Revista Brasileira de Informática na Educação – RBIE
Brazilian Journal of Computers in Education
(ISSN online: 2317-6121; print: 1414-5685)
https://sol.sbc.org.br/journals/index.php/rbie

Submission: 08/07/2025; 1st round notif.: 27/09/2025; New version: 26/11/2025; 2nd round notif.: 18/12/2025;
Camera ready: 23/03/2026; Edition review: 02/04/2026; Available online: 02/04/2026; Published: 02/04/2026;

Unveiling Patterns in Maranhão's 2023 ENEM Results Through Unsupervised Machine Learning

Anderson Amorim Alves
Programa de Pós-Graduação em
Engenharia da Computação e Sistemas (PECS)
Universidade Estadual do Maranhão (UEMA)
ORCID: 0009-0005-9785-8365
slz.anderson.ma@gmail.com

Omar Andres Carmona Cortes
Departamento de Computação (DComp)
Instituto Federal do Maranhão (IFMA)
ORCID: 0000-0002-5805-2490
omar@ifma.edu.br

Abstract

This investigation examines the performance of students from Maranhão, Brazil, in the 2023 edition of the National High School Exam (ENEM) using unsupervised machine learning to uncover latent patterns in large-scale educational data. Leveraging the CRISP-DM framework on microdata from millions of ENEM participants, we applied Recursive Feature Elimination (RFE) with a Random Forest classifier to select key socioeconomic variables, followed by association rule mining using the FP-Growth algorithm across multiple experimental configurations. The results reveal strong associations between low academic performance and factors such as parental education and occupation, lack of household technology (e.g., computers and washing machines), and gender. These findings demonstrate the utility of unsupervised learning for descriptive educational analytics, providing practical insights for targeted policy-making, resource allocation, and regional equity monitoring. This research addresses the underrepresentation of Maranhão in the educational data mining literature and proposes a scalable analytical framework applicable to other developing regions. Despite limitations in data completeness, the approach offers a replicable model for using artificial intelligence to inform public education strategies in socially vulnerable areas.

Keywords: *Educational Data Mining; Unsupervised Machine Learning; Pattern Discovery; Association Rules; ENEM*

Cite as: Alves, A. A., & Cortes, O. A. C. (2026). Unveiling Patterns in Maranhão's 2023 ENEM Results Through Unsupervised Machine Learning. Revista Brasileira de Informática na Educação, vol. 34, pp. 404–428. https://doi.org/10.5753/rbie.2026.6215.

1 Introduction

The state of Maranhão in Brazil emerges as the object of study in this analysis, with a challenging scenario depicted by student performance in the ENEM, the results achieved in the Basic Education Development Index (IDEB), and a student body served mainly by the state public education network (Inep, 2024a). Furthermore, the scarcity of educational studies that include the state of Maranhão and utilize the ENEM database may be relevant for investigating other influential variables for performance (Dutra et al., 2023). In this context, the following research question: How are socioeconomic attributes, such as family income, parental education, access to technology (computers and internet), and gender, associated with the performance of Maranhão students in the 2023 ENEM?

The study's objective is to investigate the factors related to the performance of Maranhão students in the 2023 National High School Exam (ENEM) from the perspective of Educational Data Mining using Recursive Feature Elimination (RFE) with Random Forest to select the most critical variables, followed by association rule mining via FP-Growth algorithm. Evaluation metrics, including support, confidence, and lift, ensure the interpretability and relevance of rules. This hybrid pipeline, combining supervised feature selection with unsupervised pattern discovery, enhances scalability and replicability across large educational datasets.

ENEM microdata, released on April 30, 2024 (Inep, 2024a), serve as a powerful instrument for informing public policy (Travitzki, 2021). We begin with exploratory statistical analysis to characterize dominant trends (Marconi & Lakatos, 2022), followed by association rule extraction to reveal conditional and probabilistic relationships not visible through descriptive statistics alone (L. A. Silva et al., 2016). The FP-Growth algorithm was selected over Apriori due to its efficiency in handling large transactional datasets, such as ENEM, which requires only two database scans and avoids candidate generation (Han et al., 2000).

In this context, this investigation makes the following contributions: (i) it focuses on Maranhão, a socially vulnerable and understudied region; (ii) it prioritizes unsupervised learning for descriptive analytics, contrasting with the predictive dominance in EDM; (iii) it identifies gender-specific disparities, such as female students linked to lower performance and socioeconomic disadvantage, an underexplored dimension in prior ENEM studies (Souza et al., 2022) (Dutra et al., 2023); (iv) it proposes a scalable analytical framework adaptable to other developing regions through data filtering, parameter tuning, and open-source implementation; and (v) it delivers actionable policy insights for resource allocation, digital inclusion, and equity monitoring.

Thus, this article is devised as follows: Section 2 explores the related works initially considering unsupervised machine learning, then analyzing works that deal with similar datasets; Section 3 describes the necessary theory, including association rules and the methodology; Section 4 describes how the model was obtained using all stages of CRISP-DM and discusses the results; finally, Section 5 points out the conclusions and future works.

2 Related Work

This section reviews prior research in educational data mining (EDM), with a focus on studies that utilize supervised and unsupervised machine learning, as well as those analyzing the microdata of Brazil's National High School Exam (ENEM). We organize the discussion into three areas: supervised machine learning for educational prediction, unsupervised learning for pattern discovery, and ENEM-based studies exploring socioeconomic influences.

While these studies provide valuable insights, our work differs in that we apply unsupervised learning (association rule mining via the FP-Growth algorithm) to uncover latent patterns in Maranhão's 2023 ENEM results, emphasizing gender-specific findings and policy-relevant insights in an underrepresented region, even though sometimes, studies on gender differences in the field of learning achievement are controversial (Merkys et al., 2025).

Additionally, our application of the CRISP-DM framework, combined with Recursive Feature Elimination (RFE), Random Forest, and Association Rules, offers a novel and scalable approach for descriptive analytics that can be used to guide public policies in socially vulnerable areas.

2.1 Supervised Machine Learning in Educational Analytics

Supervised machine learning is prevalent in EDM for predicting educational outcomes. Munim et al. (2025) developed a learning analytics dashboard using XGBoost to predict performance in maritime simulation training. Alalawi et al. (2024) employed five supervised machine learning algorithms (Logistic Regression, SVM, Decision Tree, k-Nearest Neighbor, and Naive Bayes) to predict student assessments, integrating pedagogical frameworks.

Vaarma and Li (2024) used supervised methods (SVM, Neural Networks, Random Forest) to predict student dropout in Finnish higher education. These studies prioritize predictive accuracy using labeled data, unlike our unsupervised approach, which uses the FP-Growth algorithm to discover patterns in unlabeled ENEM 2023 microdata from Maranhão. Additionally, their contexts (maritime training and Finnish education) differ from Maranhão's socioeconomically challenged public education system, limiting their relevance to our regional and descriptive focus.

2.2 Unsupervised Learning for Pattern Discovery

Unsupervised learning, though less common in EDM, excels at identifying hidden patterns. Ouassif and Ziani (2025) combined association rules with neural networks to recommend academic paths, while Kaur et al. (2024) utilized association rule mining to identify PhD students who dropped out.

Clustering approaches include Martin et al. (2024) work, who developed a rubric for student reasoning in chemistry; Ma et al. (2024), who grouped readers by engagement; and Alvarez-Garcia et al. (2024), who profiled PISA 2022 students, linking ICT access and gender to performance.

2.3 ENEM-Based Studies and Socioeconomic Factors

Assessment of student's competence can be performed by using various social science research methods: observation, interview, survey, testing, content analysis, fact analysis, competence portfolio, performance portfolio, reflection diary (online blog), etc.

ENEM microdata studies often explore socioeconomic influences but rarely focus on Maranhão. Lima and Brighenti (2023) analyzed 2019 ENEM data from Minas Gerais, linking private school attendance, higher income, and parental education to better scores using exploratory analysis. Souza et al. (2022) examined 2015–2019 ENEM data from Brazil's Northeast, using regression models to confirm the positive impact of parental education and access to technology, as well as the adverse effect of low income. Soares et al. (2023) applied CRISP-DM to Maranhão's IDEB and Saeb data, identifying parental education and technology access as key factors, though limited by data incompleteness. Dutra et al. (2023) reviewed ENEM studies, noting socioeconomic predictors (income, parental education, and gender) and the dominance of supervised learning.

Despite some limitations due to the absence of data for some schools, Table 1 summarizes the variables identified, along with their contributions to performance (positive or negative). A positive contribution indicates a direct association, while a negative contribution suggests an inverse relationship. As we can see, most of them contribute positively to performance, except for *Age* and *Number of Residents in the Household*.

Table 1: Variables versus Performance.

