UNIVERSIDADE FEDERAL DE MINAS GERAIS

HUDSON FERNANDES GOLINO

Modelos Complexos de Predição Aplicados na Educação

Belo Horizonte
2015

HUDSON FERNANDES GOLINO

Modelos Complexos de Predição Aplicados na Educação

Tese em formato de compilação de artigos apresentada à Universidade Federal de Minas Gerais, como parte dos requisitos para obtenção do grau de Doutor em Neurociências, pelo Programa de Pós-Graduação em Neurociências.

Área de concentração: Neurociência Clínica.

Orientador: Prof. Dr. Cristiano Mauro Assis Gomes

Belo Horizonte
2015

043 Golino, Hudson Fernandes.

Modelos complexos de predição aplicados na educação [manuscrito] / Hudson Fernandes Golino. – 2015.

127 f. : il. ; 29,5 cm.

Orientador: Cristiano Mauro Assis Gomes.

Tese (doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas.

1. Aprendizado do computador - Teses. 2. Educação - Teses. 3. Rendimento escolar - Previsão - Teses. 4. Desempenho - Teses. 5. Neurociências – Teses. I. Gomes, Cristiano Mauro Assis. II. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. IV. Título.

UNIVERSIDADE FEDERAL DE MINAS GERAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM NEUROCIÊNCIAS

UFMG

FOLHA DE APROVAÇÃO

Modelos Complexos de Predição Aplicados na Educação

HUDSON FERNANDES GOLINO

Tese submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em NEUROCIÊNCIAS, como requisito para obtenção do grau de Doutor em NEUROCIÊNCIAS, área de concentração NEUROCIÊNCIAS BÁSICAS.

Aprovada em 05 de março de 2015, pela banca constituída pelos membros:

Prof(a). Cristiano Mauro Assis Gomes - Orientador
UFMG

Prof(a). Enio Jelihovschi
UESC

Prof(a). Alessandra Gotuzo Seabra
Universidade Presbiteriana Mackenzie

Prof(a). Ana Paula Porto Noronha
Universidade São Francisco

Prof(a). Angela Maria Ribeiro
UFMG

Belo Horizonte, 5 de março de 2015.

LISTA DE FIGURAS E TABELAS

ARTIGO 1 – Publicado na Revista E-Psi, de Portugal.

ARTIGO 2 – Publicado na *Psychology*

ARTIGO 3 – Publicado na *Psychology*

ARTIGO 4 – Submetido para a *Psychological test and assessment modeling*

ARTIGO 5 – Artigo ainda não submetido

RESUMO

A presente tese apresenta a compilação de cinco artigos que empregam modelos complexos de predição na solução de problemas relacionados à educação. No primeiro artigo é apresentado os modelos de *classification and regression trees, bagging, random forest* e *boosting.* Esses modelos são empregados para montar um sistema preditivo de rendimento acadêmico de alunos do ensino superior de uma faculdade particular, tendo uma série de avaliações cognitivas como preditores. Já o segundo artigo emprega o modelo de *random forest* para predizer o desempenho escolar de alunos do primeiro ano do ensino médio da rede pública. Uma vez mais, um conjunto de avaliações cognitivas foram utilizadas. Já no terceiro artigo, apresentamos uma nova forma de visualizar a qualidade de predições realizadas utilizando a técnica de *random forest*. Essa nova técnica de visualização transforma informações estatísticas em um gráfico de redes, que possibilita o emprego de um conjunto de indicadores sobre a qualidade da predição, além dos usualmente empregados. No quarto artigo, apresentamos a técnica de *random forest* como uma nova forma de realizar imputação de dados faltantes. Investigamos o impacto da imputação no ajuste de itens de um teste cognitivo ao modelo dicotômico de Rasch, assim como na dificuldade estimada dos itens. No quinto e último artigo, comparamos o modelo de *classification trees* com o modelo *Naive Bayes* na predição de evasão acadêmica de alunos de uma faculdade pública estadual, tendo como preditores variáveis socioeconômicas. Esses artigos introduzem um conjunto de métodos quantitativos que pode auxiliar na resolução de problemas na área da educação, assim como podem levar a novas descobertas, não possibilitadas por meio de métodos usuais.

ABSTRACT

The current doctoral thesis presents a compilation of five papers employing complex predictive models to solve educational research issues. The first paper presents the classification and regression trees, as well as bagging, random forest and boosting algorithms. They are used to create an academic achievement predictive system, using a set of cognitive assessments/tests as independent variables (or predictors). The second paper, by its turn, uses the random forest algorithm to predict the academic achievement of high-school students. Once again, a set of cognitive assessments/tests were used as predictors. In the third paper, we introduce a new visualization technique that enables to visually inspect the quality of the prediction made using random forest. This technique is based on the plot of statistical information as a weighted graph, enabling the use of additional prediction quality indexes beyond total accuracy, sensitivity and specificity. The fourth paper presents the random forest algorithm as an imputation method, and investigates its impact on item fit to the dichotomous Rasch model and on their difficulty estimate. Finally, the fifth paper compares the classification tree with a Naïve Bayes classifier in the prediction of academic drop-out, using a set of socio-demographic variables as predictors. The papers presented in this doctoral dissertation introduce a set of innovative quantitative methods that have potential to solve a number of issues in the educational research field. They can also led to new discoveries, not allowed by other, more classical, methods.

Dedicatória

À memória de minha avó Carmen Fernandes, que ao tentar me fazer acreditar nas crendices populares do sul de Minas Gerais, sem querer me despertou para a investigação acerca das coisas do mundo (e principalmente a não acreditar em nada, e a desconfiar de tudo).

Agradecimentos

Agradeço infinitamente aos meus pais, Arnaldo e Dinah, que abriram mão de tantas coisas para que eu pudesse estudar e me dedicar à carreira que decidi seguir.

Agradeço imensamente ao meu orientador, Prof. Cristiano Mauro Assis Gomes, por todos os anos de trabalho duro e de grande aprendizado.

Agradeço à minha esposa, Mariana, por todo o amor, carinho e incentivo, tão importantes para minha vida pessoal e profissional.

Agradeço à todos os pesquisadores cujos trabalhos influenciaram direta ou indiretamente os meus projetos. A genialidade da maioria deles ensina não apenas a área ao qual dedicaram suas vidas, mas sobretudo que a ciência só faz sentido se o caminho for o da humildade perante as coisas. Felizmente, minha experiência mostra que a genialidade e a humildade são diretamente proporcionais.

Special thanks to Prof. Michael Lamport Commons and Prof. Patrice Marie Miller for their long-term support to my personal and academic projects.

Introdução

No ano de 2007 iniciei minhas atividades como aluno de iniciação científica sob orientação do Professor Cristiano Mauro Assis Gomes. Naquele período havia um projeto grande em andamento sobre mapeamento da arquitetura cognitiva, utilizando um conjunto de instrumentos de avaliação construídos para serem utilizados nessa pesquisa. Parte dos trabalhos envolvia avaliar um número grande de alunos da educação básica, do sexto ano à terceira série do ensino médio. Após a finalização da coleta dos dados, tivemos que dedicar um vasto tempo na elaboração de relatórios para os alunos, seus pais e para os diretores do colégio. Apesar de perceberem a importância da pesquisa e de gostarem de ver os relatórios de desempenho dos estudantes, a equipe do colégio não havia ficado plenamente satisfeita com a devolutiva dos resultados da pesquisa. Uma das reações mais comuns por parte da equipe do colégio dizia respeito ao que eles poderiam fazer com aqueles resultados. Em outras palavras, pareciam querer saber o que *concretamente* poderiam fazer tendo os resultados da pesquisa, e do desempenho dos alunos, em mãos.

Nos anos posteriores a 2007, continuamos com uma série de projetos que, invariavelmente, envolviam a coleta de dados em escolas ou faculdades. Em absolutamente todas as instituições onde nossos trabalhos foram desenvolvidos, o corpo de diretores, ou de coordenadores, eram monotemáticos. Queriam saber o que poderiam fazer com o resultado do desempenho dos estudantes em uma série de avaliações cognitivas. Explicávamos, de forma sucinta, os modelos teóricos de desenvolvimento da cognição, que serviam de fundamentação para os instrumentos utilizados, e em como eles poderiam ser úteis na condução da política educacional das instituições. Não importa a quantidade de reuniões que pudéssemos fazer, nem os artigos, livros e apresentações que enviávamos para esses profissionais. Simplesmente nada parecia estar dentro do conjunto de "ações concretas" que esses profissionais parecia querer.

Já no ano de 2012, após uma série de experiências fundamentalmente semelhantes às descritas acima, tivemos uma ideia. Poderíamos utilizar o desempenho dos estudantes nas avaliações cognitivas realizadas para construir sistemas de predição do desempenho dos mesmos no colégio ou na faculdade. Com isso, após a condução das atividades de pesquisa, poderíamos apresentar um resultado que possibilitasse ao corpo de diretores incorporar, de forma concreta, as nossas pesquisas em suas atividades institucionais. Além de melhorar o tipo de devolutiva realizada em nossas pesquisas, estaríamos contribuindo de forma significativa com a política educacional dessas instituições. Afinal de contas, ter em mãos uma forma de prever o desempenho acadêmica dos alunos, baseado em avaliações cognitivas, parecia ser uma espécie de "Santo Graal" educacional.

Quando começamos a tentar criar modelos de predição de desempenho educacional, a partir das avaliações que fazíamos, utilizamos um conjunto de ferramentas

usualmente empregadas no campo da psicometria. Dentre essas ferramentas, uma das mais amplamente utilizadas e úteis para o nosso objetivo, é a chamada *Modelagem de Equações Estruturais*. Por meio dessa técnica, é possível verificar o impacto de variáveis preditivas em uma ou mais variáveis alvo (ou variáveis dependentes). As técnicas utilizadas na modelagem de equação estrutural permitem, dentre outras coisas, verificar o percentual de variância explicada por cada variável preditiva, além dos seus intervalos de confiança. É uma ferramenta extremamente útil. No entanto, ela mudaria pouco o nosso cenário, uma vez que explicar o resultado de uma modelagem de equação estrutural para um público leigo seria pouco útil. Ademais, essa técnica tem uma série de pressupostos que são relativamente difíceis de serem satisfeitos nesse campo.

Sendo assim, começamos a estudar modelos inovadores de predição, ainda pouco utilizados na área da psicologia cognitiva, educacional e áreas afins (apesar de serem muito utilizadas em outros campos, como na ciência da computação), que além de possibilitar a construção de bons modelos preditivos, fossem relativamente simples de explicar para um público leigo. Uma das primeiras ferramentas que encontramos na literatura foi a chamada *Classification and Regression Trees* (CART). Essa ferramenta foi desenvolvida por Breiman e seus colaboradores na década de 1980 (Breiman, Friedman, Olshen, & Stone, 1984). Dentre várias vantagens da CART encontra-se o fato de que os modelos de predição desenvolvidos são facilmente explicáveis para profissionais com pouco ou nenhum conhecimento estatístico (Geurts, Irrthum, & Wehenkel, 2009), uma vez que assemelha-se a uma série de regras do tipo "se, então". Além disso, as predições realizadas utilizando-se CART geralmente levam a acurácias mais elevadas do que técnicas mais tradicionais da estatística, como regressão linear, logística, dentre outras (Geurts, Irrthum, & Wehenkel, 2009). Parecia que havíamos encontrado a ferramenta ideal, pois ela possibilitaria criar modelos preditivos relativamente simples de serem explicados, tinha boa chance de levar a acurácias minimamente aceitáveis para o campo, e ainda resolvia, como veremos mais abaixo, o problema dos pressupostos "hard" das técnicas usualmente usadas para realizar predições no campo da psicologia cognitiva e educacional (como a modelagem de equações estruturais). No entanto, como em tudo relacionado à ciência, a CART resolve alguns problemas e cria outros, como o sobreajuste e a variância (que também veremos a seguir). Essas características da CART nos fez seguir em busca de outros modelos que pudessem nos ajudar a criar modelos preditivos para as instituições de ensino.

O presente trabalho é uma compilação de cinco artigos que buscam criar modelos preditivos na área educacional trabalhando, no geral, a partir de avaliações cognitivas realizadas nos alunos. Para construir tais modelos empregamos diferentes ferramentas de uma área relativamente nova denominada de *Machine Learning*. Essa área engloba um conjunto de modelos estatísticos e computacionais concorrentes que objetivam a construção de predições, ou em outras palavras na construção de sistemas de ligação de variáveis independentes em variáveis dependentes. Nesse sentido, o primeiro artigo apresenta a *Classification and Regression Trees*, o *Bagging*, o *Random Forest* e o *Boosting*. Esses últimos três algoritmos utilizam assembleias de árvores de classificação ou regressão para realizar uma predição. São muito utilizados para lidar com o problema de

sobreajuste e de variância que a CART possui. Essas técnicas foram aplicadas tendo como variáveis preditivas uma série de avaliações cognitivas realizadas em um grupo de alunos do ensino superior, e como variável dependente o desempenho acadêmico dos mesmos.

O segundo artigo utiliza apenas o algoritmo de *Random Forest* para construir um modelo preditivo de desempenho acadêmico na primeira série do ensino médio de um colégio técnico federal. Uma vez mais, os alunos responderam à uma série de instrumentos de avaliação cognitiva, que entraram como as variáveis preditivas (ou variáveis independentes). A variável dependente utilizada foi a categoria "aprovado acima da média" ou "reprovado" no ano letivo. Uma das características do *Random Forest* é que ele constitui-se como uma "técnica de caixa preta", onde diferentemente da CART não há uma árvore típica de predição, e sim uma assembleia de árvores. Dessa forma, compreender o comportamento do algoritmo de predição, ou quão bem ele desempenha a tarefa de predizer a variável dependente, é bastante útil. Geralmente, o campo de *machine learning* baseia-se na utilização de indicadores de qualidade de predição, como a acurácia total, a sensibilidade e a especificidade. No entanto, é também possível utilizar técnicas gráficas para "visualizar" a qualidade da predição realizada. E é nesse caminho que se insere o terceiro artigo deste trabalho, no qual apresentamos uma nova forma de visualizar a qualidade de uma predição realizada via *Random Forest*, por meio do emprego de representações gráficas de rede.

Outro problema que geralmente enfrentamos na área da psicologia cognitiva e educacional, principalmente quando utilizamos instrumentos de avaliação de variáveis cognitivas, é a quantidade de dados faltantes. Eles surgem por uma série de motivos, que variam desde não conseguir responder à questão/tarefa/item, até o não querer responder, ou esquecer de dar uma resposta. Alguns modelos de análise de dados, como a família de modelos estatísticos de Rasch, da Teoria de Resposta ao Item, lida adequadamente com dados faltantes. No entanto, outros modelos são seriamente afetados na presença de dados faltantes. Por esse motivo, resolvemos realizar um estudo experimental propondo o algoritmo de *Random Forest* como uma forma de imputação de dados faltantes. O quarto artigo, portanto, mostra o impacto da imputação realizada via *Random Forest* no ajuste dos itens de um teste de raciocínio indutivo ao modelo dicotômico de Rasch, e no parâmetro de dificuldade.

Por último, tivemos acesso a um conjunto de dados socioeconômicos de uma Universidade Estadual localizada no interior da Bahia, que trazia também a situação dos alunos em seus cursos após cerca de quatro anos do ingresso na referida instituição. Já que estávamos trabalhando com a construção de modelos preditivos na área educacional, utilizando métodos inovadores de *Machine Learning*, resolvemos criar um modelo que predissesse a evasão acadêmica. Esse trabalho foi desenvolvido durante o meu período de Doutorado Sanduiche, no Instituto de Ciências Nucleares da Universidade Nacional Autônoma do México, que focou em mineração de dados.

Os cinco artigos apresentados como parte integrante do doutorado trazem uma série de inovações, seja pelo emprego de técnicas modernas de predição (artigos 1 e 2), pelo desenvolvimento de técnicas novas (artigo 3 e 4) ou pelo tipo de evidência

encontrada (artigo 5). O desenvolvimento desses trabalhos possibilitou resolver um dos problemas que vínhamos enfrentando em nossas pesquisas, pois por meio da construção de modelos de predição utilizando técnicas de *Machine Learning* é possível ou apresentá-los de forma relativamente fácil de compreender para os dirigentes das instituições de ensino (principalmente empregando a CART) ou aplicá-los de forma rápida e eficiente (por meio do chamado *model deployment*, que é a predição de novos dados baseados em modelos previamente elaborados e validados). Assim, é possível oferecer às instituições de ensino parceiras dos nossos projetos algo concreto, i.e. podemos ensiná-los a interpretar o resultado de avaliações cognitivas em termos da probabilidade do aluno ser aprovado ou reprovado no ano/semestre letivo, ou podemos aplicar as avaliações e entregar um resultado que indica se o aluno tem maior chance de ser aprovado ou reprovado. Além disso, esses trabalhos trazem um conjunto de novas possibilidades para o campo, com modelos que não se limitam, no geral, por pressupostos relativamente difíceis de serem atendidos.

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

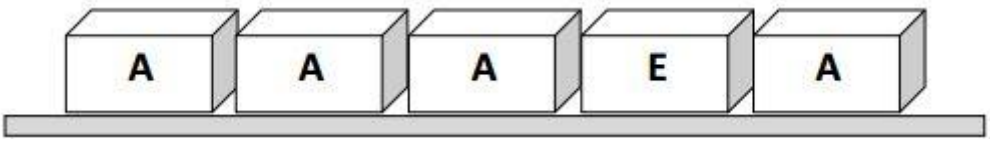
Participants

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%

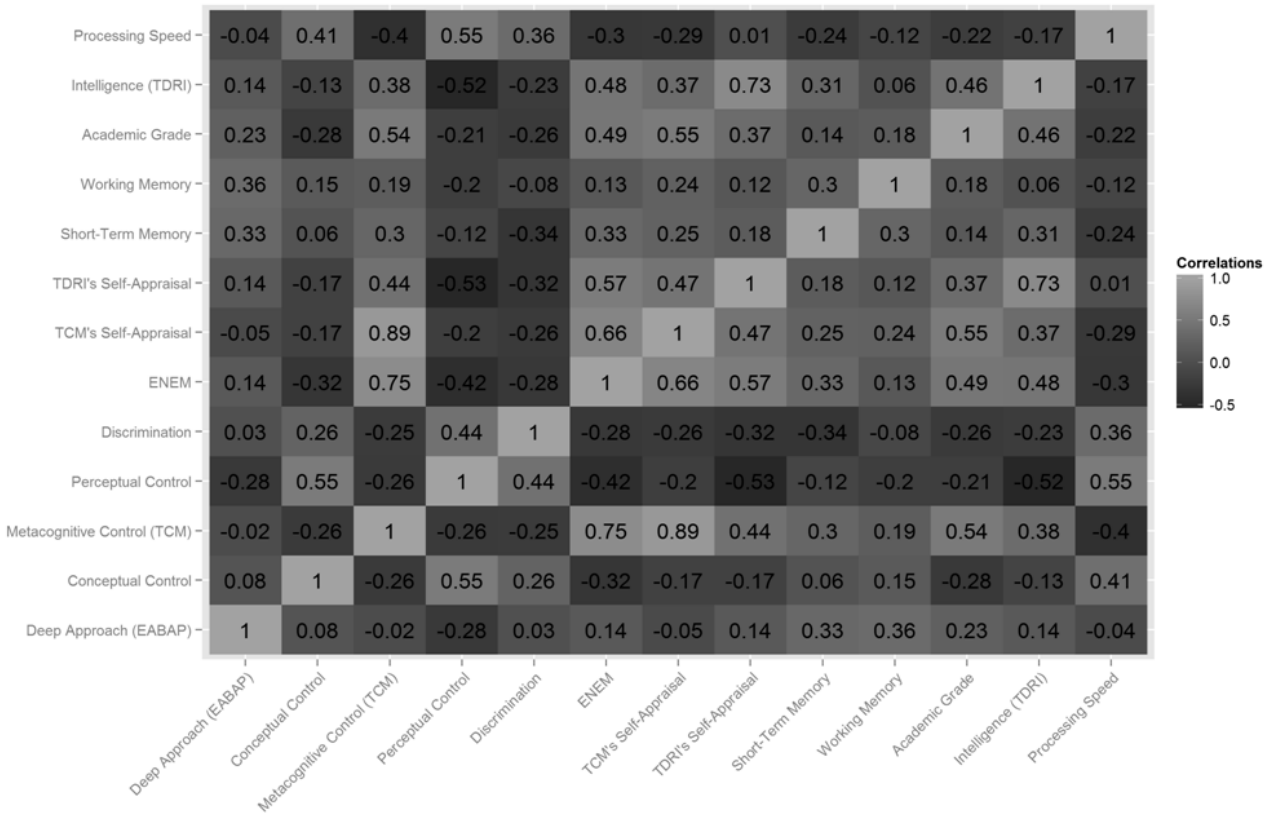
Considering the correlation matrix presented in Figure 2, the only variables with moderate correlations (greater than .30) with academic grade was the TCM (.54), the TDRI (.46), the ENEM exam (.49), the TCM Self-Appraisal Scale (.55) and the TDRI Self-Appraisal Scale (.37). The other variables presented 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.

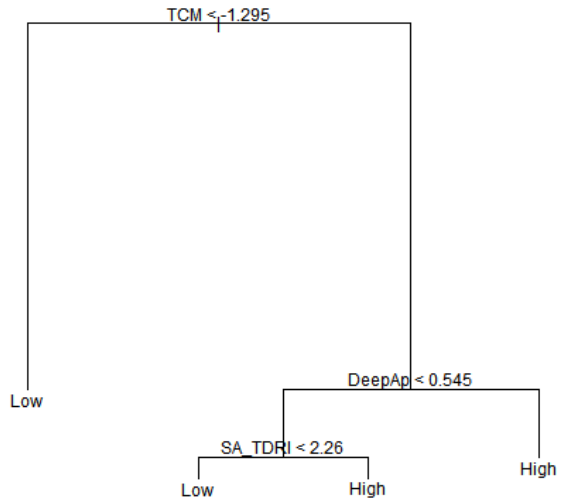


By its turn, people with TCM' Rasch score greater than -1.29 and with EABAP's Rasch score (DeepAp) greater than 0.54 are classified as being part of the high achievement class, with a probability of 60%. People are also classified as belonging to the high achievement class if they present a TCM' Rasch 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.

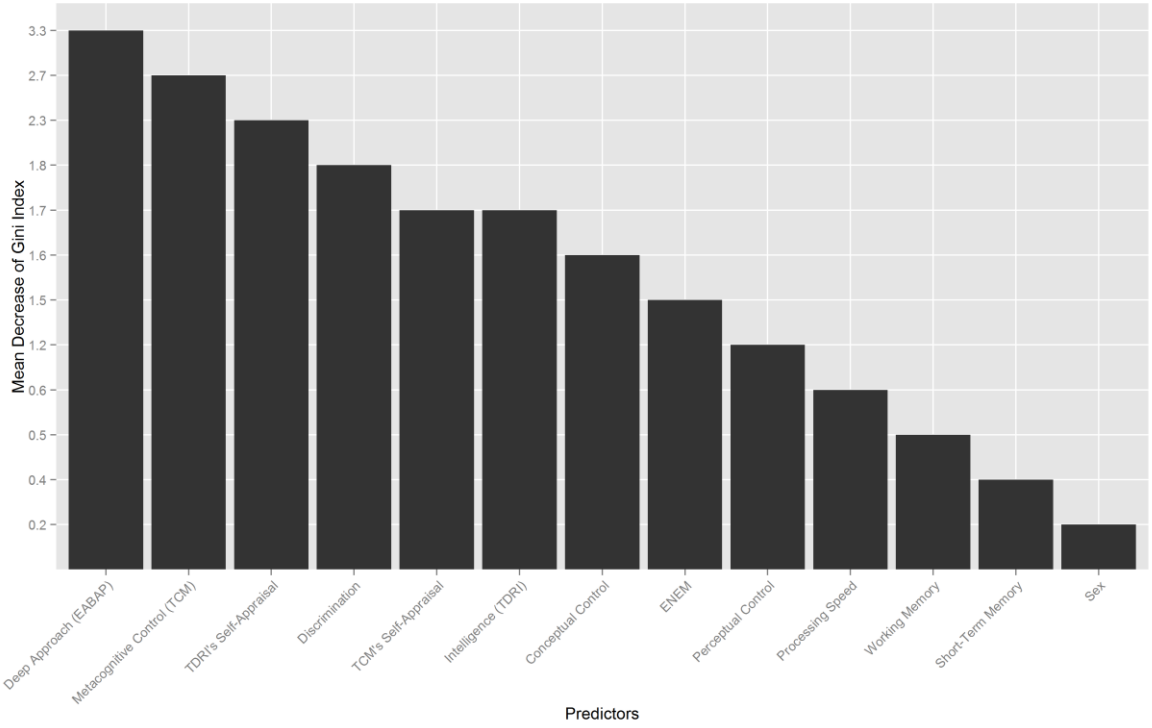
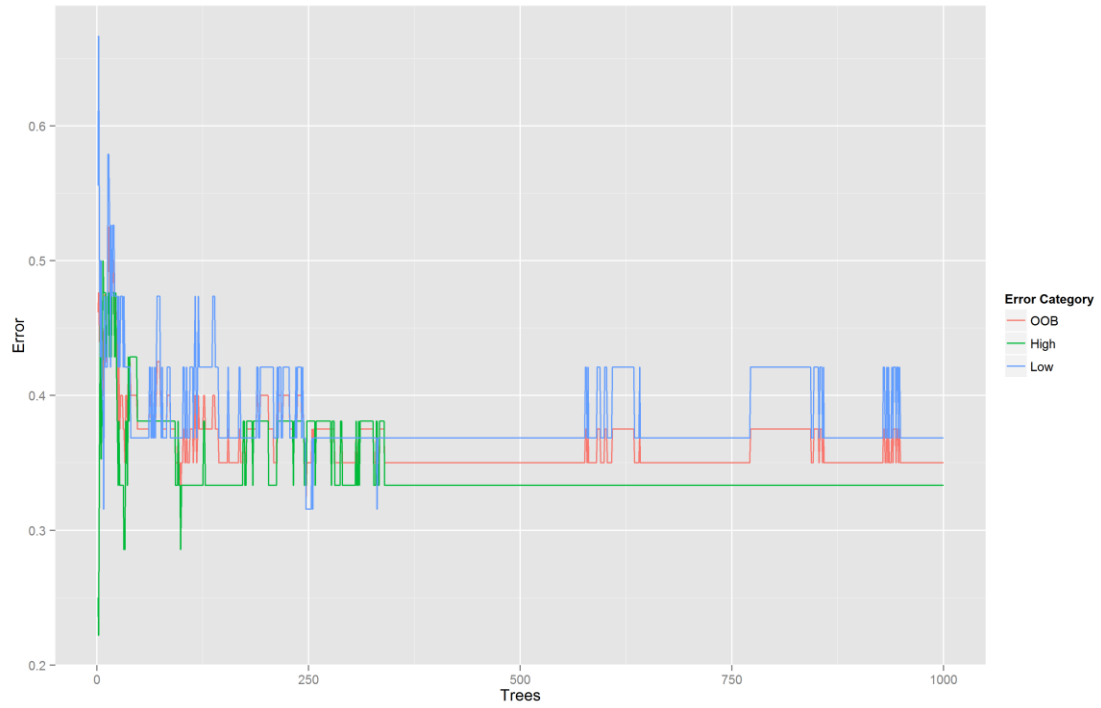


Figure 5 – Bagging's out-of-bag error (red), high achievement prediction error (green) and low achievement prediction error (blue).

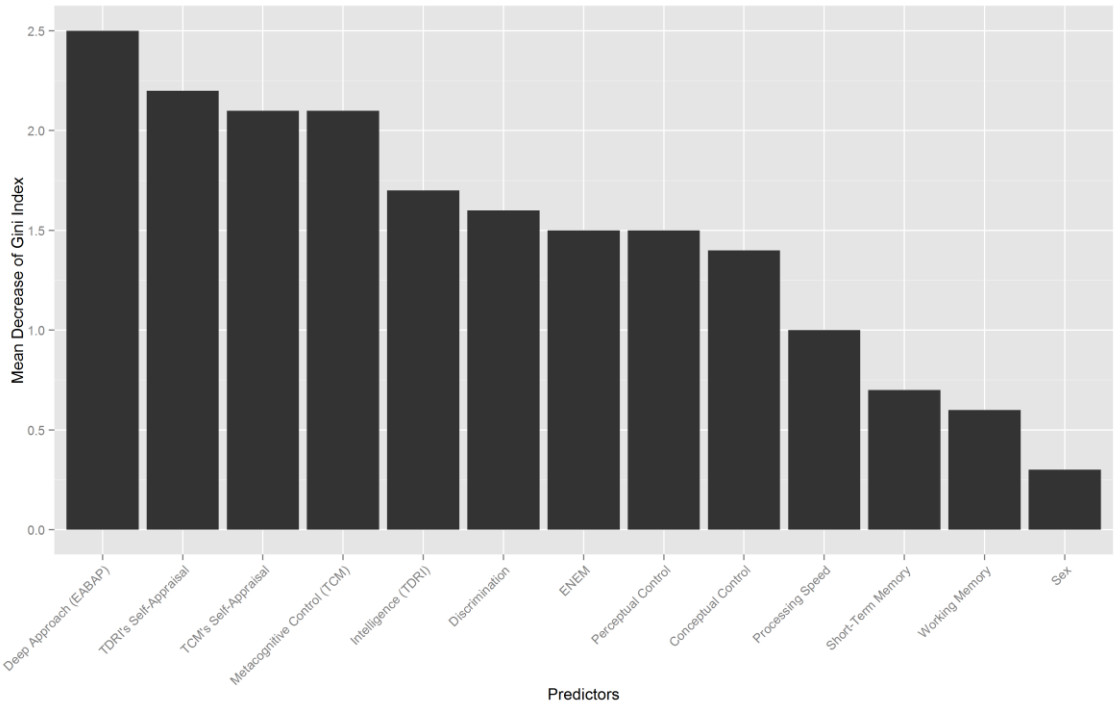


REVISTA E-PSI
REVISTA ELETRÓNICA DE PSICOLOGIA, EDUCAÇÃO E SAÚDE
ANO 4, VOLUME 1, 2014, pp.68-101.
https://www.revistaepsi.com
ISNN 2182-7591

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.

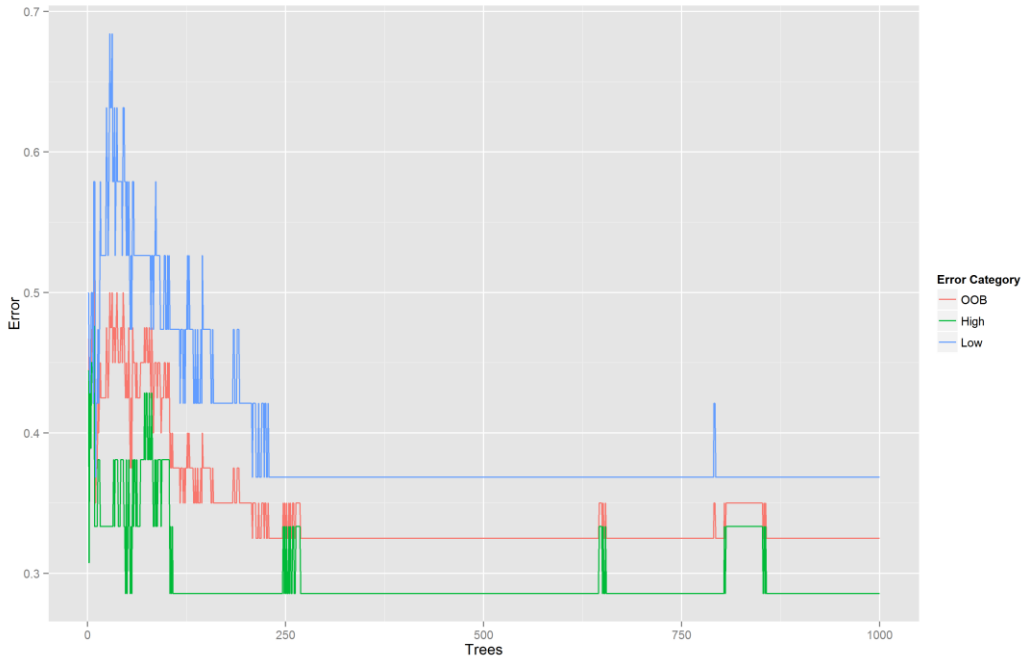


The Random Forest model was applied in the testing set for cross-validation, and presented a total accuracy of 72.97%, a sensitivity of 56.25% and a specificity of 81.71%. There was a difference of 5.47% in the total accuracy, of 6.91% in the sensitivity, and of 10.28% in the specificity.

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.

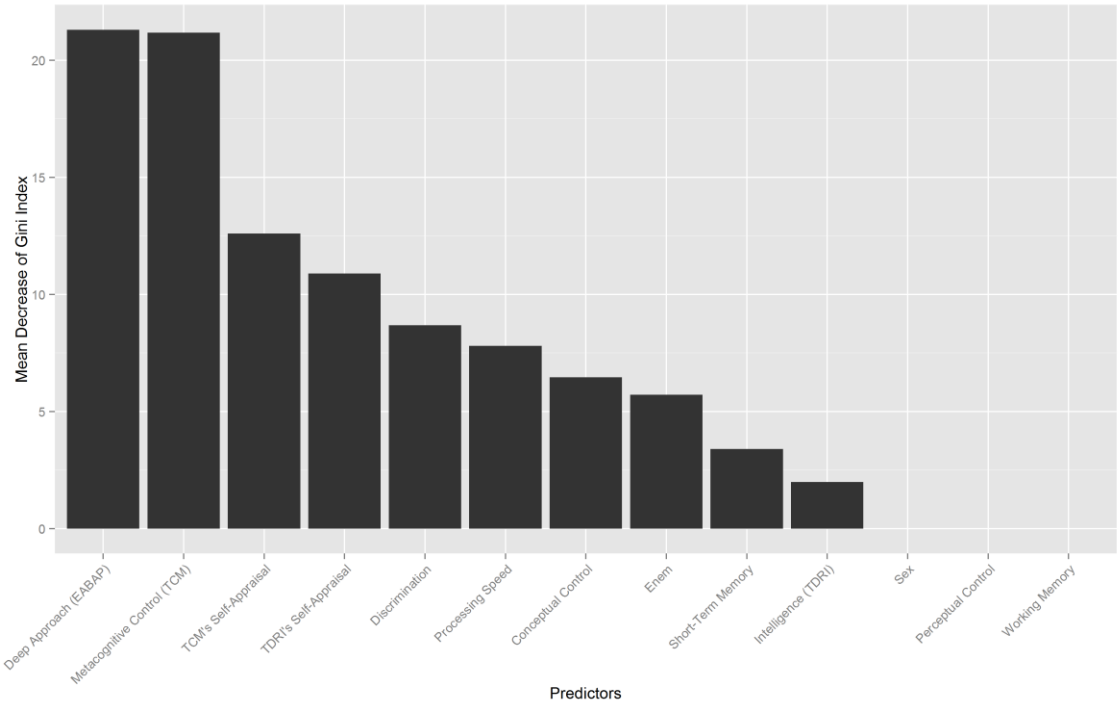
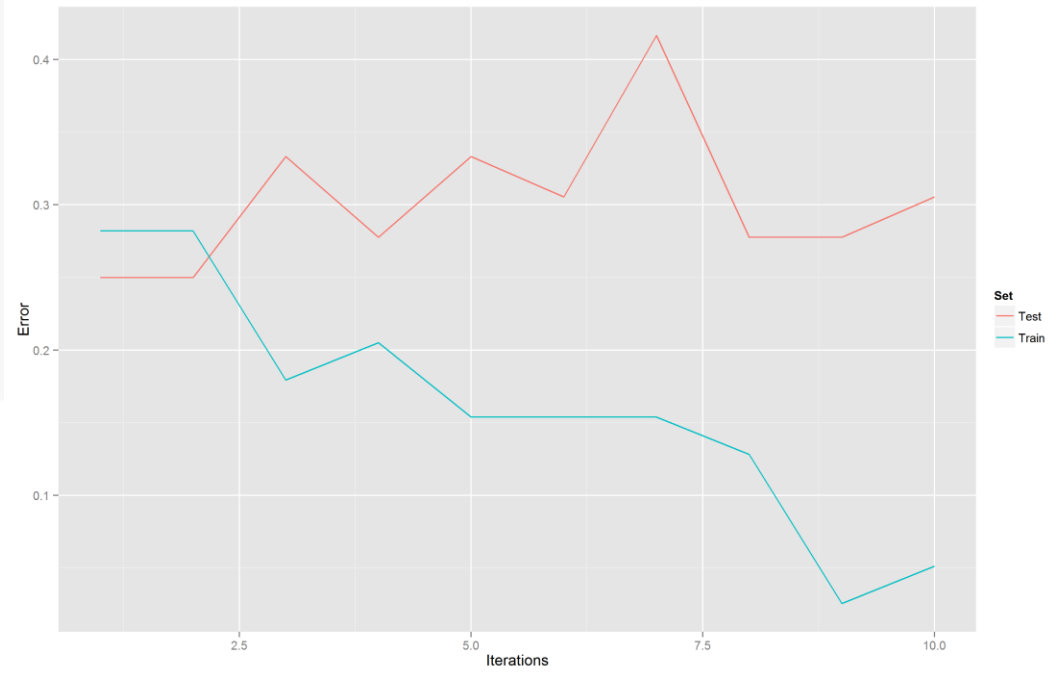


Figure 9 – Boosting's prediction error by iterations in the training and in the testing set.



