

Análise de perfis de estudantes no ENEM considerando hábitos de estudo

Analysis of student profiles in ENEM considering study habits

Janderson Ferreira Dutra¹, Damires Yluska de Souza Fernandes¹

¹Instituto Federal da Paraíba (IFPB) - Av. Primeiro de Maio, 720, Jaguaribe, João Pessoa - PB

{janderson.dutra, damires}@ifpb.edu.br

<https://orcid.org/0000-0003-4601-1461>, <https://orcid.org/0000-0001-9305-5323>

Resumo. O Exame Nacional do Ensino Médio (ENEM) constitui um relevante indicador do desempenho dos estudantes da Educação Básica no Brasil. Durante a pandemia de COVID-19, além de fatores tradicionalmente associados ao desempenho acadêmico, como renda familiar e escolaridade dos pais, aspectos comportamentais, a exemplo dos hábitos de estudo, podem ter exercido influência significativa nos resultados. Este trabalho analisa dados de estudantes participantes do ENEM 2022, com foco nesses hábitos, cujas informações foram disponibilizadas pela primeira vez em 2023. Como principal contribuição, o estudo identifica, por meio de técnicas de aprendizado de máquina não supervisionado, perfis distintos de estudantes e evidencia como diferentes padrões de hábitos de estudo se relacionam com o desempenho. Os resultados mostram que a gestão do tempo e a regularidade nos estudos impactaram positivamente o desempenho, mesmo diante de limitações no acesso a recursos tecnológicos. Conclui-se que hábitos de estudo consistentes, aliados a condições mínimas de acesso e a fatores socioeconômicos, desempenham papel central no desempenho dos estudantes.

Palavras-chave: Pandemia, hábitos de estudo, desempenho estudantil, perfis de estudantes, aprendizado não supervisionado.

Abstract. The National High School Examination (ENEM) is a relevant indicator of the academic performance of students in Basic Education in Brazil. During the COVID-19 pandemic, in addition to factors traditionally associated with academic performance, such as family income and parents' educational level, behavioral aspects, particularly study habits, may have significantly influenced students' results. This study analyzes data from students who participated in the 2022 ENEM, focusing on these habits, whose information was made available for the first time in 2023. As its main contribution, the study identifies, through unsupervised machine learning techniques, distinct student profiles and highlights how different patterns of study habits are related to performance. The results show that time management and study regularity positively impacted performance, even in the presence of limited access to technological resources. It is concluded that consistent study habits, combined with minimum access conditions and socioeconomic factors, play a central role in student performance.

Keywords: Pandemic, study habits, student performance, student profiles, unsupervised learning.

1. Introdução

O ENEM, criado pelo Ministério da Educação (MEC) para avaliar conhecimentos e habilidades desenvolvidos na Educação Básica, consolidou-se como o principal meio de acesso ao Ensino Superior no Brasil e como uma relevante fonte de dados sobre estudantes da Educação Básica (INEP, 2022).

Os microdados do ENEM são disponibilizados publicamente e possibilitam a extração de informações que subsidiam gestores e educadores no diagnóstico de problemas educacionais e no planejamento de estratégias para melhoria da infraestrutura e da qualidade do ensino. Esses dados permitem investigar fatores escolares e socioeconômicos associados ao de-

sempenho dos estudantes no exame (Argôlo, 2017).

A pandemia de COVID-19 provocou mudanças significativas na rotina de estudos, em razão do isolamento social e da suspensão das aulas presenciais (Kissler et al., 2020). A adoção emergencial do ensino remoto levantou questionamentos sobre seus impactos no desempenho no ENEM, sobretudo diante da falta de preparo de instituições e docentes, bem como das limitações enfrentadas pelos estudantes quanto ao acesso à internet e a equipamentos tecnológicos (Soares; Silva, 2020). Embora o ensino remoto tenha permitido a continuidade das atividades educacionais, também evidenciou desigualdades de acesso à tecnologia e possíveis lacunas no aprendizado, comprometendo princípios de equidade e qualidade educacional (Cunha; Silva; Silva, 2020).

Antes da pandemia, estudos já apontavam fatores associados ao desempenho dos estudantes no ENEM (Silva Filho; Adeodato, 2019; Sousa Lima et al., 2020; Franco et al., 2020). Em estudo anterior (Dutra; Firmino Júnior; Fernandes, 2023), esses fatores foram sistematizados, destacando-se renda familiar, idade, sexo, raça, tipo e dependência administrativa da escola, localização, escolaridade dos pais e notas por área de conhecimento como os mais frequentemente relacionados ao desempenho dos estudantes.

Na edição do ENEM de 2022, além dos microdados tradicionais, o INEP publicou um conjunto adicional de dados sobre hábitos de estudo, coletados por meio de um questionário voltado à compreensão da rotina de estudos e da preparação dos estudantes durante a pandemia de COVID-19 (INEP, 2022). A disponibilização desses dados ampliou as possibilidades de investigação sobre fatores associados ao desempenho de estudantes no ENEM, permitindo analisar, de forma integrada, aspectos socioeconômicos e comportamentais relacionados aos hábitos de estudo. Nesse contexto, este trabalho busca responder à seguinte questão de pesquisa: *O que a análise integrada dos dados do ENEM e dos hábitos de estudo dos estudantes durante a pandemia revela sobre perfis de desempenho e fatores associados aos resultados obtidos no exame?*

Para isso, este trabalho apresenta uma abordagem baseada em métodos de aprendizado de máquina não supervisionado, com o objetivo de identificar perfis de desempenho de estudantes e suas associações com hábitos de estudo. A análise utiliza os microdados do ENEM 2022, cujas provas ocorreram durante a pandemia, abrangendo atributos acadêmicos, socioeconômicos e comportamentais, previamente preparados para a identificação de padrões e associações relevantes.

Apesar do avanço de estudos sobre o impacto do ensino remoto no desempenho acadêmico, ainda há uma lacuna na compreensão de como fatores como gestão do tempo, condições de permanência e acesso à infraestrutura e recursos tecnológicos se relacionam no contexto de desigualdades educacionais. Nesse sentido, torna-se relevante descrever e compreender esses aspectos de forma integrada, especialmente no cenário pós-pandêmico.