Variable	**Contribution**
Level of parental education (Dutra et al., 2023; Netto & Maciel, 2021; Soares et al., 2023; Souza et al., 2022)	+
Type of school (Dutra et al., 2023; Gomes et al., 2017; L. A. Silva et al., 2016; Souza et al., 2022)	+
Administrative dependence of the school (Dutra et al., 2023; V. A. A. Silva et al., 2020)	+
Location of the school (Souza et al., 2022)	+
Location of the residence (Gomes et al., 2017)	+
Race / Ethnicity (Dutra et al., 2023; Souza et al., 2022)	+
Income (Dutra et al., 2023; Gomes et al., 2017; L. A. Silva et al., 2014; V. A. A. Silva et al., 2020; Souza et al., 2022)	+
Computer (V. A. A. Silva et al., 2020; Soares et al., 2023; Souza et al., 2022)	+
Cell phone (Souza et al., 2022)	+
Internet access (Soares et al., 2023; Souza et al., 2022)	+
Age (Souza et al., 2022)	-
Number of residents in the household (L. A. Silva et al., 2014; V. A. A. Silva et al., 2020; Souza et al., 2022)	-

Regarding the extraction of association rules in the context of Educational Data Mining (EDM) in ENEM, only the adoption of the Apriori algorithm (Srikant & Agrawal, 1997) was observed in different simulations and configurations for the minimum support and confidence hyperparameters, requiring the analyst to have greater mastery of the data to establish them and perform a qualitative analysis of the rules. Table 2 presents a summary of the experiment configurations observed for EDM of the ENEM when extracting association rules with the Apriori algorithm, in which $minsup$ is the minimum support of the antecedent and $minconf$ is the minimum confidence of the entire rule.

Thus, based on this literature review, we adopted these parameter configurations to ensure methodological consistency and comparability with prior ENEM studies. Specifically, we selected minimum support values ranging from 10% to 50% and minimum confidence thresholds between 70% and 80%. These ranges were chosen because they have been empirically proven effective in previous works such as Gomes et al. (2017), L. A. Silva et al. (2014), and V. A. A. Silva et al. (2020) for balancing the trade-off between rule quantity and the significance of the discovered patterns.

Table 2: Experiment Configurations with the Apriori Algorithm.

Base	**Observations**	**Variables**	**minsup**	**minconf**	**Rules**
ENEM 2010 (L. A. Silva et al., 2014)	310,000	5	50%	70%	2
			30%	70%	24
			25%	70%	42
			15%	70%	214
ENEM 2014 (Gomes et al., 2017)	2,456,272	24	10%	80%	-
	432,966	24	10%	80%	-
ENEM 2013 (Gomes et al., 2017)	-	24	10%	80%	-
	-	24	10%	80%	-
ENEM 2019 (V. A. A. Silva et al., 2020)	88,659	7	20%	70%	80
	46,419	7	20%	70%	-
	40,240	7	20%	70%	-

2.4 Our Proposal

Our investigation diverges from prior work in several ways. First, we prioritize unsupervised learning (FP-Growth) over the predictive, supervised approaches dominant in EDM (e.g., Munim et al. (2025), Vaarma and Li (2024)), enabling pattern discovery in unlabeled ENEM data.

Second, our focus on Maranhão addresses a gap, as most ENEM studies target broader regions, such as the Northeast, Minas Gerais or lack specificity (Dutra et al., 2023). Third, our methodological framework, which combines CRISP-DM, RFE with Random Forest, and FP-Growth, offers a novel and replicable approach for descriptive analytics, unlike the exploratory (Lima & Brighenti, 2023) or mixed-methods (Soares et al., 2023) approaches employed in prior ENEM studies. Fourth, we emphasize gender-specific findings. For instance, female participants are linked to lower performance and socioeconomic disadvantages (Dutra et al., 2023; Souza et al., 2022). However, these inequalities remain underexplored in the context of Maranhão using unsupervised learning.

Finally, our descriptive insights target policy-making, such as digital inclusion and resource allocation, in Maranhão's socioeconomically challenged context, with a scalable framework applicable to other developing regions, enhancing its practical and academic relevance.

3 Materials and Methods

3.1 Machine Learning

Machine Learning resides in algorithms' ability to extract patterns from data. This process involves obtaining mathematical models that best fit the data, which may require interventions through techniques that enhance calibration, such as data treatment, transformation, and resampling. The quality of these models can be evaluated using specific metrics, which will better describe the knowledge derived from the data (Carvalho et al., 2024). Additionally, Gabriel, 2024 defines Machine Learning as a subarea of Artificial Intelligence, which is dedicated to performing tasks previously exclusive to human skills, thus being extremely broad, encompassing a wide variety of algorithms and methodologies whose improvement depends on data processing, including Deep Learning, based on Artificial Neural Networks (ANNs), as illustrated in Figure 1.

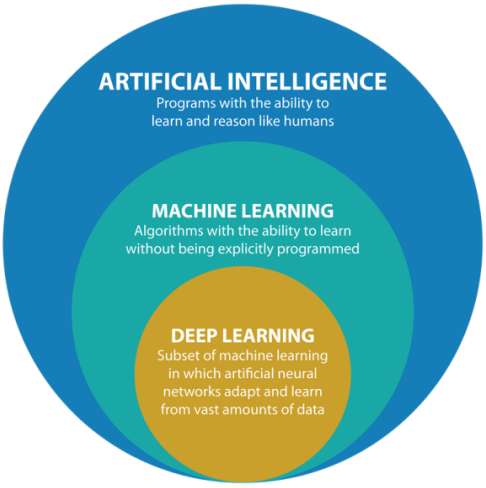


Figure 1: Artificial Intelligence and Sub-fields.

Data training is a crucial element of learning and can adopt four paradigms: supervised, unsupervised, semi-supervised, and reinforcement (Carvalho et al., 2024; Gabriel, 2024). In the first case, supervised learning, the model is trained on a dataset where the desired output is known. This search for patterns that relate the input attributes to the output variable can occur through Classification algorithms to classify or predict the class, category, or target of an instance or Regression to predict values when the attributes are continuous and dependent numerical data (Netto & Maciel, 2021).

In unsupervised learning, the model is trained on a dataset with no labels (or categories). The goal is to find patterns in the data, such as natural clusters or underlying structures. In the semi-supervised case, as only a portion of the dataset is labeled, the model is trained on this subset to predict the class of unlabeled instances. This task can result in data redundancy, which can be mitigated by the active learning strategy, in which an "oracle" labels the most imprecise instances based on specific criteria (Carvalho et al., 2024; Faceli et al., 2024). Unlike the unsupervised approach, reinforcement learning trains an agent to maximize cumulative reward through trial and error (Gabriel, 2024).

Depending on the type of task performed, algorithms can be classified as predictive and descriptive. In predictive tasks, algorithms are applied to labeled data to create models that predict the value of an attribute for new (unknown) data, using supervised learning for classification or prediction (regression) of a class. In descriptive tasks, algorithms extract patterns from data without predicting specific values using unsupervised learning. Examples of descriptive tasks include data clustering, and identification of association rules (Carvalho et al., 2024; Faceli et al., 2024). Thus, the data mining process can be summarized as follows (L. A. Silva et al., 2016):

- Classification: corresponds to correctly identifying the class or category of an observation in a dataset formed by a discrete and finite number of categories.
- Regression: consists of estimating a continuous numerical value for an output variable based on input variables, in a dataset with numerical and constant values.
- Clustering: an unsupervised strategy that organizes observations into groups (clusters) based on similarities.
- Association (or discovery of association rules): an unsupervised strategy that finds frequent relationships between attributes in a dataset.

These points are highly relevant, as they are directly associated with the choice of an algorithm according to the nature of the data and the type of problem to be solved, that is, defining the mathematical model that establishes the relationships between the underlying attributes of the data (Netto & Maciel, 2021).

3.2 Association Rules

Association rule mining is a technique used to establish relationships in the form of rules between items in a transactional database; that is, instead of the conventional rows for observations and columns for attributes, a database is worked with where each observation is converted into a transaction. The goal is to find strict rules, according to some measure of interest, differentiating itself from clustering (similarity search) and correlation (evaluation of linear dependence between items), by noting that the occurrence of one item implies the other, that is when one occurs, the other also occurs (Castro & Ferrari, 2016).

Let $A = \{a_1, \ldots, a_m\}$ be the universe of m items, I any subset of items such that $I \subset A$, and T the database converted into n transactions $T = \{t_1, \ldots, t_n\}$. Each transaction is a pair $(\text{tid}_i, \text{k-items}_i)$ where tid_i is the transaction identifier and $\text{k-items}_i \subset A$ is a set of k items. The set $K(I)$ of transactions that support the itemset I is designated as the support of the itemset. A set I of items is considered frequent if the support of I, defined as the proportion of transactions in T that contain I, is greater than or equal to a predefined minimum support value. From the frequent sets, association rules are derived that have the form "if antecedent then consequent", where both the antecedent and the consequent are frequent sets of items (Faceli et al., 2024).

