

UNIVERSIDADE ESTADUAL DO MARANHÃO
CENTRO DE CIÊNCIAS TECNOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DA COMPUTAÇÃO E
SISTEMAS

ANDERSON AMORIM ALVES

**UM OLHAR NÃO SUPERVISIONADO SOBRE O DESEMPENHO DOS
PARTICIPANTES MARANHENSES NO ENEM 2023**

São Luís - MA
2025

ANDERSON AMORIM ALVES

**UM OLHAR NÃO SUPERVISIONADO SOBRE O DESEMPENHO DOS
PARTICIPANTES MARANHENSES NO ENEM 2023**

Dissertação de mestrado apresentada como requisito para obtenção do grau de mestre no Programa de Pós-Graduação em Engenharia da Computação e Sistemas da Universidade Estadual do Maranhão.

Orientador: Prof. Dr. Omar Andres C. Cortes

São Luís - MA
2025

Alves, Anderson Amorim

Um olhar não supervisionado sobre o desempenho dos participantes maranhenses no Enem 2023. / Anderson Amorim Alves. – São Luís, MA, 2025.

101 f

Dissertação (Mestrado em Programa de Pós-Graduação em Engenharia da Computação e Sistemas) - Universidade Estadual do Maranhão, 2025.

Orientador: Prof. Dr. Omar Andres C. Cortes.

1. Ciência de Dados. 2.Mineração de Dados Educacionais. 3.Regras de Associação. 4.Enem. 5.Maranhão. I.Titulo.

CDU: 004.62:37

Elaborado por Cássia Diniz - CRB 13/910

ANDERSON AMORIM ALVES

**UM OLHAR NÃO SUPERVISIONADO SOBRE O DESEMPENHO DOS
PARTICIPANTES MARANHENSES NO ENEM 2023**

Dissertação de mestrado apresentada como requisito para obtenção do grau de mestre no Programa de Pós-Graduação em Engenharia da Computação e Sistemas da Universidade Estadual do Maranhão.

Aprovado em: 26 / 02 / 2025 .

BANCA EXAMINADORA

Prof. Dr. Omar Andres Carmona Cortes (Orientador)
Doutor em Ciência da Computação e Matemática Computacional
Instituto Federal do Maranhão (IFMA)

Prof. Dr. Antonio Fernando Lavareda Jacob Junior
Doutor em Engenharia Elétrica
Universidade Estadual do Maranhão (UEMA)

Profa. Dra. Ellen Francine Barbosa
Doutora em Ciência da Computação e Matemática Computacional
Instituto de Ciências Matemáticas e da Computação (ICMC/USP)

*A Deus, à minha família e amigos,
pelo incentivo e compreensão.*

AGRADECIMENTOS

A Deus, por conceder o dom da vida e oportunizar nosso crescimento moral, ético e espiritual.

À minha mãe, Maria das Graças Costa Amorim, por todo amor dedicado, pelo apoio incondicional, pela paciência e por acreditar no meu potencial acadêmico, em que incluo os meus irmãos, Winderson Amorim e Antônio Victor Pinheiro Costa Amorim.

À minha filha, Agatha Alves Amorim, por ser a motivação nos momentos mais delicados, cujo crescimento e desenvolvimento me encantam e inspiram.

Aos nossos professores e colegas do Programa de Pós-Graduação em Engenharia da Computação e Sistemas da UEMA, pela convivência enriquecedora, laços de amizade e companheirismo construídos durante o curso. De forma muito especial agradeço à Franciane Reis da Silva Eugenio, Georlene Lima Moraes e Marcos Henrique Vieira dos Santos.

Aos amigos que a vida me deu, pessoas formidáveis, por me ouvirem, auxiliarem, aconselharem e apoiarem em incontáveis momentos, preciso aqui citar Elio Walbert F. Ferreira, Manoel Cleber Sampaio Silva, Jackellyne Geórgia Dutra e Silva Leite, Denise Maria de M. Ribeiro, Kleber Gonçalves Costa Leite, Garleanne Gonçalves Mendes, Wadilson de Jesus Tavares Costa e Maria do Socorro Gonçalves Costa Leite.

Ao amigo Jorge Antonio Carvalho, pelo companheirismo e lutas compartilhadas, e à Nayara Alves Lindoso, pelo acolhimento, pela escuta atenta e pelo cuidado dedicado.

Ao amigo Jorge Antônio Soares Leão, pelo amor e carinho durante essas décadas de amizade que me convidam à evolução moral e à prática do bem.

À amiga Profa. Dra. Maria do Socorro Silva Lages pela amizade, apoio e por toda a experiência transmitida.

À Marcelle Maciel da Silva por acreditar nesse sonho, pelo carinho, paciência e apoio durante essa jornada.

Ao meu orientador, Prof. Dr. Omar Andres Carmona Cortes, pelos ensinamentos enquanto professor do curso, por acolher essa orientação e abraçar o projeto.

À banca examinadora pela atenção e disponibilidade em avaliar o presente trabalho, realizando importantes contribuições.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) como instituição basilar para o fomento e divulgação da produção científica em nosso país.

À UEMA por oportunizar a formação de pesquisadores e contribuir com todo o seu corpo de servidores e terceirizados para a concretização de um sonho.

*And until the ignoble and unhappy
regime
That hold our brothers in Angola
In Mozambique
South Africa
Sub-human bondage
Have been toppled
Utterly destroyed
Well, everywhere is war
Me say war
Robert Nesta Marley (1945-1981)*

*Quando a educação não é
libertadora, o sonho do oprimido é
ser o opressor.
Paulo Freire (1921-1997)*

*Lutar contra a pobreza não é um
assunto de caridade, mas de
justiça.
Nelson Mandela (1918-2013)*

*A Matemática é o alfabeto que
Deus usou para escrever o
Universo.
Galileu Galilei (1564-1642)*

RESUMO

A Ciência de Dados é uma área interdisciplinar, que pode colaborar para o enfrentamento dos desafios da humanidade para o século XXI, dentre eles a busca por uma Educação inclusiva, equitativa e de qualidade, sobretudo na realidade brasileira, em que o analfabetismo ainda atinge milhões de pessoas, de forma mais severa, os negros da Região Nordeste. O Exame Nacional do Ensino Médio (Enem) tem desempenhado importante papel para a avaliação da educação básica brasileira, bem como, servido como a principal porta de acesso à educação superior. Por outro lado, os resultados apresentados pelos estudantes brasileiros, especialmente no Estado do Maranhão, servem de alerta e apontam para a necessidade de melhorias na educação básica. O estudo justifica-se pelos resultados do Maranhão no Ideb e no Enem, colocando-o, nesta última, entre os piores resultados médios do país na edição de 2023. A problemática do estudo reside na compreensão das variáveis que assumem papel preponderante para o desempenho dos estudantes maranhenses na mais importante avaliação da educação básica, sob a perspectiva da mineração de dados educacionais. Por essa razão, o objetivo geral consiste em investigar os fatores influentes no desempenho dos estudantes maranhenses no Enem 2023, partindo das ferramentas estatísticas de análise exploratória e extraindo conhecimentos por meio da aprendizagem de máquina não supervisionada, de forma específica, a extração de regras de associação com o algoritmo *FP-Growth*. A abordagem da pesquisa é de natureza quantitativa e a metodologia empregou o processo CRISP-DM de descoberta de conhecimento em bases de dados. Os experimentos computacionais contemplaram a seleção de dez atributos por meio da aplicação da RFE combinada ao algoritmo *Random Forest*, bem como, a posterior aplicação do algoritmo *FP-Growth* em dez ensaios/configurações, sobre o subconjunto de microdados dos participantes maranhenses no Enem 2023. O estudo identificou outra variável relevante para o desempenho dos participantes, que favorece uma melhor compreensão e caracterização socioeconômica do público maranhense, a citar a natureza da ocupação do pai ou homem responsável. Além disso, conseguiu revelar padrões subjacentes aos dados, não vislumbrados na análise exploratória, como a descoberta de grupos afetados de forma cumulativa por diferentes atributos. Por exemplo, o baixo desempenho condicionado pela escolaridade e natureza da ocupação do pai, ou ainda, a caracterização do perfil socioeconômico dos participantes maranhenses a partir das variáveis selecionadas, com alto índice de confiança. Apesar de limitações inerentes à qualidade da base de dados, o estudo mostrou-se relevante socialmente e com potencial de replicabilidade para outras regiões.

Palavras-chave: Ciência de Dados. Mineração de Dados Educacionais. Regras de Associação. Enem. Maranhão.

ABSTRACT

Data Science is an interdisciplinary field that can contribute to addressing humanity's challenges in the 21st century, including the search for inclusive, equitable and quality education, particularly in the Brazilian context, where illiteracy still affects millions of people, more severely impacting black individuals in the Northeast Region. The National High School Exam (Enem) has played a significant role in evaluating Brazilian basic education and has become the primary gateway to higher education. However, the results of Brazilian students, especially in the state of Maranhão, serve as a warning and highlight the need for improvements in basic education. This study is justified by Maranhão's performance in the Ideb and the Enem, placing it among the lowest average results in the country in the 2023 edition. The research problem lies in understanding the variables that play a predominant role in the performance of Maranhão's students in the most important assessment of basic education, from the perspective of educational data mining. Therefore, the general objective is to investigate the factors influencing the performance of Maranhão's students in the 2023 Enem, using exploratory statistical analysis tools and extracting knowledge through unsupervised machine learning, specifically association rule mining with the FP-Growth algorithm. The research approach is quantitative in nature and the methodology employed the CRISP-DM process for knowledge discovery in databases. Computational experiments included the selection of ten attributes through the application of RFE combined with the Random Forest algorithm, as well as the subsequent application of the FP-Growth algorithm in ten trials/configurations on the subset of microdata from Maranhão's participants in the 2023 Enem. The study identified another relevant variable for participants' performance, which enhances the understanding and socioeconomic characterization of Maranhão's population, namely the nature of the occupation of the father or male guardian. Additionally, it revealed underlying patterns in the data not observed in the exploratory analysis, such as the discovery of groups cumulatively affected by different attributes. For example, low performance conditioned by the father's education level and occupation, or the socioeconomic profile characterization of Maranhão's participants based on the selected variables, with a high confidence index. Despite inherent limitations in the quality of the database, the study proved to be socially relevant and has the potential for replication in other regions.

Keywords: Data Science. Educational Data Mining. Association Rules. Enem. Maranhão.

LISTA DE FIGURAS

Figura 1 – Ideb do Maranhão por Etapa Educacional	18
Figura 2 – Média das Notas por Área da Região Nordeste - Enem 2023.....	18
Figura 3 – Interdisciplinaridade da Ciência de Dados.....	29
Figura 4 – Áreas que compõem a Ciência de Dados Educacionais.....	30
Figura 5 – Áreas relacionadas à Mineração de Dados Educacionais	31
Figura 6 – Etapas do processo de Mineração de Dados Educacionais	32
Figura 7 – Inteligência Artificial e suas subáreas.....	33
Figura 8 – Exemplo de diagrama de barras paralelas.....	38
Figura 9 – Algoritmo <i>FP-tree</i>	38
Figura 10 – Primeira varredura da base transacional	39
Figura 11 – Construção da FP-Tree.....	40
Figura 12 – Decomposição da árvore em vários subproblemas	41
Figura 13 – Árvore FP condicional para ‘e’	41
Figura 14 – Etapas do CRISP-DM	43
Figura 15 – Etapas do CRISP-EDM.....	44
Figura 16 – Etapas do CRISP-DM adequadas ao presente estudo	45
Figura 17 – Subconjunto Nacional de Microdados do Enem 2023 (parcial)	50
Figura 18 – Execução da RFE para seleção dos atributos.....	53
Figura 19 – Atributos selecionados com a RFE	53
Figura 20 – Árvore de Decisão com os atributos selecionados com a RFE.....	54
Figura 21 – Faixas Etárias	55
Figura 22 – Sexo.....	55
Figura 23 – Grupos Étnicos	56
Figura 24 – Situação de Conclusão do Ensino Médio.....	56
Figura 25 – Participantes por UF.....	57
Figura 26 – Escolaridade da Mãe	58
Figura 27 – Renda familiar mensal.....	59
Figura 28 – Computador na residência.....	59
Figura 29 – Acesso residencial à internet.....	59
Figura 30 – Resumo estatístico dos atributos numéricos	60
Figura 31 – Histogramas das notas.....	61
Figura 32 – <i>Boxplots</i> das notas por área	61
Figura 33 – <i>Boxplots</i> das notas por faixa etária.....	62
Figura 34 – Histograma e <i>Boxplots</i> das notas por Sexo	62
Figura 35 – <i>Boxplots</i> das notas por grupo étnico.....	62
Figura 36 – <i>Boxplots</i> das notas por Situação de Conclusão do Ensino Médio.....	63
Figura 37 – Desempenho médio por UF	63
Figura 38 – <i>Boxplots</i> das notas por UF.....	64
Figura 39 – <i>Boxplots</i> das notas por Nível de Escolaridade da Mãe	64
Figura 40 – <i>Boxplots</i> das notas por faixa de Renda Familiar Mensal	65
Figura 41 – <i>Boxplots</i> das notas de acordo com a presença de Computador na Residência.....	65

Figura 42 – <i>Boxplots</i> das notas por nível de acesso residencial à internet	65
Figura 43 – Notas por sexo.....	66
Figura 44 – Notas por grupo étnico.	66
Figura 45 – Notas por escolaridade das mães.....	66
Figura 46 – Notas por computador na residência.	66
Figura 47 – Escolaridade da Mãe (MA)	67
Figura 48 – Escolaridade do Pai (MA)	67
Figura 49 – Notas por nível de escolaridade do Pai (MA)	68
Figura 50 – Renda familiar mensal (MA)	68
Figura 51 – Notas por Grupo de ocupação do Pai	69
Figura 52 – Notas por quantidade de moradores na residência	70
Figura 53 – Subconjunto Regional de Microdados do Enem 2023	71

LISTA DE TABELAS

Tabela 1 – Variáveis influentes para o desempenho escolar segundo a literatura	25
Tabela 2 – Configurações dos experimentos das regras de associação com o Apriori	26
Tabela 3 – Extração das regras de associação	41
Tabela 4 – Avaliação das regras de associação	42
Tabela 5 – Configurações dos experimentos	46
Tabela 6 – Fontes dos dados.....	48
Tabela 7 – Variáveis com dados faltantes	49
Tabela 8 – Variáveis redundantes ou desnecessárias para o estudo	49
Tabela 9 – Variáveis selecionadas com a RFE.....	53
Tabela 10 – Categorias da variável “Q002”	57
Tabela 11 – Categorias da variável “Q006”	58
Tabela 12 – Categorias da variável “Q003”	69
Tabela 13 – Detalhamento da categorização da variável “NOTA”	71
Tabela 14 – Experimentos realizados	71
Tabela 15 – Regras de Associação do Experimento 1.....	72
Tabela 16 – Regras de Associação do Experimento 2.....	73
Tabela 17 – Regras de Associação do Experimento 4.....	75
Tabela 18 – Regras de Associação do Experimento 6.....	77
Tabela 19 – Regras de Associação do Experimento 7.....	77
Tabela 20 – Regras de Associação do Experimento 9.....	80
Tabela 21 – Regras de Associação do Experimento 10.....	82

LISTA DE ABREVIATURAS E SIGLAS

AI	Anos Iniciais [do Ensino Fundamental]
AF	Anos Finais [do Ensino Fundamental]
Apriori	Algoritmo de Construção de Regras de Associação
CRISP-DM	<i>Cross Industry Standard Process for Data Mining</i>
CRISP-EDM	<i>Cross Industry Standard Process for Educational Data Mining</i>
ENADE	Exame Nacional de Desempenho dos Estudantes
ENCCEJA	Exame Nacional para Certificação de Competências de Jovens e Adultos
ENEM	Exame Nacional do Ensino Médio
EM	Ensino Médio
FIES	Fundo de Financiamento Estudantil
FP-Growth	<i>Frequent Pattern Growth</i>
IA	Inteligência Artificial
IDEB	Índice de Desenvolvimento da Educação Básica
IDH	Índice de Desenvolvimento Humano
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
LAI	Lei de Acesso à Informação
LDB	Lei de Diretrizes e Bases da Educação Nacional
LGPD	Lei Geral de Proteção de Dados
LIBRAS	Língua Brasileira de Sinais
LMS	Sistemas para Gerenciamento da Aprendizagem
MDE	Mineração de Dados Educacionais
MEC	Ministério da Educação
ODS	Objetivos para o Desenvolvimento Sustentável
ONU	Organização das Nações Unidas
PIB	Produto Interno Bruto
PNE	Plano Nacional de Educação
PROUNI	Programa Universidade para todos
RFE	<i>Recursive Feature Elimination</i>
RNA	Rede Neural Artificial
SAEB	Sistema de Avaliação da Educação Básica
SISU	Sistema de Seleção Unificada
TRI	Teoria de Resposta ao Item
UF	Unidade Federativa

SUMÁRIO

1	INTRODUÇÃO.....	14
1.1	Motivação e Justificativa.....	17
1.2	Problema e Hipótese	19
1.3	Objetivos.....	20
1.3.1	Objetivo Geral	20
1.3.2	Objetivos Específicos	20
1.4	Enem	20
1.5	Trabalhos Correlatos.....	22
1.6	Estrutura do Trabalho	26
2	FUNDAMENTAÇÃO TEÓRICA	28
2.1	Ciência de Dados.....	28
2.2	Ciência de Dados Educacionais	29
2.2.1	Mineração de Dados Educacionais	31
2.3	Aprendizado de Máquina.....	32
2.3.1	Regras de Associação	35
2.3.1.1	<i>Algoritmo FP-Growth.....</i>	<i>38</i>
3	EXPERIMENTOS	42
3.1	Procedimentos Metodológicos	43
3.1.1	Recursos Computacionais.....	46
3.2	Compreensão do domínio.....	47
3.3	Compreensão dos Dados	47
3.4	Preparação dos Dados	48
3.4.1	Seleção dos atributos	51
4	ANÁLISE DOS RESULTADOS	55
4.1	Análise Exploratória.....	55
4.1.1	Visão geral dos participantes	55
4.1.2	Visão geral do desempenho	60
4.1.3	Visão Geral do Maranhão	66
4.2	Extração das Regras de Associação (Modelagem).....	70
4.3	Avaliação e Discussão	72
5	CONCLUSÕES, CONTRIBUIÇÕES E TRABALHOS FUTUROS	83
	REFERÊNCIAS	86
	ANEXO – DICIONÁRIO DE MICRODADOS DO ENEM 2023.....	91

1 INTRODUÇÃO

A humanidade possui grandes desafios no século XXI que envolvem a superação da pobreza, erradicação da fome, superação das desigualdades sociais, enfrentamento das mudanças climáticas, e, por mais alarmante que seja, a preservação da vida. Essas questões foram explicitamente declaradas pela Organização das Nações Unidas (ONU) através dos 17 Objetivos para o Desenvolvimento Sustentável (ODS), nos quais a Educação aparece no quarto compromisso, visando a garantia de uma educação inclusiva, equitativa e de qualidade, bem como, a eliminação das discrepâncias de gênero, sociais e étnicas (ONU Brasil, 2024).

Neste âmbito, embora os avanços tecnológicos tenham transformado diversos setores, a educação ainda enfrenta desafios que não podem mais ser ignorados. Em um cenário onde os dados são gerados em abundância, a capacidade de traduzi-los em melhorias reais no sistema educacional torna-se vital, sobretudo na realidade brasileira, em que o analfabetismo ainda atinge 11,4 milhões de pessoas, de forma mais severa, os pretos e pardos da Região Nordeste, segundo o último censo (Nery, 2024). Afinal, o que os números podem revelar sobre o desempenho dos estudantes e a eficácia das políticas públicas? O uso de ferramentas de análise e extração de conhecimentos em bases de dados, enquanto elementos da Ciência dos Dados, podem oferecer uma nova lente para compreender os fatores que moldam o aprendizado e a qualidade da educação do Brasil.

A Ciência de Dados é uma área que ganha relevância com o fenômeno do *Big Data*, que se refere ao enorme volume de dados produzidos diariamente em diversas áreas, gerando desafios significativos para as organizações (Rautenberg; Carmo, 2019). Esta produção massiva de dados também atinge o ambiente educacional a partir da adoção das tecnologias da informação e comunicação, passando a exigir das personagens do ambiente escolar, competências para extrair informações úteis e contornar problemas de aprendizagem (Ramos *et al.*, 2020).

A mineração de dados é uma etapa do processo de Ciência de Dados que vem sendo aplicada com êxito em diversos campos. Embora a metodologia empregada para descoberta do conhecimento seja semelhante, a mineração de dados irá adequar-se ao domínio de aplicação e, conseqüentemente, aos seus objetivos. Neste sentido, a Mineração de Dados Educacionais (MDE) surge para lidar com um tipo específico de dados, aqueles produzidos em ambientes educacionais (Romero; Ventura, 2013).

Em outras palavras, a MDE busca produzir conhecimento a partir dos dados brutos produzidos por alunos e professores, por meio de técnicas como classificação, agrupamento, mineração de textos (Filatro, 2021) adotando algoritmos de Aprendizado de

Máquina para realizar tarefas preditivas ou descritivas (Faceli *et al.*, 2024; Carvalho; Menezes; Bonidia, 2024).

Trata-se de uma recente área, cujo desenvolvimento deu-se a partir das conferências sobre *Educational Data Mining* (MDE) nos anos 2000 (Filatro, 2021), potencializada, inicialmente, por dois fatores essenciais: a adoção de *softwares* educacionais e de sistemas computacionais para gerenciamento da aprendizagem (LMS), os quais propiciaram a produção de dados em larga escala de maneira estruturada (Baker; Isotani; Carvalho, 2011). O desenvolvimento da MDE está diretamente relacionado à adoção das tecnologias da informação e comunicação no ambiente educacional, que promoveu não apenas a produção, mas também a disponibilização desses dados, propiciando o acesso e reduzindo custos de obtenção.

Nessa linha, a disponibilização de dados educacionais no Brasil acompanha a recente e progressiva legislação para transparência governamental, de modo a proporcionar o controle social e a participação cidadã, por meio de normas como a Lei de Acesso à Informação – LAI (nº 12.527/2011) e Lei Geral de Proteção de Dados – LGPD (nº 13.709/2018), que determinam à Administração Pública a divulgação de informações em formato estruturado, legível por máquina, possibilitando o acesso automatizado, análise e interpretação. De forma mais contundente, a instituição da Política de Dados Abertos pelo Decreto nº 8.777/2016 promoveu a divulgação de dados contidos em bases governamentais, bem como a autorização para livre utilização gratuita, tendo como detentora dos direitos autorais a União. A própria Lei de Diretrizes e Bases da Educação Nacional (LDB), em recente alteração pela Lei nº 15.001/2024, pautou em seu artigo 3º a “garantia do direito de acesso a informações públicas sobre a gestão da educação” como um princípio da educação nacional (Brasil, 1996).

A obtenção desses dados educacionais no âmbito federal perpassa pela operacionalização das avaliações, exames, pesquisas estatísticas e indicadores educacionais a cargo do Instituto Nacional de Estudos e Pesquisas Educacionais (Inep), autarquia federal vinculada ao Ministério da Educação (MEC). O Inep é responsável pela manutenção e divulgação dos dados educacionais abertos oriundos de diversas bases, como Censo Escolar, Enade ou Encceja, incluindo a divulgação anual dos microdados do Exame Nacional do Ensino Médio (Enem).

Diversos estudos têm sido embasados por estes dados educacionais abertos com o objetivo de avaliar os indicadores da qualidade educacional, compreender os fatores que impactam na aprendizagem, apoiar o processo decisório e fortalecer políticas públicas.

Em revisão sistemática de literatura Lima *et al.* (2019) constatarem 54 publicações na base do Google Acadêmico entre 2005 e 2016, que adotavam os dados do Enem ou Enade, evidenciando a Estatística Descritiva e Mineração de Dados enquanto tipos de análise adotadas, com predominância da primeira.

Já Fonseca e Namen (2016) adotaram a base do Sistema de Avaliação da Educação Básica (Saeb) para Mineração dos Dados e investigação dos fatores que relacionam o perfil de professores que lecionam Matemática com a proficiência dos alunos.

Soares *et al.* (2023), por sua vez, utilizaram a integração de duas bases de dados educacionais abertas mantidas pelo Inep: Índice de Desenvolvimento da Educação Básica - Ideb e Sistema de Avaliação da Educação Básica – Saeb, para identificar fatores que influenciam a qualidade educacional da rede pública estadual de ensino do Maranhão.

A revisão de literatura em Soares *et al.* (2021) acerca da Mineração de Dados Educacionais adotando as bases do Inep aponta para a crescente adoção dessa estratégia para abordar os problemas educacionais. O estudo apontou a tendência predominante do uso de modelos supervisionados e da metodologia para descoberta de conhecimento em bases de dados chamada CRISP-DM (*Cross-Industry Standard Process for Data Mining*). Além disso, os dados do Saeb foram os mais adotados nas pesquisas constatadas, que abordaram principalmente problemas pertinentes ao desempenho escolar.

Neste sentido, o Inep aparece no contexto brasileiro como instituição fomentadora de estudos sobre os indicadores da educação nacional, divulgando pesquisas estatísticas e dados abertos compatíveis com a análise automatizada, ao concretizar a transparência governamental ativa, em observância à LAI e LGPD.

A presente pesquisa levanta alguns estudos na área de Ciência de Dados que buscam compreender os resultados do Exame Nacional do Ensino Médio (Enem), ressaltando que esses dados, produzidos em larga escala, podem fornecer informações essenciais para orientar o desenvolvimento de estratégias e políticas voltadas para a melhoria do sistema educacional brasileiro (Gomes; Gouveia; Batista, 2017; Silva *et al.*, 2020; Elias de Sousa *et al.*, 2021; Soares *et al.*, 2023).

Neste sentido, procura-se não apenas validar os estudos relacionados, como também adotar outra abordagem para minerar dados educacionais, norteados por outras pesquisas e contribuindo para o avanço do conhecimento na área.

1.1 Motivação e Justificativa

O Maranhão, 12º estado mais populoso do Brasil e o 4º dentre os estados da região Nordeste, com 6,8 milhões de habitantes, figura dentre os estados com os mais baixos indicadores sociais e econômicos. Trata-se de um Estado rico em belezas naturais, manifestações culturais, turismo, mas sobretudo, marcado por discrepâncias sociais.

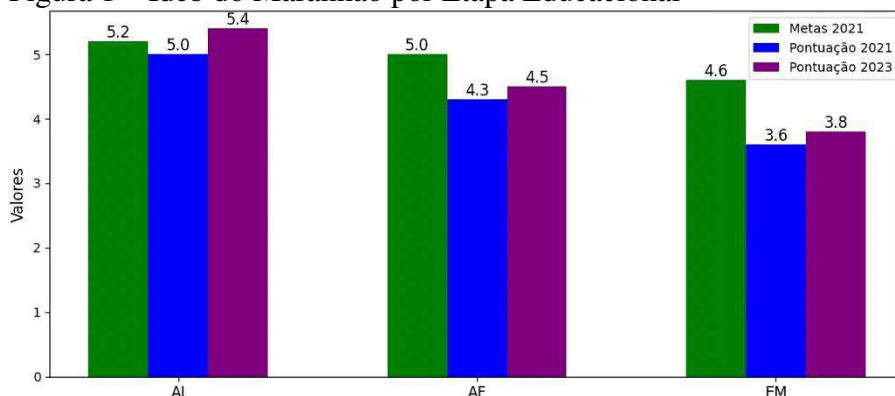
Em 2024, a Região Nordeste foi a região que mais cresceu em nível de atividade econômica no país, 3,2% em relação ao mesmo período do exercício anterior, impulsionada pelos setores de comércio e serviços (BNB, 2024). Em 2023, o Produto Interno Bruto (PIB) da região havia crescido 4,4% a mais que a média nacional. O Maranhão apresentou o maior crescimento relativo do PIB do Nordeste, alavancado pela agropecuária e serviços, segundo relatório de assessoramento do Banco do Brasil de março/2024 (Governo do Maranhão, 2024). Apesar dos avanços econômicos, o Índice de Desenvolvimento Humano – IDH e o rendimento nominal mensal domiciliar *per capita* em torno de R\$ 945,00 ainda são os mais baixos do país (IBGE, 2021; 2023).

No tocante à educação básica, o Brasil teve 7,7 milhões de matrículas no ensino médio em 2023, atendidas, em sua maioria, pela rede estadual de ensino (83,6%), seguida pela rede privada (12,8%) e federal (3,1%) (Inep, 2024).

No contexto maranhense, a rede pública estadual responde por 72,5% das unidades de ensino (784 escolas) de nível médio. Nesta etapa, 88,7% (251.434) das matrículas são atendidas pela rede estadual, enquanto a rede privada colabora com 5,8% (16.402) e a rede federal atende 4,2% (11.926), segundo dados do Censo Escolar 2023 (Inep, 2024).

Os recentes resultados divulgados pelo MEC (2024) apontaram que o Maranhão não atingiu as metas estabelecidas no Ideb para 2021 nas três etapas: anos iniciais (AI) do ensino fundamental, ao atingir 5,0 pontos, enquanto os anos finais (AF) tiveram 4,3 pontos (0,7 ponto abaixo da meta estabelecida). Já no ensino médio (EM), o resultado estadual foi de 3,6 pontos, isto é, 1 (um) ponto abaixo da meta, como ilustrado na Figura 1. Para 2023, os resultados apresentaram avanços, embora ainda apresentem condições desafiadoras. Cabe dizer, que o MEC ainda não definiu as metas para 2023.

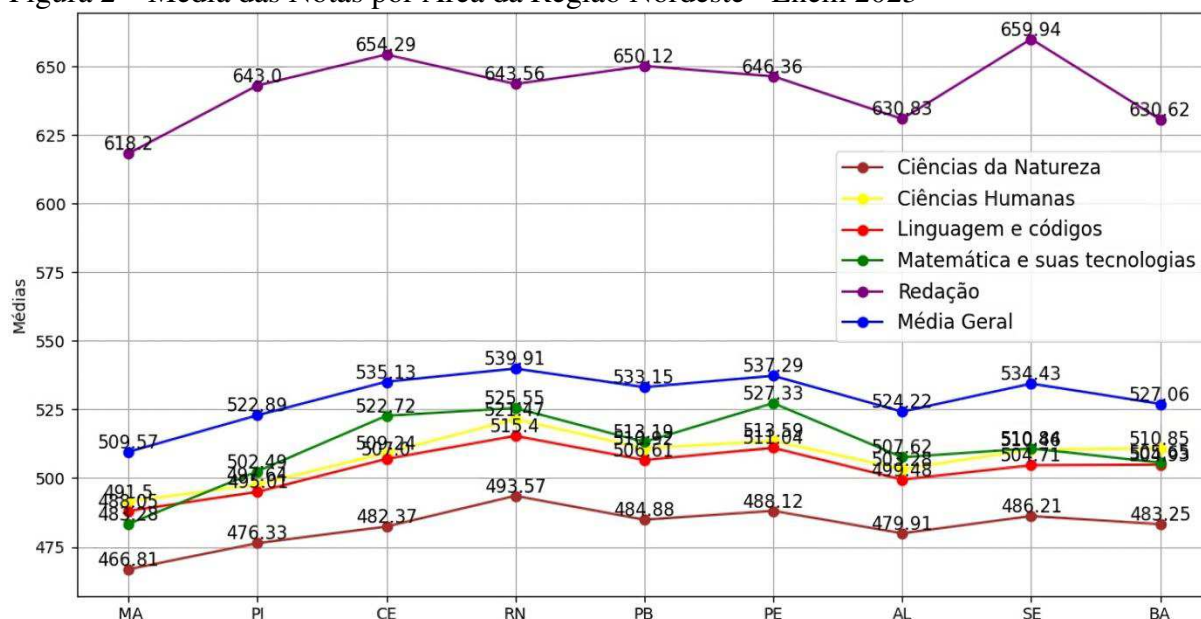
Figura 1 – Ideb do Maranhão por Etapa Educacional



Fonte: elaboração própria.

Em relação ao Enem, os resultados maranhenses figuraram como os piores resultados médios da região Nordeste na edição (2023), segundo sinopse estatística divulgada pelo Inep (2024), com as seguintes marcas: Ciências da Natureza - 466,81 pontos, Ciências Humanas - 491,5 pontos, Linguagem e códigos - 488,05 pontos, Matemática e suas tecnologias - 483,28 pontos e Redação - 618,2 pontos, como mostrado no gráfico comparativo da Figura 2.

Figura 2 – Média das Notas por Área da Região Nordeste - Enem 2023



Fonte: adaptação própria – a partir da Sinopse Estatística do Enem 2023 (Inep, 2024).

Por essa razão, o Estado do Maranhão emerge como objeto de estudo nesta análise, com um cenário desafiador retratado pelos indicadores da educação básica, pelo desempenho dos estudantes na principal avaliação da educação nacional e uma clientela atendida, em sua maioria, pela rede pública estadual de ensino (Inep, 2024).

Além disso, a escassez de estudos educacionais que contemplem o Estado do Maranhão e adotem a base do Enem pode ser relevante para investigar outras variáveis influentes para o desempenho (Dutra *et al.*, 2023).

