Analyzing the Determinant Characteristics for a Good Performance at ENADE Brazilian Exam Stratified by Teaching Modality: Face-to-face versus Online

Eric Gondran¹, Giancarlo Lucca¹, Rafael Berri¹, Helida Santos¹,² and Eduardo N. Borges¹

¹C3, Universidade Federal do Rio Grande, Av. Itália km 08, Rio Grande, Brazil
²ISC, Universidad Publica de Navarra, Campus Arrosadia, Pamplona, Spain

Keywords: ENADE, Feature Selection, Educational Data Mining, Data Science.

Abstract: The National Student Performance Exam (ENADE) annually evaluates different Brazilian higher education courses. This exam considers both face-to-face and distance learning courses. Distance learning is growing increasingly, especially during the coronavirus (COVID-19) pandemic. This study applies different techniques for selecting ENADE 2018 database characteristics, like information gain, gain rate, symmetric uncertainty, Pearson correlation, and relief F. The objective of the work is to discover which personal and socioeconomic characteristics are decisive for the student’s performance at ENADE, whether the student is in the context of Distance Education or face-to-face. It can be concluded, among other results, that: the father’s level of education directly influences performance; the higher the income, the better the performance; and white students have better performance than black and brown-skinned ones. Thus, the results obtained in this study may initiate analyzes of public policies towards improving performance at ENADE.

1 INTRODUCTION

Higher education is booming in Brazil, and according to CES (acronym in Portuguese for Higher Education Census) (Inep, 2019), from 2009 to 2019, enrollment in higher education increased by 43.7%. CES constitutes an essential instrument for obtaining data to generate information that subsidizes public policies’ feeding, monitoring, and evaluation. In 2019, students enrolled in higher education reached 8.6 million, a growth of 5.4% compared to 2018. This increase is due to distance learning (DL) (Barreto and Amaral, 2019) which, from 2009 to 2019, increased 378,9%. DL is a form of education in which learning occurs at a distance (physical and temporal), mediated by a technology tool that allows communication and interaction between participants.

Directly linked to teaching, we have the National Student Performance Exam (ENADE), an assessment that annually analyzes the different Brazilian higher education courses. This exam considers both on-site and distance learning courses. The National Institute of Educational Studies and Research Anísio Teixeira (Inep), a federal agency linked to the Ministry of Education (MEC), currently applies and elaborates this exam. ENADE microdata is generated through the examination, the minor level of granularity of collected data. They meet the demand for specific information by providing tests, templates, information about items (manual and dictionary), grades, and the student questionnaire, which contains different information regarding the candidate.

Considering the microdata from ENADE, this work aims to find which personal and socioeconomic characteristics are decisive for the student’s performance, whether he/she is a DL or a face-to-face (F2F) student. To achieve our objective, this work employs a well-established methodology called Knowledge Discovery in Database (KDD) (Tan et al., 2016). Applying data mining techniques in the field of education has shown promising results, giving rise to a new area of scientific investigation called Educational Data Mining (EDM) (Baker et al., 2011).
This article is organized as follows. In Section 2, we find the related works. In Sect. 3, feature selection techniques are presented. In Sect. 4, the adopted methodology is described. The results are presented in Sect. 5, and we have the conclusions in Sect. 6.

2 RELATED WORK

The possibilities for applying EDM techniques in Brazil are presented by (Baker et al., 2011). The study demonstrates how this area of research can contribute to a better understanding of teaching and learning processes and student motivation. (Gottardo et al., 2012) proposes the definition of a broad and generalizable set of attributes used to make inferences regarding student performance. Experiments performed showed indexes of 76% accuracy in predicting performance. (Romero et al., 2008) highlights the possibilities of using data mining to extract relevant information about students in the educational context.