O artigo está organizado da seguinte forma: a Seção 2 introduz alguns conceitos fundamentais; a Seção 3 apresenta a abordagem proposta para a investigação; a Seção 4 discute os resultados obtidos; a Seção 5 aborda trabalhos relacionados; e na Seção 6, são tecidas algumas considerações e indicados trabalhos futuros.

2. Referencial teórico

Nesta seção, são introduzidos alguns conceitos relacionados aos temas abordados neste estudo.

2.1. Hábitos de estudo e autorregulação da aprendizagem no contexto do ENEM

Em 2022, foi introduzido o questionário *Hábitos de estudo dos participantes do Enem em contexto de pandemia*, composto por 34 questões destinadas a levantar informações sobre rotinas e estratégias de preparação durante o segundo ano da pandemia (INEP, 2023). Dos 2.467.086 estudantes que consultaram o local de prova, 928.564 (37,6%) responderam ao menos uma questão do instrumento.

Os dados do questionário foram organizados pelo INEP (2023) em cinco Grupos Temáticos (GT): (i) matrícula escolar e continuidade dos estudos; (ii) gestão do tempo e planejamento das atividades de estudo; (iii) práticas de estudo e pesquisa; (iv) tecnologias, acesso remoto e problemas na rotina de estudos; e (v) dificuldades de infraestrutura, necessidade de apoio externo e avaliação da experiência de estudo no período pandêmico.

No tocante ao ensino remoto, os GTs estão relacionados ao conceito de autorregulação da aprendizagem, que envolve o planejamento, monitoramento e avaliação do próprio estudo (Zimmerman, 2000). Estudos indicam que maiores níveis de autorregulação estão associados a estratégias de estudo mais consistentes, sendo esse processo também influenciado pelo acesso a recursos tecnológicos e pelas condições de ensino (Nascimento et al., 2025). Nesse cenário, essa capacidade torna-se mais relevante, pois os estudantes dependem mais de sua autonomia para organizar rotinas, gerir o tempo e manter o engajamento acadêmico.

2.2. Mineração de dados educacionais e tarefas não supervisionadas

A Mineração de Dados é a descoberta de conhecimento em função de modelos ou padrões a partir de dados (Zaki; Meira Junior, 2020; Elmasri; Navathe, 2011). Quando os dados possuem origem ou fonte educacional, o processo pode ser denominado de Mineração de Dados Educacionais (MDE) (Ramos et al., 2020).

No contexto da MDE, destaca-se o modelo de processo CRISP-EDM (acrônimo de *Cross-Industry Standard Process for Educational Data Mining*), extensão do CRISP-DM, amplamente adotado para orientar projetos de descoberta de conhecimento em educação. O CRISP-EDM envolve as seguintes etapas: (i) entendimento do domínio em educação; (ii) entendimento dos dados educacionais; (iii) preparação dos dados para MDE; (iv) modelagem; (v) avaliação dos modelos; e (vi) implantação da solução educacional.

O Aprendizado de Máquina (AM) fornece a base técnica para a extração de modelos e

padrões na mineração de dados, sendo classificado em abordagens supervisionadas e não supervisionadas (Alpaydin, 2014). As tarefas supervisionadas incluem classificação e regressão, enquanto as não supervisionadas abrangem técnicas como agrupamento e regras de associação (Faceli et al., 2021).

As estratégias de agrupamento (*clustering*) podem ser hierárquicas, partitivas ou baseadas em densidade (Han; Kamber; Pei, 2012). Entre os métodos partitivos, destaca-se o K-means, que exige a definição prévia do número de grupos e utiliza centróides para representá-los. Esse método é amplamente adotado em estudos empíricos devido à sua eficiência computacional e simplicidade de implementação, embora apresente sensibilidade à presença de ruídos e à escolha inicial dos centróides.

A mineração de regras de associação tem como objetivo identificar tendências que auxiliem na compreensão e explicação de padrões de comportamento em conjuntos de dados (Agrawal et al., 1996). Esse processo envolve, de modo geral, duas etapas: a identificação de conjuntos de itens frequentes, a partir de limites mínimos definidos pelo usuário, e a geração das regras de associação com base em medidas de avaliação preestabelecidas (Faceli et al., 2021). As regras de associação são expressas na forma $X \rightarrow Y$, indicando que a ocorrência do conjunto de itens X (antecedente) está associada à ocorrência de Y (consequente), sendo ambos compostos por itens disjuntos. A relevância dessas regras é avaliada por medidas como suporte, confiança e lift, amplamente utilizadas na literatura (Agrawal et al., 1996).

O suporte indica a frequência de ocorrência de um conjunto de itens no conjunto de dados, enquanto a confiança expressa a força da regra, ao indicar a probabilidade de ocorrência do consequente dado o antecedente. O lift, por sua vez, avalia o grau de dependência entre antecedente e consequente, indicando se a coocorrência dos itens é maior ou menor do que a esperada ao acaso. Com o objetivo de evitar a geração de um número excessivo de regras, são adotados limiares mínimos (*minimum support threshold*) para essas medidas, restringindo a análise às associações mais relevantes (Alpaydin, 2014; Hipp; Güntzer; Nakhaeizadeh, 2000).

Entre os algoritmos mais utilizados para essa tarefa, destaca-se o Apriori, que identifica inicialmente os conjuntos de itens frequentes e, a partir deles, gera as regras de associação consideradas significativas (Agrawal et al., 1996).