A seleção desta base - microdados do Enem - considera a etapa educacional em que o Maranhão obteve o menor desempenho segundo resultados do Ideb em 2023. Acrescenta-se o fato de que este exame exerce importante papel na avaliação do ensino no Brasil, podendo ser utilizado não apenas para registrar o desempenho dos estudantes, como também, servir de instrumento de aprimoramento das políticas públicas (Travitzki, 2021).

Neste sentido, vislumbra-se a abordagem dessa problemática educacional no contexto maranhense através da mineração de dados pela eficiência no ganho de conhecimento oriundo dos dados, pela visão mais aprofundada do desempenho e a identificação de áreas que necessitem de intervenção apoiando o processo decisório, considerando que um ambiente de governança baseado em evidências pode contribuir para a melhoria da tomada de decisões e das entregas educacionais (Messias; Lúcio, 2023).

A adoção da Aprendizagem de Máquina não supervisionada por meio da extração de regras de associação, por sua vez, pode revelar nichos não vislumbrados pela análise exploratória dos dados, subsidiando o conhecimento de aspectos subjacentes aos dados em termos condicionais e probabilísticos. Além disso, consta lacuna sugerida em Silva, Morino e Sato (2014) para extensão do estudo de extração de regras de associação usando a base do Enem, para outras regiões do país.

Em suma, a investigação dos fatores que moldam os diferentes níveis de desempenho maranhense no Enem pode colaborar para a melhoria dos indicadores educacionais, contribuindo não apenas para o avanço da educação no Maranhão, mas também com potencial para servir de modelo para outras regiões do país.

Por fim, a pesquisa se insere em uma agenda global que contempla a educação como ferramenta para o desenvolvimento sustentável, reconhecendo que o acesso equitativo à educação de qualidade é crucial para o enfrentamento dos desafios sociais, econômicos e ambientais do século.

1.2 Problema e Hipótese

Neste contexto, definiu-se como problemática de estudo o desafio de compreender como os diferentes níveis de renda, escolaridade dos pais ou responsáveis, possuir computador, acesso à internet (ou outras variáveis) estão associadas ao desempenho dos estudantes maranhenses, adotando a extração de Regras de Associação enquanto estratégia de

Aprendizagem de Máquina não supervisionada para a Mineração de Dados Educacionais. Em outras palavras, busca-se uma visão mais abrangente de como as condições socioeconômicas e educacionais moldam o desempenho desses participantes nesta importante avaliação educacional investigando nichos não contemplados na análise exploratória.

Segundo Marconi e Lakatos (2022), a hipótese constitui-se como uma resposta provisória ao problema, formulada de forma explicativa ou preditiva, para fins de verificação pela pesquisa. Sendo assim, a hipótese geral do projeto é que a análise dos microdados do Enem 2023 permitirá verificar padrões no desempenho dos estudantes e identificar características ou variáveis significativas, combinadas ou não, que sugerem associação positiva ao resultado dos estudantes maranhenses nesta importante avaliação da educação básica. Para tanto, formulou-se os seguintes objetivos:

1.3 Objetivos

1.3.1 Objetivo Geral

Investigar os fatores influentes no desempenho dos estudantes maranhenses no Enem 2023, sob a perspectiva da Mineração de Dados Educacionais, partindo das ferramentas estatísticas de análise exploratória e extraíndo conhecimentos através da utilização de algoritmos de Aprendizagem de Máquina Não Supervisionada.

1.3.2 Objetivos Específicos

- Coletar, tratar e extrair um conjunto de microdados do Enem 2023 dos participantes maranhenses;
- Realizar a análise exploratória;
- Minerar os dados aplicando algoritmo de Aprendizagem de Máquina (Não Supervisionada);
- Discutir a significância dos achados pertinentes ao desempenho dos participantes.

1.4 Enem

O Exame Nacional do Ensino Médio (Enem) foi lançado em 1998 para avaliar o desempenho acadêmico dos alunos ao final do ensino básico. Trata-se de um exame com o objetivo de avaliar o domínio de competências científicas, tecnológicas e da linguagem contemporânea (Inep, 2020).

O Enem evoluiu com o passar dos anos, alterando sua abordagem metodológica, quantidade de questões e dias de aplicação. Em 2009, o exame passou a ser aplicado com 180 (cento e oitenta) questões de múltipla escolha, estruturado em quatro áreas de conhecimento: Linguagens, Códigos e suas Tecnologias; Ciências Humanas e suas Tecnologias; Ciências da Natureza e suas Tecnologias; Matemática e suas Tecnologias, além de uma prova de Redação (Inep, 2020).

Ainda em 2009, tornou-se ferramenta de acesso ao ensino superior por meio do Sistema de Seleção Unificada (Sisu) e do Programa Universidade para Todos (ProUni). Além disso, a partir de 2010, o Fundo de Financiamento Estudantil (Fies) passou a adotar o Enem como um dos critérios de seleção (Inep, 2020).

O cálculo da nota do Enem adota a Teoria de Resposta ao Item (TRI), que consiste num modelo probabilístico entre a possibilidade de acerto do participante, seu conhecimento na área e as características dos itens. Essa metodologia implica que dois candidatos com o mesmo número de acertos não terão, necessariamente, as mesmas pontuações (Inep, 2024).

Segundo o Inep, os resultados do Enem servirão como base, não apenas como mecanismo de avaliação e acesso à educação superior, mas também, para fins de acesso aos programas do governo e estudos sobre indicadores da educação nacional. Neste sentido, o Inep, por meio da Diretoria de Avaliação da Educação Básica, divulga anualmente os microdados do Enem, que reúnem os dados dos participantes por edição, de forma anonimizada, isto é, de modo a impossibilitar a identificação do candidato, em observância à LGPD (Inep, 2024).

Na edição de 2023, o Inep divulgou que o principal exame da educação nacional contou com mais de 3,9 milhões de inscritos, dos quais 48,2% já haviam concluído o ensino médio, com faixa etária predominante entre 17 e 18 anos (42,3%), sendo em maioria mulheres (61,3%). O Enem ainda materializa outras políticas públicas inclusivas, como a aplicação do exame mediante a isenção de taxa (63% na edição 2023), as pessoas e jovens em cumprimento de medida privativa de liberdade, atendimento especializado para pessoas com deficiência, videoprova em Libras aos candidatos surdos ou deficientes auditivos, dentre outros (Inep, 2023; 2024).

Em 2024, o Enem contou com 4,3 milhões de inscritos, dos quais 1,6 milhão são concluintes do ensino médio e 1,8 milhão de egressos, o que demonstra uma retomada na adesão ao exame. A maioria dos inscritos na edição é de mulheres (60,59%), e em relação à raça/cor, predominam os que se autodeclararam pardos (1.860.766), seguidos de brancos

(1.788.622) e pretos (533.861). O aumento na adesão da participação de jovens entre 16 (35,6%) e 17 anos (21,7%) foi estimulado pelo incentivo financeiro do Governo Federal através do programa Pé-de-Meia (Inep, 2024).

Em suma, o Enem é hoje a principal porta de entrada para a educação superior no Brasil, instrumento de inclusão e materialização de políticas públicas, devido também à sua integração em ações afirmativas e de compensação, atreladas à seleção (Travitizki, 2021).

1.5 Trabalhos Correlatos

A revisão de literatura buscou levantar estudos que adotavam as bases de dados abertos mantidas pelo Inep, em especial, os microdados do Enem, constatando as variáveis mais relevantes para o desempenho, ferramentas para modelagem e extração de conhecimento, e ainda, a adoção de regras de associação no contexto da MDE, em recorte das produções nos últimos 10 anos junto ao *Google Acadêmico*, Portal de Periódicos da CAPES e Anais do Congresso Brasileiro de Informática na Educação.

Lima e Brighenti (2023) buscaram avaliar o desempenho dos estudantes concluintes do ensino médio em 2019 no Enem, oriundos da rede de ensino de Minas Gerais, adotando análise exploratória para verificar o desempenho de acordo com perfil socioeconômico. As variáveis consideradas foram dependência administrativa da escolar, sexo, raça, renda e escolaridade dos pais. A estatística descritiva apontou que a maioria dos estudantes pertence ao grupo étnico de brancos, do sexo feminino, oriunda das escolas públicas estaduais, além de indicar que os melhores desempenhos estão entre os alunos da rede privada.

Dutra *et al.* (2023) realizaram uma revisão sistemática da literatura, em recorte de produções dos últimos 10 anos, buscando identificar os principais fatores que podem influenciar o desempenho dos estudantes no Enem. A pesquisa identificou fatores socioeconômicos como renda familiar, idade, sexo, raça, nível de escolaridade dos pais, tipo de escola e sua dependência administrativa como os mais influentes para o desempenho. No tocante às técnicas de mineração de dados, a revisão apontou a predominância de tarefas de Aprendizado de Máquina Supervisionado, por meio de algoritmos de classificação e regressão, além de técnicas de seleção e extração de *features*, associação e agrupamento.

Elias de Sousa *et al.* (2021) desenvolveram pesquisa com o objetivo de avaliar como os fatores sociais e econômicos influenciam o desempenho dos candidatos do Enem na Região Nordeste, entre os anos de 2015 e 2019, utilizando a base de dados mantida pelo Inep. A partir da aplicação de técnicas de mineração de dados como análise exploratória e modelos

de regressão, constatou-se que a escolaridade dos pais, o tipo de escola (privada), localização urbana da escola, raça (branca), a disponibilidade de recursos como computador, celular e acesso à internet são elementos cruciais para elevar a pontuação média, enquanto as variáveis idade, sexo, baixa renda e quantidade de moradores no domicílio podem sugerir prejuízos aos candidatos.

Soares *et al.* (2023) aplicaram diferentes técnicas de mineração de dados, como análise exploratória, de correlação, algoritmos de regressão e árvore de decisão, adotando a metodologia CRISP-DM, com o objetivo de identificar fatores que influenciam a qualidade educacional da rede pública estadual de ensino do Maranhão. Foram utilizadas as bases de dados do Ideb e Saeb, em estudo que apontou o nível de escolaridade dos pais ou responsáveis, possuir computador e acesso à internet como elementos propulsores da qualidade educacional, apesar de algumas limitações enfrentadas acerca da ausência de dados para algumas escolas.

Silva, Morino e Sato (2014) adotaram as regras de associação, enquanto técnica de MDE, para descobrir os relacionamentos existentes entre os fatores socioeconômicos e o desempenho dos participantes do Enem 2010 das capitais da região Sudeste: São Paulo, Rio de Janeiro, Belo Horizonte e Vitória. O pré-processamento dos dados e a exclusão dos faltosos resultou em uma amostra de 310 mil observações. Em relação às questões socioeconômicas, os autores selecionaram quatro variáveis para fins de investigação: quantidade de moradores na residência, escolaridade da mãe do participante, renda familiar mensal e o tipo de escola (pública ou privada). A variável desempenho considerou apenas a nota da prova objetiva categorizada em quatro intervalos: insuficiente (0 a 4 pontos), regular (4,01 a 6 pontos), bom (6,01 a 8 pontos) e excelente (8,01 a 10 pontos). A extração das regras com o algoritmo Apriori (Agrawal; Srikant, 1994) contemplou quatro simulações com o hiperparâmetro de confiança mínima fixada em 70% e o suporte mínimo definido em 50%, 30%, 25% e 15%, produzindo 2, 24, 42 e 214 regras, respectivamente. Algumas das principais regras observadas apontaram a associação dos alunos oriundos de escolas públicas e o desempenho regular, com suporte de 53% e confiança de 76%; residir com quatro a sete pessoas implica em estudar em escola pública com 74% de confiança e suporte em 31%; a escolaridade dos pais de primeiro grau implicando em o aluno estudar em escola pública e obter nota regular, com 90% de confiança, para 28% do público estudado; ou ainda, se o participante teve nota regular, escolaridade dos pais até o primeiro grau e a renda familiar mensal entre um e três salários mínimos então o participante é oriundo de escola pública com 92% de confiança, para 19% da amostra. Em suma, observa-se a baixa renda, nível primário

de escolaridade dos pais e a quantidade de pessoas no domicílio associados ao desempenho de nível regular.

Gomes, Gouveia e Batista (2017) buscaram identificar as regras de associação entre o desempenho no Enem e os fatores socioeconômicos dos participantes da região Nordeste, nas edições 2013 e 2014. Os 24 atributos selecionados consideraram os dados do participante como idade, localidade da residência (capital ou interior), tipo de escola de ensino fundamental e médio, o questionário socioeconômico e o desempenho na parte objetiva das provas. A extração das regras com o algoritmo Apriori contemplou quatro simulações considerando os dados da região Nordeste e do Estado de Pernambuco em cada edição, com os hiperparâmetros de suporte e confiança mínimos definidos em 10% e 80%, respectivamente. Algumas das principais regras geradas apontaram forte tendência entre a pontuação de 300 a 500 pontos ao fato de possuírem renda familiar mensal de até dois salários mínimos.

Silva *et al.* (2020) adotaram a clusterização e regras de associação para descobrir os relacionamentos existentes entre os grupos constatados, fatores socioeconômicos e o desempenho dos participantes do Enem 2019 do Estado de Minas Gerais. O pré-processamento dos dados considerou apenas os concluintes do ensino médio, isto é, aqueles que estavam cursando o último ano dessa etapa educacional, bem como, a exclusão dos inscritos com dados faltantes e notas zero em alguma das áreas da prova, resultando numa amostra de 88.659 observações. Baseados em análise descritiva, os autores selecionaram os atributos considerados mais relevantes para o desempenho: raça, dependência administrativa da escola, posse de computador no domicílio, quantidade de moradores na residência, escolaridade da mãe e renda familiar mensal. A variável desempenho considerou a média aritmética das cinco áreas da prova, sendo categorizada em quatro intervalos de mesma largura: abaixo de 452,9; entre 452,9 e 585,6; entre 585,6 e 718,2; e a partir de 718,2 pontos; sendo em seguida, normalizados. Os experimentos consideraram o método *k-means* para a clusterização baseada nas notas e o algoritmo Apriori para a extração das regras de associação, configurado com suporte mínimo de 20%, confiança de 70% e *lift* maior ou igual a 1, aplicado em toda a base e em cada um dos clusters obtidos, totalizando três simulações. Algumas das principais regras observadas quando do uso do Apriori em toda a base evidenciaram: a participação no cluster A de menor desempenho (452,9 e 585,6 pontos) com 72% de confiança e *lift* de 1,34, e no cluster B de maior desempenho (585,6 e 718,2 pontos) com 99,9% de confiança e *lift* de 2,15; para os inscritos autodeclarados pretos, pardos, amarelos ou indígenas e oriundos das escolas estaduais implicando em compor o primeiro

cluster (A) com uma confiança de 70,1%; os participantes cujas mães não concluíram o ensino médio e possuem renda familiar mensal abaixo de R\$ 2.000,00 integrariam o cluster A com uma confiança de 73,6%; os inscritos que possuem computador na residência e integram o cluster B implicam em mãe que concluiu o ensino médio com uma confiança de 75,7%. Além disso, a pesquisa encontrou similaridades entre o desempenho dos alunos oriundos das escolas federais e da rede privada, assim como, entre os alunos das escolas estaduais e municipais.

A Tabela 1 resume as variáveis constatadas e respectivas referências nos estudos, bem como sua contribuição ao desempenho:

Tabela 1 – Variáveis influentes para o desempenho escolar segundo a literatura

Variável	Contribuição	Referência
Nível de escolaridade dos pais	+	Dutra <i>et al.</i> (2023), Elias de Sousa <i>et al.</i> (2021), Soares <i>et al.</i> (2023), Silva, Morino e Sato (2014), Silva <i>et al.</i> (2020).
Tipo de escola	+	Dutra <i>et al.</i> (2023), Elias de Sousa <i>et al.</i> (2021), Silva, Morino e Sato (2014), Gomes, Gouveia e Batista (2017).
Dependência administrativa da escola	+	Dutra <i>et al.</i> (2023), Silva <i>et al.</i> (2020).
Localização da escola	+	Elias de Sousa <i>et al.</i> (2021).
Localização do domicílio	+	Gomes, Gouveia e Batista (2017).
Raça (branca)	+	Dutra <i>et al.</i> (2023), Elias de Sousa <i>et al.</i> (2021).
Renda	+	Dutra <i>et al.</i> (2023), Elias de Sousa <i>et al.</i> (2021), Silva, Morino e Sato (2014), Silva <i>et al.</i> (2020), Gomes, Gouveia e Batista (2017).
Computador	+	Elias de Sousa <i>et al.</i> (2021), Soares <i>et al.</i> (2023), Silva <i>et al.</i> (2020).
Celular	+	Elias de Sousa <i>et al.</i> (2021).
Acesso à internet	+	Elias de Sousa <i>et al.</i> (2021), Soares <i>et al.</i> (2023).
Sexo	-	Dutra <i>et al.</i> (2023), Elias de Sousa <i>et al.</i> (2021).
Idade	-	Elias de Sousa <i>et al.</i> (2021).
Quantidade de Moradores no Domicílio	-	Elias de Sousa <i>et al.</i> (2021), Silva, Morino e Sato (2014), Silva <i>et al.</i> (2020).

Fonte: elaboração própria.

O breve recorte de literatura observou algumas variáveis com associação ou correlação para o desempenho no Enem, ora positiva como escolaridade dos pais, renda familiar mensal, posse de computador, celular, acesso à internet, tipo de escola, ou negativa como raça (não branca), sexo, idade, baixa renda e quantidade de moradores no domicílio. Ainda evidenciou que apesar do crescente avanço de estudos na área da MDE, essa possui muitas potencialidades dentro da Aprendizagem Não Supervisionada, considerando a predominância de estudos para regressão e classificação do desempenho. Cabe dizer que as principais problemáticas investigadas dizem respeito ao desempenho e evasão escolar.

Observou-se ainda, a escassez de estudos no âmbito da MDE que contemplem o Estado do Maranhão e adotem a base de dados abertos do Enem, o que assevera a relevância da pesquisa e desperta o olhar para investigar como as variáveis influentes para o desempenho, segundo a literatura, moldam o desempenho neste público.

No tocante à extração de regras de associação no âmbito da MDE, observou-se apenas a adoção do algoritmo Apriori (Agrawal; Srikant, 1994) em diferentes simulações e configurações para os hiperparâmetros de suporte e confiança mínimos, exigindo do analista maior domínio dos dados para estabelecê-los e realizar a análise qualitativa das regras. Na Tabela 2 tem-se um resumo das configurações dos experimentos observados para a MDE do Enem, quando da extração das regras de associação com o algoritmo Apriori:

Tabela 2 – Configurações dos experimentos das regras de associação com o Apriori

Base	Observações	Variáveis	minsup	minconf	Qtde de Regras	Referência
Enem 2010	310.000	5	50%	70%	2	Silva, Morino e Sato (2014)
			30%	70%	24	
			25%	70%	42	
			15%	70%	214	
Enem 2014	2.456.272	24	10%	80%	-	Gomes, Gouveia e Batista (2017)
	432.966	24	10%	80%	-	
Enem 2013	-	24	10%	80%	-	
	-	24	10%	80%	-	
Enem 2019	88.659	7	20%	70%	80	Silva <i>et al.</i> (2020)
	46.419	7	20%	70%	-	
	40.240	7	20%	70%	-	

Fonte: elaboração própria.

Em síntese, os estudos apontaram que a referida estratégia pode colaborar ao fornecer elementos e subsídios para apoiar a tomada de decisão, fortalecer políticas públicas (Gomes; Gouveia; Batista, 2017), mapear variáveis significativas para o desempenho, constatar modelos de êxito no campo educacional e contribuir para a melhoria dos indicadores sociais (Silva *et al.* 2020).

Assim, o diferencial deste estudo está na escolha do Maranhão como objeto de análise, um estado que enfrenta desafios educacionais significativos, conferindo relevância social à investigação. Além disso, emprega-se a aprendizagem de máquina não supervisionada para conduzir uma análise descritiva, em contraste com a predominância de abordagens preditivas baseadas em algoritmos supervisionados. Quanto aos experimentos, adota-se um algoritmo de extração de regras de associação não evidenciado nos estudos correlatos, com melhor desempenho em termos de custo computacional do que o algoritmo Apriori.

1.6 Estrutura do Trabalho

A dissertação possui cinco capítulos. O primeiro capítulo apresenta a Ciência de

Dados como instrumento para enfrentar os desafios do século XXI, em especial, os problemas educacionais e colaborar para a efetivação de uma Educação inclusiva, equitativa e de qualidade. Enumera a recente legislação voltada à transparência que culmina na divulgação das bases de dados educacionais mantidas pelo governo. Introduce os aspectos que suscitam o estudo da educação estadual maranhense a partir da base de microdados do Enem, motivando e justificando a pesquisa. Ao problematizar a pesquisa, são elencados os objetivos geral e específicos. Em seguida, apresenta-se o Enem e os trabalhos correlatos que adotavam as bases de dados abertos mantidas pelo Inep, em especial, os microdados do Enem, constatando as variáveis mais relevantes para o desempenho e as estratégias de mineração de dados.

O segundo capítulo contempla a fundamentação teórica, baseada na Ciência de Dados como instrumento para conversão de dados em conhecimento, abordando a Ciência de Dados Educacionais, Aprendizagem de Máquina e, por fim, apresenta-se o modelo de extração das Regras de Associação, as métricas de avaliação e o princípio de funcionamento do algoritmo FP-Growth (Han; Pei; Yin, 2000).

O terceiro capítulo dedica-se aos experimentos computacionais, apresentando inicialmente os procedimentos metodológicos, baseados no processo CRISP-DM, declarando as seis etapas adequadas ao presente estudo, as configurações do algoritmo FP-Growth e os recursos computacionais. Em seguida, são cumpridas as etapas de compreensão do domínio, compreensão dos dados e preparação dos dados, que inclui a seleção dos atributos.

O quarto capítulo apresenta um recorte da análise exploratória nacional, com uma visão geral dos participantes, do desempenho e do Maranhão, avaliando as variáveis selecionadas e seus comportamentos em relação ao desempenho. Em seguida, procede-se à extração das Regras de Associação com as 10 (dez) variáveis selecionadas e com a variável resultante da média aritmética das cinco áreas da prova, categorizada por quartil. O capítulo é encerrado com a avaliação e discussão.

O quinto capítulo traz as conclusões, contribuições e trabalhos futuros. Ainda destaca o diferencial do trabalho, aponta as limitações inerentes à qualidade da base, bem como as oportunidades de melhoria.

O trabalho é encerrado com indicação das referências e do Anexo - Dicionário de Microdados do Enem 2023.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Ciência de Dados

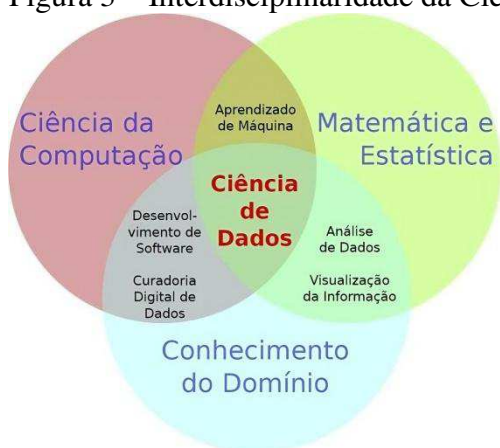
Carvalho, Menezes e Bonidia (2024) descrevem a Ciência de Dados como uma “área de conhecimento que estuda e aplica princípios e técnicas, implementados por meio de algoritmos, para extrair conhecimento novo, relevante e útil de um conjunto de dados”. Os referidos autores mencionam que, enquanto *Big Data* contempla a captação, armazenamento, processamento e transmissão de grandes volumes de dados, a Ciência de Dados lida com a extração de conhecimento de forma eficiente para apoiar o processo decisório.

Filatro (2021), por sua vez, declara que a Ciência de Dados pode ser definida como “a disciplina que fornece princípios, metodologias e orientações para transformação, validação, análise e criação de significado a partir de dados”. Segundo a autora, a Ciência de Dados é bem mais complexa do que parece, pois utiliza técnicas semelhantes à ciência da computação que contemplam a coleta de dados, armazenamento, gerenciamento, compartilhamento, proteção de dados, modelagem, análise e segurança de dados, bem como a integração de todos esses serviços.

Essa complexidade pode ser justificada pela convergência de áreas da qual se nutre a Ciência de Dados, ao integrar a – Ciência da Computação – que fornece algoritmos e técnicas para processar e analisar grandes conjuntos de dados, aproveitando os avanços em *hardware*, *software*, incluindo o Aprendizado de Máquina (*Machine Learning*), - Estatística - para identificar padrões e fazer previsões, e – Matemática – que fornece os instrumentos que conversam com os dados, como a modelagem, álgebra linear, probabilidade, etc (Filatro, 2021).

Rautenberg e Carmo (2019) afirmam que, para trabalhar com Ciência de Dados, além dos campos do conhecimento apresentados por Filatro (2021), é necessário o “Conhecimento do Domínio”, isto é, da área de aplicação na qual reside um problema – um negócio ou organização – de onde serão realizadas as análises de dados, extração de conhecimento e o fortalecimento do processo decisório. A Figura 3 reúne esse complexo de áreas destacando as possíveis intersecções e contribuições duas a duas, bem como, consolidando a Ciência de Dados como a convergência de três grandes áreas: Ciência da Computação, Matemática e Estatística e o Conhecimento do Domínio.

Figura 3 – Interdisciplinaridade da Ciência de Dados



Fonte: Rautenberg e Carmo (2019).

Dessa forma, os elementos apresentados apontam para a existência de uma metodologia, princípios e técnicas que possibilitam a produção, validação e comunicação dos conhecimentos extraídos dos conjuntos de dados, baseando-se em campos do conhecimento específicos.

Para o alcance desse objetivo são estabelecidas as etapas de um Projeto de Ciência de Dados, que, de forma resumida, começa com a definição de um problema pertinente a um domínio de aplicação, compreensão do seu contexto e recursos necessários. Em seguida, ocorre a coleta de dados, que podem ser oriundos de diversas fontes. Após a coleta, os dados são preparados, passando por transformações para correção de erros e organização. A etapa de exploração dos dados envolve análises visuais e descritivas para encontrar padrões. Com base nesses padrões e características dos dados são construídos modelos que respondem às questões do projeto. Os resultados são então comunicados de maneira clara e eficaz, utilizando diversos formatos, como relatórios e painéis. Por fim, quando necessário, a análise é automatizada para sua replicabilidade (Filatro, 2021).

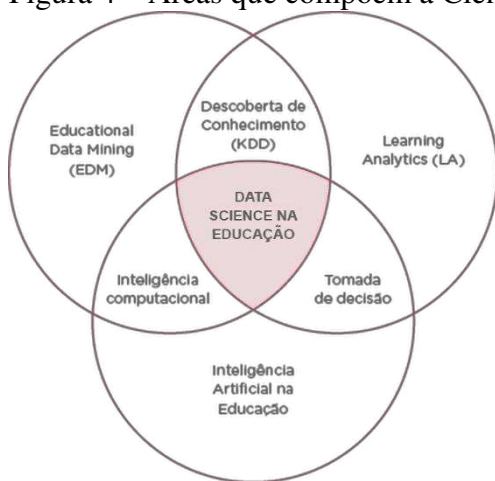
2.2 Ciência de Dados Educacionais

A Ciência de Dados, quando aplicada ao campo educacional, vai incorporar outros domínios de múltiplas disciplinas para compreender o fenômeno da educação. Essa recente área denominada *Educational Data Science*, ou Ciência de Dados Educacionais surge como subdisciplina de *Data Science* na década de 2000, a partir das conferências sobre *Educational Data Mining*, ou Mineração de Dados Educacionais (MDE). A *Learning Analytics*, ou Analítica da Aprendizagem, é outra subárea emergente a partir de 2011, assim como a *Artificial Intelligence in Education*, ou Inteligência Artificial na Educação. Esses estudos estabeleceram a Ciência de Dados Educacionais como um campo interdisciplinar que

se dedica a investigar como o *Big Data* pode influenciar a compreensão e a resolução dos desafios educacionais (Goldschmidt; Passos; Bezerra, 2015; Filatro, 2021).

Observa-se que a Ciência de Dados Educacionais já nasce subvertendo olhares acerca do mesmo objeto, possuindo mais de uma perspectiva ou abordagem para lidar com os problemas educacionais. Na Figura 4, Filatro (2021) apresenta a Ciência de Dados Educacionais como a convergência entre MDE, Analítica da Aprendizagem e Inteligência Artificial na Educação.

Figura 4 – Áreas que compõem a Ciência de Dados Educacionais



Fonte: Filatro, 2021.

Embora essas áreas estejam interconectadas compartilhando instrumentos e objetivos semelhantes, trata-se de campos distintos. Enquanto a MDE busca produzir conhecimento a partir dos dados brutos produzidos por alunos e professores, adotando técnicas como classificação, agrupamento, mineração de textos, etc, a Analítica de Aprendizagem busca compreender o processo de aprendizagem e seus ambientes, resultando em inovações no *design* instrucional de cursos, análise de evasão ou fracasso escolar, tecendo contribuições para a melhoria do processo de ensino-aprendizagem (Filatro, 2021).

Já a Inteligência Artificial (IA) é uma área da ciência da computação que se concentra no desenvolvimento de sistemas e máquinas capazes de realizar tarefas que, normalmente, exigiriam inteligência humana (Russell; Norvig, 2022).

A mudança de paradigma proporcionada pela IA na educação tem atraído o interesse cada vez maior dos sistemas educacionais. A própria forma de fazer a educação, amparada em Tecnologias da Informação e Comunicação, armazenamento de dados em nuvem, manutenção de sistemas integrados de ensino, acarreta a produção massiva de dados e influencia diretamente a própria IA. A adoção de técnicas de IA permite monitorar o desempenho dos alunos, produzir recomendações automatizadas e personalizar o ensino (Valdati, 2020).

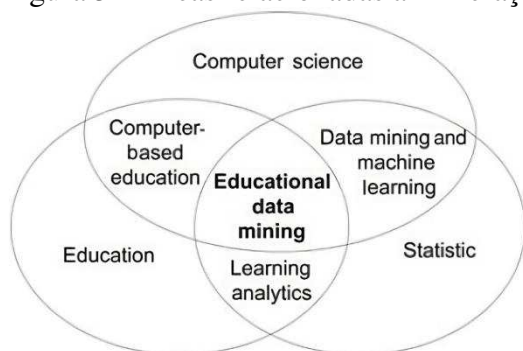
2.2.1 Mineração de Dados Educacionais

A mineração de dados vem sendo aplicada com êxito em diversos domínios. Embora a metodologia empregada para descoberta do conhecimento seja semelhante, a mineração de dados irá adequar-se ao domínio e, conseqüentemente, aos seus objetivos. Neste sentido, a MDE surge para lidar com um tipo específico de dados, aqueles produzidos em ambientes educacionais (administrativos, socioeconômicos, rendimento, sentimentos, dentre outros), fazendo uso de recursos computacionais, para contornar problemas e questões pertinentes ao domínio, como a aprendizagem, evasão, engajamento, etc (Romero; Ventura, 2013).

Goldschmidt, Passos e Bezerra (2015) também colocam a MDE como uma modalidade de mineração de dados pertinente ao domínio educacional. Na mesma trilha de entendimento, Filatro (2021) expõe a MDE como uma disciplina em ascensão que visa desenvolver métodos para analisar dados específicos provenientes de ambientes educacionais por meio dos algoritmos já existentes.

A Figura 5 ilustra a MDE como a intersecção de três grandes áreas: Ciência da Computação, Educação e Estatística, incluindo outras convergências como a Educação Apoiada por Computador, Mineração de Dados e Aprendizado de Máquina, e Analítica de Aprendizagem (Romero; Ventura, 2013).

Figura 5 – Áreas relacionadas à Mineração de Dados Educacionais

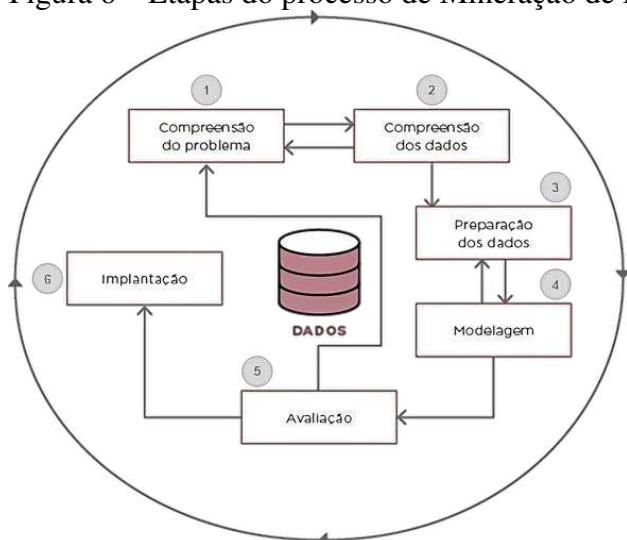


Fonte: Romero e Ventura, 2013.

Observa-se na Figura 5, que a MDE apresenta estreita relação com a Ciência de Dados ao nutrir-se dos mesmos campos de conhecimento para absorver o conhecimento advindo dos dados. De forma mais específica, edifica-se uma mineração de dados delineada para os objetivos de um domínio de aplicação específico: a Educação.

Ainda nesta trilha, Filatro (2021) admite a MDE como um processo cíclico semelhante ao ciclo de vida de um projeto de Ciência de Dados, demonstrado na Figura 6. Assim, as etapas da MDE consistem em:

Figura 6 – Etapas do processo de Mineração de Dados Educacionais



Fonte: Filatro, 2021.