Specifically, at ENADE, (Araújo et al., 2019) proposed the use of knowledge discovery techniques to develop a tool for exploring the exam data. In addition to evaluating the structure and distribution of test data, they also proposed a model based on the CART (Breiman et al., 1984) algorithm capable of predicting student performance. (Faria, 2017) work has as its primary objective the identification of the determining factors in the performance of students in Business Administration courses. They used microdata from ENADE 2012 of the Federal District. The research has predominantly quantitative analysis characteristics based on descriptive and multivariate statistical techniques. The multiple regression method was used to verify whether the student’s characteristics, such as personal and socioeconomic aspects, the institution, and the didactic-pedagogical organization, would be significant variables in explaining the test result. The results showed that the student relevant factors that aided in predicting performance were: family income, male gender, mother’s and father’s education level. Unlike the works exposed above, our proposal aims to apply different techniques for feature selection, to point out which personal and socio-economic aspects are significant characteristics in explaining the result of ENADE 2018.

3 FEATURE SELECTION

Feature Selection (FS) is the process of identifying and removing irrelevant attributes and redundant information as much as possible (Miao and Niu, 2016). FS reduces the dimensionality of the data. It improves the performance of the classifiers, as it eliminates attributes that do not add value to the classification or deteriorate the results. It contributes to a better understanding and analysis of the results obtained and allows learning algorithms to operate more quickly and effectively. FS main objective is to identify the set of attributes best representing the useful information in the data (Tasca, 2015), within a context.

3.1 Entropy

Entropy (Shannon, 1948) is defined as a form of measurement or average degree of uncertainty regarding sources of information, which consequently allows a quantification of the information present that flows in the system. In simple terms, the concept of entropy is associated with the idea that the more certain the outcome of a random experiment, the more information you get from observing its occurrence.

It can also be defined as the amount of uncertainty in a message, which decreases as the symbols are transmitted, that is, as the message becomes known, then information is obtained, which can be seen as uncertainty reduction. Entropy is calculated by:

\[ E(A) = -\sum_{a \in A} P(a) \log_2 P(a), \]

where \( E(A) \) is the entropy, \( P(a) \) is the relative frequency of values, \( A \) is the attribute to be calculated, \( a \) is the value of this attribute and \( P \) is the relative frequency of values.

3.2 Information Gain

The information gain (IG) (Hall and Smith, 1998) is defined as the amount of information obtained about a random variable or signal from the observation of another random variable. It measures the significance of the attribute with the target class; i.e., it measures the reduction of uncertainty (entropy) as a division function. As a disadvantage, it tends to prefer divisions that result in numerous partitions, each one being small but neat. Information Gain is calculated by:

\[ IG(A) = E(C) - \sum_{a \in A} \frac{n_a}{n} E(a), \]

where \( E \) denotes the entropy function, \( C \) is the class, \( A \) is the attribute to be evaluated, \( n_a \) is the number of instances of the category belonging to the attribute, \( n \) is the total number of instances and \( a \) is the attribute value.

3.3 Gain Ratio

The gain ratio (GR) (Karegowda et al., 2010) was developed to solve the IG problem. It is the ratio of IG and the attribute entropy, which is nothing more than the relative IG as an evaluation criterion. It adjusts
the IG by partitioning entropy, causing high entropy partitioning (a large number of small partitions) to be penalized. GR is defined by: $GR(A) = \frac{IG(A)}{E(A)}$, where $IG(A)$ is information gain and $E$ is the entropy.

### 3.4 Symmetric Uncertainty

Symmetric uncertainty (SU) (Yu and Liu, 2003) is a nonlinear correlation measure developed with the same purpose of GR, that is, an attempt to normalize the IG of an attribute $A$ with the class $C$. SU is defined by: $SU(A) = 2 \cdot \frac{Var(X) \cdot Var(Y)}{Var(X) + Var(Y)}$, where $IG(A)$ is Information Gain and $E$ is the entropy.

### 3.5 Pearson Correlation

Pearson Correlation (PC) (Hall, 1998), also known as a linear coefficient, measures the degree of correlation between two metric scale variables. It is a relationship degree between two quantitative attributes, and it expresses the correlation degree through values between −1 (negative or inverse correlation) and 1 (positive linear relationship). A correlation coefficient near zero indicates no relationship between the attributes. The PC is given by: $PC(A) = \frac{Cov(X,Y)}{\sqrt{Var(X) \cdot Var(Y)}}$, where $Cov$ is the covariance between the two attributes and $Var$ is the variance of each attribute. To calculate the qualitative attributes correlation, data are adapted by turning them into binary data.

