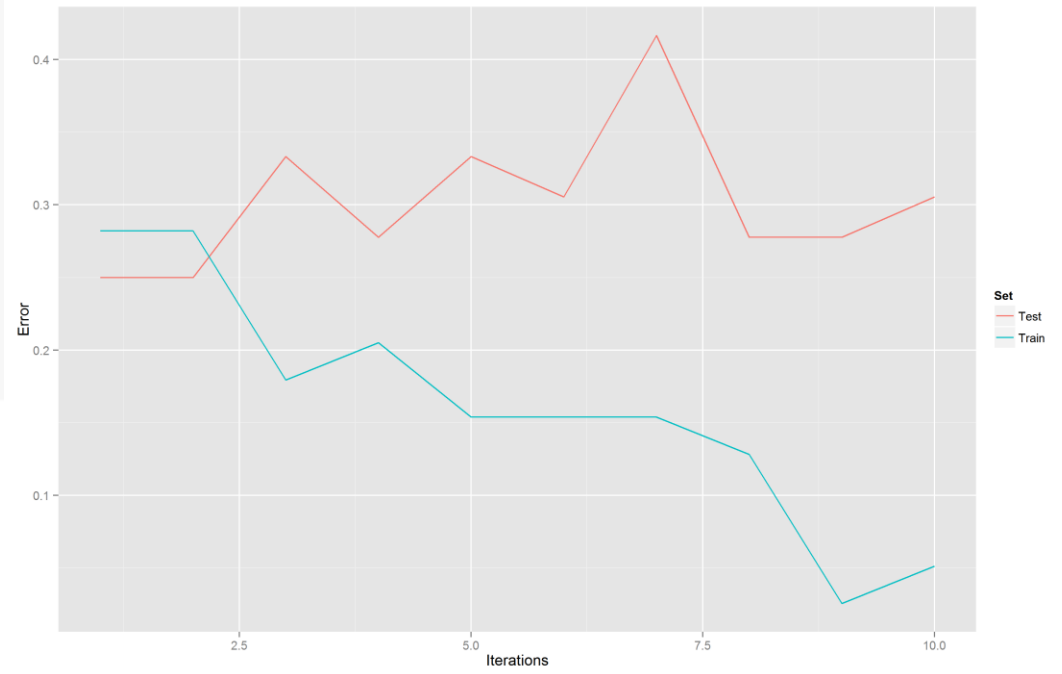


Figure 9 – Boosting's prediction error by iterations in the training and in the testing set.



REVISTA E-PSI
REVISTA ELETRÓNICA DE PSICOLOGIA, EDUCAÇÃO E SAÚDE
ANO 4, VOLUME 1, 2014, pp.68-101.
https://www.revistaepsi.com
ISNN 2182-7591

Table 4 synthesizes the results of the learning tree, bagging, random forest and boosting models. The boosting model was the most accurate, sensitive and specific in the prediction of the academic achievement class (high or low) in the training set (see Table 4 and Table 5). Furthermore, there is enough data to conclude a significant difference between the boosting model and the other three models, in terms of accuracy, sensitivity and specificity (see Table 5). However, it was also the one with the greater difference in the prediction between the training and the testing set. This difference was also statistically significant in the comparison with the other models (see Table 5).

Table 4 – Predictive Performance by Machine Learning Model.

Model	Training Set			Testing Set			Difference between the training set and testing set		
	Total Accuracy	Sensitivity	Specificity	Total Accuracy	Sensitivity	Specificity	Total Accuracy	Sensitivity	Specificity
Learning Trees	.725	.579	.857	.649	.438	.810	.076	.141	.048
Bagging	.650	.632	.667	.676	.688	.667	-.026	-.056	.000
Random Forest	.675	.632	.714	.730	.563	.817	-.055	.069	-.103
Boosting	.925	.900	.950	.694	.625	.750	.231	.275	.200

Both bagging and Random Forest presented the lowest difference in the predictive performance between the training and the testing set. Comparing the both models, there is not enough data to conclude that their total accuracy, their sensitivity and specificity are significantly different (see Table 5). In sum, both bagging and Random Forest were the more stable techniques to predict the academic achievement class.

ISSN 2182-7591

REVISTA E-PSI
REVISTA ELETRÓNICA DE PSICOLOGIA, EDUCAÇÃO E SAÚDE
ANO 4, VOLUME 1, 2014, pp.68-101

https://www.revistaepsi.com

Table 5 – Result of the Marascuilo's Procedure.