As etapas da Mineração de Dados Educacionais consistem em:

1. **Compreensão do Problema:** compreender e formular claramente o problema a ser resolvido, muitas vezes por meio de um processo iterativo de descoberta.

2. **Compreensão dos Dados:** analisar e compreender os dados disponíveis, incluindo suas limitações e pontos fortes, pois raramente correspondem diretamente ao problema em questão.

3. **Preparação dos Dados:** manipular, limpar e converter os dados em formatos adequados para as técnicas de mineração de dados, exigindo maior dispêndio de tempo e esforço. Esta etapa consome 80% do tempo de execução do projeto, em média (Hua, 2024).

4. **Modelagem:** aplicar técnicas e algoritmos de Mineração de Dados/Aprendizagem de Máquina para criar modelos ou detectar padrões.

5. **Avaliação:** testar rigorosamente os modelos para confirmar sua validade e confiabilidade.

6. **Implantação:** colocar os resultados da Mineração de Dados/Aprendizagem de Máquina à disposição dos interessados, implementando modelos em sistemas ou processos educacionais (Filatro, 2021).

Cabe destacar que o referido processo guarda relação direta com a metodologia de descoberta de conhecimento em bases de dados, intitulada de CRISP-DM (*Cross-Industry Standard Process for Data Mining*), detalhada no capítulo destinado à Metodologia.

2.3 Aprendizado de Máquina

Ao explicarem o que seria Aprendizagem de Máquina (AM), Russell e Norvig (2022) colocam a experiência como elemento basilar do que seria aprender, estabelecendo a

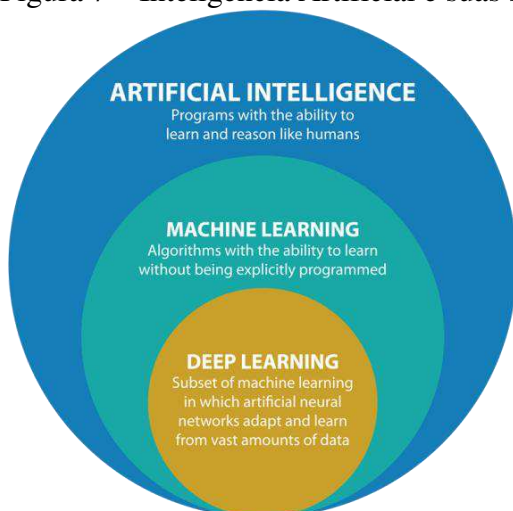
necessidade da existência de observações passadas. Partindo dessa premissa, na AM, as observações passadas seriam os dados, cujo conjunto significativo sustentaria um modelo, a partir do qual o computador, dotado de um *software*, explicaria a realidade indutivamente (Russell; Norvig, 2022).

O aprendizado reside na capacidade dos algoritmos extraírem padrões dos dados. Este processo contempla a tarefa de modelagem, isto é, a obtenção de modelos matemáticos que melhor se ajustam aos dados, na qual podem ser necessárias intervenções por meio de técnicas que irão melhorar essa “calibração”, como tratar, transformar e reamostrar os dados. A qualidade desses modelos pode ser avaliada a partir de métricas específicas, as quais irão melhor descrever o conhecimento oriundo dos dados (Carvalho; Menezes; Bonidia, 2024).

Na mesma linha, Netto e Maciel (2021) resumem a AM como métodos matemáticos usados para treinar algoritmos que extraem padrões dos dados de forma automatizada. Os autores ressaltam que a essencialidade dos dados reside no fato de constituírem a matéria-prima de construção dos modelos, sem os quais não há como treiná-los, fazer previsões ou tomar decisões.

Gabriel (2022), por sua vez, define a AM como subárea da Inteligência Artificial, a qual se dedica a realizar tarefas, até então, exclusivas das habilidades humanas. Sendo assim, a AM é extremamente abrangente, englobando uma ampla variedade de algoritmos e metodologias, cujo aprimoramento depende do processamento dos dados, incluindo o Aprendizado Profundo, do inglês *Deep Learning*, baseado em Redes Neurais Artificiais (RNA's). A Figura 7 ilustra a definição apresentada em Gabriel (2022):

Figura 7 – Inteligência Artificial e suas subáreas



Fonte: UFSM, 2021.

As RNA's são sistemas computacionais inspirados no sistema nervoso, dotados de unidades de processamento, chamadas de neurônios, estruturadas em camadas conectadas.

São considerados os modelos mais promissores atualmente, com aplicações na visão computacional, processamento de linguagem natural, previsão de séries temporais, dentre outros (Faceli *et al.*, 2024; Russell; Norvig, 2022).

Desta forma, a AM consiste numa subárea da IA que se dedica a compreender e investigar como os computadores aprendem, utilizando algoritmos para construir modelos que irão extrair conhecimento de forma automatizada, a partir do treinamento baseado num conjunto representativo de exemplos anteriores, isto é, baseado em dados.

Alguns dos conceitos acima citados abordam a palavra “treinamento” como elemento indissociável da AM. Nesse sentido, esse treinamento pode adotar quatro paradigmas ou estilos de aprendizagem: supervisionado, não supervisionado, semissupervisionado e por reforço (Carvalho; Menezes; Bonidia, 2024; Gabriel, 2022).

No primeiro caso - aprendizado supervisionado - o modelo é treinado em um conjunto de dados em que a saída desejada é conhecida. Essa busca pelos padrões que relacionam os atributos (*features*) de entrada com a variável de saída pode ocorrer através de algoritmos de Classificação, para classificar ou prever a classe, categoria ou alvo de uma instância, ou de Regressão, para prever valores quando os atributos são dados numéricos contínuos e dependentes (Netto; Maciel, 2021).

No aprendizado não supervisionado, o modelo é treinado em um conjunto de dados em que não há saídas fornecidas, ou seja, os dados não estão rotulados (ou categorizados). O objetivo é encontrar padrões nos dados, como agrupamentos naturais ou estruturas subjacentes.

No caso semissupervisionado, como apenas uma parte do conjunto de dados está rotulada, o modelo é treinado neste subconjunto para prever a classe das instâncias que não estão rotuladas. Essa tarefa pode resultar na redundância nos dados, o que pode ser contornado pela estratégia de aprendizado ativo, no qual um “oráculo” rotula as instâncias mais imprecisas a partir de critérios específicos (Faceli *et al.*, 2024; Carvalho; Menezes; Bonidia, 2024).

No aprendizado por reforço, com inspiração na psicologia comportamental, o modelo interage com um ambiente dinâmico para alcançar um objetivo particular, recebendo um sinal ou estímulo (*feedback*) em termos de recompensas ou penalidades. Trata-se de um sistema que não se baseia em exemplos, como no aprendizado supervisionado, tampouco no não supervisionado, já que possui um modelo definido, mas sim incentivando o aprendizado de comportamentos que levam a recompensas (Russel; Norvig, 2022; Gabriel, 2022).

A depender do tipo de tarefa executada, os algoritmos de AM podem ser classificados em preditivos e descritivos. Nas tarefas preditivas, os algoritmos são aplicados a dados rotulados para criar modelos que preveem o valor de um atributo para novos dados (desconhecidos), fazendo uso do aprendizado supervisionado para classificação ou predição (regressão) de uma classe. Já nas tarefas descritivas, os algoritmos extraem padrões dos dados sem prever valores específicos, usando aprendizado não supervisionado. Exemplos de tarefas descritivas incluem agrupamento de dados e identificação de regras de associação (Faceli *et al.*, 2024; Carvalho; Menezes; Bonidia, 2024).

Silva, Peres e Boscarioli (2016) resumem as principais tarefas de mineração de dados da seguinte forma:

- **Classificação:** corresponde a identificar corretamente a classe ou categoria de uma observação, numa base formada por um número discreto e finito de categorias.
- **Regressão:** consiste em estimar um valor numérico contínuo para uma variável de saída com base em variáveis de entrada, numa base de dados numérica e de valores contínuos.
- **Agrupamento (clusterização):** estratégia não supervisionada que organiza as observações em grupos (clusters) com base em similaridades.
- **Associação (ou descoberta de regras de associação):** estratégia não supervisionada que encontra relações frequentes entre atributos de um conjunto de dados.

Tais pontos são extremamente relevantes, pois estão diretamente associados à escolha do algoritmo de acordo com a natureza dos dados e tipo de problema que se deseja resolver, isto é, definir o modelo matemático que realizará a tarefa de constatar as relações entre os atributos e que estão subjacentes aos dados (Netto; Maciel, 2021).

2.3.1 Regras de Associação

A mineração de regras de associação é uma técnica usada para estabelecer relações sob a forma de regras entre os itens de uma base de dados transacional, isto é, ao invés da estrutura convencional de linhas para as observações e colunas para os atributos, trabalha-se com uma base em que cada observação é convertida numa transação. O objetivo é encontrar regras fortes, de acordo com alguma medida de interesse, diferenciando-se do agrupamento (busca de similaridade) e da correlação (avaliação da dependência linear entre

itens), ao constatar que a ocorrência de um item implica no outro, ou seja, quando um ocorre, o outro também ocorre (de Castro; Ferrari, 2016).

As regras de associação possuem aplicação em diversos campos, sendo mais comumente utilizadas na análise de compras para entender padrões de comportamento de consumidores, ao identificar que a compra de um produto pode estar associada à compra de outro, auxiliando empresas no desenvolvimento de estratégias de marketing e vendas. A sua aplicabilidade se estende a áreas como a análise do mercado financeiro, à melhoria do treinamento e desempenho físico, comportamento eleitoral, serviços bancários e até na área da saúde, ajudando a prever doenças a partir de hábitos dos pacientes. Essas associações têm como característica positiva a sua interpretabilidade, fato não evidenciado em boa parte das tarefas de mineração de dados, e, uma vez mineradas, podem auxiliar no aprimoramento de processos em diferentes domínios (Silva; Peres; Boscarioli, 2016).

Matematicamente, considera-se $A = \{a_1, \dots, a_m\}$ o universo de m itens, I qualquer subconjunto de itens (*itemset*) tal que $I \subset A$ e T a base de dados convertida em n transações, isto é, $T = \{t_1, \dots, t_n\}$. Cada transação é um par $(tid_i, k\text{-}items_i)$ em que tid_i é a identificação da transação e $k\text{-}item_i \subset A$ é um conjunto de k itens. O conjunto $K(I)$ de transações que suporta o *itemset* I é designado por suporte do *itemset*. Um conjunto I de itens é considerado frequente se o suporte de I , definido como a proporção de transações em T que contêm I , é maior ou igual a um valor mínimo de suporte pré-definido. A partir dos conjuntos frequentes, são derivadas as regras de associação que possuem a forma de “se antecedente então consequente”, em que tanto o antecedente quanto o consequente são conjuntos frequentes de itens (Faceli *et al.*, 2024).

Para lidar com a grande quantidade de regras de associação que podem ser derivadas, é importante restringir a geração de regras de acordo com transações relevantes quantitativamente (suporte), bem como, a proporção da regra em relação às transações (confiança), estabelecendo valores mínimos de suporte (minsup) e de confiança (minconf) adequados (de Castro; Ferrari, 2016).

A ideia de Suporte está associada à frequência do elemento na transação, seja do item ou da regra. Seja $\sigma(A)$ a contagem do suporte de itens do conjunto A , bem como $\sigma(A \cup C)$ a contagem do suporte da regra, o Suporte da regra $A \rightarrow C$, descreve a frequência em que todos os itens dos dois conjuntos de itens, A e C , aparecem juntos em transações individuais da base (Silva; Peres; Boscarioli, 2016). O suporte da regra $A \rightarrow C$ pode ser dado em relação ao total de transações da base $\sigma(T)$ conforme Equação 1:

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

A confiança da regra $A \rightarrow C$ diz respeito à possibilidade de ocorrência da regra em relação ao suporte do seu antecedente (A), que também pode ser descrita como a probabilidade condicional $P(C|A)$ (Silva; Peres; Boscarioli, 2016), conforme Equação 2 a seguir:

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

A análise ou seleção das regras de associação ainda pode ser refinada pela medida de correlação chamada *lift*, que consiste na relação entre a Confiança da Regra e seu consequente, conforme Equação 3:

$$Lift(A \rightarrow C) = \frac{Conf(A \rightarrow C)}{\sigma(C)} = \frac{\sigma(A \cup C)}{\sigma(A) \cdot \sigma(C)} \quad (3)$$

Essa medida (*lift*) assume valores de 0 a infinito e indica a importância da regra de associação. Para valores abaixo 1 tem-se correlação negativa entre os itens de A e C , correlação positiva para valores acima de 1 e ausência de correlação para valores próximos a 1 (Silva; Peres; Boscarioli, 2016).

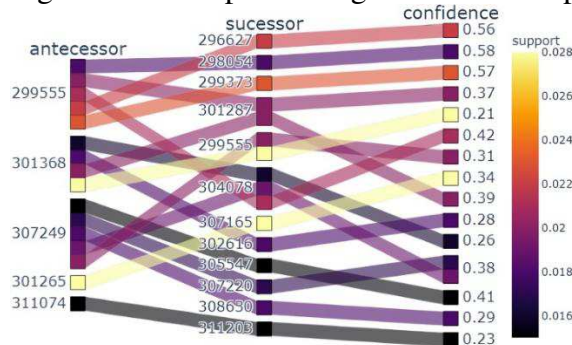
Dois dos principais algoritmos de regras de associação são o Apriori (Agrawal; Srikant, 1994) e o *FP-Growth* (Han; Pei; Yin, 2000), os quais utilizam atributos categóricos nominais e/ou codificados na forma binária (Silva; Peres; Boscarioli, 2016) e contemplam as etapas de levantamento dos itens ou conjuntos de itens mais frequentes, para, em seguida, produzir as regras (Faceli *et al.*, 2024).

Um exemplo da adoção de regras de associação para fins comerciais tem-se em Silva e Guedes (2023), que propuseram uma abordagem para geração e visualização de regras de associação em portal de notícias, com o objetivo de compreender os padrões nas preferências dos usuários e sugerir melhorias no sistema de recomendação de conteúdos, em percurso metodológico que envolveu quatro etapas: a análise exploratória, pré-processamento dos dados, geração das regras com o algoritmo Apriori e a visualização.

Dois pontos sensíveis foram contemplados neste estudo: a) a definição dos hiperparâmetros do algoritmo; e b) a visualização das regras geradas. Para o estabelecimento do hiperparâmetro de suporte mínimo foi considerado aquilo que a análise exploratória apontou como sendo uma publicação muito acessada (1,5% em relação ao universo de 147.155 entradas), com 20% de chances de ocorrência (confiança mínima) e *lift* > 1, prosseguindo com outros dois experimentos cujos ajustes mais restritivos reduziam e

refinavam a quantidade de regras. Para o segundo ponto, adotou-se o diagrama de coordenadas ou barras paralelas disponível no pacote *plotly.express* do Python, ilustrado na Figura 8, em que cada coluna/barra corresponde a uma categoria, facilitando a comparação de frequências relativas segundo as alturas das faixas.

Figura 8 – Exemplo de diagrama de barras paralelas



Fonte: Silva e Guedes (2023).

Assim, a adoção de regras de associação pode ser estendida a diversos domínios, mas requer que o analista tenha um profundo entendimento da natureza dos dados para mapear e identificar padrões subjacentes de maneira eficaz, utilizando configurações adequadas para extração e visualização das regras.

Adota-se neste estudo o algoritmo FP-Growth (Han; Pei; Yin, 2000), mais eficiente que o Apriori (Agrawal; Srikant, 1994), devido à abordagem de compressão da base de dados numa estrutura de árvore (*FP-Tree*), reduzindo o espaço de busca e os candidatos aos padrões frequentes (de Castro; Ferrari, 2016).

2.3.1.1 Algoritmo FP-Growth

O algoritmo FP-Growth é uma técnica eficiente para encontrar padrões frequentes em grandes conjuntos de dados, baseado em duas etapas principais: compressão da base de dados na FP-Tree (*Frequent Pattern Tree*) varrendo a base de dados duas vezes para, em seguida, realizar a Extração de Padrões Frequentes (Faceli *et al.*, 2024). Na ilustração da Figura 9 apresenta-se o pseudocódigo de construção da árvore:

Figura 9 – Algoritmo *FP-tree*

Entrada: Um Banco de Dados de Transações *DB*;

Um limiar do suporte mínimo σ ;

Saída: Todos os *itemsets* com suporte maior que σ

- 1 Varrer o banco de dados *DB* uma vez;
- 2 Coletar *F*, o conjunto de itens frequentes e o suporte de cada item;
- 3 Ordenar *F* em ordem decrescente de suporte como *Flist*;
- 4 Criar a raiz FP-tree, *T*, e rotulá-la como *null*;
- 5 **para cada** transação *t* $\in$ *DB* **faça**
- 6 Selecionar os itens frequentes em *t*;

```

7      Ordená-los de acordo com Flist;
8      Seja [ilIts] os itens frequentes ordenados em t;
9      Chamar insere_arvore([ilIts], T);
10 fim
11 Função insere_arvore([ilIts], T);
12 se T tem um filho N rotulado i então
13     Incremente os contadores de N com 1;
14 fim
15 senão
16     Crie um nó, N, com seu contador inicializando com 1, seu pai ligado a T, e
        seu nó ligado aos nós com o mesmo rótulo i;
17     se Its é não vazio então
18         Chamar insere_arvore(Its, N);
19     fim
20 fim

```

Fonte: Faceli *et al.* (2024, p. 172).

A primeira varredura tem por objetivo contabilizar os itens frequentes de dimensão unitária, isto é, os *1-itemsets*, os quais serão adotados como os nodos da estrutura, após ordená-los de forma decrescente numa matriz. Este procedimento faz com que as categorias de itens mais frequentes sejam alocadas mais próximas da raiz da árvore, reduzindo o espaço de busca para a posterior extração dos itens frequentes (Faceli *et al.*, 2024).

A Figura 10 ilustra a primeira varredura em uma base com 10 transações, descartando os itens não frequentes segundo um suporte mínimo pré-definido de 20%, para em seguida, ordená-los dentro de cada transação:

Figura 10 – Primeira varredura da base transacional

TID	Itens	Item	Suporte	Suporte(%)	TID	Itens
1	{b, a, f}	a	8	80%	1	{a, b}
2	{b, d, c}	b	7	70%	2	{b, c, d}
3	{a, c, e, d}	c	6	60%	3	{a, c, d, e}
4	{a, d, e}	d	5	50%	4	{a, d, e}
5	{b, c, a}	e	3	30%	5	{a, b, c}
6	{a, b, c, d}	f	1	10%	6	{a, b, c, d}
7	{a}	item 'f' descartado			7	{a}
8	{a, c, b}				8	{a, b, c}
9	{a, b, d}				9	{a, b, d}
10	{b, e, c}				10	{b, c, e}

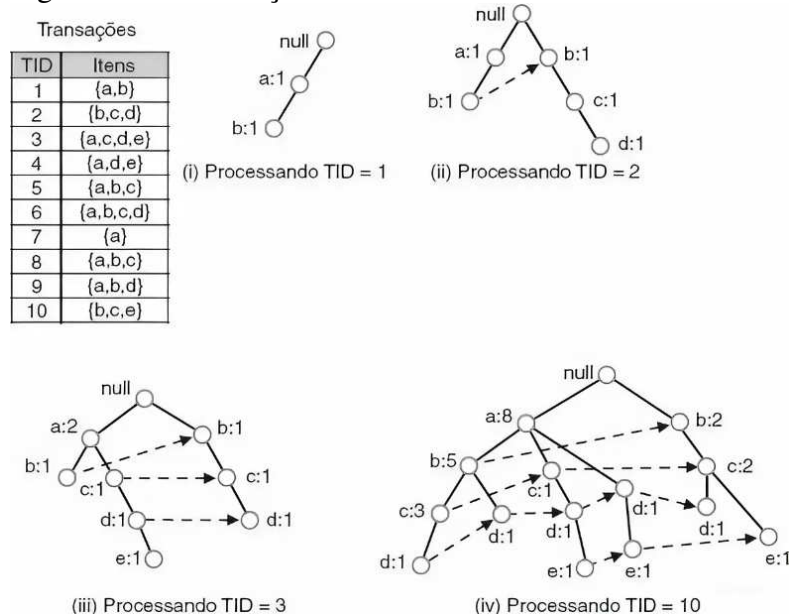
Fonte: adaptado de Fonseca, 2024.

A Figura 11 ilustra a segunda varredura para construção da árvore, em que cada transação é inserida priorizando os itens mais frequentes, isto é, segundo a ordem decrescente do suporte dos *1-itemsets*, a partir do nó raiz (*null*). Cada *1-itemset* corresponderá a um nodo inserido e, quando não sobrepostos, serão criados novos ramos recebendo ponteiros (setas tracejadas), conectando-os em diferentes caminhos.

Na citada ilustração, a inserção da primeira transação (TID=1) promove o caminho *null*→a→b, criando os nodos 'a' e 'b' e inicializando seus respectivos contadores na

etapa (i). A inserção da segunda transação (TID=2) apresenta um caminho distinto do primeiro, já que não possuem o mesmo prefixo (item 'a'), sendo criado novo percurso a partir de *null* na etapa (ii) com seus respectivos contadores. Neste momento é inserido um ponteiro entre os nodos 'b'. A inserção da terceira transação (TID=3), por sua vez, possui o prefixo 'a' em comum com TID=1, sendo este revisitado e tendo sua frequência incrementada na etapa (iii). Já os nodos 'c', 'd' e 'e' têm ponteiros com os nodos coincidentes de TID=2, para fins de contabilização. Esse processo iterativo é realizado em toda a base transacional até TID=10, indicada na etapa (iv).

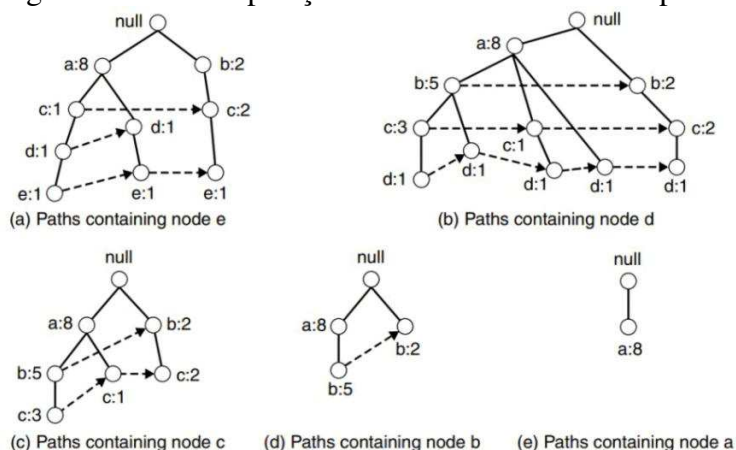
Figura 11 – Construção da FP-Tree



Fonte: Faceli *et al.*, 2024.

A fase de extração de padrões frequentes adota uma abordagem baseada em sufixos, isto é, analisando a árvore de forma ascendente e mapeando todos os caminhos até aquele nodo, auxiliado pelos ponteiros que irão registrar cada ocorrência em diferentes transações. A Figura 12 ilustra a busca por conjuntos de itens frequentes por meio da decomposição da árvore em vários problemas menores que terminam em 'e', depois em 'd', 'c', 'b' e, por fim, em 'a' (Tan; Steinbach; Kumar, 2009).

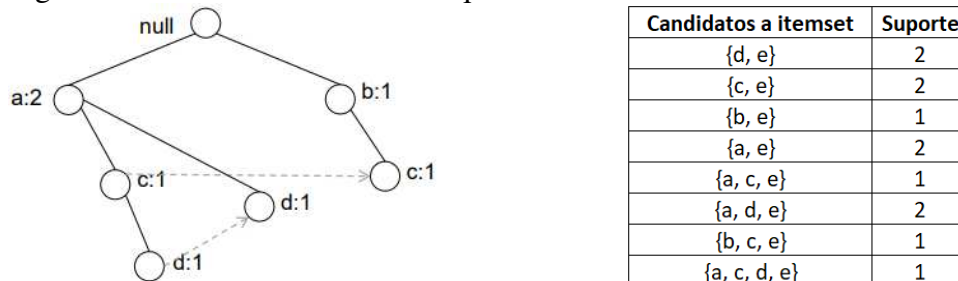
Figura 12 – Decomposição da árvore em vários subproblemas



Fonte: Tan, Steinbach e Kumar, 2009.

Para cada subproblema da Figura 12 é possível extrair as árvores condicionais para a ocorrência de cada nodo, para fins de exploração de todas as possibilidades de conexão, seja para *1-itemset*, *2-itemsets* ou *3-itemsets*, etc, atualizando-se os respectivos suportes. Na Figura 13 tem-se a ilustração da Árvore FP condicional em ‘e’, obtida a partir da eliminação de todos os nodos ‘e’ na Figura 12(a), com a indicação dos caminhos (*itemsets*) e suportes, observando-se a contagem do nodo a fim de que não haja perda de informação:

Figura 13 – Árvore FP condicional para ‘e’



Fonte: adaptado de Fonseca, 2024.

A Tabela 3 demonstra a extração de todos os candidatos a *itemset* a partir das árvores condicionais de cada um dos nodos (‘e’, ‘d’, ‘c’, ‘b’, ‘a’), eliminando-se aqueles que não atendam ao suporte mínimo de 20%, culminando na extração das regras de associação.

Tabela 3 – Extração das regras de associação

Item	Base Padrão Condicional	Contagem do item	Candidatos a itemset	Itemset Frequente	Regras de Associação
e	{{a,c,d:1}, {a,d:1}, {b,c:1}}	a:2, b:1, c:2, d:2	d,e:2 / c,e:2 / b,e:1 / a,e:2 /a,c,e:1 / a,d,e:2 / b,c,e:1 / a,c,d,e:1 / e:3	d,e:2 / c,e:2 / a,e:2 / a,d,e:2 / e:3	<a,e>, <c,e>, <d,e>, <a,d,e>, <e>
d	{{a,b,c:1, {a,b:1}, {a:1}, {a,c:1} {b,c:1}}	a:4, b:3, c:3	c,d:3 / b,d:3 / a,d:4 / b,c,d:2 / a,c,d:2 / a,b,d:2 / a,b,c,d:1	c,d:3 / b,d:3 / a,d:4 / b,c,d:2 / a,c,d:2 / a,b,d:2	<a,d>, <b,d>, <c,d> <a,b,d>, <a,c,d>,<b,c,d>
c	{{a,b: 3}, {a:1}, {b:2}}	a:4, b:5	c:6 / b,c:5 / a,c:4 / a,b,c:3	c:6 / a,c:4 / b,c:5 / a,b,c:3	<c>, <a,c>, <b,c>, <a,b,c>
b	{{a:5}}	a:5	a,b:5 / b:7	a,b:5 / b:7	<a,b>,
a	-	-	a:8	a:8	<a>

Fonte: adaptado de Fonseca, 2024 e Tech and Man, 2020.

A Tabela 4 exemplifica a análise de cada regra de associação, considerando as diferentes configurações para antecedentes e consequentes, e posteriormente, a seleção das regras segundo uma confiança mínima estabelecida em 60%, descartando os conjuntos frequentes de tamanho 1.

Tabela 4 – Avaliação das regras de associação

Regra	Possíveis configurações	Confiança	Situação da Regra
<c,e>	$\{c\} \rightarrow \{e\}$	$Conf(c \rightarrow e) = \frac{\sigma(c \cup e)}{\sigma(c)} = \frac{2}{6} = 0,33$	descartada
	$\{e\} \rightarrow \{c\}$	$Conf(e \rightarrow c) = \frac{\sigma(e \cup c)}{\sigma(e)} = \frac{2}{3} = 0,66$	válida
<d,e>	$\{d\} \rightarrow \{e\}$	$Conf(d \rightarrow e) = \frac{\sigma(d \cup e)}{\sigma(d)} = \frac{2}{5} = 0,40$	descartada
	$\{e\} \rightarrow \{d\}$	$Conf(e \rightarrow d) = \frac{\sigma(e \cup d)}{\sigma(e)} = \frac{2}{3} = 0,66$	válida
<a,d,e>	$\{a,d\} \rightarrow \{e\}$	$Conf(a, d \rightarrow e) = \frac{\sigma(a, d \cup e)}{\sigma(a, d)} = \frac{2}{4} = 0,50$	descartada
	$\{a,e\} \rightarrow \{d\}$	$Conf(a, e \rightarrow d) = \frac{\sigma(a, e \cup d)}{\sigma(a, e)} = \frac{2}{2} = 1,00$	válida
	$\{d,e\} \rightarrow \{a\}$	$Conf(d, e \rightarrow a) = \frac{\sigma(d, e \cup a)}{\sigma(d, e)} = \frac{2}{2} = 1,00$	válida
	$\{a\} \rightarrow \{d,e\}$	$Conf(a \rightarrow d, e) = \frac{\sigma(a \cup d, e)}{\sigma(a)} = \frac{2}{8} = 0,25$	descartada
	$\{d\} \rightarrow \{a,e\}$	$Conf(d \rightarrow a, e) = \frac{\sigma(d \cup a, e)}{\sigma(d)} = \frac{2}{5} = 0,40$	descartada
	$\{e\} \rightarrow \{a,d\}$	$Conf(e \rightarrow a, d) = \frac{\sigma(e \cup a, d)}{\sigma(e)} = \frac{2}{3} = 0,66$	válida
...	...	...	...

Fonte: adaptado de Fonseca, 2024.

O FP-Growth evita a geração explícita de candidatos a itens frequentes, partindo do sufixo padrão e avaliando quais outros itens ocorrem simultaneamente de forma frequente com esse item, minerando recursivamente a árvore. A estratégia de divisão e conquista ao decompor a extração de regras de associação em subproblemas menores reduz o espaço de busca. Essas características fazem do FP-Growth um algoritmo eficiente para a mineração de padrões frequentes em grandes bases de dados (de Castro; Ferrari, 2016).

3 EXPERIMENTOS

Nesta seção, estão descritos os procedimentos metodológicos e os detalhes de cada etapa do seu desenvolvimento. Inicialmente, é importante destacar a abordagem

quantitativa da pesquisa, dispensando análise que mapeie elementos intrínsecos ao comportamento do público em estudo, nutrindo-se, tão somente, das ferramentas estatísticas para compreender as características predominantes ou padrões em grandes conjuntos de dados e amostras (Marconi; Lakatos, 2022).

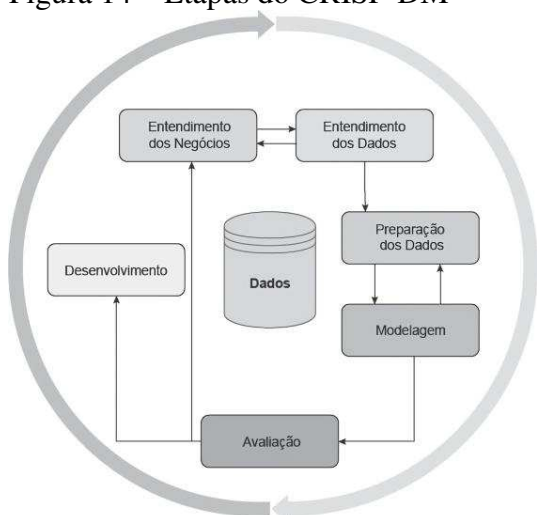
Para atingir os objetivos da pesquisa, foi realizada uma pesquisa de caráter exploratório visando maior conhecimento e familiaridade com o problema, além de constatar as discussões atuais e a evolução do conhecimento na temática, o que fomenta a elaboração de hipóteses (Gil, 2022) na seção de estudos correlatos.

Para o objeto em estudo, adotou-se o método CRISP-DM (*Cross Industry Standard Process for Data Mining*) considerando sua abrangência (compreensão do negócio e dos dados), popularidade (Sharda; Delen; Turban, 2019; Carvalho; Menezes; Bonidia, 2024), e ainda, sua adoção em larga escala para diferentes contextos, incluindo o ambiente educacional (Ramos *et al.*, 2020).

3.1 Procedimentos Metodológicos

No ano de 1996, foi proposto o CRISP-DM (*Cross-Industry Standard Process for Data Mining*) enquanto modelo promissor para extração de conhecimento em base de dados, em resposta às crescentes demandas das indústrias, modelo este ilustrado na Figura 14 (Goldschmidt; Passos; Bezerra, 2015; Carvalho; Menezes; Bonidia, 2024).

Figura 14 – Etapas do CRISP-DM



Fonte: Carvalho, Menezes e Bonidia, 2024.

As seis etapas do Processo CRISP-DM consistem em:

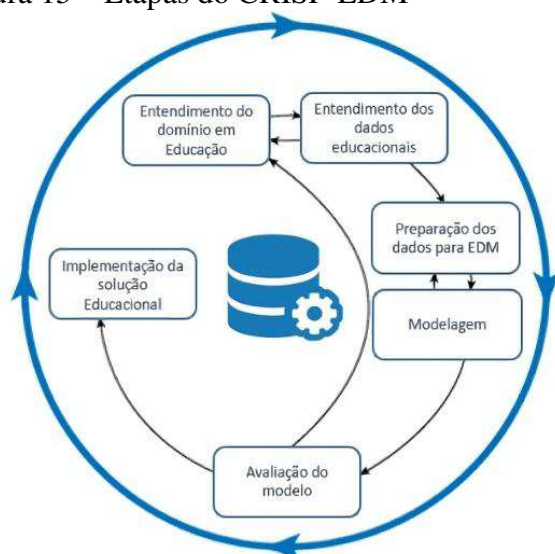
- a) Compreensão do negócio: que contempla informações e recursos do domínio de aplicação;

- b) Compreensão dos dados: verificar as fontes e bases disponíveis, bem como a qualidade dos dados;
- c) Preparação dos dados: contempla a seleção, tratamento e transformação dos dados;
- d) Modelagem: aplicação de técnicas e algoritmos para identificação dos padrões;
- e) Avaliação: verificar a qualidade dos modelos usando métricas específicas;
- f) Desenvolvimento: entrega de resultados ou produtos (relatórios, *dashboards*, etc) aos clientes e interessados (Chapman *et al.*, 2000).