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

Psychology, 2014, 5, 2046-2057
Published Online November 2014 in SciRes. http://www.scirp.org/journal/psych
http://dx.doi.org/10.4236/psych.2014.518207

Scientific Research

Predicting Academic Achievement of High-School Students Using Machine Learning

Hudson F. Golino[1], Cristiano Mauro Assis Gomes[2], Diego Andrade[2]

[1]Núcleo de Pós-Graduação, Pesquisa e Extensão, Faculdade Independente do Nordeste, Vitória da Conquista, Brazil
[2]Department of Psychology, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
Email: hfgolino@gmail.com, cristianogomes@ufmg.br

Received 27 September 2014; revised 23 October 2014; accepted 12 November 2014

Abstract

The present paper presents a relatively new non-linear method to predict academic achievement of high school students, integrating the fields of psychometrics and machine learning. A sample composed by 135 high-school students (10th grade, 50.34% boys), aged between 14 and 19 years old (M = 15.44, DP = 1.09), answered to three psychological instruments: the Inductive Reasoning Developmental Test (TDRI), the Metacognitive Control Test (TCM) and the Brazilian Learning Approaches Scale (BLAS-Deep Approach). The first two tests have a self-appraisal scale attached, so we have five independent variables. The students' responses to each test/scale were analyzed using the Rasch model. A subset of the original sample was created in order to separate the students in two balanced classes, high achievement (n = 41) and low achievement (n = 47), using grades from nine school subjects. In order to predict the class membership a machine learning non-linear model named Random Forest was used. The subset with the two classes was randomly split into two sets (training and testing) for cross validation. The result of the Random Forest showed a general accuracy of 75%, a specificity of 73.69% and a sensitivity of 68% in the training set. In the testing set, the general accuracy was 68.18%, with a specificity of 63.63% and with a sensitivity of 72.72%. The most important variable in the prediction was the TDRI. Finally, implications of the present study to the field of educational psychology were discussed.

Keywords

Machine Learning, Assessment, Prediction, Intelligence, Learning Approaches, Metacognition

How to cite this paper: Golino, H. F., Gomes, C. M. A., & Andrade, D. (2014). Predicting Academic Achievement of High-School Students Using Machine Learning. *Psychology, 5,* 2046-2057. http://dx.doi.org/10.4236/psych.2014.518207

1. Introduction

Machine learning is a relatively new science field composed by a broad class of computational and statistical methods to make predictions, inferences, and to discover new relations in data (Flach, 2012; Hastie, Tibshirani, & Friedman, 2009). There are two main areas within the machine learning field. The unsupervised learning focuses in the discovery and detection of new relationships, patterns and trends in data. The supervised learning area, by the other side, focuses in the prediction of an outcome using a given set of predictors. If the outcome is categorical, then the task to be accomplished is named classification, if it is numeric then the task is called regression.

There are several types of algorithms to perform classification and regression (Hastie et al., 2009). Among these algorithms, the tree based models are supervised learning techniques of special interest to the psychology and to the education research field. It can be used to discover which variable, or combination of variables, better predicts a given outcome, e.g. high or low academic achievement. It can identify the cutoff points for each variable that maximally predict the outcome, and can also be applied to study the non-linear interaction effects of the independent variables and its relation to the quality of the prediction (Golino & Gomes, 2014). Within psychology, there are a growing number of applications of the tree-based models in different areas, from ADHA diagnosis (Eloyan et al., 2012; Skogli et al., 2013) to perceived stress (Scott, Jackson, & Bergeman, 2011), suicidal behavior (Baca-Garcia et al., 2007; Kuroki & Tilley, 2012), adaptive depression assessment (Gibbons et al., 2013), emotions (Tian et al., 2014; van der Wal & Kowalczyk, 2013) and education (Blanch & Aluja, 2013; Cortez & Silva, 2008; Golino & Gomes, 2014; Hardman, Paucar-Caceres, & Fielding, 2013).

The main benefit of using the tree-based models in psychology is that they do not make any assumption regarding normality, linearity of the relation between variables, homoscedasticity, collinearity or independency (Geurts, Irrthum, & Wehenkel, 2009). The tree-based models also do not demand a high sample-to-predictor ratio and are more suitable to interaction effects (especially non-linearity) than the classical techniques, such as linear and logistic regression, ANOVA, MANOVA, structural equation modelling and so on. Finally, the tree-based models, especially the ensemble techniques, can lead to high prediction accuracy, since they are known as the state-of-the-art methods in terms of prediction accuracy (Flach, 2012; Geurts et al., 2009). The current paper focuses on the methodological aspects of the classification tree (Breiman, Friedman, Olshen, & Stone, 1984) and its most famous ensemble technique, Random Forest (Breiman, 2001a). To illustrate the use of tree-based models in educational psychology, the Random Forest algorithm will be used to predict levels of academic achievement of high school students (low vs. high). Finally, we will discuss the limits and possibilities of this new predictive method to the field of educational psychology.

Recursive Partitioning and Ensemble Techniques

A classification tree partitions the feature space into several distinct mutually exclusive regions (non-overlapping). Each region is fitted with a specific model that designates one of the classes to that particular space. The class is assigned to the 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 partition 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 (predictor) that leads to the most purity configuration. Then, each region of the tree is modeled accordingly to the majority class. One or two original nodes are also split into more nodes, using some of the given predictors that provide the best fit possible. This splitting process continues until the feature space achieves the most purity configuration possible, with Rm regions or nodes classified with a distinct Ck class. 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. In a classification tree, the first split indicates the most important variable, or feature, in the prediction. Let's take a look in **Figure 1** to see how a classification tree looks like.

Figure 1 shows the classification tree presented by Golino and Gomes (2014) with three predictors of the academic achievement (high and low) of medicine students: The Metacognitive Control Test (TCM), Deep Learning Approach (DeepAp) and the Self-Appraisal of the Inductive Reasoning Developmental Test (SA_TDRI). The most important variable in the prediction was TCM, since it was the predictor located at the first

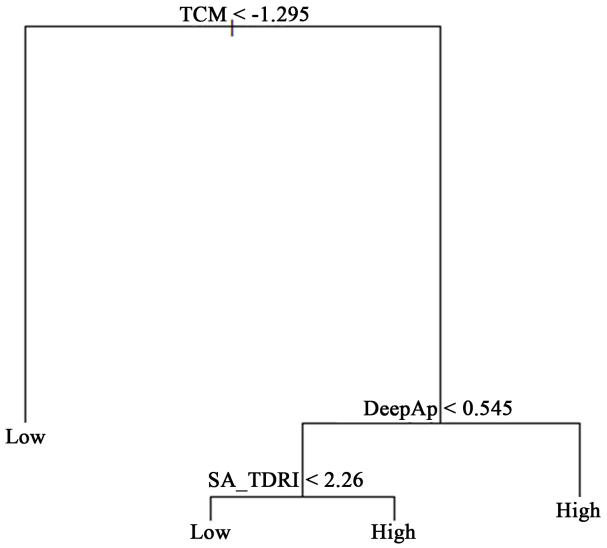


Figure 1. A classification tree from Golino and Gomes (2014).

split of the classification tree. The first split indicates the variable that separates the feature space into two purest nodes. In the case shown in **Figure 1**, 52.50% of the sample used to grow the tree had a TCM score smaller than −1.295, and were classified as having a low academic achievement. The remaining 47.5% had a TCM score greater than −1.295, and were classified in the low or in the high achievement class accordingly their scores on the DeepAp and on the SA_TDRI. Those with a TCM score greater than −1.295 and a DeepAp score greater than .545 were classified as belonging to the high achievement class. The same occurred to those with a TCM score greater than −1.295, a DeepAp score lower than .545 and a SA_TDRI score greater than 2.26. Finally, the participants with a TCM score greater than −1.295, a DeepAp score lower than .545 but with a SA_TDRI score smaller than 2.26 were classified as belonging to the low achievement group. This classification tree presented a total accuracy of 72.50%, with a sensitivity of 57.89% and a specificity of 85.71% (Golino & Gomes, 2014).

Geurts, Irrthum and Wehenkel (2009) argue that learning trees are among the most popular algorithms of machine learning due to its interpretability, flexibility and ease of use. Interpretability referrers to its easiness of understanding. It means that the model constructed to map the feature space (predictors) into the output space (dependent variable) is easy to understand, since it is a roadmap of if-then rules. The description of **Figure 1** above shows exactly that. 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. 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, it suffers 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. Golino and Gomes (2014) showed that in spite of having a total accuracy of 72.50% in the training sample, the tree presented in **Figure 1** behaved poorly in a testing set, with a total accuracy of 64.86%. The difference between the two data sets is due to the overfit of the tree to the training set.

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, as shown by the results of Golino and Gomes (2014).

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 trees can be used.

The ensemble trees are simply the junction of several models to perform the classification task based on the prediction made by every single tree. The most famous ensemble tree algorithm is the Random Forest (Breiman, 2001a), that is used to increase the prediction accuracy, decrease the variance between data sets and to avoid overfitting.

The procedure takes a random subsample of the original data set (with replacement) and of the feature space to grow the trees. The number of the selected features (variables) is smaller than the number of total elements of the feature space. Each tree assigns a single class to the each region of the feature space for every observation. Then, each class of each region of every tree grown 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 trees. As the Random Forest does not use the entire observations (only a subsample of it, usually 2/3), the remaining observations (known as out-of-bag, or OOB) is used to verify the accuracy of 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).

As pointed by Breiman (2001a), the number of selected variables is held constant during the entire procedure for growing the forest, and usually is set to square-root of the total number of variables. Since the Random Forest subsamples the original sample and the predictors, it is considered an improvement over other ensemble trees, as the bootstrap aggregating technique (Breiman, 2001b), or simply bagging. Bagging is similar to Random Forest, except for the fact that does not subsample the predictors. Thus, bagging creates correlated trees (Hastie et al., 2009), which may affect the quality of the prediction. The Random Forest algorithm decorrelates the trees grown, and as a consequence it also decorrelates the errors made by each tree, yielding a more accurate prediction.

Why decorrelating the trees is so important? Following the example created by James et al. (2013), 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 classes available in our data. By consequence, the bagged trees will be very similar to each other, making the predictions and the errors highly correlated. This may 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 variables 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. Therefore, the resulting trees will present less variance in the classification error and in the OOB error, leading to a more reliable prediction. In sum, the Random Forest is an ensemble of trees that improves the prediction accuracy, decreases variance and avoids overfitting by using only a subsample of the observations and a subsample of predictors. It has two main tuning parameters. The first is the size of the subsample of features used in each split (*mtry*), which is mandatory to be smaller than the total number of features, and is usually set as the square root of the total number of predictors. The second tuning parameter is the number of trees to grow (n_{tree}).

The present paper investigates the prediction of academic achievement of high-school students (high achievement vs. low achievement) using two psychological tests and one educational scale: the Inductive Reasoning Developmental Test (TDRI), the Metacognitive Control Test (TCM) and the Brazilian Learning Approaches Scale (BLAS-Deep approach). The first two tests have a self-appraisal scale attached, so we have five independent variables. In the next section will be presented the participants, instruments used and the data analysis procedures.

2. Method

2.1. Participants

The sample is composed by 135 high-school students (10th grade, 50.34% boys), aged between 14 and 19 years old (M = 15.44, DP = 1.09), from a public high-school from [omitted as required by the review process]. The sample was selected by convenience, and represents approximately 90% of the students of the 10th grade. The students received a letter inviting them to be part of the study. Those who agreed in participating signed a inform consent, and confirmed they would be present in the schedule days to answer all the instruments.

2.2. Measures and Procedures

2.2.1. The Inductive Reasoning Developmental Test (TDRI) and Its Self-Appraisal Scale (SA_TDRI)

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, 2008; Commons & Pekker, 2008; Commons & Richards, 1984) 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 (see **Figure 2**), 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.

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* mean = .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 [χ^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.

2.2.2. The Metacognitive Control Test (TCM) and Its Self-Appraisal Scale (SA_TCM)

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* mean = 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.

2.2.3. The Brazilian Learning Approaches Scale (EABAP)

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 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 only the deep learning approach items (DeepAp) were used. We will analyze only the nine deep approach items using the partial credit Rasch model.

Figure 2. Example of TDRI's item 1 (from the first developmental stage assessed).

2.3. Data Analysis

2.3.1. Estimating the Students' Ability in Each Test/Scale

The student's ability estimates on the inductive reasoning developmental test, on the metacognitive control test, on the Brazilian learning approaches scale, and on the self-appraisal scales were computed using the original data set of each test/scale, through the software Winsteps (Linacre, 2012). This procedure was followed in order to achieve reliable estimates, since only 135 students answered the tests. The mixture of the original data set from each test to the high-school students' answers did not significantly 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 (after adding the data from the high-school students) and the statistical model used is provided in **Table 1**.

2.3.2. Defining the Achievement Classes (High vs. Low)

The final grade in the following nine school subjects was provided by the school at the end of the academic year: arts, philosophy, physics, history, informatics, math, chemistry, sociology and Brazilian Portuguese. The final grades ranged from 0 to 10, and the students were considered approved in the academic year in each school subject only if he/she had a grade equal to or above seven. Students with grades lower than seven in a particular school subject are submitted to an additional assessment. Finally, those with an average grade of seven or more are considered able to proceed to the next school grade (11th grade). Otherwise, the students need to re-do the current grade (10th grade). From the total sample, only 65.18% (n = 88) were considered able to proceed to the next school year and 34.81% (n = 47) were requested to re-do the 10th grade. These two groups could be used to compose the high and the low achievement classes. However, since the tree-based models require balanced classes (i.e., classes with approximately the same number of cases) we needed to subset the high achievement class (those who proceeded to the next school grade) in order to obtain a subsample closer to the low achievement class size (those who would need to re-do the 10th grade). Therefore, we computed the mean final grade over all nine grades for every student, and verified the mean of each group of students. Those who passed to the next school grade had a mean final grade of 7.69 (SD = .48), while those who would need to re-do the 10th grade had a mean final grade of 6.06 (SD = 1.20). We select every student with a mean final grade equals to or higher than 7.69 (n = 41) and called them the "high achievement" group. The 47 students that would need to re-do the 10th grade formed the "low achievement" group. Finally, we had 88 students divided in two balanced classes.

2.3.3. Machine Learning Procedures

The sample was randomly split in two sets with equal sizes, training and testing, for cross-validation. 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.

Since the single trees usually lead to overiftting and to high variance between datasets, we used only the Random Forest algorithm through the random Forest package (Liaw & Wiener, 2012) of the R software (R Development Core Team, 2011). As pointed in the introduction, the Random Forest has two main tuning parameters: the number of trees (n_{tree}) and the number of variables used (*mtry*). We set *mtry* as two, because is the

Table 1. Item reliability, item fit, person reliability, person fit and model used by instrument.

Test	Item reliability	Item *INFIT* (mean, SD)	Person reliability	Person *INFIT* (mean, SD)	Model
Inductive reasoning developmental test (TDRI)	1.00	.98, .17	.85	.98, .91	Dichotomous Rasch Model
TDRI's self-appraisal scale (SA_TDRI)	.98	.98, .11	.79	.97, .31	Dichotomous Rasch Model
Metacognitive control test (TCM)	.99	1.00, .13	.80	.99, .31	Dichotomous Rasch Model
TCM's self-appraisal scale (SA_TCM)	.98	1.02, .26	.74	.98, .20	Dichotomous Rasch Model
Brazilian learning approaches scale—Deep learning items (DeepAp)	.99	1.00, .08	.80	1.01, .69	Partial Credit Rasch Model
Inductive reasoning developmental test (TDRI)	1.00	.98, .17	.85	.98, .91	Dichotomous Rasch Model

integer closest to the square root of the total number of predictors (5), and n_{tree} as 10,000. 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. Total accuracy is the proportion of observations correctly classified:

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

where $n|T_E|$ is the number of observations in the testing set. In spite of being an important indicator of the general prediction's quality, the total accuracy is not an informative measure of the errors in each class. For example, a general accuracy of 80% can represent an error-free prediction for the C1 class, and an error of 40% for the C2 class. In the educational scenario, it is preferable to have lower error in the prediction of the low achievement class, since students at risk of academic failure compose this class. So, the sensitivity will be preferred over general accuracy and specificity. The sensitivity is the rate of observations correctly classified in a target class, e.g. C1 = low achievement, over the number of observations that belong to that class:

$$Sens = \frac{\sum_{x \in T_E} I\left(y_i = C_1\right)}{\sum_{x \in T_E} I\left(C_1\right)}$$

Specificity, on the other hand, is the rate of correctly classified observations of the non-target class, e.g. C2 = high achievement, over the number of observations that belong to that class:

$$Spec = \frac{\sum_{x \in T_E} I\left(y_i = C_2\right)}{\sum_{x \in T_E} I\left(C_2\right)}$$

Finally, the model construct in the training set will be applied in the testing set for cross-validation. Since the Random Forest is a black box technique—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 one importance measure will be used: the mean decrease of accuracy. It 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).

2.3.4. Descriptive Analysis Procedures

After estimating the student's ability in each test or scale the Shapiro-Wilk test of normality will be conducted in order to discover which variables presented a normal distribution. To verify if there is any statistically significant difference between the students' groups (high achievement vs. low achievement) the two-sample T test will be conducted in the normally distributed variables and the Wilcoxon Sum-Rank test in the non-normal variables, both at the .05 significance level. In order to estimate the effect sizes of the differences, the R's compute.es package (Del Re, 2013) is 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. McGraw and Wong (1992) developed the CLES indicator as a more intuitive tool than the other effect size indicators. It converts an effect into a probability that a score taken at random from one distribution will be greater than a score taken at random from another distribution (McGraw & Wong, 1992). In other words, it 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).

3. Results

3.1. Descriptive

The Brazilian Learning Approaches Scale (Deep Learning) presented a normal distribution (W = .99, *p*-value = .64), while all the other four variables presented a *p*-value smaller than .001. There was a statistically significant difference at the 99% level between the high and the low achievement groups in the median Rasch score of the Inductive Reasoning Developmental ($\tilde{x}_{High} = 2.14$, $\sigma^2 = 5.80$, $\tilde{x}_{Low} = -1.47$, $\sigma^2_{Low} = 15.52$, W = 1359, $p < .01$), in the median Rasch score of the Metacognitive Control Test ($\tilde{x}_{High} = -1.03$, $\sigma^2 = 7.29$, $\tilde{x}_{Low} = -3.40$, $\sigma^2_{Low} = 4.37$, W = 928, $p < .01$), in the median Rasch score of the TDRI's self-appraisal scale ($\tilde{x}_{High} = 2.03$, $\sigma^2 =$

3.01, $\tilde{x}_{Low} = 1.16$, $\sigma^2_{Low} = 4.66$, W = 1152, $p < .001$), in the median Rasch score of the TCM's self-appraisal scale ($\tilde{x}_{High} = 1.07$, $\sigma^2 = 4.18$, $\underline{x}_{Low} = -1.08$, $\sigma^2_{Low} = 2.45$, W = 954, $p < .01$) and in the mean Rasch score of the Brazilian learning approaches scale-deep approach ($\bar{x}_{High} = 1.13$, $\sigma^2 = .80$, $\bar{x}_{Low} = .50$, $\sigma^2_{Low} = .61$, t(37) = 3.32, $p < .01$). The effect sizes, its 95% confidence intervals, variance, significance and common language effect sizes are described in **Table 2**.

According to Cohen (1988), the effect size is considered small when it is between .20 and .49, moderate between .50 and .79 and large when values are over .80. Only the difference in the Rasch score of the inductive reasoning developmental test presented a large effect size (d = .88, $p < .05$).

As pointed before, the common language effect size indicates how often a score sampled from one distribution is greater than the score sampled from the other distribution if both are randomly selected (McGraw & Wong, 1992). Then, considering the common language effect size, the probability that a TDRI score taken at random from the high achievement group is greater than a TDRI score taken at random of the low achievement group is 73.41%. It means that out of 100 TDRI scores from the high achievement group, 73.41 will be greater than the TDRI scores of the low achievement group. The Rasch scores of the other tests have moderate effect sizes. Their common language effect size varied from 64.92% to 70.10%, meaning that the probability of a score taken at random at the high achievement group be greater than a score taken at random in the low achievement group is at least 64.92% and at most 70.10%. **Figure 3** shows the mean score for each test and its 95% confidence interval by both classes (low and high).

3.2. Machine Learning Results

The result of the Random Forest model with 10,000 trees showed an out-of-bag error rate of .29, a total accuracy of 75.00%, a sensitivity of 68.00% and a specificity of 73.69%. The mean decrease accuracy showed the inductive reasoning developmental stage (TDRI) as the most important variable in the prediction, since when it is left out of the prediction the accuracy decreases 66.22% in average. The second most important variable is the deep learning approach, which is associated with a mean decrease accuracy of 28.45% when is not included in the predictive model. In third place is the metacognitive control test (19.68%); in the fourth position is the TDRI self-appraisal scale (19.50%), followed by the TCM self-appraisal scale (5.78%). **Figure 4** shows the high achievement prediction error (green line), the out-of-bag error (red line) and the low achievement prediction error (black line) per tree. The errors become more stable with approximately more than 1700 trees.

The predictive model constructed in the training set was applied in the testing set for cross-validation. It presented a total accuracy of 68.18%, a sensitivity of 72.72% and a specificity of 63.63%. There was a difference of 6.82% in the total accuracy, of 2.28% in the sensitivity, and of 10.06% in the specificity.

4. Discussion

The present paper briefly introduced the concept of recursive partitioning used in the tree-based models of machine learning. The tree-based models are very useful to study the role of psychological and educational constructs in the prediction of academic achievement. Unlike the most classical approaches, such as linear and logistic regression, as well as the structural equation modeling, the tree-based models do not make assumptions about the normality of data, the linearity of the relation between the variables, neither requires homoscedasticity, collinearity or independence (Geurts, Irrthum, & Wehenkel, 2009). A high predictor-to-sample ratio can be used

Table 2. Tests, effect sizes and common language effect size (CLES).

Test	Effect size of the difference (d)	95% C.I. (d)	σ^2 (d)	*p*-value (d)	CLES
Inductive reasoning developmental test (TDRI)	.88	.43, 1.34	.05	.00	73.41%
Metacognitive control test (TCM)	.59	.11, 1.06	.06	.02	66.05%
TDRI' self-appraisal scale (SA_TDRI)	.54	.10, .99	.05	.02	64.92%
TCM' self-appraisal scale (SA_TCM)	.65	.17, 1.12	.06	.01	67.62%
EABAP (DeepAp)	.75	.27, 1.22	.06	.00	70.10%

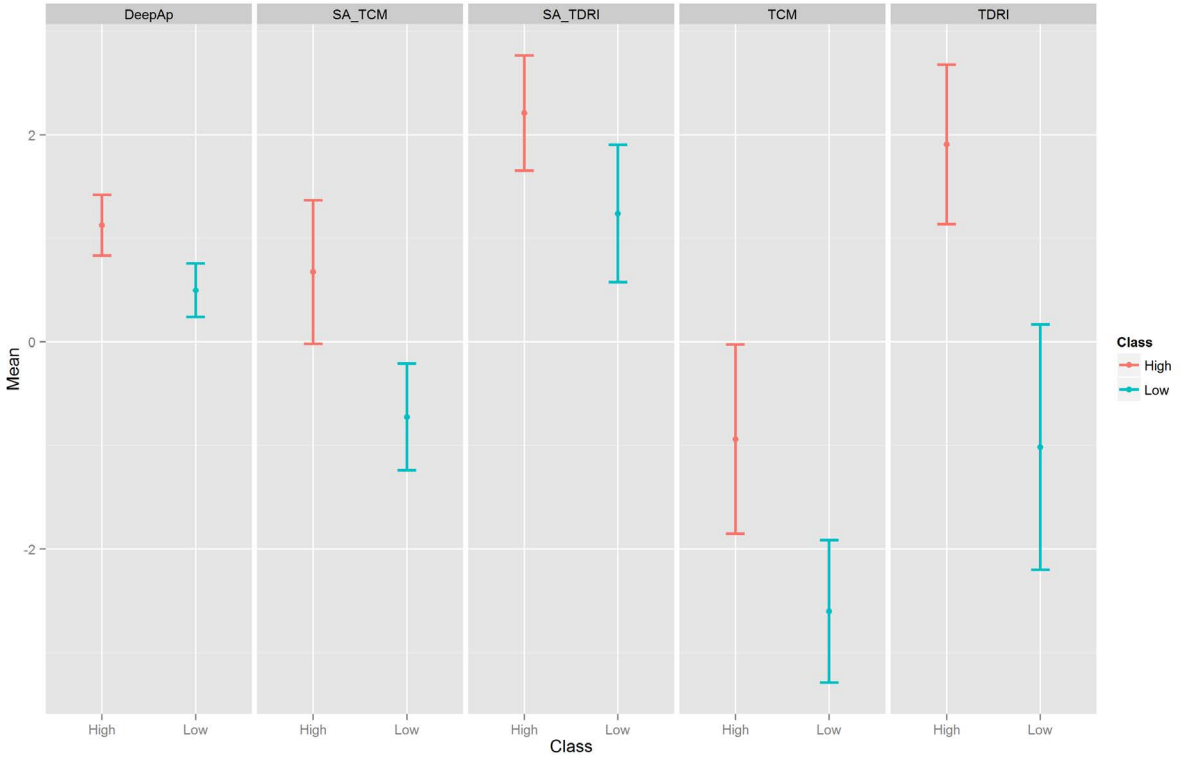


Figure 3. Score means and its 95% confidence intervals for each test, by class (high vs. low academic achievement).

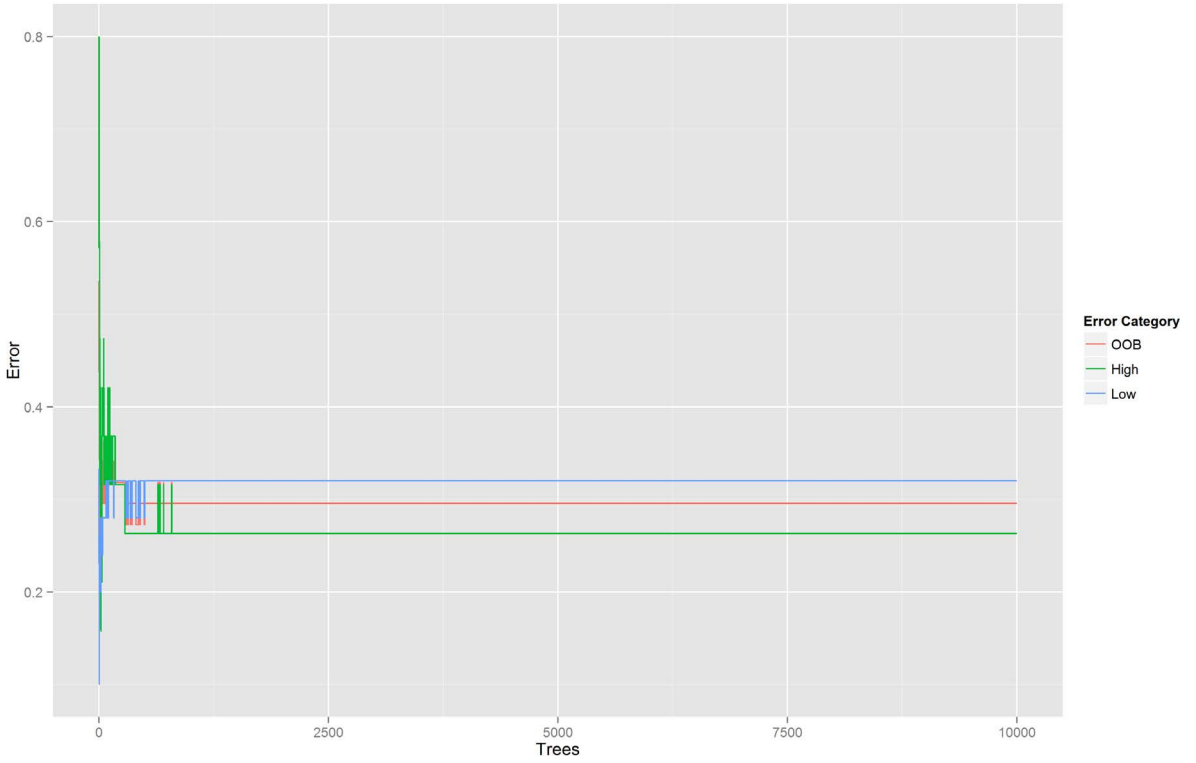


Figure 4. Random Forest's out-of-bag error (red), high achievement prediction error (green) and low achievement prediction error (blue).

without harm to the quality of the prediction, and missingness is well handled by the prediction algorithms. The tree-based models are also more suitable to non-linear interaction effects than the classical techniques. When several trees are ensemble to perform a prediction it generally leads to a high accuracy (Flach, 2012; Geurts et al., 2009), decreasing the chance of overfitting and diminishing the variance between datasets. The focus of the current paper was the application of this relatively new predictive method in the educational psychology field.

Psychology is taking advantage of the tree-based models in a broad set of applications (Baca-Garcia et al., 2007; Eloyan et al., 2012; Gibbons et al., 2013; Kuroki & Tilley, 2012; Scott, Jackson, & Bergeman, 2011; Skogli et al., 2013; Tian et al., 2014; van der Wal & Kowalczyk, 2013). Within education, Blanch and Aluja (2013), Cortes and Silva (2008) and Golino and Gomes (2014) applied the tree-based models to predict the academic achievement of students from the secondary and tertiary levels using a set of psychological and socio-demographic variables as predictors. The discussion of their methods and results are beyond the scope of the current paper, since we focused on the methodological aspects of machine learning, and how it can be applied in the educational psychology field.

In the present paper we showed the Rasch scores of the tests and scales used significantly differentiated the high achievement from the low achievement 10th grade students. Inductive reasoning presented a large effect size, while the deep learning approach, metacognitive control and self-appraisals presented moderate effect sizes. The random forest prediction lead to a total accuracy of 75%, a sensitivity of 68% and a specificity of 73.69% in the training set. The testing set result was a little bit worse, with a total accuracy of 68.18%, a sensitivity of 72.72% and a specificity of 63.63%. The most important variable in the prediction was the inductive reasoning that was associated with a mean decrease accuracy of 66.22% when left out of the prediction bag. The deep learning approach was the second most important variable (mean decrease accuracy of 28.45%), followed by metacognitive control (19.68%), TDRI self-appraisal (19.50%) and TCM self-appraisal (5.78%). This result reinforces previous findings that showed incremental validity of the learning approaches in the explanation of academic performance beyond intelligence, using traditional techniques (Chamorro-Premuzic & Furnham; 2008; Furnham Monsen, & Ahmetoglu, 2009; Gomes & Golino, 2012). It also reinforces the incremental validity of metacognition, over intelligence, in the explanation of academic achievement (van der Stel & Veenman, 2008; Veenman & Beishuizen, 2004).

5. Conclusion

The application of machine learning models in the prediction of academic achievement/performance, especially the tree-based models, represents an innovative complement to the traditional techniques such as linear and logistic regression, as well as structural equation modelling (Blanch & Aluja, 2013). More than the advantages pointed earlier, the tree-based models can help us to understand the non-linear interactions between psycho-educational variables in the prediction of academic outcomes. These machine learning models not only represent an advance in terms of prediction accuracy, but also represent an advance in terms of inference. Future studies could benefit from employing a larger and broader sample, involving students from different schools. It would also be interesting to investigate, in the future, the impact of varying the tuning parameters of the random forest model in the accuracy, sensitivity, specificity and variability of the prediction.

Acknowledgements

The current research was financially supported by a grant provided by the Fundação de Amparo à Pesquisa de Minas Gerais (FAPEMIG) to the authors. The authors also receive grants provided by the Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) and by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) of the Brazil's Ministry of Science, Technology and Innovation.

References

Baca-Garcia, E., Perez-Rodriguez, M., Saiz-Gonzalez, D., Basurte-Villamor, I., Saiz-Ruiz, J., Leiva-Murillo, J. M., & de Leon, J. (2007). Variables Associated with Familial Suicide Attempts in a Sample of Suicide Attempters. *Progress in Neuro-Psychopharmacology & Biological Psychiatry, 31,* 1312-1316. http://dx.doi.org/10.1016/j.pnpbp.2007.05.019

Blanch, A., & Aluja, A. (2013). A Regression Tree of the Aptitudes, Personality, and Academic Performance Relationship. *Personality and Individual Differences, 54,* 703-708. http://dx.doi.org/10.1016/j.paid.2012.11.032

Breiman, L. (2001a). Random Forests. *Machine Learning, 1,* 5-32. http://dx.doi.org/10.1023/A:1010933404324

Breiman, L. (2001b). Bagging Predictors. *Machine Learning, 24,* 123-140. http://dx.doi.org/10.1007/BF00058655

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

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.), Hillsdale, NJ: Lawrence Erlbaum Associates.

