https://dx.doi.org/10.22161/ijaers.6769 ISSN: 2349-6495(P) | 2456-1908(O)

Investigating the Efficacy of Brazilian Public Policies for Ethnic-Racial issues in Higher Education: The Case of Tocantins State in ENADE 2014

Carlos Barros, Tayla Pinto, George Brito, Andreas Kneip and David Prata

Department of PPGMCS, Federal University of Tocantins, BRAZIL

Abstract*— *This paper aims to investigate the demand for public policies in the light of data science evidence, namely, from a technical perspective. In order to reach this goal, we analyzed the data set of the Brazilian government on the ENADE, the national examination of student performance, for 2014, considering socioeconomic and ethnic-racial issues. The results showed that socioeconomic factors could influence the student performance in the tests. However, from the point of view of ethnic-racial elements, it was not possible to perceive statistically significant evidences that could influence the student's performance in the ENADE exam. The result suggests that there is no differences in intellectual capacity between the race ethnicities studied.

***Keywords*— *education, public, policies, data mining, ENADE*.**

I. INTRODUCTION

The continuous and fast technological advances have contributed to the acquisition and increase in the volume of data, from different segments, stored in databases, data warehouse and other types of data repositories. Potential valuable information is hidden in these data (PANDA e PATRA, 2008). The huge volume makes it very difficult, if not impossible for humans, to extract information without the help of appropriate computational tools (ZHOU, 2003). Thus, the so-called data science is dedicated to the extraction of knowledge from large databases with the help of computational tools, to deals with these situations.

Data science uses information technology techniques to identify useful information in large masses of data and may be applied in information management, information retrieval processing, decision making, process controls and in several other applications (GOLDSCHMIDT and PASSOS, 2005). The data science process was normalized by a group of researchers in the late 1980s (DIAS, 2008), and can be described, according to Fayyad (FAYYAD et al., 1996), as "a multi-step, non-trivial, interactive and iterative process for identifying understandable, valid, new, and possibly useful patterns from large data sets."

The area of Education can benefit from this process, due to the generation of a great amount of information. In Brazil, INEP (National Institute of Educational Studies and Research Anísio Teixeira) conducts regular evaluations on the quality of education in the country, from basic education to higher education. In this way, it has several open databases available for downloads, with microdata generated by these evaluations, that can be explored by education researches and by investigatory entities (INEP, 2016).

The constant evaluations carried out by INEP on education in Brazil, generates bases with giant volumes of data, which are constantly publicized. These data can be used for specific studies generating information with a high degree of relevance. Therefore, INEP data was analyzed concerning on the performance of undergraduate students in the State of Tocantins for the ENADE (National Exam of Student Performance) 2014 exam.

This article intends to first present the material and methodology applied to operate the data science on the ENADE database and, then, to analyze the results of the study at the final topic.

II. ENADE

In recent years, higher education in Brazil have been widely expanded, increasing the number of institutions and courses, and the access of the students to the universities has been facilitated. The Institutions of Higher Education had a significant increase in the number of students, and this growth can have consequences, for example, may affect the quality of teaching. Such issues

https://dx.doi.org/10.22161/ijaers.6769 ISSN: 2349-6495(P) | 2456-1908(O)

show the importance of conducting researches that may contribute to the advancement of higher education, especially in what concerns the quality of teaching, enabling universities to meet the greater demands of country education.

INEP conducts periodic research on education in Brazil; one of these surveys is ENADE. This examination is one of the procedures of evaluation of the National System of Evaluation of Higher Education (Sinaes), carried out by INEP; according to guidelines established by the National Commission for the Evaluation of Higher Education (Conaes), a collegiate body for coordination and supervision of Sinaes (INEP, 2016).

The purpose of ENADE is to follow the students' learning process and academic performance in relation to the programmatic contents provided in the curricular guidelines of the respective undergraduate course. The skills has to be adjusted to the requirements arising from the evolution of knowledge and their competences to understand subjects outside the specific scope of profession, linked to the Brazilian and worldwide reality, and other areas of knowledge.

The ENADE surveys help to formulate the objective of this work, that is, to use the ENADE database to verify the performance of undergraduate students from the State of Tocantins in this evaluation, and to discover if socioeconomic factors, such as family income, and ethnic-racial factors could influence the performance of these students on the exam.

III. MATERIALS

The ENADE microdata are available for consultation on the INEP Portal, and can be downloaded via ASCII format. The exam consists of general training issues (common to all areas of knowledge) and specific training issues. For the all knowledge areas evaluated in ENADE 2014, the analysis considered only records that corresponded to students who took the test in the state of Tocantins.