A metodologia CRISP-DM é um processo não rígido que pode ser revisitado ou retroalimentado pelas etapas anteriores, adequando-se ao contexto de aplicação (Chapman *et al.*, 2000).

Apoiado em levantamento de estudos que adotam a CRISP-DM para minerar dados educacionais, Ramos *et al.* (2020) propuseram o modelo CRISP-EDM (*Cross Industry Standard Process for Educational Data Mining*), uma adaptação da metodologia para o contexto educacional, conforme ilustração da Figura 15:

Figura 15 – Etapas do CRISP-EDM



Fonte: Ramos, 2020.

O modelo proposto mantém as seis etapas da metodologia “original”, ressaltando que a principal mudança reside na etapa de “implementação da solução educacional”, que exige a validação dos resultados ou do conhecimento gerado pelas personagens do ambiente educacional: professores, gestores e especialistas (Ramos *et al.*, 2020).

Cabe dizer que há outras metodologias de mineração de dados, as quais não diferenciam ou restringem sua adoção pelo tipo de negócio, dimensão do conjunto de dados ou recurso computacional adotado. Trata-se de processos centrados em metodologias

padronizadas para lidar com problemas em diferentes domínios, tendo como pressuposto a existência de uma matéria-prima: dados (Sharda; Delen; Turban, 2019).

Considerando o detalhamento das etapas do CRISP-DM enquanto processo de extração de conhecimento de base de dados não rígida e sua adequação ao campo educacional, sintetiza-se os procedimentos metodológicos do presente estudo no diagrama apontado na Figura 16:

Figura 16 – Etapas do CRISP-DM adequadas ao presente estudo



Fonte: elaboração própria.

A seguir, detalha-se cada etapa da Figura 16:

- Compreensão do domínio:** contempla as informações gerais do tema e as perspectivas que transformam o desempenho dos participantes maranhenses no Enem 2023 em um problema de ciência de dados, bem como, os objetivos do estudo.
- Compreensão dos dados:** fonte e natureza dos dados, coleta dos microdados do Enem 2023, análise do arquivo de metadados (dicionário), identificação dos atributos e tipos de variáveis.
- Preparação dos dados:** limpeza, tratamento, transformação dos dados, categorização de acordo com o arquivo de metadados, extração do subconjunto dos participantes maranhenses, bem como, a seleção dos atributos relevantes.
- Análise Exploratória:** busca uma visão aprofundada dos dados, constatar o perfil dos participantes e propiciar uma visão geral do desempenho.
- Modelagem:** envolve a aplicação da técnica de Aprendizagem de Máquina Não Supervisionada de Regras de Associação com objetivo de identificar nichos e constatar padrões não vislumbrados da Análise Exploratória, por meio do algoritmo *FP-Growth*, conforme configurações dos experimentos elencados na Tabela 5, para $lift > 1$:

Tabela 5 – Configurações dos experimentos

EXP.	minsup	minconf	antecedentes	EXP.	minsup	minconf	antecedentes
1	3%	5%	[2,4]	7	3%	5%	[3,6]
2	10%	5%	[2,4]	8	10%	5%	[3,6]
3	10%	10%	[2,4]	9	10%	10%	[3,6]
4	20%	10%	[2,4]	10	20%	10%	[3,6]
5	20%	20%	[2,4]				
6	25%	20%	[2,4]				

Fonte: elaboração própria.

e) Avaliação: avaliação da qualidade das regras geradas segundo as métricas de suporte, confiança e *lift* e interpretação dos resultados.

f) Desenvolvimento: discussão dos achados pertinentes ao desempenho dos participantes maranhenses segundo as regras processadas.

Desse modo, o processo CRISP-DM, adequado à análise dos dados educacionais dos participantes maranhenses do Enem 2023, irá nortear a extração de conhecimentos úteis que podem orientar a melhoria da educação no Estado.

3.1.1 Recursos Computacionais

A análise de grandes volumes de dados torna-se impraticável sem o uso de ferramentas computacionais adequadas, sendo necessária sua adoção para tratar, analisar, interpretar e relacionar esses dados, traçando estratégias específicas em diferentes domínios de aplicação (Goldschmidt; Passos; Bezerra, 2015).

A linguagem de programação escolhida para a pesquisa é o *Python*, pelas características que a colocam dentre as mais utilizadas na Ciência de Dados, com as ricas ferramentas e bibliotecas de análise e visualização de dados: *NumPy*, *SciPy*, *Pandas*, *Matplotlib*, *Seaborn* e *Plotly*. Quando o assunto é *Machine Learning* destacam-se as bibliotecas como *Scikit-Learn*, *TensorFlow*, *Keras* e *PyTorch* (Netto; Maciel, 2021).

A importação, tratamento, visualização dos dados foi realizada com a aplicação *Jupyter Notebook* – versão 3.11.7, integrante da plataforma *Anaconda*, assim como, a implementação dos pacotes com os algoritmos de aprendizagem de máquina (Netto; Maciel, 2021). De modo auxiliar, utilizou-se o *Google Colaboratory*, plataforma gratuita que permite executar códigos diretamente no navegador.

O hardware utilizado é composto de processador (CPU) Intel Core i5-8500, 3.00 GHz (até 4.10 GHz), 6 núcleos; memória RAM de 8GB DDR4, armazenamento SSD 256GB, com placa gráfica (GPU) Intel UHD Graphics 630.

3.2 Compreensão do domínio

A educação está elencada no rol dos direitos sociais previstos na Constituição Federal de 1988, sendo dever do Estado ofertar uma educação pública, gratuita, com padrões mínimos de qualidade e que promova o acesso aos níveis mais elevados do ensino (Brasil, 1996).

O Plano Nacional de Educação (PNE) é o instrumento norteador dos investimentos em educação, consignando as diretrizes, objetivos, metas e estratégias para a educação nacional para um período de dez anos, que contempla importantes indicadores como o Ideb. Os últimos ciclos de avaliação do Ideb apresentaram avanços em diversas frentes, mas o Brasil ainda permanece abaixo das metas estabelecidas para a educação básica (MEC, 2024).

Os recentes resultados divulgados pelo MEC (2024) apontaram que o Maranhão não atingiu as metas do Ideb para 2021 nas três etapas da educação básica, tendo o ensino médio o pior desempenho.

Já na principal avaliação da educação básica, o Enem, o Maranhão apresentou o pior resultado médio da região Nordeste na edição 2023, segundo sinopse estatística divulgada pelo Inep (2024).

Diante desse cenário, o Maranhão torna-se um objeto de estudo relevante para a análise de indicadores educacionais, especialmente por meio de dados educacionais abertos, os quais têm embasado diversos estudos com o objetivo de avaliar os indicadores da qualidade educacional, compreender os fatores que impactam na aprendizagem, apoiar o processo decisório e fortalecer políticas públicas.

Dado o baixo desempenho do Maranhão no Ideb, especialmente no ensino médio, adota-se a base de microdados do Enem 2023, como fonte para a Mineração dos Dados Educacionais.

3.3 Compreensão dos Dados

A obtenção dos microdados do Enem ocorre por meio da divulgação anual, a cargo do Inep, em observância à LAI, LGPD e LDB. Esses dados prescindem de autorização para uso, considerando a autorização expressa para livre utilização gratuita, conforme Decreto nº 8.777/2016.

Trata-se de um conjunto de dados secundários, isto é, que foram coletados originalmente por outro ente, no caso concreto o Inep, para fins além do propósito atual de análise (Malhotra, 2012). Os dados estão estruturados em uma planilha eletrônica no formato

“.csv” (valores separados por vírgula), o que facilita a identificação, armazenagem e processamento (Carvalho; Menezes; Bonidia, 2024). Segue na Tabela 6, a indicação dos arquivos utilizados com uma breve descrição:

Tabela 6 – Fontes dos dados

Pasta	Arquivo	Descrição
DADOS	MICRODADOS_ENEM_2023.csv	Informações gerais sobre a realização das provas, a caracterização do participante, notas das provas objetivas e da redação; respostas das 25 (vinte e cinco) questões socioeconômicas.
DICIONÁRIO	Dicionário_Microdados_Enem_2023.xlsx	Informações gerais sobre as variáveis contidas na base de dados.
LEIA-ME E DOCUMENTOS TÉCNICOS	Leia_Me_Enem_2023	Breve descrição do Enem, bem como das informações sobre as bases e arquivos disponibilizados nos Microdados.

Fonte: Inep, 2024.

O arquivo “MICRODADOS_ENEM_2023.csv” reúne 76 atributos pertinentes a cada um dos participantes do Enem 2023. O detalhamento dessas variáveis, com a descrição, categorização e tipo, estão declaradas no arquivo de metadados “Dicionário_Microdados_Enem_2023.xlsx”, em Anexo, divididos em seis grupos:

- 1) Dados do Participante: reúne as primeiras 12 (doze) variáveis pertinentes à caracterização do participante (faixa etária, sexo, cor/raça, etc).
- 2) Dados da Escola: reúne 7 (sete) variáveis pertinentes à caracterização da escola do participante (município, unidade federativa, dependência administrativa, etc).
- 3) Dados do Local de Aplicação da Prova: reúne 4 (quatro) variáveis pertinentes à caracterização do local de aplicação (município, unidade federativa, etc)
- 4) Dados da Prova Objetiva: reúne 21 (vinte e uma) variáveis pertinentes à execução da prova objetiva (presença por área de conhecimento, tipo de prova, notas por área de conhecimento, respostas, gabarito, etc).
- 5) Dados da Redação: reúne 7 (sete) variáveis pertinentes à execução da prova discursiva (situação do candidato, notas por competência e nota final da redação);
- 6) Dados do Questionário Socioeconômico: reúne 25 (vinte e cinco) variáveis pertinentes à caracterização socioeconômica dos participantes (escolaridade dos pais, renda mensal, moradia, transporte, computador, acesso à internet, etc).

3.4 Preparação dos Dados

As etapas de importação dos dados, tratamento prévio e criação de um novo subconjunto de dados apenas com os participantes maranhenses foram realizadas no *Jupyter Notebook* – versão 3.11.7.

A importação do arquivo “MICRODADOS_ENEM_2023.csv”, de tamanho 1,65 GB (1.777.162.429 bytes), deu-se por meio de dois argumentos, um para usar padrão de codificação latino “ISO-8859-1” e o outro para o tipo de extensão “.csv”, no caso concreto “;”, armazenando-o numa estrutura de dados bidimensional chamada *dataframe*, com as seguintes dimensões: 3.933.955 linhas por 76 colunas.

Em seguida, todos os dados do *dataframe* foram convertidos para o tipo categórico, em observância ao arquivo de metadados “Dicionário_Microdados_Enem_2023.xlsx”, excetuando-se as notas das provas objetiva e da redação do tipo numérico (*float*).

Na Tabela 7 constam os atributos pertinentes aos Dados da Escola, os quais trouxeram apenas as informações pertinentes aos estudantes concluintes do ensino médio, isto é, daqueles no último ano do ensino médio, o que reduziu o conjunto de variáveis investigadas:

Tabela 7 – Variáveis com dados faltantes

Variável	Descrição	Dados Faltantes	
TP_ENSINO	Tipo de instituição que concluiu ou concluirá o Ensino Médio	2.594.874	66,0%
CO_MUNICIPIO_ESC	Código do município da escola	2.975.449	75,6%
NO_MUNICIPIO_ESC	Nome do município da escola	2.975.449	75,6%
CO_UF_ESC	Código da Unidade da Federação da escola	2.975.449	75,6%
SG_UF_ESC	Sigla da Unidade da Federação da escola	2.975.449	75,6%
TP_DEPENDENCIA_ADM_ESC	Dependência administrativa (Escola)	2.975.449	75,6%
TP_LOCALIZACAO_ESC	Localização (Escola)	2.975.449	75,6%
TP_SIT_FUNC_ESC	Situação de funcionamento (Escola)	2.975.449	75,6%

Fonte: adaptado do arquivo de metadados “Dicionário_Microdados_Enem_2023.xlsx”, (Inep, 2024)

A variável “TP_ESCOLA” (Tipo de escola do Ensino Médio – Dados do Participante), que continha as categorias “Pública” e “Privada”, por sua vez, não possuía dados faltantes, mas contava com 64,4% das observações na categoria de “não respondentes”, tornando a mesma inconsistente para o estudo.

Na Tabela 8 constam as variáveis que diziam respeito apenas à codificação de outro atributo ou informações técnicas da aplicação da prova, e que por essa razão, foram eliminadas da análise:

Tabela 8 – Variáveis redundantes ou desnecessárias para o estudo

N	Variável	Descrição
01	NU_INSCRICAO	Número de inscrição
02	NU_ANO	Ano do Enem
03	CO_MUNICIPIO_PROVA	Código do município da aplicação da prova

04	CO_UF_PROVA	Código da Unidade da Federação da aplicação da prova
05	CO_PROVA_CN	Código do tipo de prova de Ciências da Natureza
06	CO_PROVA_CH	Código do tipo de prova de Ciências Humanas
07	CO_PROVA_LC	Código do tipo de prova de Linguagens e Códigos
08	CO_PROVA_MT	Código do tipo de prova de Matemática
09	TX_RESPOSTAS_CN	Vetor com as resp. da parte objetiva da prova de C. da Natureza
10	TX_RESPOSTAS_CH	Vetor com as resp. da parte objetiva da prova de C. Humanas
11	TX_RESPOSTAS_LC	Vetor com as resp. da parte objetiva da prova de Linguagens e Códigos
12	TX_RESPOSTAS_MT	Vetor com as resp. da parte objetiva da prova de Matemática
13	TX_GABARITO_CN	Vetor com o gabarito da parte objetiva da prova de C. da Natureza
14	TX_GABARITO_CH	Vetor com o gabarito da parte objetiva da prova de C. Humanas
15	TX_GABARITO_LC	Vetor com o gab. da parte objetiva da prova de Linguagens e Códigos
16	TX_GABARITO_MT	Vetor com o gab. da parte objetiva da prova de Matemática
17	TP_STATUS_REDACAO	Situação da redação do participante
18	NU_NOTA_COMP1	Nota da competência 1 - Redação
19	NU_NOTA_COMP2	Nota da competência 2 - Redação
20	NU_NOTA_COMP3	Nota da competência 3 - Redação
21	NU_NOTA_COMP4	Nota da competência 4 - Redação
22	NU_NOTA_COMP5	Nota da competência 5 - Redação

Fonte: adaptado do arquivo de metadados “Dicionário_Microdados_Enem_2023.xlsx”, (Inep, 2024)

Para os dados numéricos, procedeu-se à eliminação das observações com dados faltantes e com notas nulas, observando a ordem decrescente de ocorrência em cada atributo numérico.

Em seguida, foi adicionada uma nova variável intitulada “NU_NOTA_FINAL”, resultante da média aritmética das notas das cinco áreas de conhecimento armazenadas nas variáveis: “NU_NOTA_CN”, “NU_NOTA_CH”, “NU_NOTA_LC”, “NU_NOTA_MT”, e “NU_NOTA_REDACAO”.

Por fim, cada atributo teve suas categorias substituídas pela descrição correspondente no arquivo de metadados, para facilitar a análise exploratória. A Figura 17 apresenta uma visualização parcial do subconjunto de dados nacionais, um *dataframe* com 2.569.190 linhas (na forma transposta), ou seja, uma redução de 34,7% no número de observações:

Figura 17 – Subconjunto Nacional de Microdados do Enem 2023 (parcial)

	2	3	4	9	10
TP_FAIXA_ETARIA	21	17	18	26 a 30	23
TP_SEXO	F	F	F	M	M
TP_COR_RACA	Branca	Parda	Parda	Branca	Parda
TP_ST_CONCLUSAO	Egressos	Concluintes	Concluintes	Egressos	Egressos
SG_UF_PROVA	RS	CE	CE	SP	RN
NU_NOTA_CN	502.0	459.0	402.5	564.7	644.9
NU_NOTA_CH	498.9	508.5	379.2	630.3	620.2
NU_NOTA_LC	475.6	507.2	446.9	610.4	626.9
NU_NOTA_MT	363.2	466.7	338.3	680.2	736.3
NU_NOTA_REDACAO	700.0	880.0	560.0	600.0	860.0
Q002	Não completou E.S.	Não completou E.M.	Não completou 4º/5º Ano do E.F.	Não completou E.S.	Não completou 8º/9º Ano do E.F.
Q006	Até 1.980	Até 1.980	Até 1.320	Até 3.960	Até 1.320
Q024	Não	Não	Não	Sim, um	Sim, um
Q025	Sim	Sim	Não	Sim	Sim
NU_NOTA_FINAL	507.94	564.28	425.38	617.12	697.66

Fonte: elaboração própria.

Neste ponto, procedeu-se à Análise Exploratória Nacional, detalhada em tópico específico.

3.4.1 Seleção dos atributos

A seleção de atributos busca identificar um subconjunto de variáveis que mantenha ou melhore o desempenho preditivo de um modelo, reduzindo custos computacionais e melhorando a interpretabilidade. Essa importante etapa, pode ser feita por ordenação, rordenando os atributos preditivos e selecionando-os segundo o poder discriminativo em relação ao alvo, ou por complementaridade, quando se busca um grupo de atributos que, de forma combinada, conferem melhor poder preditivo (Carvalho; Menezes; Bonidia, 2024).

Apesar de aparentemente melhor, a seleção por complementaridade possui alto custo computacional, o que pode torná-la inviável em conjuntos de dados com muitos atributos. Cabe dizer, que a seleção pode ser dependente ou independente do algoritmo de modelagem utilizado, podendo ocorrer diferentes abordagens, como filtro, que utiliza propriedades gerais dos atributos, como medidas estatísticas para identificar e descartar atributos redundantes ou irrelevantes, ou empacotamento, que integram a seleção de atributos ao próprio processo de modelagem, avaliando diretamente a qualidade daquele subconjunto para o desempenho do modelo (Carvalho; Menezes; Bonidia, 2024).

Dentro da abordagem por empacotamento, há alguns métodos específicos, como a seleção avançada (parte de um conjunto vazio adicionando iterativamente o melhor dentre os atributos restantes), eliminação reversa (remove iterativamente o pior recurso presente em

todo o conjunto), seleção progressiva e eliminação regressiva (combina a seleção do melhor e a remoção do pior atributo simultaneamente) e a eliminação recursiva de atributos (Sayak, 2020).

A Eliminação Recursiva de Atributos, do inglês *Recursive Feature Elimination* (RFE), é uma técnica gulosa que busca identificar o subconjunto de atributos mais relevante, partindo de todo o conjunto e removendo, em cada repetição, aqueles menos importantes, até obter a quantidade de atributos requerida. Ao avaliar múltiplas combinações, a RFE pode resultar em alto custo computacional para grandes conjuntos de dados (Sayak, 2020).

No presente estudo adota-se a RFE combinada à modelagem com o algoritmo *Random Forest* (Marteletto, 2022), um poderoso recurso para lidar com grandes bases de dados pela sua capacidade de avaliar a importância de cada variável para a construção do modelo. Esse algoritmo apoia-se em um conjunto de árvores de decisão treinadas com diferentes amostras do conjunto de dados, que considera a classificação final da variável alvo como a combinação das diferentes árvores individuais, melhorando assim a qualidade do poder preditivo (Carvalho; Menezes; Bonidia, 2024).

A seleção dos atributos contemplou:

1) Remoção manual de atributos inconsistentes ou desnecessários para o estudo (Faceli *et al.*, 2024), que incluem as variáveis “TP_ESCOLA”, as listadas nas Tabelas 7 e 8, bem como, os dados numéricos pertinentes às variáveis das notas por área – “NU_NOTA_CN”, “NU_NOTA_CH”, “NU_NOTA_LC”, “NU_NOTA_MT”, “NU_NOTA_REDACAO” e “NU_NOTA_FINAL” – em virtude da categorização desta última na variável “NOTA”.

2) Aplicação do algoritmo RFE para seleção dos 10 (dez) atributos dentre os 33 (trinta e três) restantes, com o algoritmo *Random Forest*, mediante a eliminação de 1 (um) atributo a cada repetição, tendo a variável “NOTA” como alvo, conforme código declarado na Figura 18:

Figura 18 – Execução da RFE para seleção dos atributos

```
# Seleção Manual dos atributos relevantes, eliminando colunas redundantes e com dados faltantes.
subset_df = df[['TP_FAIXA_ETARIA', 'TP_SEXO', 'TP_ESTADO_CIVIL', 'TP_COR_RACA', 'TP_ST_CONCLUSAO', 'TP_ANO_CONCLUSAO', 'NO_MUNICIPIO_PROVA',
                'Q001', 'Q002', 'Q003', 'Q004', 'Q005', 'Q006', 'Q007', 'Q008', 'Q009', 'Q010', 'Q011', 'Q012', 'Q013', 'Q014', 'Q015',
                'Q016', 'Q017', 'Q018', 'Q019', 'Q020', 'Q021', 'Q022', 'Q023', 'Q024', 'Q025', 'NOTA']]

/usr/local/lib/python3.10/dist-packages/ipykernel/ipkernel.py:283: DeprecationWarning: `should_run_async` will not call `transform_cell`
and should_run_async(code)

X = subset_df.drop(columns=['NOTA'])
y = subset_df['NOTA']

/usr/local/lib/python3.10/dist-packages/ipykernel/ipkernel.py:283: DeprecationWarning: `should_run_async` will not call `transform_cell`
and should_run_async(code)

# OneHotEncoding das variáveis categóricas
X = pd.get_dummies(X)

/usr/local/lib/python3.10/dist-packages/ipykernel/ipkernel.py:283: DeprecationWarning: `should_run_async` will not call `transform_cell`
and should_run_async(code)

# Seleção de 10 atributos
n_features = 10
clf = RandomForestClassifier()
rfe = RFE(clf, n_features_to_select=n_features, step=1)
# Ajuste do modelo para os objetos e seleção dos atributos
rfe = rfe.fit(X, y)
#columns_retained_RFE = X.columns[trans.get_support()].values
```

Fonte: adaptado de Carvalho; Menezes; Bonidia, 2024.

3) A aplicação retornou as seguintes variáveis, conforme Figura 19:

Figura 19 – Atributos selecionados com a RFE

```
selected_features = X.columns[rfe.support_]
print(f"Atributos selecionados: {selected_features}")

Atributos selecionados: Index(['TP_FAIXA_ETARIA_18', 'TP_SEXO_M', 'TP_ST_CONCLUSAO_Concluintes',
                              'Q001_Não completou E.M.', 'Q003_Grupo 2', 'Q005_3', 'Q009_Sim, 2',
                              'Q011_Sim, 1', 'Q014_Sim, 1', 'Q024_Não'],
                              dtype='object')
```

Fonte: elaboração própria.

A Tabela 9 detalha as variáveis selecionadas com a RFE:

Tabela 9 – Variáveis selecionadas com a RFE

N	Variável	Descrição	Categorias
1	TP_FAIXA_ETARIA	Faixa etária	20 categorias com valores numéricos inteiros de 1 a 20.
2	TP_SEXO	Sexo	2 categorias com valores alfanuméricos 'M' ou 'F'.
3	TP_ST_CONCLUSAO	Situação de conclusão do Ensino Médio	4 categorias com valores numéricos inteiros de 1 a 4.
4	Q001	Nível de escolaridade do pai ou o homem responsável pelo participante.	5 categorias com valores alfanuméricos de 'A' a 'E'.
5	Q003	Grupo de natureza da ocupação do pai ou homem responsável pelo participante.	6 categorias com valores alfanuméricos de 'A' a 'F'.
6	Q005	Quantidade de moradores na residência (incluindo o participante).	20 categorias com valores numéricos inteiros de 1 a 20.
7	Q009	Quantidade de quartos para dormir.	5 categorias com valores alfanuméricos de 'A' a 'E'.
8	Q011	Quantidade de motocicletas.	5 categorias com valores alfanuméricos de 'A' a 'E'.
9	Q014	Quantidade de máquinas de lavar roupa.	5 categorias com valores alfanuméricos de 'A' a 'E'.
10	Q024	Quantidade de computadores.	5 categorias com valores alfanuméricos de 'A' a 'E'.

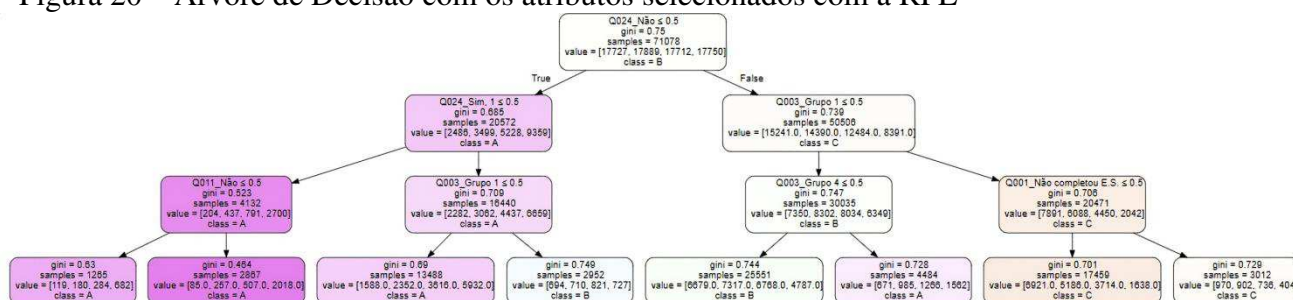
Fonte: adaptado do arquivo de metadados "Dicionário_Microdados_Enem_2023.xlsx", (Inep, 2024).

A seleção de 10 (dez) atributos na presente etapa justifica-se pela redução do custo computacional e interpretabilidade dos resultados, considerando a quantidade de variáveis estudadas nos trabalhos correlatos de extração das regras de associação com o algoritmo Apriori usando a base do Enem, estudos esses, já elencados na Tabela 2.

Após a extração das variáveis, adotou-se a “Árvore de Decisão” com o intuito de exibir os relacionamentos entre os atributos com a variável de desempenho. A “Árvore de Decisão” refere-se à estrutura em forma de árvore invertida dos modelos gerados, onde cada nó corresponde a um atributo preditivo do conjunto de dados e cada caminho da raiz até um nó folha, define uma regra que estabelece o rótulo ou classe daquela observação (Carvalho; Menezes; Bonidia, 2024).

A Figura 20 ilustra a árvore de decisão com os atributos selecionados até o terceiro nível. O nó raiz consiste na variável “Q024” (quantidade de computadores), cuja ramificação no primeiro nível, alcança a variável “Q003” (natureza da ocupação do pai) à direita, se a condição não é atendida, ou a própria variável “Q024” quando se atender. No segundo nível, por sua vez, o lado esquerdo ramifica-se nas variáveis “Q011” (quantidade de motocicletas) e “Q003”, enquanto o lado direito subdivide-se em “Q003” e “Q001” (nível de escolaridade do pai). Esses caminhos levam aos nós folhas na base da árvore e suas respectivas classes. Esse recurso facilita a interpretação e a identificação dos relacionamentos e das regras para decidir a que classe pertence determinado objeto (Carvalho; Menezes; Bonidia, 2024).

Figura 20 – Árvore de Decisão com os atributos selecionados com a RFE



Fonte: elaboração própria.

A ilustração acima revela a importância de variáveis como posse do computador (Elias de Sousa *et al.*, 2021; Soares *et al.*, 2023; Silva *et al.*, 2020), escolaridade dos pais (Dutra *et al.*, 2023; Elias de Sousa *et al.*, 2021; Soares *et al.*, 2023; Silva, Morino e Sato 2014; Silva *et al.*, 2020) e a natureza da ocupação do pai para o desempenho dos participantes maranhenses.

4 ANÁLISE DOS RESULTADOS

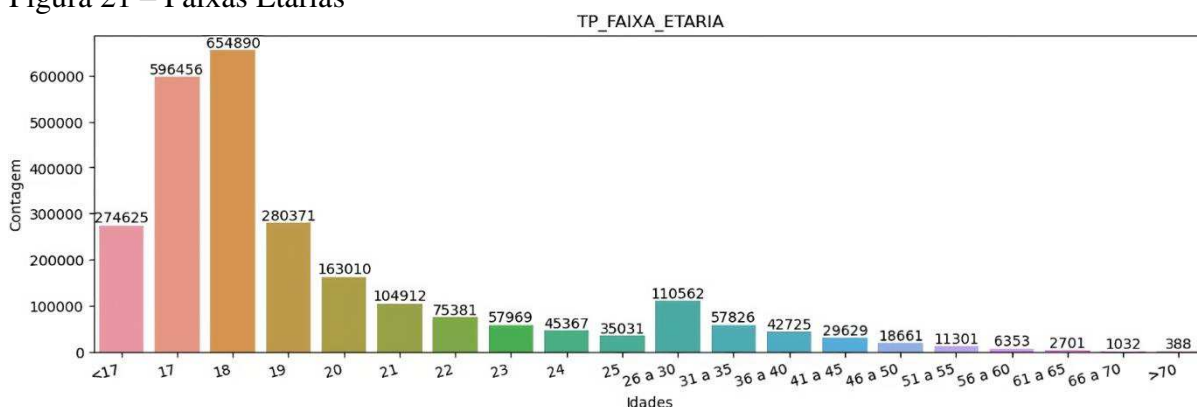
4.1 Análise Exploratória

O estudo de estatística tem dois ramos principais: estatística descritiva e estatística inferencial. A estatística descritiva envolve a análise exploratória de dados, no tocante à sua redução, análise e interpretação, a ser utilizada numa fase posterior, a análise confirmatória de dados (ou inferência estatística), a qual tem por objetivo fazer previsões para uma população, a partir dos dados de uma determinada amostra (Morettin; Bussab, 2023).

4.1.1 Visão geral dos participantes

Na Figura 21, inicia-se a análise exploratória do Subconjunto Nacional de Microdados do Enem 2023, apresentando o perfil etário dos participantes que, em sua maioria, é formado por jovens e adultos com até 19 anos de idade, considerando a idade do inscrito em 31.12.2023, conforme arquivo de metadados.

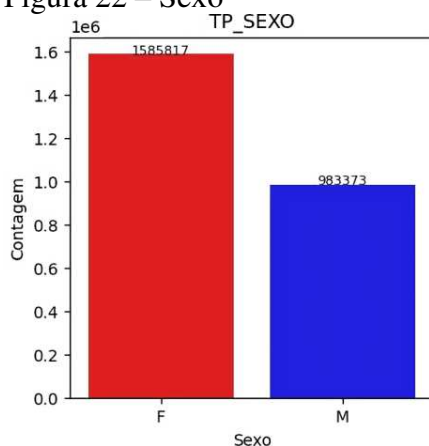
Figura 21 – Faixas Etárias



Fonte: elaboração própria.

A Figura 22 demonstra que a maioria dos participantes é do sexo feminino, em torno de 62%.

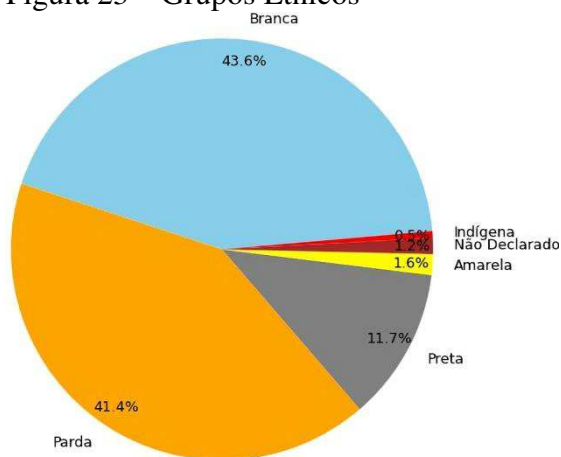
Figura 22 – Sexo



Fonte: elaboração própria.

Apesar da predominância de pessoas brancas (43,6%), o grupo de negros, evidenciado pelas categorias de pretos (11,7%) e pardos (41,4%), representa a maioria étnica de participantes do Enem, conforme ilustração na Figura 23.

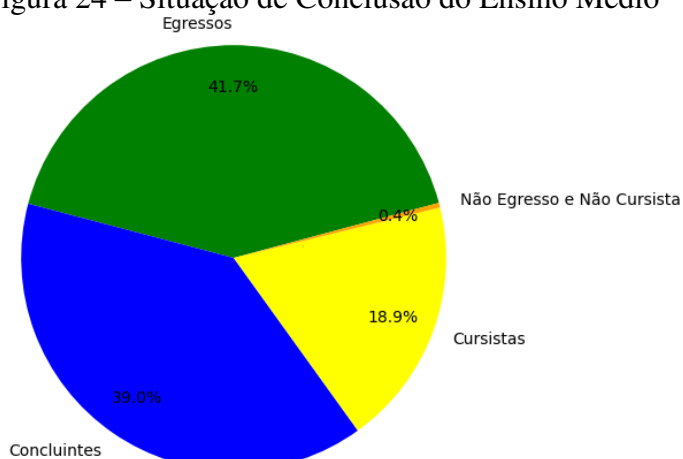
Figura 23 – Grupos Étnicos



Fonte: elaboração própria.