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. http://dx.doi.org/10.1080/02604020802301105

Commons, M. L., & Pekker, A. (2008). Presenting the Formal Theory of Hierarchical Complexity. *World Futures, 64,* 375-382. http://dx.doi.org/10.1080/02604020802301204

Cortez, P., & Silva, A. M. G. (2008). Using Data Mining to Predict Secondary School Student Performance. In A. Brito, & J. Teixeira (Eds.), *Proceedings of 5th Annual Future Business Technology Conference*, Porto, 5-12.

Del Re, A. C. (2013). compute.es: Compute Effect Sizes. R Package Version 0.2-2. http://cran.r-project.org/web/packages/compute.es

Eloyan, A., Muschelli, J., Nebel, M., Liu, H., Han, F., Zhao, T., Caffo, B. et al. (2012). Automated Diagnoses of Attention Deficit Hyperactive Disorder Using Magnetic Resonance Imaging. *Frontiers in Systems Neuroscience, 6,* 61. http://dx.doi.org/10.3389/fnsys.2012.00061

Fischer, K. W. (1980). A Theory of Cognitive Development: The Control and Construction of Hierarchies of Skills. *Psychological Review, 87,* 477-531. http://dx.doi.org/10.1037/0033-295X.87.6.477

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: Psychology Press.

Flach, P. (2012). *Machine Learning: The Art and Science of Algorithms That Make Sense of Data.* Cambridge: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511973000

Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives, 19,* 25-42. http://dx.doi.org/10.1257/089533005775196732

Geurts, P., Irrthum, A., & Wehenkel, L. (2009). Supervised Learning with Decision Tree-Based Methods in Computational and Systems Biology. *Molecular BioSystems, 5,* 1593-1605. http://dx.doi.org/10.1039/b907946g

Gibbons, R. D., Hooker, G., Finkelman, M. D., Weiss, D. J., Pilkonis, P. A., Frank, E., Moore, T., & Kupfer, D. J. (2013). The Computerized Adaptive Diagnostic Test for Major Depressive Disorder (CAD-MDD): A Screening Tool for Depression. *Journal of Clinical Psychiatry, 74,* 669-674. http://dx.doi.org/10.4088/JCP.12m08338

Golino, H. F., & Gomes, C. M. A. (2012). 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, 7-8 July 2012.

Golino, H. F., & Gomes, C. M. A. (2013). 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.

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,* 68-101.

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. 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,* 623-633. http://dx.doi.org/10.1590/S0102-79722012000400001

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,* 19-27. http://dx.doi.org/10.1590/S0102-79722011000100004

Hardman, J., Paucar-Caceres, A., & Fielding, A. (2013). Predicting Students' Progression in Higher Education by Using the Random Forest Algorithm. *Systems Research and Behavioral Science, 30,* 194-203. http://dx.doi.org/10.1002/sres.2130

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction* (2nd ed.). New York: Springer. http://dx.doi.org/10.1007/978-0-387-84858-7

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R.* New York: Springer. http://dx.doi.org/10.1007/978-1-4614-7138-7

Kuroki, Y., & Tilley, J. L. (2012). Recursive Partitioning Analysis of Lifetime Suicidal Behaviors in Asian Americans. *Asian American Journal of Psychology, 3,* 17-28. http://dx.doi.org/10.1037/a0026586

Liaw, A., & Wiener, M. (2012). Random Forest: Breiman and Cutler's Random Forests for Classification and Regression. R Package Version 4.6-7. http://cran.r-project.org/web/packages/randomForest/

Linacre, J. M. (2012). *Winsteps® Rasch Measurement Computer Program*. Beaverton, OR: Winsteps.com.

McGraw, K. O., & Wong, S. P. (1992). A Common Language Effect Size Statistic. *Psychological Bulletin, 111,* 361-365. http://dx.doi.org/10.1037/0033-2909.111.2.361

Scott, S. B., Jackson, B. R., & Bergeman, C. S. (2011). What Contributes to Perceived Stress in Later Life? A Recursive Partitioning Approach. *Psychology and Aging, 26,* 830-843. http://dx.doi.org/10.1037/a0023180

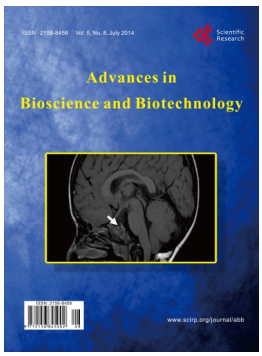
Skogli, E., Teicher, M. H., Andersen, P., Hovik, K., & Øie, M. (2013). ADHD in Girls and Boys—Gender Differences in Co-Existing Symptoms and Executive Function Measures. *BMC Psychiatry, 13,* 298. http://dx.doi.org/10.1186/1471-244X-13-298

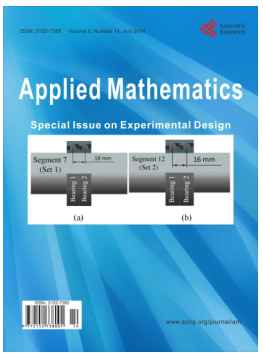
Tian, F., Gao, P., Li, L., Zhang, W., Liang, H., Qian, Y., & Zhao, R. (2014). Recognizing and Regulating e-Learners' Emotions Based on Interactive Chinese Texts in e-Learning Systems. *Knowledge-Based Systems, 55,* 148-164. http://dx.doi.org/10.1016/j.knosys.2013.10.019

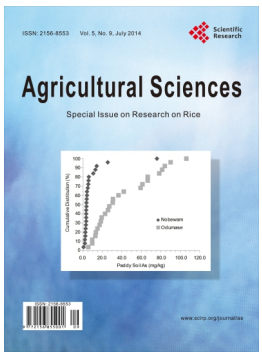
van der Wal, C., & Kowalczyk, W. (2013). Detecting Changing Emotions in Human Speech by Machine and Humans. *Applied Intelligence, 39,* 675-691. http://dx.doi.org/10.1007/s10489-013-0449-1

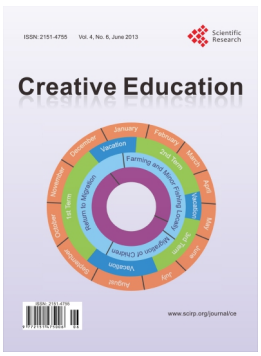
Scientific Research Publishing (SCIRP) is one of the largest Open Access journal publishers. It is currently publishing more than 200 open access, online, peer-reviewed journals covering a wide range of academic disciplines. SCIRP serves the worldwide academic communities and contributes to the progress and application of science with its publication.

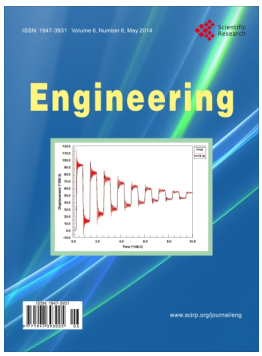
Other selected journals from SCIRP are listed as below. Submit your manuscript to us via either submit@scirp.org or Online Submission Portal.

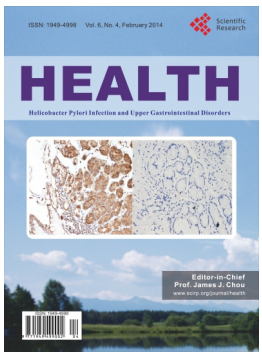


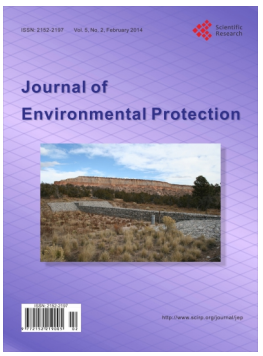


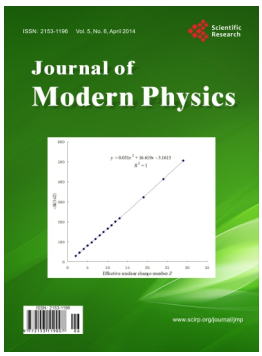


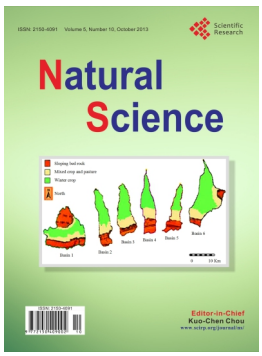


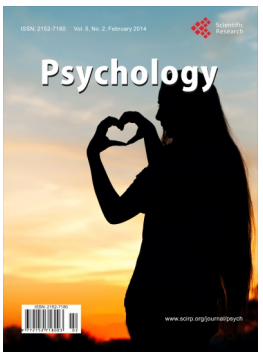












Psychology, 2014, 5, 2084-2098
Published Online December 2014 in SciRes. http://www.scirp.org/journal/psych
http://dx.doi.org/10.4236/psych.2014.519211

Scientific Research

Visualizing Random Forest's Prediction Results

Hudson F. Golino[1], Cristiano Mauro Assis Gomes[2]

[1]Núcleo de Pós-Graduação, Pesquisa e Extensão, Faculdade Independente do Nordeste, Vitória da Conquista, Brazil
[2]DepartmentofPsychology, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
Email: hfgolino@gmail.com, cristianogomes@ufmg.br

Received 17 October 2014; revised 8 November 2014; accepted 1 December 2014

Abstract

The current paper proposes a new visualization tool to help check the quality of the random forest predictions by plotting the proximity matrix as weighted networks. This new visualization technique will be compared with the traditional multidimensional scale plot. The present paper also introduces a new accuracy index (proportion of misplaced cases), and compares it to total accuracy, sensitivity and specificity. It also applies cluster coefficients to weighted graphs, in order to understand how well the random forest algorithm is separating two classes. Two datasets were analyzed, one from a medical research (breast cancer) and the other from a psychology research (medical student's academic achievement), varying the sample sizes and the predictive accuracy. With different number of observations and different possible prediction accuracies, it was possible to compare how each visualization technique behaves in each situation. The results pointed that the visualization of random forest's predictive performance was easier and more intuitive to interpret using the weighted network of the proximity matrix than using the multidimensional scale plot. The proportion of misplaced cases was highly related to total accuracy, sensitivity and specificity. This strategy, together with the computation of Zhang and Horvath's (2005) clustering coefficient for weighted graphs, can be very helpful in understanding how well a random forest prediction is doing in terms of classification.

Keywords

Machine Learning, Assessment, Prediction, Visualization, Networks, Cluster

1. Introduction

There is an old Danish phrase that synthetizes the challenge of constructing predictive models in any area of re-

search: It is hard to make predictions, especially about the future (Steincke, 1948). This is an inconvenient facet of the predictive sciences, because a good prediction can improve the quality of our choices and actions, leading to evidence-based decisions. According to Kuhn and Johnson (2013) the main culprits behind the difficulty of making predictions are the inadequate pre-processing of data, the non-appropriate model validation, the naive or incorrect extrapolation of the prediction and the bitter overfitting. In spite of its difficulties, the process of developing predictive models is less painful today than it was in the past, due to the development of algorithms that can lead to high accuracy levels and to the advance of high-speed personal computers to run these algorithms (Kuhn & Johnson, 2013). These algorithms are still subjected to the issues pointed by Kuhn and Johnson (2013), but they present many advantages over the most classical predictive techniques (James, Tibshirani & Friedman, 2009).

Machine learning is the field providing the majority of the predictive algorithms currently applied in a broad set of areas, from systems biology (Geurts, Irrthum, & Wehenkel, 2009) to ADHD diagnosis (Eloyan et al., 2012; Skogli et al., 2013) and education (Blanch & Alucha, 2013; Cortez & Silva, 2008; Golino & Gomes, 2014; Hardman, Paucar-Caceres, & Fielding, 2013). Among the vast number of algorithms available, the classification and regression trees or CART (Breiman, Friedman, Olshen, & Stone, 1984) are some of the most used due to a triplet pointed by Geurts, Irrthum and Wehenkel (2009): interpretability, flexibility and ease of use. The first item on the triplet regards the understandability of the CART results, since it is a roadmap of if-then rules. James, Witten, Hastie and Tibshirani (2013) point the tree models are easier to explain to people than linear regression, since it mirrors the human decision-making more than other predictive models. The second item on the triplet, flexibility, refers to the applicability of the CART to a wide range of problems, handling different types of variables (nominal, ordinal, interval and ratio), with no assumptions regarding normality, linearity, independency, collinearity or homoscedasticity (Geurts, et al., 2009). CART is also more appropriate than the c statistic to study the impact of additional variables to the predictive model (Hastie, Tibshirani, & Friedman, 2009), being especially relevant to the study of incremental validity. Finally, the third item on the triplet, ease of use, refers to the somehow computational facility of implementing the CART algorithm, to the low number of tuning parameters and to the widely available software and packages to apply it.

In spite of the CART qualities it suffers from two issues: overfitting and variance (Geurts et al., 2009). 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. When it becomes sample dependent, in the sense that the partitioning is too suitable to the training set, it will tend to behave poorly in new data sets. The variance issue is exactly a consequence of the overfitting. 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 and regression trees do not have the same predictive accuracy as the other machine learning algorithms (James et al., 2013).

A strategy known as ensemble can be used to deal with overfitting and variance, and to increase the predictive accuracy of classification and regression trees. In the CART context, several trees can be combined (or ensemble) to perform a task based on the prediction made by every single tree (Seni & Elder, 2010). Among the ensemble models using CART, Random Forest (Breiman, 2001) is one of the most widely applied (Seni & Elder, 2010). In the classification scenario, the Random Forest algorithm takes a random subsample of the original data set (with replacement) and of the feature space to grow the trees. The number of the selected features (variables) is smaller than the number of total elements of the feature space. Each tree assigns a single class to the each region of the feature space for every observation. Then, each class of each region of every tree grown 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 trees. As the Random Forest does not use the entire observations (only a subsample of it, usually 2/3), the remaining observations (known as out-of-bag, or OOB) are used to verify the accuracy of 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). The Random Forest algorithm leads to less overfitting than single trees, and is an appropriate technique to decrease the variance in a testing set and to increase the accuracy of the prediction. However, the Random Forest result is not easily understandable as the result of single CARTs. There is no "typical tree" to look at in order to understand the prediction roadmap (or the if-then rules). Therefore, the Random Forest algorithm is known as a "black-box" approach.

Improving Random Forest's Result Interpretability Using Visualization Techniques

In order to make the Random Forest's results more understandable and interpretable, two main approaches can be used: variable importance plots and multidimensional scaling. In the first approach, an importance measure is plotted for each variable used in the prediction. Predictions using random forest usually employ two importance measures: the mean decrease accuracy and the Gini index. The former indicates how much in average the accuracy decreases in the out-of-bag samples when a given variable is excluded from the predictive 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 second approach is a data visualization tool primarily used to identify clusters (Borg & Groenen, 2005; Quach, 2012). Multidimensional scaling has a number of methods, being the most widely applied the classical scaling. The classical scaling enables to compare the n observations visually by taking the first two or three principal components of the data matrix provided. In the random forest context, the matrix can be the proximity measures for the n observations of the dataset. As pointed before, the random forest compute a number of trees using a subsample of the predictors and of the participants. The remaining observations, out-of-bag, are used to verify the accuracy of the prediction. Each tree will lead to a specific prediction, and this prediction can be applied in the total sample. Every time two observations, lets say k and j, are in the same terminal node, they receive a proximity score of one. The proximity measure is the total proximity score between any two observations divided by the total number of trees grown.

In spite of being useful to interpret the result of the random forest algorithm, the variable importance measures are not suitable to indicate the quality of the prediction. Variable importance can help only in the identification of each variable's relevance to the prediction of the outcome. In the multidimensional scaling method, by the other side, the closer the n observations from a class are in a two-dimensional space, and the further from the observations of another class, the highest the quality of the prediction. However, the distance between the points are no direct representation of the algorithm's accuracy performance. It is a representation of the first two or three principal components' score making the interpretability of the cases' distribution not very intuitive. In the current paper we propose an alternative method to improve random forest's result interpretability that provides a more intuitive plot than the multidimensional scaling plot. Our proposition is to represent the proximity matrix as weighted graphs.

Briefly, a graph (G) is defined as a set of vertices (V) connected by a set of edges (E) with a given relation, or association. A graph can be directed or undirected. A direct graph is a special type of graph in which the elements of E are ordered pairs of V. In other words, in a direct graph the edges have directions, usually represented by arrows. An indirect graph is a graph in which the elements of E is an unordered pairs of V. In an indirect graph, the edges have no directions. Finally, a graph can be unweighted or weighted. In the case of an unweighted graph, the edges connecting the vertices have no weights. The edges just indicate if two vertices are connected or not. In the case of weighted graphs (also called networks), every edge is associated to a specific weight that represents the strength of the relation between two vertices. Vertices can represent entities such as variables or participants from a study. The edges, on the other hand, can represent the correlation, covariance, proximity or any other association measure between two entities.

The representation of relationships between variables (e.g. proximity measures) as weighted edges can enable the identification of important structures that could be difficult to identify by applying other techniques (Epskamp, Cramer, Waldorp, Schmittmann, & Borsboom, 2012). In the context of interpreting and understanding a random forest result, the representation of the proximity matrix as a network can facilitate the visual inspection of the prediction accuracy, turning complex patterns in easily understandable images without using any dimension reduction methods. The distance between two nodes in a network became a representation of the nodes' proximity score. Considering the proximity scores as the average number of times two observations were classified in the same terminal node, the distance turns into a direct representation of the random forest's performance. This direct representation of the prediction accuracy makes the interpretability of the graph very straightforward. The closer the observations from a class are from each other, and the further they are from the observations of another class, the highest the quality of the prediction. This interpretation of the weighted graph is the same as the interpretation of the multidimensional scale plot. The difference lies in what constitutes the distance in the two-dimensional space. In the multidimensional scale plot the distance between two points is the difference of the principal components' scores, while in the weighted graph it is the proximity measure itself! So, in the

weighted graph approach, the closer two observations are in the two-dimensional space the higher the average number of times they were classified in the same terminal node in the random forest classification procedure. In the same line, the closer all observations from a single class are from each other, the better the predictive accuracy of the model for that class. Therefore, plotting the proximity measure as weighted networks provides a direct representation of the prediction accuracy, which improves the interpretability of the random forest result. Furthermore, the representation of statistical information using networks are more memorable than using common graphs, such as points, bar, lines, and so on (Borkin et al., 2013). The importance of using memorable graphs lies in the increment of visualization effectiveness and engagement they provide, facilitating the communication of research results.

In the next sections, we describe the analysis of two datasets, one from a breast cancer research (N = 699) and the other from an academic achievement research of college students (N = 77). The use of two datasets is justifiable since it enables the comparison of predictive models under different conditions. The breast cancer dataset is used in a number of technical studies in the machine learning and pattern recognition field (Mangasarian & Wolberg, 1990; Wolberg & Mangasarian, 1990; Mangasarian, Setiono, & Wolberg, 1990; Bennett & Mangasarian, 1992) and provides a relatively large number of observations from a research area in which the predictive accuracy is generally high (breast cancer diagnosis). The academic achievement dataset, by the other hand, presents a few number of observations from a field in which the predictive accuracy is generally low (educational achievement prediction). In addition, both datasets separates their observations in two classes and provides a number of predictors. So, comparing the visualization techniques in both datasets will help to demonstrate the applicability of our method in predictions made under different conditions.

2. Method

2.1. Datasets

Two datasets are used in the current study. The first one came from a study on breast cancer assessing tumors from 699 patients of the University of Wisconsin Hospitals, available at the MASS package (Venables & Ripley, 2002). The second dataset comes from a paper showing the accuracy of different tree-based models in the prediction of Medicine students' academic achievement (Golino & Gomes, 2014). This dataset, called Medical Students hereafter, presents data from 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 dataset contains the score of each participant in 12 psychological/educational tests and scales. Except for those tests evaluating response speed (in milliseconds), the scores in all the other tests were computed as the ability estimate calculated using the Rasch models of the Item Response Theory field. We employ both datasets for three main reasons, as pointed before: 1) they present different sample sizes; 2) there are huge differences in terms of predictive accuracy for each field (breast cancer diagnosis presents higher accuracy than educational achievement prediction, in general); and 3) both deals with classification problems, involving only two classes. With different number of observations and different possible prediction accuracies, it will be possible to compare how each visualization technique behaves in each situation.

2.2. Random Forest Procedure

The random forest will be applied using the random Forest package (Liaw & Wiener, 2012) of the R statistical software (R Development Core Team, 2011). Since our goal is to show a new visualization tool to increase random forest's result interpretability, we do not split each dataset in training and testing sets, as is usually required. Separating the datasets into training and testing would double the number of plots and analysis, significantly increase the number of pages with no direct benefit for our goal. For each dataset, four different models will be computed, each one with a different number of predictors (mtry parameter: 1 or 3) and trees (ntree parameter: 10 or 1000). The proximity matrix will be recorded in each model, for each dataset.

2.3. Visualization Procedures

Two visualization procedures will be used in each one of the four models fitted, for each dataset. The multidimensional scaling plot will be generated using the ladder plot function of the plotrix package (Lemon, 2006), while the weighted graphs will be plotted using the q graph package (Epskamp et al., 2012). The layout of the

weighted graphs will be computed using a modified version of the Fruchterman-Reingold algorithm (Fruchterman & Reingold, 1991). This algorithm computes a graph layout with the edges depending on their absolute weights (Epskamp et al., 2012), i.e. the stronger the relationship between two vertices, the closest in space it is represented in the network (shorter edges for stronger weights). On the other hand, the weakest the connection between two vertices, the further apart they are represented in the space. The *qgraph* package also plots the width of the edges and its color intensity according to its weight. So the highest the weight of an edge, the highest its width and the more intense its color. Finally, a Venn-diagram for each group will be plotted to show the 95% confidence region for each class. A confidence region is a multidimensional generalization of a confidence interval, represented by an ellipsoid containing a set of points, or in our case, nodes from a network.

2.4. Indicators of the Prediction Quality: Total Accuracy, Sensitivity, Specificity, Cluster Coefficients for Weighted Networks and Proportion of Misplaced Cases

The quality of the random forest prediction can be checked using total accuracy, sensitivity and specificity. Total accuracy is the proportion of observations correctly classified by the predictive model. It is an important indicator of the general prediction's quality, but is not a very informative measure of the errors in each class. For example, a general accuracy of 80% can represent an error-free prediction for the C_1 class, and an error of 40% for the C_2 class. In order to verify within class errors, two other measures can be used: sensitivity and specificity. Sensitivity is the rate of observations correctly classified in a target class, e.g. C_1 = low achievement, over the number of C_1 observations. Specificity, on the other hand, is the rate of correctly classified observations of the non-target class, e.g. C_1 = high achievement, over the number of C_2 observations.

In the current paper we introduce a new index of the random forest's prediction quality, named proportion of misplaced cases (PMC). It can be calculated as follows. In a weighted graph it is possible to represent a 95% confidence region of the nodes for each class using Venn-diagrams. The PMC index is the sum of cases (or observations) from a C_k class that is within the 95% confidence region of the C_w class (represented below as CR_w) plus the sum of cases from the C_w that is within the 95% confidence region of the C_k class (represented below as CR_k), divided by the sample size. The PMC index can be mathematically written as:

$$NMC = \frac{1}{n}\sum\left[\left(y_i \in C_k\right) \subset CR_w\right] + \sum\left[\left(y_i \in C_w\right) \subset CR_k\right] \quad (1)$$

In order to gather additional evidences of the prediction quality, three clustering coefficients for weighted networks will be computed. One of the first clustering measures for weighted networks was proposed by Barrat et al. (2004), combining topological information of a network with its weights and considering only the adjacent weights of a node (Kalna & Higham, 2007). The Barrat et al. (2004) clustering coefficient has an interesting property discovered by Antoniou and Tsompa (2008): the clustering measure is independent of the size of the network and of the weights' distribution, both for a fully connected network and for a not-fully connected one. Onnela et al. (2005) proposed the second weighted clustering coefficient used in the present paper and, contrarily to Barrat et al.'s (2004) measure, take into account the weights of all edges (Kalna & Higham, 2007). The third weighted clustering coefficient to be employed was proposed by Zhang and Horvath (2005), and is the only relying exclusively on the network weights (Kalna & Higham, 2007). Not surprisingly, both Onnela et al. (2005) and Zhang and Horvath's (2005) coefficients are almost linearly related to the network weights, i.e. the clustering values increases when the weights' values increases (Antoniou & Tsompa, 2008). This finding may suggest that both coefficients are more informative of the random forest's classification performance than Barrat et al.'s (2004) weighted clustering coefficient. The cluster coefficients were calculated using the *qgraph* package (Epskamp et al., 2012). The Wilcoxon Sum Rank test will be calculated to verify if the distribution of the clustering coefficients for each class, in every model from both datasets, were identical or not. In order to verify which clustering coefficient better separated the classes, the Hodges-Lehmann estimator will be employed, since it estimates the median of the difference between the clustering coefficients from a sample of each class. So, the cluster technique presenting the highest Hodges-Lehmann estimator can be considered the best index of class separation.

3. Results

The results are divided into three parts. The first one presents the random forest result and the plots of the breast

cancer dataset. The second part shows the random forest result and the plots of the medical students dataset. The third part shows the comparison of total accuracy, sensitivity, specificity, percentage of misplaced cases and clustering coefficients for each one of the four models computed for each dataset.

3.1. Breast Cancer Results

The result of random forest's model 1, using one variable at each split (i.e. mtry = 1) and ensemble 10 trees to classify benign and malignant breast tumors, shows a total accuracy of 95.08%. The sensitivity of model 1 is 93% and the specificity is 96%. Adding two more predictors at each split (mtry = 3) and holding the number of trees as 10 slightly increases the total accuracy (96%), the sensitivity (94%) and the specificity (97%) in model 2. The same occurs in model 3, which increases the number of trees to 1000 and uses only one variable at each split, resulting in a total accuracy of 97%, in a sensitivity of 98% and in a specificity of 97%. The fourth model held the same number of trees used in model 3 and increases the number of variables used at each split to three. It results in a total accuracy of 97%, in a sensitivity of 96% and in a specificity of 97%. This result can be checked in **Table 1**.

The multidimensional scale plot (**Figure 1**) shows quite nicely the separation of the benign (red dots) and the malignant (blue dots) classes. It is possible to see that people with the same class membership are mostly classified in the same terminal nodes. It occurred because the total accuracy, sensitivity and specificity were high and led to a correct prediction of the out-of-bag samples most of the time. As a result, the proximity scores were higher for people with the same class membership, and lower for people with different class membership. When the proximity matrix is plotted using the multidimensional scale plot, it is really easy to visually discriminate two clusters, one composed by the people with benign tumors and the other composed by people with malignant tumors.

In spite of being a powerful tool to visualize the random forest's prediction performance, the interpretation of the cases' distribution in each class in the multidimensional scale plot is not straightforward. The multidimensional scale plot first reduces the number of dimensions using principal component analysis of the proximity matrix, and then plot the first two component scores in a two-dimensional plane. The plot generated using the weighted network technique (**Figure 2**) is much simpler to understand than the multidimensional scale plot. It is easy to verify the separation of the classes, but is also easy to understand the distribution of the cases in each class. The closer two nodes are in the two dimensional space, the higher their proximity score! Following the same line of reasoning, the further two nodes are from each other, the lower their proximity score. Since the proximity scores indicates the average number of times two observations were classified in the same terminal node, the nodes' distance in the weighted network is a direct representation of the random forest's performance. Furthermore, no dimension reduction is required, and that is what makes the weighted network representation easier to understand and interpret than the multidimensional scale plot.

Table 1. Datasets, models, sample size (N), number of trees (ntree), number of predictors (mtry), total accuracy, sensitivity, specificity, proportion of misplaced cases (PMC) and mean clustering coefficients for the entire sample, for the target class only and for the non-target class only.

Dataset	Models	N	ntree	mtry	Total Accuracy	Sensitivity	Specificity	PMC	Zhang Cluster	Onnela Cluster	Barrat Cluster	Zhang Cluster Target	Onnela Cluster Target	Barrat Cluster Target	Zhang Cluster Non-Target	Onnela Cluster Non-Target	Barrat Cluster Non-Target
Breast Cancer	1	699	10	1	0.95	0.93	0.96	1.57%	0.60	0.56	0.72	0.67	0.63	0.77	0.46	0.42	0.63
	2	699	10	3	0.96	0.94	0.97	1.14%	0.65	0.62	0.76	0.72	0.69	0.79	0.53	0.48	0.68
	3	699	1000	1	0.97	0.98	0.97	0.57%	0.51	0.27	0.93	0.67	0.36	0.97	0.22	0.09	0.85
	4	699	1000	3	0.97	0.96	0.97	0.86%	0.66	0.47	0.94	0.81	0.62	0.97	0.37	0.19	0.88
Medical Students	1	77	10	1	0.70	0.60	0.79	37.66%	0.32	0.31	0.47	0.33	0.33	0.50	0.30	0.28	0.45
	2	77	10	3	0.67	0.62	0.71	33.77%	0.40	0.39	0.57	0.40	0.40	0.58	0.40	0.38	0.57
	3	77	1000	1	0.70	0.63	0.76	16.88%	0.17	0.08	0.91	0.18	0.08	0.91	0.17	0.08	0.92
	4	77	1000	3	0.74	0.66	0.81	16.88%	0.20	0.10	0.92	0.20	0.09	0.91	0.19	0.10	0.93

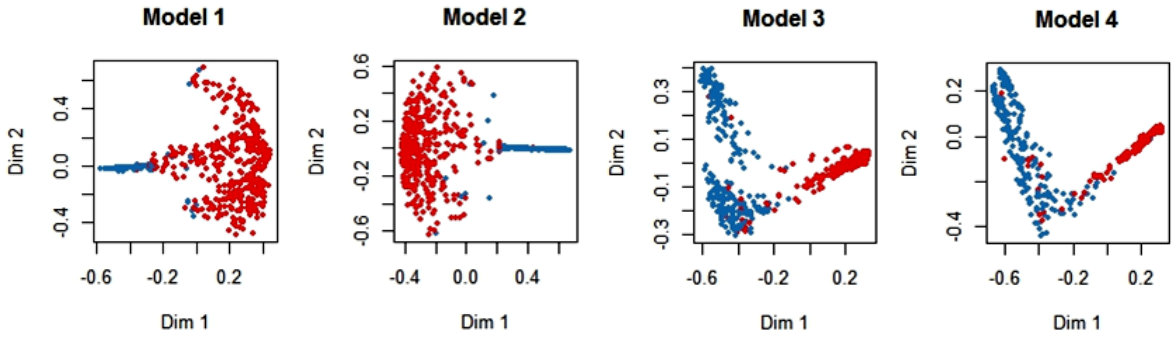


Figure 1. Multisimensional scale plot of the breast cancer's random forest proximity matrix.

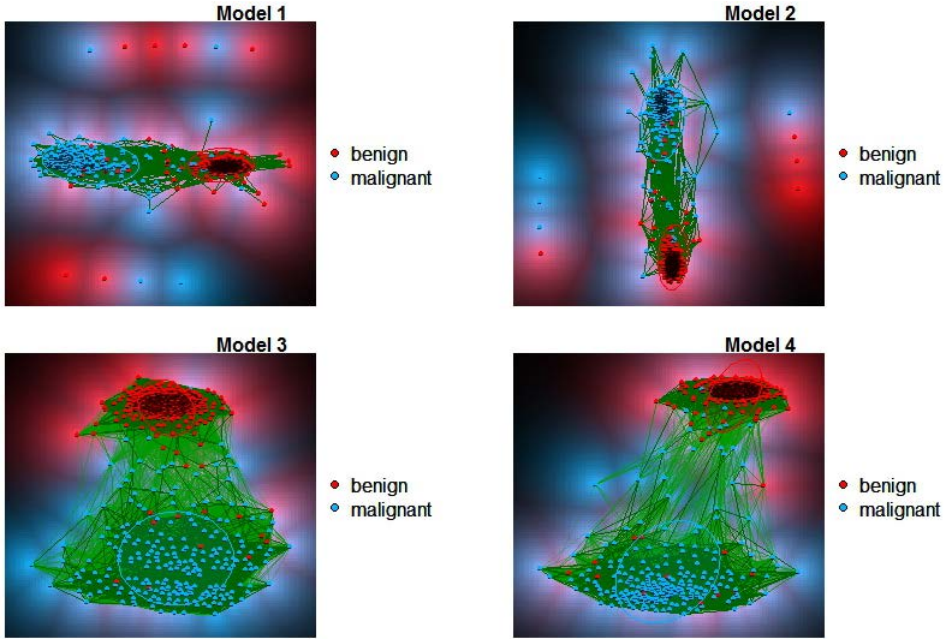


Figure 2. Weighted network representation of the breast cancer's random forest proximity matrix.

The Venn-diagram shows the 95% confidence region for each class. There is no interpolation of the confidence region in all models, suggesting a very good separation between the benign and malignant classes. The proportion of misplaced cases (PMC), i.e. the number of cases from the benign class within the 95% confidence region of the malignant class plus the number of cases from the malignant class within the 95% confidence region of the benign class divided by n, was as follows. The PMC in model one is 1.57%, in model two is 1.14%, 0.57% in model three and 0.86% in model four. These values are also displayed in **Table 1**. The PMC presents a correlation of −0.96 with total accuracy, −0.97 with sensitivity and −0.83 with specificity.

The results of the weighted clustering coefficients for the breast cancer dataset are displayed in **Table 1**. Considering the values of the entire sample, the Zhang and Horvath's (2005) coefficient mean value ranges from 0.51 (model three) to 0.66 (model four), while Onnela et al.'s (2005) varies from 0.27 (model three) to 0.62 (model two). The coefficient of Barrat et al. (2004) ranges from 0.72 (model one) to 0.94 (model four). Considering the cluster coefficient mean value for the observations belonging to the target class only (malignant tumor), the Zhang and Horvath's (2005) coefficient ranges from 0.67 (models one and three) to 0.81 (model four). The Onnela et al.'s (2005) cluster coefficient mean value varies from 0.36 (model three) to 0.69 (model two) and the Barrat et al.'s (2004) mean coefficient ranges from 0.77 (model one) to 0.97 (models three and four). Taking into account only the mean cluster coefficient for the observations belonging to the non-target class (benign tumor), the Zhang and Horvarth's coefficient varies from 0.22 (model three) to 0.53 (model two). The Onnela et al.'s (2005) mean value varies from 0.09 (model three) to 0.48 (model two) and the Barrat et al.'s (2004) mean

coefficient ranges from 0.63 (model one) to 0.88 (model four).

The density distribution of the weighted clustering coefficients for each class is displayed in **Figure 3**. The distribution of the Zhang and Horvarth's (2005) weighted clustering coefficients of the benign and malignant classes are non-identical at the 0.05 significance level in model one (W = 98001.5, *p*-value < 0.001, Hodges-Lehmann = 0.24), in model two (W = 100,447, *p*-value < 0.001, Hodges-Lehmann = 0.19), in model three (W = 108,013, *p*-value < 0.001, Hodges-Lehmann = 0.47) and in model four (W = 107,791, *p*-value < 0.001, Hodges-Lehmann = 0.44).