The socioeconomic survey questionnaire consists of 81 questions, of which thirteen are exclusive questions for undergraduate students. The "gross test grade", the "nt_ger" attribute, can vary from 0 to 100, which is the weighted average of the general training grade (25%) and the specific component grade (75%).

Data preparation was performed in Microsoft Office Access 2016. The file "microdados_enade_2014.csv" has 481,720 records.

For the study, a data cleaning was performed selecting students from the State of Tocantins, resulting in 6,488 records. In order to avoid interferences in the result, students with "Present" status in the test and the "gross test" variable through the "nt_ger (general grade)" attribute with “not null” were considered, leaving 3,182 records remaining.

The original database consists of 156 attributes, which correspond to the variables of higher education institution: variables of the course, variables of the enrollee, variables of presence, performance variables, perception of the test variables, and variables of the socioeconomic questionnaire. According to the objective of the research, the variables that compose the socioeconomic questionnaire were selected as study variables, besides the variable "gross grade of the test", resulting in 87 attributes considered for the study.

One of the objectives of this work is to find out if socioeconomic factors influence the performance of Tocantins students in ENADE. One way to achieve this goal is to apply the J48 algorithm (SAHU and MEHTRE, 2015) to categorize each record into a "performance range", to find out if family income can influence the performance of the student in the test.

For the application of the J48 algorithm, it was necessary to categorize the numerical variable "gross grade of assessment", the target attribute, into categories, since this is a prerequisite for the algorithm. Thus, four classes were created for the variable "grade level": class "[D]", with grades between 0 and 24.9, class "[C]" with grades between 25 and 49.9, class " [B] ", with grades between 50 and 74.9, and class " [A] ", with grades between 75 and 100 inclusive.

The result of the cleanup and preparation of the data, generated by Access 2016, was the file “Enade Students present.csv.” This file was used in the modeling step into WEKA tool, converting the file with extension ".csv" to a file with extension ".arff".

IV. DATA MODELING AND EVALUATION

After the data general study and groundwork, the data-modeling step is started. In this step, the data mining algorithms are chosen and applied in the database, in order to find valid and understandable patterns that can be used in the study. To support the data mining process, the WEKA computational tool has several algorithms to mine data, and it is an open source.

Data mining techniques (CABENA, 1998) are commonly categorized as supervised (predictive) learning and unsupervised (descriptive) learning (LAROSE, 2014). The limits that define the differences between the techniques are subtle, since some descriptive methods can be predictive or vice versa.

In data analysis, as well as data mining, variables can be

https://dx.doi.org/10.22161/ijaers.6769 *ISSN: 2349-6495(P) | 2456-1908(O)*

classified into one of two types: quantitative or categorical. Quantitative variables assume numerical values and represent some kind of measurement. On the other hand, categorical variables, take a category or labeled values and place the observation of the individual in one of several groups. Unsupervised learning does not require a preliminary categorization of its records, and it is not necessary to give a semantic meaning to the data, that is, the data are treated by their similarities of values (SILVA et al., 2014). The difference of the supervised learning method is the fundamental existence of a predetermined target attribute, so the algorithm can learn which target attribute values are associated with it. In this work, supervised learning was applied to semantically classify data into categories.

For the analysis of this work, the method of classification by Decision Tree (WITTEN et al., 2011) using the algorithm C4.5 (in its implementation J48) was adopted. Decision tree is a well-known supervised learning technique, which is suitable for data analysis involving continuous and discrete qualitative variables presented in the databases. It may promote accuracy, speed and ease of understanding of results. Decision trees uses information gain (entropy) to generate tree, which brings the extremely important concern on which are the optimal attributes to be considered in the analysis. The attribute evaluator algorithm "InfoGainAttributeEval" of the Weka tool was used in order to assess the attributes, and the search method "Ranker", to classify the individual attributes. The attributes considered in order to meet the objectives of this study are shown in Table 1.

Table 1: Attributes considered for the study.

Attribute	Description
qe_i8	Item 8 of the student questionnaire. "What is the total income of your family, including your income?"
qe_i13	Item 13 of the student questionnaire. "Throughout your academic career, have you received any kind of academic scholarship? In case there is more than one option, mark only the longest scholarship) "
qe_i15	Item 15 of the student questionnaire. "Did you enter the undergraduate course through affirmative action or social inclusion policies?"
qe_i17	Item 17 of the student questionnaire. "What kind of school did you attend high school?

https://dx.doi.org/10.22161/ijaers.6769 *ISSN: 2349-6495(P) | 2456-1908(O)*

In order to analyze the student's grade in ENADE, in relation to the classification of performance levels, a confidence factor of 0.25 was used. The statistics information about the generated decision tree shows that the model created for the student performance obtained an excellent result, with the classification of correct instances of 99.9686%, with only 0.0314% of instances classified incorrectly.