Também é relevante caracterizar os participantes segundo a situação de conclusão do ensino médio. Os participantes no último ano do ensino médio, integrantes da categoria de Concluintes, representam 39%, enquanto os Egressos representam a maioria, com 41,7%. Já os participantes que não concluiriam o ensino médio em 2023, integrantes da categoria de cursistas, representaram 18,9%, como resume o gráfico da Figura 24. Cabe dizer que esta última categoria, consiste no perfil de treineiro, isto é, daqueles candidatos que realizam o exame com o fim específico de testar seus conhecimento e preparar-se para as edições futuras (Inep, 2024).

Figura 24 – Situação de Conclusão do Ensino Médio

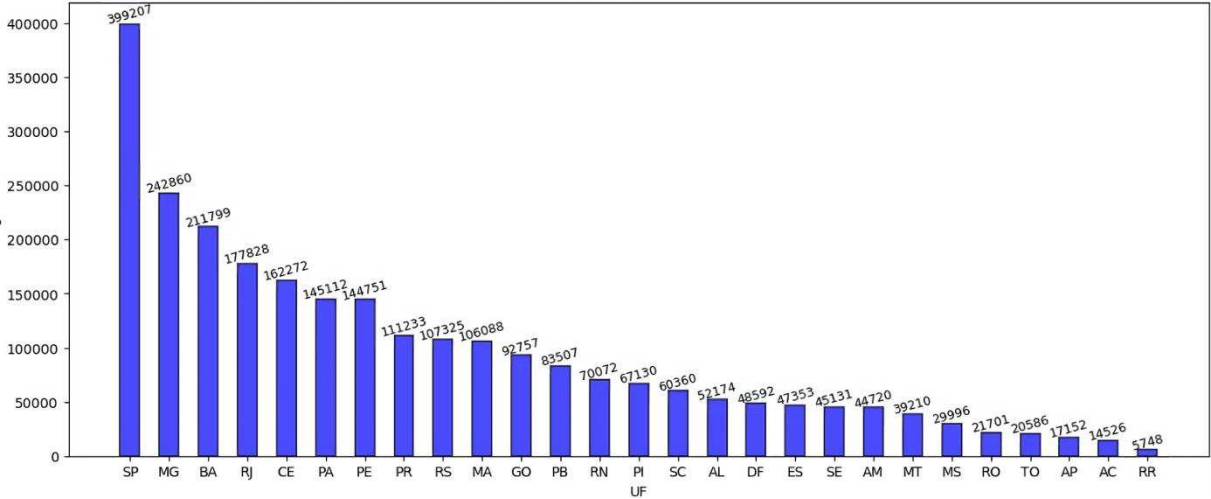


Fonte: elaboração própria.

Os microdados do Enem permitem evidenciar o índice de participação do Enem por Unidade Federativa (UF), demonstrada na Figura 25, com destaque para a região Sudeste, liderada pelos estados de São Paulo e Minas Gerais, seguidos pelo estado da Bahia. O

Maranhão aparece na 10ª colocação nacional e 4º dentre os estados nordestinos, considerando os participantes que não faltaram em nenhum dos dias de aplicação e não tiveram notas 0 (zero) nas provas.

Figura 25 – Participantes por UF



Fonte: elaboração própria.

Foram analisados quatro aspectos pertinentes ao questionário socioeconômico: escolaridade das mães dos participantes (Q002), renda familiar mensal (Q006), posse de computador na residência (Q024) e o acesso residencial à internet (Q025).

A variável “Q002” buscou evidenciar até que série a mãe, ou a mulher responsável pelo participante, estudou, estruturando as respostas nas categorias elencadas na Tabela 10:

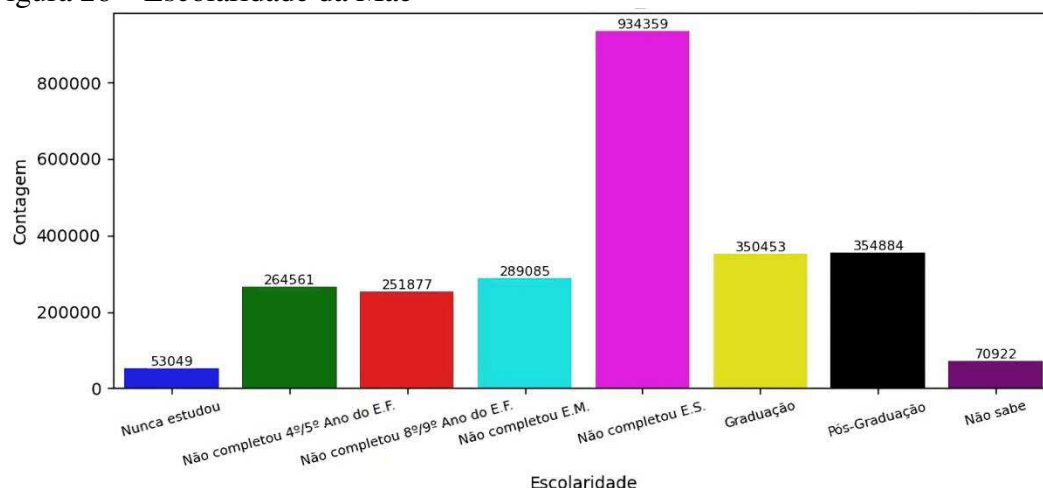
Tabela 10 – Categorias da variável “Q002”

Categoria	Descrição	Representatividade
A	Nunca estudou.	2,1%
B	Não completou a 4ª série/5º ano do Ensino Fundamental.	10,3%
C	Completou a 4ª série/5º ano, mas não completou a 8ª série/9º ano do Ensino Fundamental.	9,8%
D	Completou a 8ª série/9º ano do Ensino Fundamental, mas não completou o Ensino Médio.	11,2%
E	Completou o Ensino Médio, mas não completou a Faculdade.	36,4%
F	Completou a Faculdade, mas não completou a Pós-graduação.	13,6%
G	Completou a Pós-graduação.	13,8%
H	Não sei.	2,8%

Fonte: adaptado do arquivo de metadados “Dicionário_Microdados_Enem_2023.xlsx”, (Inep, 2024)

A Figura 26 evidencia que a categoria de maior ocorrência é a de mulheres que possuem o ensino médio completo mas não concluíram a faculdade (36,4%). Por outro lado, também é possível constatar nas quatro primeiras categorias, um número considerável de mulheres que não concluíram a educação básica.

Figura 26 – Escolaridade da Mãe



Fonte: elaboração própria.

Já a variável ‘Q006’ buscou evidenciar a renda familiar, estruturando as respostas em categorias (aglutinadas para melhor visualização) elencadas na Tabela 11:

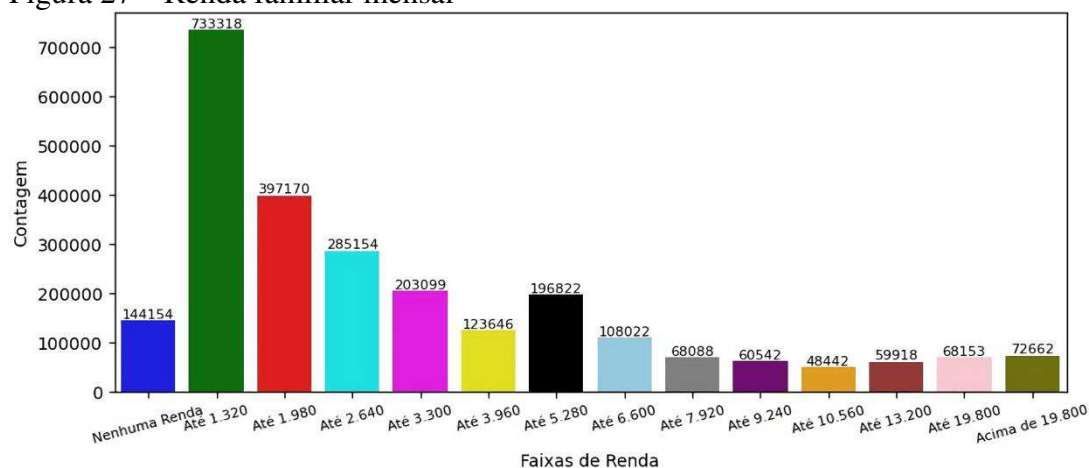
Tabela 11 – Categorias da variável “Q006”

Categoria	Descrição	Representatividade
A	Nenhuma Renda	5,6%
B	Até R\$ 1.320,00	28,5%
C	De R\$ 1.320,01 até R\$ 1.980,00.	15,5%
D	De R\$ 1.980,01 até R\$ 2.640,00.	11,1%
E	De R\$ 2.640,01 até R\$ 3.300,00.	7,9%
F	De R\$ 3.300,01 até R\$ 3.960,00.	4,8%
G	De R\$ 3.960,01 até R\$ 5.280,00.	7,7%
H	De R\$ 5.280,01 até R\$ 6.600,00.	4,2%
I	De R\$ 6.600,01 até R\$ 7.920,00.	2,7%
J	De R\$ 7.920,01 até R\$ 9.240,00.	2,4%
K	De R\$ 9.240,01 até R\$ 10.560,00.	1,9%
L	De R\$ 10.560,01 até R\$ 13.200,00.	2,3%
M	De R\$ 13.200,01 até R\$19.800,00.	2,7%
N	Acima de R\$ 19.800,00	2,8%

Fonte: adaptado do arquivo de metadados “Dicionário_Microdados_Enem_2023.xlsx”, (Inep, 2024)

A Figura 27 indica que a categoria de maior frequência é a de famílias vivendo com até um salário mínimo (valor de referência vigente em 2023).

Figura 27 – Renda familiar mensal



Fonte: elaboração própria.

O registro da quantidade de computadores na residência dos participantes, através da variável “Q024”, permitiu constatar que a maioria dos participantes não possui o dispositivo (47,2%), conforme ilustração da Figura 28.

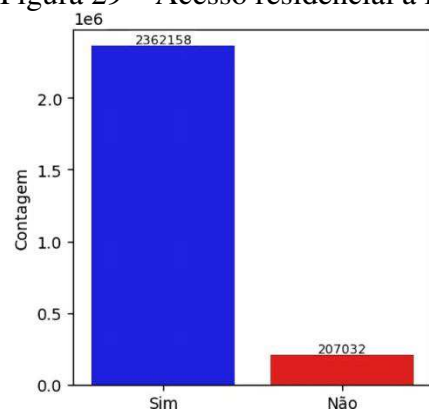
Figura 28 – Computador na residência



Fonte: elaboração própria.

O acesso à internet foi outro elemento pesquisado no questionário. Na Figura 29, tem-se o resultado da análise descritiva da variável “Q025”, que apontou o acesso à internet na residência pela grande maioria dos participantes do Enem 2023.

Figura 29 – Acesso residencial à internet



Fonte: elaboração própria.

4.1.2 Visão geral do desempenho

A avaliação dos resultados do exame é um ponto sensível, utilizado para fins de análises e comparações, não apenas de forma univariada, mas também, verificando o comportamento das notas relacionado a outras variáveis de interesse.

Na Figura 30, o resumo estatístico dos atributos numéricos destaca algumas informações importantes, como a média nacional do Enem 2023, qualificada pela exclusão dos faltosos e notas nulas, atingindo 548,85 pontos, com um desvio padrão de 87,4 pontos.

As notas da Redação possuem a maior amplitude dentre todas as áreas, isto é, distância entre os valores mínimo e máximo (600 pontos) e a maior variabilidade, com um desvio padrão de 174,9 pontos. Além disso, os maiores e menores desempenhos medianos ocorreram nas áreas de Redação (640 pontos) e Ciências da Natureza (497 pontos), respectivamente.

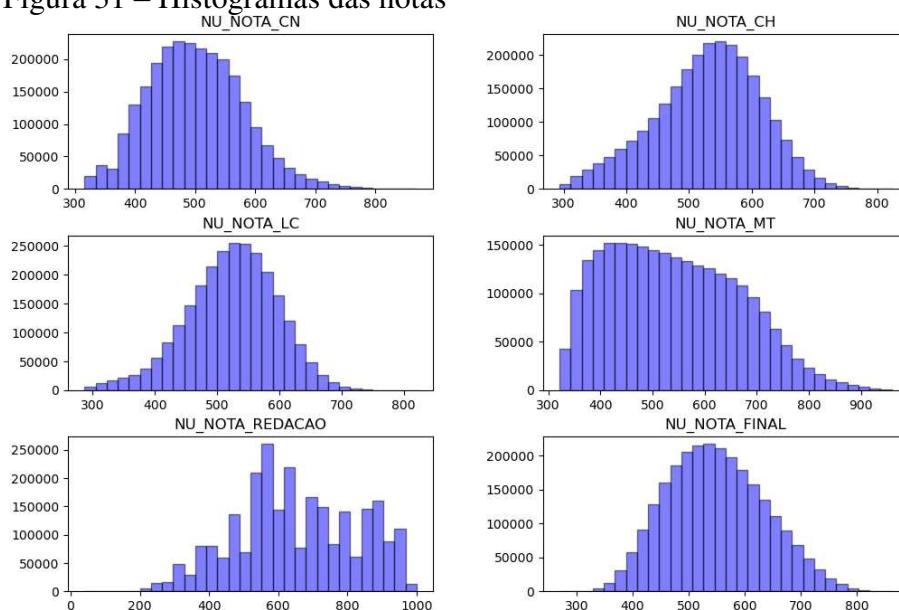
Figura 30 – Resumo estatístico dos atributos numéricos

	NU_NOTA_CN	NU_NOTA_CH	NU_NOTA_LC	NU_NOTA_MT	NU_NOTA_REDACAO	NU_NOTA_FINAL
count	2.569190e+06	2.569190e+06	2.569190e+06	2.569190e+06	2.569190e+06	2.569190e+06
mean	5.009274e+02	5.297365e+02	5.231669e+02	5.405983e+02	6.498059e+02	5.488470e+02
std	7.876968e+01	8.392717e+01	7.225752e+01	1.249133e+02	1.748823e+02	8.737762e+01
min	3.161000e+02	2.932000e+02	2.870000e+02	3.213000e+02	4.000000e+01	2.691600e+02
25%	4.439000e+02	4.757000e+02	4.772000e+02	4.369000e+02	5.200000e+02	4.839600e+02
50%	4.968000e+02	5.357000e+02	5.273000e+02	5.294000e+02	6.400000e+02	5.441400e+02
75%	5.533000e+02	5.889000e+02	5.735000e+02	6.342000e+02	8.000000e+02	6.095600e+02
max	8.684000e+02	8.230000e+02	8.208000e+02	9.586000e+02	1.000000e+03	8.625800e+02

Fonte: elaboração própria.

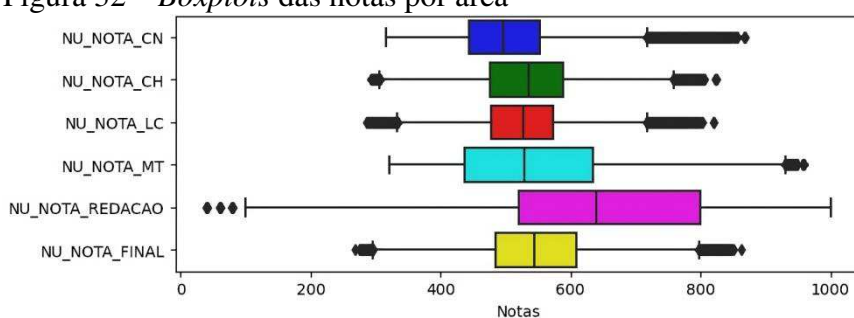
A Figura 31, por sua vez, apresenta os histogramas das notas por área e nota final. Os histogramas constituem um tipo de gráfico construído a partir da distribuição de frequências, composto de retângulos contíguos que sintetizam as frequências das respectivas classes ou intervalos de classes (Morettin; Singer, 2022). Os referidos gráficos apontam certo nível de assimetria das notas nas cinco áreas da prova. Na área de Matemática e suas Tecnologias é mais latente a concentração de notas à esquerda, isto é, abaixo da média, apontando uma assimetria positiva (Mattos; Konrath; Azambuja, 2017).

Figura 31 – Histogramas das notas



Fonte: elaboração própria.

A Figura 32 traz outra forma de representação para os mesmos atributos numéricos, os diagramas de caixa ou *boxplot*, para resumir a distribuição dos dados baseado nas medidas de posição chamadas de quartis, permitindo a identificação da posição mediana, entre o primeiro e terceiro quartis, bem como, a ocorrência de valores discrepantes (Morettin; Singer, 2022). Na citada figura, pode-se ratificar a área de maior amplitude e variabilidade das notas (Redação), bem como, evidenciar a ocorrência de valores atípicos, também chamados de *outliers*, em todas as áreas. Além disso, a característica de assimetria das distribuições também pode ser observada.

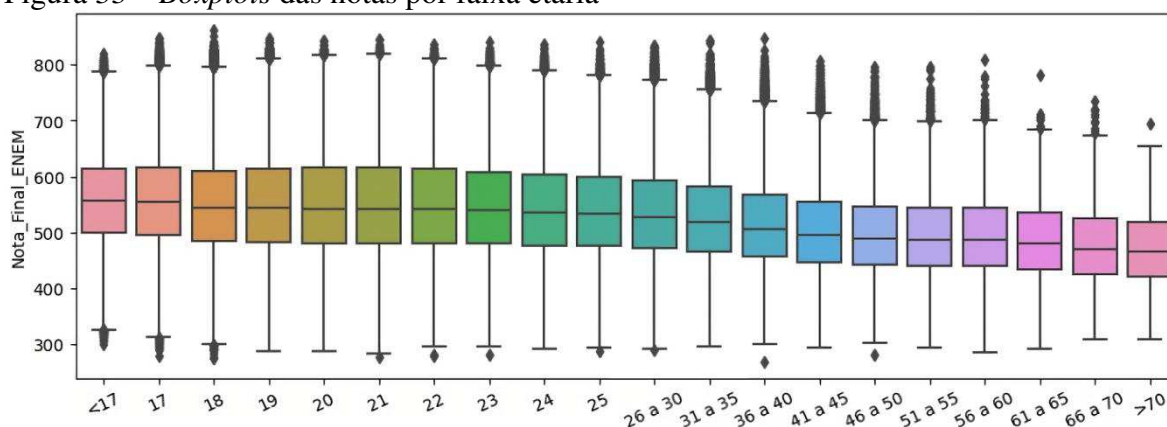
Figura 32 – *Boxplots* das notas por área

Fonte: elaboração própria.

Neste ponto da análise exploratória, constata-se o comportamento das Notas Finais (média dentre as cinco áreas) de acordo com a caracterização do participante (primeiro grupo de variáveis da base).

Na figura 33, os *boxplots* das notas por faixa etária sugerem um comportamento de queda de desempenho à medida que a idade avança.

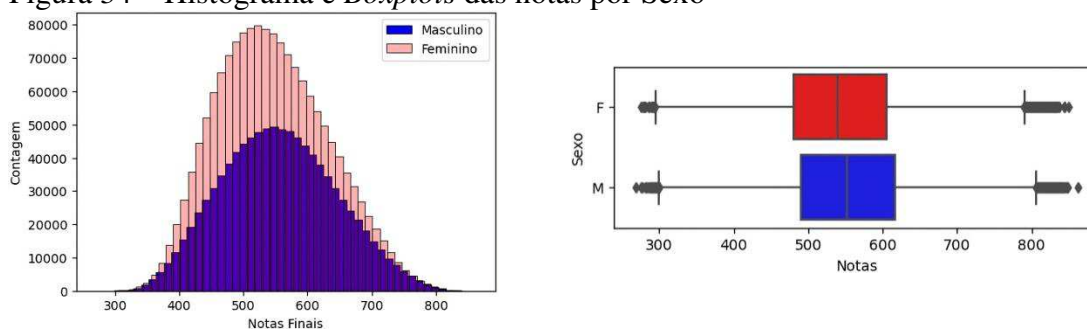
Figura 33 – *Boxplots* das notas por faixa etária



Fonte: elaboração própria.

Já na figura 34, os histogramas e *boxplots* das notas por sexo apontam que, apesar da maioria feminina, os melhores desempenhos são de participantes do sexo masculino.

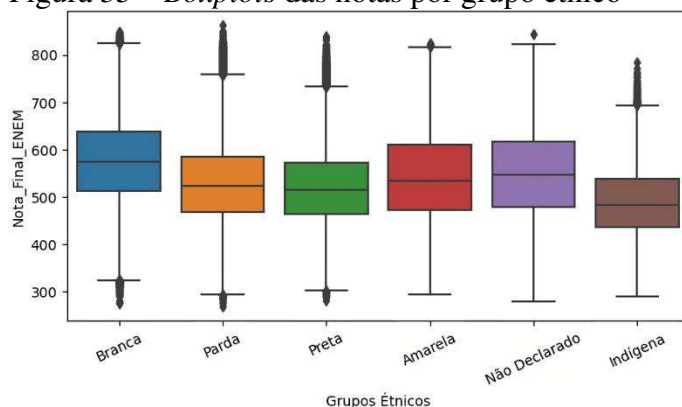
Figura 34 – Histograma e *Boxplots* das notas por Sexo



Fonte: elaboração própria.

Na Figura 35, a representação dos diagramas de caixa por grupo étnico aponta que as notas dos pretos e pardos estão distribuídas numa região de valores mais reduzidos da amostra, isto é, intervalo de menor desempenho, seguido pelo grupo de indígenas.

Figura 35 – *Boxplots* das notas por grupo étnico

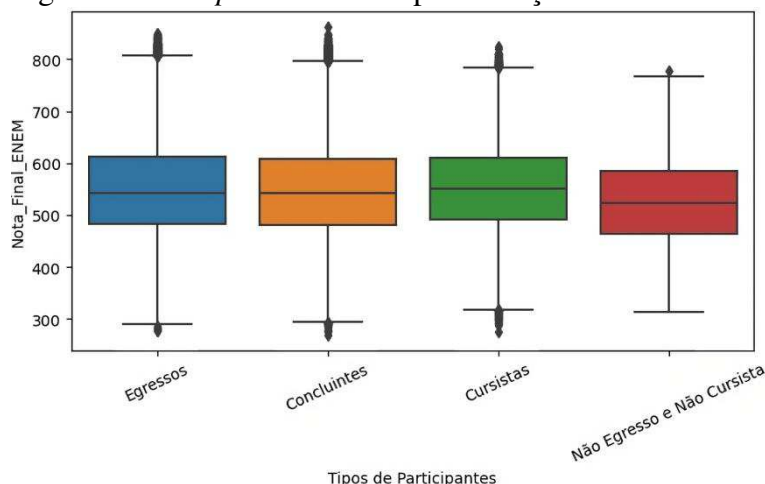


Fonte: elaboração própria.

Na Figura 36, os *boxplots* das notas por situação de conclusão do ensino médio evidenciam uma queda gradativa de rendimento daqueles que estão no último ano do ensino médio (Concluintes), seguida pelos participantes Cursistas do primeiro ou segundo ano

(também chamados de treineiros), em relação ao grupo de Egressos. Observa-se que as notas passam a ocupar intervalos de menores valores dentro da amostra.

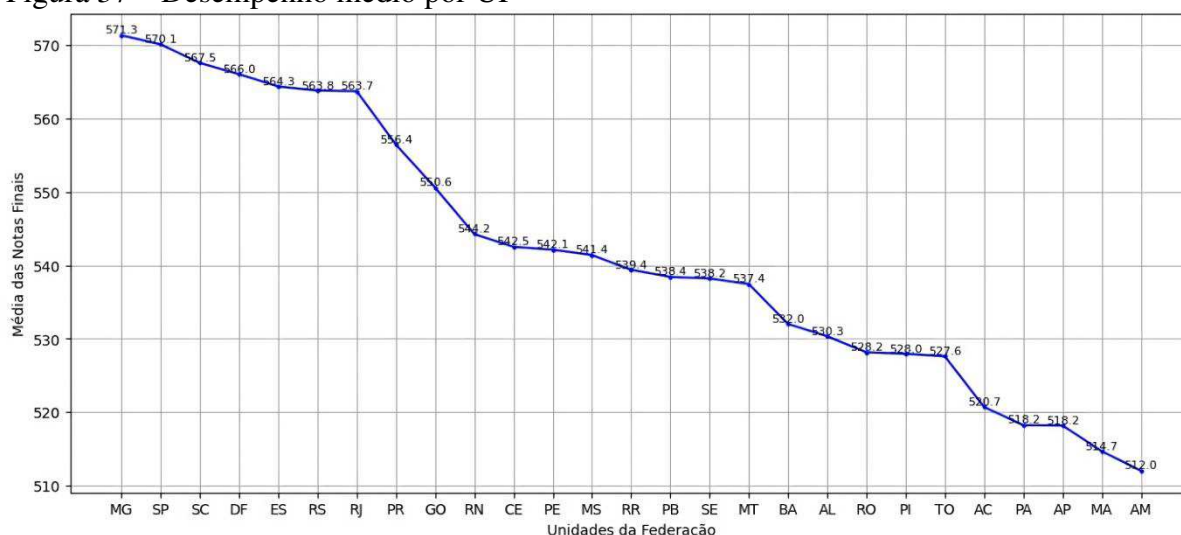
Figura 36 – *Boxplots* das notas por Situação de Conclusão do Ensino Médio



Fonte: elaboração própria.

O desempenho dos participantes do Enem 2023 por localidade também é um achado relevante para direcionar esforços para melhoria do ensino e do impulso socioeconômico. Adotando a média “qualificada” dos participantes como índice de ranqueamento, a Figura 37 aponta que os estados das regiões Sudeste e Sul dominam as dez primeiras posições. Por outro lado, o Estado do Maranhão, 10º colocado em quantidade de participantes, aparece na penúltima colocação nacional e último, dentre os estados nordestinos.

Figura 37 – Desempenho médio por UF



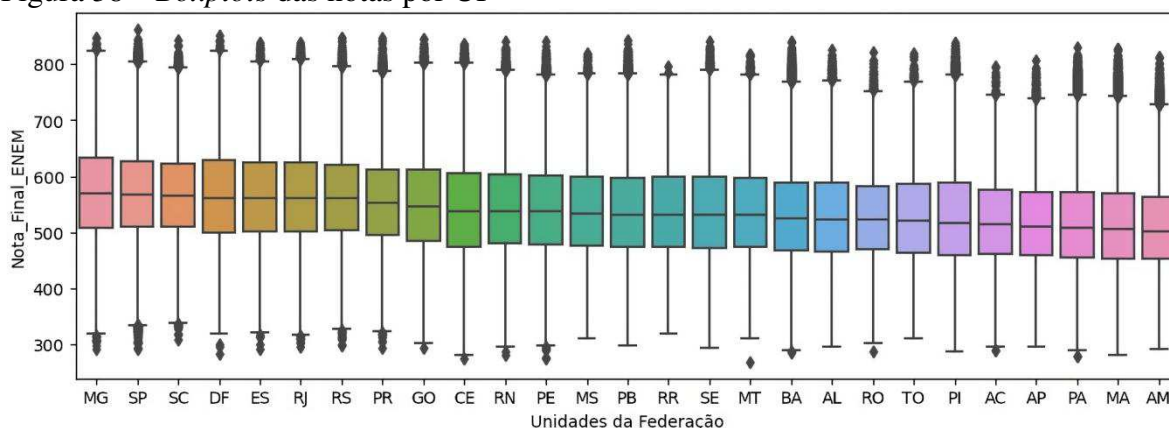
Fonte: elaboração própria.

O presente achado destaca o Maranhão como objeto de estudo, a fim de compreender como as características socioeconômicas estruturadas e divulgadas pelo Inep,

por meio dos microdados do Enem, influenciam o desempenho desses participantes, sejam eles, egressos, concluintes ou cursistas do Ensino Médio.

Na Figura 38, os *boxplots* destacam os intervalos de distribuição das notas por UF, evidenciando que as características do desempenho maranhense possuem maior similaridade aos estados da região Norte (Acre, Amapá, Pará e Amazonas).

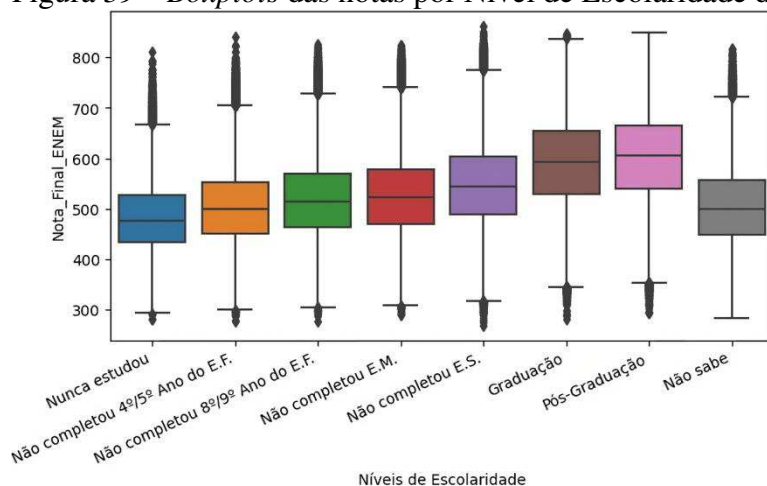
Figura 38 – *Boxplots* das notas por UF



Fonte: elaboração própria.

Na Figura 39, os *boxplots* das notas por nível de escolaridade da mãe, ou da mulher responsável pelo participante, demonstra o incremento do desempenho associado ao avanço na escolaridade dessas mulheres.

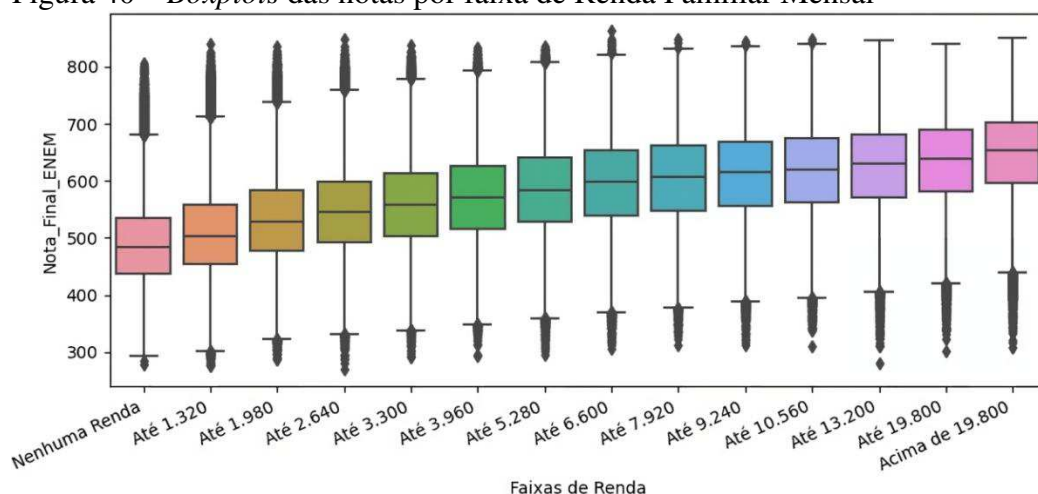
Figura 39 – *Boxplots* das notas por Nível de Escolaridade da Mãe



Fonte: elaboração própria.

De acordo com a Figura 40, a renda familiar mensal emergiu como um indicador relevante, sugerindo que estudantes com condições financeiras mais favoráveis tendem a apresentar um desempenho superior no Enem.

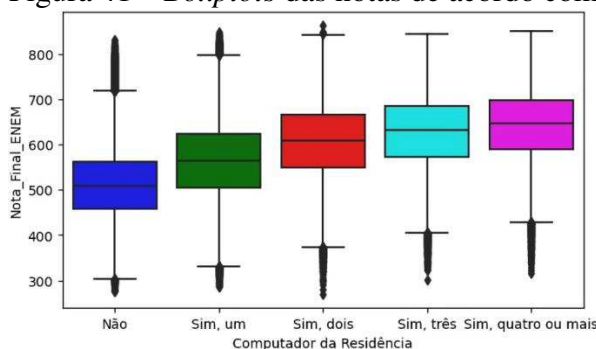
Figura 40 – *Boxplots* das notas por faixa de Renda Familiar Mensal



Fonte: elaboração própria.

A ausência de computador na residência, fato evidenciado pela maioria dos participantes (47,2%), sugere prejuízos ao desempenho. A Figura 41 destaca a disparidade existente entre o desempenho dos grupos que possuem pelo menos um computador na residência daqueles que não possuem.

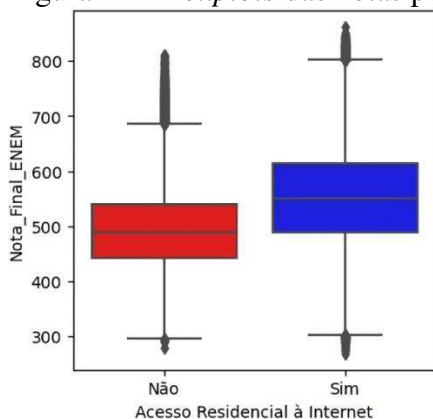
Figura 41 – *Boxplots* das notas de acordo com a presença de Computador na Residência



Fonte: elaboração própria.