The same occurs to the distribution of Onnela et al.'s (2005) weighted clustering coefficient of both classes, being non-identical in all four models ($W_{Model\ 1}$ = 92,536.5, $p\text{-value}_{Model\ 1}$ < 0.001, $Hodges\text{-}Lehmann_{Model\ 1}$ = 0.23; $W_{Model\ 2}$ = 96,000.5, $p\text{-value}_{Model\ 2}$ < 0.001, $Hodges\text{-}Lehmann_{Model\ 2}$ = 0.22; $W_{Model\ 3}$ = 107,907, $p\text{-value}_{Model\ 3}$ < 0.001, $Hodges\text{-}Lehmann_{Model\ 3}$ = 0.30; $W_{Model\ 4}$ = 106,024, $p\text{-value}_{Model\ 4}$ < 0.001, $Hodges\text{-}Lehmann_{Model\ 4}$ = 0.45). Finally, the distribution of Barrat et al.'s (2004) cluster coefficient of both classes are non-identical for every model ($W_{Model\ 1}$ = 85,034.5, $p\text{-value}_{Model\ 1}$ < 0.001, $Hodges\text{-}Lehmann_{Model\ 1}$ = 0.17; $W_{Model\ 2}$ = 84,151, $p\text{-value}_{Model\ 2}$ < 0.001, $Hodges\text{-}Lehmann_{Model\ 2}$ = 0.10; $W_{Model\ 3}$ = 104,920, $p\text{-value}_{Model\ 3}$ < 0.001, $Hodges\text{-}Lehmann_{Model\ 3}$ = 0.14; $W_{Model\ 4}$ = 99584, $p\text{-value}_{Model\ 4}$ < 0.001, $Hodges\text{-}Lehmann_{Model\ 4}$ = 0.03).

In our results, the Hodges-Lehmann estimator gives the median of the difference between the clustering coefficients from a sample of the benign class and from a sample of the malignant class. By the Hodges-Lehmann estimator, the best weighted clustering coefficient separating the benign and the malignant classes of the breast cancer dataset is the Zhang cluster from model four.

3.2. Medical Students' Results

As happened with the breast cancer results, the multidimensional scale plots (**Figure 4**) provides a clear way of inspecting clusters and patterns than the parallel coordinate plots. The cases' distribution in model two, for example, seems to be a little bit more heterogeneous between classes than in model three or four, which does re-

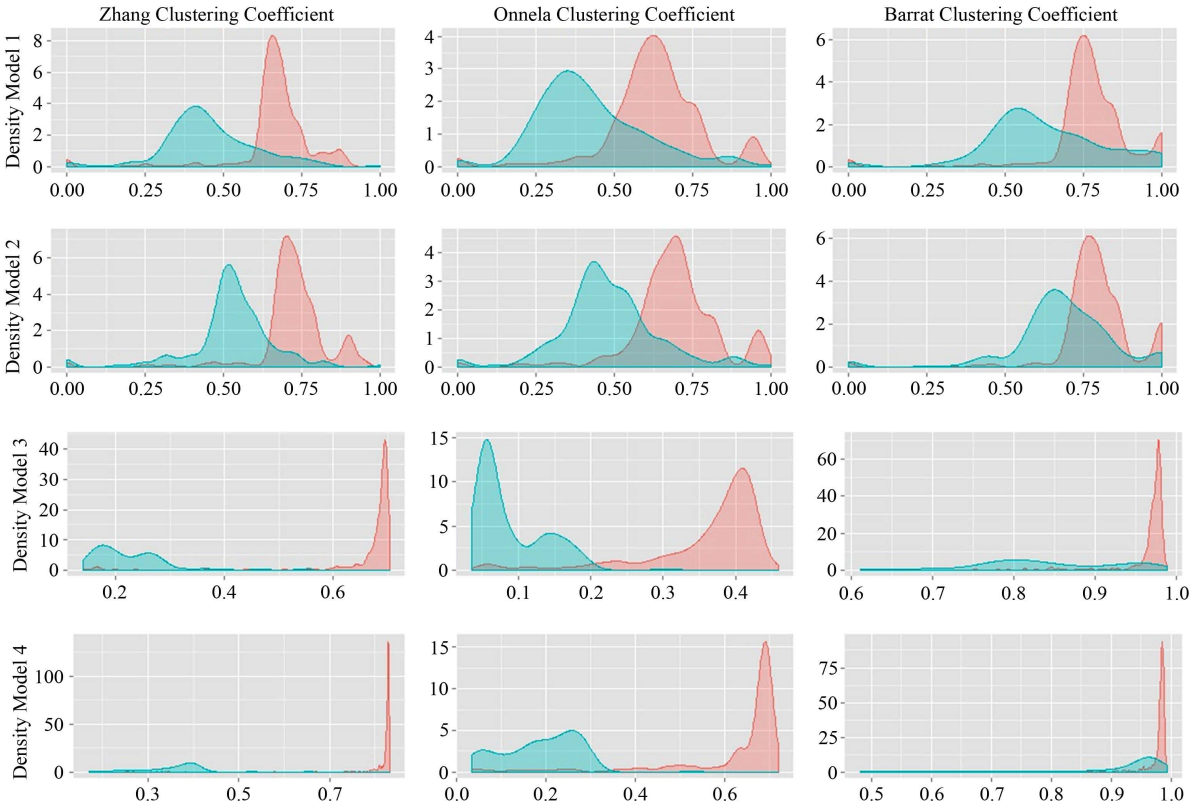


Figure 3. Density distribution of three weighted clustering coefficients (Zhang, Onnela and Barrat) of the malignant class (blue) and the benign class (red), from models 1 to 4.

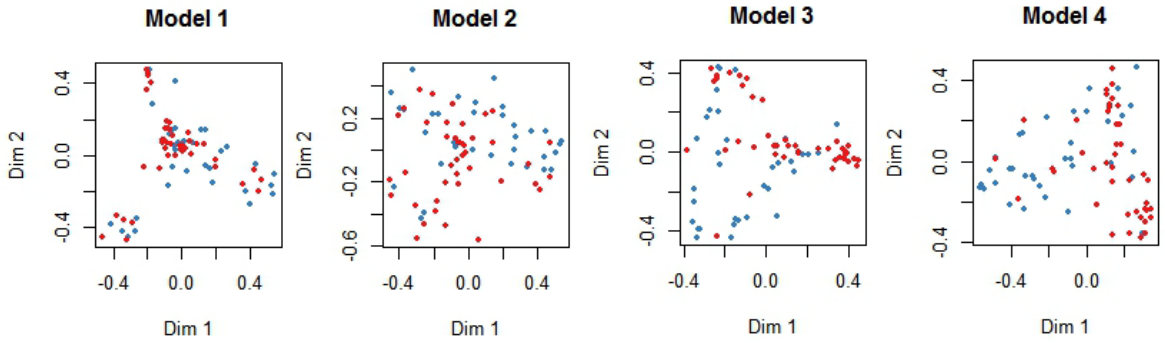


Figure 4. Multidimensional scale plots of the medical students' random forest proximity matrix.

flects the performance of the models. However, due to the accuracy of the models, only by consulting the accuracy indexes is possible to tell which model is the best one. Model one presents a total accuracy of 70.13%, a sensitivity of 60% and a specificity of 78.58%. Model two shows a worse performance than model one, with a total accuracy of 67.11%, a sensitivity of 61.77% and a specificity of 71.53%. The third model presents a total accuracy of 70.13%, a sensitivity of 62.86% and a specificity of 76.20%, while the fourth model shows a total accuracy of 74.03%, a sensitivity of 65.72% and a specificity of 80.96%.

Again, the plot generated using the weighted network technique (**Figure 5**) offers a clearer picture than the multidimensional scale plots. It is easy to verify the separation of the classes, to understand the distribution of the cases in each class, as well as to compare the prediction performance between the models. The 95% confidence regions is more superposed in models one and two than in models three and four. The *proportion of misplaced cases* is equal to 37.66% in model one, 33.76% in model two and 16.88% in models three and four.

Table 1 shows the results of the weighted clustering coefficients for the medical students' dataset. The Zhang and Horvath's (2005) mean clustering coefficients of the entire sample ranges from 0.17 (model three) to .40 (model two), while the Onnela et al.'s (2005) varies from 0.27 (model three) to 0.39 (model two). The coefficient of Barrat et al. (2004) ranges from 0.47 (model one) to 0.92 (model four). Taking into consideration the cluster coefficient mean value for the observations belonging to the target class only (low achievement), the Zhang and Horvath's (2005) ranges from 0.18 (model three) to 0.40 (model two). The Onnela et al.'s (2005) mean coefficient varies from 0.50 (model one) to 0.91 (model four) and the Barrat et al.'s (2004) from 0.77 (model one) to 0.97 (models three and four). Finally, considering only the mean cluster coefficient for the observations belonging to the non-target class (high achievment), the Zhang and Horvarth's varies from 0.17 (model three) to 0.40 (model two). The Onnela et al.'s (2005) varies from 0.08 (model three) to 0.38 (model two) and the Barrat et al.'s (2004) ranged from 0.45 (model one) to 0.93 (model four).

The density distribution of the weighted clustering coefficients for each class of the medical students dataset is displayed in **Figure 6**. Only the distribution of the Barrat et al.'s (2005) weighted clustering coefficient of the low and high achievement classes are non-identical at the 0.05 significance level ($W_{Model\ 3} = 495$, $p\text{-value}_{Model\ 3} < 0.05$, $\text{Hodges-Lehmann}_{Model\ 3} = -0.01$; $W_{Model\ 4} = 466$, $p\text{-value}_{Model\ 4} < 0.01$, $\text{Hodges-Lehmann}_{Model\ 4} = -0.02$). The Barrat et al.'s (2004) coefficient from model four was the best to separate the low and the high achievement classes of the medical students' dataset according to the Hodges-Lehmann estimator.

3.3. Relating the Indicators of the Prediction Quality

Figure 7 shows the correlation pattern (left) and the *p*-values (right) of the following variables: total accuracy (T.A), sensitivity (Sns), specificity (Spc), sample size (N), number of trees (ntr), number of predictors (mtr), Zhang cluster (Zh.C), Zhang cluster of the target class (Z.C.T), Zhang cluster of the non-target class (Z.C.N), Onnela cluster (On.C), Onnela cluster of the target class (O.C.T), Onnela cluster of the non-target class (O.C.N), Barrat cluster (Br.C), Barrat cluster of the target class (B.C.T), Barrat cluster of the non-target class (B.C.N) and proportion of misplaced cases (NMC). The correlations were plot using the *qgraph* package (Epskamp et al., 2012) and the layout was computed through the Fruchterman-Reingold algorithm. The higher the correlations, the closer the variables in the two-dimensional space. Positive correlations are represented by green edges, and negative correlations by red edges. The *p*-values are also plotted, with different intensity of blue representing different levels of significance.

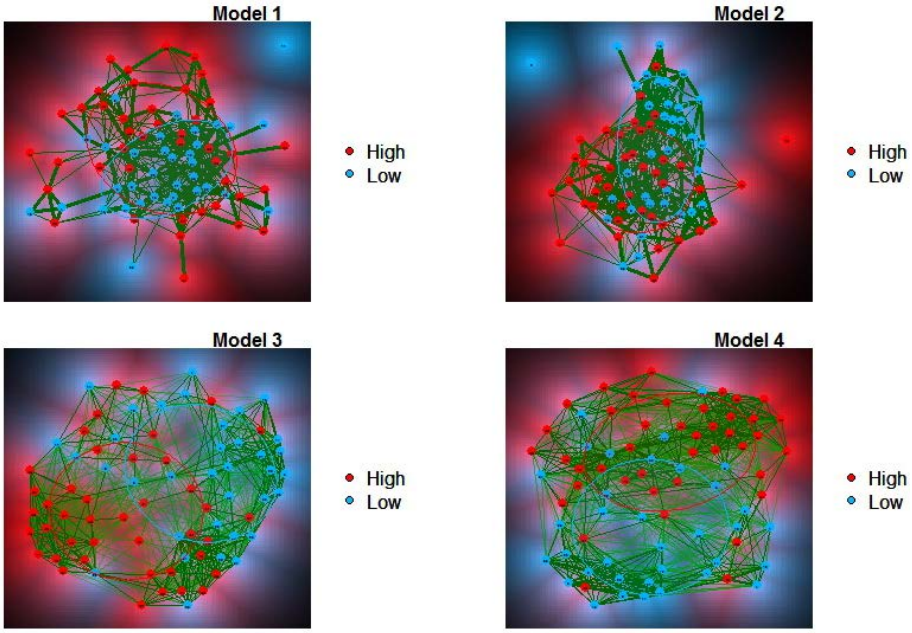


Figure 5. Weighted network representation of the medical students' random forest proximity matrix.

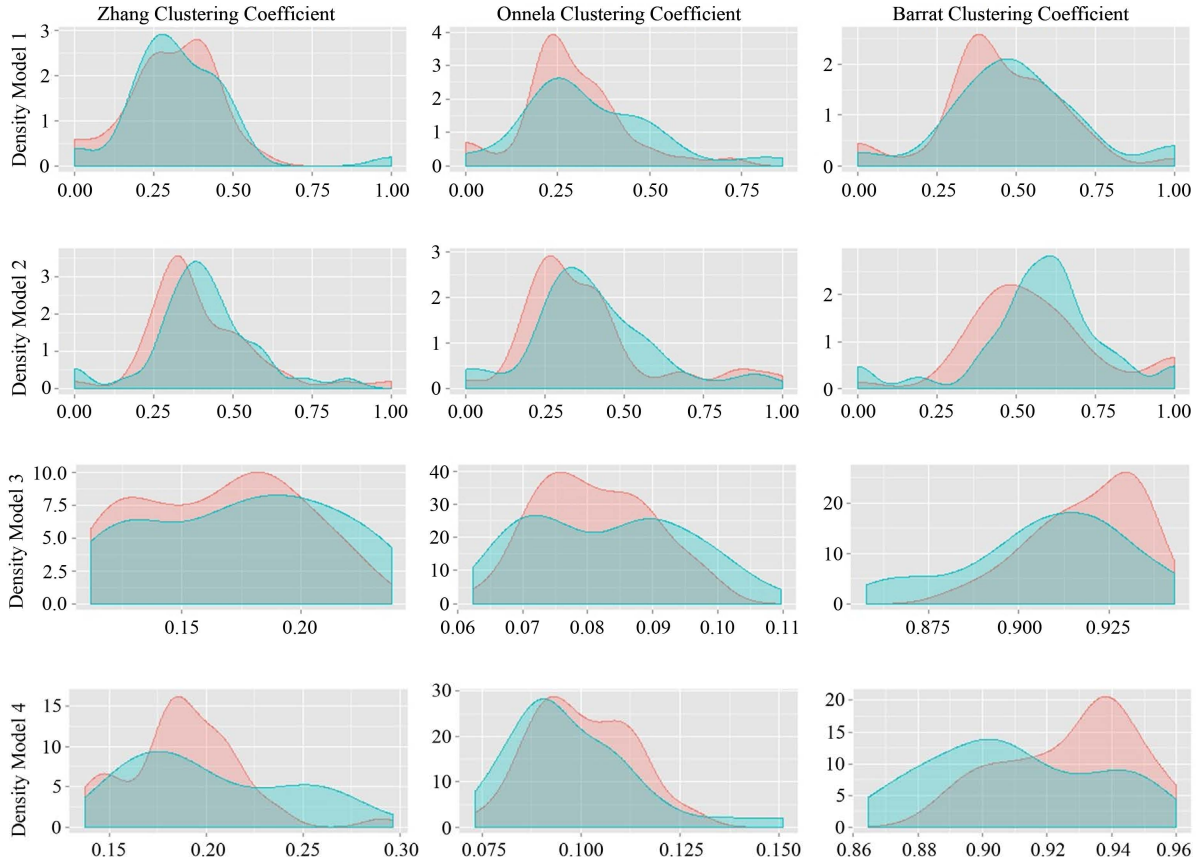


Figure 6. Density distribution of three weighted clustering coefficients (Zhang, Onnela and Barrat) of the low achievement class (blue) and the high achievement class (red), from models 1 to 4.

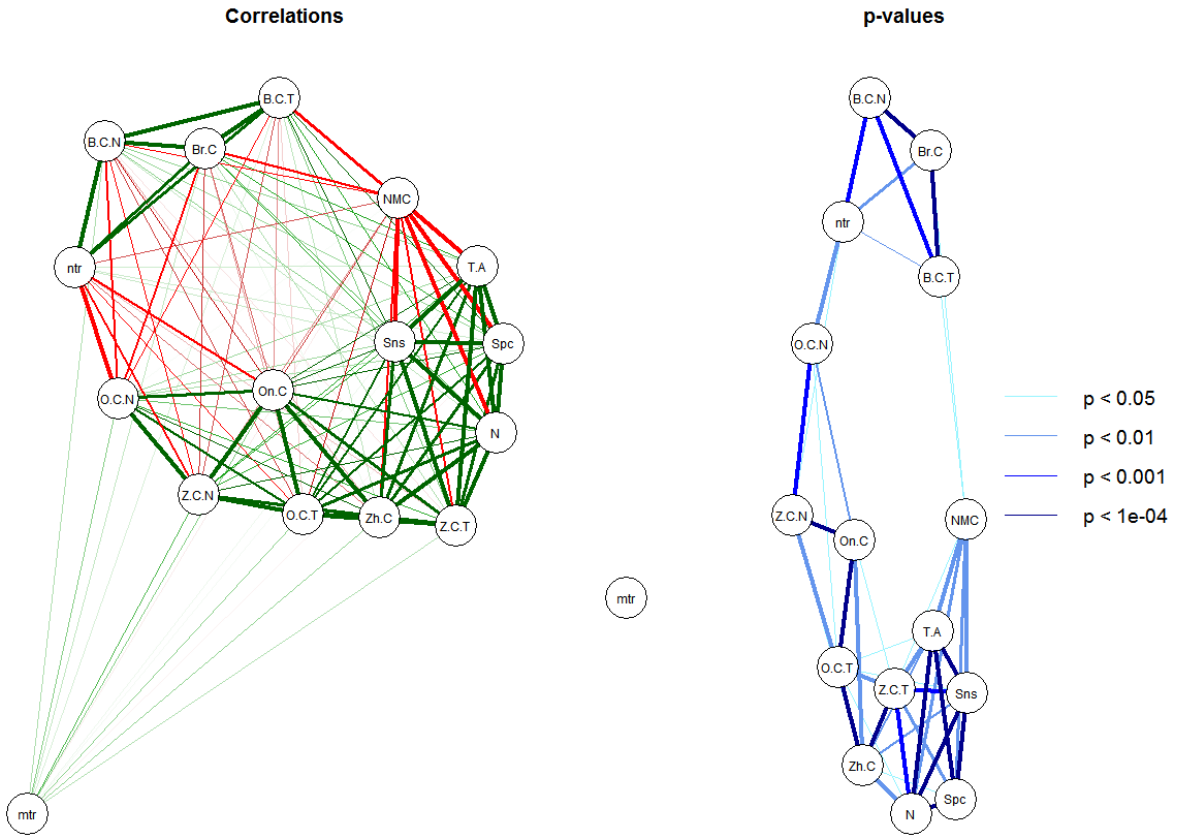


Figure 7. Plot of the quality indicators' correlation matrix.

The number of predictors are poor and non-significantly related to any other indicators. The correlation between the number of trees and the mean Barrat et al.'s (2004) clustering coefficient for the entire sample, for the target class and for the non-target class are high and significant. When the number of trees in the random forest increases, the weighted clustering coefficient of Barrat et al. (2004) also increases. The Barrat et al. (2004) coefficient is not significantly related to the total accuracy, sensitivity and specificity of the predictive model, but has a negative and significant relation to the proportion of misplaced cases. By its turn, the proportion of misplaced cases has a strong, negative and significant correlation with total accuracy, sensibility, specificity and sample size. The Zhang and Horvath's (2005) weighted clustering coefficient for the entire sample and for the target class has a strong, positive and significant correlation with total accuracy, sensitivity, specificity and sample size. However, the Zhang and Horvath's (2005) coefficient for the non-target class is not significantly related to total accuracy, sensitivity or specificity. Finally, Onnela et al.'s (2004) coefficient is not significantly related to the accuracy indexes, but its mean clustering value for the non-target class is strong, negative and significantly related to the number of trees. The correlations are displayed in **Table 2**.

4. Discussion

In the last decades, the development of algorithms that can lead to high predictive accuracy and the advance of high-speed personal computers to run these algorithms led the data science to a completely new level (Kuhn & Johnson, 2013). Machine learning is the field providing the majority of the predictive algorithms currently applied in a broad set of areas (Geurts, Irrthum, & Wehenkel, 2009; Eloyan et al., 2012; Skogli et al., 2013; Blanch & Alucha, 2013; Cortez & Silva, 2008; Golino & Gomes, 2014; Hardman, Paucar-Caceres, & Fielding, 2013). Among the techniques and algorithms available, the classification and regression trees or CART (Breiman, Friedman, Olshen, & Stone, 1984) is amongst the most used ones (Geurts, Irrthum, & Wehenkel, 2009). However, in spite of the CART qualities it suffers from the two related issues of overfitting and variance (Geurts et al., 2009). Ensemble tree techniques were developed to deal with both issues and to increase the predictive ac-

Table 2. Correlation between sample size (N), number of trees (ntree), number of predictors (mtry), total accuracy, sensitivity, specificity, proportion of misplaced cases (PMC) and mean clustering coefficients for the entire sample, for the target class only and for the non-target class only.

	N	ntree	mtry	Total Acc.	Sen.	Spe.	Zh.C	On.C	Br.C	Z.C.T	O.C.T	B.C.T	Z.C.N	O.C.N	B.C.N	PMC
N	1	0.00	0.00	0.99	0.99	0.97	0.91	0.7	0.35	0.95	0.79	0.45	0.53	0.28	0.13	−0.88
ntree	0.00	1	0.00	0.10	0.11	0.10	−0.29	−0.65	0.87	−0.14	−0.51	0.84	−0.75	−0.92	0.92	−0.34
mtry	0.00	0.00	1	0.02	0.03	−0.02	0.21	0.24	0.12	0.15	0.23	0.08	0.34	0.23	0.16	−0.04
Total Acc.	0.99	0.10	0.02	1	0.99	0.99	0.86	0.63	0.44	0.91	0.73	0.53	0.45	0.19	0.22	−0.91
Sen.	0.99	0.11	0.03	0.99	1	0.97	0.87	0.62	0.45	0.92	0.73	0.54	0.44	0.19	0.23	−0.91
Spe.	0.97	0.10	−0.02	0.99	0.97	1	0.82	0.59	0.42	0.87	0.69	0.51	0.41	0.17	0.20	−0.90
Zh.C	0.91	−0.29	0.21	0.86	0.87	0.82	1	0.91	0.04	0.98	0.96	0.13	0.78	0.56	−0.18	−0.65
On.C	0.70	−0.65	0.24	0.63	0.62	0.59	0.91	1	−0.31	0.82	0.98	−0.23	0.97	0.86	−0.49	−0.39
Br.C	0.35	0.87	0.12	0.44	0.45	0.42	0.04	−0.31	1	0.16	-0.18	0.99	−0.41	−0.63	0.97	−0.71
Z.C.T	0.95	−0.14	0.15	0.91	0.92	0.87	0.98	0.82	0.16	1	0.9	0.26	0.66	0.41	−0.06	−0.72
O.C.T	0.79	−0.51	0.23	0.73	0.73	0.69	0.96	0.98	−0.18	0.9	1	−0.09	0.91	0.75	−0.38	−0.49
B.C.T	0.45	0.84	0.08	0.53	0.54	0.51	0.13	−0.23	0.99	0.26	−0.09	1	−0.35	−0.59	0.94	−0.78
Z.C.N	0.53	−0.75	0.34	0.45	0.44	0.41	0.78	0.97	−0.41	0.66	0.91	−0.35	1	0.94	−0.55	−0.24
O.C.N	0.28	−0.92	0.23	0.19	0.19	0.17	0.56	0.86	−0.63	0.41	0.75	−0.59	0.94	1	−0.73	0.01
B.C.N	0.13	0.92	0.16	0.22	0.23	0.20	−0.18	−0.49	0.97	−0.06	−0.38	0.94	−0.55	−0.73	1	−0.53
PMC	−0.88	−0.34	−0.04	−0.91	−0.91	−0.90	−0.65	−0.39	−0.71	−0.72	−0.49	−0.78	−0.24	0.01	−0.53	1

curacy of classification and regression trees, being the random forest (Breiman, 2001) one of the most widely applied (Seni & Elder, 2010). In spite of the random forest's useful characteristics, its result is not easily understandable as the result of single CARTs. There is no "typical tree" to look at in order to understand the prediction roadmap (or the if-then rules). In order to increase random forests understandability and interpretability variance importance plots and multidimensional scaling plots can be used. The first strategy will unveil how important each predictor is in the classification, given all trees generated. It is of low to no use if one is interested in assessing the quality of the prediction. The second strategy is well suited to visually inspect how well the prediction is behaving, in the sense of separating two or more classes. However, the interpretation of the multidimensional scaling plot is not straightforward, since it plots the first two or three principal components' score of the proximity matrix. Multidimensional scaling plot does not generate a direct representation of the models' predictive accuracy, and thus makes the interpretability of the cases' distribution not very intuitive.

When dealing with predictive models, visualization techniques can be a very helpful tool to learn more about how the algorithms operate (Borg & Groenen, 2005; Wickham, Caragea, & Cook, 2006), to understand the results (Quack, 2012; Wickham, 2006) and to seek patterns (Honarkhah & Caers, 2010). It is also very informative to apply visualization techniques to compare predictions under different conditions by looking how groups are arranged in the space (specially their boundaries) (Wickham, Caragea, & Cook, 2006). Recent researches have shown that the kind of plot used in a research interferes in its comprehension and memorability (Borkin et al., 2013). Borkin and colleagues (2013) showed that representing statistical information using networks are more memorable than using common graphs, such as points, bar, lines, and so on (Borkin et al., 2013). The relevance producing memorable graphs lies in its capacity to increment the visualization effectiveness and engagement, therefore facilitating the results communication.

In the current paper we proposed the representation of the proximity matrix as weighted networks, a new method to improve random forest's result interpretability. The use of weighted networks as a representation of the relationships between variables enables the identification of important structures that could be difficult to iden-

tify using other techniques (Epskamp et al., 2012). This approach is a direct representation of the random forest's performance, which facilitates the visual inspection of the predictive accuracy, turning complex patterns into easily understandable images without using any dimension reduction methods.

Our results suggest that representing the proximity matrix as networks indeed facilitates the visualization of the random forest classification performance, especially in cases where the accuracy is not very high. The approach we propose also seems to be more intuitive than the multidimensional scaling plot, especially when comparing model performance under different conditions. In the breast cancer dataset, **Figure 2** shows that while in models one and two the observations are closer in the space than in models three and for, indicating higher proximity scores, the number of misplaced cases is also higher. This can be seen looking at the colors of the observations within the 95% confidence region for each class. So, it is pretty straightforward to compare the four predictive models, and to conclude that models three and four are better than models one and two. The same analysis is impossible to do in **Figure 1**, which shows the multidimensional scaling plot. There is no visual clue to determine which model is the best one, in terms of total accuracy, sensibility and specificity. Furthermore, when analyzing the new performance index, proportion of misplaced cases, we can confirm our visual analysis. While model one and two have 1.57% and 1.14% of misplaced cases, models three and four have only 0.57% and 0.86%, respectively. The proportion of misplaced cases presents a very high negative correlation with total accuracy (−0.96) and sensitivity (−0.97), indicating the lower the number of misplaced cases, the higher both indexes.

The same interpretation holds for the medical students' dataset. While is very difficult to tell which predictive model is the best one by inspecting the multidimensional scale plot in **Figure 4**, it is easier to do so by inspecting the network in **Figure 5**. It is very easy to see that the number of misplaced cases is higher in models one and two, than in models three and four. Again, this can be done by looking at the colors of the observations within the 95% confidence region for each class. The visual inspection is confirmed by the performance indexes, since models three and four have higher sensitivity (63% and 66%), and model four have higher total accuracy (74%) and specificity (81%) than the other models. The proportion of misplaced cases differentiates only models one and two from models three and four.

Since we are dealing with weighted networks, computing clustering coefficients for this kind of graph can also be informative. Barrat et al. (2004) presented the first clustering coefficient developed specifically for weighted graphs, and has the nice property of being independent of the size of the network and of the weights' distribution (Antoniou & Tsompa, 2008). Contrarily to Barrat's coefficient, the second clustering measure adopted in this paper, proposed by Onnela et al. (2005) take into account the weights of all edges. The only clustering coefficient adopted here that relies exclusively on the network weights is the one proposed by Zhang & Horvath (2005). Since both Onnela et al. (2005) and Zhang and Horvath's (2005) coefficients are almost linearly related to the network weights (Antoniou & Tsompa, 2008), we expected that these two clustering measures would be more informative of the random forest's classification performance than Barrat et al.'s (2004) coefficient.

Our findings suggest that Barrat et al.'s clustering measure is sensitive to the number of trees grown in the random forest procedure, since it presented a strong and significant correlation with the number of trees used ($r = 0.87$, $p < 0.001$). The same occurs when the Barrat et al.'s coefficient is used to the measure the clustering of the target classes ($r = 0.84$, $p < 0.01$) and of the non-target classes ($r = 0.92$, $p < 0.001$). Barrat et al.'s (2005) coefficient is not significantly related to the total accuracy, sensitivity and specificity of the predictive models, and cannot be regarded as an index of the prediction quality. However, it presents a negative and significant relation to the proportion of misplaced cases ($r = -0.71$, $p < 0.05$), and can be considered an indicator of the misplacement within the 95% confidence region of the classes: the higher the Barrat value, the lower the proportion of misplacements.

By the other side, Zhang and Horvath's (2005) weighted clustering coefficient for the entire sample presents a strong, positive and significant correlation with total accuracy ($r = 0.86$, $p < 0.01$), sensitivity ($r = 0.87$, $p < 0.05$) and specificity ($r = 0.82$, $p < 0.05$). It also presents a high and positive correlation with sample size ($r = 0.91$, $p < 0.001$). The same scenario occurs for the Zhang and Horvath's (2005) target class coefficient, since it presented a positive and significant correlation with total accuracy ($r = 0.91$, $p < 0.001$), sensitivity ($r = 0.92$, $p < 0.001$), specificity ($r = 0.87$, $p < 0.01$) and sample size ($r = 0.97$, $p < 0.001$). As predicted, Zhang and Horvath's (2005) coefficient can be considered a good indicator of the prediction quality.

However, unlike we expected, Onnela et al.'s (2004) coefficient was not significantly related to the accuracy

indexes (see **Table 2**). Its mean clustering value for the non-target class was strong, negative and significantly related to the number of trees ($r = -0.92$, $p < 0.01$). On the other hand, the indicator we propose, proportion of misplaced cases, have a strong, negative and significant correlation with total accuracy ($r = -0.91$, $p < 0.01$), sensibility ($r = -0.91$, $p < 0.01$), specificity ($r = -0.90$, $p < 0.01$). It seems that this new index is a good indicator of the random forest prediction accuracy, since the lower the proportion of misplaced cases within the 95% confidence region in each class, the higher the total accuracy, sensibility and specificity.

5. Conclusion

In sum, the current paper proposed a new visualization tool to help check the quality of the random forest prediction, as well as introduced a new accuracy index (proportion of misplaced cases) that was highly related to total accuracy, sensitivity and specificity. This strategy, together with the computation of Zhang and Horvath's (2005) clustering coefficient for weighted graphs, can be very helpful in understanding how well a random forest prediction is doing in terms of classification. It adds new tools to help in the accuracy-interpretability trade-off.

Acknowledgements

The current research was financially supported by a grant provided by the Fundação de Amparo à Pesquisa de Minas Gerais (FAPEMIG) to the two authors, as well as by the INEP foundation. The second author is also funded by the CNPQ council.

References

Antoniou, I. E., & Tsompa, E. T. (2008). Statistical Analysis of Weighted Networks. *Discrete Dynamics in Nature and Society, 2008,* 16. http://dx.doi.org/10.1155/2008/375452

Barrat, A., Barthelemy, M., Pastor-Satorras, R., & Vespignani, A. (2004). The Architecture of Complex Weighted Networks. *Proceedings of the National Academy of Sciences of the United States of America, 101,* 3747-3752. http://dx.doi.org/10.1073/pnas.0400087101

Bennett, K. P., & Mangasarian, O. L. (1992) Robust Linear Programming Discrimination of Two Linearly Inseparable Sets. *Optimization Methods and Software, 1,* 23-34. http://dx.doi.org/10.1080/10556789208805504

Blanch, A., & Aluja, A. (2013). A Regression Tree of the Aptitudes, Personality, and Academic Performance Relationship. *Personality and Individual Differences, 54,* 703-708. http://dx.doi.org/10.1016/j.paid.2012.11.032

Borg, I. & Groenen, P. (2005). *Modern Multidimensional Scaling: Theory and Applications* (2nd ed.). New York: Springer-Verlag.

Borkin, M. A., Vo, A. A., Bylinskii, Z., Isola, P., Sunkavalli, S., Oliva, A., & Pfister, H. (2013). What Makes a Visualization Memorable? *Visualization and Computer Graphics IEEE Transactions, 12,* 2306-2315. http://dx.doi.org/10.1109/TVCG.2013.234

Breiman, L. (2001). Random Forests. *Machine Learning, 1,* 5-32. http://dx.doi.org/10.1023/A:1010933404324

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

Cortez, P., & Silva, A. M. G. (2008). Using Data Mining to Predict Secondary School Student Performance. In A. Brito, & J. Teixeira (Eds.), *Proceedings of 5th Annual Future Business Technology Conference,* Porto, 5-12.

Eloyan, A., Muschelli, J., Nebel, M., Liu, H., Han, F., Zhao, T., Caffo, B. et al. (2012). Automated Diagnoses of Attention Deficit Hyperactive Disorder Using Magnetic Resonance Imaging. *Frontiers in Systems Neuroscience, 6*. http://dx.doi.org/10.3389/fnsys.2012.00061

Epskamp, S., Cramer, A. O. J., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012). Qgraph: Network Visualizations of Relationships in Psychometric Data. *Journal of Statistical Software, 48,* 1-18. http://www.jstatsoft.org/v48/i04/

Fruchterman, T. M. J., & Reingold, E. M. (1991). Graph Drawing by Force-Directed Placement. *Software: Practice and Experience, 21,* 1129-1164. http://dx.doi.org/10.1002/spe.4380211102

Geurts, P., Irrthum, A., & Wehenkel, L. (2009). Supervised Learning with Decision Tree-Based Methods in Computational and Systems Biology. *Molecular Biosystems, 5,* 1593-1605. http://dx.doi.org/10.1039/b907946g

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,* 68-101.