It can be noted that the accuracy distribution per class was 0.9999%, and the consistent confusion matrix shows that the algorithm performed optimally. In order to validate each of these relations, the test of variance was applied. The variance test (ANOVA) was performed in the SAS environment using the SAS Studio tool, with the procedure PROC ANOVA. The hypotheses considered for the Anova test are as follows:

• H0 = null hypothesis, all means are equal:

- $\mu a = \mu b = \mu c = \mu d$

• H1 = Alternative hypothesis, at least one of the averages is different from the others.

When entering SAS Studio you must upload the file, in .csv format, of the database that will be carried out the test containing only the response variable and the independent variable.

V. RESULTS

The results obtained in the accomplishment of this work were composed by the last phase of the CRISP-DM (BOSNJAK et al., 2009) methodology, the deployment. This phase aims to organize the knowledge acquired by analyzing the results in order to be presented in an understandable way so the client can use to support decision-making. The objective of this study is to analyze the performance of Tocantins students in the ENADE 2014 exam and to determine if certain socioeconomic factors may exert some influence on this performance. Thus, the original database had to be prepared and adapted, being composed only of data that were inherent in the study.

The database collected on the INEP website presents the students' grades through the variable "gross test score (nt_ger)", composed of the weighted average of 75% of the specific components and 25% of the general components, represented by a numerical type data varying from 0 - 100 points. This variable of the note was discretized for the attribute "level of notes (nt_nable)" and allocated in classes, "A" being the highest level grouping notes from 75 to 100 points, "B" grouping notes from 50 to 74.9, "C" with notes from 25 to 49.9, and "D" with grades ranging from 0 to 24.9.

The first result obtained with the modeling stage was the distribution of the performance of the students from the higher education of the State of Tocantins by level, through the application of the decision tree algorithm J48. Based on Figure 1, 3182 students who took the ENADE exam in Tocantins in 2014, 63.17% had "C" level of performance [24.9 - 50] and 13.39% had "D" level of performance [0 - 24.9], resulting in 76.56% of students with a level below the ENADE score average of 50 points. Only 23.44% of the students had scores above the average of 50 points, with 22.72% having B [50-74.9] performance and only 0.72% having A [75-100] performance.

In order to accomplish the objectives of this work, the performance of the students (target attribute) considering some socioeconomic factors, were obtained through the selection of attributes. The attribute evaluator algorithm "InfoGainAttributeEval" was applied by the combination with the search method "Ranker", to find out if these related socioeconomic factors can influence students' performance in the exam.

The first socioeconomic factor analyzed to confront with the student's grade was the family income, reported by the students in the socioeconomic questionnaire, Figure 2.

A) Up to 1.5 minimum salary (up to R $ 1,086.00);

B) From 1.5 to 3 minimum wages (R $ 1,086.01 to R $ 2,172.00);

C) From 3 to 4.5 minimum wages (R $ 2,172.01 to R $ 3,258.00);

D) From 4.5 to 6 minimum wages (R $ 3,258.01 to R $ 4,344.00);

E) From 6 to 10 minimum wages (R $ 4,344.01 to R $ 7,240.00);

F) From 10 to 30 minimum wages (R $ 7,240.01 to R $ 21,720.00);

G) Above to 30 minimum wages (more than R $ 21,720.01).

https://dx.doi.org/10.22161/ijaers.6769 *ISSN: 2349-6495(P) | 2456-1908(O)*

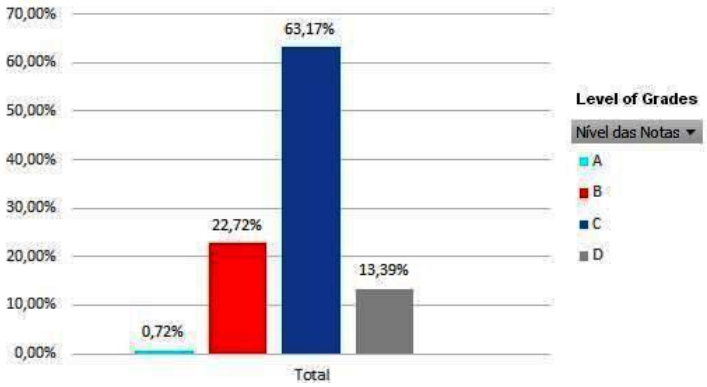


Fig.1: Distribution of students by performance level.