Pairwise Comparisons	Comparison between sample sets									Comparison between models (prediction in the training set)								
	Total Accuracy ($Acc_{T_R} - Acc_{T_E}$)			Sensitivity ($Sens_{T_R} - Sens_{T_E}$)			Specificity ($Spec_{T_R} - Spec_{T_E}$)			Total Accuracy			Sensitivity			Specificity		
	Value	Critical Range	Difference Significant?	Value	Critical Range	Difference Significant?	Value	Critical Range	Difference Significant?	Value	Critical Range	Difference Significant?	Value	Critical Range	Difference Significant?	Value	Critical Range	Difference Significant?
Learning Tree – Bagging	.051	.055	No	.086	.074	Yes	.048	.038	Yes	.075	.116	No	.053	.123	No	.19	.104	Yes
Learning Tree – Random Forest	.022	.062	No	.072	.077	No	.055	.066	No	.05	.115	No	.053	.123	No	.143	.102	Yes
Learning Tree – Boosting	.154	.089	Yes	.134	.101	Yes	.152	.081	Yes	.2	.092	Yes	.321	.103	Yes	.093	.073	Yes
Bagging – Random Forest	.029	.049	No	.013	.061	No	.103	.054	Yes	.025	.119	No	0	.121	No	.048	.116	No
Bagging – Boosting	.205	.080	Yes	.219	.089	Yes	.200	.071	Yes	.275	.097	Yes	.268	.101	Yes	.283	.092	Yes
Random Forest – Boosting	.176	.085	Yes	.206	.091	Yes	.097	.089	Yes	.25	.096	Yes	.268	.101	Yes	.236	.089	Yes

ISNN 2182-7591

Discussion

The studies exploring the role of psychological and educational constructs in the prediction of academic performance can help to understand how the human being learns, can lead to improvements in the curriculum designs, and can be very helpful to identify students at risk of low academic achievement (Musso & Cascallar, 2009; Musso et al., 2013). As pointed before, the traditional techniques used to verify the relationship between academic achievement and its psychological and educational predictors suffers from a number of assumptions and from not providing high accurate predictions. The field of Machine Learning, on the other hand, provides several techniques that lead to high accuracy in the prediction of educational and academic outcomes. Musso et al. (2013) showed the use of a Machine Learning model in the prediction of academic achievement with accuracies above 90% in average. The model they adopted, named artificial neural networks, in spite of providing very high accuracies are not easily translated into a comprehensive set of predictive rules. The relevance of translating a complex predictive model into a comprehensive set of relational rules is that professionals can be trained to make the prediction themselves, given the result of psychological and educational tests. Moreover, a set of predictive rules involving psycho-educational constructs may help in the construction of theories regarding the relation between these constructs in the learning or academic outcome, filling the gap pointed by Edelsbrunner and Schneider (2013).

In the present paper we introduced the basics of single learning trees, bagging, Random Forest and Boosting in the context of academic achievement prediction (high achievement *vs* low achievement). These techniques can be used to achieve higher accuracy rates than the traditional statistical methods, and its result are easily understood by professionals, since a classification tree is a roadmap of rules for predicting a categorical outcome.

In order to predict the academic achievement level of 59 Medical students, thirteen variables were used, involving sex and measures of intelligence, metacognition, learning approaches, basic high school knowledge and basic cognitive processing indicators. About 46% of the predictors were statistically significant to differentiate the low and the high achievement group, presented a moderately high (above .70) effect

ISNN 2182-7591

size: ENEM; the Inductive Reasoning Developmental Test; the Metacognitive Control Test; the TDRI's Self-Appraisal Scale; the TCM's Self-Appraisal Scale and the Discrimination indicator. In exception of the perceptual discrimination indicator, all the variables pointed before presented correlation coefficients greater than .30. However the two predictors with the highest correlation with academic achievement presented only moderate values (TCM=.54; TCM's Self-Appraisal Scale=.55).