The concept of support is associated with the frequency of the element in the transaction, whether it is the item or the rule. Let $\sigma(A)$ the support count of items in set A, as well as the support count of the rule, support the rule and describe the frequency with which all items of the

two sets of items, A and C, appear together in individual database transactions (L. A. Silva et al., 2016). The support of the rule $A \rightarrow C$ can be given concerning the total number of transactions in the database T according to Equation 1:

$$\text{Sup}(A \rightarrow C) = P(A \cup C) = \frac{\sigma(A \cup C)}{\sigma(T)} = \frac{\sigma(A \cup C)}{n} \tag{1}$$

The confidence of the rule $A \rightarrow C$ refers to the possibility of the rule occurring concerning the support of its antecedent A, which can also be described as the conditional probability $P(C|A)$ (L. A. Silva et al., 2016), according to Equation 2 below:

$$\text{Conf}(A \rightarrow C) = P(C|A) = \frac{\sigma(A \cup C)}{\sigma(A)} \tag{2}$$

The analysis or selection of association rules can still be refined by the correlation measure called *lift*, which consists of the relationship between the *Confidence* of the *Rule* and its consequent, according to Equation 3:

$$\text{Lift}(A \rightarrow C) = \frac{\text{Conf}(A \rightarrow C)}{P(C)} = \frac{P(A \cup C)}{P(A)P(C)} \tag{3}$$

The measure, (*lift*), assumes values from 0 to infinity and indicates the importance of the association rule. For values below 1, there is a negative correlation between the items of A and C, a positive correlation for values above 1, and no correlation for values close to one (L. A. Silva et al., 2016).

The two classic association-rule algorithms are *Apriori* (Srikant & Agrawal, 1997), which repeatedly scans the database to generate candidate itemsets, and *FP-Growth* (Han et al., 2000), which builds a compact FP-tree and mines rules in only two scans—ideal for the million-row ENEM dataset. Both algorithms require nominal attributes (or one-hot encoding) (L. A. Silva et al., 2016) and follow the frequent-itemset $\rightarrow$ rule-generation pipeline (Faceli et al., 2024).

This study adopts the *FP-Growth* algorithm (Han et al., 2000), which is more efficient than Apriori (Srikant & Agrawal, 1997) due to the approach of compressing the database into a tree structure (*FP-Tree*), reducing the search space and candidates for frequent patterns (Castro & Ferrari, 2016).

3.3 Methodology

The CRISP-DM (Cross-Industry Standard Process for Data Mining) was proposed as a promising model for knowledge extraction from databases in response to industries' growing demands, as illustrated in Figure 2 (Carvalho et al., 2024; Goldschmidt et al., 2015), in 1996.

Considering the detailing of the CRISP-DM phases as a non-rigid knowledge extraction process from databases and its suitability for the educational field, the methodological procedures appropriate for this study are summarized as follows:

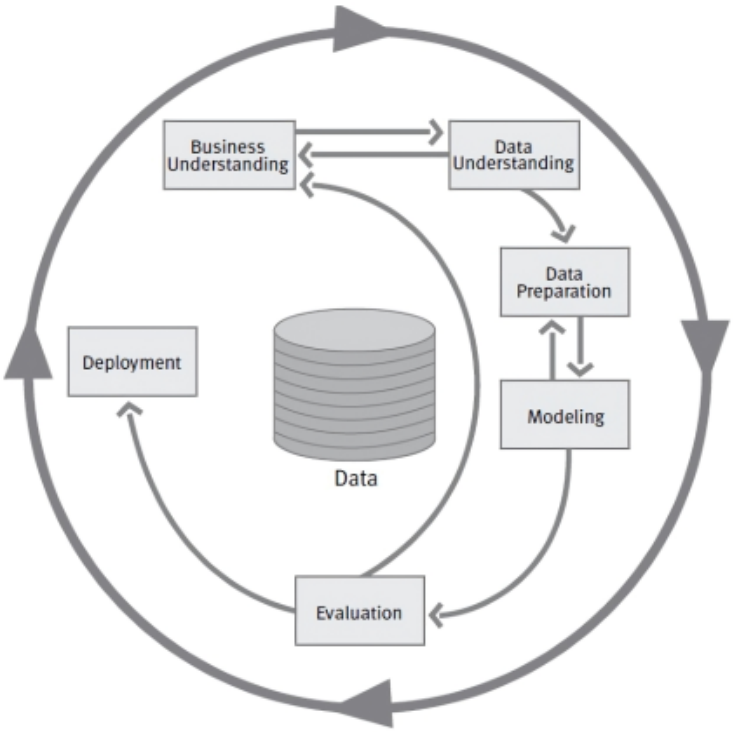


Figure 2: CRISP-DM (Chapman et al., 2000).

- Domain Understanding: encompasses the general information about the topic and the perspectives that transform the performance of Maranhão's participants in the 2023 ENEM into a data science problem, as well as the study's objectives.
- Data Understanding: source and nature of the data, collection of the 2023 ENEM microdata (Inep, 2024a), analysis of the metadata file (dictionary), identification of attributes and variable types.
- Data Preparation: cleaning, treatment, data transformation, categorization according to the metadata file, extraction of the subset of Maranhão's participants, as well as the selection of relevant attributes, through the manual removal of inconsistent or unnecessary attributes for the study (Faceli et al., 2024), and the subsequent application of the RFE algorithm for the selection of 10 (ten) attributes out of the remaining 33 (thirty-three), using the Random Forest algorithm, by eliminating 1 (one) attribute at each iteration, with the performance variable (score) as the target (Carvalho et al., 2024).
- Exploratory Analysis: seeks an in-depth view of the data, ascertaining the profile of the participants and providing an overview of performance.
- Modeling: involves the application of unsupervised machine learning techniques for association rules to identify niches and detect patterns not observed in the exploratory analysis, using the FP-Growth algorithm (Han et al., 2000), according to the experiment configurations listed in Table 3, for $lift > 1$, in which *minsup* is the minimum support of the antecedent and $minconf$ is the minimum confidence of the entire rule, as previously stated.

Table 3: Experiment Setup.

EXP.	minsup	minconf	antecedents
1	3%	5%	[2,4]
2	10%	5%	[2,4]
3	20%	10%	[2,4]
4	25%	20%	[2,4]
5	3%	5%	[3,6]
6	10%	10%	[3,6]
7	20%	10%	[3,6]

- Evaluation: assessment of the quality of the generated rules based on the metrics of support, confidence, and lift, and interpretation of the results.
- Deployment: Discussion of the findings related to the performance of Maranhão's participants based on the processed rules.

Thus, the CRISP-DM process, adapted to analyze educational data from Maranhão's participants in the 2023 ENEM, will guide the extraction of valuable knowledge that can inform improvements in education in the state.

4 Computational Experiments

4.1 Domain Understanding

Education is listed among the social rights provided in the 1988 Brazilian Federal Constitution, making it the State's duty to provide public, free education with minimum quality standards and promote access to higher levels of education (Brazil, 1996).

Additionally, the National Education Plan (PNE) serves as the guiding instrument for investments in education. It outlines the guidelines, objectives, goals, and strategies for national education over ten years, including important indicators such as the IDEB. Recent IDEB evaluation cycles have shown progress in various areas, but Brazil still falls short of the targets set for basic education (Inep, 2024b).

Recent results of the Ministry of Education (MEC) indicated that Maranhão did not meet the IDEB targets for 2021 in the three stages of basic education, with high school showing the worst performance (Ministério da Educação, 2024). In the 2023 edition of the primary assessment of basic education, the ENEM, Maranhão had the lowest average result in the Northeast region (Inep, 2024a).

Given this scenario, Maranhão becomes a relevant object of study for analyzing educational indicators, primarily through open educational data, which has supported various studies aimed at evaluating educational quality indicators, understanding the factors that impact learning, supporting decision-making, and strengthening public policies.

4.2 Data Gathering

The ENEM microdata is obtained through INEP's annual disclosure in an anonymized format, in compliance with the General Data Protection Law (LGPD). Given the express authorization for free use, this data does not require authorization to use (Brazil, 2016). This is a set of secondary, structured data compiled in a ".csv" spreadsheet format, facilitating identification, storage, and processing (Carvalho et al., 2024). The files used were: 1) `MICRODADOS_ENEM_2023.csv`; 2) `Dicionario_Microdados_Enem_2023.xlsx`; and 3) `Leia_Me_Enem_2023` (glossary of technical terms) (Inep, 2024a).