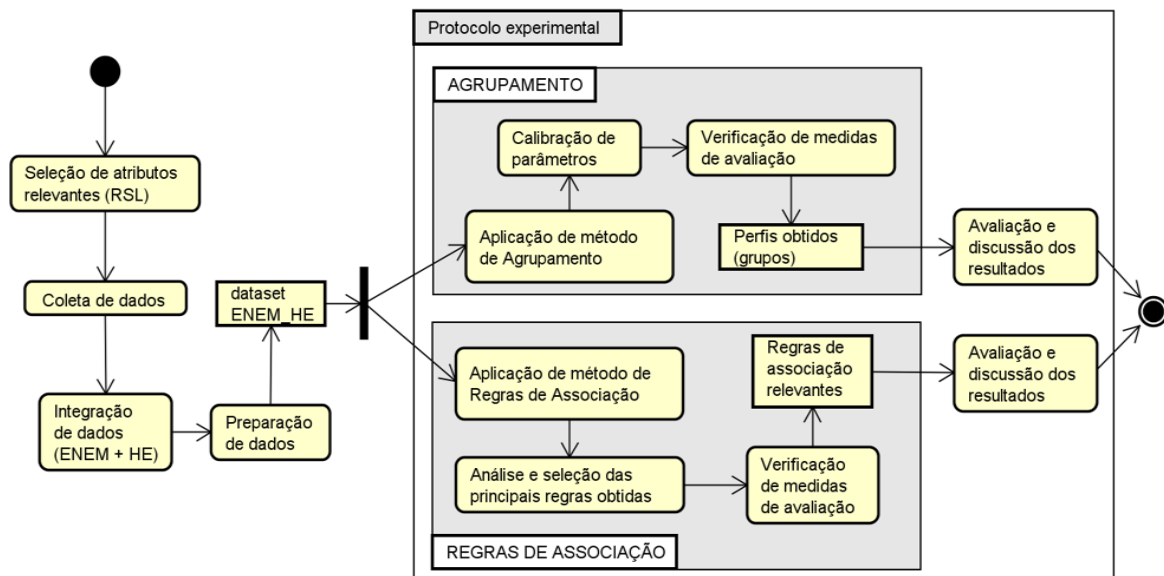
3. Metodologia

A abordagem metodológica proposta neste trabalho se baseia nas etapas definidas pelo CRISP-EDM, com adaptações ao contexto deste trabalho norteado pela questão de pesquisa formulada. Uma visão geral da abordagem proposta está representada na Figura 1.

ção dos dados, foram realizadas atividades de eliminação de instâncias duplicadas, criação de variáveis, padronização de nomes, normalização e transformação de variáveis.

O dataset gerado após a integração dos dados, denominado de ENEM_HE, foi usado como entrada nas tarefas de AM não supervisionado (agrupamento e regras de associação) definidas por meio de um protocolo experimental. Aspectos essenciais de especificação e de decisões relacionados às etapas de seleção, preparação dos dados e de construção dos modelos são descritos a seguir.

Figura 1 – Etapas da abordagem proposta



A abordagem inclui um conjunto de etapas com o intuito de obter insights sobre como os hábitos de estudo dos estudantes juntamente com fatores já comumente considerados relevantes puderam impactar em seus desempenhos no ENEM durante um período da pandemia. Para isso, busca-se identificar perfis de estudantes a partir de atributos relevantes ao entendimento do desempenho do estudante com base nos microdados comuns obtidos no ENEM e na agregação de dados relacionados aos hábitos de estudo durante a pandemia. Após a seleção dos atributos relevantes obtidos da Revisão Sistemática da Literatura (RSL) realizada no trabalho de Dutra, Firmino Júnior e Fernandes (2023), foram coletados e integrados os dois datasets (ENEM e hábitos de estudo) para o ano 2022, considerando apenas estudantes que responderam ao questionário de hábitos de estudo. Em seguida, na prepara-

3.1. Identificação de atributos mais relevantes

Os microdados do ENEM da edição 2022 são compostos por 76 atributos, organizados e distribuídos por dados sobre o(a): participante, escola, local de aplicação da prova, prova objetiva e questionário socioeconômico. O conjunto de dados possui 2.467.086 instâncias de estudantes de todo Brasil. Ressalta-se que os microdados disponibilizados pelo INEP são previamente anonimizados, não permitindo a identificação dos participantes.

A seleção de atributos relevantes para compor o dataset ENEM_HE deste estudo foi realizada em duas etapas: (i) seleção de atributos padrão do ENEM; e (ii) seleção de atributos de hábitos de estudo.

A lista de 17 atributos considerados relevantes para a análise de desempenho no ENEM, conforme identificada por Dutra, Firmino Júnior e Fernandes (2023), foi utilizada como base para a seleção de atributos (primeira etapa) na composição do dataset ENEM_HE. Esses atributos foram classificados nas seguintes categorias: dados socioeconômicos, dados de localização, notas em todas as áreas de conhecimento e perfil das escolas. Cada atributo recebeu uma contagem correspondente ao número de estudos que o identificaram como relevante. A Tabela 1, adaptada de Dutra, Firmino Júnior e Fernandes (2023), apresenta a distribuição dos atributos por categorias, seguida pela contagem de estudos que os utilizaram. Por exemplo, no caso da categoria Socioeconômicos, 13 trabalhos identificados na RSL fizeram uso do atributo renda familiar mensal.

Tabela 1 – Atributos relevantes por categoria de dados.

Categoria	Atributos relevantes
Socioeconômicos	renda familiar mensal (13), idade (10), sexo (10), raça (9), escolaridade dos pais (9), número de pessoas que moram no domicílio (4) e estado civil (4).
Localização	região/uf/município da escola (8), região/uf/município de domicílio do estudante (6) e região de nascimento (2).
Notas	notas por área (13), nota da redação (8), nota média (8), nota em língua estrangeira (5).
Perfil das escolas	tipo de escola (9) e tipo administrativo da escola (3).

A seleção dos atributos padrão do ENEM foi realizada com base no número de ocorrências e citações nos estudos analisados, conforme levantamento apresentado na Tabela 1. Em seguida, esses atributos foram analisados quanto à sua relação com o desempenho e com os hábitos de estudo durante o período de aulas remotas na pandemia. Os atributos selecionados foram: *tp_faixa_etaria* (idade), *tp_sexo* (sexo), *tp_cor_raca* (raça), Q001 (escolaridade do pai), Q002 (escolaridade da mãe), Q006 (renda familiar), notas por área de conhecimento, nota da redação e nota média. As notas das áreas e da redação foram incluídas por subsidiarem a construção do atributo *NOTA_MEDIA*, utilizado como indicador do desempenho geral do estudante no ENEM.

