STUDY ON SOCIOECONOMIC RELATIONS AND STUDENT PERFORMANCE ON THE ENADE 2018 EXAM

Maíra Matos Araújo[1], Caique Augusto Cardoso de Moraes [a], Jander Almeida Silva[a] , Lucas Souza Nogueira Santos[a] , Ruan Nilton dos Santos Azevedo[a] , Yasmim Thasla Santos Ferreira[a], Márcio Renê Brandão Soussa[a]

[a] SENAI CIMATEC University Center

Abstract: The National Student Performance Exam (ENADE) was proposed by INEP (The National Institute of Educational Studies and Research Anísio Teixeira) in 2004 with the aim of evaluating the performance of graduates of Brazilian Higher Education Institutions (HEIs). For this reason, the present work aims to understand possible correlations between socioeconomic, educational factors and student performance by applying data mining and visualization techniques on the 2018 ENADE dataset. The results obtained in this paper are similar to other works, suggesting that students who attended high school in private schools and/or study at a public college/university perform better on the exam. However, there are several points of improvement for future work.

Keywords: Data Mining, Enade, Classification Tree, Education.

ESTUDO SOBRE AS RELAÇÕES SOCIOECONÔMICAS E O DESEMPENHO DOS ALUNOS NA PROVA DO ENADE 2018

Resumo: O Exame Nacional de Desempenho dos Estudantes (ENADE) foi proposto pelo INEP (Instituto Nacional de Pesquisas Educacionais Anísio Teixeira) em 2004 com o foco de avaliar o desempenho dos graduandos das Instituições de Ensino Superior (IES) brasileiras. Por esse motivo, o presente trabalho tem como objetivo compreender as possíveis correlações entre fatores socioeconômicos, educacionais e o desempenho dos alunos, por meio da aplicação de técnica de mineração de dados e visualização de dados no conjunto de dados do ENADE de 2018. Os resultados obtidos neste artigo se assimilam a outros trabalhos,onde alunos que cursaram o ensino médio em uma escola particular e/ou estudam em uma faculdade pública possuem um desempenho melhor no ENADE. Contudo, existem vários pontos de melhoria para trabalhos futuros.

Palavras-chave: Data Mining, Enade, Classification Tree, Education.

1. INTRODUCTION

Higher Education Institutions (HEIs) play an important role in the social growth and economic development of societies [1], as they are responsible for training professionals committed to ethical, social and economic aspects [2] and, therefore, the Ministry of Education created metrics to assess the quality of teaching at HEIs in Brazil.

In this context, in 2004, the National Institute of Educational Studies and Research Anísio Teixeira (INEP) instituted the National Student Performance Examination (ENADE), composing the National Higher Education Assessment System (SINAES) [3], which aims at assessing the quality of undergraduate courses in public and private Higher Education Institutions in Brazil [4]. Since then, ENADE has become a mandatory curriculum component in undergraduate courses, as determined by Law No. 10,861/2004 [5].

ENADE occurs annually, however, each year it is applied only to graduates of a group of undergraduate courses that make up a framework area. With each application of ENADE, socioeconomic and performance information is collected from participants, as well as information about educational institutions, forming a large database that can be used as study material to identify and understand possible important relationships between socioeconomic and student performance.

Therefore, the use of data analysis tools and algorithms is essential, as can be seen in the project by Machado and Francelino [6], who applied the k-means algorithm on the data of 2013 computing course and the results suggest a relationship between family income and parents' education with student performance, so that the groups that performed better are associated with higher family income and parents' education level.

Nogueira and Tsunoda [7] applied the C4.5 decision tree algorithm to the 2012 database and concluded that economic aspects influence more student performance than ethnic-racial elements, so they developed a discussion about the adoption of affirmative action policies related to social quotas in public universities based on income and scholarship opportunities in private universities.