The file `MICRODADOS_ENEM_2023.csv` (Inep, 2024a) contains 76 attributes regarding each ENEM 2023 participant. The details of these variables, including their description, categorization, and type, are declared in the metadata file `Dicionario_Microdados_Enem_2023.xlsx` divided into six groups: 1) Participant Data (12 variables), 2) School Data (7 variables), 3) Test Location Data (4 variables), 4) Objective Test Data (21 variables), 5) Essay Data (7 variables), and 6) Socioeconomic Questionnaire Data (25 variables).

4.3 Data Pre-processing

The steps of data import, preliminary treatment, and creation of a new subset of data containing only Maranhão's participants were carried out in Jupyter Notebook version 3.11.7. The import of the file `MICRODADOS_ENEM_2023.csv`, with a size of 1.65 GB (1,777,162,429 bytes), was done using two arguments: the first one to use the Latin encoding standard "ISO-8859-1" and the second one for the ".csv" file extension, explicitly using ";" as a separator. The data was stored in a two-dimensional data structure called *dataframe* with the following dimensions: 3,933,955 rows by 76 columns.

Subsequently, all data in the dataframe were converted to the categorical type, following the metadata file `Dicionario_Microdados_Enem_2023.xlsx`, except for the scores of the objective test and essay, which remained as numerical (float) data. Table 4 lists the attributes with missing data related to School Data, such as school municipality (location), which included only information relevant to high school graduates, *i.e.*, those in their final year of high school, which reduced the set of variables investigated.

Table 4: Variables Missing Data.

Variable	**Meaning**	**Missing Data (%)**
TP_ENSINO	Type of education offered	2,594,874 (66.0%)
CO_MUNICIPIO_ESC	IBGE city code of the school	2,975,449 (75.6%)
NO_MUNICIPIO_ESC	Name of the municipality	2,975,449 (75.6%)
CO_UF_ESC	IBGE state code of the school	2,975,449 (75.6%)
SG_UF_ESC	State acronym	2,975,449 (75.6%)
TP_DEPENDENCIA_ADM_ESC	Administrative dependency	2,975,449 (75.6%)
TP_LOCALIZACAO_ESC	Urban or rural area	2,975,449 (75.6%)
TP_SIT_FUNC_ESC	School operating status	2,975,449 (75.6%)

Table 5 presents the variables that participate only in encoding another attribute or technical information about the test application and, for this reason, were removed from the analysis. Observations with missing numerical data and null scores were also removed. Moreover, it is

important to clarify that missing values reflect data that were not provided in the original public dataset.

A variable named NU_NOTA_FINAL was added, resulting from the arithmetic average of the scores in the five knowledge areas of the test: NU_NOTA_CN (Natural Sciences), NU_NOTA_CH (Human Sciences), NU_NOTA_LC (Languages and Codes), NU_NOTA_MT (Mathematics and its technologies), and NU_NOTA_REDACAO (Essay). This last variable was categorized into quartiles, following a strategy observed in (L. A. Silva et al., 2014; V. A. A. Silva et al., 2020), as shown in Table 6.

Table 5: Redundant or Unnecessary Variables.

N	Variable	Description
01	NU_INSCRICAO	Registration number
02	NU_ANO	Year of the ENEM
03	CO_MUNICIPIO_PROVA	Municipality code of the test location
04	CO_UF_PROVA	State code of the test location
05	CO_PROVA_CN	Code for the Natural Sciences test type
06	CO_PROVA_CH	Code for the Human Sciences test type
07	CO_PROVA_LC	Code for the Languages and Codes test type
08	CO_PROVA_MT	Code for the Mathematics test type
09	TX_RESPOSTAS_CN	Vector with the answers to the objective part of the Natural Sciences test
10	TX_RESPOSTAS_CH	Vector with the answers to the objective part of the Human Sciences test
11	TX_RESPOSTAS_LC	Vector with the answers to the objective part of the Languages and Codes test
12	TX_RESPOSTAS_MT	Vector with the answers to the objective part of the Mathematics test
13	TX_GABARITO_CN	Vector with the answer key for the objective part of the Natural Sciences test
14	TX_GABARITO_CH	Vector with the answer key for the objective part of the Human Sciences test
15	TX_GABARITO_LC	Vector with the answer key for the objective part of the Languages and Codes test
16	TX_GABARITO_MT	Vector with the answer key for the objective part of the Mathematics test
17	TP_STATUS_REDACAO	Essay status of the participant
18	NU_NOTA_COMP1	Score for competency 1 - Essay
19	NU_NOTA_COMP2	Score for competency 2 - Essay
20	NU_NOTA_COMP3	Score for competency 3 - Essay
21	NU_NOTA_COMP4	Score for competency 4 - Essay
22	NU_NOTA_COMP5	Score for competency 5 - Essay

Table 6: Categorization of Score Variables.

	count	mean	std	min	25%	50%	75%	max
A	26,533	415.6	26.8	281.7	398.5	420.6	437.3	452.3
B	26,512	479.26	15.3	452.3	466.0	479.3	492.4	505.9
C	26,522	535.66	17.9	505.9	520.2	534.9	550.7	568.6
D	26,521	628.5	48.1	568.6	589.2	616.6	656.9	828.9

Thus, the categories A, B, C, and D of the score variable had the following intervals: ()A) from 281.7 to 452.3 points, (B): from 452.3 to 505.9 points, (C): from 505.9 to 568.6 points, and (D): from 568.6 to 828.9 points. The selection of attributes included:

- Manual removal of inconsistent or unnecessary attributes for the study, which include the variables `TP_ESCOLA` (inconsistent data), those listed in Tables 4 and 5, and numerical data.
- Application of the Recursive Feature Elimination (RFE) algorithm to select 10 (ten) attributes out of the remaining 33 (thirty-three), using the Random Forest algorithm, by eliminating 1 (one) attribute at each iteration, with the "SCORE" variable as the target.

RFE is a greedy technique that seeks to identify the most relevant subset of attributes, starting from the entire set and removing the least important ones at each iteration until the required number of attributes is obtained (Sayak, 2020). In this study, RFE is combined with the Random Forest algorithm, a powerful tool for handling large datasets due to its ability to assess the importance of each variable for model construction. Table 7 details the variables selected with RFE.

Table 7: Selected Variable Using RFE.

Variable	**Description**
TP_FAIXA_ETARIA	Age group
TP_SEXO	Gender
TP_ST_CONCLUSAO	High school completion status
Q001	Level of education of the father or male guardian of the participant
Q003	Occupational group of the father or male guardian of the participant
Q005	Number of residents in the household (including the participant)
Q009	Number of bedrooms
Q011	Number of motorcycles
Q014	Number of washing machines
Q024	Number of computers

The selection of ten attributes at this stage is justified by the reduction in computational cost and interpretability of the results, considering the number of variables studied in related works on association rule extraction using the Apriori algorithm with the ENEM database.

4.4 Exploratory Analysis

The national exploratory analysis revealed that the age profile of participants is predominantly young adults, up to 19 years old, based on the age of the registrant as of 12/31/2023, with approximately 62% being female. Although whites predominate (43.6%), blacks, evidenced by the categories of blacks (11.7%) and mixed-race individuals (41.4%), represent the majority ethnic group of ENEM participants.

Participants in their final year of high school, belonging to the graduating (Concluintes) category, represent 39%, while dropouts (Egressos) represent the majority, with 41.7%. Participants

who would not complete high school in 2023, belonging to the high school (Cursistas) category, represent 18.9%.

The participation rate by Federative Unit, shown in Figure 3, highlights the states of the Southeast region, led by São Paulo and Minas Gerais. Maranhão (MA) ranks 10th nationally and 4th among the Northeastern states, considering participants who did not miss any test days and had no zero scores on the tests.