The single learning tree model showed that the Metacognitive Control Test was the best predictor of the academic achievement class, and together with the Brazilian Learning Approaches Scale and the TDRI's Self-Appraisal scale, explained 67% of the outcome's variance. The total accuracy in the training set was 72.5%, with a sensitivity of 57.9% and a specificity of 85.7%. However, when the single tree model was applied in the testing set, the total accuracy decreased 7.6%, while the sensitivity dropped 14.1% and the specificity 4.8%. This result suggests an overfitting of the single tree model. Interestingly, one of the variables that contributed in the prediction of the academic achievement in the single tree model (learning approach) was not statistically significant to differentiate the high and the low achievement group. Furthermore, the Brazilian Learning Approaches Scale presented a correlation of only .23 with academic achievement. Even tough, the learning approach together with metacognition (TCM and TDRI's Self-Appraisal Scale) explained 67% of the academic achievement variance. The size of a correlation and the non-significance in differences between groups are not indicators of a bad prediction from one variable over another.

The bagging model, by its turn, presented a lower total accuracy, sensitivity and specificity in the training phase if compared to the single tree model. However this difference was only significant in the specificity (a difference of .048). Comparing the prediction made in the two sample sets, the bagging model outperformed the single tree model, since it resulted in more stable predictions (see Table 3 and Table 4). The out-of-bag error was .35, and the mean difference from the training set performance (accuracy, sensitivity and specificity) to the test set performance was only -.027. The total accuracy of the bagging model was 65% in the training set and 67.6% in the testing set, while the sensitivity and specificity was 63.2% and 66.7% in the former, and 68.8% and 66.7% in the latter. The classification of the bagging model became more pure when the Brazilian Learning Approaches Scale, the Metacognitive Control Test or the TDRI's Self-

ISNN 2182-7591

Appraisal Scale was used in the split, as pointed by the decrease in the Gini index of Figure 4. The three more important variables in the prediction of the academic achievement pointed by the bagging model matched the variables selected by the single tree algorithm.

The Random Forest model showed a small decrease in the out-of-bag error if compared to the bagging model, but the overall performance of the two models was basically the same, with no statistically significant difference in the training set prediction. If compared with bagging, the Random Forest deviation from the performance in the training set in relation to the testing set was only significantly different in the sensitivity. The mean difference of the Random Forest model from the training set performance to the test set performance was 2.9%. Only the sensitivity in the training set phase was significantly lower in the Random Forest in the comparison with the single learning trees. However, the Random Forest model was also more stable in the prediction performance than the single learning tree model. The classification of the Random Forest model became more pure when the Brazilian Learning Approaches Scale, the TDRI's Self-Appraisal Scale, the TCM's Self-Appraisal Scale or the the Metacognitive Control Test was used in the split, as pointed by the decrease in the Gini index. The variable importance measure of the Random Forest basically matched the result of the bagging and of the single tree algorithm.

Finally, the boosting model was the one presenting the higher accuracy, sensitivity and specificity, being statistically different from all other models. This model achieved a total accuracy of 92.50% in the training set, with sensitivity of 90% and specificity of 95%. However, it was the model with the greater difference in the prediction performance from the training set to the testing set. So, we can argue that in spite of the great performance in the training set, this was due to over fit, since in the testing set the accuracy dropped 23%.

In sum, both the bagging and the Random Forest model were the better models to predict high and low academic achievement of college students, since they presented the most stable predictions between the training and testing sample sets. Moreover, these models presented an overall accuracy close to 70%. Three variables were consistently pointed as important in the prediction (the Metacognitive Control Test, the Brazilian Learning Approach Scale and the TDRI's Self-Appraisal Scale). This result goes in the

ISNN 2182-7591

same direction of other studies showing the relevance of metacognition (Musso, Kyndt, Cascallar, & Dochy, 2012) and learning approaches (Norton & Crowley, 1995; Kyndt, 2011) in the explication of academic achievement in higher education.

Acknowledgments

The authors thank the Foundation for Research Support of the State of Minas Gerais (FAPEMIG) for financing the research, and the Faculdade Independente do Nordeste for financial support.

ISNN 2182-7591

References

Breiman, L. (2001a). Random forests. *Machine Learning, 1*(45), 5-32. Doi10.1023/A:1010933404324.

Breiman, L. (2001b). Bagging predictors. *Machine Learning*, *24*(2), 23-140.

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

Commons, M.L., & Richards, F.A. (1984). Applying the general stage model. In M.L. Commons, F.A. Richards, & C. Armon (Eds.), *Beyond formal operations. Late adolescent and adult cognitive development: Late adolescent and adult cognitive development* (Vol.1, pp.141-157). New York: Praeger.