Adicionalmente, os atributos Q022 (celular na residência), Q024 (computador na residência) e Q025 (internet na residência) do questionário socioeconômico foram integrados ao dataset ENEM_HE, por sua relação direta com o acesso

às aulas remotas. Outros atributos, como *tp_dependencia_adm_esc* e *cod_uf_esc*, foram avaliados, mas excluídos devido à alta incidência de valores nulos e à ausência de relação direta com os hábitos de estudo, considerando o escopo nacional da análise.

A segunda etapa da seleção concentrou-se nos atributos de hábitos de estudo mais representativos. Para isso, foi realizada uma análise da correlação com a nota média dos estudantes. Observou-se correlação positiva forte entre o desempenho e os atributos dos grupos temáticos Práticas de estudo e pesquisa e Gestão do tempo e planejamento de estudos, enquanto correlações negativas mais expressivas foram identificadas nos grupos Matrícula escolar e Dificuldades de infraestrutura/ajuda de terceiros.

Optou-se por não utilizar subquestões dos hábitos de estudo devido à baixa relevância, ao elevado número de valores nulos e ao excesso de detalhamento. Assim, foram considerados apenas atributos agregados, como recebimento de ajuda de terceiros e dificuldades de infraestrutura.

3.2. Integração e preparação de dados

Após a coleta dos microdados gerais do ENEM e daqueles relacionados aos hábitos de estudo, os conjuntos de dados foram integrados por meio da variável comum *NU_INSCRICAO*. Embora essa variável represente uma máscara e não o número real de inscrição, sua correspondência com os microdados de hábitos de estudo confirmou que as instâncias se referem exclusivamente a estudantes participantes da edição de 2022 do ENEM (INEP, 2022). Em seguida, aplicou-se um filtro para selecionar apenas os estudantes que responderam integralmente ao questionário de hábitos de estudo. Foram removidas instâncias incompletas ou duplicadas, bem como registros de estudantes reprovados ou com nota zero em ao menos uma prova objetiva ou na redação.

Após a integração, os atributos do dataset ENEM_HE tiveram sua nomenclatura padronizada, de modo a torná-los mais representativos e permitir a identificação dos respectivos grupos temáticos. A Figura 2 mostra um fragmento do dataset, composto por 38 atributos,

sendo 9 provenientes de Dutra, Firmino Júnior e Fernandes (2023) e 29 oriundos do questionário de hábitos de estudo, ainda sem tratamento.

Figura 2 – Fragmento do dataset ENEM_HE.

TP_FAIXA_ETARIA	TP_SEXO	TP_COR_RACA	NIVEL_ESC_PAI	NIVEL_ESC_MAE	REND_FAMILIAR	NU_CELULAR	NU_COMPUTADOR	ACESSO_INTERNET	APR_MATR_SIT_MEDIO	...
0	6	F	2	E	E	B	E	A	B	C ...
1	4	F	3	E	E	B	B	A	B	C ...
2	4	M	3	E	E	E	B	A	B	A ...
3	3	F	3	E	E	G	D	C	B	A ...
4	4	F	3	E	E	C	B	A	B	A ...
...										
EST_TEMP_ASSID_AULA_ONLINE	PRAT_EST_REV_ANOT	PRAT_EST_REV_VIDEOAULA	PROB_ROT_EST_PANDEMIA	DIF_INFRA_PANDEMIA	AJUD_TERC_PANDEMIA	AUTOAV_PREPARACAO_APRENDIZ	NOTA_MEDIA			
B	B	B	B	B	B	B	B	414.10		
D	C	D	A	A	C	B	438.10			
B	D	B	A	A	A	C	605.58			
C	C	B	A	B	A	C	630.62			
B	C	B	A	A	A	B	406.86			

32793 rows x 38 columns

As variáveis categóricas foram transformadas em valores binários por meio da técnica de *one-hot encoding* (PEDREGOSA et al., 2011). As variáveis numéricas TP_FAIXA_ETARIA, TP_COR_RACA e NOTA_MEDIA foram normalizadas pelo método *MinMaxScaler*, com valores no intervalo [0,1]. Além disso, algumas variáveis, especialmente as numéricas, tiveram sua granularidade reduzida por discretização, incluindo TP_FAIXA_ETARIA, TP_COR_RACA, NIVEL_ESC_PAI, NIVEL_ESC_MAE, RENDA_FAMILIAR, NU_CELULAR, NU_COMPUTADOR e NOTA_MEDIA.

A redução da granularidade contribuiu para facilitar a interpretação dos perfis no processo de agrupamento e para reduzir a complexidade dos itens candidatos na mineração de regras de associação. Ao final da etapa de preparação, o dataset ENEM_HE resultante incluía 332.793 instâncias (46,69% do total inicial) e 136 atributos, representados por valores booleanos (0 – falso; 1 – verdadeiro), indicando a presença ou ausência de cada item.

Os dados e códigos da pesquisa estão disponíveis em acesso aberto no repositório Zenodo, sob a licença Creative Commons Attribution 4.0 International (CC BY 4.0), incluindo o dataset anonimizado em formato CSV, o dicionário de dados em PDF e um notebook Jupyter (.ipynb) com o código de análise. O acesso é realizado por meio do DOI: <https://doi.org/10.5281/zenodo.18270416>.

3.3. Protocolo experimental

O protocolo experimental compreende duas etapas voltadas à construção, avaliação e análise de modelos de aprendizado não supervisionado,

com o objetivo de responder à questão de pesquisa.

Na etapa de agrupamento, foi adotado o algoritmo K-means, em função de sua ampla utilização na literatura e de sua eficiência em conjuntos de dados de alta dimensionalidade, como o ENEM_HE. A escolha do número ideal de clusters depende também do entendimento dos dados e do problema a ser solucionado. O dataset apresenta três categorias de desempenho (baixo, médio e alto), definidas com base na nota média de cada estudante. Considerando as três categorias de desempenho definidas a partir da nota média dos estudantes, foi estabelecido $k=3$ para os experimentos. Após a aplicação do algoritmo, foi criada a variável CLUSTER para identificar os grupos de estudantes, os quais foram analisados com base nos atributos do ENEM e nos hábitos de estudo, com ênfase no desempenho no exame.