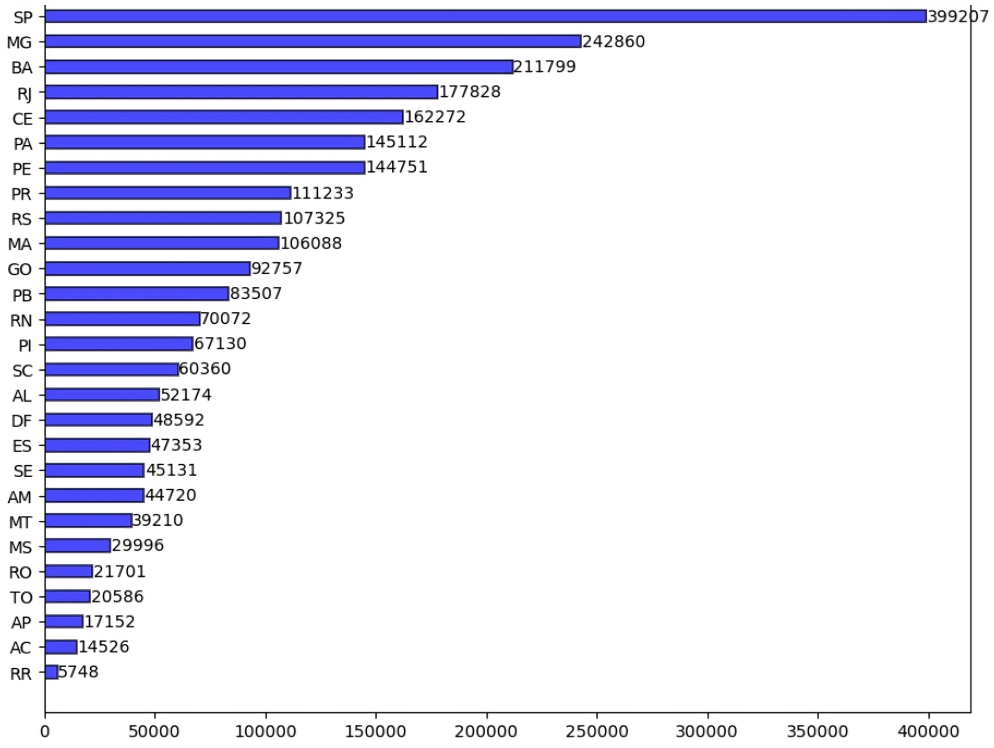


Figure 3: Participants by Federation Unit.

Using the “qualified” mean of participants as a ranking index, Figure 4 shows Maranhão in the second-to-last position nationally, with an average score of 514.7 points. This finding highlights Maranhão as a key case for understanding how the socioeconomic characteristics structured and disclosed by INEP through the ENEM microdata are associated with the performance of these participants.

In the Maranhão context, the state-level exploratory analysis revealed that the majority of participants are between 17 and 19 years old (57%), with around 63% being female. Regarding ethnic characterization, the majority are Black, as evidenced by the categories of Black (13.8%) and mixed-race (61.3%) participants, with 22.3% being ”White“, which diverges from the national predominance.

Graduates represented 33.5%, while former students comprised the majority at 47.3%. Meanwhile, high school students accounted for 18.9%. Regarding socioeconomic variables, most mothers did not complete higher education (39.6%), and a significant number had not completed basic education (37.3%). The monthly family income is concentrated in the lowest brackets: no income (11.2%), up to R$1,320.00 (46.3%), and R$1,320.00 up to R$1,980.00 (15.4%).

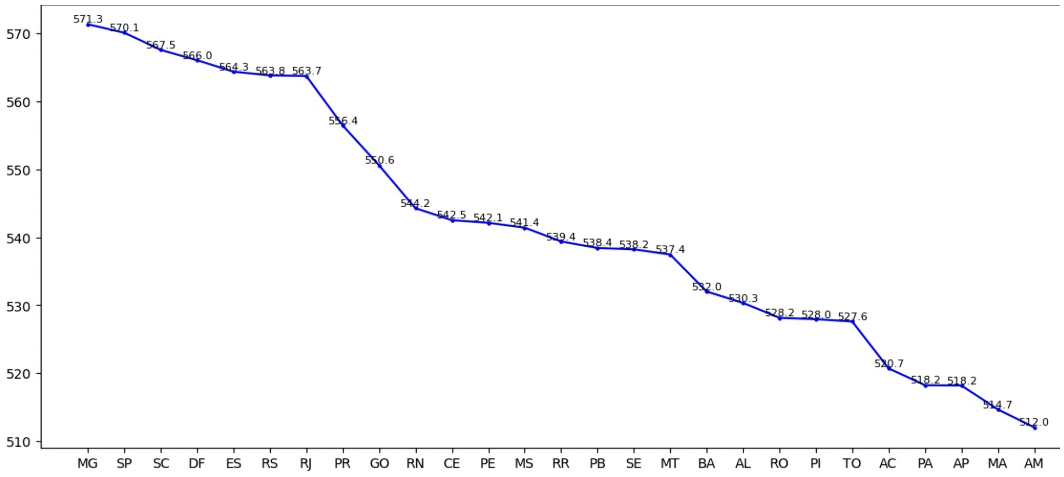
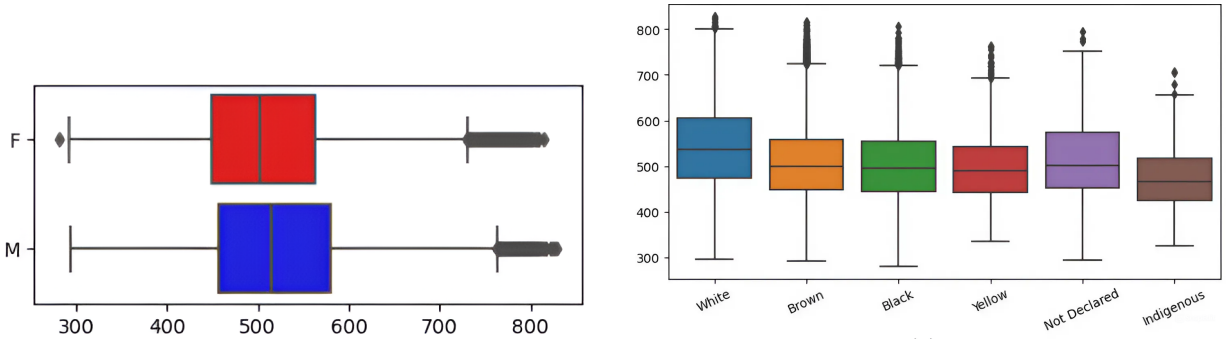


Figure 4: Mean Performance by Federation Unit.

The exploratory analysis also revealed that the majority (71%) do not own a computer at home, although a significant number have internet access at home (81.6%). When performance behavior is compared according to gender, as shown in Figure 5a, or ethnicity, as shown in Figure 5b, male and White participants achieve better results.



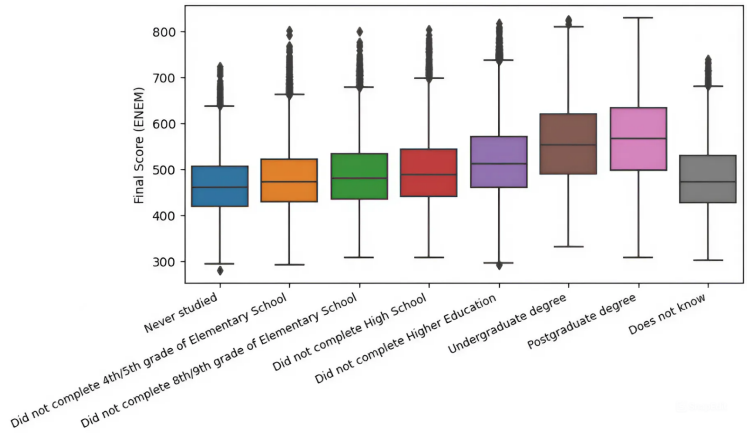
(a) Gender (b) Ethnic

Figure 5: Scores by Gender and Ethnicity.

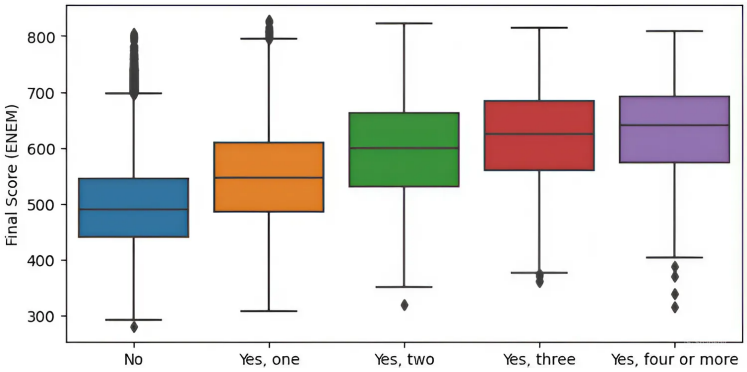
Similarly, maternal education and computer ownership positively correlated with score improvement, as illustrated in Figures 6a and 6b.

The nature of the participant's father's or male guardian's occupation appeared among the variables selected by RFE. Descriptive analysis indicates that the predominant occupation of fathers is related to agricultural or extractive activities, such as farming, livestock, and fishing, which are classified as Group 1 occupations, accounting for approximately 33%. This variable also showed a positive association with performance.

In summary, the exploratory analysis of socioeconomic factors aligns with findings positively related to participant performance in related studies, such as age, gender, ethnicity, parental education level, monthly family income (Dutra et al., 2023), computer ownership, and internet access at home (Soares et al., 2023; Souza et al., 2022).