Hardman, J., Paucar-Caceres, A., & Fielding, A. (2013). Predicting Students' Progression in Higher Education by Using the Random Forest Algorithm. *Systems Research and Behavioral Science, 30,* 194-203. http://dx.doi.org/10.1002/sres.2130

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference and Prediction* (2nd ed.). New York: Springer. http://dx.doi.org/10.1007/978-0-387-84858-7

Honarkhah, M., & Caers, J. (2010). Stochastic Simulation of Patterns Using Distance-Based Pattern Modeling. *Mathematical Geosciences, 42,* 487-517. http://dx.doi.org/10.1007/s11004-010-9276-7

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R.* New York: Springer. http://dx.doi.org/10.1007/978-1-4614-7138-7

Kalna, G., & Higham, D. J. (2007). A Clustering Coefficient for Weighted Networks, with Application to Gene Expression Data. *Journal of AI Communications-Network Analysis in Natural Sciences and Engineering, 20,* 263-271.

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. New York: Springer. http://dx.doi.org/10.1007/978-1-4614-6849-3

Lemon, J. (2006). Plotrix: A Package in the Red Light District of R. *R-News, 6,* 8-12.

Liaw, A., & Wiener, M. (2012). *Random Forest: Breiman and Cutler's Random Forests for Classification and Regression*. R Package Version 4.6-7.

Mangasarian, O. L., & Wolberg, W. H. (1990). Cancerdiagnosis via Linear Programming. *SIAM News, 23,* 1-18.

Mangasarian, O. L., Setiono, R., & Wolberg, W. H. (1990). Pattern Recognition via Linear Programming: Theory and Application to Medical Diagnosis. In T. F. Coleman, & Y. Y. Li (Eds.), *Large-Scale Numerical Optimization* (pp. 22-30). Philadelphia, PA: SIAM Publications.

Onnela, J. P., Saramaki, J., Kertesz, J., & Kaski, K. (2005). Intensity and Coherence of Motifs in Weighted Complex Networks. *Physical Review E, 71,* Article ID: 065103. http://dx.doi.org/10.1103/PhysRevE.71.065103

Quach, A. T. (2012). *Interactive Random Forests Plots*. All Graduate Plan B and Other Reports, Paper 134, Utah State Univesity.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. http://www.R-project.org

Seni, G., & Elder, J. F. (2010). *Ensemble Methods in Data Mining: Improving Accuracy through Combining Predictions*. Morgan & Claypool Publishers. http://dx.doi.org/10.2200/S00240ED1V01Y200912DMK002

Skogli, E., Teicher, M. H., Andersen, P., Hovik, K., & Øie, M. (2013). ADHD in Girls and Boys—Gender Differences in Co-Existing Symptoms and Executive Function Measures. *BMC Psychiatry, 13,* 298. http://dx.doi.org/10.1186/1471-244X-13-298

Steincke, K. K. (1948). *Farvelogtak: Ogsaaen Tilvaerelse IV*. København: Fremad.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (4th ed.). New York: Springer. http://dx.doi.org/10.1007/978-0-387-21706-2

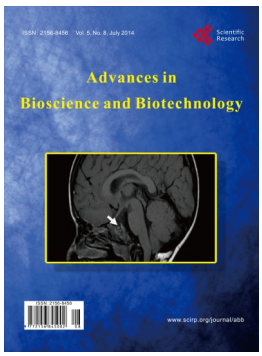
Wickham, H., Caragea, D., & Cook, D. (2006). Exploring High-Dimensional Classification Boundaries. *Proceedings of the 38th Symposium on the Interface of Statistics, Computing Science, and Applications—Interface 2006: Massive Data Sets and Streams*, Pasadena, May 24-27 2006.

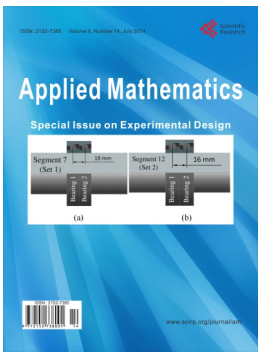
Wolberg, W. H., & Mangasarian, O. L. (1990) Multisurface Method of Pattern Separation for Medical Diagnosis Applied to Breast Cytology. *Proceedings of the National Academy of Sciences of the United States of America, 87,* 9193-9196.

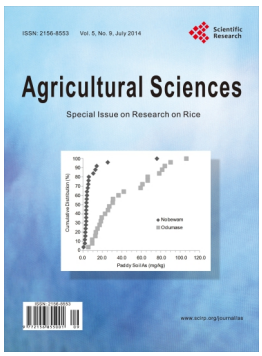
Zhang, B., & Horvath, S. (2005). A General Framework for Weighted Gene Co-Expression Network Analysis. *Statistical Applications in Geneticsand Molecular Biology, 4*. http://dx.doi.org/10.2202/1544-6115.1128

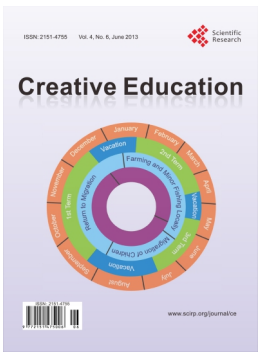
Scientific Research Publishing (SCIRP) is one of the largest Open Access journal publishers. It is currently publishing more than 200 open access, online, peer-reviewed journals covering a wide range of academic disciplines. SCIRP serves the worldwide academic communities and contributes to the progress and application of science with its publication.

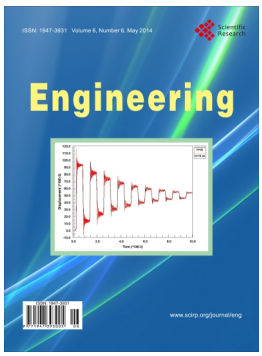
Other selected journals from SCIRP are listed as below. Submit your manuscript to us via either submit@scirp.org or Online Submission Portal.

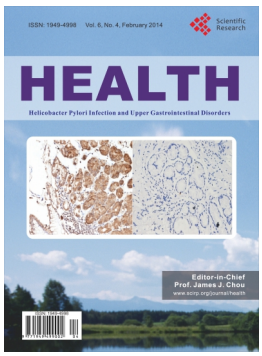


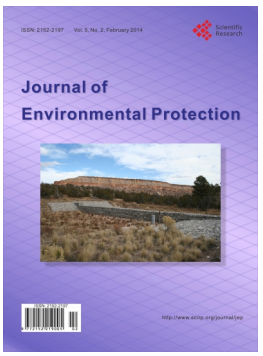


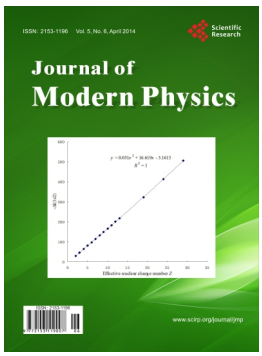


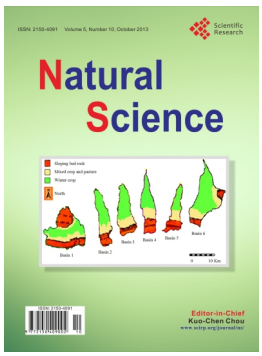


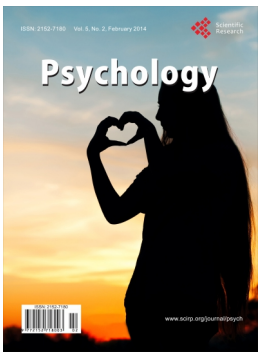












Random Forest as an imputation method for psychology research: its impact on item fit and difficulty of the Rasch Model.

Hudson F. Golino
Universidade Estadual de Feira de Santana, Bahia, Brasil
E-mail: hfgolino@gmail.com

Cristiano Mauro Assis Gomes
Universidade Federal de Minas Gerais, Brasil
E-mail: cristianogomes@ufmg.br

Abstract: The present paper presents a new non-parametric imputation technique, named random forest, from the machine learning field. The random forest procedure has two tuning parameters, the number of trees grown in the prediction and the number of predictors used. Fifty experimental conditions were created in the imputation procedure, with different combinations of predictors (from 1 to 10) and number of trees (10, 50, 100, 500 and 1000). We examined how each experimental condition affected the fit of the items of an inductive reasoning test to the dichotomous Rasch Model, as well as its difficulty. The results point that using random forest to impute missing values is a reliable technique to be used in psychological researches, since it led to statistically significant differences in the infit's median only in 4% of the experimental conditions investigated, compared to the original missing values dataset result. However, researchers should be aware that in 32% of the experimental conditions the imputation procedure significantly increased the estimated items' difficulty median, compared to the original dataset.

Keywords: Imputation, Machine Learning, Random Forest, Testing, Assessment.

Introduction

In general, there are three main types of missing data: missing at random (MAR), missing completely at random (MCAR) and missing not at random (MNAR). In the first case, the missing is independent from the variable containing it, while in the second case the missing is independent of the variable containing the missing and of all observed variables (Enders, 2010; Hohensinn & Kubinger, 2011). Finally, in the third type of missing data (MNAR) the occurrence of missing is dependent on the variable that contains the missing and is known as the "not ignorable missing" type, since it causes much of the problems in

data analysis (Enders, 2010; Hohensinn & Kubinger, 2011; Schafer & Graham, 2002). Usually, researches using large scale educational assessments and tests have to deal with missing values before running an analysis (e.g. Weirich, Haag, Hecht, Böhme, Siegle, & Lüdtke, 2014). However, this is not an easy task, due to a number of reasons as the magnitude of missing data, the type of missing data and the possible bias it can create in the analysis, among other issues (Koretz, McCaffrey, & Sullivan, 2001). In general, there is a number of procedures to deal with missing values, from listwise deletion to multiple imputation methods, each one having advantages and pitfalls, and being more suitable for MAR, MCAR or MNAR situations (see Allison, 2001; Barnard, Rubin, & Schenker, 1998; Little & Rubin, 2002; Schafer & Graham, 2002). The majority of the imputation methods available depend on tuning parameters or on the stipulation of a parametric model, making assumptions about the distribution of the data (Stekhoven & Buehlmann, 2012).

Table 1. Missing data handling methods and imputation techniques: pros and cons.

Method	Pros	Cons
Listwise deletion	Is a quick and easy-to-apply method, being the default in a number of statistical softwares. It is suported for a large number of procedures.	It simply deletes the observations with missing values, decreasing the sample size. Is a missing data handling strategy that can compromise the statistical analysis if a larger sample size is demanded (as in confirmatory factor analysis).
Pairwise deletion	It takes advantage of the available data to compute the statistical techniques.	Have limited application in education researches because many general statistical packages do not suport this kind of missing data handling strategy.
Mean imputation	Is easy to implement, since it imputes the missing values with the mean of the available data for that variable.	Decreases the standard error of the mean, increasing the risk of type I error.
Conditional mean imputation	Predicts the missing values based on the available through regression. Can be easily implemented in most of the statistical softwares.	Produces biased standard errors of parameter estimates, requires a multivariate normal distribution and all the other linear regression assumptions.
Maximum likelihood expectation-maximization imputation	Produces unbiased standard errors of parameter estimates.	Only can be used for linear and log-linear models.
Multiple imputation	Simulates the natural variation in missing data, leading to unbiased standard errors of parameter estimates.	It is somehow challenging to use, may lead to different values, for

		the same dataset using the same method.
Hot Deck Imputation	Uses the available data, imputing a missing value based on an observed value that is closer in terms of distance.	The theory behind it is not as well developed as the other methods.
Dummy Variable Adjustment	Simply to understand and to apply.	Produce biased parameter estimates.
Zero Imputation	Simply to understand and to apply. Can be very useful when the missing occurs because the respondent did not know the correct answer.	Limited usefulness. Can lead to biased estimates.
Single random imputation	Can be viewed as a compromise between regression imputation and multiple imputation.	Underestimates parameter estimates.
Last Observed Value Carried Forward (LOCF)	Suited for longitudinal studies.	Produce biased parameter estimates with lower standard errors.

In the psychological and educational testing field, missing values can also be treated as incorrect responses or as not administered items. The reason to treat missing data as incorrect responses is straightforward, since one may judge the test incompleteness as a lack of ability to solve the items presented to the examinee. The second strategy, by its turn, is possible due to an important characteristic of one of the most famous contemporary psychometric model, the Simple Logistic Model (SLM, a.k.a. the dichotomous Rasch Model) developed by Georg Rasch (1960/1980): the separation of item and person parameters. Briefly, the SLM establishes that the right/wrong scored response X_{pi}, which emerges from the encounter between the person p and the item i, depends upon the ability β of that person and on the difficulty δ of the item, and can be expressed as the following probabilistic function:

$$P\left\{X_{pi} = x\right\} = \frac{e^{x(\beta_p - \delta_i)}}{1 + e^{(\beta_p - \delta_i)}} \quad (1)$$

If X_{pi} is correct (i.e. equals one), then:

$$P\left\{X_{pi} = 1\right\} = \frac{e^{(\beta_p - \delta_i)}}{1 + e^{(\beta_p - \delta_i)}} \quad (2)$$

and If X_{pi} is incorrect (i.e. equals zero), then:

$$P\left\{X_{pi} = 0\right\} = \frac{1}{1 + e^{(\beta_p - \delta_i)}} \quad (3)$$

Andrich (1988) shows the probability of the first person correctly answering an item and the second person incorrectly answering an item, given that only one of them answers it correctly is:

$$P\left\{(x_{1i}=1\wedge x_{2i}=0)\,\middle|\,(x_{1i}=1\wedge x_{2i}=0)\vee(x_{1i}=0\wedge x_{2i}=1)\right\}=$$
$$\left(\frac{\frac{e^{(\beta_1-\delta_i)}}{1+e^{(\beta_1-\delta_i)}}\frac{1}{1+e^{(\beta_2-\delta_i)}}}{\left(\frac{e^{(\beta_1-\delta_i)}}{1+e^{(\beta_1-\delta_i)}}\frac{1}{1+e^{(\beta_1-\delta_i)}}\right)+\left(\frac{1}{1+e^{(\beta_2-\delta_i)}}\frac{e^{(\beta_2-\delta_i)}}{1+e^{(\beta_2-\delta_i)}}\right)}\right)=\frac{e^{\beta_1}}{e^{\beta}1+e^{\beta_2}}$$

(4)

Therefore, the probability of the first person correctly answering an item, when only one of them give the right answer depends only on the relative ability of the persons, and not on the difficulty of the item itself! If we change the example to a person answering two items (1 and 2), the same holds, i.e. the probability of the first item being correctly answered, when only one of them is correctly answered depends only on the relative difficulty of the items, and not on the ability of the persons.

The separation of person and item parameters was a very important development in psychometrics. It enabled, for example, much more flexibility in data collection than was possible by classical test theory or by other latent variable models. Once you can compare the ability of two persons independently of the items they answer, you can give different numbers and combinations of items to different samples, and analyze them all together (Bond & Fox, 2007). In sum, the Simple Logistic Model of Georg Rasch (1960/1980) provides in its very mathematical foundations a kind of solution for missing data: you can treat them as not administered items, without affecting the comparison of people. Indeed, the psychometric literature points that the strategy of treating missing data as not administered items leads to lower bias than the strategy to handle it as incorrect responses (Hohensinn & Kubinger, 2011; Shin, 2009).

When someone is making an imputation procedure in psychological and/or educational testing, the specification of the model is crucial to both the method's performance and to the estimation of the item and/or person parameters. As pointed before, the imputation methods available generally make assumptions about how the data is distributed (Stekhoven

& Buehlmann, 2012). There is a lack of non-parametric imputation models. Since one of the underlying interests of the psychological testing field lies in imputing values that are as close as possible to the "real value", i.e. the value people would have choose or obtained if she/he had answered to the item, the application of 'modern' prediction models is a possible strategy to be followed. In the present paper, we will present a new non-parametric approach to deal with missing data that allows interactive and non-linear effects: Random Forest (Breiman, 2001). Random forest is a machine learning method that is used for prediction (both classification and regression). We will focus on the classification procedures, since in psychological testing the general interest lies in imputing categorical responses (dichotomous or polytomous). The main benefits of using random forest (or any other tree-based models) is that it does not make any assumption regarding normality, linearity of the relation between variables, homoscedasticity, collinearity or independency (Geurts, Irrthum, & Wehenkel, 2009). Random forest also does not demand a high sample-to-predictor ratio and is more suitable to interaction effects (especially non-linearity) than the classical techniques. Finally, it can lead to high accuracies, since it is known as one of the state-of-the-art methods in terms of prediction (Flach, 2012; Geurts et al., 2009). The next section will introduce the basic notions of the machine learning models, in particular the tree-based models (random forest is one of them). Then, we will apply this new method in a real dataset. The goal is to evaluate how an imputation based on random forest (with different experimental conditions) alters the estimation of items' fit to the simple logistic model of Georg Rasch (1960/1980) and its difficulty parameter, compared to the fit and difficulty provided by treating the missings as not administered items.

Machine Learning Models:

Machine learning is a broad class of computational and statistical methods 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 is used to accomplish different kind of tasks for predictions, inferences, and knowledge discovery (Flach, 2012; Hastie et al., 2009). The machine learning techniques are divided in two main areas, each one accomplishing different kind of tasks: unsupervided and supervised learning.

The former is used to discover, to detect or to learn relationships, structures, trends or patterns in data. In an unsupervised learning task, there is a d-vector of observations or measurements of features, $\mathfrak{X} = \mathfrak{F}_1 \times \mathfrak{F}_2 \times \mathfrak{F}_3 \times ... \times \mathfrak{F}_d$, but no previously known outcome or associated response (Flach, 2012; James, Witten, Hastie, & Tibshirani, 2013). The supervised learning field, on the other hand, deals with tasks where there is an associated response or outcome y_i, $y_i \in \mathfrak{Y}$, for each observation of a predictor x_i, $i = 1, \dots, n$. The d-vector $\mathfrak{X}$ is called the feature space and the vector $\mathfrak{Y}$ is called the output space. The difference between the unsupervised learning tasks and the supervised one relies in the data structure. While the former is composed only by the d-vector $\mathfrak{X}$, or the feature space, the latter is composed by $\mathfrak{X}$ and by the output space $\mathfrak{Y}$

In the case where there is an associated response or outcome for each observation, 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 two classes), the goal is to construct a labeling function (l) that maps the feature space into the output space composed by a small and finite set of classes $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_k\}$, so that $l: \mathfrak{X} \to \mathcal{C}$. In this case the output space is a vector containing the allocation of each individual in one of the $\mathcal{C}_k$ class: $\mathfrak{Y} \equiv \mathcal{C}$. In sum, in the classification problem a categorical outcome (right answer vs wrong answer), is predicted using a set of features (or predictors, independent variables). The present paper deals with a classification problem: predicting right and wrong responses in particular items, based on the responses on other items.

The machine learning literature presents a wide range of computational and statistical models to solve classification problems (Hastie et al., 2009). The tree-based models are amongst the most used ones. A classification tree (Breiman, Friedman, Olshen, & Stone, 1984) partitions the feature space into several distinct mutually exclusive regions (Figure 1). Each region is fitted with a specific model that designates one of the classes to that particular space.

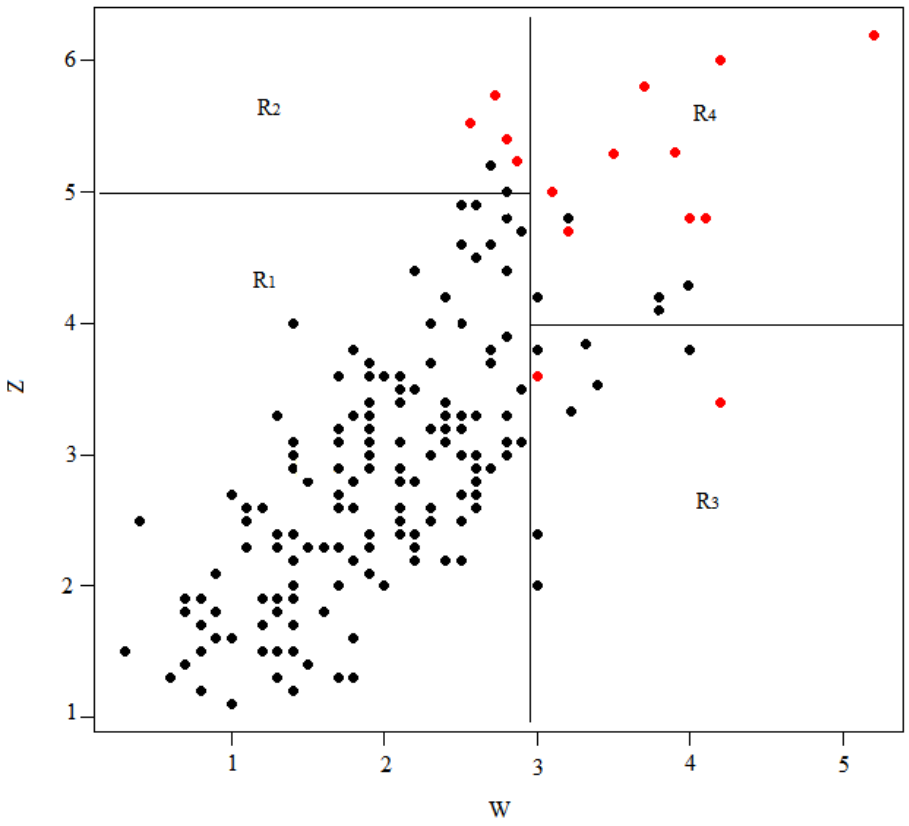


Figure 1. Partitioning of 2-dimensional feature space into four non-overlapping regions R_1, R_2, R_3 and R_4.

A class is assigned to the region of the feature space by identifying the majority class in that region. Figure 1 shows the distribution of observations in two variables, *W* and *Z*, colored according to two classes: C_1 (black dots) and C_2 (red dots). A person s_n with a value of *W* smaller than three and a value of *Z* smaller than five belongs to the region R_1 of the feature space: $s_n \in R_1 \leftrightarrow \{W < 3 \ \wedge Z < 5\}$. Since C_1 is the majority class in R_1, those falling within this region will be estimated as pertaining to the class C_1: $R_1 \rightarrow \widehat{C_1}$. In the same line, people with a value of *W* smaller than three and a value of *Z* greater than five belongs to the region R_2 of the feature space: $s_n \in R_2 \leftrightarrow \{W < 3 \ \wedge Z > 5\}$. Since C_2 is the majority class in R_2, those falling within this region will be estimated as belonging to the class C_2: $R_2 \rightarrow \widehat{C_2}$. People with *W* greater than three and a value of *Z* smaller than four belongs to the region R_3 and $R_3 \rightarrow \widehat{C_1}$ because class C_1 is majoritarian in R_3. Finally, those with *W* greater than three and *Z* greater than four belongs to the region R_4, $s_n \in R_4 \leftrightarrow \{W > 3 \ \wedge Z > 4\}$, and $R_4 \rightarrow \widehat{C_4}$.

This set of if-then rules can be easily understood by inspecting the structure of the classification tree (see Figure 2).

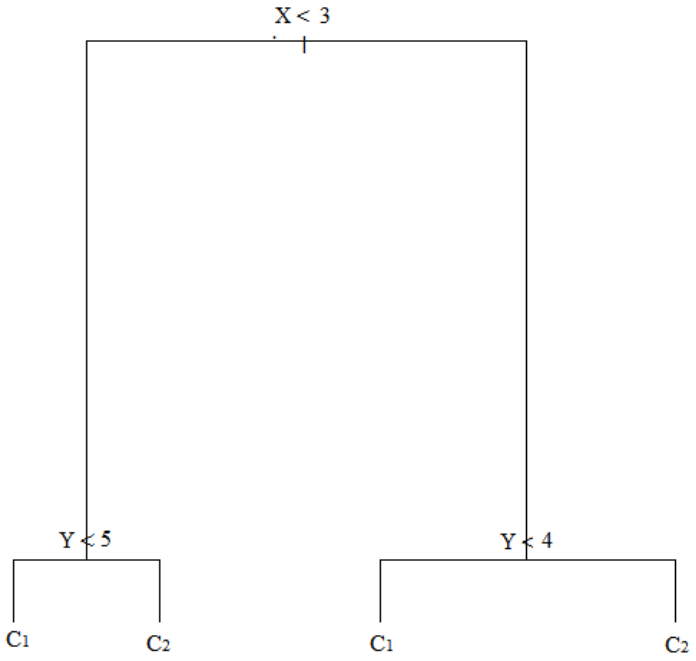


Figure 2. Classification Tree.

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, *right answer*. A node with 90% of *right answer*s and 10% of *wrong answers* 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 (predictor) that leads to the most purity configuration (top split in Figure 2). Then, each region of the tree is modeled accordingly to the majority class. One or two original nodes are also split into more nodes, using some of the given predictors that provide the best fit possible (bottom split in Figure 2). This splitting process continues until the feature space achieves the most purity configuration possible, with R_m regions or nodes classified with a distinct C_k class. 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. In a classification tree, the first split indicates the most important variable, or feature, in the prediction.

The classification trees have two main basic tuning parameters (for more fine grained tuning parameters see Breiman, Friedman, Olshen & 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 $\alpha|T|$. Geurts, Irrthum and Wehenkel (2009) argue that classification 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, does not make any assumption regarding normality, linearity or independency and can be applied in datasets with a large *p* low *n* characteristic (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. 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 splitting, 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 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 statistical learning approaches (James et al., 2013). In order to prevent overfitting, the variance issue and also to increase the accuracy of the regression trees, a strategy named *ensemble techniques* can be used.

Ensemble techniques are simply the junction of several models to perform the classification task. The main technique that ensemble classification trees is called random forest (Breiman, 2001). Random forest increases prediction accuracy and decreases variance between data sets as well as avoid overfitting (James et al., 2013) by bootstrapping both the sample and the predictors (Flach, 2012, Hastie et al., 2009; James et al., 2013). The procedure generates *B* different bootstraps from the training set, taking a random subsample n of the original data set with replacement and a subsample $\mathfrak{X}_m$ of the feature space $\mathfrak{X}$ at each node to growing the trees. It, then, assigns a value to the R_j regions of the feature space for every b. Lastly, the resulting predictions are averaged in order to make a more accurate prediction, based on the multiple bootstrap (Hastie et al., 2009; James et al., 2013). As the random forest does not use the entire observations (only a subsample of it, usually 2/3), the remaining observations (known as *out-of-bag*, or OOB) is used to verify the quality of the prediction. The out-of-bag error can be computed as a "valid estimate of the test error for the random forest model, since the response for each observation is predicted using only the trees that were not fit using that observation" (p. 323, James et al., 2013). Random Forests have two main basic tuning parameters: 1) the size of the subsample of features 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.

Classification trees have the advantage of being easy to interpret, but can lead to important issues, such as overfitting and variance. Random forest is an ensemble of learning trees that deal with both issues by bootstrapping the sample and the predictors. Random forest is also one of the state-of-the-art methods in terms of accuracy (Flach, 2012; Geurts et al., 2009). Since it is a method used for prediction, it can be easily applied in the context of missing data for imputing, or predicting, values. In the present paper we show how an imputation based on random forest (with different experimental conditions) alters the estimation of item and person parameters, and its fit to the simple logistic model of Georg Rasch (1960/1980), compared to the estimation and fit provided by treating the missings as not administered items.

Methods

Dataset

In the current paper we analyzed data from the *Inductive Reasoning Developmental Test – 3rd version* (IRDT - Golino & Gomes, 2012), collected over the past two years in 1810 Brazilian people. 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. The IRDT was created to assess developmental stages of reasoning based on Common's Hierarchical Complexity Model (Commons, 2008; Commons & Pekker, 2008; Commons & Richards, 1984) and on Fischer's Dynamic Skill Theory (Fischer, 1980; Fischer & Yan, 2002). The dataset was published by Golino and Gomes (2014) for reproducible purposes.

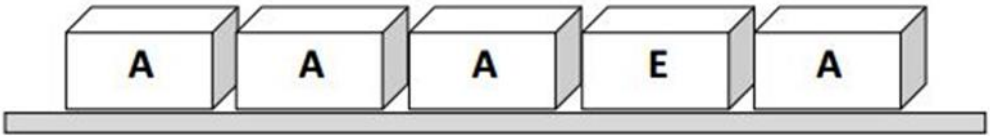


Figure 3. Inductive Reasoning Developmental Test item example (from the lowest difficulty level).

Imputation

In order to apply the random forest approach to impute missing values, we used the package *missforest* (Stekhoven & Buehlmann, 2012) of the R software (R Core Team, 2013). This package addresses the missing data issue using an iterative imputation scheme: first it trains a random forest classifier on observed values, then it predicts the missing values and finally proceed iteratively (Stekhoven & Buehlmann, 2012). For categorical data the *missforest* package provides the proportion of falsely classified entries (PFC) over the categorical missing values, which is an index ranging from 0 to 1, indicating the quality of the imputation procedure. Fifty experimental conditions were used in the imputations, varying the number of trees grown (*ntree*: 10, 50, 100, 500 and 1000) and the number of predictors (*mtry:* from 1 to 10). Each imputation had the maximum number of iterations fixed to ten.

Rasch Model

The *eRm* package (Mair & Hatzinger, 2007a) of the R statistical software were used to apply the simple logistic model of Georg Rasch (1960/1980) for dichotomous responses. This package uses the conditional maximum likelihood to estimate item parameters (Mair & Hatzinger, 2007b). The item fit statistic used was the information weighted (infit) mean-

square statistic. It represents "the amount of distortion of the measurement system" (Linacre, 2002. p.1). Values between 0.5 and 1.5 logits are considered productive for measurement, and <0.5 and between 1.5 and 2.0 are not productive for measurement, but do not degrade it (Wright & Linacre, 1994). We also compared the item difficulty parameter under the fifty imputation conditions.

Comparing the experimental conditions

The Shapiro-Wilk normality test was used to verify if the error distribution, as well as the item fit and parameter distribution, came from a population with normal distribution, using the 5% significance level. The normally distributed variables can be compared in terms of the imputation conditions (number of trees and number of variables used in the prediction) via ANOVA and the Tukey's post-hoc test. The non-normally distributed variables can be compared using the multiple comparison version of the Kruskal-Wallis test implemented in the *pgirmess* package (Giraudoux, 2014).

RESULTS

Random Forest Imputation Results

The Shapiro-Wilk normality test showed that is possible to refute the hypothesis that the imputation errors (falsely classified entries) came from a population with normal distribution (W = 0.70, p = 6.901e-09). The median proportion of falsely classified entries varied from .10 to .19 (Median = .11, Mean = .12, SD = .03) when the number of trees (*ntree*) was set to 10. When *ntree* was set to 50 it ranged from .08 to .13 (Median = .08, Mean = .09, SD = .01) and from .08 to .12 (Median = .08, Mean = .08, SD = .01) when it was set to 100. When the number of trees used in the prediction increased to 500, the falsely classified entries varied from .08 to .11 (Median = .08, Mean = .08, SD = .01), and when set to 1000 it varied from .07 to .11 (Median = .08, Mean = .08, SD = .01). The Kruskal-Wallis multiple comparison (see Table 1) showed that there was a statistically significant difference in the proportion of falsely classified entries between 10 and 100 (Obs. Dif. = 20.60, Critical Diff. = 18.30, p <.05), 500 (Obs. Dif. = 25.50, Critical Diff. = 18.30, p <.05) and 1000 trees (Obs. Dif. = 27.40, Critical Diff. = 18.30, p <.05).

Table 1. Kruskal-Wallis Multiple Comparison for different number of trees

Number of trees	Observed Difference	Critical Difference	p-value
10-50	13.00	18.30	p >.05
10-100	20.60	18.30	p <.05*
10-500	25.50	18.30	p <.05*
10-1000	27.40	18.30	p <.05*
50-100	7.60	18.30	p >.05
50-500	12.50	18.30	p >.05
50-1000	14.40	18.30	p >.05
100-500	4.90	18.30	p >.05
100-1000	6.80	18.30	p >.05
500-1000	1.90	18.30	p >.05

The median proportion of falsely classified entries varied from .11 to .19 (Median = .12, Mean = .13, SD = .03) when the number variables used in the prediction (*mtry*) was set to one. When *mtry* was set to two it ranged from .09 to .13 (Median = .09, Mean = .10, SD = .02) and from .08 to .12 (Median = .08, Mean = .09, SD = .01) when it was set to three. When the number of predictors used was from four to six, the falsely classified entries varied from .08 to .11 (Median = .08, Mean = .09, SD = .01), and when set to seven it varied from .07 to .10 (Median = .08, Mean = .08, SD = .01). When *mtry* was set from eight to ten, the proportion of falsely classified entries varied from .08 to .10 (Median = .08, Mean = .08, SD = .01).

The Kruskal-Wallis multiple comparison (see Table 2) showed that there was a statistically significant difference in the proportion of falsely classified entries only between 1 and 9 predictors (Obs. Dif. = 30.60, Critical Diff. = 30.06, p <.05), and between 1 and 10 (Obs. Dif. = 31.20, Critical Diff. = 30.06, p <.05).

Table 2. Kruskal-Wallis Multiple Comparison for different number of predictors

Number of predictors	Observed Difference	Critical Difference	p-value
1-2	10.00	30.06	p >.05
1-3	13.40	30.06	p >.05
1-4	18.20	30.06	p >.05
1-5	21.40	30.06	p >.05
1-6	25.00	30.06	p >.05
1-7	28.20	30.06	p >.05
1-8	29.00	30.06	p >.05
1-9	30.60	30.06	p <.05*
1-10	31.20	30.06	p <.05*
2-3	3.40	30.06	p >.05
2-4	8.20	30.06	p >.05
2-5	11.40	30.06	p >.05

2-6	15.00	30.06	p >.05
2-7	18.20	30.06	p >.05
2-8	19.00	30.06	p >.05
2-9	20.60	30.06	p >.05
2-10	21.20	30.06	p >.05
3-4	4.80	30.06	p >.05
3-5	8.00	30.06	p >.05
3-6	11.60	30.06	p >.05
3-7	14.80	30.06	p >.05
3-8	15.60	30.06	p >.05
3-9	17.20	30.06	p >.05
3-10	17.80	30.06	p >.05
4-5	3.20	30.06	p >.05
4-6	6.80	30.06	p >.05
4-7	10.00	30.06	p >.05
4-8	10.80	30.06	p >.05
4-9	12.40	30.06	p >.05
4-10	13.00	30.06	p >.05
5-6	3.60	30.06	p >.05
5-7	6.80	30.06	p >.05
5-8	7.60	30.06	p >.05
5-9	9.20	30.06	p >.05
5-10	9.80	30.06	p >.05
6-7	3.20	30.06	p >.05
6-8	4.00	30.06	p >.05
6-9	5.60	30.06	p >.05
6-10	6.20	30.06	p >.05
7-8	0.80	30.06	p >.05
7-9	2.40	30.06	p >.05
7-10	3.00	30.06	p >.05
8-9	1.60	30.06	p >.05
8-10	2.20	30.06	p >.05
9-10	0.60	30.06	p >.05

Figure 4 shows the imputation error behavior for every experimental condition used. For each number of trees used, the median of falsely classified entries (errors) decreased exponentially with the increase in the number of predictors (*mtry*). The colored solid lines in Figure 4 shows the loess smoothed decline tendency and the shaded gray area shows the 95% confidence interval on the fitted values (solid line).