Para a tarefa de regras de associação, foi utilizado o algoritmo Apriori, escolhido com base em estudos que analisam o desempenho de estudantes no ENEM por meio dessa abordagem, evidenciando sua adequação ao contexto educacional investigado. Na etapa inicial, destinada à obtenção dos itemsets frequentes, foram analisados diferentes limiares de suporte, variando entre 20% e 100%. Observou-se que o valor de 25% apresentou um equilíbrio adequado entre representatividade e volume de itemsets, resultando em 4.291 itens frequentes

em um total de 332.793 transações do dataset ENEM_HE. Valores de suporte mais elevados levariam à redução significativa do número de itemsets, excluindo variáveis potencialmente relevantes para a análise dos perfis de estudantes.

Na etapa de geração das regras de associação, considerando a alta dimensionalidade do dataset ENEM_HE e a consequente geração de um grande volume de regras, foram definidos limiares mínimos para as métricas de suporte, confiança e lift, de modo a garantir a viabilidade da análise. Para a seleção e interpretação dos resultados, foram consideradas apenas regras com lift superior a 1, uma vez que essa medida indica correlação positiva entre antecedentes e consequentes, evidenciando associações relevantes no contexto investigado.

três grupos com características distintas associadas ao desempenho acadêmico.

4. Resultados e discussões

Esta seção apresenta os resultados obtidos a partir dos modelos construídos e da avaliação experimental realizada. Para isso, foram definidos dois experimentos, descritos a seguir. Os experimentos foram executados no ambiente *Google Colaboratory*, utilizando bibliotecas da linguagem Python amplamente empregadas em estudos de mineração e análise de dados, como *numpy*, *pandas*, *matplotlib* e *scikit-learn* (Pedregosa et al., 2011).

4.1. Experimento 1

O objetivo do Experimento 1 é identificar perfis de estudantes segundo a abordagem proposta, por meio da tarefa de agrupamento. Para isso, os atributos relevantes foram organizados por temas relacionados aos hábitos de estudo, resultando na identificação de três perfis de estudantes. O Cluster 0 corresponde a estudantes de alto desempenho (149.458 instâncias), o Cluster 1 a estudantes de desempenho médio (53.862 instâncias) e o Cluster 2 a estudantes de baixo desempenho (129.473 instâncias).

A avaliação qualitativa dos perfis considerou os dados padrão do ENEM e as variáveis de hábitos de estudo, sendo a distribuição dos clusters apresentada na Figura 3, que evidencia

Figura 3 - Distribuição de grupos obtidos.

K-MEANS (K=3)				
RSL + GRUPOS TEMÁTICOS DE HE	VARIÁVEL	CLUSTER 0	CLUSTER 1	CLUSTER 2
Atributos relevantes da RSL	TP_SEXO	femino	femino	femino
	TP_COR_RACA	pretaPardaIndigena	pretaPardaIndigena	pretaPardaIndigena
	NIVEL_ESC_PAI	nãoCompletoMédio	nãoCompletoMédio	nãoCompletoMédio
	NIVEL_ESC_MAE	completoAtéMédio	completoAtéMédio	completoAtéMédio
	REND_FAMILIAR	até1818	até1818	até1818
	NU_CELULAR	tem	tem	tem
	NU_COMPUTADOR	tem	tem	nãoTem
	ACESSO_INTERNET	tem	tem	tem
Situação de matrícula escolar e percepção da própria aprendizagem	APR_MATR_SIT_MEDIO	ensinoRegular	ensinoRegular	ensinoRegular
	APR_MATR_VINCULO	nãoInterrompeu	nãoInterrompeu	nãoInterrompeu
	APR_MATR_TP_ESTUDO	híbrido	híbrido	híbrido
	APR_MATR_APREND_PANDEMIA	aprendeuMaisPresen	aprendeuMaisPresen	aprendeuMaisPresen
Gestão do tempo e planejamento de estudos (1ª parte)	GEST_TEMP_ATV_CRONOGR	muitasVezes	nenhumaVez	poucasVezes
	GEST_TEMP_ATV_TEMPO	muitasVezes	nenhumaVez	poucasVezes
	GEST_TEMP_ATV_MATERIAL	muitasVezes	nenhumaVez	poucasVezes
	GEST_TEMP_ATV_HORA_PROG	muitasVezes	nenhumaVez	poucasVezes
Práticas de estudo e pesquisa (1ª parte)	PRAT_EST_LER	muitasVezes	nenhumaVez	poucasVezes
	PRAT_EST_RESUM_TEXTO	muitasVezes	nenhumaVez	poucasVezes
	PRAT_EST_RESUM_VIDEO	muitasVezes	nenhumaVez	poucasVezes
	PRAT_EST_ATV_FIXACAO	muitasVezes	nenhumaVez	poucasVezes
	PRAT_EST_ATV_AVALIACAO	muitasVezes	nenhumaVez	poucasVezes
	PRAT_EST_DISTRACOES	muitasVezes	nenhumaVez	poucasVezes
	PRAT_EST_ANOT_DUV_VIDEO	muitasVezes	nenhumaVez	poucasVezes
	PRAT_EST_ANOT_DUV_VIDEO_COMPL	muitasVezes	nenhumaVez	poucasVezes
	PRAT_EST_ANOT_DUV_PROF	muitasVezes	nenhumaVez	poucasVezes
	PRAT_EST_ESTRUT_IDEIA_REDACAO	muitasVezes	nenhumaVez	poucasVezes
	PRAT_EST_TREINAR_REDACAO	muitasVezes	poucasVezes	poucasVezes
	PRAT_EST_PARTICIPAR_FORUM	poucasVezes	nenhumaVez	poucasVezes
Gestão do tempo e planejamento de estudos (2ª parte)	GEST_TEMP_PONTUAL_AULA_ONLINE	muitasVezes	nenhumaVez	poucasVezes
	GEST_TEMP_ASSID_AULA_ONLINE	muitasVezes	nenhumaVez	poucasVezes
Práticas de estudo e pesquisa (2ª parte)	PRAT_EST_REV_ANOT	muitasVezes	nenhumaVez	poucasVezes
	PRAT_EST_REV_VIDEOAULA	muitasVezes	nenhumaVez	poucasVezes
Problemas na rotina de estudos	PROB_ROT_EST_PANDEMIA	sim	sim	sim
Dificuldades de infraestrutura	DIF_INFRA_PANDEMIA	não	não	não
Ajuda de terceiros	AJUD_TERC_PANDEMIA	sim	ninguémAuxiliou	sim
Avaliação sobre a própria experiência	AUTOAV_PREPARACAO_APRENDIZ	bemPreparado	poucoPreparado	poucoPreparado
DESEMPENHO	NOTA_MEDIA	alto	médio	baixo

Os valores apresentados nos Clusters 0, 1 e 2 correspondem às respostas mais frequentes dos atributos analisados e caracterizam perfis distintos de estudantes em relação ao desempenho no ENEM, classificados como alto, médio e baixo desempenho, conforme mostra a Figura 3.