(a) Maternal Education



(b) Computer

Figure 6: Scores by Maternal Education and Computer Ownership.

4.5 Modeling

For data modeling, the dataset was restricted to the state of Maranhão, based on the variable `SG_UF_PROVA` with a total of 106,088 (one hundred six thousand and eighty-eight) observations, considering the variables resulting from the Attribute Selection stage, listed in Table 7, and the variable `SCORE`.

The categorization of variables was adjusted slightly to facilitate identification. Table 8 summarizes the experiments conducted, according to the parameters of minimum support and confidence, number of antecedents, and rules generated, with $lift > 1$. As we can see, the lower the support and confidence, the higher the number of rules. Therefore, exploratory analysis, verification of frequent items, and the proper use of evaluation metrics are crucial for inspecting the generated rules.

4.6 Results and Discussion

The large number of generated rules can be addressed by refining them based on the selected metrics, particularly lift, which helps verify whether the relationship is significant (L. A. Silva et al., 2016). While Experiment 1 produced excessive rules, other experiments contain redundancy,

Table 8: Experiment Setup.

EXP.	minsup	minconf	antecedents
1	3%	5%	[2,4]
2	10%	5%	[2,4]
3	20%	10%	[2,4]
4	25%	20%	[2,4]
5	3%	5%	[3,6]
6	10%	10%	[3,6]
7	20%	10%	[3,6]

demonstrating that the support hyperparameter is decisive for pruning associations. In this context, we present the most relevant ones in Table 8.

In Experiment 1, the analysis of the top 150 most relevant rules based on the lift, which is related to the variable SCORE, is listed in Table 9, in which Rule id01 reveals a significant pattern with a lift of 2.88. It indicates that candidates whose fathers belong to occupational Group 4, such as teachers, technicians, small business owners, and who do not own a motorcycle, are strongly associated with achieving performance level “D” and owning a washing machine. In practical terms, this rule suggests that a specific socioeconomic stratum, characterized by middle-income occupations and possession of certain household appliances, is positively associated with academic outcomes, differentiating these students from lower-performing groups.

Similarly, Rule id03 reinforces the relationship between digital inclusion and academic performance, in which owning a computer is strongly linked to better performance. This result aligns with our exploratory analysis, which showed that 71% of participants in Maranhão lack computer access, highlighting the digital divide as a key dimension in educational inequality.

Table 9: Association Rules From Experiment 1.

id	antecedents	consequents	support	conf.	lift
01	(Group 4, No Motorcycle)	(D, One Washing Machine)	0.032	0.347	2.877
02	(D, One Washing Machine)	(Group 4, No Motorcycle)	0.032	0.268	2.877
03	(Graduates, No Computer)	(A, No Washing Machine, 18)	0.033	0.127	2.793
04	(A, No Washing Machine, 18)	(Graduates, No Computer)	0.033	0.714	2.793

Table 10 refers to Experiment 2, in which the analysis of the top 150 rules based on lift revealed rules related to the variable SCORE and the gender of the candidates, listing rules that associated the group of female participants with low performance (categories ”A“ or ”B“) or less favorable socioeconomic conditions, such as the father belonging to Group 1 of occupational nature, the absence of a computer, and the absence of a washing machine.

Table 10: Association Rules From Experiment 2.

id	antecedents	consequents	support	conf.	lift
01	(No Washing Machine, A)	(Group 1)	0.102	0.537	1.616
02	(No Computer, A)	(Group 1)	0.111	0.517	1.559
03	(Group 1, No Computer)	(A)	0.111	0.387	1.545
04	(Group 1, No Washing Machine)	(A)	0.102	0.371	1.481

Rules id01 and id02 associated the lowest performance level (category ”A“) with Group 1 of the father's occupational nature, with support of around 11%, confidence of 54% and 52%, and lift of 1.62 and 1.56, respectively, as shown in Table 10. These subsets occur in approximately 11% of transactions. This also suggests that when the candidate does not own a computer and has a performance “A”, the father belongs to Group 1 of occupational nature in more than half of the cases. The lower performance (category “A”) is related to the occupational nature of Group 1, possibly indicating a relationship with socioeconomic factors. Rules id03 and id04 indicate that the father belongs to Group 1 of occupational nature, suggesting the candidate's low performance (level ”A“) in about 38% of cases, with a lift of 1.54 and 1.48, respectively.

Results of Experiment 3 are shown in Table 11, in which none of the 87 (eighty-seven) generated rules were associated with the variable `SCORE`. Instead, the rules indicate less favorable socioeconomic conditions involving female participants.

Table 11: Association Rules From Experiment 3.

id	antecedents	consequents	support	conf.	lift
01	(No Washing Machine, No Computer)	(Group 1)	0.251	0.445	1.339
02	(Group 1, No Computer)	(No Washing Machine)	0.251	0.873	1.292
03	(No Motorcycle, No Computer)	(No Washing Machine, Female)	0.261	0.577	1.281
04	(No Washing Machine, Female)	(No Motorcycle, No Computer)	0.261	0.580	1.281
05	(No Washing Machine, No Motorcycle)	(No Computer, Female)	0.261	0.604	1.280
06	(No Computer, Female)	(No Washing Machine, No Motorcycle)	0.261	0.554	1.280
07	(Group 1, No Washing Machine)	(No Computer)	0.251	0.904	1.272

The results of Experiment 4 are presented in Table 12, in which none of the 43 (forty-three) generated rules were associated with the variable `SCORE`. Similarly to Experiment 3, the strongest rules pointed to less favorable social conditions involving female participants. Additionally, rules id05, id06, and id07 associate the lack of a motorcycle, computer, and washing machine with female participants, with a confidence above 81% and a lift of 1.2. These rules suggest that a significant portion of transactions reveal associations reflecting the socioeconomic conditions of ENEM 2023 participants, especially females, with a high level of confidence.

Table 12: Association Rules from Experiment 4.

id	antecedents	consequents	support	conf.	lift
01	(No Motorcycle, No Computer)	(No Washing Machine, Female)	0.261	0.577	1.281
02	(No Washing Machine, Female)	(No Motorcycle, No Computer)	0.261	0.581	1.281
03	(No Washing Machine, No Motorcycle)	(No Computer, Female)	0.261	0.604	1.280
04	(No Computer, Female)	(No Washing Machine, No Motorcycle)	0.261	0.554	1.280
05	(No Motorcycle, No Computer, Female)	(No Washing Machine)	0.261	0.844	1.249
06	(No Computer, Female)	(No Washing Machine)	0.387	0.819	1.212
07	(No Washing Machine, Female)	(No Computer)	0.387	0.859	1.209

In Experiment 5, shown in Table 13, the low support and confidence levels returned rules with higher lift values and the most relevant rules considering the variables `SCORE` and gender. The top three rules ranked by lift linked female participants to less favorable socioeconomic conditions. For example, rule id01 indicates that a participant who does not own a computer and is under 17 years old is 6.4 times more likely to be the case when the participant does not own a washing machine, is a high school student, and is female.

Table 13: Association Rules from Experiment 5.

id	antecedents	consequents	support	confidence	lift
01	(No Washing Machine, High School Students, Female)	(No Computer, <17)	0.033	0.426	6.395
02	(High School Students, No Computer, Female)	(No Washing Machine, <17)	0.033	0.392	6.384
03	(<17, No Computer, Female)	(High School Students, No Washing Machine)	0.033	0.708	6.344
04	(A, No Washing Machine, 18)	(Graduates, No Computer)	0.033	0.715	2.794
05	(No Washing Machine, Graduates, No Computer)	(A, 18)	0.033	0.163	2.687
06	(No Washing Machine, Did Not Complete 4th/5th Grade, A)	(Group 1, No Computer)	0.035	0.671	2.336
07	(No Washing Machine, Did Not Complete 4th/5th Grade, A)	(Group 1, No Computer)	0.035	0.671	2.336
08	(D, One Washing Machine, No Motorcycle)	(Group 4)	0.033	0.405	2.638
09	(Group 4, One Washing Machine, No Motorcycle)	(D)	0.033	0.589	2.357

Also, in Table 13, rules id04, id05, and id06 introduced other variables related to low performance, such as age, high school completion status, and the father's or male guardian's education level. These rules point to vulnerable profiles, where socioeconomic factors are systematically associated with low academic performance of younger participants, *i.e.*, those in their final year of high school. The high lift in these rules (above 2.5) reflects a significant association between the analyzed variables. Rule id07 identifies a group of participants affected by both low education levels and the occupational nature of the father or male guardian, despite the low support. The confidence level of 67.1% and the lift of 2.34 reinforce the link between economic, social, and educational conditions associated with low performance. Rules id08 and id09 included rules that considered the best performance category in the ENEM – level "D" (between 568.58 and 828.86 points) – associated with Group 4 of the father's occupational nature.