### 3.6 Relief F

Over the years, a Relief extension called Relief F (Kononenko, 1994; Kira and Rendell, 1992; Urbanowicz et al., 2018) has been developed, aiming to improve the original algorithm by estimating probabilities more reliably. It handles multiclass and incomplete datasets, while the complexity remains the same. It is calculated using a function $W$ defined by: $W(A) = W(A) - \frac{\text{diff}(A,R,H)}{m} \cdot 5 + \frac{\text{diff}(A,R,M)}{m}$, where $A$ is the attribute, $W(A)$ is a vector with each attribute score, $R_i$ is the target instance, $H$ is the closest instance of the same class, $M$ is the closest instance to the other class, $m$ is the number of random instances selected to be part of the calculation, and the function $\text{diff}$ calculates the difference between attributes.

### 4 METHODOLOGY

In this section, we present the methodology used for this study. We emphasize that the work is supported by the KDD process, which comprises five stages.

1. **Selection.** This work takes into account ENADE 2018 microdata. They have 548,127 instances and 137 attributes of the numeric or character type. The attributes are divided, among others, into the institution and course information, student information, the number of items in the objective part, types of presence (participant present, absent or canceled test), test perception questionnaire, and student questionnaire. The original database was divided into online students (96,927 instances) and F2F students (451,200 instances). After analyzing all database attributes, we focus on the personal, socioeconomic aspects and participant’s course. We emphasize that at this point, 23 attributes were kept in each database.

2. **Preprocessing.** The first preprocessing operation was the application of a filter to select only those participants who had actually taken the test. We removed 32,285 participants from the online modality and 115,765 F2F students. The criteria for removing attributes include: candidates with a blank test in the objective and discursive part of general education, candidates with a blank test in the objective and discursive part of the specific component, participation with a result disregarded by the Applicator. The second step verified null or incomplete data, including blank test notes and the blank part of the questionnaire. We excluded 15 online cases and 103 F2F. Online databases had 64,627 instances, and F2F had 335,332.

3. **Transformation.** The first operation was to rename the attributes. At this stage, 23 attributes had names referring to the student’s questionnaire number (QE_101 to QE_In). The nominal values of the attributes (A, B, etc.) were also renamed, for example, the father’s level of schooling was renamed to (None, Elementary 1, Elementary 2, High school, Undergraduate, Graduate). The courses were also grouped according to their primary areas, according to the tables provided by CNPq and CAPES, Brazilian funding agencies. ENADE’s exam occurs every three years in a specific set of courses. Not all courses took the test in 2018. The scores obtained by the candidates were also categorized, with their values discretized into three frequency categories (low, medium, and high performance), keeping the original distribution. Discretized online student grades performance: Low ($\leq 30$), Medium ($30 < \text{grade} \leq 60$).
and High (> 60). Discretized grades of face-to-face students: Low (≤ 31), Medium (31 < grade ≤ 62) and High (> 62).

4. Data Mining. In this step, we apply the five distinct methods for FS. We consider the database (online and F2F), taking into account 23 pre-selected attributes. In Sect. 3 we present the applied FS algorithms in detail. For each algorithm, the most relevant characteristics are selected according to the ranking generated by the algorithms. We also perform an exploratory analysis of the most frequent attributes in all methods.

5. Interpretation and Evaluation. After applying the five FS methods and analyzing the ranking generated by the algorithms, the top-10 most frequent attributes are considered in at least four of the FS algorithms, both for the students in online and F2F modality. After choosing the best attributes, we perform an exploratory analysis of each attribute. We generate graphics for each modality and compare online and face-to-face modality profiles. In Sect. 5, we present and discuss the results of the exploratory analysis.

5 RESULTS

This section presents the results of applying the different FS algorithms.

5.1 Online Modality Results

This subsection presents the results obtained for the FS techniques in the online modality data. Such results are available in Table 1, where the lines are the 22 attributes considered and the columns are the different FS methods. Each cell values refer to the result of the operation and the ranking obtained by the attribute. The top ten values for each method are underlined. In addition, we highlight in bold the attributes selected in the first ten positions in at least four of the algorithms used. At the end of the table, the column Rank contains the sum of the four best positions obtained by the different SA methods. To define the most relevant characteristics to explain the online students’ performance at ENADE 2018, we considered the top-10 most frequent attributes in at least four of the feature selection algorithms.