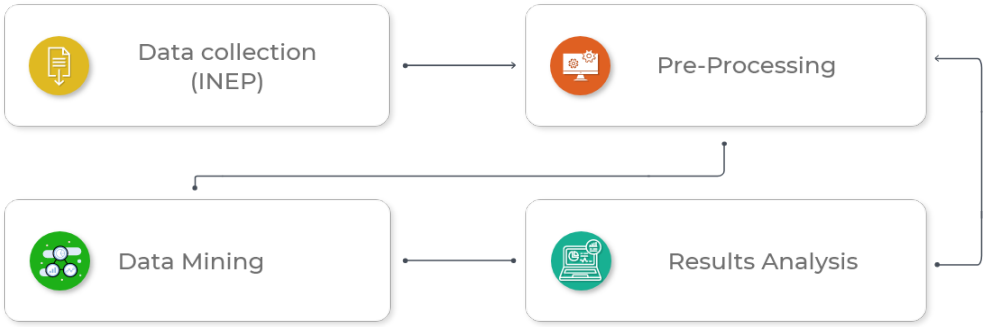
Notwithstanding the work presented above, there is still a need to continually seek new data analysis strategies that can contribute to improve the quality of higher education in Brazil and, therefore, this work aims to apply the Classification. Regression Tree (CART) algorithm, proposed by [8] on an ENADE database to investigate possible correlations between socioeconomic and educational factors and student performance, presenting the results through an easy-to-interpret dashboard tool.

2. METHODOLOGY

The methodological path of this work can be seen in the schematic drawing in Figure 1. It started with obtaining the ENADE database, then a pre-processing step

was carried out, in sequence, the application of the data mining technique , resulting in a classifier (model) and finally, an analysis of the results was performed.

Figure 1. Project execution steps



Source: Own

2.1 Obtaining the Database

The database chosen for this work was ENADE 2018, freely available in CSV (Comma-Separated Values) format on the INEP website [9]. Originally, the database contained 1.2 million records and 137 columns (variables), with information regarding higher education institutions, courses and students.

2.2 Pre-processing

During the pre-processing stage, some tasks were carried out, such as the elimination of records of students who abstained from taking the exam and invalid records, reducing the number of records in the database to 397,928.

Then, it was necessary to carry out some transformations in the database, including creating a categorical variable (CATEGORY), categorizing student performance as follows: category A for those with a performance < 25; category B for those who scored between 25 (inclusive) and 50; category C, between 50 (inclusive) and 75 and category D, for grades greater than or equal to 75 In Table 1, it is possible to visualize the structure of the database used.

Table 1. Database Structure

Variable Name	Description	Categories
CATEGORIA	**Category based on student performance**	A - overall grade less than 25 B - overall score between 25 (inclusive) and 50 C - overall score between 50 (inclusive) and 75 D - overall score greater than or equal to 75

NOTA_GERAL	**Overall grade - Weighted average of general training (25%) and specific component (75%) (value from 0 to 100)**	Min = 0 Max = 93.7 (Maximum for the year 2018)
MODALIDADE	**Teaching Mode Code**	Classroom teaching or E-learning
UF	**Federative Unit of the Course**	Acre (AC) , Alagoas (AL), Amapá (AP), Amazonas (AM), Bahia (BA), Ceará (CE), Distrito Federal (DF), Espírito Santo (ES), Goiás (GO), Maranhão (MA), Mato Grosso (MT), Mato Grosso do Sul (MS), Minas Gerais (MG), Pará (PA), Paraíba (PB), Paraná (PR), Pernambuco (PE), Piauí (PI), Rio de Janeiro (RJ), Rio Grande do Norte (RN), Rio Grande do Sul (RS), Rondônia (RO), Roraima (RR), Santa Catarina (SC), São Paulo (SP), Tocantins (TO)
REGIÃO	**Course operating region**	Região Norte (NO), Região Nordeste (NE), Região Sudeste (SE), Região Sul (SUL), Região Centro-Oeste (CO)
SEXO	**Sex**	Male(M) or Female(F)
IDADE	**Age of registrant on 11/25/2018**	Age of enrollee on 11/25/2018
ETNIA	**Color or race?**	White, Black, Yellow, Brown, Indigenous or not declared.
RENDA	**Total income of your family**	Up to 1.5 minimum wage (up to R$1,431.00).
		From 1.5 to 3 minimum wages (R$ 1,431.01 to R$ 2,862.00).
		From 3 to 4.5 minimum wages (R$2,862.01 to R$4,293.00).
		From 4.5 to 6 minimum wages (R$ 4,293.01 to R$ 5,724.00).
		From 6 to 10 minimum wages (R$ 5,724.01 to R$ 9,540.00).
		From 10 to 30 minimum wages (R$ 9,540.01 to R$ 28,620.00).
		Above 30 minimum wages (more than R$28,620.00).