Grupo 0 (alto desempenho): este grupo reúne estudantes com melhores notas. Do ponto de vista socioeconômico, os estudantes apre-

sentam maior acesso a computadores e celulares, renda familiar mais elevada e predominância do sexo feminino. Em geral, a mãe possui ensino médio completo, enquanto o pai não o concluiu.

Quanto à situação escolar, eles mantêm vínculo estável com a instituição e frequentam o ensino regular, com predominância do modelo híbrido, indicando maior adaptação a diferentes formatos.

Nos hábitos de estudo, destacam-se pela frequente utilização de estratégias de gestão do tempo e planejamento, como organização de cronogramas. Também realizam diversas práticas de estudo, como leitura, resumos, revisão e participação em atividades acadêmicas, além de apresentarem maior assiduidade. A ausência de dificuldades de infraestrutura e o apoio de terceiros durante a pandemia contribuíram para maior disponibilidade de estudo. Em termos de autopercepção, consideram-se bem preparados.

Grupo 1 (médio desempenho): composto por estudantes de desempenho intermediário. Caracteriza-se por renda familiar de até 1,5 salário mínimo e predominância de mulheres pretas, pardas ou indígenas. A escolaridade dos pais é, em geral, limitada, com apenas um deles tendo concluído o ensino médio.

Embora os estudantes deste grupo tenham acesso a dispositivos e internet, o engajamento nas atividades é menor. Mantêm vínculo escolar regular e indicam preferência pelo ensino híbrido, porém apresentam baixa frequência na organização do tempo e no planejamento dos estudos.

As práticas de estudo, como leitura, resumo e produção textual, são pouco realizadas, assim como o treino de redação. A maioria não recebeu apoio de terceiros durante a pandemia e se autoavalia como pouco preparada. Esses fatores contribuem para um desempenho intermediário.

Grupo 2 (baixo desempenho): Este grupo concentra estudantes com menores notas e maior vulnerabilidade socioeconômica, com renda de até R\$ 1.818,00. Há predominância de mulheres e estudantes de grupos raciais historicamente desfavorecidos, com baixa escolaridade dos pais.

Apesar do acesso à internet e celular, o acesso a computadores é limitado, o que pode comprometer o aprendizado. Apresentam maior dificuldade de adaptação ao ensino híbrido e maior instabilidade no vínculo escolar.

Quanto aos hábitos de estudo, há baixa frequência de planejamento e uso de estratégias eficazes, como leitura, resumos e busca por apoio docente. Também relatam menor supor-

te externo durante a pandemia e se consideram pouco preparados para o exame.

Os resultados indicam que os hábitos de estudo, especialmente relacionados à gestão do tempo, planejamento e práticas de estudo, são os principais fatores na diferenciação dos perfis de desempenho. Em contraste, variáveis como situação de matrícula, autopercepção e aspectos socioeconômicos apresentaram menor poder discriminativo entre os grupos.

4.2. Experimento 2

O Experimento 2 teve como objetivo aprofundar a análise integrada dos dados do ENEM e dos hábitos de estudo dos estudantes, por meio da extração de regras de associação que pudessem corroborar os perfis de desempenho identificados na etapa de agrupamento e evidenciassem os fatores mais fortemente associados aos resultados obtidos no exame.

A partir da aplicação de um limiar de suporte de 25% no dataset ENEM_HE, foram analisadas regras de associação no conjunto global e nos perfis de alto, médio e baixo desempenho, priorizando associações positivas ($\text{lift} > 1$). No cenário global, predominam regras associadas a variáveis socioeconômicas mais frequentes e a aspectos dos hábitos de estudo, como gestão do tempo, organização e aproveitamento da preparação durante a pandemia.

A Tabela 2 apresenta uma amostra das regras selecionadas para discussão, considerando o dataset ENEM_HE completo, indicando as regras no formato X (antecedente) $\rightarrow$ Y (consequente), bem como os respectivos valores de Confiança e Lift.

Tabela 2 - Exemplos de regras de associação obtidas por dataset completo.

DATASET COMPLETO	Conf.	Lift
GEST_TEMP_ATV_MATERIAL=poucasVezes, NU_CELULAR=tem $\square$ CLUSTER=2	75,4%	1.94
PRAT_EST_REV_ANOT=poucasVezes, APR_MATR_VINCULO=nãoInterrompeu $\square$ CLUSTER=0	79,9%	1.78
APR_MATR_VINCULO=nãoInterrompeu, NU_CELULAR=tem, NU_COMPUTADOR=nãoTem $\square$ RENDA_FAMILIAR=até1818	75,8%	1.51
TP_COR_RACA=branca, APR_MATR_SIT_MEDIO=ensinoRegular, NU_COMPUTADOR=tem, NU_CELULAR=tem $\square$ APR_MATR_VINCULO=nãoInterrompeu	96,9%	1.10
RENDA_FAMILIAR=até1818, APR_MATR_SIT_MEDIO=ensinoRegular $\square$ TP_FAIXA_ETARIA=faixaProfissional, NU_CELULAR=tem	60,4%	1.10
APR_MATR_APREND_PANDEMIA=aprendeu-MaisPresencial, GEST_TEMP_ATV_TEMPO =nenhumaVez $\square$ CLUSTER=1	63,4%	1.63
NIVEL_ESC_MAE=completouSuperior $\square$ APR_MATR_VINCULO=nãoInterrompeu, NU_CELULAR=tem, ACESSO_INTERNET=tem	93,5%	1.08
TP_SEXO=femino, PROB_ROT_EST_PANDEMIA=sim, APR_MATR_VINCULO =nãoInterrompeu, DIF_INFRA_PANDEMIA=não $\square$ ACESSO_INTERNET=tem	96,9%	1.05
APR_MATR_TP_ESTUDO=apenasRemoto, NU_CELULAR=tem, ACESSO_INTERNET=tem $\square$ PROB_ROT_EST_PANDEMIA=sim	80,5%	1.03