Commons, M.L. (2008). Introduction to the model of hierarchical complexity and its relationship to postformal action. *World Futures, 64*, 305-320.

Commons, M.L., & Pekker, A. (2008). Presenting the formal theory of hierarchical complexity. *World Futures*, *64*, 375-382.

Del Re, A.C. (2013). Compute.es: Compute effect sizes (R package version 0.2-2.) [computer software manual]. Retrieved from: http://cran.r-project.org/web/packages/compute.es.

Demetriou, A., Mouyi, A., & Spanoudis, G. (2008). Modeling the structure and development of g. *Intelligence*, *5*, 437-454.

Edelsbrunner, P., & Schneider, M. (2013). Modelling for prediction vs. modelling for understanding: Commentary on Musso et al. (2013). *Frontline Learning Research*, *2*, 99-101.

Fischer, K.W. (1980). A theory of cognitive development: The control and construction of hierarchies of skills. *Psychological Review*, *87*, 477-531.

Fischer, K.W., & Yan, Z. (2002). The development of dynamic skill theory. In R. Lickliter, & D. Lewkowicz (Eds.), *Conceptions of development: Lessons from the laboratory*. Hove, UK: Psychology Press.

Flach, P. (2012). *Machine Learning: The art and science of algorithms that make sense of data*. Cambridge: Cambridge University Press.

Fox, J. (2010). Polycor: Polychoric and polyserial correlations (R package version 0.7-8.) [computer software manual]. Retrieved from: http://cran.r-project.org/web/packages/polycor.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 24-42.

Freund, Y., & Schapire, R.E. (1997). A decision-theoretic generalization of on-line learning and an application to Boosting. *Journal of Computer and System Sciences*, *55*(1), 119-139.

ISNN 2182-7591

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.

Golino, H.F., & Gomes, C.M.A. (2014): Dataset from medical students: E-psi study, http://dx.doi.org/10.6084/m9.figshare.973012.

Golino, H.F., & Gomes, C. M.A. (2012, July). The structural validity of the Inductive Reasoning Developmental Test for the measurement of developmental stages. In K. Stålne (Chair), *Adult development: Past, present and new agendas of research.* Symposium conducted at the Meeting of the European Society for Research on Adult Development, Coimbra, Portugal.

Golino, H.F., & Gomes, C.M.A. (2013, October). Controlando pensamentos intuitivos: O que o pão de queijo e o café podem dizer sobre a forma como pensamos. In C.M.A. Gomes (Chair), *Neuroeconomia e neuromarketing*. Symposium conducted at the VII Simpósio de Neurociências da Universidade Federal de Minas Gerais, Belo Horizonte, Brasil.

Golino, H.F., Gomes, C.M.A., & Demetriou, A. (2012, July). The development of hierarchical processes: Processing efficiency and memory from children to older adults. In K. Stålne (Chair), *Adult development: Past, present and new agendas of research.* Symposium conducted at the Meeting of the European Society for Research on Adult Development, Coimbra, Portugal.

Gomes, C.M.A., & Golino, H.F. (2009). Estudo exploratório sobre o Teste de Desenvolvimento do Raciocinio Indutivo (TDRI). In D. Colinvaux (Ed.), *Anais do VII Congresso Brasileiro de Psicologia do Desenvolvimento: Desenvolvimento e Direitos Humananos* (pp.77-79). Rio de Janeiro: UERJ. Retrieved from: http://www.abpd.psc.br/files/congressosAnteriores/AnaisVIICBPD.pdf.

Gomes, C.M.A. (2010). Perfis de estudantes e a relação entre abordagens de aprendizagem e rendimento escolar. *Psico*, *41*, 503-509.

Gomes, C.M.A., & Golino, H.F. (2012). Validade incremental da Escala de Abordagens de Aprendizagem. *Psicologia: Reflexão e Crítica*, *25*(4), 623-633.

Gomes, C.M.A., Golino, H.F., Pinheiro, C.A.R., Miranda, G.R., & Soares, J.M.T. (2011). Validação da Escala de Abordagens de Aprendizagem (EABAP) em uma amostra brasileira. *Psicologia: Reflexão e Crítica*, *24*(1), 19-27.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference and prediction* (2nd ed.). New York, NY: Springer.

Hutchinson, S., & Lovell, C. (2004). A review of methodological characteristics of research published in key journals in higher education. *Research in Higher Education*, *45*, 383-403.