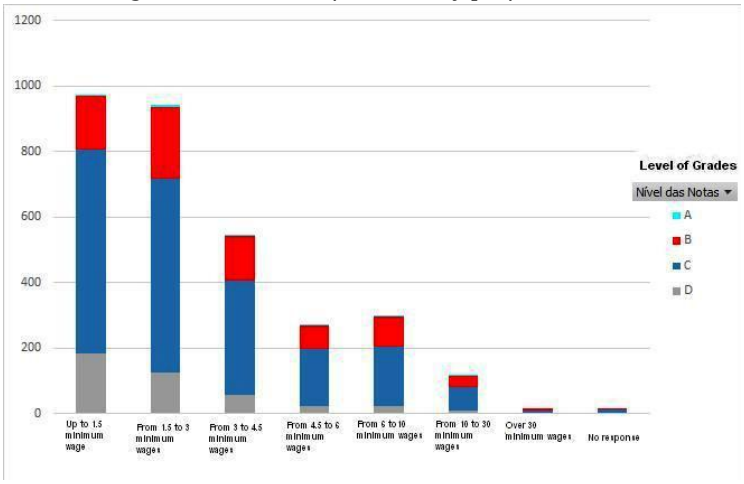


Fig.2: Distribution of students' performance by family's income.

It can be observed that the highest percentage of students who had a "D" level of performance, the lowest level of the scale, has a family income up to 1.5 minimum salary. This amount corresponds to 42.94% of the total of "D" level students. The 31.04% of the students who obtained "C" level (63.17% represented in Figure 1) also have income up to 1.5 minimum wages. Levels C and D represent grades up to 49.9 points of the exam, and it means grades below the ENADE rate.

Considering the students with scores above the ENADE average (50 points), the largest portion of the students has income ranging from 1.5 to 3 minimum salaries. The percentage of students in "B" level was 30.15% (of the overall total of 22.72%, Figure 1) and 34.78% (of the overall total of 0.72%, Figure 1) are on "A" level.

According to the absolute values shown in Figure 1, it can be seen a gradually improvement in the exam performance regarding the growth of the family income. To prove this hypothesis, the higher the performance the higher the family income, the variance analysis test was performed. For this, the file containing the occurrences of the general grade (nt_ger) attribute, and the family student income (qe_i8), the independent variable, was submitted. The command "ODS graphics on" of SAS was also used to generate the boxplot, Figure 3.

https://dx.doi.org/10.22161/ijaers.6769 *ISSN: 2349-6495(P) | 2456-1908(O)*

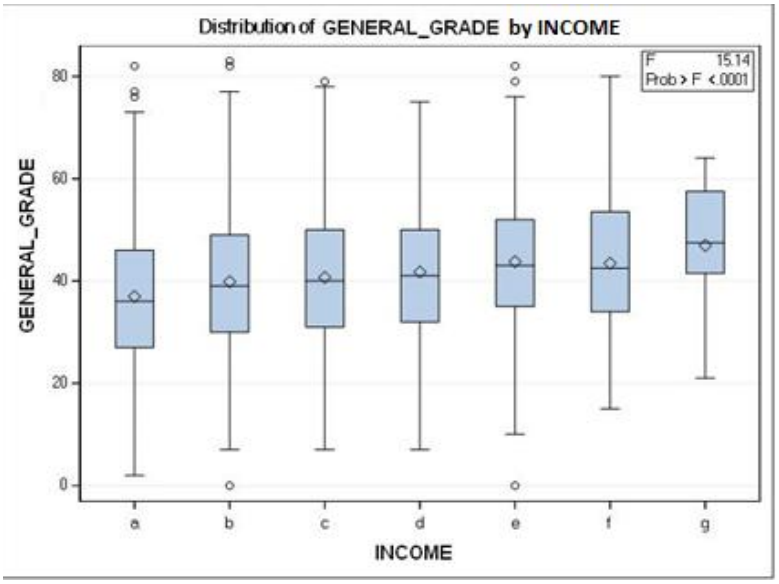


Fig.3: Anova test box plot.

In the box plot of Figure 3, it is possible to see an increase from box "a" to box "g", a proportion relation between student performance and income, suggesting an improvement in student performance with family incomes raise. The "a" box represents the family income of 1.5 minimum salaries. The data distribution of "a" box is between 27 and 46. From this, there is a growing trend as family income increases, e.g., the "f" box represents incomes from 10 to 30 minimum salaries, with distribution between 35 and 53. In this way, the alternative hypothesis that income can be an influence factor to the student's performance is accepted, $F = 15.14$, $p < .0001$.

It is also possible to visualize in Figure 4, the number of "observations read" and "observations used". It is noticed that the value used is less than the amount available in the database. This is due to the null values, namely, students who did not answer the family income-question in the questionnaire.

The ANOVA Procedure

Class Level Information		
Class	Levels	Values
RENDA	7	a b c d e f g

Number of Observations Read	3182
Number of Observations Used	3167

Fig.4: Number of observations for the income family.