O comportamento do desempenho de acordo com o nível de acesso à internet no domicílio, evidenciado na Figura 42, mostrou uma correlação positiva, o que sugere a importância da conectividade para a preparação dos estudantes.

Figura 42 – *Boxplots* das notas por nível de acesso residencial à internet



Fonte: elaboração própria.

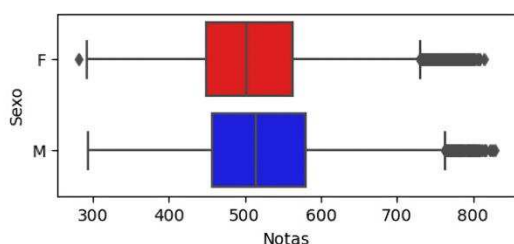
4.1.3 Visão Geral do Maranhão

No contexto maranhense, a análise exploratória estadual apontou que a maioria dos participantes possui entre 17 e 19 anos de idade (57%), sendo em torno de 63% do sexo feminino. No tocante à caracterização étnica, a maioria é de negros, evidenciada pelas categorias de pretos (13,8%) e pardos (61,3%), com 22,3% de participantes brancos, destoando da predominância nacional.

Os participantes concluintes representaram 33,5%, enquanto os egressos foram maioria, com 47,3%. Já os treineiros, isto é, os cursistas do ensino médio, representaram 18,9%.

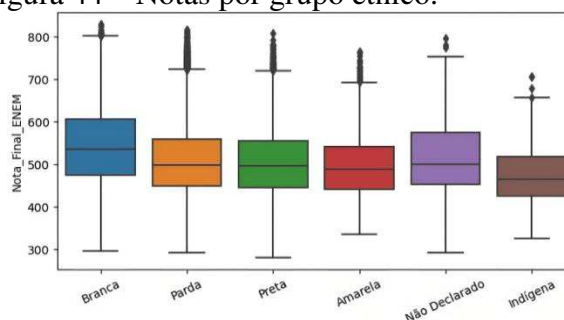
Ao relacionar o comportamento do desempenho de acordo as principais variáveis, como sexo na Figura 43 ou etnia na Figura 44, observa-se que os participantes do sexo masculino ou brancos possuem melhores resultados.

Figura 43 – Notas por sexo.



Fonte: elaboração própria.

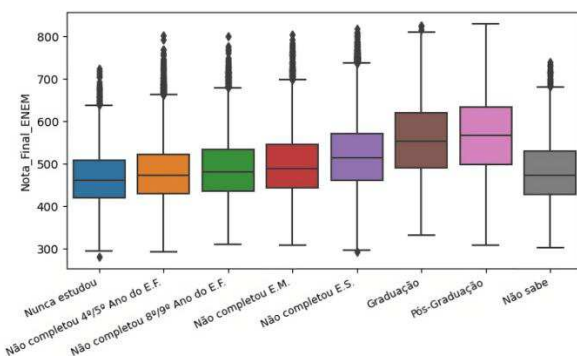
Figura 44 – Notas por grupo étnico.



A análise exploratória ainda evidenciou que a maioria não possui computador na residência (71%), embora número significativo possua acesso à internet no domicílio (81,6%).

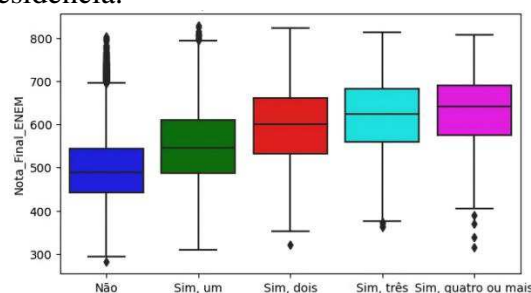
Do mesmo modo, a escolaridade das mães e a posse de computador apontou correlação positiva com o incremento das notas, conforme ilustrações das Figuras 45 e 46.

Figura 45 – Notas por escolaridade das mães.



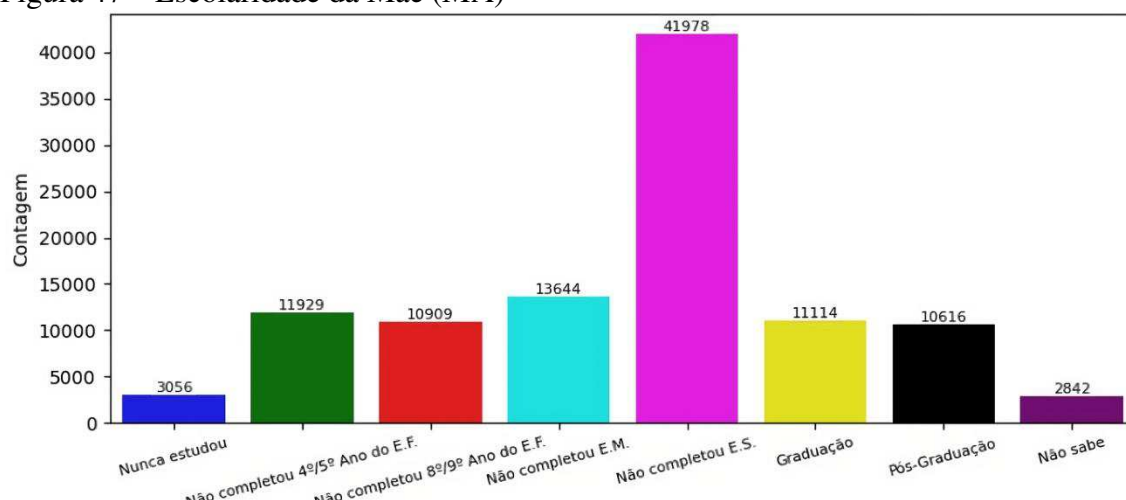
Fonte: elaboração própria.

Figura 46 – Notas por computador na residência.



Em relação à escolaridade das mães, observou-se que a maioria não completou a educação superior (39,6%), bem como, número significativo não havia concluído a educação básica (37,3%), como ilustrado na Figura 47.

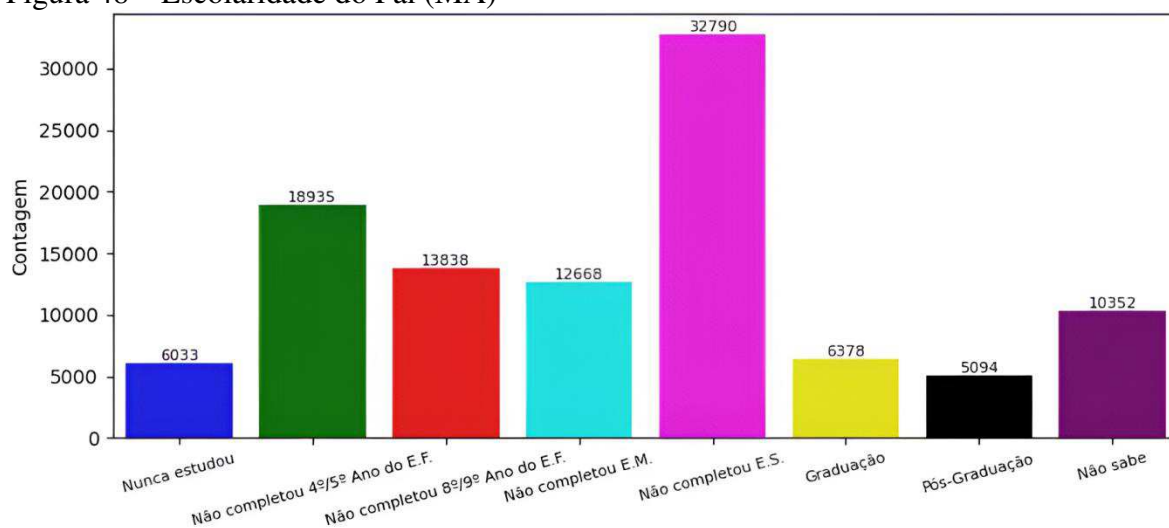
Figura 47 – Escolaridade da Mãe (MA)



Fonte: elaboração própria.

Já escolaridade do pai apresenta-se concentrada nos menores níveis, quando comparada à escolaridade das mães, isto é, as mulheres maranhenses apresentaram-se melhor qualificadas, já que 48,52% daqueles não completaram a educação básica e 31% não completou a educação superior, como se verifica na Figura 48.

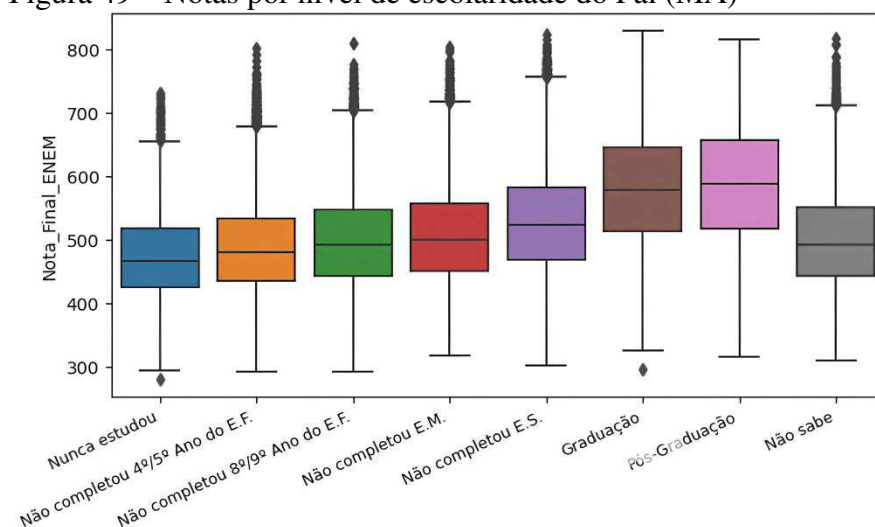
Figura 48 – Escolaridade do Pai (MA)



Fonte: elaboração própria.

A escolaridade do pai ou homem responsável pelo participante apresentou uma associação positiva em relação ao desempenho, como se verifica na Figura 49.

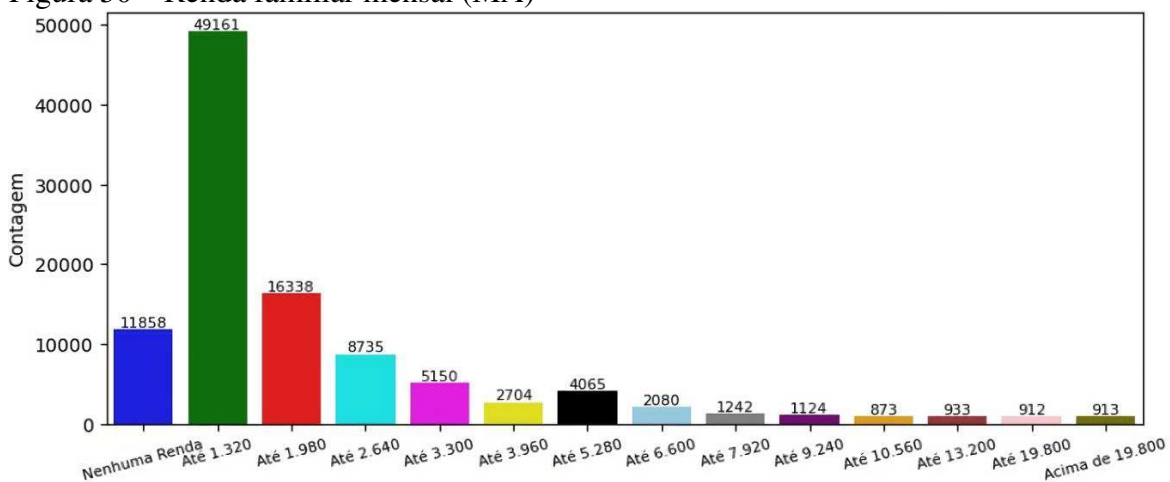
Figura 49 – Notas por nível de escolaridade do Pai (MA)



Fonte: elaboração própria.

A renda familiar mensal está concentrada nas menores faixas: nenhuma renda (11,2%), acima de R\$ 0,00 até R\$ 1.320,00 (46,3%) e acima de R\$ 1.320,00 até R\$ 1.980,00 (15,4%), conforme comportamento demonstrado na Figura 50.

Figura 50 – Renda familiar mensal (MA)



Fonte: elaboração própria.

A natureza da ocupação do pai ou homem responsável pelo participante apareceu dentre as variáveis selecionadas pela RFE, sendo categorizada pelo Dicionário de Microdados 2023 na Tabela 12. A análise descritiva aponta que a predominância da ocupação dos pais está relacionada às atividades campesinas de agricultura, pecuária e pesca.

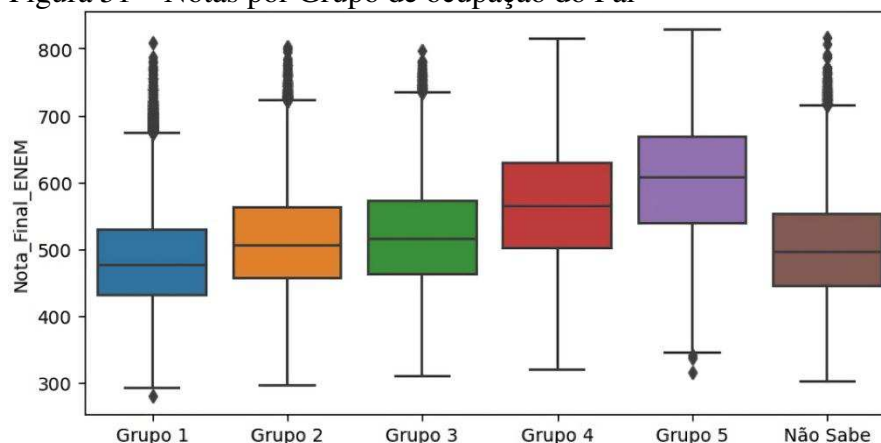
Tabela 12 – Categorias da variável “Q003”

Categoria	Descrição	Qtde	Percentual
Grupo 1	Lavrador, agricultor sem empregados, bóia fria, criador de animais (gado, porcos, galinhas, ovelhas, cavalos etc.), apicultor, pescador, lenhador, seringueiro, extrativista.	35.215	33,2
Grupo 2	Diarista, empregado doméstico, cuidador de idosos, babá, cozinheiro (em casas particulares), motorista particular, jardineiro, faxineiro de empresas e prédios, vigilante, porteiro, carteiro, office-boy, vendedor, caixa, atendente de loja, auxiliar administrativo, recepcionista, servente de pedreiro, repositor de mercadoria.	20.190	19,0
Grupo 3	Padeiro, cozinheiro industrial ou em restaurantes, sapateiro, costureiro, joalheiro, torneiro mecânico, operador de máquinas, soldador, operário de fábrica, trabalhador da mineração, pedreiro, pintor, eletricista, encanador, motorista, caminhoneiro, taxista.	18.848	17,8
Grupo 4	Professor (de ensino fundamental ou médio, idioma, música, artes etc.), técnico (de enfermagem, contabilidade, eletrônica etc.), policial, militar de baixa patente (soldado, cabo, sargento), corretor de imóveis, supervisor, gerente, mestre de obras, pastor, microempresário (proprietário de empresa com menos de 10 empregados), pequeno comerciante, pequeno proprietário de terras, trabalhador autônomo ou por conta própria.	16.301	15,4
Grupo 5	Médico, engenheiro, dentista, psicólogo, economista, advogado, juiz, promotor, defensor, delegado, tenente, capitão, coronel, professor universitário, diretor em empresas públicas ou privadas, político, proprietário de empresas com mais de 10 empregados.	3.905	3,7
Não Sabe	Não sei.	11.629	11,0

Fonte: adaptado do arquivo de metadados “Dicionário_Microdados_Enem_2023.xlsx”, (Inep, 2024).

Ao relacionar a referida variável com o desempenho, observa-se que a gradação da Tabela 12 associa-se positivamente ao desempenho, conforme demonstrado na Figura 51. Em outras palavras, o Grupo de Natureza da Ocupação mais elevada, que sugere melhores faixas de remuneração, aponta para o incremento do desempenho dos participantes maranhenses.

Figura 51 – Notas por Grupo de ocupação do Pai

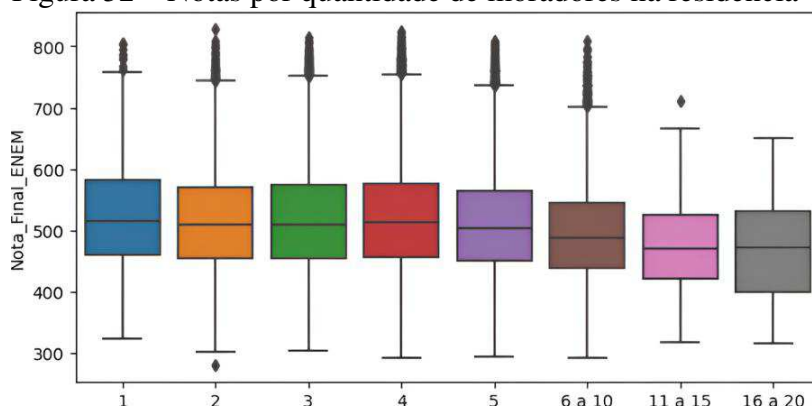


Fonte: elaboração própria.

Já a tendência das notas segundo a quantidade de moradores na residência aponta para uma associação negativa em relação ao desempenho, o que vai ao encontro dos achados

apontados pela literatura em Elias de Sousa *et al.* (2021), Silva, Morino e Sato (2014) e Silva *et al.* (2020), como se observa na Figura 52.

Figura 52 – Notas por quantidade de moradores na residência



Fonte: elaboração própria.

As demais variáveis selecionadas pela RFE (quantidade de quartos para dormir, motocicletas ou máquinas de lavar) também tiveram uma associação positiva para o desempenho, o que sugere um significado agregado em relação à caracterização socioeconômica dos participantes.

Em suma, a análise exploratória por questão socioeconômica coaduna com os achados relacionados positivamente ao desempenho dos participantes nos estudos correlatos, como sexo, etnia, nível de escolaridade dos pais, renda familiar mensal (Dutra *et al.*, 2023), possuir computador, acesso à internet na residência (Elias de Sousa *et al.*, 2021; Soares *et al.*, 2023) e quantidade de moradores no domicílio (Elias de Sousa *et al.*, 2021, Silva, Morino e Sato, 2014; Silva *et al.*, 2020).

4.2 Extração das Regras de Associação (Modelagem)

Para a modelagem dos dados, o *dataset* foi restringido para o Estado do Maranhão, segundo a variável “SG_UF_PROVA”, com um total de 106.088 (cento e seis mil e oitenta e oito) observações, considerando as variáveis resultantes da etapa de Seleção dos Atributos, elencadas na Tabela 9.

Já a variável “NU_NOTA_FINAL” foi categorizada em quartis, sendo criadas as categorias ‘A’, ‘B’, ‘C’ e ‘D’ na variável ‘NOTA’, adotando estratégia observada em Silva, Morino e Sato (2014) e em Silva *et al.* (2020), quando da extração de regras de associação com o algoritmo Apriori. A distribuição das categorias está evidenciada na Tabela 13:

Tabela 13 – Detalhamento da categorização da variável “NOTA”

NOTA	count	mean	std	min	25%	50%	75%	max
A	26533.0	415.550037	26.765272	281.26	398.48	420.56	437.34	452.30
B	26512.0	479.202083	15.320076	452.32	466.04	479.32	492.42	505.86
C	26522.0	535.629091	17.868057	505.88	520.18	534.88	550.68	568.58
D	26521.0	628.403849	48.045101	568.58	589.18	616.58	656.94	828.86

Fonte: elaboração própria.

Desse modo, as categorias ‘A’, ‘B’, ‘C’ e ‘D’ da variável ‘NOTA’ tiveram os seguintes intervalos:

- ‘A’: de 281,26 a 452,30 pontos.
- ‘B’: de 452,32 a 505,86 pontos.
- ‘C’: de 505,88 a 568,58 pontos.
- ‘D’: de 568,58 a 828,86 pontos.

A Figura 53 exibe a visualização parcial do subconjunto regional adotado para a extração das regras de associação. Cabe dizer que a categorização das variáveis teve pequenos ajustes para facilitar a identificação.

Figura 53 – Subconjunto Regional de Microdados do Enem 2023

	TP_FAIXA_ETARIA	TP_SEXO	TP_ST_CONCLUSAO	Q001	Q003	Q005	Q009	Q011	Q014	Q024	NOTA
0	26 a 30	Feminino	Egressos	Não completou 4º/5º Ano do E.F.	Grupo 2	3 Mors	Um Qt	Não Mt	Não ML	Não PC	B
1	23	Feminino	Egressos	Não completou 8º/9º Ano do E.F.	Grupo 3	4 Mors	Dois Qts	Não Mt	Não ML	Não PC	B
2	36 a 40	Feminino	Egressos	Não completou 4º/5º Ano do E.F.	Grupo 2	3 Mors	Um Qt	Não Mt	Não ML	Não PC	C
3	18	Masculino	Egressos	Não completou 4º/5º Ano do E.F.	Grupo 1	4 Mors	Dois Qts	Uma Mt	Não ML	Não PC	D
5	19	Masculino	Egressos	Não completou E.S.	Grupo 3	2 Mors	Um Qt	Não Mt	Não ML	Não PC	D

Fonte: elaboração própria.

A Tabela 14 apresenta o resumo dos experimentos realizados, conforme parâmetros de suporte e confiança mínimos, quantidade de antecedentes e regras geradas, com $lift > 1$.

Tabela 14 – Experimentos realizados

EXP.	minsup	minconf	antecedentes	Regras
1	3%	5%	[2,4]	22.887
2	10%	5%	[2,4]	846
3	10%	10%	[2,4]	846
4	20%	10%	[2,4]	87
5	20%	20%	[2,4]	87
6	25%	20%	[2,4]	43

EXP.	minsup	minconf	antecedentes	Regras
7	3%	5%	[3,6]	11.050
8	10%	5%	[3,6]	244
9	10%	10%	[3,6]	244
10	20%	10%	[3,6]	12

Fonte: elaboração própria.

Observa-se como os parâmetros, em especial o limiar de suporte, afetam a quantidade de regras geradas. Daí a importância da análise exploratória, verificação dos itens frequentes e do uso adequado das métricas de avaliação para inspecionar as regras geradas.

4.3 Avaliação e Discussão

O grande número de regras geradas pode ser contornado pelo refinamento segundo às métricas selecionadas, a especificar a *lift*, que colabora para verificar se a relação é de fato significativa (Silva; Peres; Boscarioli, 2016). Enquanto o experimento 1 apresentou demasiadas regras, os experimentos 3, 5 e 9 guardaram redundâncias em relação aos experimentos 2, 4 e 8, respectivamente, demonstrando que o hiperparâmetro de suporte é determinante para a poda das associações.

No experimento 1, a análise das 150 primeiras regras mais relevantes segundo a *lift*, que relacionaram a variável “NOTA”, estão elencadas na Tabela 15.

Tabela 15 – Regras de Associação do Experimento 1

id	antecedents	consequents	support	confidence	lift
01	(Grupo 4, Não Mt)	(D, Uma ML)	0.032398	0.347558	2.877905
02	(D, Uma ML)	(Grupo 4, Não Mt)	0.032398	0.268264	2.877905
03	(Concluintes, Não PC)	(A, Não ML, 18)	0.032699	0.127856	2.793827
04	(A, Não ML, 18)	(Concluintes, Não PC)	0.032699	0.714521	2.793827

Fonte: elaboração própria.

Id01) Se o pai integra o Grupo 4 de natureza da ocupação e não possuem motocicleta (Mt) então o candidato possui rendimento “D” (entre 568,58 e 828,86 pontos) e possui uma máquina de lavar no domicílio (ML), ocorre em 3,24% das transações com uma confiança de 34,8% e *lift* de 2,88. Em 34,76% das vezes em que pai integrante do Grupo 4 e não possuir motocicleta ocorrem rendimento “D” e possuir uma máquina de lavar no domicílio também ocorrem. Assim como, a probabilidade do candidato ter rendimento “D” e possuir uma máquina de lavar é 2,88 vezes maior quando o pai integra o Grupo 4 de natureza da ocupação e não possuem motocicleta.

A regra id01 aponta para a melhoria do desempenho acadêmico atenuado pelas condições socioeconômicas da família, isto é, pelo perfil de renda dos pais integrantes do Grupo 4 de natureza da ocupação, onde estão inclusas categorias como professores, técnicos, microempresários, comerciantes, pequenos proprietários de terras, autônomos, etc.

Id02) Se o candidato possui rendimento “D” (entre 568,58 e 828,86 pontos) e possui uma máquina de lavar no domicílio então o pai integra o Grupo 4 de natureza da ocupação e não possuem motocicleta, com um suporte de 3,24%, confiança de 26,8% e *lift* de 2,88.

Id03) Se o candidato é concluinte do ensino médio e não possui computador no domicílio (Não PC) então possui rendimento “A” (pontuação entre 281,26 e 452,30), não possui máquina de lavar no domicílio e tem 18 anos de idade, com um suporte de 3,27%, confiança de 12,8% e *lift* de 2,79. Neste caso, apenas 3,27% das transações incluem

“Concluintes, Não PC” seguidas de “A, Não ML, 18”. Em 12,79% das vezes em que “Concluintes, Não PC” ocorrem “A, Não ML, 18” também ocorrem. Apesar da baixa confiança, a probabilidade de “A, Não ML, 18” ocorrerem é 2,79 vezes maior quando “Concluintes, Não PC” ocorrem.

A regra id03 fortalece a influência das condições socioeconômicas, em especial, possuir computador na residência, para o desempenho dos alunos (Elias de Sousa *et al.*, 2021; Soares *et al.*, 2023; Silva *et al.*, 2020). Cabe dizer, que a análise exploratória evidenciou que 71% dos participantes maranhenses não possuem o citado dispositivo, bem como apontou a discrepância entre o desempenho daqueles que possuem pelo menos um computador na residência em comparação àqueles que não possuem.

A regra id04, por sua vez, apontou que em 71,45% das vezes em que “A, Não ML, 18” ocorrem, “Concluintes, Não PC” também ocorrem. Além disso, a probabilidade de “Concluintes, Não PC” ocorrerem é 2,79 vezes maior quando “A, Não ML, 18” ocorrem. Essa regra sugere uma associação mais significativa entre os itens do antecedente para a ocorrência do consequente.

Na Tabela 16 foram elencadas as regras do experimento 2, que associaram o grupo de participantes do sexo feminino ao baixo desempenho (categorias “A” ou “B”) ou às condições socioeconômicas menos favoráveis, como pai integrante do Grupo 1 de natureza da ocupação, ausência de computador e máquina de lavar.

Tabela 16 – Regras de Associação do Experimento 2

id	antecedents	consequents	support	confidence	lift
01	(Não ML, A)	(Grupo 1)	0.102820	0.536653	1.616709
02	(Não PC, A)	(Grupo 1)	0.111059	0.517390	1.558678
03	(Grupo 1, Não PC)	(A)	0.111059	0.386549	1.545553
04	(Grupo 1, Não ML)	(A)	0.102820	0.370567	1.481652
05	(Não PC, Feminino)	(B, Não ML)	0.114056	0.241599	1.313051
06	(B, Não ML)	(Não PC, Feminino)	0.114056	0.619877	1.313051
07	(A, Feminino)	(Não ML, Não PC)	0.121079	0.739493	1.310734
08	(Não ML, Não PC)	(A, Feminino)	0.121079	0.214609	1.310734
09	(Não Mt, Não PC)	(B, Não ML)	0.108881	0.240175	1.305314
10	(B, Não ML)	(Não Mt, Não PC)	0.108881	0.591752	1.305314
11	(A, Não ML, Feminino)	(Não PC)	0.121079	0.924833	1.302140
12	(Dois Qts, Não ML, A)	(Não PC)	0.105224	0.917256	1.291472
13	(Grupo 1, A)	(Não PC)	0.111059	0.914965	1.288246
14	(Dois Qts, A)	(Não ML, Não PC)	0.105224	0.724776	1.284648
15	(Não ML, Não PC)	(Dois Qts, A)	0.105224	0.186507	1.284648
16	(Não ML, A)	(Não PC)	0.174647	0.911542	1.283427
17	(B, Não ML, Não Mt)	(Não PC)	0.108881	0.909098	1.279986
18	(Não ML, A)	(Dois Qts, Não PC)	0.105224	0.549198	1.278798

Fonte: elaboração própria.

As regras id01 e id02 associaram o menor nível de desempenho (categoria “A”) ao Grupo 1 de natureza da ocupação dos pais, com um suporte em torno de 11%, confiança de 54% e 52% e *lift* de 1,62 e 1,56, respectivamente. Esses subconjuntos ocorrem em torno de 11% das transações. Significa dizer, também, que quando o candidato não possui moto ou computador e tem desempenho “A” implica no fato de o pai integrar o Grupo 1 de natureza da ocupação em mais da metade dos casos. O desempenho mais baixo (categoria “A”) está relacionado à natureza ocupacional do Grupo 1, possivelmente indicando uma relação com fatores socioeconômicos.

Já as regras id03 e id04 indicam que o fato de o pai integrar o Grupo 1 de natureza da ocupação implica no baixo desempenho do candidato (nível “A”) em torno de 38% dos casos, com *lift* de 1,54 e 1,48, respectivamente.

O desempenho de nível “B” (entre 452,32 e 505,86) apareceu com mais força nas regras id05 e id06 associado ao fato de não possuir computador na residência:

Id05) Se o candidato não possui computador na residência e é do sexo feminino então teve desempenho de nível “B” e não possui máquina de lavar com uma confiança de 24,2% e *lift* de 1,31. A regra sugere uma associação fraca, mas interessante, entre restrições materiais (como falta de computador e máquina de lavar) e desempenho “B” no caso de mulheres.

Id06) Se o candidato teve desempenho de nível “B” e não possui máquina de lavar então não possui computador na residência e é do sexo feminino com um suporte de 11,4%, confiança de 62% e *lift* de 1,31.

A regra id07 indica que se o candidato teve desempenho de nível “A” e é do sexo feminino então não possui máquina de lavar e computador na residência, com um suporte de 12,1%, confiança de 74% e *lift* de 1,31. Já a regra recíproca id08, isto é, se o candidato não possui máquina de lavar e computador na residência então possui desempenho nível “A” e é do sexo feminino, tem suporte de 12,1%, confiança de 21,5% e *lift* de 1,31.

As regras id’s 11, 12, 13 e 16 associam o desempenho de nível “A” (entre 281,26 e 452,3 pontos) ao fato de não possuir computador na residência com uma confiança acima de 91% e *lift* em torno de 1,30, o que ratifica a importância do dispositivo (Elias de Sousa *et al.*, 2021; Soares *et al.*, 2023; Silva *et al.*, 2020):

Id11) Se o candidato possui desempenho de nível “A”, não possui máquina de lavar e é do sexo feminino então não possui computador na residência, com suporte de 12,1%, confiança de 93% e *lift* de 1,30.

Id12) Se a residência do candidato possui dois quartos para dormir, não possui máquina de lavar e o desempenho foi nível “A”, então não possui computador na residência com suporte de 10,5%, confiança de 92% e *lift* de 1,29.

Id13) Se o pai do candidato está no Grupo 1 de natureza da ocupação e o desempenho foi nível “A” então não possui computador na residência, com suporte de 11,1%, confiança de 91,5% e *lift* de 1,29.

Id16) Se o candidato não possui máquina de lavar e teve desempenho de nível “A” então não possui computador na residência, com suporte de 17,5%, confiança de 91% e *lift* de 1,28.

A regra id11 aponta que probabilidade condicional de não possuir computador dado que o candidato teve desempenho de nível “A” (entre 281,26 e 452,3 pontos), não possui máquina de lavar e é do sexo feminino, é de 91,5%. Já a regra id13 indica que a probabilidade condicional de não possuir computador dado que o pai integra o Grupo 1 de natureza da ocupação e o desempenho foi nível “A” (entre 281,26 e 452,3 pontos) é de 91,5%.

Essas as regras apontam para a falta de recursos materiais como elementos influentes para o baixo desempenho, dada a confiança elevada, assim como, sinalizam para o perfil de vulnerabilidade do participante do sexo feminino. A ausência de computadores na residência aparece mais consistentemente associada ao baixo desempenho, destacando a importância do dispositivo para a melhoria os resultados educacionais (Elias de Sousa *et al.*, 2021; Soares *et al.*, 2023; Silva *et al.*, 2020).

No experimento 4, nenhuma das 87 (oitenta e sete) regras geradas associaram a variável “NOTA”, destacando-se aquelas que apontam para condições socioeconômicas menos favoráveis envolvendo participantes do sexo feminino, elencadas na Tabela 17.