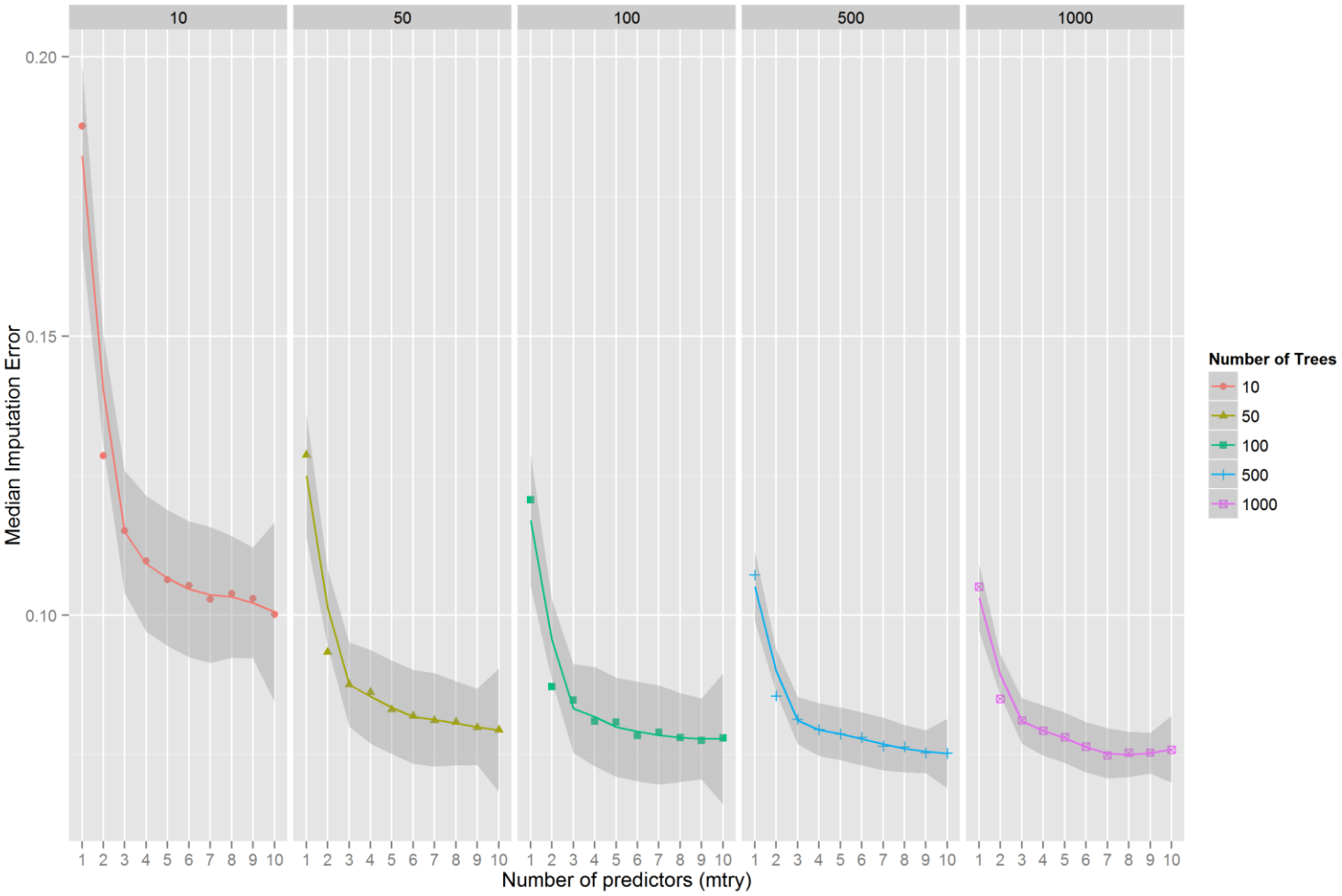


Figure 4. Median imputation error by experimental condition

Figure 4 shows that when the number of trees are set to 10, the median imputation error is above 10%. It is also above 10% when the number of predictors are set to one, irrespective of the number of trees used in the random forest procedure. The proportion of falsely classified entries achieved its lowest value (.07) when the numbers of predictors was seven and the number of trees was 1000. The maximum error value (.19) was achieved with *ntree* of 10 and *mtry* of 1.

Rasch Model Results

In the dataset with missing values the infit varied from 0.68 to 1.21 (Median = 0.86, Mean = 0.96, SD = .15). The items' difficulty parameter ranged from -0.57 to 11.57 logits (Median = 6.32, Mean = 5.93, SD = 3.7). Figure 5 shows the item-person map from the not imputed dataset.

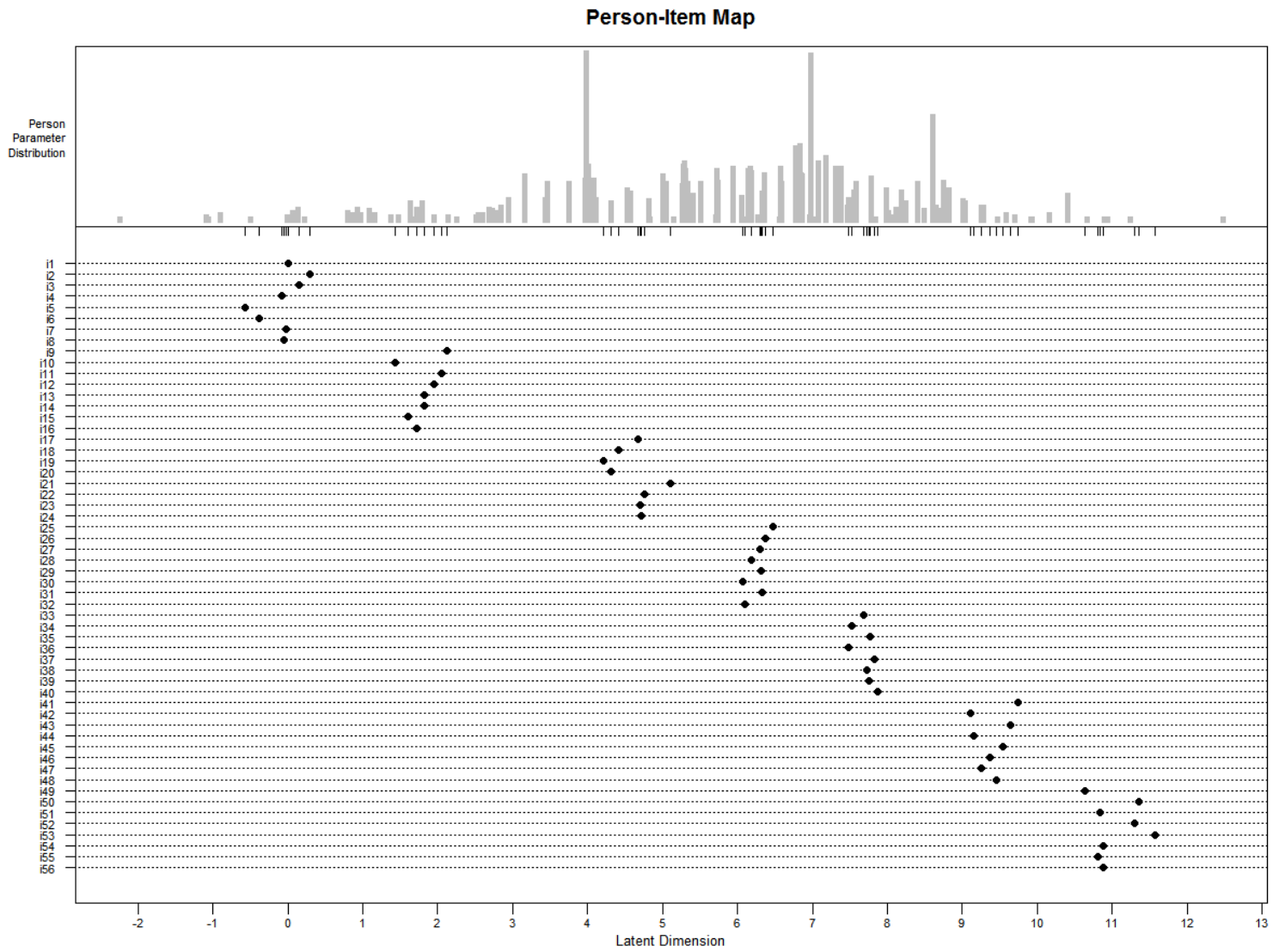


Figure 5. Person-item map from the not-imputed dataset.

The Shapiro-Wilk normality test showed that is possible to refute the hypothesis that the *infit* from all experimental conditions came from a population with normal distribution (W = 0.96, p = 2.2e-16). When the number of trees was set to 10 the median infit varied from 0.82 to 0.91 (Median = 0.85, Mean = 0.86, SD = .03). When *ntree* was set to 50 the median infit ranged from 0.82 to 0.88 (Median = 0.86, Mean = 0.85, SD = .02) and from 0.81 to 0.89 (Median = 0.86, Mean = 0.86, SD = .02) when *ntree* was set to 100. When the number of trees used in the prediction increased to 500 and to 1000, the median infit varied from 0.84 to 0.88 (Median = 0.86, Mean = 0.86, SD = .01). The Kruskal-Wallis multiple comparison

(see Table 3) showed that there was no statistically significant difference in the infit, for any pairwise comparison made between the number of trees and the not imputed dataset.

Table 3. Kruskal-Wallis Multiple Comparison for different number of trees and for the not imputed dataset (INFIT)

Pairwise comparisons: Number of trees and Not imputed dataset	Observed Difference	Critical Difference	p-value
10-50	9.34	144.64	p >.05
10-100	8.36	144.64	p >.05
10-500	30.51	144.64	p >.05
10-1000	11.81	144.64	p >.05
10-Not Imputed	85.97	339.22	p >.05
50-100	17.71	144.64	p >.05
50-500	21.16	144.64	p >.05
50-1000	21.15	144.64	p >.05
50-Not Imputed	95.31	339.22	p >.05
100-500	38.87	144.64	p >.05
100-1000	3.45	144.64	p >.05
100-Not Imputed	77.61	339.22	p >.05
500-1000	42.31	144.64	p >.05
500-Not Imputed	116.47	339.22	p >.05
1000-Not Imputed	74.16	339.22	p >.05

The variability of the infit's median, for each number of predictors used can be viewed in Table 4.

Table 4. Variability of the infit's median, for each number of predictors

Number of predictors	Minimum	Median	Mean	Maximum	Standard Deviation
1	0.81	0.84	0.84	0.91	0.04
2	0.85	0.86	0.86	0.86	0.00
3	0.86	0.87	0.87	0.89	0.01
4	0.86	0.87	0.87	0.88	0.01
5	0.84	0.86	0.86	0.86	0.01
6	0.83	0.85	0.85	0.86	0.01
7	0.84	0.85	0.85	0.87	0.01
8	0.82	0.85	0.86	0.88	0.02
9	0.86	0.86	0.86	0.87	0.00
10	0.85	0.85	0.86	0.87	0.01

The Kruskal-Wallis multiple comparison (see Table 5) showed that there was statistically significant infit difference according to the number of predictors (*mtry*) only between one and two predictors, and between one and four to ten predictors.

Table 5. Kruskal-Wallis Multiple Comparison for different number of predictors and for the not imputed dataset (INFIT)

Pairwise comparison between the number of predictors and the not imputed dataset	Observed Difference	Critical Difference	p-value
1-2	302.66	231.18	p < .05*
1-3	161.02	231.18	p > .05
1-4	256.68	231.18	p < .05*
1-5	266.77	231.18	p < .05*
1-6	265.18	231.18	p < .05*
1-7	272.73	231.18	p < .05*
1-8	290.92	231.18	p < .05*
1-9	292.17	231.18	p < .05*
1-10	284.79	231.18	p < .05*
1-Not Imputed	329.20	400.42	p > .05
2-3	141.65	231.18	p > .05
2-4	45.99	231.18	p > .05
2-5	35.90	231.18	p > .05
2-6	37.49	231.18	p > .05
2-7	29.94	231.18	p > .05
2-8	11.75	231.18	p > .05
2-9	10.50	231.18	p > .05
2-10	17.88	231.18	p > .05
2-Not Imputed	26.53	400.42	p > .05
3-4	95.66	231.18	p > .05
3-5	105.75	231.18	p > .05
3-6	104.16	231.18	p > .05
3-7	111.71	231.18	p > .05
3-8	129.90	231.18	p > .05
3-9	131.15	231.18	p > .05
3-10	123.77	231.18	p > .05
3-Not Imputed	168.18	400.42	p > .05
4-5	10.09	231.18	p > .05
4-6	8.50	231.18	p > .05
4-7	16.05	231.18	p > .05
4-8	34.24	231.18	p > .05
4-9	35.49	231.18	p > .05
4-10	28.11	231.18	p > .05
4-Not Imputed	72.52	400.42	p > .05
5-6	1.59	231.18	p > .05
5-7	5.96	231.18	p > .05

5-8	24.15	231.18	p > .05
5-9	25.40	231.18	p > .05
5-10	18.02	231.18	p > .05
5-Not Imputed	62.43	400.42	p > .05
6-7	7.55	231.18	p > .05
6-8	25.74	231.18	p > .05
6-9	26.99	231.18	p > .05
6-10	19.61	231.18	p > .05
6-Not Imputed	64.02	400.42	p > .05
7-8	18.19	231.18	p > .05
7-9	19.44	231.18	p > .05
7-10	12.06	231.18	p > .05
7-Not Imputed	56.47	400.42	p > .05
8-9	1.25	231.18	p > .05
8-10	6.13	231.18	p > .05
8-Not Imputed	38.28	400.42	p > .05
9-10	7.38	231.18	p > .05
9-Not Imputed	37.03	400.42	p > .05
10-Not Imputed	44.41	400.42	p > .05

Figure 6 shows the median infit meansquare of items for each experimental condition, plus the median infit from the not imputed dataset (dark gray line) and its 95% confidence interval (gray rectangle). It is possible to see that the majority of medians falls within the infit's 95% confidence interval of the dataset with missing values. For those falling outside this confidence interval, only two are statistically significant: 1) 10 trees and 1 predictor; 2) 100 trees and 1 predictor.

.

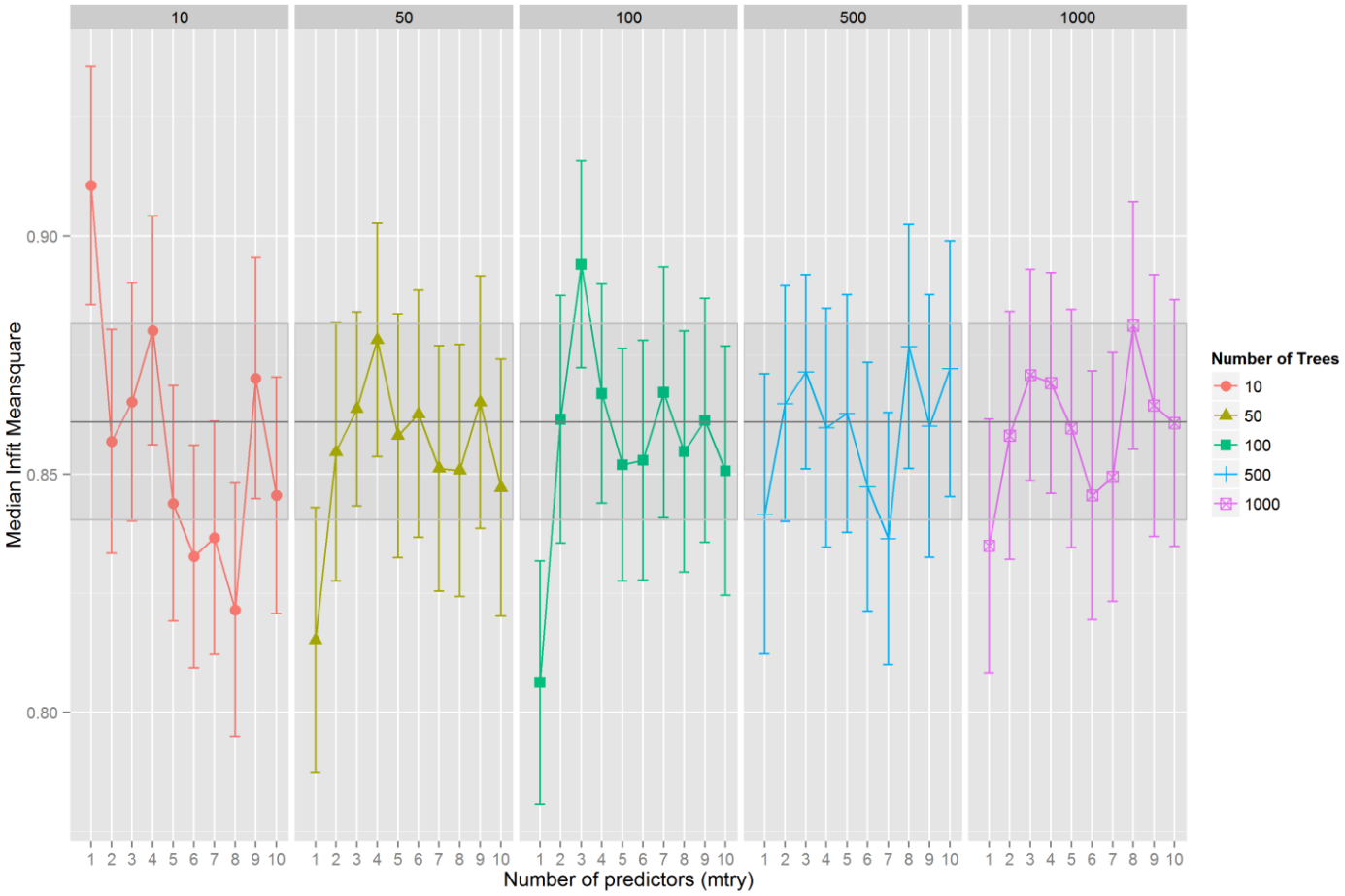


Figure 6. Median Infit MSQ by experimental condition. The dark gray line represents the median infit of the not imputed dataset, and the gray rectangle its 95% confidence interval.

In terms of difficulty, the number of trees alone did not significantly affected its estimation, since for all pairwise comparison (using Kruskal-Wallis multiple comparison) the observed difference was lower than the critical difference (see Table 6).

Table 6. Kruskal-Wallis Multiple Comparison for different number of trees and for the not imputed dataset (item's difficulty median)

Pairwise comparisons: Number of trees and Not imputed dataset	Observed Difference	Critical Difference	p-value
10-50	67.45	143.35	p >.05
10-100	67.06	143.35	p >.05
10-500	68.98	143.35	p >.05
10-1000	64.08	143.35	p >.05
10-Not Imputed	20.68	336.18	p >.05
50-100	0.39	143.35	p >.05
50-500	1.52	143.35	p >.05

50-1000	3.37	143.35	p >.05
50-Not Imputed	88.13	336.18	p >.05
100-500	1.91	143.35	p >.05
100-1000	2.98	143.35	p >.05
100-Not Imputed	87.74	336.18	p >.05
500-1000	4.89	143.35	p >.05
500-Not Imputed	89.66	336.18	p >.05
1000-Not Imputed	84.76	336.18	p >.05

The number of predictors, by the other side, significantly affected the estimation of items' difficulty when comparing one predictor with six and seven predictors (see Table 7).

Table 7. Kruskal-Wallis Multiple Comparison for different number of predictors and for the not imputed dataset (item's difficulty median)

Pairwise comparison between the number of predictors and the not imputed dataset	Observed Difference	Critical Difference	p-value
1-2	121,14	229,11	p > .05
1-3	108,01	229,11	p > .05
1-4	131,31	229,11	p > .05
1-5	183,08	229,11	p > .05
1-6	231,55	229,11	p < .05*
1-7	276,88	229,11	p < .05*
1-8	219,79	229,11	p > .05
1-9	179,40	229,11	p > .05
1-10	171,31	229,11	p > .05
1-Not Imputed	88,05	396,83	p > .05
2-3	13,13	229,11	p > .05
2-4	10,17	229,11	p > .05
2-5	61,93	229,11	p > .05
2-6	110,40	229,11	p > .05
2-7	155,74	229,11	p > .05
2-8	98,65	229,11	p > .05
2-9	58,26	229,11	p > .05
2-10	50,17	229,11	p > .05
2-Not Imputed	33,09	396,83	p > .05
3-4	23,29	229,11	p > .05
3-5	75,06	229,11	p > .05
3-6	123,53	229,11	p > .05
3-7	168,87	229,11	p > .05
3-8	111,78	229,11	p > .05
3-9	71,39	229,11	p > .05
3-10	63,30	229,11	p > .05

3-Not Imputed	19,97	396,83	p > .05
4-5	51,77	229,11	p > .05
4-6	100,24	229,11	p > .05
4-7	145,57	229,11	p > .05
4-8	88,48	229,11	p > .05
4-9	48,09	229,11	p > .05
4-10	40,00	229,11	p > .05
4-Not Imputed	43,26	396,83	p > .05
5-6	48,47	229,11	p > .05
5-7	93,81	229,11	p > .05
5-8	36,72	229,11	p > .05
5-9	3,68	229,11	p > .05
5-10	11,76	229,11	p > .05
5-Not Imputed	95,03	396,83	p > .05
6-7	45,34	229,11	p > .05
6-8	11,75	229,11	p > .05
6-9	52,15	229,11	p > .05
6-10	60,23	229,11	p > .05
6-Not Imputed	143,50	396,83	p > .05
7-8	57,09	229,11	p > .05
7-9	97,48	229,11	p > .05
7-10	105,57	229,11	p > .05
7-Not Imputed	188,84	396,83	p > .05
8-9	40,39	229,11	p > .05
8-10	48,48	229,11	p > .05
8-Not Imputed	131,75	396,83	p > .05
9-10	8,09	229,11	p > .05
9-Not Imputed	91,35	396,83	p > .05
10-Not Imputed	83,27	396,83	p > .05

Figure 7 shows the median difficulty of items for each experimental condition, plus the median difficulty of items from the not imputed dataset (pink line) and its 95% confidence interval (gray rectangle). It is possible to see the following experimental conditions falling outside the items' difficulty median 95% confidence interval of the not imputed dataset: 1) 10 trees and seven or 10 predictors; 2) 50 trees and three, five, six , seven or eight predictors; 3) 100 trees and four or seven predictors; 4) 500 trees and from four to nine predictors; and 5) 1000 trees and four, seven, eight or nine predictors.

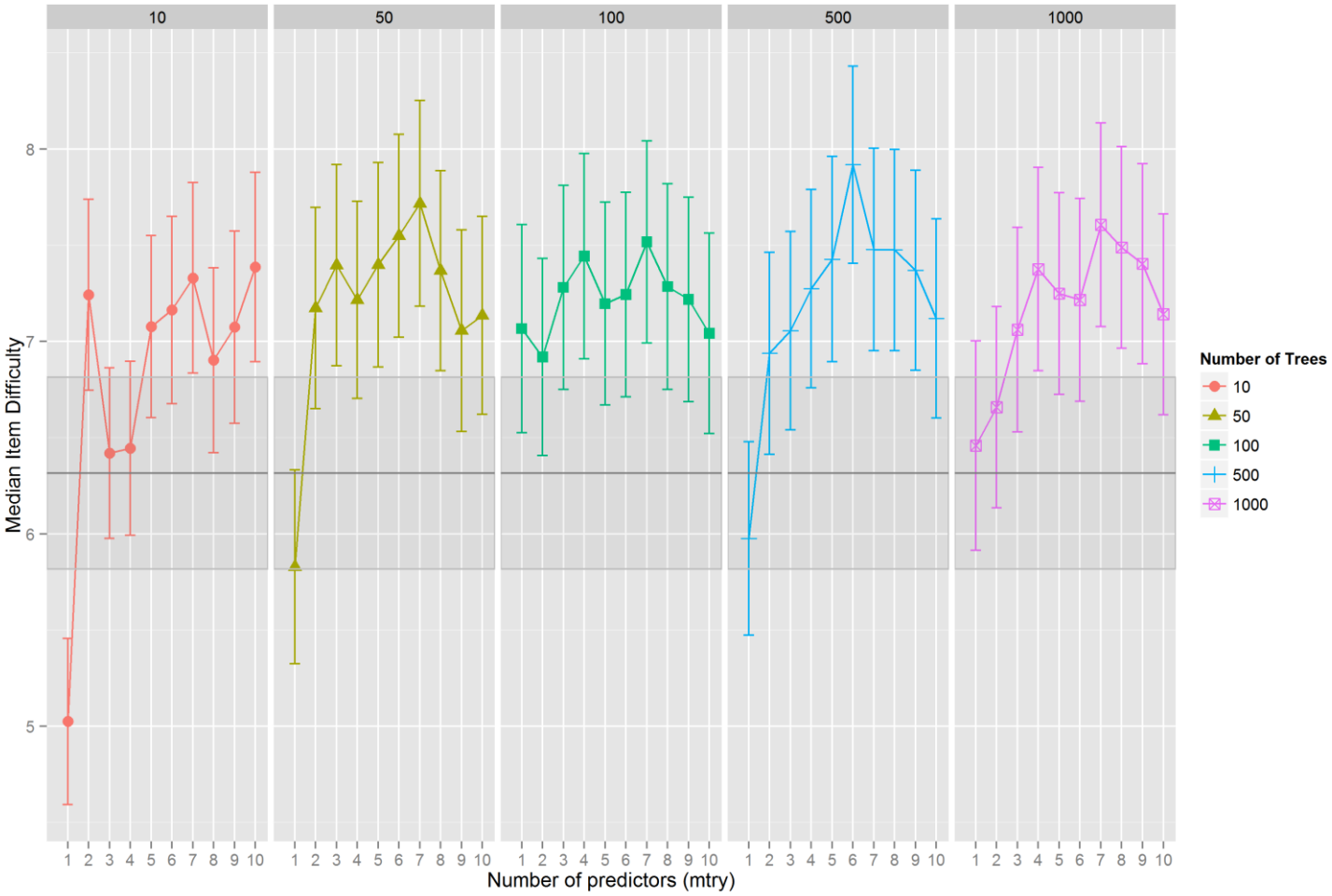


Figure 7. Median item difficulty by experimental condition. The dark gray line represents the median difficulty of items from the not imputed dataset, and the gray rectangle its 95% confidence interval.

Discussion

The majority of the imputation methods available depends on tuning parameters or on the stipulation of a parametric model, being suitable for MAR, MCAR or MNAR situations (Stekhoven & Buehlmann, 2012). In the present paper, we introduced random forest (Breiman, 2001), a new non-parametric machine learning approach that can be applied to deal with missing data (Stekhoven & Buehlmann, 2012). As pointed before, the main benefits of using random forest is that it does not make any assumption regarding normality, linearity of the relation between variables, homoscedasticity, collinearity or independency (Geurts et al., 2009), does not demand a high sample-to-predictor ratio and is more suitable to interaction effects (especially non-linearity) than the classical prediction techniques. It is

also one of the state-of-the-art methods in terms of predictive accuracy (Flach, 2012; Geurts et al., 2009), and can helps the educational assessment field to deal with the missing data issue.

In 2012 the package *missforest* (Stekhoven & Buehlmann, 2012) of the R software (R Core Team, 2013) was released to implement an imputation algorithm based on the random forest model. Among the benefits of using the *missforest* package, it provides the proportion of falsely classified entries (PFC) over the categorical missing values, which is an index ranging from 0 to 1 indicating the quality of the imputation procedure. A previous research showed that the random forest imputation (via *missforest*) led to one of the lowest absolute bias in recovering the true difficulty parameter of a multidimensional item response theory model compared to listwise deletion, forward imputation and multivariate imputation by chained equations (Andreis & Ferrari, 2012).

The present paper investigated how fifty experimental conditions affected the fit of items to the Simple Logistic Model of Rasch (1960/1980), by varying the number of trees (*ntree*: 10, 50, 100, 500 and 1000) and the number of predictors (*mtry:* from 1 to 10) used in the random forest imputation. The result of the Rasch model was compared in each experimental condition, contrasting it to the result obtained when the Rasch model was applied in the dataset with missingness. The choice to compare the experimental imputation conditions with the original missing values dataset is due to the mathematical property of parameters' separation of the simple logistic model of Georg Rasch (1960/1980). So, we treated the missing values as not administered items, and fitted the Rasch model. The result showed the original dataset presented infit values ranging from 0.68 to 1.21 (Median = 0.86, Mean = 0.96, SD = .15), and items' difficulty from -0.57 to 11.57 logits (Median = 6.32, Mean = 5.93, SD = 3.7). The imputation procedures led to errors (proportion of falsely classified entries) ranging from .07 to .19. The lowest infit was 0.45, obtained after setting the number of predictors as one and the number of trees as 50 in the random forest imputation procedure. The highest infit was 1.59, resulted from the use of three predictors and ten trees in the random forest imputation. The lowest infit median was 0.81 (*mtry* = 1, *ntree* = 100), and the highest was 0.91 (*mtry* = 1, *ntree* = 10).

The number of trees alone did not changed items' infit or difficulty median at the 95% significance level, in any pairwise comparison (computed via Kruskal-Wallis multiple

comparison) involving the experimental conditions plus the original dataset results. On the other hand, the number of predictors affected both items' infit and difficulty in some situations. There was a statistically significant infit's median difference between the experimental conditions using only one predictor and those using two or those using between four and ten predictors. By its turn, the items' difficulty median difference was statistically significant when the experimental conditions using only one predictor was compared to those using six or seven predictors. It is important to note, however, that no experimental condition manipulating only the number of predictors (*mtry*) resulted in statistically significant infit or difficulty median differences from the original dataset results.

Comparing the interaction effects between the number of trees and the number of predictors in the random forest imputation procedure, only two conditions led to differences in the infit's median from the original dataset: 1) one predictor and ten trees; 2) one predictor and 100 trees. Their infit's median 95% confidence interval are outside the 95% confidence interval of the original dataset's infit. At the same time, several experimental conditions led to differences in the items' difficulty median from the original dataset: 1) 10 trees and seven or 10 predictors; 2) 50 trees and three, five, six , seven or eight predictors; 3) 100 trees and four or seven predictors; 4) 500 trees and from four to nine predictors; and 5) 1000 trees and four, seven, eight or nine predictors.

In sum, the random forest imputation procedure applied using the *missforest* package (Stekhoven & Buehlmann, 2012) is reliable to be used before fitting the simple logistic model of Rasch (1960/1980), since it led to statistically significant differences in the infit's median only in 4% of the experimental conditions investigated, compared to the original missing values dataset's result. However, researchers should be aware that in 32% of the experimental conditions used in the current paper the imputation procedure significantly increased the estimated items' difficulty median, compared to the original dataset.

References:

Allison, P. D. (2001). *Missing Data*. Sage University Papers Series on Quantitative Applications in the Social Sciences. Thousand Oaks: Sage.

Andreis, F. & Ferrari, P.A. (2012). Missing data and parameters estimates in multidimensional item response models. *Electronic Journal of Applied Statistical Analysis, 5*(3), 431-437.

Andrich, D. (1988). *Rasch models for measurement*. Sage series on quantitative applications in the Social Sciences, Beverly Hills.

Barnard, J., Rubin, D.B., & Schenker, N. (1998). Multiple imputation methods. In: P. Armitage and T. Colton (Eds.), Encyclopedia of Biostatistics (pp. 2772-2780). New York: Wiley.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences (2nd ed.)*. Mahwah, NJ US: Lawrence Erlbaum Associates Publishers.

Breiman, L. (2001). Random Forests. *Machine Learning, 1*(45), 5-32. doi10.1023/A:1010933404324

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*, 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.

Cheema, J. R. (2014). A Review of Missing Data Handling Methods in Education Research. *Review of Educational Research*, *84*(4), 487-508.

Enders, C. (2010). *Applied Missing Data Analysis*. Guilford Press: New York.

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, U.K.: Psychology Press.

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

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.

Giraudoux, P. (2014). *pgirmess: Data analysis in ecology*. R package version 1.5.9. http://CRAN.R-project.org/package=pgirmess

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, PT.

Golino, H.F., & Gomes, C.M.A. (2014): Dataset used in the paper "Random Forest as an imputation method for psychology research: its impact on item fit and difficulty of the Rasch Model". figshare. http://dx.doi.org/10.6084/m9.figshare.1202194

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning:* Data Mining, Inference and Prediction (2nd Edition). New York, NY: Springer.

Hohensinn, C. & Kubinger, K.D. (2011). On the impact of missing values on item fit and the model validness of the Rasch model. *Psychological Test and Assessment Modeling, 53*(3), 380-393.

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

Koretz, D., McCaffrey, D., & Sullivan, T. (2001). Using TIMSS to Analyze Correlates of Performance Variation in Mathematics. *U.S. Department of Education, National Center for Education Statistics.* Working Paper No. 2001-05, Washington, DC.

Linacre J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16* (2), 878.

Little, R.J. & Rubin, D. (2002). *Statistical Analysis with Missing Data*. John Wiley & Sons, Inc: Hoboken.

Mair, P. & Hatzinger, R. (2007a). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software, 20*(9), 1-20.

Mair, P. & Hatzinger, R. (2007b). CML based estimation of extended Rasch models with the eRm package in R. *Psychology Science, 49*, 26-43.

R Development Core Team (2011). R: A Language and Environment for Statistical Computing. [Computer software manual]. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. (Copenhagen, Danish Institute for Educational Research). Expanded edition (1980) with foreword and afterword by B.D. Wright, (1980).Chicago: MESA Press.

Schafer, J.L. & Graham., J.W. (2002). Missing Data: Our View of the State of the Art. *Psychological Methods, 7*, 142-177.

Shin, S.H. (2009). How to Treat Omitted Responses in Rasch Model-Based Equating. *Pratical Assessment, Research & Evaluation, 14*(1).

Stekhoven, D.J. and Buehlmann, P. (2012). MissForest - nonparametric missing value imputation for mixed-type data, *Bioinformatics, 28*(1), 112-118, doi: 10.1093/bioinformatics/btr597

Weirich, S., Haag, N., Hecht, M., Böhme, K., Siegle, T., & Lüdtke, O. (2014). Nested multiple imputation in large-scale assessments. *Large-scale Assessments in Education, 2*(9), 1:18.

Wright, B. D. & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370.

Predicting Academic Dropout using two statistical learning approaches: Classification Trees and Naïve Bayes.

Hudson F. Golino[1,2,4]

Enio Jelihovschi[3]

Cristiano Mauro Assis Gomes[4]

Christopher Rhodes Stephens Stevens[2]

[1]Universidade Estadual de Feira de Santana, Bahia, Brasil
[2]Instituto de Ciências Nucleares, Universidad Nacional Autonoma de Mexico
[3]Departamento de Ciências Exatas e Tecnológicas, Universidade Estadual de Santa Cruz, Brasil
[4]Universidade Federal de Minas Gerais, Brasil

Abstract:
Keywords:

Introduction:

Statistical Learning Models:

Statistical learning are a broad class of computational and statistical methods to extract a model from a system of observations or measurements (Geurts, Irrthum, & Wehenkel, 2009; Hastie, Tibshirani & Friedman, 2009). The extraction of a model from the sole observations is used to accomplish different kind of tasks for predictions, inferences and knowledge discovery (Flach, 2012; Hastie et al., 2009). The statistical learning techniques are divided in two main areas, each one accomplishing different kind of tasks: unsupervided and supervised learning. The former is used to discover, to detect or to learn relationships, structures, trends or patterns in data. In an unsupervised learning task, there is a d-vector of observations or measurements of features, $\mathfrak{X} = \mathfrak{F}_1 \times \mathfrak{F}_2 \times \mathfrak{F}_3 \times ... \times \mathfrak{F}_d$, but no previously known outcome or associated response (Flach, 2012; James, Witten, Hastie, & Tibshirani, 2013). The supervised learning field, on the other hand, deals with tasks where there is an associated response or outcome y_i, $y_i \in \mathfrak{Y}$, for each observation of a predictor x_i, $i = 1, ..., n$. The d-vector $\mathfrak{X}$ is called the feature space and the vector $\mathfrak{Y}$ is called the output space. The difference between the unsupervised learning tasks and the supervised one relies in the

data structure. While the former is composed only by the d-vector $\mathfrak{X}$, or the feature space, the latter is composed by $\mathfrak{X}$ and by the output space $\mathfrak{Y}$ (see Table 1).