Considerando o dataset completo, a regra mais forte indica que estudantes que realizaram poucas vezes atividades de gestão de tempo para preparação de material, mesmo tendo o celular à disposição, apresentam praticamente o dobro de chance de pertencer ao Cluster 2 (Baixo Desempenho). Por outro lado, estudantes que realizaram com maior frequência práticas de revisão de anotações e que não interromperam os estudos durante a pandemia apresentam forte associação com o Cluster 0 (Alto Desempenho). Já aqueles que indicaram ter aprendido mais na modalidade presencial e que não realizaram atividades de gestão do tempo tendem a pertencer ao Cluster 1 (Médio Desempenho). De modo geral, observa-se que estudantes de baixo desempenho concentram-se entre aqueles com menor frequência

de práticas de gestão do tempo e maior preferência pelo ensino presencial, sendo também influenciados por fatores socioeconômicos, como renda familiar, escolaridade dos pais e acesso às aulas remotas.

Como ilustração, conforme a Tabela 2, algumas regras evidenciam a relação entre condições socioeconômicas e continuidade dos estudos. Por exemplo, estudantes cujas mães possuem Ensino Superior completo tendem a não interromper os estudos e apresentam acesso a celular e internet. Observa-se ainda que estudantes de cor/raça branca aparecem em maior proporção entre aqueles que mantiveram os estudos durante a pandemia. Além disso, estudantes com renda familiar de até R\$ 1.818,00 e que cursaram o ensino médio regular concentram-se na faixa etária adequada e possuem, em sua maioria, acesso ao celular.

A Tabela 3 apresenta uma amostra das regras selecionadas para o cenário correspondente ao Cluster 0 (alto desempenho).

Tabela 3 - Exemplos de regras de associação obtidas para o perfil de alto desempenho.

CLUSTER 0 (alto desempenho)	Conf.	Lift
APR_MATR_SIT_MEDIO=ensinoRegular, APR_MATR_VINCULO=nãoInterrompeu, NU_CELULAR=tem, NU_COMPUTADOR=tem, GEST_TEMP_ASSID_AULA_ONLINE =muitasVezes $\square$ ACESSO_INTERNET=tem	99,3%	1.06
DIF_INFRA_PANDEMIA=não, APR_MATR_SIT_MEDIO=ensinoRegular, PRAT_EST_TREINAR_REDACAO=muitasVezes $\square$ APR_MATR_VINCULO=nãoInterrompeu	99,3%	1.06
TP_COR_RACA=pretaPardaIndígena, NU_CELULAR=tem $\square$ APR_MATR_VINCULO=não Interrompeu, ACESSO_INTERNET=tem, PRAT_EST_RESUM_VIDEO=muitasVezes	53,5%	1.02
APR_MATR_VINCULO=nãoInterrompeu, ACESSO_INTERNET=tem, PRAT_EST_TREINAR_REDACAO=muitasVezes, APR_MATR_SIT_MEDIO=ensinoRegular $\square$ PRAT_EST_ES-TRUT_IDEIA_REDACAO=muitasVezes	73,1%	1.31
ACESSO_INTERNET=tem, PRAT_EST_ANOT_DUV_VIDEO_COMPLEMENTAR=muitas Vezes, PROB_ROT_EST_PANDEMIA=sim $\square$ PRAT_EST_ATV_FIXACAO=muitasVezes	59,5%	1.17

Para o Perfil de Alto Desempenho (Cluster 0),

foram geradas 89.491 regras de associação, considerando o filtro da medida de lift. As regras associadas a esse perfil envolvem predominantemente a realização frequente de hábitos de estudo. Esses estudantes, em geral, não relataram dificuldades de infraestrutura, possuem matrícula no ensino regular e praticaram com frequência atividades como treinamento de redação, elaboração de resumos e uso de materiais complementares, contribuindo para a continuidade dos estudos e para o melhor desempenho no exame.

Uma renda familiar mais elevada está associada à maior probabilidade de o estudante realizar o exame na faixa etária adequada. A dedicação aos estudos, a frequência ao ensino regular e a gestão do tempo para acompanhar as aulas online também se relacionam com melhor desempenho acadêmico. Destaca-se ainda a importância da prática de redação e da realização de atividades como resumos, anotações e exercícios de fixação para o aprendizado.

A Tabela 4 apresenta uma amostra das regras selecionadas para o cenário correspondente ao Perfil de Médio Desempenho (Cluster 1).

Tabela 4 - Exemplos de regras de associação obtidas para o perfil de médio desempenho.

CLUSTER 1 (médio desempenho)	Conf.	Lift
NU_CELULAR=tem, ACESSO_INTERNET=tem, GEST_TEMP_ATV_TEMPO=nenhumaVez, GEST_TEMP_ATV_MATERIAL=nenhumaVez, PRAT_EST_ANOT_DUV_PROF=poucasVezes □ GEST_TEMP_ATV_HORA_PROG=nenhumaVez	95,0%	1.37
GEST_TEMP_ATV_HORA_PROG=nenhumaVez, PRAT_EST_RESUM_TEXTO=poucasVezes, PRAT_EST_RESUM_VIDEO=poucasVezes □ PRAT_EST_LER=nenhumaVez, PRAT_EST_ANOT_DUV_PROF=nenhumaVez	81,6%	1.56
PRAT_EST_PARTICIPAR_FORUM=poucasVezes, AJUD_TERC_PANDEMIA=ninguémAuxiliou, NU_CELULAR=tem, ACESSO_INTERNET=tem □ PROB_ROT_EST_PANDEMIA=sim	94,7%	1.13
GEST_TEMP_ATV_TEMPO=nenhumaVez, PRAT_EST_LER=nenhumaVez, PRAT_EST_RESUM_TEXTO=nenhumaVez, PRAT_EST_RESUM_VIDEO=nenhumaVez, PRAT_EST_ATV_FIXACAO=nenhumaVez, PRAT_EST_REV_ANOT=nenhumaVez □ GEST_TEMP_ATV_HORA_PROG=nenhumaVez	94,8%	1.37