ISNN 2182-7591

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

Kyndt, E. (2011). *Investigating students' approaches to learning.* Doctoral dissertation, Katholieke Universiteit Leuven, Leuven, Belgium.

Leek, J.T. (2013). *Predicting with trees.* Coursera's data analysis class material. Retrieved from: https://github.com/jtleek/dataanalysis.

Liaw, A., & Wiener, M. (2012). RandomForest: Breiman and Cutler's random forests for classification and regression (R package version 4.6-7.) [computer software manual]. Retrieved from: http://cran.r-project.org/web/packages/randomForest/.

Linacre, J.M. (2012). Winsteps® Rasch measurement computer program [computer software manual]. Beaverton, Oregon: Winsteps.com.

Marascuilo, L.A. (1966). Large-sample multiple comparisons. *Psychological Bulletin*, *65*(5), 280-290. Doi: 10.1037/h0023189.

Musso, M., Kyndt, E., Cascallar, E., & Dochy, F. (2012). Predicting mathematical performance: The effect of cognitive processes and self-regulation factors. *Education Research International*, Article ID 250719 (13 pages). Doi:10.1155/2012/250719.

Musso, M.F., Kyndt, E., Cascallar, E.C., & Dochy, F. (2013). Predicting general academic performance and identifying the differential contribution of participating variables using artificial neural networks. *Frontline Learning Research*, *1*, 42-71. http://dx.doi.org/10.14786/flr.v1i1.13.

Norton, L., & Crowley, C. (1995). Can students be helped to learn how to learn? An evaluation of an approaches to learning programme for first year degree students. *Higher Education*, *29*, 307-328.

R Development Core Team (2011). R: A language and environment for statistical computing [computer software manual]. R Foundation for Statistical Computing, Vienna, Austria, Retrieved from: http://www.R-project.org.

Ripley, B.D. (2013). Package "tree": Classification and regression trees (R package version 1.0-33.) [computer software manual]. Retrieved from: http://cran.r-project.org/web/packages/tree.

REVISTA E-PSI
REVISTA ELETRÓNICA DE PSICOLOGIA, EDUCAÇÃO E SAÚDE
ANO 4, VOLUME 1, 2014, pp.68-101.
https://www.revistaepsi.com
ISNN 2182-7591

Quatro Métodos de Machine Learning para Predizer o Desempenho Acadêmico de Estudantes Universitários: Um Estudo Comparativo

Resumo

O presente trabalho investiga a predição de desempenho academico (alto vs. baixo) por meio de quatro técnicas de machine learning (learning trees, bagging, Random Forest, e Boosting), usando um conjunto de testes e escalas psicológicas e educacionais nas seguintes áreas: inteligência, metacognição, conhecimento educacional básico prévio, abordagens de aprendizagem e processamento cognitivo básico. A amostra foi composta por 77 estudantes universitários (55% mulheres) matriculados no 2º e 3º ano de uma Escola de Medicina particular do estado de Minas Gerais, Brasil. A amostra foi dividida aleatoriamente em dois conjuntos, treino e teste, para realizar-se uma validação cruzada. No conjunto de treino, a acurácia total da predição variou entre 65% (bagging model) e 92.5% (boosting model), enquanto a sensibilidade variou entre 57.9% (learning tree) e 90% (boosting model) e a especificidade entre 66.7% (bagging model) e 95% (boosting model). A diferença no desempenho preditivo dos modelos, comparando-se o conjunto de treino e o de teste, variou entre -2.6% e 23.1% em termos da acuracia total, entre -5.6% e 27.5% na sensibilidade e entre 0% e 20% na especificidade, para os modelos bagging e boosting respectivamente. Esse resultado evidencia que esses modelos de machine learning podem atingir altos níveis de acuracia na predição do desempenho academico, mas a diferença na capacidade preditiva entre os conjuntos de treino e de teste indica que alguns modelos são mais estáveis que outros na predição. As vantagens dos modelos de árvore de machine learning na predição do desempenho acadêmico serão apresentadas e discutidas ao longo do texto.

Palavras-chave: Ensino Superior; *Machine Learning*; desempenho acadêmico, predição.

Como citar este artigo: Golino, H.F., & Gomes, C.M.A. (2014). Four machine learning methods to predict academic achievement of college students: A comparison study. *Revista E-Psi*, *4*(1), 68-101.

Received: November 12, 2013 Revision received: March 6, 2014 Accepted: April 1, 2014