Table 1. The data structure in the supervised and unsupervised learning fields.

	Feature space ($\mathfrak{X}$)				Output space
Subjects	$\mathfrak{F}_1$	$\mathfrak{F}_2$	…	$\mathfrak{F}_d$	$\mathfrak{Y}$
s_1	x_{11}	x_{12}	…	x_{1d}	y_1
s_2	x_{21}	x_{21}	…	x_{2d}	y_2
…	…	…	…	…	…
s_n	x_{n1}	x_{n2}	…	x_{nd}	y_n
Unsupervised learning data structure: $\mathfrak{X}$					
Supervised learning data structure: $\mathfrak{X} \wedge \mathfrak{Y}$					

In the case where there is an associated response or outcome for each observation, 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 two classes: dropped from university and regularly enrolled in the university), the goal is to construct a labeling function (l) that maps the feature space into the output space composed by a small and finite set of classes $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \dots, \mathcal{C}_k\}$, so that l: $\mathfrak{X} \to \mathcal{C}$. In this case the output space is a vector containing the allocation of each individual in one of the $\mathcal{C}_k$ class: $\mathfrak{Y} \equiv \mathcal{C}$. In sum, in the classification problem a categorical outcome is predicted using a set of features (or predictors). The present paper deals with a classification problem: predicting academic dropout of university students using a set of socio-economic variables.

The statistical learning literature presents a wide range of computational and statistical models to solve classification problems (Hastie et al., 2009). These models can be divided into two main categories: discriminative and generative (Ulusoy & Bishop, 2006, Mitchell, 1997). The discriminative approach for classification aims to learning (or discovering) the boundary dividing two or more classes. It focuses on directly estimating the conditional probability of an outcome or response C given the predictors X, so that: $f(X) = \arg\max_C P(C|X)$. In other words, the discriminative models directly model the class posterior

probability (Mitchell, 1997). Hence, the most likely class considering X is chosen. Generative models, by the other side, model the joint probability of features and the correspondent class: $P(C, X)$. Then, generative models perform the classification task by applying Bayes rule to compute the posterior probability of the response variable (Stephens et al., 2009). In the next paragraphs, we will present two models: the discriminative model of classification trees and the generative model known as Naïve Bayes.

Classification Trees:

As pointed before, the discriminative approach provides classification splits, given a set of preditors, in order to construct a boundary to separate two or more classes. One example of the discriminative approach is the classification trees (Breiman, Friedman, Olshen, & Stone, 1984). A classification tree partitions the feature space into several distinct mutually exclusive regions (Figure 1). Each region is fitted with a specific model that designates one of the classes to that particular space.

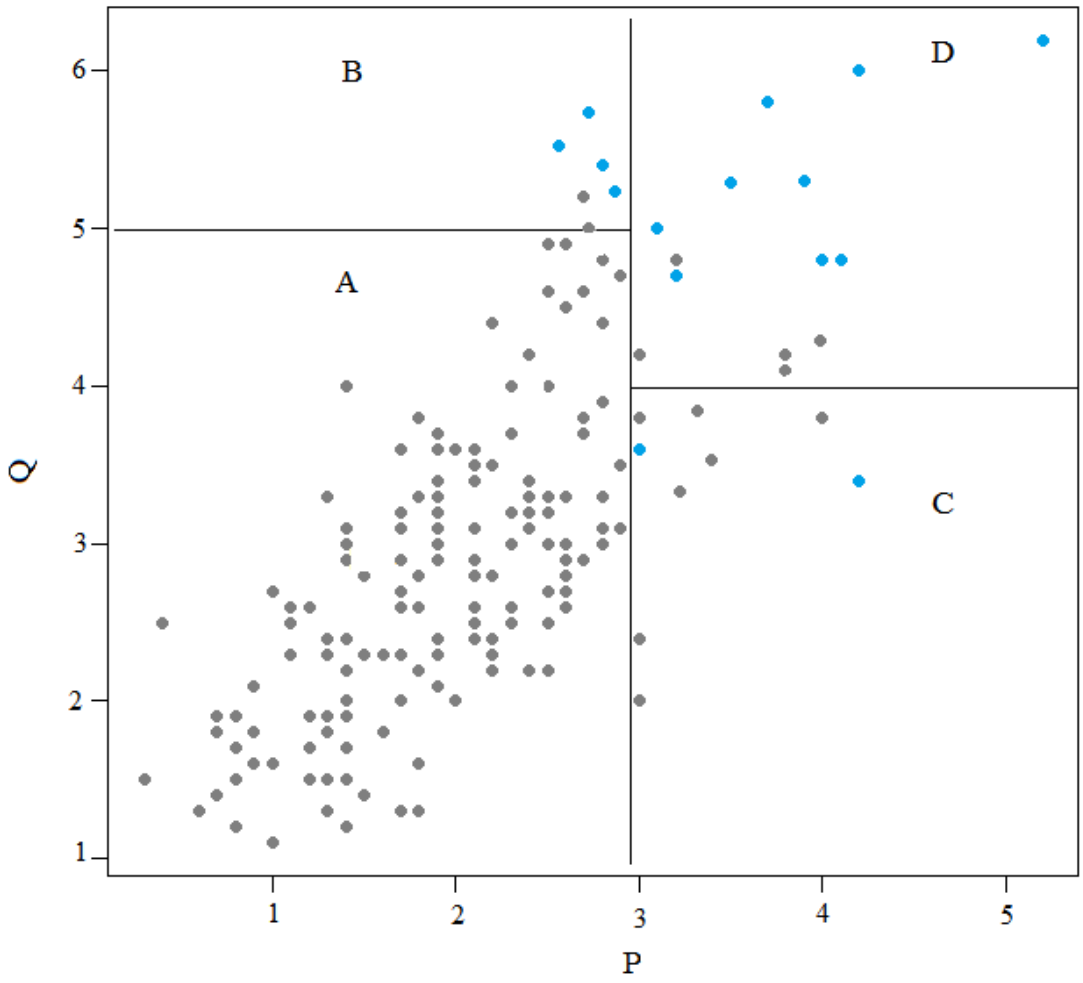


Figure 1. Partitioning of 2-dimensional feature space into four non-overlapping regions R_1, R_2, R_3 and R_4.

A class is assigned to the region of the feature space by identifying the majority class in that region. Figure 1 shows the distribution of observations in two variables, *P* and *Q*, colored according to two classes: C_1 (gray dots) and C_2 (blue dots). A person s_n with a value of *P* smaller than three and a value of *Q* smaller than five belongs to the region *A* of the feature space: $s_n \in A \leftrightarrow \{P < 3 \wedge Q < 5\}$. Since C_1 is the majority class in *A*, those falling within this region will be estimated as pertaining to the class C_1: $A \rightarrow \widehat{C_1}$. In the same line, people with a value of *P* smaller than three and a value of *Q* greater than five belongs to the region *B* of the feature space: $s_n \in B \leftrightarrow \{P < 3 \wedge Q > 5\}$. Since C_2 is the majority class in *B*, those falling within this region will be estimated as belonging to the class C_2: $B \rightarrow \widehat{C_2}$.

People with *P* greater than three and a value of *Q* smaller than four belongs to the region *C* and $C \rightarrow \widehat{C_1}$ because class C_1 is majoritary in *C*. Finally, those with *P* greater than three and *Q* greater than four belongs to the region *D*, $s_n \in D \leftrightarrow \{P > 3 \wedge Q > 4\}$, and $D \rightarrow \widehat{C_2}$. This set of if-then rules can be easily understood by inspecting the structure of the classification tree (see Figure 2).

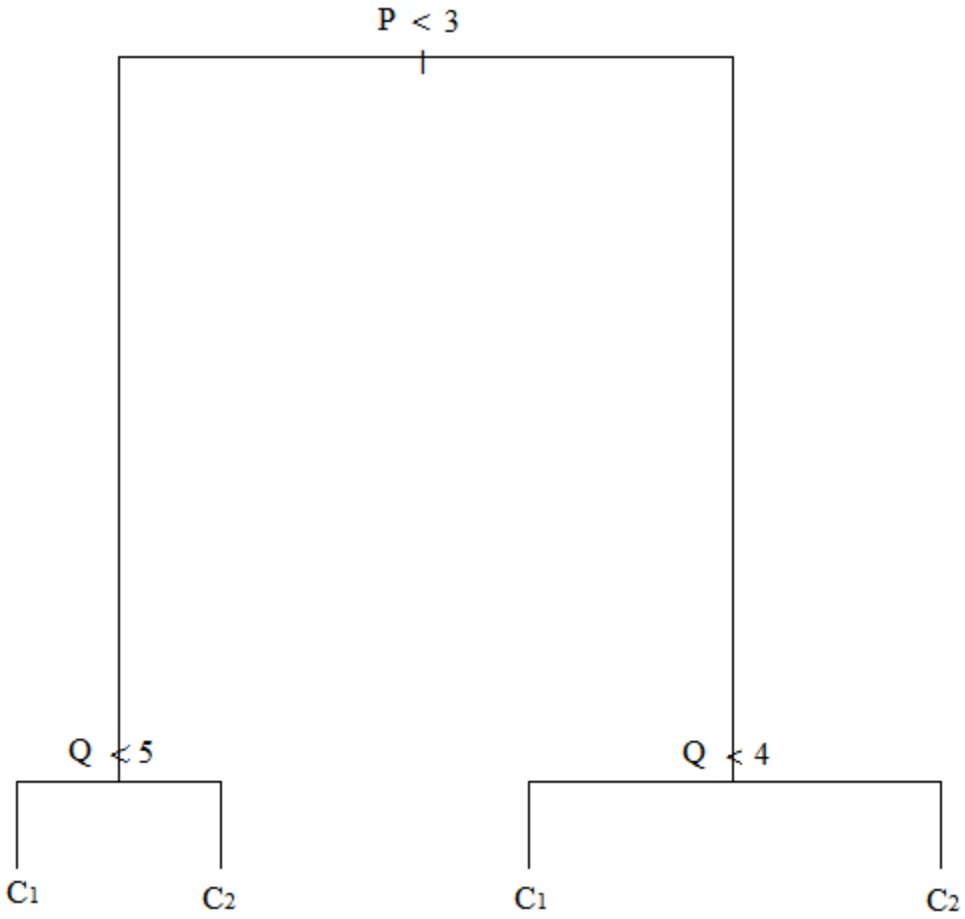


Figure 2. Classification Tree.

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, *dropped from university*. A node with 90% of students that dropped from university and 10% of regular enrolled 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 (predictor) that leads to the most purity configuration (top split in Figure 2). Then, each

region of the tree is modeled accordingly to the majority class. One or two original nodes are also split into more nodes, using some of the given predictors that provide the best fit possible (bottom split in Figure 2). This splitting process continues until the feature space achieves the most purity configuration possible, with R_m regions or nodes classified with a distinct C_k class. 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. In a classification tree, the first split indicates the most important variable, or feature, in the prediction.

The classification trees have two main basic tuning parameters (for more fine grained tuning parameters see Breiman, Friedman, Olshen & 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 $\alpha|T|$. Geurts, Irrthum and Wehenkel (2009) argue that classification 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, does not make any assumption regarding normality, linearity or independency and can be applied in datasets with a large *p* low *n* characteristic (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. Finally, the ease of use means that the tree based techniques are computationally simple, yet powerful.

Naïve Bayes:

Generative models are interested in the joint probability of features and the correspondent class, P(C,X), contrasting with the discriminative models that directly models the posterior probability P(C|X). Understanding the difference between the discriminative and the generative models involves understanding how P(C,X) and P(C|X) differ, and it can be done using a simple example. Consider four subjects S_n, n= (1,2,3,4), one predictor X_j, j

= {(1 = Parental pressure to choose a university carrier), (2 = No parental pressure to choose a university carrier)} and a class C_k, $k =$ {(0 = Enrolled in the university), (1 = Dropped from the university)}. The ordered pair (C, X) for each subject is: (C, X) = {(1,1), (1,1), (1,2), (0,2)}. Table 2 presents the data of our example. There are four possible pairs of C,X. In our example, there was no subject that dropped the university and did not feel parental pressure to choose a university carrier (C = 0, X = 1). Two subjects dropped and felt parental pressure (C = 1, X = 1), one dropped and did not feel parental pressure (C = 1, X = 2) and one that neither dropped nor felt parental pressure (C = 0, X = 2). Calculating the probability of each possible combination of (C,X) gives the picture presented in Table 2 at column P(C,X). This is an oversimplification of what happens when one uses a generative models. By contrast, if we estimate the probability of the response or outcome C given the predictors X, hence using the logic of the discriminative models, we arrive in a very different scenario (see the far right column in Table 2). The probability of dropping the university given a person felt parental pressure to choose a university carrier is $P(C = 1|X = 1) = 1$, because every person that felt parental pressure dropped from the university! In the same line, the probability of dropping from the university given the person did not feel parental pressure to pursue a university carrier is 50%, since one out of two have X = 2 and also have C = 1.

Table 2. Example of the difference between generative and discriminative models towards the probability estimation.

Subjects	C	X	(C, X)	Possible combinations of (C,X)	Number of occurrences for each combination of (C,X)	P(C,X)	P(C|X)
S_1	1	1	(1,1)	(0,1)	0	0	0
S_2	1	1	(1,1)	(1,1)	2	½	1
S_3	1	2	(1,2)	(1,2)	1	¼	½
S_4	0	2	(0,2)	(0,2)	1	¼	½

Directly modeling the class posterior probability, mapping the predictors into the outcome variable, yields different results than explicitly modeling the actual distribution of each class. As pointed before, generative models perform the classification task by applying Bayes rule to compute the posterior probability of the response variable assuming the independence of the predictors (Pernkopf & Bilmes, 2005). So, if we are interested in

predicting a particular class C_k, using a vector of predictors $\boldsymbol{X}$, $\boldsymbol{X} = \{X_1, X_2, \dots, X_m\}$, their joint probability distribution is:

$$P(C_k, \boldsymbol{X}) = P(C_k|\boldsymbol{X})P(\boldsymbol{X}).$$

Applying Bayes' rule, leads to:

$$\frac{P(\boldsymbol{X}|C_k)P(C_k)}{P(\boldsymbol{X})}$$

And assuming the predictors are conditionally independent given C_k, leads to the Naïve Bayes approximation (Stephens et al., 2009):

$$\prod_{m=1}^{M} \frac{P(X_m|C_k)P(C_k)}{P(X)}$$

Stephens et al. (2009) presented a test statistic to assess the statistical significance of $P(C_k|X_m)$. The relevance of assessing the statistical significance of the posterior probability is that probabilities does not accounts for sample size. Let us return to the example presented in Table 2. The probability of evading the university given a person felt parental pressure to choose a university carrier equals one. This can be a coincidence of only one case, as in Table 2, or can be a result of 1,000 observations. Obviously, the last case is more significant than the first! To solve this issue, let's consider that a class C_k, k = (1,0) is a random variable following a binomial distribution, so that every observation has a probability $P(C_k)$ of having a class C_k. The null hypothesis is that the distribution of the class C_k is independent of the predictors X_m, and is randomly distributed in the population. We can calculate the significance of the dependence of C_k on X_m, using the following equation:

$$\varepsilon(C_k|X_m) = \frac{N_{X_j}\big(P(C_k|X_m) - P(C_k)\big)}{\Big(N_{X_j}\big(P(C_k)(1 - P(C_k))\big)\Big)^{1/2}}$$

Where N_{X_j} is the total number of observations with $X_m = j$. The numerator is the difference between the actual number of co-occurrences of C_k and X_m, relative to the expected number if the class distribution were obtained by a binomial with sampling probability of $P(C_k)$. To put the numerator into appropriate units, the denominator of the

equation above is the standard deviations of the binomial distribution (Stephens et al., 2009). The values of $\varepsilon(C_k|X_m)$ can be interpreted as follows. Considering a normal approximation for the binomial distribution, $\varepsilon(C_k|X_m) = 2$ represents the 95% significance level. So, predictors with an epsilon equals to or greater than two are considered statistically significant.

In the last paragraph, we have presented a way to check the significance of a variable in the prediction of a discrete class. Now we present a score function to generate predictions given a set of predictors ***X'***, that works as a proxy for the Naïve Bayes approximation presented earlier:

$$S(C_k|X') = \sum_{m=1}^{N} S(C_k|X'_m) = \sum_{m=1}^{N} ln\left(\frac{P(X'_m|C_k)}{P(X'_m|\overline{C_k})}\right)$$

where $\overline{C_k}$ is the complement of the set C_k. An overall zero score $S(C_k|X')$ means that the probability to find C_k is the same as would be found if C_k were randomly distributed. In other words, a zero score indicates that conditioned on the predictor, the probability of C_k remains the same. A positive score means that the probability of C_k is higher than chance given the predictor X'_m. A negative score, by the other side, indicates the probability of C_k smaller than chance given the predictor. In other words, a positive score indicates a risk factor for the occurrence of the class C_k, while a negative score indicates a protective factor.

The Naïve Bayes algorithm is a classification algorithm based on Bayes rule, that assume the predictors as conditionally independent given the target outcome (C_k). This assumption is rarely met in reality, therefore the name "Naïve Bayes". However, even with this somehow unrealistic assumption, the Naïve Bayes classifier have shown to be very efficient in terms of prediction in different fields (Bishop, 2006; Stephens et al., 2009; Ulusoy & Bishop, 2006). Two of its greatest characteristics are that it is not sensitive to irrelevant features and it is a powerful tool to make inferences and to understand the data! Now that we have introduced the basic of two discriminative models, classification trees and random forests, and one generative model (Naïve Bayes), we will use them to predict evasion from University.

Methods:

Sample:

Composed the sample 1,318 Brazilian students (51.51% Male) that entered the Universidade Estadual de Santa Cruz (UESC), located in the state of Bahia, Brazil, in 2008. These students answered a socio-economic questionnaire when entering the University, and four years later its situation in the course was recorded as dropped or enrolled in the course. The distribution of students by each of the three main study fields offered by UESC is: exact sciences (22.76%), life sciences (25.79%) and humanities (51.44%). Nearly half of the UESC's vacancies are quotas to afrodescendants, native indians, direct descendants of slaves known as *Quilombolas* and students that completed their high school education in public schools. From the 1,318 students enrolled at UESC in 2008, 49.39% came from private high schools and competed for the regular vacancies (without quotas), 39.45% came from public high schools or are afrodescendants, 10.62% are native indians and only 0.5% are quilombolas.

Data analysis:

The sample was split into a training (70%) and testing set (30%). The training set was used in the learning phase, while the testing set was used to verify the quality of the classifier developed in the learning phase (cross-validation). The classification tree was applied using the package *tree* (Ripley, 2013) of the free and open source *R* software for statistical computing (R Development Core Team, 2011). The complexity of the tree, i.e. the number of terminal nodes was chosen using a 10-fold cross validation in the training set. Different numbers of terminal nodes is set, and its impact in the deviance index is recorded. The three complexity tuning parameter is set by the number of terminal nodes that leads to the smaller deviance in the 10-fold cross validation procedure. In order to calculate the deviance index, first is necessary to calculate the proportion of class $\mathcal{C}_k$ in a node m of the region R_m, with N_m people:

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

The labeling function that will assign a $\mathcal{C}_k$ class to a node m is: $max_{\mathcal{C}_k} \hat{p}_{m\mathcal{C}_k}$.

And the residual mean deviance is given by the following formulae:

$$-2\sum_{m}\sum_{\mathcal{C}_k} n_{m\mathcal{C}_k}\log\hat{p}_{m\mathcal{C}_k}$$

where $n_{m\mathcal{C}_k}$ is the number of people (or cases/observations) from the $\mathcal{C}_k$ class in the m region (James et al., 2013).

The Naïve Bayes approximation was implemented in R, and the test set predictions were made using the score presented in the last section. The quality of the three classifiers' prediction was assessed in the test set via total accuracy, sensitivity and specificity. Total accuracy is the proportion of observations correctly classified:

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

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. $\mathcal{C}_1 = evaded$, 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. $\mathcal{C}_2 = not\ evaded$, over the number of observations that belong to that class:

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

A receiver operating characteristic curve was plotted with the sensitivity and specificity of each classifier using the *pROC* package (Robin et al., 2011). The area under the curve was calculated and its 95% confidence interval was estimated using DeLong's method (DeLong, DeLong & Clarke-Pearson, 1988). The variable names, its meaning, as well as its posterior probability and the epsilon are displayed in the appendix A.

Results:

Discriminative Model:

The first classification tree, constructed without limiting the number of terminal nodes (tree size or complexity) resulted in a tree with 12 terminal nodes, generated from

splits over six predictors. The 10-fold cross validation showed that a tree with only four terminal nodes would lead to the lowest deviance (Figure 3).

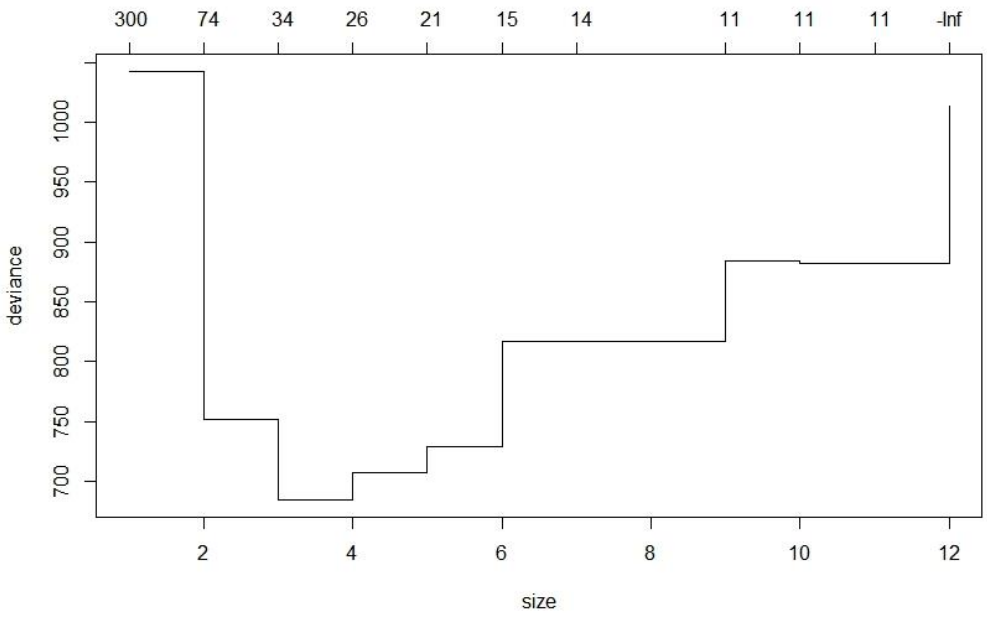


Figure 3. Effect of tree size (number of terminal nodes) in the deviance index, estimated using a 10-fold cross validation of the training set.

Thus, the tree was pruned in order to have only four terminal nodes, accordingly to the 10-fold cross validation result (Figure 3) and resulted in the structure showed in Figure 4. Only two variables were used to split the feature space: course and the pass rate's group (percentage of passing rate in all the semesters). Those in the first pass rate's group (mean of 26.74%) were classified as "dropped" (left branch of the tree in Figure 4). The remaining people in the second, third and fourth pass rate group were classified as follows. Those in the pass rate group three and four were classified as "enrolled", as well as those in the second group, irrespective of the course (a =ADT, BBI, BIO, BMA, CCO, COS, DRT, EFE, HIS, LBI, LEA, LEF, LEP, LFI, LIP, LMA, LQU, PDG; b = AGR, BFI, CIC, ECN, EPS, FLS, GEO, MEV).

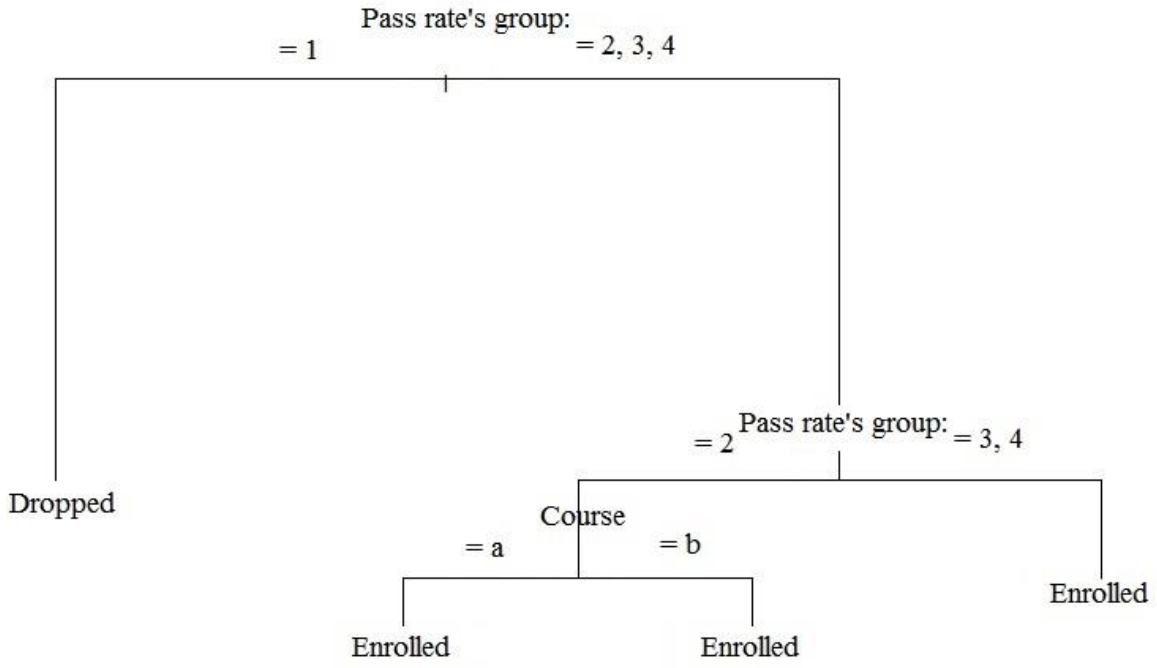


Figure 4. Classification tree with four terminal nodes, constructed after pruning the first tree.

Table 3. Tree splits and it's posterior probability.

Variables	Leaf (left to right)	n	Deviance	Majority Class	Splits cut left	Splits cut right	$P(C_{Dropped}\|X)$	$P(C_{Enrolled}\|X)$
Pass rate group		893	1035.46	Enrolled	= 1	=2, 3, 4	0.27	0.73
<leaf>	1	224	264.27	Dropped			0.72	0.28
Pass rate group		669	473.63	Enrolled	= 2	= 3, 4	0.11	0.89
Course		227	262.20	Enrolled	= a	= b	0.26	0.74
<leaf>	2	132	177.84	Enrolled			0.40	0.60
<leaf>	3	95	49.98	Enrolled			0.07	0.93
<leaf>	4	442	137.61	Enrolled			0.04	0.96

The tree classifier (Figure 4) was used to predict academic evasion in the test set. It resulted in a sensitivity of 70.96% and in a specificity of 90.75%. The area under the curve (AUC) was 80.64% (95% C.I.: 75.72% - 85.56%). The ROC curve and the 95% confidence interval are displayed in Figure 5.

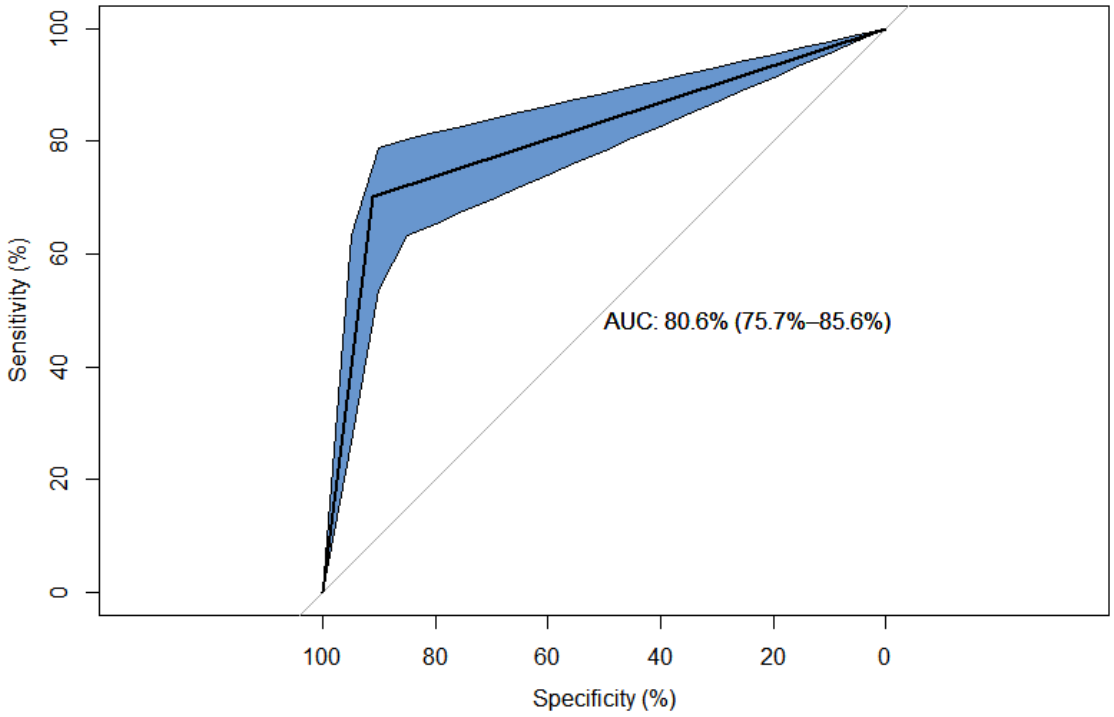


Figure 5. ROC curve of the tree classifier and its 95% confidence interval (blue).

Naïve Bayes:

The first step of the Naïve Bayes approach is to calculate the joint probability (C, X). Once this probability is calculated, the next step is to calculate the *Epsilon*, $\varepsilon(C_k|X_m)$, in order to discover the variables that significantly alter the probability of the class. Finally, the score $S(C_k|X')$ is calculated in order to function as a proxy of the Naïve Bayes approximation showed in the introduction. Appendix A shows the name of all the predictors (variables), its values, descriptions as well as its probabilities, scores $S(C_k|X')$ and epsilons.

The area under the curve (AUC) of the Naïve Bayes classifier was 74.8% (95% C.I.: 69.7% - 80.02%). The ROC curve and the 95% confidence interval are displayed in Figure 6.

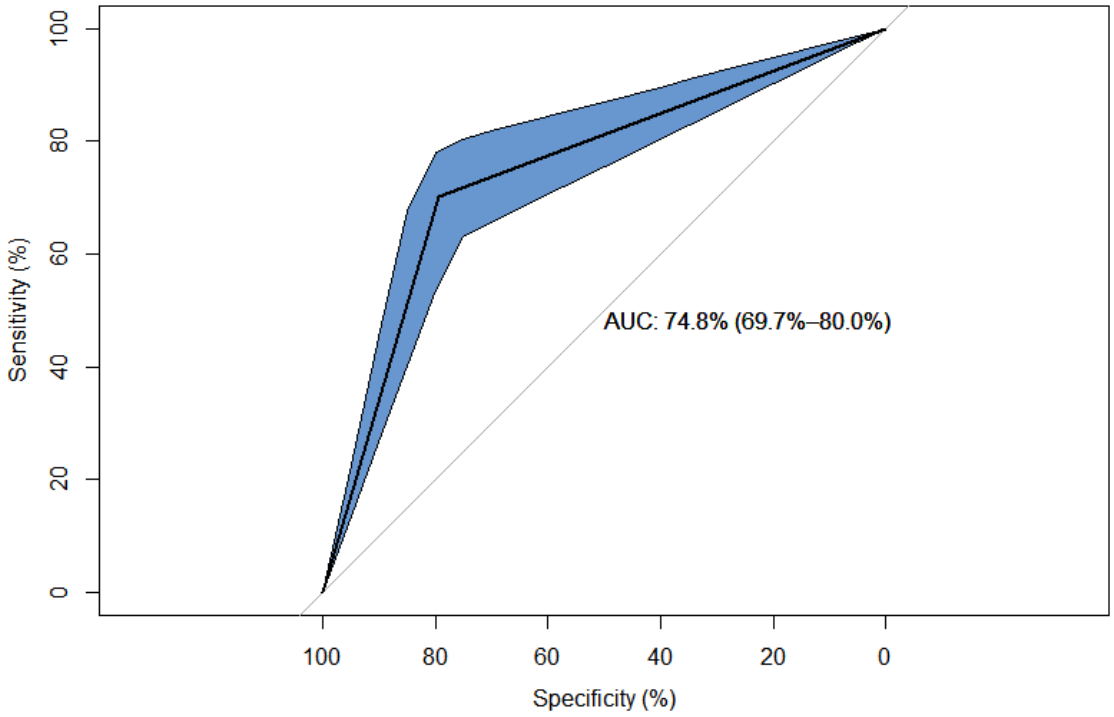


Figure 6. ROC curve of the Naïve Bayes classifier and its 95% confidence interval (blue).

Figure 6 shows the AUC from the learning tree classifier and from the Naïve Bayes classifier. There is no evidence to refute the hypothesis that both AUC are identical (D = 1.5904, p-value = 0.1121).