The next Anova test intends to validate the relationship between students' performance by the type of school they finished the high school.

A) All in public school;
B) All in private school;
C) All out of the country;
D) The majority in public school;
E) Most in private school;
F) Partially in Brazil and partially out of the country.

It is possible to verify, in Figure 5, that the database has been correctly read through the alphabetical list of variables and attributes. The level information of the class displays the values of the attribute CONCLUSAO_EM, which represents the 6 alternatives of the question for the socioeconomic questionnaire. Each letter corresponds to an answer about the type of school that the student completed in high school.

The ANOVA Procedure

Class Level Information		
Class	Levels	Values
CONCLUSAO_EM	6	a b c d e f

Number of Observations Read	3182
Number of Observations Used	3167

Fig.5: Number of observations for the type of school the student accomplished in high school.

In Figure 6, it is shown the proportional values for the distribution of student's performance relatively to the type of school, which the student had completed the high school. The "b" box corresponds to the students who finished high school in a private school. It is highlighted by having achieved scores above to 80 points. In addition, its distribution obtained a score between 35 And 55

https://dx.doi.org/10.22161/ijaers.6769 *ISSN: 2349-6495(P) | 2456-1908(O)*

points. While the "a" box representing students who completed high school entirely in public school scored 30 to 49.

It is also observed that the majority of the students who finished high school in public school, 63.73% obtained below-average performance, and the highest percentage had a C-level performance. In order to endorse that the type of secondary school accomplished by the student is one of the factors that can influence the performance of the students in the ENADE exam, the test of variance analysis was performed with F = 12.26, and p < .0001.

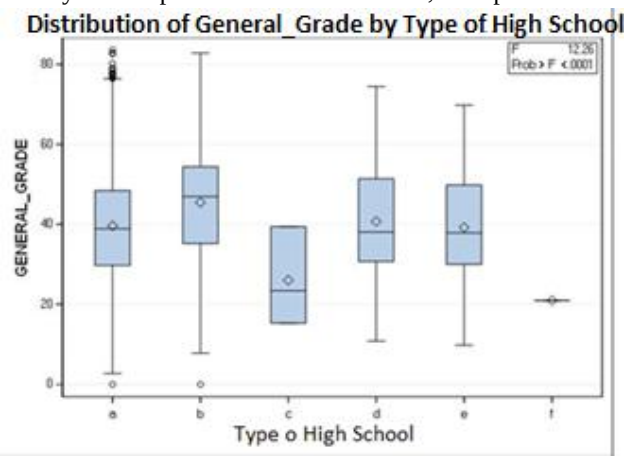


Fig.6: Box plot of the Anova test for the relationship between student performance in ENADE and the type of high school attended by the student.

Seeking to better understand which of the categories have statistically significant results between them, and, in a special highlight, between private and public high schools, the Anova test was performed for the two.

In Figure 7, it is possible to perceive that the students who had the opportunity to study in private school, showed a statistically significant result to be more qualified and to achieve superior performance compared to those who studied in public school. From the test of the variance analysis, F = 55.04 and p < .0001, the hypothesis that the type of high school accomplished by the student can influence the students' performance in ENADE exam is accepted.

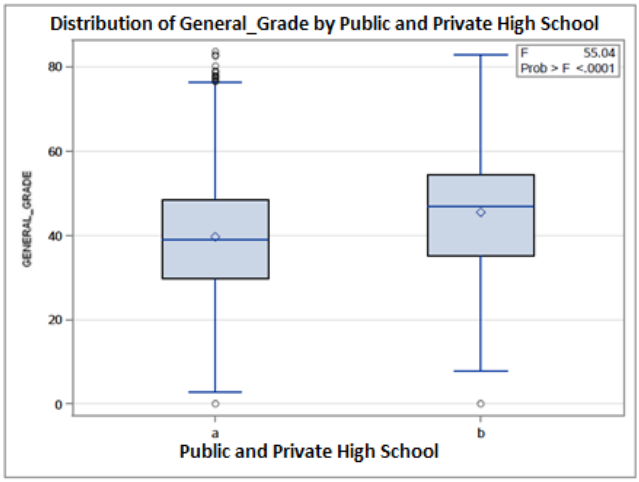


Fig.7: Statistically significant result of the Anova test to the relationship of the student's performance in ENADE exam to public and private high school attended by the student.

The following Anova test aims to validate the relation between student's perform in ENADE exam and the type of criterion used to enter into the undergraduate degree. The correct reading of the informed base is shown in Figure 8, through the alphabetical list of variables and the attributes of GENERAL_GRADE and INGRESS__IES. The information level of the class displays the values of the INGRESS_IES attribute, which represents the 6 alternatives of the question in the socioeconomic questionnaire. Each letter corresponds to an answer about the criteria used to enter to the undergraduate course.