In Experiment 6, Table 14, with 245 (two hundred and forty-five) rules, it was possible to identify Group 1 of the father's occupational nature, as well as different performance levels associated with female participants, with emphasis on rules id02 and id04, where confidence exceeds 80%.

Table 14: Association Rules from Experiment 6.

id	antecedents	consequent	support	confidence	lift
01	(No Washing Machine, No Computer, Female)	(Group 1, No_Mt)	0.111	0.287	1.526
02	(Group 1, No_Mt, Female)	(No Washing Machine, No Computer)	0.111	0.839	1.487
03	(Group 1, No_Mt, No_PC)	(No Washing Machine, Female)	0.111	0.653	1.451
04	(Group 1, Two Bedrooms, Female)	(No Washing Machine, No Computer)	0.107	0.807	1.431

Finally, only 12 rules were generated in the seventh experiment presented in Table 15. It is observed that the more rigorous parameterization returned rules with increasingly lower values for the lift metric. On the other hand, the confidence in the highlighted rules proved quite significant. According to $lift$, the first six rules presented confidence levels ranging from 83% to 86%. The remaining rules, in turn, returned lift values very close to 1, suggesting little or no influence of the antecedents on the consequent.

Table 15: Association Rules from Experiment 7.

id	antecedents	consequents	support	confidence	lift
01	(No Motorcycle, No Computer, Female)	(No Washing Machine)	0.261	0.843995	1.249
02	(No Washing Machine, No Motorcycle, Female)	(No Computer)	0.261	0.884537	1.245
03	(Two Bedrooms, No Computer, Female)	(No Washing Machine)	0.239	0.831133	1.230
04	(Two Bedrooms, No Motorcycle, No Computer)	(No Washing Machine)	0.229	0.830329	1.229
05	(Two Bedrooms, No Washing Machine, No Motorcycle)	(No Computer)	0.229	0.866700	1.220
06	(Two Bedrooms, No Washing Machine, Female)	(No Computer)	0.239	0.866564	1.220

Among other influential variables for performance, the number of residents in the household, a variable with an inverse association with performance (L. A. Silva et al., 2014; V. A. A. Silva et al., 2020; Souza et al., 2022), appeared among those selected by RFE. However, no relevant rules were found regarding the association with the variable `SCORE` only rules that confirmed the socioeconomic profile of the participants.

4.7 Practical Implications

We identified patterns linking student performance to socioeconomic variables, including parental education, occupation, household assets (such as computers, washing machines, and motorcycles), gender, age, and high school completion status. Thus, we can identify four areas with practical implications: Policy Targeting and Resource Allocation, School-Level Planning and Support, Curriculum and Pedagogical Design, and Regional Equity Monitoring.

- **Policy targeting and resource allocation** - The association rules provide actionable insights for public education managers in Maranhão. For instance, students whose fathers work in agriculture and lack access to a computer or a washing machine are consistently associated with lower performance. This pattern supports considering targeted digital inclusion policies or resource subsidies for these households, such as computer or Internet access grants, which aligns with the theory that the digital divide is a new border where social inequalities are shaping and being shaped by the latest development of technology (Chen & Wellman, 2007). Furthermore, the policy-maker can, for instance, develop conditional cash transfer policies tied to educational engagement for families in occupational groups most closely associated with underperformance.
- **School-Level Planning and Support** - School administrators could use a similar rule-mining approach with local data to proactively identify students who, despite not yet underperforming, match socioeconomic profiles associated with low performance. This proactive stance enables preemptive academic or emotional support, such as identifying at-risk students early for mentoring or tutoring programs and informing school counseling by flagging students who match high-risk profiles for extra support.
- **Curriculum and Pedagogical Design** - Educators can tailor instructional design by recognizing that many students, especially those without home computers, may face digital literacy gaps. Classroom strategies should address these disparities by incorporating offline alternatives and guided technology use, and by adjusting pedagogical strategies to accommodate students with limited home access to educational technology. Further, educators can emphasize offline learning materials or provide extra lab time for digital skill development.

- **Regional Equity Monitoring** - Beyond Maranhão, the method applied here can be used in other Brazilian states to surface localized inequalities and support comparative analyses of educational equity, enabling regionally tailored policy responses.

5 Conclusions

This study examines the factors that influence Maranhão students' performance in the 2023 ENEM, utilizing educational data mining on INEP's public microdata. By employing the FP-Growth algorithm for association rule mining, the research reveals latent patterns not detected in exploratory analysis, linking socioeconomic variables such as parental occupation, age, gender, high school completion status, and the absence of household technology such as computers and washing machines, to lower performance brackets. Notably, the father's or male guardian's occupation, strongly correlated with family income, emerges as a critical factor, addressing a research gap in the socioeconomic characterization of Maranhão's student population. While the study identifies patterns rather than causal relationships, these associations highlight broader structural inequalities, with low performance being tied to systemic issues such as limited access to technology and agricultural occupations.

Moreover, the paper's focus on Maranhão, an underrepresented state with significant educational challenges, underscores its social relevance. Its novel use of unsupervised learning via FP-Growth, contrasting with the prevalent supervised methods in educational data mining, offers a fresh perspective. Furthermore, the proposed framework, which integrates CRISP-DM, Recursive Feature Elimination (RFE) with Random Forest, and FP-Growth, provides a scalable and replicable approach for analyzing educational data in other Brazilian regions by changing only a filter such as *SG_UF_PROVA == "PI"*, in which now the query is targeting the Piauí state, then running the rest of the pipeline: RFE with Random Forest $\rightarrow$ FP-Growth ($sup \geq 0.10$, $conf \geq 0.50$), thereby enhancing its academic and practical impact.

5.1 Comparison with ENEM Works

While these studies provide valuable insights into socioeconomic predictors of ENEM performance, a direct comparison with our findings in Maranhão's 2023 ENEM data reveals both convergences and regional nuances. For instance, Lima and Brighenti (2023) identified parental education and private school attendance as positive contributors to higher scores in Minas Gerais, aligning closely with our association rules where low parental education, such as "no formal education" antecedent, yields high-confidence (conf > 70%) links to below-average performance across subjects. Similarly, Souza et al., 2022 confirmed the positive impact of technology access in Brazil's Northeast, mirroring our results: lack of computers/internet shows strong associations (sup = 25%, lift = 1.5) with low scores, exacerbated in Maranhão's public schools.

Soares et al. (2023), using CRISP-DM on IDEB/SAEB data for Maranhão, highlighted parental education and technology as key factors, echoing our RFE-selected variables, yet noted data incompleteness. Our FP-Growth approach overcomes this by mining complete ENEM microdata, uncovering gender-specific rules, for instance, female students + low income $\rightarrow$ low performance, conf=80%, underexplored in prior works. Dutra et al. (2023) review underscores income

and race as universal predictors (+ contribution), consistent with our $lift > 1.0$ rules, but our unsupervised pipeline uniquely reveals probabilistic interactions, such as occupation vs gender, tailored to Maranhão's vulnerabilities. These alignments validate the generalizability of socioeconomic influences. At the same time, our gender-disaggregated patterns and scalable framework address gaps in descriptive analytics for underrepresented regions, informing targeted interventions like digital inclusion programs.

5.2 Limitations and Future Direction

The investigation's limitations highlight broader challenges in leveraging data for the greater good. Incomplete graduate-reported variables, such as school administrative dependence, underscore the need for integrated data systems that merge institutional records with survey responses to reduce gaps. Emerging tools, such as natural language processing (NLP), could automate the extraction of missing metadata from text-based responses, while blockchain-based credentialing might enhance data traceability.

An evolution toward supervised machine learning could strengthen predictive validity, particularly if models incorporate spatial analytics to assess regional disparities, such as rural vs. urban divides in resource allocation. For instance, geospatial clustering algorithms could identify underserved areas where interventions, such as targeted funding or teacher training, are most needed.

Finally, while income and parental education remain critical determinants of educational outcomes, future research should adopt intersectional frameworks to examine how access to technology, regional infrastructure disparities, and policy variations interact with these factors. To strengthen methodological rigor and scalability, subsequent studies could focus on:

- Optimizing model performance through metaheuristic techniques, such as Genetic Algorithms, for hyperparameter tuning, enhancing predictive accuracy while reducing computational costs.
- Validating decision rules via unsupervised clustering algorithms, such as k-means or DBSCAN, to identify latent patterns in the data and assess the robustness of inferred relationships.
- Expanding the geographical scope by replicating the analysis using national-level databases would improve the generalizability of the findings across diverse regional contexts and socioeconomic conditions.
- Explainable AI (XAI) methods could clarify how specific variables, such as regional infrastructure, influence outcomes, ensuring transparency for policymakers.