Analyzing the gender (Figure 1), most (60.8%) of the participants are female, and however, they obtained inferior results compared to the males. Analyzing the skin color (Fig. 2), white people are predominant, with 33,845 participants (52.4% of them) having the most outstanding high-performance rates (9%) and the least low-performance rates (25%). Brown-skinned people is the second-highest rate of participants (35.5%), having the worst high-performance (5%) and one of the highest low-performance indices (32%). Similarly, we find low-performance indices for black, yellow, and indigenous people. Finally, the best overall performance came from those who did not declare their skin color.

5.2 Face-to-face Modality Results

This subsection presents the results obtained when the FS techniques are applied in the base that considers the F2F modality. Such results are available in Table 2, using the same approach seen in Table 1.

The same criterion used in online education was used to define the most important characteristics. Most participants (289,580/86.4%) did not receive any academic scholarship; 10% and 20% of them presented high and low performances, respectively. Those with a scholarship (Fig. 3) show a notable superior performance compared to the ones without it. We highlight that scientific research and PET (acronym in Portuguese for Tutorial Education Program) participants have the most elevated high-performance (32% and 35%) and the least low-performance (7% and 8%). Analyzing the weekly study time (Fig. 4), the
Table 1: Online modality results obtained with different feature selection methods.

<table>
<thead>
<tr>
<th># characteristics</th>
<th>PC</th>
<th>IG</th>
<th>GR</th>
<th>SU</th>
<th>RF</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge Area</td>
<td>1</td>
<td>0.1253</td>
<td>2</td>
<td>0.0404</td>
<td>1</td>
<td>0.0372</td>
</tr>
<tr>
<td>Family Income</td>
<td>3</td>
<td>0.0415</td>
<td>1</td>
<td>0.0243</td>
<td>3</td>
<td>0.0177</td>
</tr>
<tr>
<td>Gender</td>
<td>2</td>
<td>0.0510</td>
<td>1</td>
<td>0.0108</td>
<td>4</td>
<td>0.0117</td>
</tr>
<tr>
<td>High School Education</td>
<td>6</td>
<td>0.0967</td>
<td>3</td>
<td>0.0139</td>
<td>3</td>
<td>0.0311</td>
</tr>
<tr>
<td>Scholarship/Funding</td>
<td>2</td>
<td>0.0346</td>
<td>7</td>
<td>0.0098</td>
<td>6</td>
<td>0.0050</td>
</tr>
<tr>
<td>Father’s Level of Education</td>
<td>17</td>
<td>0.0163</td>
<td>14</td>
<td>0.0113</td>
<td>3</td>
<td>0.0051</td>
</tr>
<tr>
<td>Chosen Course</td>
<td>13</td>
<td>0.029</td>
<td>6</td>
<td>0.0105</td>
<td>7</td>
<td>0.0042</td>
</tr>
<tr>
<td>Skin Color</td>
<td>7</td>
<td>0.0373</td>
<td>10</td>
<td>0.0083</td>
<td>10</td>
<td>0.0040</td>
</tr>
<tr>
<td>Mother’s Level of Education</td>
<td>18</td>
<td>0.0128</td>
<td>8</td>
<td>0.0070</td>
<td>13</td>
<td>0.0031</td>
</tr>
<tr>
<td>Weekly Study Time</td>
<td>9</td>
<td>0.0333</td>
<td>11</td>
<td>0.0058</td>
<td>14</td>
<td>0.0031</td>
</tr>
<tr>
<td>Age</td>
<td>8</td>
<td>0.0369</td>
<td>12</td>
<td>0.0057</td>
<td>12</td>
<td>0.0034</td>
</tr>
<tr>
<td>Financial Status</td>
<td>15</td>
<td>0.0197</td>
<td>9</td>
<td>0.0069</td>
<td>16</td>
<td>0.0029</td>
</tr>
<tr>
<td>High School Modality</td>
<td>12</td>
<td>0.0244</td>
<td>13</td>
<td>0.0046</td>
<td>11</td>
<td>0.0037</td>
</tr>
<tr>
<td>Family’s Undergraduate Degree</td>
<td>14</td>
<td>0.0197</td>
<td>15</td>
<td>0.0040</td>
<td>8</td>
<td>0.0042</td>
</tr>
<tr>
<td>Work Status</td>
<td>10</td>
<td>0.0114</td>
<td>14</td>
<td>0.0046</td>
<td>15</td>
<td>0.0229</td>
</tr>
<tr>
<td>Marital Status</td>
<td>5</td>
<td>0.0402</td>
<td>18</td>
<td>0.0029</td>
<td>17</td>
<td>0.0019</td>
</tr>
<tr>
<td>People Living in your Household</td>
<td>19</td>
<td>0.0123</td>
<td>16</td>
<td>0.0039</td>
<td>20</td>
<td>0.0015</td>
</tr>
<tr>
<td>Social Inclusion Program</td>
<td>16</td>
<td>0.0163</td>
<td>17</td>
<td>0.0038</td>
<td>9</td>
<td>0.0041</td>
</tr>
<tr>
<td>Household Location and People in it</td>
<td>11</td>
<td>0.0310</td>
<td>19</td>
<td>0.0024</td>
<td>19</td>
<td>0.0017</td>
</tr>
<tr>
<td>Chosen Education Institution</td>
<td>21</td>
<td>0.0058</td>
<td>20</td>
<td>0.0016</td>
<td>22</td>
<td>0.0008</td>
</tr>
<tr>
<td>Academic Scholarship</td>
<td>20</td>
<td>0.0074</td>
<td>21</td>
<td>0.0004</td>
<td>21</td>
<td>0.0012</td>
</tr>
<tr>
<td>Student financial aid</td>
<td>22</td>
<td>0.0020</td>
<td>22</td>
<td>0.0001</td>
<td>18</td>
<td>0.0017</td>
</tr>
</tbody>
</table>