A) No;

B) Yes, by ethno-racial criteria;

C) Yes, by income criteria;

D) Yes, for having studied in public or private school with scholarship;

E) Yes, by the combination of two or more previous criteria;

F) Yes, but a different system from the previous ones.

https://dx.doi.org/10.22161/ijaers.6769 *ISSN: 2349-6495(P) | 2456-1908(O)*

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat
2	INGRESSO_IES	Char	1	$1.	$1.
1	NOTA_GERAL	Num	8	BEST12.	BEST32.

The ANOVA Procedure

Class Level Information

Class	Levels	Values
INGRESSO_IES	6	a b c d e f

Number of Observations Read	3182
Number of Observations Used	3167

Fig.8: Criteria used by the student to be admitted to the undergraduate course.

The box blot, Figure 9, shows the distribution of the types of criteria used for entry into undergraduate course by the student's performance in ENADE exam. There is a great variation between the means of the criteria type alternatives.

In Figure 10, the Duncan test for ANOVA shows ethno-racial criteria in "A" group of Duncan, which has the higher-grade means. In the other side, the income criteria is in "C" group of Duncan, which is part of the minor-grade of means.

The last Anova test validated the relationship between students' performance and the receipt of an academic scholarship during the undergraduate course. The correct reading of the informed database, Figure 11, through the alphabetical list of variables and attributes shows the attributes of GENERAL_NOTE and ACADEMIC_SCHOLARSHIP. The class level information displays the values of the ACADEMIC SCHOLARSHIP attribute, which represents the 6 alternatives of the question in the socioeconomic questionnaire. Each letter corresponds to a response on receiving an academic scholarship during the undergraduate course.

A) None;
B) Scientific initiation scholarship;
C) Monitoring / mentoring scholarship;
D) Extension scholarship;
E) PET scholarship;
F) Another type of academic scholarship.

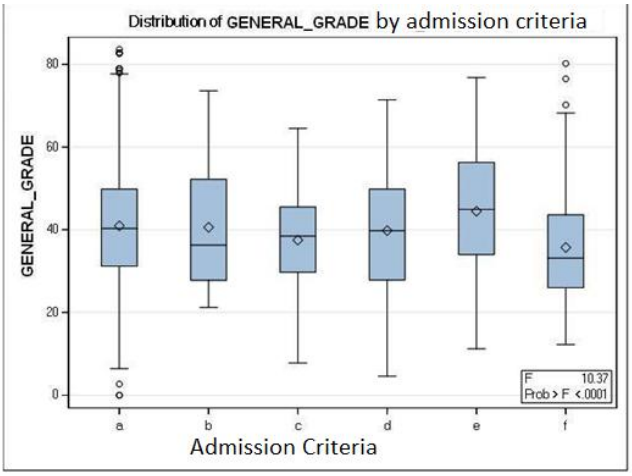
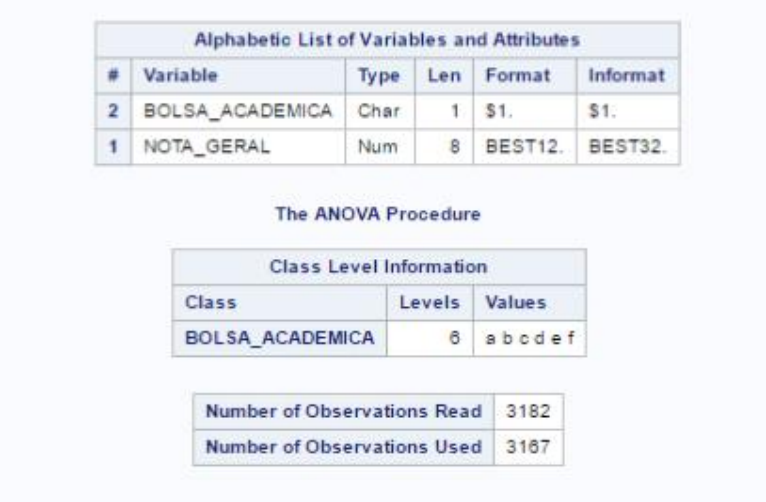


Fig.9: Anova box plot for student performance vs. criteria types used for admission in the undergraduate course.

https://dx.doi.org/10.22161/ijaers.6769 *ISSN: 2349-6495(P) | 2456-1908(O)*

Means with the same letter are not significantly different.				
Duncan Grouping		Mean	N	INGRESSO_IES
	A	44.477	35	e
	A			
B	A	40.947	2569	a
B	A			
B	A	40.602	44	b
B				
B	C	39.791	80	d
B	C			
B	C	37.502	109	c
	C			
	C	35.710	330	f

Fig.10: Duncan Test of ANOVA for Student Performance vs. Criteria Types Used for Admission to Higher Education.



Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat
2	BOLSA_ACADEMICA	Char	1	$1.	$1.
1	NOTA_GERAL	Num	8	BEST12.	BEST32.

The ANOVA Procedure

Class Level Information		
Class	Levels	Values
BOLSA_ACADEMICA	6	a b c d e f

Number of Observations Read	3182
Number of Observations Used	3167

Fig.11: Student's performance in ENADE exam versus receipt of an academic scholarship during undergraduate course.

The box plot, Figure 12, shows the distribution of the alternatives for receiving an academic scholarship during graduation. There is a small variation between the averages of boxes b, c, d and e, which refers to each of the discriminate alternatives of scholarships: scientific initiation; monitoring / tutorial, extension, and PET scholarships, respectively.

https://dx.doi.org/10.22161/ijaers.6769 ISSN: 2349-6495(P) | 2456-1908(O)

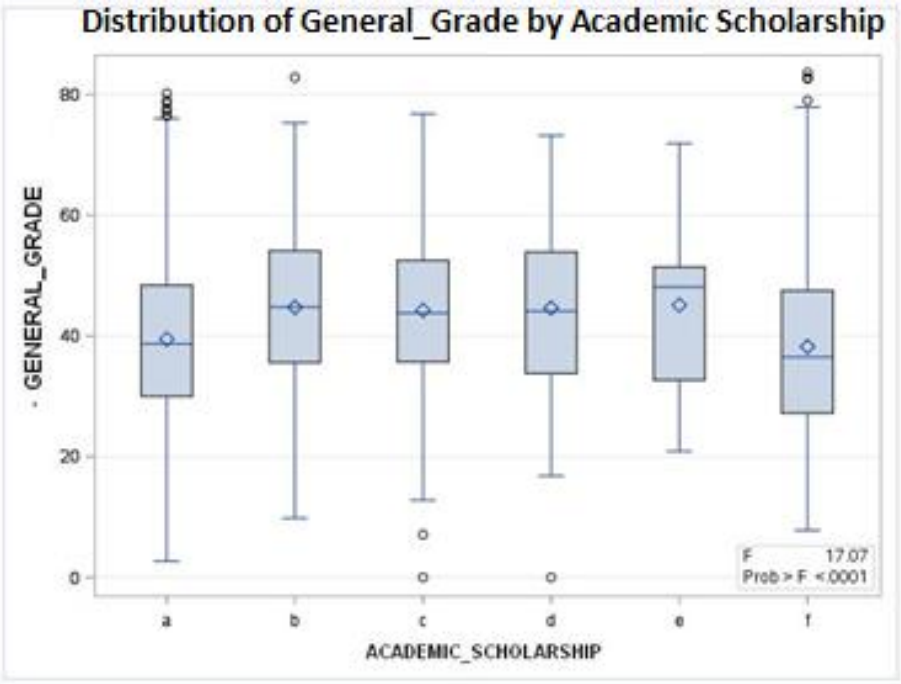


Fig.12: Anova Test for the student's performance versus receipt of an academic scholarship during the undergraduate course.

In Figure 13, the Duncan test for ANOVA shows a statistically significant result for the student performance to whom received discriminated academic scholarship and who not received, distinguished by "A" and "B" groups of Duncan test.

Means with the same letter are not significantly different.			
Duncan Grouping	Mean	N	BOLSA_ACADEMICA
A	45.114	29	e
A			
A	44.772	326	b
A			
A	44.614	137	d
A			
A	44.263	134	c
B	39.489	1980	a
B			
B	38.247	561	f

Fig.13: Duncan Test of ANOVA for Student's Performance in ENADE exam versus Receipt of Academic Scholarship during undergraduate course.

VI. CONCLUSION

Considering the factors selected to be analyzed, the results obtained in this study indicate that the socioeconomic factors can influence the performance of Tocantins' students in the ENADE exam. This assertion is based on the analysis of the selected attributes with the

https://dx.doi.org/10.22161/ijaers.6769 *ISSN: 2349-6495(P) | 2456-1908(O)*

target attribute (student general grades performance). The results were obtained through the analysis for the distribution of the absolute values, and also proved by the distribution of the proportional values in the graphs generated by the variance analysis test of ANOVA. According to the results obtained, it can be inferred that the students with higher family income have higher grades in the exam.