Tabela 17 – Regras de Associação do Experimento 4

id	antecedents	consequents	support	confidence	lift
01	(Não ML, Não PC)	(Grupo 1)	0.250830	0.444589	1.339361
02	(Grupo 1, Não PC)	(Não ML)	0.250830	0.873031	1.292377
03	(Não Mt, Não PC)	(Não ML, Feminino)	0.261406	0.576621	1.280725
04	(Não ML, Feminino)	(Não Mt, Não PC)	0.261406	0.580605	1.280725
05	(Não ML, Não Mt)	(Não PC, Feminino)	0.261406	0.604078	1.279584
06	(Não PC, Feminino)	(Não ML, Não Mt)	0.261406	0.553721	1.279584
07	(Grupo 1, Não ML)	(Não PC)	0.250830	0.903995	1.272801
08	(Não Mt, Não PC, Feminino)	(Não ML)	0.261406	0.843995	1.249393
09	(Não ML, Não Mt, Feminino)	(Não PC)	0.261406	0.884537	1.245405
10	(Não ML, Feminino)	(Dois Qts, Não PC)	0.238557	0.529855	1.233758
11	(Dois Qts, Não ML, Feminino)	(Não PC)	0.238557	0.866564	1.220099
12	(Não ML, Não Mt)	(Não PC)	0.374020	0.864316	1.216934
13	(Não PC, Feminino)	(Não ML)	0.386566	0.818841	1.212156
14	(Não ML, Feminino)	(Não PC)	0.386566	0.858596	1.208881

Fonte: elaboração própria.

As regras id's 03, 04 e 05 associam o fato de não possuir motocicleta, computador e máquina de lavar ao sexo feminino com suporte de 26,1% uma confiança em torno de 58% e *lift* de 1,28:

Id03) Se o participante não possui motocicleta e computador na residência, então não possui máquina de lavar e é do sexo feminino, com suporte de 26,1%, confiança de 58% e *lift* de 1,28.

Id04) Se o participante não possui máquina de lavar e é do sexo feminino, então não possui motocicleta e computador na residência, com suporte de 26,1%, confiança de 58% e *lift* de 1,28.

Id05) Se o candidato não possui máquina de lavar e motocicleta na residência, então não possui computador na residência e é do sexo feminino, com suporte de 26,1%, confiança de 60,4% e *lift* de 1,28.

Já as regras id's 08, 09, 11, 13 e 14 associam tais questões (não possuir motocicleta, computador e máquina de lavar) ao sexo feminino com uma confiança acima de 81% e *lift* em torno de 1,21:

Id08) Se o participante não possui motocicleta e computador na residência, sendo do sexo feminino, então não possui máquina de lavar, com suporte de 26,1%, confiança de 84,4% e *lift* de 1,25.

Id09) Se o participante não possui máquina de lavar e motocicleta na residência, sendo do sexo feminino, então não possui computador na residência, com suporte de 26,1%, confiança de 88,5% e *lift* de 1,25.

Id11) Se a residência do participante possui dois quartos para dormir, não possui máquina de lavar e é do sexo feminino, então não possui computador, com suporte de 23,9%, confiança de 86,7% e *lift* de 1,22.

Id13) Se o participante não possui computador na residência e é do sexo feminino, então não possui máquina de lavar, com suporte de 38,7%, confiança de 81,9% e *lift* de 1,21.

Id14) Se o participante não possui máquina de lavar e é do sexo feminino, então não possui computador na residência, com suporte de 38,7%, confiança de 86% e *lift* de 1,21.

Essas regras reforçam o impacto cumulativo de múltiplas limitações socioeconômicas aos participantes do sexo feminino (Dutra *et al.*, 2023; Elias de Sousa *et al.*, 2021).

No experimento 6, nenhuma das 43 (quarenta e três) regras geradas associaram a variável “NOTA”. Assim como no experimento 4, as regras de maior de maior força apontam

para condições sociais menos favoráveis envolvendo participantes do sexo feminino, conforme resultados elencados na Tabela 18 (Dutra *et al.*, 2023; Elias de Sousa *et al.*, 2021).

Tabela 18 – Regras de Associação do Experimento 6

id	antecedents	consequents	support	confidence	lift
01	(Não Mt, Não PC)	(Não ML, Feminino)	0.261406	0.576621	1.280725
02	(Não ML, Feminino)	(Não Mt, Não PC)	0.261406	0.580605	1.280725
03	(Não ML, Não Mt)	(Não PC, Feminino)	0.261406	0.604078	1.279584
04	(Não PC, Feminino)	(Não ML, Não Mt)	0.261406	0.553721	1.279584
05	(Não Mt, Não PC, Feminino)	(Não ML)	0.261406	0.843995	1.249393
06	(Não PC, Feminino)	(Não ML)	0.386566	0.818841	1.212156
07	(Não ML, Feminino)	(Não PC)	0.386566	0.858596	1.208881
08	(Dois Qts, Feminino)	(Não ML)	0.275290	0.759795	1.124749
09	(Dois Qts, Feminino)	(Não PC)	0.287026	0.792185	1.115375
10	(Não ML, Não Mt, Não PC)	(Feminino)	0.261406	0.698909	1.111232

Fonte: elaboração própria.

As regras id's 05, 06 e 07 associam o fato de não possuir motocicleta, computador e máquina de lavar ao sexo feminino com uma confiança acima de 81% e *lift* de 1,2:

Id05) Se o participante não possui motocicleta e computador na residência, sendo do sexo feminino, então não possui máquina de lavar, com suporte de 26,1%, confiança de 84,4% e *lift* de 1,25.

Id06) Se o participante não possui computador na residência e é do sexo feminino, então não possui máquina de lavar, com suporte de 38,7%, confiança de 81,9% e *lift* de 1,21.

Id07) Se o participante não possui máquina de lavar e é do sexo feminino, então não possui computador na residência, com suporte de 38,7%, confiança de 85,9% e *lift* de 1,21.

Essas regras evidenciam que uma parcela significativa das transações revelam associações que refletem as condições socioeconômicas dos participantes do Enem 2023, em especial, do sexo feminino, com um alto nível de confiança (Dutra *et al.*, 2023; Elias de Sousa *et al.*, 2021).

Já no experimento 7, os baixos índices de suporte e confiança retornaram regras com maiores valores para a métrica *lift*. Na Tabela 19 foram reunidas as regras julgadas mais relevantes, considerando as variáveis “NOTA” e sexo (Dutra *et al.*, 2023; Elias de Sousa *et al.*, 2021).

Tabela 19 – Regras de Associação do Experimento 7

id	antecedents	consequents	support	confidence	lift
01	(Não ML, Cursistas, Feminino)	(Não PC, <17)	0.032925	0.425612	6.394612
02	(Cursistas, Não PC, Feminino)	(Não ML, <17)	0.032925	0.392296	6.384086
03	(<17, Não PC, Feminino)	(Cursistas, Não ML)	0.032925	0.708376	6.343926
04	(Não ML, <17, Feminino)	(Cursistas, Não PC)	0.032925	0.765505	6.314019
05	(Não ML, Não PC, <17)	(Cursistas, Feminino)	0.032925	0.662306	5.418577
06	(A, Não ML, 18)	(Concluintes, Não PC)	0.032699	0.714521	2.793827

07	(Não ML, Concluintes, Não PC)	(A, 18)	0.032699	0.162528	2.686969
08	(A, Não PC, 18)	(Não ML, Concluintes)	0.032699	0.611062	2.660853
09	(D, Uma ML, Não Mt)	(Grupo 4)	0.032398	0.405259	2.637452
10	(Não ML, Concluintes, A)	(Não PC, 18)	0.032699	0.469291	2.633492
11	(Não ML, Não PC, 18)	(Concluintes, A)	0.032699	0.232788	2.524892
12	(Concluintes, Não PC, A)	(Não ML, 18)	0.032699	0.399286	2.477888
13	(Masculino, Não ML, 18)	(Concluintes, Não PC)	0.034236	0.625668	2.446403
14	(Masculino, Não ML, Concluintes)	(Não PC, 18)	0.034236	0.424646	2.382961
15	(Masculino, Não PC, 18)	(Não ML, Concluintes)	0.034236	0.545673	2.376118
16	(Grupo 4, Uma ML, Não Mt)	(D)	0.032398	0.589133	2.356620
17	(Não ML, Não completou 4º/5º Ano do E.F., A)	(Grupo 1, Não PC)	0.034698	0.671103	2.335826

Fonte: elaboração própria.

As cinco primeiras regras ranqueadas segundo a *lift* relacionaram o sexo feminino às condições socioeconômicas menos favoráveis. A regra id01, por exemplo, indica que o fato de o participante não possuir computador e ter menos de 17 anos é 6,4 vezes mais provável quando o participante não possui máquina de lavar, é cursista do ensino médio e é do sexo feminino.

As regras id's 06, 07, 08, 11 e 12 trouxeram outras variáveis relacionadas ao baixo desempenho, como idade (Elias de Sousa *et al.*, 2021), situação de conclusão do ensino médio e nível de escolaridade do pai ou homem responsável pelo participante:

Id06) Se o candidato tem desempenho de nível “A”, não possui máquina de lavar e tem 18 anos, então é concluinte do ensino médio e não possui computador na residência, com suporte de 3,3%, confiança de 71,5% e *lift* de 2,8.

Id07) Se o participante não possui máquina de lavar, é concluinte do ensino médio e não possui computador, então obteve desempenho de nível “A” e tem 18 anos de idade, com suporte de 3,3%, confiança de 16,3% e *lift* de 2,7.

Id08) Se o participante tem desempenho de nível “A”, não possui computador na residência e tem 18 anos, então não possui máquina de lavar e é concluinte do ensino médio, com suporte de 3,3%, confiança de 61,1% e *lift* de 2,7.

Id11) Se o participante não possui máquina de lavar e computador, com 18 anos de idade, então é concluinte do ensino médio e teve desempenho de nível “A”, com suporte de 3,3%, confiança de 23,3% e *lift* de 2,5.

Id12) Se o participante é um concluinte do ensino médio, não possui computador e teve desempenho de nível “A”, então não possui máquina de lavar e tem 18 anos, com suporte de 3,3%, confiança de 40% e *lift* de 2,5.

Tais regras apontam perfis de vulnerabilidade, onde fatores socioeconômicos contribuem para o baixo desempenho acadêmico dos participantes mais jovens, isto é,

daqueles no último ano do ensino médio. A *lift* elevada nessas regras (acima de 2,5) reflete um grau de associação relevante entre as variáveis analisadas.

Neste ponto, fica evidente um provável perfil de participante com menores chances de ingresso à educação superior devido ao desempenho, e que, devido às condições sociais em quem está inserido é levado a buscar alternativas após a conclusão da educação básica, como a inserção no mercado de trabalho.

A regra id17 acrescentou outra variável influente para o desempenho – o nível de escolaridade do pai:

Id17) Se o participante não possui máquina de lavar, teve desempenho de nível “A” e o pai não completou o 4º/5º Ano do Ensino Fundamental, então o pai pertence ao Grupo 1 de natureza da ocupação e não há computador na residência, com suporte de 3,5%, confiança de 67,1% e *lift* de 2,34.

Cabe dizer, que a análise exploratória apontou uma associação positiva entre o nível de escolaridade dos pais (Dutra *et al.*, 2023; Elias de Sousa *et al.*, 2021; Soares *et al.*, 2023; Silva, Morino e Sato, 2014; Silva *et al.*, 2020), bem como, da renda, ao desempenho dos participantes maranhenses (Dutra *et al.*, 2023, Elias de Sousa *et al.*, 2021; Silva, Morino e Sato, 2014; Silva *et al.*, 2020; Gomes, Gouveia e Batista, 2017).

A regra id17 identifica um grupo de participantes afetados tanto pela baixa escolaridade quanto pela natureza da ocupação do pai ou homem responsável, apesar do baixo suporte. A confiança de 67,1% e *lift* de 2,34 reforçam a interdependência entre as condições econômicas, sociais e educacionais para o baixo desempenho (Dutra *et al.*, 2023; Elias de Sousa *et al.*, 2021; Soares *et al.*, 2023; Silva, Morino e Sato, 2014; Silva *et al.*, 2020).

As regras id's 09 e 16 trouxeram regras que contemplaram a categoria de melhor desempenho no Enem – nível “D” (entre 568,58 e 828,86 pontos) – associada ao Grupo 4 de natureza da ocupação do pai.

Id09) Se o participante tem desempenho de nível “D” (entre 568,58 e 828,86 pontos), possui uma máquina de lavar e não possui motocicleta, então o pai integra o Grupo 4 de natureza da ocupação, com suporte de 3,2%, confiança de 40,5% e *lift* de 2, 6.

Id16) Se o pai do participante pertence ao Grupo 4 de natureza da ocupação, possui uma máquina de lavar e não possui motocicleta, então teve desempenho de nível “D”, com suporte de 3,2%, confiança de 59% e *lift* de 2,4.

O Grupo 4 de natureza ocupação do pai está associado a melhores condições educacionais e financeiras, com maior probabilidade de desempenho acadêmico elevado, indicando que as condições socioeconômicas e profissionais do responsável exercem

influência no sucesso educacional (Dutra *et al.*, 2023, Elias de Sousa *et al.*, 2021; Silva, Morino e Sato, 2014; Silva *et al.*, 2020; Gomes, Gouveia e Batista, 2017).

Já o experimento 9, com 245 (duzentas e quarenta e cinco) regras, permitiu identificar o Grupo 1 de natureza da ocupação do pai, bem como, diferentes níveis de desempenho associados aos participantes do sexo feminino. Na Tabela 20 foram elencadas algumas dessas regras:

Tabela 20 – Regras de Associação do Experimento 9

id	antecedents	consequents	support	confidence	lift
01	(Não ML, Não PC, Feminino)	(Grupo 1, Não Mt)	0.111115	0.287442	1.525700
02	(Grupo 1, Não Mt, Feminino)	(Não ML, Não PC)	0.111115	0.838705	1.486585
03	(Grupo 1, Não Mt, Não PC)	(Não ML, Feminino)	0.111115	0.653364	1.451179
04	(Grupo 1, Dois Qts, Feminino)	(Não ML, Não PC)	0.107232	0.807267	1.430860
05	(Não ML, Não Mt, A)	(Não PC)	0.116696	0.935469	1.317116
06	(A, Não ML, Feminino)	(Não PC)	0.121079	0.924833	1.302140
07	(Dois Qts, Não ML, A)	(Não PC)	0.105224	0.917256	1.291472
08	(B, Não ML, Não Mt)	(Não PC)	0.108881	0.909098	1.279986
09	(Não Mt, Não PC, A)	(Não ML)	0.116696	0.854382	1.264770
10	(B, Não ML, Feminino)	(Não PC)	0.114056	0.897693	1.263928
11	(Não ML, Não PC, Feminino)	(A)	0.121079	0.313216	1.252346
12	(Não ML, Não Mt, Não PC)	(A)	0.116696	0.312004	1.247498
13	(B, Não Mt, Não PC)	(Não ML)	0.108881	0.838366	1.241059
14	(A, Não PC, Feminino)	(Não ML)	0.121079	0.835882	1.237383
15	(B, Não PC, Feminino)	(Não ML)	0.114056	0.832817	1.232846
16	(Dois Qts, Não PC, A)	(Não ML)	0.105224	0.824629	1.220724
17	(Dois Qts, Não ML, Não PC)	(A)	0.105224	0.302578	1.209809
18	(Não ML, Não PC, Feminino)	(B)	0.114056	0.295050	1.180645
19	(Não ML, Não Mt, Não PC)	(B)	0.108881	0.291111	1.164884
20	(Masculino, Não Mt, Não PC)	(Não ML)	0.112614	0.784130	1.160772
21	(Masculino, Não ML, Não Mt)	(Não PC)	0.112614	0.820761	1.155610
22	(B, Não ML, Não PC)	(Feminino)	0.114056	0.703447	1.118448
23	(Não ML, Não PC, A)	(Feminino)	0.121079	0.693275	1.102274
24	(B, Não ML, Não PC)	(Não Mt)	0.108881	0.671531	1.079725
25	(Não ML, Não PC, A)	(Não Mt)	0.116696	0.668178	1.074334
26	(Não ML, Não PC, A)	(Dois Qts)	0.105224	0.602494	1.062880
27	(Masculino, Não ML, Não PC)	(Não Mt)	0.112614	0.634029	1.019427

Fonte: elaboração própria.

As quatro primeiras regras mostraram a relação do Grupo 1 da natureza da ocupação do pai, condições socioeconômicas (ausência de computador) aos participantes do sexo feminino. No caso das regras id's 02 e 04, a confiança foi acima de 80%:

Id02) Se o participante pertence ao Grupo 1 de natureza da ocupação, não possui motocicleta e é do sexo feminino, então não possui máquina de lavar e computador na residência, com suporte de 11,1%, confiança de 84% e *lift* de 1,5.

Id04) Se o pai pertence ao Grupo 1 de natureza da ocupação, a residência possui dois quartos para dormir e o participante é do sexo feminino, então não possui máquina de lavar e computador, com suporte de 10,7%, confiança de 80,7% e *lift* de 1,4.

Ambas as regras apontam para a influência da natureza da ocupação do pai como um fator determinante para as condições materiais do domicílio, principalmente para participantes do sexo feminino.

As regras id's 06, 10, 14, 16, 22 e 23 revelaram, mais uma vez, a influência das condições socioeconômicas para o baixo desempenho dos participantes do sexo feminino. Neste caso, as regras evidenciaram a associação aos dois menores níveis de desempenho, isto é, categorias "A" (entre 281,26 e 452,3 pontos) e "B" (entre 452,32 e 505,86 pontos), com uma confiança acima de 70%:

Id06) Se o participante obteve desempenho de nível "A", não possui máquina de lavar e é do sexo feminino, então não possui computador na residência, com suporte de 12,1%, confiança de 92,5% e *lift* de 1,3.

Id10) Se o participante obteve desempenho de nível "B", não possui máquina de lavar e é do sexo feminino, então não possui computador na residência, com suporte de 11,4%, confiança de 89,8% e *lift* de 1,3.

Id14) Se o participante obteve desempenho de nível "A", não possui computador na residência e é do sexo feminino, então não possui máquina de lavar, com suporte de 12,1%, confiança de 83,6% e *lift* de 1,2.

Id15) Se o participante obteve desempenho de nível "B", não possui computador na residência e é do sexo feminino, então não possui máquina de lavar, com suporte de 11,4%, confiança de 83,3% e *lift* de 1,2.

Id22) Se o participante obteve desempenho de nível "B", não possui máquina de lavar e computador na residência, então é do sexo feminino, com suporte de 11,4%, confiança de 70,3% e *lift* de 1,1.

Id23) Se o participante não possui máquina de lavar e computador na residência e obteve desempenho de nível "A", então é do sexo feminino, com suporte de 12,1%, confiança de 69,3% e *lift* de 1,1.

O décimo experimento, por sua vez, gerou apenas 12 regras, elencadas na Tabela 21. Observa-se que a parametrização mais rigorosa, retornou regras com valores para a métrica *lift* cada vez menores. Por outro lado, a confiança das regras evidenciadas mostrou-se bastante significativa. As primeiras seis regras, segundo a *lift*, apresentaram confiança entre 83% e 86% como se evidencia na Tabela 21:

Tabela 21 – Regras de Associação do Experimento 10

id	antecedents	consequents	support	confidence	lift
01	(Não Mt, Não PC, Feminino)	(Não ML)	0.261406	0.843995	1.249393
02	(Não ML, Não Mt, Feminino)	(Não PC)	0.261406	0.884537	1.245405
03	(Dois Qts, Não PC, Feminino)	(Não ML)	0.238557	0.831133	1.230353
04	(Dois Qts, Não Mt, Não PC)	(Não ML)	0.228848	0.830329	1.229163
05	(Dois Qts, Não ML, Não Mt)	(Não PC)	0.228848	0.866700	1.220291
06	(Dois Qts, Não ML, Feminino)	(Não PC)	0.238557	0.866564	1.220099
07	(Não ML, Não Mt, Não PC)	(Feminino)	0.261406	0.698909	1.111232
08	(Dois Qts, Não ML, Não PC)	(Feminino)	0.238557	0.685984	1.090682
09	(Não ML, Não PC, Feminino)	(Dois Qts)	0.238557	0.617118	1.088679
10	(Não ML, Não PC, Feminino)	(Não Mt)	0.261406	0.676225	1.087273
11	(Não ML, Não Mt, Não PC)	(Dois Qts)	0.228848	0.611860	1.079404
12	(Dois Qts, Não ML, Não PC)	(Não Mt)	0.228848	0.658065	1.058075

Fonte: elaboração própria.

As demais regras, por sua vez, retornaram regras com valores da métrica lift muito próximos a 1, sugerindo baixa ou ausência de influência dos antecedentes sobre os consequentes.

Dentre outras variáveis influentes para o desempenho, a quantidade de moradores no domicílio – variável que possui associação negativa para o desempenho (Elias de Sousa *et al.*, 2021, Silva; Morino; Sato, 2014, Silva *et al.*, 2020) – apareceu dentre aquelas selecionadas pela RFE, entretanto, não se vislumbrou regras relevantes quanto à associação com a variável “NOTA”, apenas regras que confirmavam o perfil socioeconômico dos participantes.

5 CONCLUSÕES, CONTRIBUIÇÕES E TRABALHOS FUTUROS

O presente trabalho investigou os fatores influentes no desempenho dos estudantes maranhenses no Enem 2023, utilizando Regras de Associação para Mineração de Dados Educacionais da base pública de microdados mantida pelo Inep.

A análise exploratória estadual trouxe à tona o comportamento dos dados diferenciando-os da caracterização nacional, evidenciando o Maranhão como objeto de estudo pelas suas características étnicas, econômicas e sociais, bem como pelo desempenho no exame em comparação aos demais Estados da Federação.

A mineração de dados educacionais realizada mediante extração de regras de associação com o algoritmo FP-Growth colaborou para descobrir como essas variáveis ou seus subgrupos, combinados ou não, consolidam nichos não vislumbrados na Análise Exploratória.

Os nichos descobertos com maior força de associação atrelaram a natureza da ocupação do pai ou homem responsável pelo participante, elemento que possui forte correlação com a renda familiar mensal, bem como, idade, sexo, situação de conclusão do ensino médio, posse de computador, dentre outras questões materiais que caracterizaram o perfil socioeconômico dos participantes maranhenses, às menores faixas de desempenho no Enem.

Além disso, as regras constatadas forneceram indicativos críticos sobre o desempenho feminino, que, apesar de ser maior número, mostrou-se inferior aos resultados apresentados pelos candidatos do sexo masculino. A inquietação sugerida no referido achado pode servir como limiar para uma compreensão detalhada do contexto socioeconômico e cultural das meninas maranhenses e nortear a idealização de políticas públicas.

Essas descobertas contêm elementos que podem direcionar estudos mais aprofundados sobre variáveis influentes para desempenho, não apenas no contexto maranhense, ampliando o olhar sobre como as condições socioeconômicas e educacionais moldam o sucesso dos estudantes nessa importante avaliação nas diferentes regiões do país.

Desse modo, pode-se inferir que o diferencial desse trabalho reside na escolha do Maranhão como objeto de estudo, um estado com desafios educacionais significativos, tornando a investigação socialmente relevante. Além disso, adota-se a aprendizagem de máquina não supervisionada para realizar uma tarefa descritiva, contrapondo a predominância das análises preditivas com algoritmos de aprendizagem de máquina supervisionados. No tocante aos experimentos, optou-se por um algoritmo de extração de regras de associação não

evidenciado nos estudos correlatos, que apresenta melhor desempenho em termos de custo computacional do que o algoritmo Apriori.

O estudo consegue inserir outra variável relevante para o desempenho dos participantes, que permite uma melhor compreensão e caracterização socioeconômica do público maranhense, destacando-se a natureza da ocupação do pai ou do homem responsável, em atenção às lacunas de pesquisa.

As contribuições do presente estudo contemplam:

- a) A abordagem de problemas educacionais a partir da mineração de dados;
- b) A aplicação de MDE no Enem 2023 permitiu revelar padrões ocultos nos dados. Por exemplo, grupos que são afetados de forma cumulativa por diferentes atributos, como o baixo desempenho condicionado pela natureza da ocupação e escolaridade do pai, ou ainda, a caracterização do perfil socioeconômico dos participantes maranhenses a partir das variáveis estudadas com alto nível de confiança;
- c) A relevância social do Maranhão enquanto objeto de estudo;
- d) Abordagem quantitativa bem estruturada com o CRISP-DM, com descrição detalhada da metodologia, permitindo sua reprodutibilidade;
- e) Redução de dimensionalidade das características mais relevantes para o desempenho dos participantes maranhenses no Enem, por meio da Aprendizagem de Máquina, reduzindo o custo computacional e facilitando a interpretabilidade do modelo;
- f) A descoberta da natureza da ocupação do pai ou do homem responsável como variável influente para o desempenho dos participantes maranhenses no Enem 2023.
- g) A discussão das regras de associação relacionadas ao desempenho, obtidas em diferentes configurações do algoritmo FP-Growth, avaliando toda a base;
- h) O alinhamento do trabalho aos ODS da ONU, inserido assim, numa agenda global que visa à educação inclusiva e equitativa.
- i) A submissão do artigo para o *IEEE Transaction on Education*.

O trabalho contou com limitações em relação à qualidade dos dados, referente às variáveis que possuíam preenchimento apenas pela categoria de concluintes, o que deixou, por exemplo, a dependência administrativa da escola fora da análise.

Além disso, a inserção de uma abordagem supervisionada poderia fortalecer as conclusões. O estudo consegue ratificar a importância de fatores como renda e escolaridade

dos pais para o desempenho, mas não explora possíveis diferenças entre as regiões do Estado, por exemplo, rural *versus* urbano.

Por fim, os trabalhos futuros incluem: (i) refinar a utilização dos hiperparâmetros através do uso de metaheurísticas, como por exemplo, Algoritmos Genéticos; (ii) testar algoritmos de clusterização para validação de regras, para identificar padrões nos dados e verificar a robustez das regras extraídas; e (iii) estender o estudo utilizando a base nacional, permitindo uma avaliação mais ampla da aplicabilidade dos modelos em diferentes contextos regionais e garantindo maior representatividade dos resultados.

REFERÊNCIAS

- AGRAWAL, R.; SRIKANT, R. *Fast algorithms for mining association rules in large data bases*. In: Proceedings of the 20th International Conference on Very Large Data Bases (VLDB), Santiago, Chile, 1994, p. 487-499. **Anais**. Disponível em: <<https://www.vldb.org/conf/1994/P487.PDF>>. Acesso em 08 jul. 2024.
- BAKER, Ryan Shaun Joazeiro de; ISOTANI, Seiji; CARVALHO, Adriana Maria Joazeiro Baker de. Mineração de dados educacionais: oportunidades para o Brasil. **Revista Brasileira de Informática na Educação (RBIE)**, v. 19, n. 2, p. 3-13, 2011. Disponível em: <<http://milanesa.ime.usp.br/rbie/index.php/rbie/article/view/1301/1172>>. Acesso em: 04 set. 2024.
- _____. **Institui a Política de Dados Abertos do Poder Executivo Federal**. Decreto nº 8.777, de 11 de maio de 2016.. Brasília: Presidência da República, [2020]. Disponível em: <http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2016/decreto/d8777.htm>. Acesso em: 10 set. 2024.
- _____. **Lei de Acesso à Informação**. Lei nº 12.527 de 18 de novembro de 2011. Regula o acesso a informações previsto no inciso XXXIII do art. 5º, no inciso II do § 3º do art. 37 e no § 2º do art. 216 da Constituição Federal; altera a Lei no 8.112, de 11 de dezembro de 1990; revoga a Lei no 11.111, de 5 de maio de 2005, e dispositivos da Lei no 8.159, de 8 de janeiro de 1991; e dá outras providências. Diário Oficial da União, Brasília, 18 nov. 2011a. Disponível em: <http://www.planalto.gov.br/ccivil_03/_Ato2011-2014/2011/Lei/L12527.htm>. Acesso em: 10 set. 2024.
- BRASIL. **Lei de Diretrizes e Bases da Educação Nacional**. Lei nº 9.394, de 20 de dezembro de 1996. Estabelece as diretrizes e bases da educação nacional. Disponível em: <https://www.planalto.gov.br/ccivil_03/LEIS/L9394.htm>. Acesso em: 10 set. 2024.
- _____. **Lei Geral de Proteção de Dados Pessoais**. Lei nº 13.709, de 14 de agosto de 2018. Diário Oficial da União, Brasília, DF, 15 ago. 2018. Disponível em: <https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm>. Acesso em: 10 set. 2024.
- CARVALHO, André C. P. L. F. de; MENEZES, Angelo Garangau; BONIDIA, Robson Parmezan. **Ciência de Dados: fundamentos e aplicações**. 1. ed. Rio de Janeiro: LTC, 2024.
- CHAPMAN, P. *et al.* **CRISP-DM 1.0: Step-by-step data mining guide**. SPSS inc, 2000. Disponível em: <<https://mineracaodedados.files.wordpress.com/2012/12/crisp-dm-1-0.pdf>>. Acesso em: 25 abr. 2024.
- DE CASTRO, Leandro Nunes; FERRARI, Daniel Gomes. **Introdução à mineração de dados: conceitos básicos, algoritmos e aplicações**. 1. ed. São Paulo: Saraiva, 2016.
- DUTRA, J. F.; FIRMINO JÚNIOR, J. B.; FERNANDES, D. Y. de S. Fatores que podem interferir no desempenho de estudantes no ENEM: uma revisão sistemática da literatura. **Revista Brasileira de Informática na Educação**, [S. l.], v. 31, p. 323–351, 2023. Disponível em <<https://journals-sol.sbc.org.br/index.php/rbie/article/view/3087>>. Acesso em: 16 maio. 2024.
- ELIAS DE SOUZA, A.; SÁ SANTOS, L. M. de; MENEZES LARRUSCAIM, I. de; NOBREGA BESARRIA, C da. Determinantes do Desempenho no ENEM na Região Nordeste: Uma Análise de Dados em Painel do Período de 2015 a 2019. **Revista Brasileira de Estudos Regionais e Urbanos**, [S. l.], v. 15, n. 4, p. 690–711, 2022. Disponível em: <<https://revistaaber.org.br/rberu/article/view/915>>. Acesso em: 16 maio. 2024.

Enem 2024 tem 4,3 milhões de inscritos confirmados. **Inep**, 2024. Disponível em: <<https://www.gov.br/inep/pt-br/assuntos/noticias/enem/enem-2024-tem-4-3-milhoes-de-inscritos-confirmados>>. Acesso em: 12 dez. 2024.

Estatísticas do Censo Escolar da Educação Básica. **Inep**, 2024. Disponível em: <<https://www.gov.br/inep/pt-br/acao-a-informacao/dados-abertos/inep-data/estatisticas-censo-escolar>>. Acesso em: 06 set. 2024.

Exame Nacional do Ensino Médio. 2020. **Inep**, 2020. Disponível em: <<https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem/historico>>. Acesso em: 4 abr. 2024.

FACELI, Katti, *et al.* **Inteligência Artificial: uma abordagem de Aprendizado de Máquina**. 2. ed. Rio de Janeiro: LTC, 2024.

FILATRO, Andrea. **Data Science na Educação: presencial, a distância e corporativa**. 1. ed. São Paulo: Saraiva Educação, 2021.

FONSECA, George H. G. **Mineração de Dados: Associação II**. UFOP. 2024. Notas de Aula. 2024. Disponível em: <https://professor.ufop.br/sites/default/files/george/files/a15_mineracao_de_dados_-_associacao_ii.pdf>. Acesso em: 2 out. 2024.

GABRIEL, Martha. **Inteligência Artificial: do zero ao metaverso**. 1. ed. Barueri: Atlas, 2024.

GIL, Antonio Carlos. **Como elaborar projetos de pesquisa**. 7. ed. Barueri/SP: Atlas, 2022.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel; BEZERRA, Eduardo. **Data mining: conceitos, técnicas, algoritmos, orientações e aplicações**. Rio de Janeiro: Elsevier, 2015.

GOMES, Tancicleide; GOUVEIA, Roberta; BATISTA, Maria Conceição. Dados Educacionais Abertos: Associações em dados dos inscritos do Exame Nacional do Ensino Médio. In: Workshop de Informática na Escola (WIE), 23. , 2017, Recife. **Anais [...]**. Porto Alegre: Sociedade Brasileira de Computação, 2017 . p. 895-904. Disponível em: <<https://doi.org/10.5753/cbie.wie.2017.895>>. Acesso em: 2 set. 2024.

Google Colaboratory. Disponível em: <<https://colab.research.google.com/>>. Acesso em: 3 mar. 2024.

HAN, J.; PEI, J.; YIN, Y. *Mining frequent patterns without candidate generation*. **Newsletter ACM SIGMOD Record**, v. 29, n. 2, 2000, p. 1-12. Disponível em: <<https://doi.org/10.1145/335191.335372>>. Acesso em: 08 jul. 2024.

HUA, Chew C. **Inteligência Artificial, Análise e Ciência de Dados: Conceitos Fundamentais e Modelos**. São Paulo: Cengage Learning Brasil, 2024.

Ideb: Maranhão avança nos anos iniciais do ensino fundamental. **MEC**, 2024. Disponível em: <<https://www.gov.br/mec/pt-br/assuntos/noticias/2024/agosto/ideb-maranhao-avanca-nos-anos-iniciais-do-ensino-fundamental>>. Acesso em: 06 set. 2024.

Informe Macroeconômico, ano 4, n.139, 10 a 14 jun. 2024. **Banco do Nordeste do Brasil (BNB)**. Disponível em: <https://bnb.gov.br/s482-dspace/bitstream/123456789/2003/1/2024_INMA_139.pdf>. Acesso em: 06 set. 2024.