Table 2: Face-to-face modality results obtained with different feature selection methods.

<table>
<thead>
<tr>
<th># characteristics</th>
<th>PC</th>
<th>IG</th>
<th>GR</th>
<th>SU</th>
<th>RF</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scholarship/Funding</td>
<td>3</td>
<td>0.0341</td>
<td>3</td>
<td>0.0166</td>
<td>1</td>
<td>0.0206</td>
</tr>
<tr>
<td>Academic Scholarship</td>
<td>2</td>
<td>0.0360</td>
<td>1</td>
<td>0.0130</td>
<td>2</td>
<td>0.0111</td>
</tr>
<tr>
<td>High School Education</td>
<td>1</td>
<td>0.0196</td>
<td>2</td>
<td>0.0168</td>
<td>5</td>
<td>0.0065</td>
</tr>
<tr>
<td>Family Income</td>
<td>1</td>
<td>0.0551</td>
<td>6</td>
<td>0.0124</td>
<td>4</td>
<td>0.0106</td>
</tr>
<tr>
<td>Knowledge Area</td>
<td>10</td>
<td>0.0205</td>
<td>5</td>
<td>0.0158</td>
<td>4</td>
<td>0.0056</td>
</tr>
<tr>
<td>Chosen Course</td>
<td>8</td>
<td>0.0218</td>
<td>9</td>
<td>0.0104</td>
<td>2</td>
<td>0.0053</td>
</tr>
<tr>
<td>Weekly Study Time</td>
<td>14</td>
<td>0.0155</td>
<td>7</td>
<td>0.0123</td>
<td>8</td>
<td>0.0052</td>
</tr>
<tr>
<td>Father’s Level of Education</td>
<td>12</td>
<td>0.0199</td>
<td>13</td>
<td>0.0083</td>
<td>13</td>
<td>0.0035</td>
</tr>
<tr>
<td>Mother’s Level of Education</td>
<td>21</td>
<td>0.0077</td>
<td>19</td>
<td>0.0039</td>
<td>22</td>
<td>0.0011</td>
</tr>
<tr>
<td>Social Inclusion Program</td>
<td>20</td>
<td>0.0064</td>
<td>20</td>
<td>0.0026</td>
<td>14</td>
<td>0.0031</td>
</tr>
<tr>
<td>Family’s Undergraduate Degree</td>
<td>19</td>
<td>0.0124</td>
<td>21</td>
<td>0.0015</td>
<td>12</td>
<td>0.0044</td>
</tr>
<tr>
<td>Age</td>
<td>18</td>
<td>0.0135</td>
<td>22</td>
<td>0.0014</td>
<td>21</td>
<td>0.0015</td>
</tr>
</tbody>
</table>