Acknowledgment

This work was supported by the Brazilian Coordination for the Improvement of Higher Education Personnel (CAPES) – Finance Code 001,

Authors' CRediT

A.A.A: Software, Data curation, Investigation, Writing – original draft, Formal analysis; O. A. C. C.: Conceptualization, Methodology, Investigation, Supervision, Validation, Writing – review, editing, and final version.

References

Alalawi, K., Athauda, R., & Chiong, R. (2024). An extended learning analytics framework integrating machine learning and pedagogical approaches for student performance prediction and intervention. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-024-00429-7 [GS Search].

Alvarez-Garcia, M., Arenas-Parra, M., & Ibar-Alonso, R. (2024). Uncovering student profiles. an explainable cluster analysis approach to PISA 2022. *Computers & Education*, *223*, 105166. https://doi.org/10.1016/j.compedu.2024.105166 [GS Search].

Brazil. (1996). Law of guidelines and bases for national education, lei de diretrizes e bases [Accessed on Sept. 10, 2024]. https://www.planalto.gov.br/ccivil_03/LEIS/L9394.htm

Brazil. (2016). Establishes the open data policy of the federal executive branch [Accessed on Sept. 10, 2024]. http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2016/decreto/d8777.htm

Carvalho, A. C. P. L. F., Menezes, A. G., & Bonidia, R. P. (2024). *Data science: Fundamentals and applications* (1st). LTC.

Castro, L. N., & Ferrari, D. G. (2016). *Introduction to data mining: Basic concepts, algorithms, and applications* (1st). Saraiva.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. https://mineracaodedados.files.wordpress.com/2012/12/crisp-dm-1-0.pdf [GS Search].

Chen, W., & Wellman, B. (2007). Minding the cyber-gap: The internet and social inequality. In A. Lareau & D. Conley (Eds.), *The blackwell companion to social inequalities* (pp. 523–545). Blackwell Publishing. https://doi.org/10.1002/9780470996973.ch23 [GS Search].

Dutra, J. F., Firmino Júnior, J. B., & Fernandes, D. Y. S. (2023). Fatores que podem interferir no desempenho de estudantes no ENEM: Uma revisão sistemática da literatura. *Rev. Bras. Informática Educ.*, *31*, 323–351. https://doi.org/https://doi.org/10.5753/rbie.2023.3087 [GS Search].

Faceli, K., Lorena, A. C., Gama, J., Almeida, T. A., & Carvalho, A. C. P. L. F. (2024). *Artificial intelligence: A machine learning approach* (2nd). LTC.

Gabriel, M. (2024). *Artificial intelligence: From zero to the metaverse* (1st). Atlas.

Goldschmidt, R., Passos, E., & Bezerra, E. (2015). *Data mining: Concepts, techniques, algorithms, guidelines, and applications*. Elsevier.

Gomes, T., Gouveia, R., & Batista, M. C. (2017). Dados educacionais abertos: Associações em dados dos inscritos do exame nacional do ensino médio. *Proc. Workshop de Informática na Escola (WIE)*, *23*, 895–904. https://doi.org/10.5753/cbie.wie.2017.895 [GS Search].

Han, J., Pei, J., & Yin, Y. (2000). Mining frequent patterns without candidate generation. *SIGMOD Rec.*, *29*(2), 1–12. https://doi.org/10.1145/335191.33537 [GS Search].

Inep. (2024a). ENEM 2023 microdata [Accessed on Apr. 30, 2024]. https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem

Inep. (2024b). MEC and Inep release results of the 2023 school census [Accessed on Sept. 6, 2024]. https://www.gov.br/inep/pt-br/assuntos/noticias/censo-escolar/mec-e-inep-divulgam-resultados-do-censo-escolar-2023

Kaur, M., Singh, M., & Saini, M. (2024). Analyzing the relation among different factors leading to Ph.D. dropout using numerical association rule mining. *Education and Information Technologies*, *29*, 375–399. https://doi.org/10.1007/s10639-023-12260-z [GS Search].

Lima, C. C. V., & Brighenti, C. R. G. (2023). Performance of students from Minas Gerais in the national high school exam considering socioeconomic variables. *Educação e Pesquisa*, *49*, e253303. https://doi.org/10.1590/S1678-4634202349253303 [GS Search].

Ma, Y., Cain, K., & Ushakova, A. (2024). Application of cluster analysis to identify different reader groups through their engagement with a digital reading supplement. *Computers & Education*, *214*, 105025. https://doi.org/10.1016/j.compedu.2024.105025 [GS Search].

Marconi, M. d. A., & Lakatos, E. M. (2022). *Scientific methodology* (8th). Atlas.

Martin, P. P., Kranz, D., & Graulich, N. (2024). Revealing rubric relations: Investigating the interdependence of a research-informed and a machine learning-based rubric in assessing student reasoning in chemistry. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-024-00440-y [GS Search].

Merkys, G., Vaitkevičius, S., Bubelienė, D., & Brazdeikis, V. (2025). The influence of social conditionality on the results in computer science test of graduates. *Informatics in Education*, *24*(1), 145–173. https://doi.org/10.15388/infedu.2025.04 [GS Search].

Ministério da Educação. (2024). Ideb: Maranhão advances in early years of elementary education [Accessed on Sept. 6, 2024]. https://www.gov.br/mec/pt-br/assuntos/noticias/2024/agosto/ideb-maranhao-avanca-nos-anos-iniciais-do-ensino-fundamental

Munim, Z. H., Kjeldsberg, F., Bustgaard, M., Bhagat, S., Haavardtun, P., Kim, T.-E., Lindroos, E., Thorvaldsen, H., Nyairo, F., & Lampiola, J. (2025). Predictive performance assessment in simulation training using machine learning. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-025-00464-y [GS Search].

Netto, A., & Maciel, F. (2021). *Python for data science and machine learning: Simplified*. Alta Books.

Ouassif, K., & Ziani, B. (2025). Predicting university major selection and academic performance through the combination of apriori algorithm and deep neural network. *Education and Information Technologies*, *30*, 333–346. https://doi.org/10.1007/s10639-024-13022-1 [GS Search].

Sayak, P. (2020). Python feature selection tutorial: A beginner's guide [Accessed on Dec. 12, 2024]. https://www.datacamp.com/tutorial/feature-selection-python

Silva, L. A., Morino, A. H., & Sato, T. M. C. (2014). Prática de mineração de dados no exame nacional do ensino médio. *Proc. Workshops of the Brazilian Congress on Informatics in Education (WCBIE)*, *3*, 651. https://doi.org/10.5753/cbie.wcbie.2014.651 [GS Search].

Silva, L. A., Peres, S. M., & Boscarioli, C. (2016). *Introduction to data mining: With applications in R* (1st). Elsevier.

Silva, V. A. A., Moreno, L. L. O., Gonçalves, L. B., Soares, S. S. R. F., & Souza Júnior, R. R. (2020). Identificação de desigualdades sociais a partir do desempenho dos alunos do ensino médio no ENEM 2019 utilizando mineração de dados. *Proc. Brazilian Symposium on*

Informatics in Education (SBIE), *31*, 72–81. https://doi.org/10.5753/cbie.sbie.2020.72 [GS Search].

Soares, R. C., Weber Neto, N., Coutinho, L. R., Santos, D. V., Silva, F. J. S., & Teles, A. S. (2023). Minerando dados para entender os fatores de influência da qualidade educacional do Maranhão. *Revista Brasileira de Informática na Educação*, *31*, 378–406. https://doi.org/10.5753/rbie.2023.2831 [GS Search].

Souza, A. E., Santos, L. M. S., Larrucaim, I. M., & Besarria, C. N. (2022). Determinantes do desempenho no enem na região nordeste: Uma análise de dados em painel do período de 2015 a 2019. *Rev. Bras. Estud. Reg. Urbanos*, *15*(4), 690–711. https://doi.org/https://doi.org/10.54766/rberu.v15i4.915 [GS Search].

Srikant, R., & Agrawal, R. (1997). Mining generalized association rules [Data Mining]. *Future Generation Computer Systems*, *13*(2), 161–180. https://doi.org/https://doi.org/10.1016/S0167-739X(97)00019-8 [GS Search].

Travitzki, R. (2021). Possíveis contribuições do Enem para a democratização do acesso à educação superior no Brasil. *Em Aberto*, *34*(112). https://doi.org/10.24109/2176-6673.emaberto.34i112.4993 [GS Search].

Vaarma, M., & Li, H. (2024). Predicting student dropouts with machine learning: An empirical study in finnish higher education. *Technology in Society*, *76*, 102474. https://doi.org/10.1016/j.techsoc.2024.102474 [GS Search].