Instituto Brasileiro de Geografia e Estatística. **IBGE**, 2021; 2023. Disponível em: <<https://cidades.ibge.gov.br/brasil/ma/panorama>>. Acesso em: 06 set. 2024.

Introdução à *Machine Learning*. UFSM, 2021. Disponível em: <<https://ufsm.br/r-791-3117>>. Acesso em: 11 maio 2024.

Jupyter Notebook – versão 3.11.7. *Anaconda*. Disponível em: <<https://www.anaconda.com/>>. Acesso em: 3 mar. 2024.

LIMA, Cayo César Viana de; BRIGHENTI, Carla Regina Guimarães. Desempenho de Estudantes de Minas Gerais no Exame Nacional do Ensino Médio considerando variáveis socioeconômicas. **Educação E Pesquisa**, vol. 49, dez. 2023, p. e253303. Disponível em <<https://doi.org/10.1590/S1678-4634202349253303>>. Acesso em 16 maio. 2024.

MALHOTRA, Naresh K. **Pesquisa de marketing**: uma orientação aplicada. 6. ed. Porto Alegre: Bokman, 2012.

MARCONI, Marina de Andrade; LAKATOS, Eva Maria. **Metodologia Científica**. 8. ed. Barueri: Atlas, 2022.

MARTELETTO, Sergio Reinaldo. **Técnicas de seleção de atributos através de *Random Forests***: um estudo de caso para detecção de tendências em séries temporais financeiras. 2022. 85 f. Dissertação (Mestrado em Ciências) - Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2022. Disponível em: <https://www.teses.usp.br/teses/disponiveis/100/100132/tde-08032023-134543/publico/dissertacao_mestrado_SM-versao_corrigida.pdf>. Acesso em: 15 dez. 2024.

MATTOS, Viviane Leite Dias de; KONRATH, Andréa C.; AZAMBUJA, Ana Maria Volkmer de. **Introdução à Estatística**: aplicações em ciências exatas. 1. ed. Rio de Janeiro: LTC, 2017.

MEC e Inep divulgam resultados do Censo Escolar 2023. **Inep**, 2024. Disponível em: <<https://www.gov.br/inep/pt-br/assuntos/noticias/censo-escolar/mec-e-inep-divulgam-resultados-do-censo-escolar-2023>>. Acesso em: 06 set. 2024.

MESSIAS, J. R. A.; LÚCIO, M. de L. Gestão e governança pública para o desenvolvimento - educação básica orientada por evidências. **REVISTA DA AGU**, [S. l.], v. 22, n. 03, 2023. DOI: 10.25109/2525-328X.v.22.n.03.2023.3306. Disponível em: <<https://revistaagu.agu.gov.br/index.php/AGU/article/view/3306>>. Acesso em: 06 set. 2024.

Microdados do Enem 2023. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira – **Inep**. Brasília: Inep, 2024. Disponível em: <<https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem>>. Acesso em: 30 abr. 2024.

MORETTIN, Pedro Alberto; SINGER, Julio da Motta. **Estatística e Ciência de Dados**. 1. ed. Rio de Janeiro: LTC, 2022.

MORETTIN, Pedro A.; BUSSAB, Wilton de Oliveira. **Estatística básica**. 10. ed. São Paulo: Saraiva Uni, 2023.

NERY, Carmen. Censo 2022: Taxa de analfabetismo cai de 9,6% para 7,0% em 12 anos, mas desigualdades persistem. **Agência IBGE Notícias**, 17 maio 2024. Disponível em: <agenciadenoticias.ibge.gov.br/agencia-noticias/2012-agencia-de-noticias/noticias/40098-censo-2022-taxa-de-analfabetismo-cai-de-9-6-para-7-0-em-12-anos-mas-desigualdades-persistem>. Acesso em 20 set. 2024.

NETTO, Amilcar; MACIEL, Francisco. **Python para Data Science e Machine Learning**: Descomplicado. Rio de Janeiro: Alta Books, 2021.

ONU BRASIL. Os Objetivos de Desenvolvimento Sustentável no Brasil. **Nações Unidas Brasil**. 2024. Disponível em: <<https://brasil.un.org/pt-br/sdgs>>. Acesso em: 20 set. 2024.

Participantes do Pé-de-Meia podem fazer o Enem de graça. **Inep**, 2024. Disponível em: <<https://www.gov.br/inep/pt-br/assuntos/noticias/enem/participantes-do-pe-de-meia-podem-fazer-o-enem-de-graca>>. Acesso em: 12 dez. 2024.

PIB do Maranhão foi o que mais cresceu no Nordeste, aponta Banco do Brasil. **Governo do Maranhão**. Disponível em: <<https://www.ma.gov.br/noticias/pib-do-maranhao-foi-o-que-mais-cresceu-no-nordeste-aponta-banco-do-brasil>>. Acesso em: 06 set. 2024.

RAMOS, Jorge Luis Cavalcanti; RODRIGUES, Rodrigo Lins; SILVA, João Carlos Sedraz; OLIVEIRA, Pamella Letícia Silva de. CRISP-EDM: uma proposta de adaptação do Modelo CRISP-DM para mineração de dados educacionais. In: Simpósio Brasileiro de Informática na Educação (SBIE), 31., 2020. **Anais eletrônicos** [...]. Porto Alegre: Sociedade Brasileira de Computação, 2020. p. 1092-1101. Disponível em: <<https://doi.org/10.5753/cbie.sbie.2020.1092>>. Acesso em: 25 abr. 2024.

RAUTENBERG, Sandro; CARMO, Paulo Ricardo V. Big Data e Ciência de Dados: complementariedade conceitual no processo de tomada de decisão. **Brazilian Journal of Information Science: Research Trends**, São Paulo, v. 13, n. 1, p. 56-67, mar. 2019. Disponível em <doi.org/10.36311/1981-1640.2019.v13n1.06.p56>. Acesso em: 15 abr. 2024.

ROMERO, C.; VENTURA, S. *Data mining in education*. **Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery**, v. 3, n. 1, p. 12–27, jan./fev., 2013. Disponível em: <<https://doi.org/10.1002/widm.1075>>. Acesso em: 26 abr. 2024

RUSSELL, Stuart J.; NORVIG, Peter. **Inteligência Artificial: uma abordagem moderna**. 4. ed. Rio de Janeiro: LTC, 2022.

SAYAK, Paul. *Python Feature Selection Tutorial: A Beginner's Guide*. **Datacamp**, 2020. Disponível em: <<https://www.datacamp.com/tutorial/feature-selection-python>>. Acesso em: 12 dez. 2024.

SHARDA, Ramesh; DELEN, Dursun; TURBAN, Efraim. **Business Intelligence e Análise de Dados para Gestão do Negócio**. Porto Alegre: Bookman, 2019.

SILVA, Leandro Augusto; PERES, Sarajane Marques; BOSCARIOLI, Clodis. **Introdução à mineração de dados: com aplicações em R**. 1. ed. Rio de Janeiro: Elsevier, 2016.

SILVA, Leandro A.; MORINO, Anderson Hideki; SATO, Thiago Massahiro Conti. Prática de mineração de dados no Exame Nacional do Ensino Médio. In: Anais dos Workshops do Congresso Brasileiro de Informática na Educação (WCBIE), 3. , 2014. **Anais** [...]. Porto Alegre: Sociedade Brasileira de Computação, 2014. p. 651. Disponível em: <<http://milanesa.ime.usp.br/rbie/index.php/wcbie/article/view/3289>>. Acesso em: 2 set. 2024.

SILVA, T. A. C. da; GUEDES, L. A. Uma Abordagem para Geração e Visualização de Regras de Associação de Acesso a Conteúdos de Portal de Notícias. In: XVI Congresso Brasileiro de Inteligência Computacional (CBIC 2023), Salvador, Brasil, 2023. **Anais**. Disponível em: <https://sbic.org.br/eventos/cbic_2023/cbic2023-039/>. Acesso em 12 set. 2024.

SILVA, Vinicius Alberto Alves da; MORENO, Lorenza Leão Oliveira; GONÇALVES, Luciana Brugiolo; SOARES, Stênio São Rosário Furtado; SOUZA JÚNIOR, Robson Rocha. Identificação de Desigualdades Sociais a partir do desempenho dos alunos do Ensino Médio no ENEM 2019 utilizando Mineração de Dados. In: Simpósio Brasileiro de Informática na Educação (SBIE), 31., 2020, Online. **Anais** [...]. Porto Alegre: Sociedade Brasileira de Computação, 2020 . p. 72-81. Disponível em: <<https://doi.org/10.5753/cbie.sbie.2020.72>>. Acesso em: 2 set. 2024.

Sinopses Estatísticas do Exame Nacional do Ensino Médio. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira – **Inep**. Brasília: Inep, 2024. Disponível em: <<https://www.gov.br/inep/pt-br/aceso-a-informacao/dados-abertos/sinopses-estatisticas/enem>>. Acesso em: 30 abr. 2024.

SOARES, R. de C.; WEBER NETO, N.; COUTINHO, L. R.; SANTOS, D. V. dos; SILVA, F. J. da S. e; TELES, A. S. Mineração de dados da educação básica brasileira usando as bases do INEP: Uma revisão sistemática da literatura. **Revista Novas Tecnologias na Educação (RENOTE)**, Porto Alegre, v. 19, n. 1, p. 361–370, 2021. DOI: 10.22456/1679-1916.118526. Disponível em: <<https://seer.ufrgs.br/index.php/renote/article/view/118526>>. Acesso em: 5 set. 2024.

SOARES, R. de C.; WEBER NETO, N.; COUTINHO, L. R.; SANTOS, D. V. dos; SILVA, F. J. da S. e; TELES, A. S. Minerando Dados para Entender os Fatores de Influência da Qualidade Educacional do Maranhão. **Revista Brasileira de Informática na Educação**, v. 31, p. 378–406, 2023. Disponível em: <<https://journals-sol.sbc.org.br/index.php/rbie/article/view/2831>>. Acesso em: 16 maio. 2024.

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. **Introduction to Data Mining**. Boston: Pearson, 2009.

TECH AND MAN. *FP Growth Algorithm: Association Rule Mining*. **YouTube**, 17 de mar. de 2020. 17min53s. Disponível em: <www.youtube.com/watch?v=GcgfSJAaBto>. Acesso em: 2 out. 2024.

TRAVITZKI, Rodrigo. Possíveis contribuições do Enem para a democratização do acesso à educação superior no Brasil. **Em Aberto**, v. 34, n. 112, 2021. Disponível em <<https://doi.org/10.24109/2176-6673.emaberto.34i112.4993>>. Acesso em: 20 maio. 2024.

3,9 milhões estão inscritos no Enem 2023. **Inep**, 2023. Disponível em: <<https://www.gov.br/inep/pt-br/assuntos/noticias/enem/3-9-milhoes-estao-inscritos-no-enem-2023>>. Acesso em: 15 abr. 2024.

VALDATI, Aline de Brittos. **Inteligência artificial - IA**. 1. ed. São Paulo: Contentus, 2020.

ANEXO – DICIONÁRIO DE MICRODADOS DO ENEM 2023

DICIONÁRIO DE VARIÁVEIS - ENEM 2023					
NOME DA VARIÁVEL	Descrição	Variáveis Categóricas		Tamanho	Tipo
		Categoria	Descrição		
DADOS DO PARTICIPANTE					
NU_INSCRICAO	Número de inscrição ¹			12	Numérica
NU_ANO	Ano do Enem			4	Numérica
TP_FAIXA_ETARIA	Faixa etária ²	1	Menor de 17 anos	2	Numérica
		2	17 anos		
		3	18 anos		
		4	19 anos		
		5	20 anos		
		6	21 anos		
		7	22 anos		
		8	23 anos		
		9	24 anos		
		10	25 anos		
		11	Entre 26 e 30 anos		
		12	Entre 31 e 35 anos		
		13	Entre 36 e 40 anos		
		14	Entre 41 e 45 anos		
		15	Entre 46 e 50 anos		
		16	Entre 51 e 55 anos		
		17	Entre 56 e 60 anos		
		18	Entre 61 e 65 anos		
		19	Entre 66 e 70 anos		
		20	Maior de 70 anos		
TP_SEXO	Sexo	M	Masculino	1	Alfanumérica
		F	Feminino		
TP_ESTADO_CIVIL	Estado Civil	0	Não informado	1	Numérica
		1	Solteiro(a)		
		2	Casado(a)/Mora com companheiro(a)		
		3	Divorciado(a)/Desquitado(a)/Separado(a)		
		4	Viúvo(a)		
TP_COR_RACA	Cor/raça	0	Não declarado	1	Numérica
		1	Branca		
		2	Preta		
		3	Parda		
		4	Amarela		
		5	Indígena		
		6	Não dispõe da informação		
TP_NACIONALIDADE	Nacionalidade	0	Não informado	1	Numérica

		1	Brasileiro(a)		
		2	Brasileiro(a) Naturalizado(a)		
		3	Estrangeiro(a)		
		4	Brasileiro(a) Nato(a), nascido(a) no exterior		
TP_ST_CONCLUSAO	Situação de conclusão do Ensino Médio	1	Já concluí o Ensino Médio	1	Numérica
		2	Estou cursando e concluirei o Ensino Médio em 2023		
		3	Estou cursando e concluirei o Ensino Médio após 2023		
		4	Não concluí e não estou cursando o Ensino Médio		
TP_ANO_CONCLUIU	Ano de Conclusão do Ensino Médio	0	Não informado	1	Numérica
		1	2022		
		2	2021		
		3	2020		
		4	2019		
		5	2018		
		6	2017		
		7	2016		
		8	2015		
		9	2014		
		10	2013		
		11	2012		
		12	2011		
		13	2010		
		14	2009		
		15	2008		
		16	2007		
		17	Antes de 2007		
TP_ESCOLA	Tipo de escola do Ensino Médio	1	Não Respondeu	1	Numérica
		2	Pública		
		3	Privada		
TP_ENSINO	Tipo de instituição que concluiu ou concluirá o Ensino Médio	1	Ensino Regular	1	Numérica
		2	Educação Especial - Modalidade Substitutiva		
IN_TREINEIRO	Indica se o inscrito fez a prova com intuito de apenas treinar seus conhecimentos ³	1	Sim	1	Numérica
		0	Não		
DADOS DA ESCOLA					
CO_MUNICIPIO_ESC	Código do município da escola			7	Numérica
	1º dígito: Região				
	1º e 2º dígitos: UF				
	3º, 4º, 5º e 6º dígitos: Município				
	7º dígito: dígito verificador				
NO_MUNICIPIO_ESC	Nome do município da escola			150	Alfanumérica

CO_UF_ESC	Código da Unidade da Federação da escola			2	Numérica
SG_UF_ESC	Sigla da Unidade da Federação da escola			2	Alfanumérica
TP_DEPENDENCIA_ADM_ESC	Dependência administrativa (Escola)	1	Federal	1	Numérica
		2	Estadual		
		3	Municipal		
		4	Privada		
TP_LOCALIZACAO_ESC	Localização (Escola)	1	Urbana	1	Numérica
		2	Rural		
TP_SIT_FUNC_ESC	Situação de funcionamento (Escola)	1	Em atividade	1	Numérica
		2	Paralisada		
		3	Extinta		
		4	Escola extinta em anos anteriores.		
DADOS DO LOCAL DE APLICAÇÃO DA PROVA					
CO_MUNICIPIO_PROVA	Código do município da aplicação da prova			7	Numérica
	1º dígito: Região				
	1º e 2º dígitos: UF				
	3º, 4º, 5º e 6º dígitos: Município				
	7º dígito: dígito verificador				
NO_MUNICIPIO_PROVA	Nome do município da aplicação da prova			150	Alfanumérica
CO_UF_PROVA	Código da Unidade da Federação da aplicação da prova			2	Alfanumérica
SG_UF_PROVA	Sigla da Unidade da Federação da aplicação da prova			2	Alfanumérica
DADOS DA PROVA OBJETIVA					
TP_PRESENCA_CN	Presença na prova objetiva de Ciências da Natureza	0	Faltou à prova	1	Numérica
		1	Presente na prova		
		2	Eliminado na prova		
TP_PRESENCA_CH	Presença na prova objetiva de Ciências Humanas	0	Faltou à prova	1	Numérica
		1	Presente na prova		
		2	Eliminado na prova		
TP_PRESENCA_LC	Presença na prova objetiva de Linguagens e Códigos	0	Faltou à prova	1	Numérica
		1	Presente na prova		
		2	Eliminado na prova		
TP_PRESENCA_MT	Presença na prova objetiva de Matemática	0	Faltou à prova	1	Numérica
		1	Presente na prova		
		2	Eliminado na prova		
CO_PROVA_CN	Código do tipo de prova de Ciências da Natureza	1221	Azul	4	Numérica
		1222	Amarela		
		1223	Rosa		
		1224	Cinza		
		1225	Rosa - Ampliada		
		1226	Rosa - Superampliada		

		1227	Laranja - Braile		
		1228	Laranja - Adaptada Ledor		
		1229	Verde - Videoprova - Libras		
		1301	Azul (Reaplicação)		
		1302	Amarela (Reaplicação)		
		1303	Cinza (Reaplicação)		
		1304	Rosa (Reaplicação)		
CO_PROVA_CH	Código do tipo de prova de Ciências Humanas	1191	Azul	4	Numérica
		1192	Amarela		
		1193	Branca		
		1194	Rosa		
		1195	Rosa - Ampliada		
		1196	Rosa - Superampliada		
		1197	Laranja - Braile		
		1198	Laranja - Adaptada Ledor		
		1199	Verde - Videoprova - Libras		
		1271	Azul (Reaplicação)		
		1272	Amarela (Reaplicação)		
		1273	Branca (Reaplicação)		
		1274	Rosa (Reaplicação)		
CO_PROVA_LC	Código do tipo de prova de Linguagens e Códigos	1201	Azul	4	Numérica
		1202	Amarela		
		1203	Rosa		
		1204	Branca		
		1205	Rosa - Ampliada		
		1206	Rosa - Superampliada		
		1207	Laranja - Braile		
		1208	Laranja - Adaptada Ledor		
		1209	Verde - Videoprova - Libras		
		1281	Azul (Reaplicação)		
		1282	Amarela (Reaplicação)		
		1283	Rosa (Reaplicação)		
		1284	Branca (Reaplicação)		
CO_PROVA_MT	Código do tipo de prova de Matemática	1211	Azul	4	Numérica
		1212	Amarela		
		1213	Rosa		
		1214	Cinza		
		1215	Rosa - Ampliada		
		1216	Rosa - Superampliada		
		1217	Laranja - Braile		
		1218	Laranja - Adaptada Ledor		
		1219	Verde - Videoprova - Libras		
		1291	Azul (Reaplicação)		
		1292	Amarela (Reaplicação)		
		1293	Rosa (Reaplicação)		

		1294	Cinza (Reaplicação)		
NU_NOTA_CN	Nota da prova de Ciências da Natureza			9	Numérica
NU_NOTA_CH	Nota da prova de Ciências Humanas			9	Numérica
NU_NOTA_LC	Nota da prova de Linguagens e Códigos			9	Numérica
NU_NOTA_MT	Nota da prova de Matemática			9	Numérica
TX_RESPOSTAS_CN	Vetor com as respostas da parte objetiva da prova de Ciências da Natureza ⁴		A,B,C,D, E, * (dupla marcação), . (em branco)	45	Alfanumérica
TX_RESPOSTAS_CH	Vetor com as respostas da parte objetiva da prova de Ciências Humanas ⁴		A,B,C,D, E, * (dupla marcação), . (em branco)	45	Alfanumérica
TX_RESPOSTAS_LC	Vetor com as respostas da parte objetiva da prova de Linguagens e Códigos ⁵		A,B,C,D, E, * (dupla marcação), . (em branco), 9 (Item não apresentado)	45	Alfanumérica
TX_RESPOSTAS_MT	Vetor com as respostas da parte objetiva da prova de Matemática ⁴		A,B,C,D, E, * (dupla marcação), . (em branco)	45	Alfanumérica
TP_LINGUA	Língua Estrangeira	0	Inglês	1	Numérica
		1	Espanhol		
TX_GABARITO_CN	Vetor com o gabarito da parte objetiva da prova de Ciências da Natureza ⁶			45	Alfanumérica
TX_GABARITO_CH	Vetor com o gabarito da parte objetiva da prova de Ciências Humanas ⁶			45	Alfanumérica
TX_GABARITO_LC	Vetor com o gabarito da parte objetiva da prova de Linguagens e Códigos ⁷			50	Alfanumérica
TX_GABARITO_MT	Vetor com o gabarito da parte objetiva da prova de Matemática ⁶			45	Alfanumérica
DADOS DA REDAÇÃO					
TP_STATUS_REDACAO	Situação da redação do participante	1	Sem problemas	1	Numérica
		2	Anulada		
		3	Cópia Texto Motivador		
		4	Em Branco		
		6	Fuga ao tema		
		7	Não atendimento ao tipo textual		
		8	Texto insuficiente		
		9	Parte desconectada		
NU_NOTA_COMP1	Nota da competência 1 - Demonstrar domínio da modalidade escrita formal da Língua Portuguesa.			9	Numérica

NU_NOTA_COMP2	Nota da competência 2 - Compreender a proposta de redação e aplicar conceitos das várias áreas de conhecimento para desenvolver o tema, dentro dos limites estruturais do texto dissertativo-argumentativo em prosa.			9	Numérica
NU_NOTA_COMP3	Nota da competência 3 - Selecionar, relacionar, organizar e interpretar informações, fatos, opiniões e argumentos em defesa de um ponto de vista.			9	Numérica
NU_NOTA_COMP4	Nota da competência 4 - Demonstrar conhecimento dos mecanismos linguísticos necessários para a construção da argumentação.			9	Numérica
NU_NOTA_COMP5	Nota da competência 5 - Elaborar proposta de intervenção para o problema abordado, respeitando os direitos humanos.			9	Numérica
NU_NOTA_REDACAO	Nota da prova de redação			9	Numérica

DADOS DO QUESTIONÁRIO SOCIOECONÔMICO

Q001	Até que série seu pai, ou o homem responsável por você, estudou?	A	Nunca estudou.	1	Alfanumérica
		B	Não completou a 4ª série/5º ano do Ensino Fundamental.		
		C	Completou a 4ª série/5º ano, mas não completou a 8ª série/9º ano do Ensino Fundamental.		
		D	Completou a 8ª série/9º ano do Ensino Fundamental, mas não completou o Ensino Médio.		
		E	Completou o Ensino Médio, mas não completou a Faculdade.		
		F	Completou a Faculdade, mas não completou a Pós-graduação.		
		G	Completou a Pós-graduação.		
		H	Não sei.		
Q002	Até que série sua mãe, ou a mulher responsável por você, estudou?	A	Nunca estudou.	1	Alfanumérica
		B	Não completou a 4ª série/5º ano do Ensino Fundamental.		
		C	Completou a 4ª série/5º ano, mas não completou a 8ª série/9º ano do Ensino Fundamental.		
		D	Completou a 8ª série/9º ano do Ensino Fundamental, mas não completou o Ensino Médio.		
		E	Completou o Ensino Médio, mas não completou a Faculdade.		

		F	Completo a Faculdade, mas não completo a Pós-graduação.		
		G	Completo a Pós-graduação.		
		H	Não sei.		
Q003	A partir da apresentação de algumas ocupações divididas em grupos ordenados, indique o grupo que contempla a ocupação mais próxima da ocupação do seu pai ou do homem responsável por você. (Se ele não estiver trabalhando, escolha uma ocupação pensando no último trabalho dele).	A	Grupo 1: Lavrador, agricultor sem empregados, bóia fria, criador de animais (gado, porcos, galinhas, ovelhas, cavalos etc.), apicultor, pescador, lenhador, seringueiro, extrativista.	1	Alfanumérica
		B	Grupo 2: Diarista, empregado doméstico, cuidador de idosos, babá, cozinheiro (em casas particulares), motorista particular, jardineiro, faxineiro de empresas e prédios, vigilante, porteiro, carteiro, office-boy, vendedor, caixa, atendente de loja, auxiliar administrativo, recepcionista, servente de pedreiro, repositor de mercadoria.		
		C	Grupo 3: Padeiro, cozinheiro industrial ou em restaurantes, sapateiro, costureiro, joalheiro, torneiro mecânico, operador de máquinas, soldador, operário de fábrica, trabalhador da mineração, pedreiro, pintor, eletricista, encanador, motorista, caminhoneiro, taxista.		
		D	Grupo 4: Professor (de ensino fundamental ou médio, idioma, música, artes etc.), técnico (de enfermagem, contabilidade, eletrônica etc.), policial, militar de baixa patente (soldado, cabo, sargento), corretor de imóveis, supervisor, gerente, mestre de obras, pastor, microempresário (proprietário de empresa com menos de 10 empregados), pequeno comerciante, pequeno proprietário de terras, trabalhador autônomo ou por conta própria.		
		E	Grupo 5: Médico, engenheiro, dentista, psicólogo, economista, advogado, juiz, promotor, defensor, delegado, tenente, capitão, coronel, professor universitário, diretor em empresas públicas ou privadas, político, proprietário de empresas com mais de 10 empregados.		
		F	Não sei.		
Q004	A partir da apresentação de algumas ocupações divididas em grupos ordenados, indique o grupo que contempla a ocupação mais próxima da ocupação da sua mãe ou da mulher responsável por você. (Se ela não estiver trabalhando, escolha uma ocupação pensando no último trabalho dela).	A	Grupo 1: Lavradora, agricultora sem empregados, bóia fria, criadora de animais (gado, porcos, galinhas, ovelhas, cavalos etc.), apicultora, pescadora, lenhadora, seringueira, extrativista.	2	Numérica
		B	Grupo 2: Diarista, empregada doméstica, cuidadora de idosos, babá, cozinheira (em casas particulares), motorista particular, jardineira, faxineira de empresas e prédios, vigilante, porteira, carteira, office-boy, vendedora, caixa, atendente de loja, auxiliar administrativa, recepcionista, servente de pedreiro, repositora de mercadoria.		
		C	Grupo 3: Padeira, cozinheira industrial ou em restaurantes, sapateira, costureira, joalheira, torneira mecânica, operadora de máquinas, soldadora, operária de fábrica, trabalhadora da mineração, pedreira, pintora, eletricista, encanadora, motorista, caminhoneira, taxista.		

		D	Grupo 4: Professora (de ensino fundamental ou médio, idioma, música, artes etc.), técnica (de enfermagem, contabilidade, eletrônica etc.), policial, militar de baixa patente (soldado, cabo, sargento), corretora de imóveis, supervisora, gerente, mestre de obras, pastora, microempresária (proprietária de empresa com menos de 10 empregados), pequena comerciante, pequena proprietária de terras, trabalhadora autônoma ou por conta própria.		
		E	Grupo 5: Médica, engenheira, dentista, psicóloga, economista, advogada, juíza, promotora, defensora, delegada, tenente, capitã, coronel, professora universitária, diretora em empresas públicas ou privadas, política, proprietária de empresas com mais de 10 empregados.		
		F	Não sei.		
Q005	Incluindo você, quantas pessoas moram atualmente em sua residência?	1	1, pois moro sozinho(a).	2	Numérica
		2	2		
		3	3		
		4	4		
		5	5		
		6	6		
		7	7		
		8	8		
		9	9		
		10	10		
		11	11		
		12	12		
		13	13		
		14	14		
		15	15		
		16	16		
		17	17		
		18	18		
		19	19		
		20	20		
Q006	Qual é a renda mensal de sua família? (Some a sua renda com a dos seus familiares.)	A	Nenhuma Renda	1	Alfanumérica
		B	Até R\$ 1.320,00		
		C	De R\$ 1.320,01 até R\$ 1.980,00.		
		D	De R\$ 1.980,01 até R\$ 2.640,00.		
		E	De R\$ 2.640,01 até R\$ 3.300,00.		
		F	De R\$ 3.300,01 até R\$ 3.960,00.		
		G	De R\$ 3.960,01 até R\$ 5.280,00.		
		H	De R\$ 5.280,01 até R\$ 6.600,00.		
		I	De R\$ 6.600,01 até R\$ 7.920,00.		
		J	De R\$ 7.920,01 até R\$ 9.240,00.		
		K	De R\$ 9.240,01 até R\$ 10.560,00.		

		L	De R\$ 10.560,01 até R\$ 11.880,00.		
		M	De R\$ 11.880,01 até R\$ 13.200,00.		
		N	De R\$ 13.200,01 até R\$ 15.840,00.		
		O	De R\$ 15.840,01 até R\$19.800,00.		
		P	De R\$ 19.800,01 até R\$ 26.400,00.		
		Q	Acima de R\$ 26.400,00.		
Q007	Em sua residência trabalha empregado(a) doméstico(a)?	A	Não.	1	Alfanumérica
		B	Sim, um ou dois dias por semana.		
		C	Sim, três ou quatro dias por semana.		
		D	Sim, pelo menos cinco dias por semana.		
Q008	Na sua residência tem banheiro?	A	Não.	1	Alfanumérica
		B	Sim, um.		
		C	Sim, dois.		
		D	Sim, três.		
		E	Sim, quatro ou mais.		
Q009	Na sua residência tem quartos para dormir?	A	Não.	1	Alfanumérica
		B	Sim, um.		
		C	Sim, dois.		
		D	Sim, três.		
		E	Sim, quatro ou mais.		
Q010	Na sua residência tem carro?	A	Não.	1	Alfanumérica
		B	Sim, um.		
		C	Sim, dois.		
		D	Sim, três.		
		E	Sim, quatro ou mais.		
Q011	Na sua residência tem motocicleta?	A	Não.	1	Alfanumérica
		B	Sim, uma.		
		C	Sim, duas.		
		D	Sim, três.		
		E	Sim, quatro ou mais.		
Q012	Na sua residência tem geladeira?	A	Não.	1	Alfanumérica
		B	Sim, uma.		
		C	Sim, duas.		
		D	Sim, três.		
		E	Sim, quatro ou mais.		
Q013	Na sua residência tem freezer (independente ou segunda porta da geladeira)?	A	Não.	1	Alfanumérica
		B	Sim, um.		
		C	Sim, dois.		
		D	Sim, três.		
		E	Sim, quatro ou mais.		
Q014	Na sua residência tem máquina de lavar roupa? (o tanquinho NÃO deve ser considerado)	A	Não.	1	Alfanumérica
		B	Sim, uma.		
		C	Sim, duas.		
		D	Sim, três.		
		E	Sim, quatro ou mais.		

Q015	Na sua residência tem máquina de secar roupa (independente ou em conjunto com a máquina de lavar roupa)?	A	Não.	1	Alfanumérica
		B	Sim, uma.		
		C	Sim, duas.		
		D	Sim, três.		
		E	Sim, quatro ou mais.		
Q016	Na sua residência tem forno micro-ondas?	A	Não.	1	Alfanumérica
		B	Sim, um.		
		C	Sim, dois.		
		D	Sim, três.		
		E	Sim, quatro ou mais.		
Q017	Na sua residência tem máquina de lavar louça?	A	Não.	1	Alfanumérica
		B	Sim, uma.		
		C	Sim, duas.		
		D	Sim, três.		
		E	Sim, quatro ou mais.		
Q018	Na sua residência tem aspirador de pó?	A	Não.	1	Alfanumérica
		B	Sim.		
Q019	Na sua residência tem televisão em cores?	A	Não.	1	Alfanumérica
		B	Sim, uma.		
		C	Sim, duas.		
		D	Sim, três.		
		E	Sim, quatro ou mais.		
Q020	Na sua residência tem aparelho de DVD?	A	Não.	1	Alfanumérica
		B	Sim.		
Q021	Na sua residência tem TV por assinatura?	A	Não.	1	Alfanumérica
		B	Sim.		
Q022	Na sua residência tem telefone celular?	A	Não.	1	Alfanumérica
		B	Sim, um.		
		C	Sim, dois.		
		D	Sim, três.		
		E	Sim, quatro ou mais.		
Q023	Na sua residência tem telefone fixo?	A	Não.	1	Alfanumérica
		B	Sim.		
Q024	Na sua residência tem computador?	A	Não.	1	Alfanumérica
		B	Sim, um.		
		C	Sim, dois.		
		D	Sim, três.		
		E	Sim, quatro ou mais.		
Q025	Na sua residência tem acesso à Internet?	A	Não.	1	Alfanumérica
		B	Sim.		

1. Referente ao Enem 2023, trata-se de uma máscara e não o seu número de inscrição original no Enem. O mesmo NU_INSCRICAO para anos diferentes não identifica o mesmo participante no exame, não permite o acesso aos dados cadastrais como nome, endereço, RG etc, nem identifica o mesmo participante em microdados de pesquisas diferentes.

2. A partir da Idade do inscrito em 31/12/2023.

3. Foi considerado treineiro o inscrito que não havia concluído o ensino médio e não o concluiria em 2023

4. As 45 primeiras posições deste campo são referentes as respectivas respostas. O asterisco (*) indica dupla marcação e o ponto (.)

resposta em branco.

5. As 45 primeiras posições deste campo são referentes as respectivas respostas, das quais as 5 primeiras correspondem a parte de língua estrangeira. O asterisco (*) indica dupla marcação e o ponto (.) resposta em branco.

6. As 45 primeiras posições deste campo são referentes ao respectivo gabarito

7. As 50 primeiras posições deste campo são referentes ao respectivo gabarito, das quais, para as 10 primeiras, as 5 primeiras correspondem à prova de Língua Inglesa e as outras 5 à prova de Língua Espanhola.

Fonte: Inep, 2024.