majority (46.3%) studies from 1 to 3 hours a week, with a high performance of 8% and a low performance of 22%. Clearly, when the number of weekly study hours increases, high-performance increases and low-performance decreases. Considering the mother’s education level (Fig. 5), those whose mothers have no education or have completed just elementary school 1 and 2 have the worst high-performance rates and the highest low-performance rates. Mothers with undergraduate and graduate degree studies imply better performance on student rates.

Figure 3: Relation of the type of academic scholarship and the participant performance in the face-to-face modality.
5.3 Comparing Results between Online and Face-to-face Modalities

In this section, we sought to compare the characteristics in common between online and F2F participants, in which we seek to understand the difference between these two profiles. Comparisons are shown in Figures 6 to 11. For each figure (attribute), we will perform an exploratory analysis.

Analyzing Fig. 6, we can see a slight growth of applied social sciences, going from 66.5% in online modality to 74.9% in F2F. There is a better distribution in the F2F data in the other categories. Regarding performance, there is an improvement in the F2F students in applied social sciences. Among all categories in the online modality, human sciences have the best overall performance with rates of 21% (high) and 11% (low). In the F2F modality, the highlight is Linguistics, Languages, and Arts with an efficiency of 18% (high) and 8% (low).

Observing the students who receive scholarships or funding grants (Fig. 7), there is an increase in the number of students in free courses and FIES (acronym in Portuguese for Finance Fund for Higher Education Students), in F2F modality. Better overall performance can be seen in all on-site students, highlighting free courses and fully funded ProUni students, who rose from 10%, 6% to 21%, 17% respectively. Sadly, those with some scholarship or funding grants are the worst performers in both modalities.

Considering the school type attended in high
school (Fig. 8), we see most students are from public schools. It is also noticeable that the proportion of face-to-face students from private schools (26.8%) is twice as much compared to online students, 11.7%. Students’ performance from private schools presents the best results for high and low performance.

Regarding the parents’ level of education (Fig. 9), in online modality, elementary school 1 is the majority with 39.4%, whereas, in F2F modality, high school predominates with 32.6%. There is an increase in the F2F modality for those with a father having an undergraduate and graduate degree education. It is seen that as the level of education increases, the high-performance index also increases.

The course choice by job market inclusion is mostly seen in online and F2F modalities, as seen in Fig. 10. Focusing on F2F participants’ performance, we notice an increase in performance compared to online students. We note that those who opt for an online course have the best results in this same modality.

Analyzing the family income (Fig. 11), most students have an income lower than 4.5 minimum wages (mw). There is an increase in F2F students with a family income above 10 mw. In performance, we observe F2F students with an income less than 6 mw increase high performance. In both modalities, as income increases, the high performance also increases.
the factors that influence the performance of undergraduate students significantly contributes to a better understanding of the national education panorama. Besides, this study can help authorities make decisions and propose new public policies concerning Higher Education.

6 CONCLUSIONS

In this paper, we applied educational data mining in ENADE 2018 data set to find the main characteristics related to performance, both in distance learning and F2F modality. We used five different algorithms for selecting attributes to 23 pre-selected personal and socioeconomic characteristics. The essential characteristics selected in both modalities were: knowledge area, family income, public or private high school, scholarships and funding, father’s level of education, and reason for choosing the course. Gender and skin color were also important for online modality. And for the F2F modality, the type of academic scholarship, weekly study time, and the mother’s schooling.

Parents’ education directly influenced the results, and the higher the level of education, the better the performance. Public universities were protagonists, as students who paid no fees performed better than those in private institutions, even if financed by government programs or scholarships from the institutions. We see a direct and proportional relationship between family income and student performance. The higher the income, the better the high-performance index and the lower the low-performance index. Participants who attended (fully or partially) private high schools have an advantage over those who attended public schools. Finally, we conclude that analyzing

ACKNOWLEDGEMENTS

This study was supported by CNPq (305805/2021-5) and PNP/D/CAPES (464880/2019-00).

REFERENCES