In this scenario, a higher family income can somehow provide the student with better conditions to carry out his or her activities during their graduation journey. On the other hand, when the analysis is made from the point of view of the ethnic-racial elements, we do not perceive a relevant influence for the race-ethnicity of the student to a low performance in the exam. It was verified by the ANOVA and Duncan tests performed between general grades and the criteria used to enter to the undergraduate course of Institutions. Conversely, the ethnic-racial element had statistically significant to a better performance of the students as well as higher incomes, as we can see in the results for the "A" group of Duncan in the same test. This result also suggests that there is no differences in intellectual capacity between race ethnicities.

The results indicate that, for the application of affirmative action policies related to social quotas and the distribution of scholarships, it is fairer to provide socially economically disadvantaged groups, putting in practice, in this way, distributive justice.

The research was based on the study of data from the ENADE 2014. It was considered only records of students from the state of Tocantins. Nonetheless, it is necessary to have a more comprehensive study and to find out if this is the reality only of students from federal state of Tocantins or if this result can be extended to the rest of the country.

REFERENCES

[1] BOSNJAK, Z.; GRLJEVIC, O.; BOSNJAK, S. Crisp-dm as a framework for discovering knowledge in small and medium sized enterprises' data. In: IEEE. Applied Computational Intelligence and Informatics, 2009. SACI'09. 5th International Symposium on. [S.l.], 2009. p. 509–514.

[2] CABENA, P. Discovering Data Mining: From Concept to Implementation.[S.l.]: Prentice Hall PTR, 1998. (An IBM Press Book Series). ISBN 9780137439805.

[3] CHAPMAN, P. et al. Crisp-dm 1.0 step-by-step ddata mining guide. 2000.

[4] DIAS, M. M. Parâmetros na escolha de técnicas e ferramentas de mineração de dados. Acta Scientiarum. Technology, v. 24, p. 1715-1725, 2008.

[5] FAYYAD, U.; Piatetsky-Shapiro, G.; Smyth, P. From data mining to knowledge discovery in databases. AI magazine, v. 17, n. 3, p. 37, 1996.

[6] GOLDSCHMIDT, R.; Passos, E. Data Mining: Um Guia Prático Conceitos, Técnicas, Ferramentas, Orientações e Aplicações. [S.l.: s.n.], 2005.

[7] INEP - Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. <http://portal.inep.gov.br/basica-levantamentos-acessar/>. "Último acesso em 18/10/2016".

[8] LAROSE, D. T. Discovering Knowledge in Data: An Introduction to DataMining. [S.l.]: John Wiley Sons, 2014.

[9] PANDA, M.; PATRA, M. R. A comparative study of data mining algorithms fornetwork intrusion detection. In: IEEE. 2008 First International Conference onEmerging Trends in Engineering and Technology. [S.l.], 2008. p. 504–507

[10] QUINLAN, J. R. Generating production rules from decision trees. In: CITESEER. IJCAI. [S.l.], 1987. v. 87, p. 3040-307.

[11] REZENDE, S. O. et al. Mineração de dados. Sistemas inteligentes: Fundamentose Aplicações, v. 1, p. 307–335, 2003.

[12] RODRIGUES, R. L. et al. A literatura brasileira sobre mineração de dadoseducacionais. In: Anais dos Workshops do Congresso Brasileiro de Informática. na Educação. [S.l.: s.n.], 2014. v. 3, n. 1, p. 621.

[13] SAHU, S.; MEHTRE, B. Network intrusion detection system using j48 decisiontree. In: IEEE. Advances in Computing, Communications and Informatics (ICACCI), 2015 International Conference on. [S.l.], 2015. p. 2023–2026.

[14] SAS Products - SAS® Studios. <http://support.sas.com/software/products/sasstudio /index.html>. "Último acesso em 26/10/2016".

[15] SILVA, M. d. A. O Pré-Processamento em Mineração de Dados como método de suporte à modelagem algorítmica. Dissertação (Mestrado) — UniversidadeFederal do Tocantins, Curso de Pós-Graduação em Modelagem Computacional deSistemas da Universidade Federal do Tocantins, Palmas, 2014.

[16] Weka 3: Data Mining Software in Java. <http:www.cs.waikato.ac.nz/ml/weka/>. "Último acesso em 20/10/2016".

[17] WITTEN I.; FRANK, E. H. M. Data Mining Practical Machine LearningTools and Techniques. Third Edition. [S.l.]: Elsevier, 2011. (The Morgan KaufmannSeries in Data Management Systems). ISBN 9780080890364.

[18] YUN, Z.; WEIHUA, L.; YANG, C. Applying balanced scordcard strategicperformance management to crisp-dm. In: IEEE. Information Science, Electronicsand Electrical Engineering (ISEEE), 2014 International Conference on. [S.l.],2014. v. 3, p. 2009–2014.

[19] ZHOU, Z.-H. Three perspectives of data mining. Elsevier, 2003.