ISSN: 2357-7592

SITUACAO_TRABALHO	**Work situation (except internship or scholarships)**	I'm not working.
		Work eventually.
		I work up to 20 hours a week.
		I work from 21 to 39 hours a week.
		I work 40 hours a week or more.
TP_BOLSA	**Type of scholarship or course funding you received to cover all or most tuition fees**	None, my course is free.
		None, although my course is not free.
		ProUni integral.
		Partial ProUni only.
		FIES, only.
		ProUni Partial and FIES.
		Scholarship offered by state, district or municipal government.
		Scholarship offered by the institution itself.
		Scholarship offered by another entity (company, NGO, other).
		Financing offered by the institution itself.
		Bank financing.
TP_ENSINO_MEDIO	**Type of high school**	All in public school.
		All in private school (private).
		All abroad.
		Mostly in public school.
		Mostly in a private (private) school.
		Part in Brazil and part abroad.
SUPERIOR_FAMILIA	**Some family member finished college**	Yes
		No
HORAS_ESTUDO	**Hours per week dedicated to study**	None, I just attend classes.
		From one to three.
		From four to seven.
		From eight to twelve.
		More than twelve.

2.3 Data Mining

ISSN: 2357-7592

This step aimed to create a classification model that would allow identifying possible relationships between socioeconomic data of students and educational institutions with the students' performance. For this, we chose to use a supervised technique, Classification and Regression Tree (CART) algorithm, proposed by [8], implemented in Python, using the Sklearn library.

After a careful analysis of the attributes present in the database (Table 1), the variable CATEGORY was selected as the dependent variable and as predictor variables (independent) the following attributes were selected: gender, age, income, employment status, hours of study, overall grade, family's level of education and type of highschool of the student as well as federative unit, type of scholarship, region of the course and modality of the course. The other columns were discarded for containing redundant information or of less relevance to the model.

Some versions of the tree were created, varying the depth, selected attributes and division of the training and test data in order to achieve the best accuracy. The field importance values were also obtained, a value that denominates the weight of the field for the model.

3. RESULTS AND DISCUSSION

During the first simulations with the algorithm, it was noticed that the occurrence of categories was very unbalanced, that is, categories A and D appeared in a much smaller number than categories B and C, causing classes A and D to remain few and therefore, producing inaccurate results. To mitigate this problem, an attempt was made to change the weight of the categories, but even so, the models produced showed low precision, classifying, for example, more than 80% of students belonging to classes A and D with false positives. Finally, it was decided to generate two classification models (two trees), one for categories A and D and the other for categories B and C because both pairs have similar amounts of data, seeking to increase the accuracy of the models.

Tree 1 comprises students with categories A and D and presented an accuracy of 77%. In tree 1, category A had as its main profile people who do not study in public colleges, studied in non-private schools, do not participate in the PROUNI scholarship program, in addition to not earning between 10 and 30 minimum wages. Class A students who attend a public college have an income of less than 1.5 minimum wages and work 40 hours a week or more. public school, study more than one hour a week and earn between 10 and 30 minimum wages, while the second profile studied in private school and public college, do not work more than 40 hours a week and do not have an income of less than 1.5 minimum wages. Tree 2 (Figures 5 and 6) comprises students from categories B and C and presented an accuracy of 60%, having 4 levels of depth.

Class B students had only one profile found, having students who attended secondary schools in schools outside the private network, study at private colleges and do not have an income between 10 and 30 minimum wages. However, students in class C attend a private school and study at a public college, do not work more than 40 hours per week and do not have an income lower than 1.5 minimum wages

ISSN: 2357-7592

for the most part. Another very important profile for class C are students who studied in private schools and attend public college, work more than 40 hours a week and earn between 10 and 30 minimum wages. As can be seen, both trees 1 and 2 had as the most important features: studied in private schools, studying in a public college and family income. In Table 3 we can see the distribution of the profiles more clearly.

Table 2. Student Profiles

Tree	Category	Profile
1	A (Note < 25)	1. Studied at a school outside the private network 2. Did not study in private college 3. Not participating in the PROUNI scholarship program 4. Does not earn between 10 and 30 minimum wages
		1. Study at public college 2. Studied at a school outside the private network 3. Works 40 hours a week or + 4. Income less than 1.5 minimum wages
	D (Note > =75)	1. Study in private 2. Go to college outside the public network 3. Study more than an hour a week 4. Has income between 10 and 30 minimum wages
		1. Study in private 2. Study at public college 3. Do not work more than 40 hours a week 4. Do not have an income of less than 1.5 minimum wages
2	B (25 <= Note < 50)	1. Studied at a school outside the private network 2. Study at a private college 3. Do not receive scholarships 4. Does not earn between 10 and 30 minimum wages
	C (50 <= Note <75)	1. Study in a private network 2. Study at public college 3. Do not work 40 hours a week or more 4. Do not earn less than 1.5 minimum wages
		1. Study in a private network 2. Study at public college 3. Do not work 40 hours a week or more 4. Do not earn less than 1.5 minimum wages

Finally, Recall values were measured (the frequency at which the model encountered a certain class); Accuracy, which is the ratio between the number of True Positive by the sum of False Positive and True Positive; and F1-Score, which

ISSN: 2357-7592

represents the relationship between Recall and Accuracy. These model evaluation metrics are expressed in Table 2.