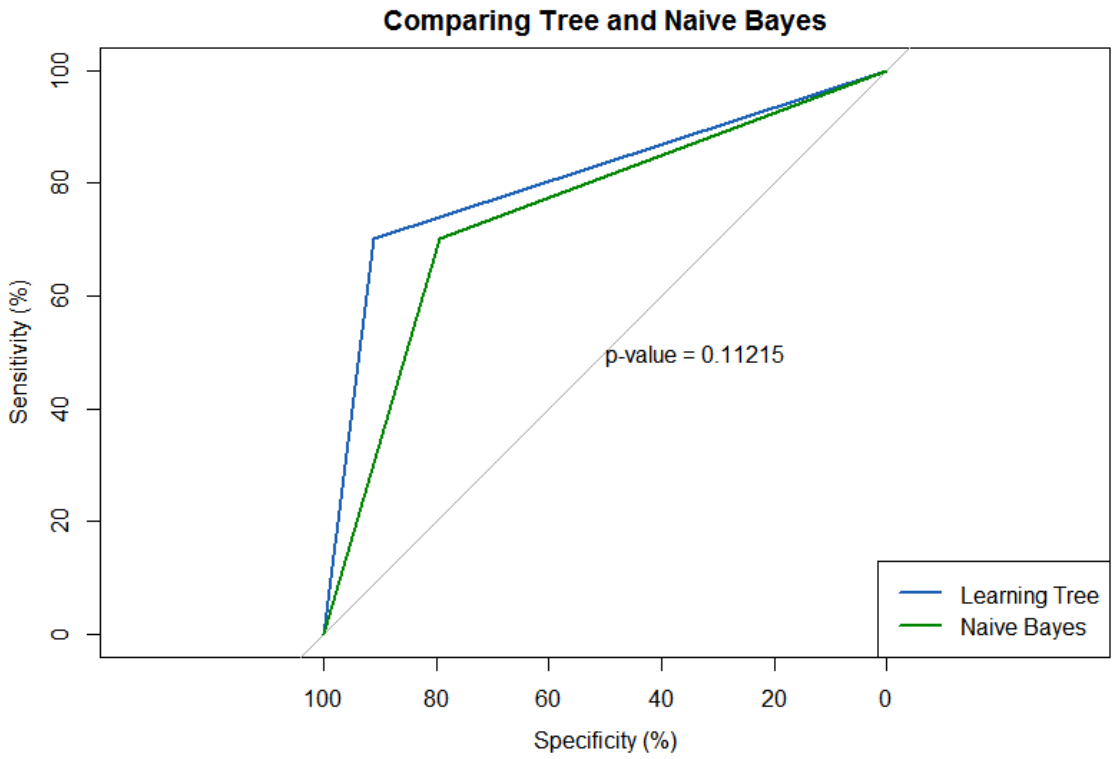


Figure 7. Comparing the AUC from the learning tree and from the Naïve Bayes classifiers.

The most important predictors of academic evasion, accordingly to the epsilon value are displayed in Table 5. Students in the first pass rate group (mean = 26.74%) have an epsilon of 13.89, with a probability of evading the university equals to 64%, against a 26% prior probability of academic evasion. Students enrolled in the Bachelor Degree in Mathematics have an epsilon of 7.11, with an 88% probability of dropping from the university. In general, those enrolled at exact sciences' courses have a 45% probability of dropout, with an epsilon of 6.73. Those enrolled in the *licenciatura* in Physics [P(C|X) = 0.73, Epsilon = 4.21] and in the *licenciatura* in Mathematics [P(C|X) = 0.57, Epsilon = 3.86] also have an dropout probability that is statistically significant at 95% confidence level. The same scenario is present to those who are married [P(C|X) = 0.40, Epsilon = 3.62], male [P(C|X) = 0.33, Epsilon = 3.38], enrolled at the Bachelor Degree in Physics [P(C|X) = 0.62, Epsilon = 2.94]. The students who did not attend a pre-university course before going through the university selection exam presented a dropout probability of 32% (Epsilon = 2.85). Students that are not afrodescendants presented 33% probability of dropping the university (Epsilon = 2.62), while those whose family's income are from eleven to twenty times the

minimum wage presented a probability of 40% (Epsilon = 2.61). Students whose father or responsible completed a college level had a dropout probability of 37% (Epsilon = 2.50), and those who entered the university without quotas had 31% probability of dropping out (Epsilon = 2.39). Students that had difficulty in being approved in the intended course presented a dropout probability of 54% (Epsilon = 2.31), those with two people living with the family income had a probability of 36% (Epsilon = 2.29), those who expect from the course only the college degree had a probability of 41% (Epsilon = 2.29) and those enrolled at the Philosophy course had an evasion probability of 40% (Epsilon = 2.17). Finally, presented a statistically significant probability of dropping out, at the 95% confidence level, students who were working part time [P(C|X) = 0.35, Epsilon = 2.15], enrolled in *licenciatura* in Chemistry [P(C|X) = 0.45, Epsilon = 4.11] and those whose high-school years were coursed mostly at private schools [P(C|X) = 0.41, Epsilon = 2.05]. Two other predictors were pretty close to the two standard deviation units required to statistical significance at the 95% confidence level: Applied to the university before to another university but were not approved [P(C|X) = 0.37, Epsilon = 1.94] and have 1 to 2 children [P(C|X) = 0.35, Epsilon = 1.92].

Table 4. Variable importance from the Naïve Bayes classifier (or risk factors)

Variable	Value	Description	P(C|X)	P(C)	Epsilon
Pass Rate	1	Group 1 (Mean = 26.74%)	0.64	0.26	13.89
Course	BMANA	Bachelor in Mathematics	0.88	0.26	7.11
Field	Exatas	Exact Sciences	0.45	0.26	6.73
Course	LFINA	Licenciatura in Physics	0.73	0.26	4.21
Course	LMANA	Licenciatura in Mathematics	0.57	0.26	3.86
Married	2	Yes	0.40	0.26	3.62
Sex	M	Male	0.33	0.26	3.38
Course	BFINA	Bachelor in Physics	0.62	0.26	2.94
Pre-University Course	1	No	0.32	0.26	2.85
Afrodescendants	1	No	0.33	0.26	2.62
Family income	7	From eleven to twenty times the minimum wage	0.40	0.26	2.61
Scholarity: father or responsible	7	Complete college level	0.37	0.26	2.50
Quota	1	No Quota	0.31	0.26	2.39
Why did you choose your course?	6	Difficulty passing the intended course	0.54	0.26	2.31
How many people live with your family's income?	2	2 people	0.36	0.26	2.29
What do you expect from your course?	2	College degree	0.41	0.26	2.29
Course	FLSNA	Philosophy	0.40	0.26	2.17
Are you currently working?	3	Yes, part time	0.35	0.26	2.15
Course	LQUNA	Licenciatura in Chemistry	0.45	0.26	2.11
High School Period	2	Mostly at private school	0.41	0.26	2.05
Have you ever applied to a University before?	3	Yes, Other, Not approved	0.37	0.26	1.94
Children	2	1 to 2	0.35	0.26	1.92

While Table 4 presented the most important predictors of academic dropout, or in other words the risk factors, Table 5 presents the protective factors. Students with significantly less probability of dropout were those in the third pass rate group [Mean = 92.13%, P(C|X) = 0.02, Epsilon = -8.61] and at the fourth passing rate group [Mean = 99.22%, P(C|X) = 0.05, Epsilon = -7.75], as well as those enrolled in Life Sciences courses [P(C|X) = 0.15, Epsilon = -3,91], females [P(C|X) = 0.19, Epsilon = -3.46], enrolled in medicine [P(C|X) = 0.00, Epsilon = -3.18], nursing [P(C|X) = 0.07, Epsilon = -2.77], law [P(C|X) = 0.11, Epsilon = -2.75], agronomy [P(C|X) = 0.06, Epsilon = -2.66] and in *licenciatura* in Pedagogy [P(C|X) = 0.13, Epsilon = -2.28]. The same occurred for those who entered the university through quotas for afrodescendants [P(C|X) = 0.21, Epsilon = -2.20],

Table 5. Protective factors of academic evasion.

Variable	Value	Description	P(C|X)	P(C)	Epsilon2
Pass Rate	3	Group 3 (Mean = 92,13%)	0,02	0,26	-8,61
Pass Rate	4	Group 4 (Mean = 99,22%)	0,05	0,26	-7,75
Field	Vida	Life Sciences	0,15	0,26	-3,91
Sex	F	Female	0,19	0,26	-3,46
Course	MED	Medicine	0,00	0,26	-3,18
Course	EFENA	Nursing	0,07	0,26	-2,77
Course	DRTNA	Law	0,11	0,26	-2,75
Course	AGRNA	Agronomy	0,06	0,26	-2,66
Course	PDGNA	Licenciatura in Pedagogy	0,13	0,26	-2,28
Quota	2	Public School OR Afrodescendants	0,21	0,26	-2,20
Which option do you use most often to stay informed?	1	TV	0,21	0,26	-2,11
Scholarity: mother or responsible	3	Complete primary education	0,13	0,26	-2,10
Have you ever applied to a University before?	2	Yes, UESC, Not approved	0,21	0,26	-2,05
Pre-University Course	3	Yes, Private	0,19	0,26	-1,98
Field	Humanas	Human Sciences	0,22	0,26	-1,94

Discussion

Discriminative machine learning classifiers directly model the posterior probability $P(C|X)$, creating a map from the predictors X to the class C (Ng & Jordan, 2002). On the other hand, generative models computes the joint probability of C and X, P(C, X), and uses Bayes' rule to calculate $P(C|X)$. Some authors argues the discriminative models are better to perform a classification problem (Vapnik, OLHAR LIVRO), while others have provided evidences that they are very similar in terms of accuracy (Ng & Jordan, 2002). In the present paper we compared the performance of a learning tree classifier against a Naïve Bayes model in the prediction of academic dropout. The tree's area under the curve was 80.64% (95% C.I.: 75.72% - 85.56%), while the Naïve Bayes' AUC was 74.8% (95% C.I.: 69.7% - 80.02%). Comparing both AUCs there was no evidence to refute the hypothesis that they are identical (D = 1.5904, p-value = 0.1121).

The solution that led to the lowest deviance in the pruned tree model presented the percentage of passing rate in all the semesters (pass rate) as the most important predictor of academic dropout. The posterior probability of dropping the university given the student was

in the first passing rate group was 72%. On the other hand, the probability of being enrolled in the last semester of the course (i.e. not dropping out) given the student was in the second, third or fourth pass rate groups was 89%. This means that students that presents lower academic achievement in general, represented by the percentage of passing rate in all the semesters, have higher probability of dropping out then the students with higher academic achievement. The tree classifier also pointed that given the second pass rate group, the course enrolled made no difference in the prediction: 60% of students from the courses *a* (a =ADT, BBI, BIO, BMA, CCO, COS, DRT, EFE, HIS, LBI, LEA, LEF, LEP, LFI, LIP, LMA, LQU, PDG) did not dropped the university, and 93% of students from the courses *b* (b = AGR, BFI, CIC, ECN, EPS, FLS, GEO, MEV) did not dropped the university. Of course the difference in the probabilities are huge (60% vs. 93%), meaning that a lot more students from courses *a* drop the university, comparing with students from courses *b*. However, the tree classifiers makes the prediction based on the majority class of the feature space region!

The Naïve Bayes classifier showed a similar scenario, but much more fine grained then the tree classifier. It was possible to identify 21 risk factors, i.e. predictors that significantly increases the dropout probability. In general, the risk factors point to students in the first pass rate group, enrolled at exact sciences' courses, married, with high monthly income, not afrodescendant, that did not attend a pre-university course and that did not enter the university through the quotas system. Also in the dropout risk are those who father or responsible have college education, who chose the undergraduate course because had difficulty to be approved in the intended course, who are enrolled in the Philosophy course, who are enrolled in the Chemistry course (*licenciatura*) and those who studied mostly at private school during the high school period. Finally, are at risk students that have applied to a university before but were not approved, as well as those with one or two children.

The results of our study showed that both discriminative and generative models were good at predicting academic dropout. The generative model of Naïve Bayes is a more fine grained method to understand the role of every variable in leveraging the probability of dropout. Thus, is a more informative technique to generate inferences regarding the phenomena studied. Discovering the risk factors of academic dropout can help the educational institutions to improve their selection methods, as well as to decrease the number

of dropouts by applying an intervention program directed to those in risk of dropping the university course.

References

Bishop, C.M. (2006). *Pattern Recognition and Machine Learning.* Springer-Verlag.

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

DeLong, E.R., DeLong, D.M., & Clarke-Pearson, D.L. (1988). Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics, 44*, 837-845.

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

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.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning:* Data Mining, Inference and Prediction (2nd Edition). New York, NY: Springer.

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

Mitchell, T. (1997). *Machine Learning*. McGraw-Hill.

Ng, A.Y, & Jordan, M. (2002). On Discriminative vs. Generative Classifiers: A comparison of logistic regression and Naive Bayes. *NIPS, 14*, 2002.

Pernkopf, F. & Bilmes, J. (2005). Discriminative versus generative parameter and structure learning of Bayesian Network Classifiers. International Conference on Machine Learning, 657-664.

R Development Core Team (2011). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

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

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics, 12*, 77-. DOI: 10.1186/1471-2105-12-77

Stephens, C.R., Heau, J.G., González, C., Ibarra-Cerdeña, C.N., Sánchez-Cordero, V., et al. (2009) Using Biotic Interaction Networks for Prediction in Biodiversity and Emerging Diseases. *PLoS ONE 4*(5): e5725. doi: 10.1371/journal.pone.0005725

Ulusoy, I. & Bishop, C.M. (2006). Comparison of Generative and Discriminative Techniques for Object Detection and Classification, Toward Category-Level Object Recognition. LNCS, 4170, pp 173-195, 2006. Available at http://research.microsoft.com/en-us/um/people/cmbishop/downloads/Bishop-Sicily-05.pdf

APENDIX A

Variable	Value	Description	Variable Name	P(C|X)	P(C)	P(X|C)	S(C|X)	Epsilon
Region	1	Urban	q3	0.26	0.26	0.97	0.04	0.46
Region	2	Rural	q3	0.15	0.26	0.03	-0.69	-1.85
Afrodescendants	1	No	q6	0.33	0.26	0.36	0.35	2.62*
Afrodescendants	2	Yes	q6	0.23	0.26	0.64	-0.15	-1.63
Married	1	No	q7	0.24	0.26	0.80	-0.12	-1.39
Married	2	Yes	q7	0.40	0.26	0.20	0.66	3.62*
Children	1	No	q8	0.25	0.26	0.85	-0.06	-0.73
Children	2	1 to 2	q8	0.35	0.26	0.13	0.43	1.92
Children	3	3 to 5	q8	0.30	0.26	0.01	0.21	0.30
Children	4	> 5	q8	0.50	0.26	0.01	1.06	1.11
Primary School Period	1	Mostly at public school	q11	0.33	0.26	0.11	0.36	1.55
Primary School Period	2	Entirely at public school	q11	0.22	0.26	0.47	-0.19	-1.79
Primary School Period	3	Mostly at private school	q11	0.33	0.26	0.09	0.36	1.36
Primary School Period	4	Entirely at private school	q11	0.28	0.26	0.33	0.12	0.93
High School Period	1	Mostly at public school	q12	0.34	0.26	0.07	0.39	1.32
High School Period	2	Mostly at private school	q12	0.41	0.26	0.06	0.70	2.05*
High School Period	3	Entirely at public school	q12	0.24	0.26	0.55	-0.12	-1.21
High School Period	4	Entirely at private school	q12	0.27	0.26	0.32	0.06	0.43
Pre-University Course	1	No	q15	0.32	0.26	0.52	0.31	2.85*
Pre-University Course	2	Yes, Public	q15	0.22	0.26	0.37	-0.19	-1.61
Pre-University Course	3	Yes, Private	q15	0.19	0.26	0.11	-0.42	-1.98
Pre-University Course Duration	1	0 to 6 months	q16	0.22	0.26	0.19	-0.18	-1.10
Pre-University Course Duration	2	7 to 12 months	q16	0.21	0.26	0.17	-0.26	-1.49
Pre-University Course Duration	3	13 to 24 months	q16	0.19	0.26	0.08	-0.39	-1.56
Pre-University Course Duration	4	25 to 36 months	q16	0.13	0.26	0.01	-0.89	-1.49
Pre-University Course Duration	5	> 36 months	q16	0.33	0.26	0.02	0.36	0.60
Pre-University Course Duration	N	No	q16	0.32	0.26	0.53	0.32	2.92*

Have you ever applied to a University before?	1	No, never applied before	q17	0.28	0.26	0.25	0.13	0.84
Have you ever applied to a University before?	2	Yes, UESC, Not approved	q17	0.21	0.26	0.36	-0.25	-2.05
Have you ever applied to a University before?	3	Yes, Other, Not approved	q17	0.37	0.26	0.10	0.50	1.94
Have you ever applied to a University before?	4	Yes, but never enrolled	q17	0.26	0.26	0.04	-0.01	-0.02
Have you ever applied to a University before?	5	Yes, but I'm changing the course	q17	0.26	0.26	0.05	-0.01	-0.04
Have you ever applied to a University before?	6	Yes, finished and want to do another course	q17	0.37	0.26	0.04	0.53	1.33
Have you ever applied to a University before?	7	Yes, as an experience	q17	0.30	0.26	0.14	0.20	0.98
Have you ever applied to a University before?	8	Yes, other motive	q17	0.19	0.26	0.02	-0.39	-0.71
How many times did you try to enter UESC?	1	0 times	q18	0.28	0.26	0.37	0.13	1.06
How many times did you try to enter UESC?	2	1 time	q18	0.27	0.26	0.30	0.06	0.44
How many times did you try to enter UESC?	3	2 times	q18	0.22	0.26	0.18	-0.18	-1.05
How many times did you try to enter UESC?	4	3 times	q18	0.26	0.26	0.10	-0.01	-0.06
How many times did you try to enter UESC?	5	4 times	q18	0.17	0.26	0.03	-0.52	-1.17
How many times did you try to enter UESC?	6	5 times	q18	0.24	0.26	0.02	-0.12	-0.22
How many times did you try to enter UESC?	7	> 5 times	q18	0.15	0.26	0.01	-0.65	-0.86
Why did you choose UESC?	1	Labour market guaranteed	q19	0.24	0.26	0.15	-0.11	-0.55
Why did you choose UESC?	2	Low demand and easy completion	q19	0.27	0.26	0.25	0.04	0.25
Why did you choose UESC?	3	Personal Affinity, vocation, personal achievement	q19	0.21	0.26	0.12	-0.25	-1.16
Why did you choose UESC?	4	Lack of required travel	q19					
Why did you choose UESC?	5	Permits reconciliation of the profession with good pay	q19	0.15	0.26	0.01	-0.65	-0.86

Why did you choose UESC?	6	Difficulty passing the intended course	q19	0.28	0.26	0.37	0.12	0.96
Why did you choose UESC?	7	Allows you to reconcile with other chores	q19	0.23	0.26	0.04	-0.18	-0.48
Why did you choose UESC?	8	Important for the development of the country	q19	0.24	0.26	0.02	-0.12	-0.22
Why did you choose UESC?	9	Other reason	q19	0.33	0.26	0.04	0.36	0.89
Why did you choose your course?	1	Labour market guaranteed	q20	0.32	0.26	0.09	0.32	1.20
Why did you choose your course?	2	Low demand and easy completion	q20	0.25	0.26	0.00	-0.04	-0.04
Why did you choose your course?	3	Personal Affinity, vocation, personal achievement	q20	0.25	0.26	0.70	-0.05	-0.55
Why did you choose your course?	4	Lack of required travel	q20	0.17	0.26	0.03	-0.55	-1.35
Why did you choose your course?	5	Permits reconciliation of the profession with good pay	q20	0.26	0.26	0.04	0.03	0.09
Why did you choose your course?	6	Difficulty passing the intended course	q20	0.54	0.26	0.03	1.21	2.31*
Why did you choose your course?	7	Allows you to reconcile with other chores	q20	0.40	0.26	0.03	0.65	1.26
Why did you choose your course?	8	Important for the development of the country	q20	0.22	0.26	0.03	-0.20	-0.49
Why did you choose your course?	9	Other reason	q20	0.29	0.26	0.05	0.14	0.41
What do you expect from your course?	1	Increased knowledge and general culture	q21	0.26	0.26	0.21	0.00	0.02
What do you expect from your course?	2	College degree	q21	0.41	0.26	0.07	0.71	2.29*
What do you expect from your course?	3	Critical consciousness that allows the interaction in society	q21	0.23	0.26	0.14	-0.14	-0.70
What do you expect from your course?	4	Theoretical training geared for future employment	q21	0.24	0.26	0.40	-0.09	-0.77
What do you expect from your course?	5	Higher activity level to improve the already developed	q21	0.30	0.26	0.07	0.19	0.64

What do you expect from your course?	6	Theoretical training focused on teaching and research	q21	0.26	0.26	0.06	0.03	0.09
What do you expect from your course?	7	Other things	q21	0.30	0.26	0.04	0.22	0.59
What most influenced your choice of course?	1	The family	q22	0.31	0.26	0.05	0.25	0.76
What most influenced your choice of course?	2	Friends and teachers	q22	0.19	0.26	0.03	-0.41	-0.91
What most influenced your choice of course?	3	Career counselor, vocational testing	q22	0.25	0.26	0.04	-0.04	-0.11
What most influenced your choice of course?	4	Information obtained by media	q22	0.22	0.26	0.07	-0.21	-0.75
What most influenced your choice of course?	5	personal interest in the course	q22	0.26	0.26	0.71	0.00	0.04
What most influenced your choice of course?	6	It is the only one possible to me	q22	0.37	0.26	0.03	0.52	1.10
What most influenced your choice of course?	7	I need to prepare for a job	q22	0.27	0.26	0.08	0.05	0.20
Are you decided or undecided with the course choice of made?	1	Decided	q23	0.25	0.26	0.93	-0.03	-0.36
Are you decided or undecided with the course choice of made?	2	Undecided	q23	0.36	0.26	0.07	0.50	1.60
Scholarity: father or responsible	1	illiterate	q24	0.22	0.26	0.03	-0.20	-0.49
Scholarity: father or responsible	2	Incomplete primary education	q24	0.23	0.26	0.25	-0.18	-1.21
Scholarity: father or responsible	3	Complete primary education	q24	0.25	0.26	0.08	-0.03	-0.10
Scholarity: father or responsible	4	Incomplete secondary education	q24	0.24	0.26	0.08	-0.11	-0.42
Scholarity: father or responsible	5	Complete secondary education	q24	0.27	0.26	0.33	0.06	0.42
Scholarity: father or responsible	6	Incomplete college level	q24	0.23	0.26	0.07	-0.17	-0.62
Scholarity: father or responsible	7	Complete college level	q24	0.37	0.26	0.16	0.50	2.50*

Scholarity: mother or responsible	1	illiterate	q25	0.30	0.26	0.03	0.19	0.45
Scholarity: mother or responsible	2	Incomplete primary education	q25	0.26	0.26	0.21	-0.02	-0.10
Scholarity: mother or responsible	3	Complete primary education	q25	0.13	0.26	0.03	-0.83	-2.10
Scholarity: mother or responsible	4	Incomplete secondary education	q25	0.28	0.26	0.07	0.10	0.37
Scholarity: mother or responsible	5	Complete secondary education	q25	0.26	0.26	0.37	0.01	0.10
Scholarity: mother or responsible	6	Incomplete college level	q25	0.26	0.26	0.10	0.00	0.00
Scholarity: mother or responsible	7	Complete college level	q25	0.29	0.26	0.18	0.14	0.78
Number of books read in the last year (not counting school books)	1	None	q26	0.35	0.26	0.03	0.45	0.89
Number of books read in the last year (not counting school books)	2	1 to 2	q26	0.25	0.26	0.19	-0.05	-0.29
Number of books read in the last year (not counting school books)	3	3 to 5	q26	0.26	0.26	0.45	0.02	0.14
Number of books read in the last year (not counting school books)	4	6 to 10	q26	0.23	0.26	0.18	-0.14	-0.84
Number of books read in the last year (not counting school books)	5	11 to 20	q26	0.32	0.26	0.11	0.31	1.29
Number of books read in the last year (not counting school books)	6	> 20	q26	0.23	0.26	0.05	-0.13	-0.38
Which option do you use most often to stay informed?	1	TV	q27	0.21	0.26	0.39	-0.25	-2.11
Which option do you use most often to stay informed?	2	Radio	q27	0.38	0.26	0.01	0.54	0.76
Which option do you use most often to stay informed?	3	Newspaper	q27	0.27	0.26	0.06	0.05	0.17

Which option do you use most often to stay informed?	4	Magazine	q27	0.29	0.26	0.14	0.16	0.80
Which option do you use most often to stay informed?	5	Conversations	q27	0.27	0.26	0.03	0.07	0.16
Which option do you use most often to stay informed?	6	Internet	q27	0.31	0.26	0.37	0.24	1.86
In addition to studies with which activities you most occupies your time?	1	Literature	q28	0.24	0.26	0.21	-0.08	-0.47
In addition to studies with which activities you most occupies your time?	2	Theater/Dance	q28	0.18	0.26	0.02	-0.45	-0.82
In addition to studies with which activities you most occupies your time?	3	Cinema/Videos	q28	0.23	0.26	0.10	-0.14	-0.58
In addition to studies with which activities you most occupies your time?	4	Sports	q28	0.30	0.26	0.19	0.23	1.31
In addition to studies with which activities you most occupies your time?	5	Religion	q28	0.25	0.26	0.12	-0.05	-0.25
In addition to studies with which activities you most occupies your time?	6	Music	q28	0.27	0.26	0.13	0.07	0.35
In addition to studies with which activities you most occupies your time?	7	Craft/Painting	q28	0.27	0.26	0.01	0.07	0.11
In addition to studies with which activities you most occupies your time?	8	Hanging out	q28	0.33	0.26	0.01	0.36	0.52
In addition to studies with which activities you most occupies your time?	9	Other	q28	0.25	0.26	0.21	-0.03	-0.18
Do you have internet access?	1	No	q29	0.34	0.26	0.05	0.41	1.11
Do you have internet access?	2	Yes, in my home	q29	0.29	0.26	0.57	0.14	1.35
Do you have internet access?	3	Yes, in other places	q29	0.22	0.26	0.38	-0.21	-1.76
Are you currently working?	1	No	q32	0.25	0.26	0.62	-0.06	-0.59
Are you currently working?	2	Yes, once in a while	q32	0.22	0.26	0.05	-0.22	-0.68

Are you currently working?	3	Yes, part time	q32	0.35	0.26	0.14	0.45	2.15
Are you currently working?	4	Yes, full time	q32	0.25	0.26	0.19	-0.02	-0.12
Registered Worker?	1	No	q33	0.25	0.26	0.77	-0.02	-0.24
Registered Worker?	2	Yes (max 1 year)	q33	0.21	0.26	0.05	-0.25	-0.80
Registered Worker?	3	Yes (from 1 to 3 years)	q33	0.28	0.26	0.08	0.13	0.48
Registered Worker?	4	Yes (from 3 years to 5 years)	q33	0.35	0.26	0.03	0.43	0.98
Registered Worker?	5	Yes (> 5 years)	q33	0.29	0.26	0.06	0.18	0.59
Does your family receive any Social Program help?	1	Yes	q35	0.26	0.26	0.88	-0.01	-0.11
Does your family receive any Social Program help?	2	Nobody	q35	0.27	0.26	0.12	0.07	0.30
How many people live with you?	1	Alone	q36	0.35	0.26	0.03	0.44	0.94
How many people live with you?	2	2 people	q36	0.31	0.26	0.29	0.25	1.71
How many people live with you?	3	3 people	q36	0.23	0.26	0.25	-0.17	-1.18
How many people live with you?	4	4 to 6 people	q36	0.25	0.26	0.39	-0.05	-0.45
How many people live with you?	5	7 or more people	q36	0.24	0.26	0.03	-0.12	-0.30
Family income	1	Half the minimum wage	q37	0.25	0.26	0.00	-0.04	-0.04
Family income	2	from half to one minimum wage	q37	0.35	0.26	0.06	0.44	1.33
Family income	3	from one to two minimum wages	q37	0.20	0.26	0.14	-0.32	-1.69
Family income	4	From two to three minimum wages	q37	0.25	0.26	0.24	-0.02	-0.13
Family income	5	From three to five minimum wages	q37	0.21	0.26	0.23	-0.28	-1.88
Family income	6	From five to ten minimum wages	q37	0.32	0.26	0.19	0.31	1.74
Family income	7	From eleven to twenty times the minimum wage	q37	0.40	0.26	0.11	0.65	2.61*
Family income	8	Over twenty minimum wages	q37	0.42	0.26	0.02	0.72	1.25
What's your role in your family's income?	1	Do not work, I get financial help from family	q38	0.25	0.26	0.62	-0.06	-0.64

What's your role in your family's income?	2	Work and receive financial help from family	q38	0.21	0.26	0.08	-0.26	-0.98
What's your role in your family's income?	3	Work and do not receive financial help from family	q38	0.27	0.26	0.04	0.06	0.17
What's your role in your family's income?	4	Work and contribute in part to the support of family	q38	0.29	0.26	0.18	0.17	0.93
What's your role in your family's income?	5	Work and I am the breadwinner of the family	q38	0.37	0.26	0.08	0.51	1.75
How many people live with your family's income?	1	1 person	q39	0.34	0.26	0.06	0.40	1.22
How many people live with your family's income?	2	2 people	q39	0.36	0.26	0.14	0.49	2.29*
How many people live with your family's income?	3	3 people	q39	0.25	0.26	0.20	-0.06	-0.34
How many people live with your family's income?	4	4 people	q39	0.25	0.26	0.31	-0.05	-0.39
How many people live with your family's income?	5	5 people	q39	0.20	0.26	0.15	-0.30	-1.62
How many people live with your family's income?	6	6 people	q39	0.28	0.26	0.11	0.12	0.54
How many people live with your family's income?	7	> 6 people	q39	0.23	0.26	0.03	-0.16	-0.40
Do you Intends to work while attenting the university?	1	No	q40	0.20	0.26	0.04	-0.30	-0.81
Do you Intends to work while attenting the university?	2	Yes, only basic internships	q40	0.25	0.26	0.33	-0.06	-0.46
Do you Intends to work while attenting the university?	3	Yes, in the last 2 years of University	q40	0.18	0.26	0.02	-0.45	-0.82
Do you Intends to work while attenting the university?	4	Yes, since the first year, but only part time	q40	0.28	0.26	0.44	0.12	1.03
Do you Intends to work while attenting the university?	5	Yes, since the first year, full time	q40	0.25	0.26	0.17	-0.03	-0.19
Housing	1	Owned	q41	0.25	0.26	0.78	-0.02	-0.20
Housing	2	Rented	q41	0.29	0.26	0.18	0.16	0.90
Housing	3	Other	q41	0.21	0.26	0.04	-0.28	-0.79
Main transport type	1	Bus	q42	0.26	0.26	0.77	0.01	0.15
Main transport type	2	Ride	q42	0.22	0.26	0.05	-0.20	-0.60

Main transport type	3	Motorcycle	q42	0.27	0.26	0.03	0.04	0.11
Main transport type	4	Car	q42	0.26	0.26	0.15	0.00	-0.02
Where are you going to live after entering the University?	1	With my parents	q43	0.25	0.26	0.67	-0.07	-0.74
Where are you going to live after entering the University?	2	With relatives	q43	0.28	0.26	0.07	0.13	0.45
Where are you going to live after entering the University?	3	Alone	q43	0.30	0.26	0.03	0.19	0.45
Where are you going to live after entering the University?	4	Pension	q43	0.28	0.26	0.18	0.10	0.55
Where are you going to live after entering the University?	5	Other	q43	0.36	0.26	0.04	0.47	1.20
Course Type	Bacharelado	Bachelor	Tipo	0.23	0.26	0.55	-0.15	-1.47
Course Type	Licenciatura	Licenciatura	Tipo	0.30	0.26	0.45	0.22	1.88
Shift	Diurno	Day	Turno	0.25	0.26	0.66	-0.07	-0.74
Shift	Noturno	Night	Turno	0.28	0.26	0.33	0.13	0.95
Course	ADTNA	Business Administration	Curso	0.24	0.26	0.06	-0.09	-0.29
Course	AGRNA	Agronomy	Curso	0.06	0.26	0.01	-1.72	-2.66
Course	BBINA	Bachelor in Biological Sciences	Curso	0.18	0.26	0.02	-0.45	-0.82
Course	BFINA	Bachelor in Physics	Curso	0.62	0.26	0.03	1.53	2.94*
Course	BIONA	Biomedicine	Curso	0.21	0.26	0.02	-0.27	-0.47
Course	BMANA	Bachelor in Mathematics	Curso	0.88	0.26	0.09	3.05	7.11*
Course	CCONA	Accounting	Curso	0.25	0.26	0.02	-0.04	-0.08
Course	CICNA	Computer Sciences	Curso	0.33	0.26	0.06	0.36	1.15
Course	COSNA	Media (Journalism)	Curso	0.26	0.26	0.04	0.03	0.07
Course	DRTNA	Law	Curso	0.11	0.26	0.03	-0.99	-2.75
Course	ECNNA	Economy	Curso	0.25	0.26	0.06	-0.04	-0.14
Course	EFENA	Nursing	Curso	0.07	0.26	0.01	-1.51	-2.77
Course	EPS	production and systems engineering	Curso	0.21	0.26	0.04	-0.28	-0.79
Course	FLSNA	Philosophy	Curso	0.40	0.26	0.08	0.65	2.17*
Course	GEONA	Geography	Curso	0.34	0.26	0.04	0.41	1.07
Course	HISNA	History	Curso	0.25	0.26	0.05	-0.04	-0.12
Course	LBINA	Licenciatura in Bio, Sciences	Curso	0.21	0.26	0.03	-0.24	-0.53

Course	LEA	Licenciatura in Foreign Languages	Curso	0.24	0.26	0.02	-0.12	-0.22
Course	LEF	Licenciatura in Physical Education	Curso	0.23	0.26	0.03	-0.13	-0.31
Course	LEPNA	Licenciatura in Spanish and Portuguese Languages	Curso	0.12	0.26	0.02	-0.96	-1.87
Course	LFINA	Licenciatura in Physics	Curso	0.73	0.26	0.05	2.07	4.21*
Course	LIPNA	Licenciatura in English and Portuguese Languages	Curso	0.26	0.26	0.02	0.03	0.05
Course	LMANA	Licenciatura in Mathematics	Curso	0.57	0.26	0.07	1.32	3.86*
Course	LQUNA	Licenciatura in Chemistry	Curso	0.45	0.26	0.04	0.87	2.11*
Course	MED	Medicine	Curso	0.00	0.26	0.00	NA	-3.18
Course	MEVNA	Veterinary Medicine	Curso	0.27	0.26	0.03	0.04	0.11
Course	PDGNA	Licenciatura in Pedagogy	Curso	0.13	0.26	0.03	-0.89	-2.28
Field	Exatas	Exact Sciences	Área	0.45	0.26	0.43	0.87	6.73*
Field	Humanas	Human Sciences	Área	0.22	0.26	0.42	-0.22	-1.94
Field	Vida	Life Sciences	Área	0.15	0.26	0.14	-0.71	-3.91
Sex	F	Female	SEXO	0.19	0.26	0.35	-0.42	-3.46
Sex	M	Male	SEXO	0.33	0.26	0.65	0.33	3.38*
Quota	1	No Quota	Cota	0.31	0.26	0.60	0.24	2.39*
Quota	2	Public School OR Afrodescendants	Cota	0.21	0.26	0.31	-0.29	-2.20
Quota	3	Indians (natives)	Cota	0.21	0.26	0.09	-0.26	-1.05
Quota	4	Quilombola	Cota	0.33	0.26	0.00	0.36	0.30
Pass Rate	1	Group 1 (Mean pass rate of 26,74%)	Pass rate	0.64	0.26	0.68	1.63	13.89*
Pass Rate	2	Group 2 (Mean pass rate of 69,97%)	Pass rate	0.23	0.26	0.25	-0.13	-0.87
Pass Rate	3	Group 3 (Mean pass rate of 92,13%)	Pass rate	0.02	0.26	0.02	-3.03	-8.61
Pass Rate	4	Group 4 (Mean pass rate of 99,22%)	Pass rate	0.05	0.26	0.05	-1.96	-7.75