The accuracy of each class is influenced both by the diversity of student profiles and the amount of information. It can also be observed that the model classifies many false positives for classes with a very large number of data. The variables with the greatest impact on the models were public college (0.24), prouni scholarship (0.19) and study at a private school (0.42). The presence of these fields mark the main nodes of the trees.

Table 3. Classification Report

	Tree 1		**Tree 2**	
	Note < 25 (A)	**Note > 75 (D)**	**25 <= Note < 50(C)**	**50<= Note < 75(B)**
Precision	0.82	0.69	0.65	0.57
F1-Score	0.73	0.78	0.65	0.56
Recall	0.74	0.77	0.66	0.55

In order to present the data in a clear, objective and easy-to-interpret manner, a dashboard tool was developed, whose main page can be seen in Figure 2. It allows the visualization of general data, i.e., of all courses, as well as allowing the desired courses to be filtered. The information is presented through cards, graphs and interactive classification trees, and in addition, it is possible to generate custom reports in CSV or XLS format.

ISSN: 2357-7592

Figure 2. Dashboard sections that demonstrate interactive graphs and classifications trees, generated from the data obtained and analyzed.

Source: Own

4. CONCLUSION

The main justification of this project is to make the ENADE feedback data accessible to any public, serving as an aid to government agencies and education professionals in the process of identifying points for improvement in the application of the test, with the aim of improving student performance, resulting in an improvement in the country's education.

According to INEP's president during a press conference in 2019, he highlights: "If we dive into the studies of returns, we can make important advances. It is important to go deeper into these numbers with the objective of improving higher education in Brazil". Highlighting the relevance of the Enade results for the improvement of higher education.

The developed application should serve as a national aid tool in the preparation and subsequent study of the application of ENADE tests and assessments, since it proposes to provide a coherent data analysis based on the data mining technique implemented in this work. The aim is to positively influence the quality evaluation methods of Higher Education Institutions in Brazil, and thus be a vector of improvement for Brazilian education.

With this, it can be concluded that, from the point of view of a public policy, this is the most correct measure to benefit students with an inclusion policy to increase the number of professionals with higher education WAINER, Jacques[8].

5. REFERENCES

[1] FERREIRA, E. C. C.. The importance of Higher Education Institutions in Regional Development in Portugal. **Évora**, March, 2019. <https://www.redalyc.org/pdf/2030/203016893005.pdf>

[2] REIS, A.L.; BANDOS, MF C. The Social Responsibility of Higher Education Institutions: A Systemic Reflection for Development. **Revista Gestão & Ciência,** Special Edition, Nov, 2012.

[3] BRITO, M. R. F. de. SINAES and ENADE: from conception to implementation. **Evaluation: Journal of Evaluation of Higher Education (Campinas)**, vol. 13, no. 3, p. 841-850, 2008.

[4] Inep. Federal Government, 2018. **National Student Performance Examination (Enade)**. Available at: <https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enade>. Accessed on: December 2, 2020.

[5] BRAZIL. Presidency of the Republic, Civil House. Deputy Chief of Legal Affairs. **Establishes the National Higher Education Assessment System – SINAES and other measures**. Brasília, DF, 2004.

[6] MACHADO, L. S.; FRANCELINO, W. L.. Data mining in Enade computing microdata. **Computer Science-Shark**, 2020.

[7] ANDRINO, E. D. N.; TSUNODA, Denise Fukumi. Data Mining To Analyze The Relationship Between The Socio-economic Characteristics Of Higher Education Graduates And The Performance Of These Students At Enade 2012. **Revista Percurso**, vol. 15, no. 1, 2015.

[8] BREIMAN, L.et al. **Classification and Regression Trees**. 1984.

[9] Inep. Federal Government, 2020. **Microdata from the National Student Performance Exam**. Available at: <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enade>. Accessed on: December 2, 2020.

ISSN: 2357-7592