NIVEL_ESC_MAE=nãoCompletoMédio, RENDA_FAMILIAR=até3636, APR_MATR_TP_ESTUDO=híbrido □ GEST_TEMP_PONTUAL_AULA_ONLINE=nenhumaVez, GEST_TEMP_ASSID_AULA_ONLINE=nenhumaVez	88,1%	1.33
---	-------	------

Neste perfil foram geradas 3.603.659 regras de associação, sendo a amostra analisada composta majoritariamente por respostas negativas quanto à realização dos hábitos de estudo. As regras indicam estudantes com baixa realização de atividades de gestão do tempo, organização e esclarecimento de dúvidas, além de menor frequência em práticas como leitura, resumos e revisões, o que contribui para o desempenho intermediário.

Por fim, conforme amostra apresentada na Tabela 5, no Perfil de Baixo Desempenho (Cluster 2) foram identificadas 1.098.373 regras de associação, caracterizadas pela baixa frequência de hábitos de estudo. As regras indicam maior proporção de estudantes com limitações de infraestrutura, menor acesso a recursos tecnológicos, renda familiar mais baixa e menor escolaridade dos pais, além de baixa realização de práticas como leitura, resumos, anotações e treinamento de redação.

Tabela 5 - Exemplos de regras de associação obtidas para o perfil de baixo desempenho.

CLUSTER 2 (baixo desempenho)	Conf.	Lift
NU_CELULAR=tem, ACESSO_INTERNET=tem, APR_MATR_SIT_MEDIO=ensinoRegular, APR_MATR_VINCULO=nãoInterrompeu, PROB_ROT_EST_PANDEMIA=sim, PRAT_EST_ANOT_DUV_VIDEO=poucasVezes □ APR_MATR_APREND_PANDEMIA=aprendeuMaisPresencial	83,1%	1.03
NU_CELULAR=tem, ACESSO_INTERNET=tem, PRAT_EST_ATV_FIXACAO=poucasVezes, PRAT_EST_RESUM_TEXTO=poucasVezes, PRAT_EST_RESUM_VIDEO=poucasVezes □ PRAT_EST_LER=poucasVezes	90,4%	1.23
GEST_TEMP_ATV_CRONOGR=poucasVezes, EST_TEMP_ATV_TEMPO=poucasVezes, PRAT_EST_DISTRAcoes=poucasVezes, PRAT_EST_ANOT_DUV_VIDEO=poucasVezes □ RENDA_FAMILIAR=até1818, APR_MATR_SIT_MEDIO=ensinoRegular, DIF_INFRA_PANDEMIA=sim	77,1%	1.09

NU_CELULAR=tem, ACESSO_INTERNE- T=tem, PROB_ROT_EST_PANDEMIA=sim, APR_MATR_VINCULO=nãoInterrompeu, PRAT_EST_ANOT_DUV_PROF=poucasVezes ☐ AUTOAV_PREPARACAO_APRENDIZ=poucoPrepa- rado, APR_MATR_APREND_PANDEMIA =apren- deuMaisPresencial	67,8%	1.02
TP_SEXO=masculino, APR_MATR_SIT_ME- DIO=ensinoRegular, PRAT_EST_TREINAR_RE- DACAO=poucasVezes, GEST_TEMP_ATV_TEM- PO=poucasVezes, PRAT_EST_LER=poucasVezes ☐ DIF_INFRA_PANDEMIA=sim, AUTOAV_PREPA- RACAO_APRENDIZ=poucoPreparado	80,4%	1.06

O nível de escolaridade dos pais mostra-se como fator relevante, influenciando o ambiente doméstico e o incentivo aos estudos. De modo geral, os resultados indicam que a baixa frequência de realização de hábitos de estudo está associada a menores níveis de desempenho no ENEM, especialmente quando combinada com condições socioeconômicas menos favoráveis.

5. Trabalhos Relacionados

Diversos estudos têm investigado os impactos da pandemia de COVID-19 na educação brasileira, especialmente no desempenho dos estudantes no ENEM, a partir de abordagens baseadas em análise de dados educacionais. De modo geral, esses trabalhos evidenciam a influência de fatores socioeconômicos e de acesso a recursos no desempenho dos participantes, embora ainda apresentem limitações quanto à incorporação de aspectos comportamentais, como hábitos de estudo.

Weber Neto et al. (2022a; 2022b; 2022c) analisaram os efeitos da pandemia em diferentes recortes geográficos e temporais do ENEM. Os resultados apontam redução significativa na participação dos estudantes e aumento das desigualdades educacionais, especialmente associadas à renda, acesso à tecnologia e tipo de escola. Apesar da abrangência das análises, os estudos concentram-se predominantemente em variáveis estruturais, com menor ênfase em fatores relacionados ao comportamento de estudo dos estudantes.

Souza e Klug (2025) utilizaram os microdados do ENEM 2023 para desenvolver modelos preditivos de desempenho estudantil com base em variáveis socioeconômicas. Os resultados

evidenciaram que fatores como escolaridade dos pais, ocupação dos responsáveis e renda familiar exercem forte influência na predição do desempenho, indicando maior probabilidade de melhores resultados entre estudantes de contextos mais favorecidos. Embora os achados reforcem o papel das condições socioeconômicas e o potencial de técnicas de aprendizagem de máquina, a análise permanece centrada em variáveis tradicionais, sem incorporar aspectos comportamentais, como a organização do estudo e a autorregulação da aprendizagem, que podem contribuir de forma complementar para a compreensão do desempenho acadêmico.

Nascimento et al. (2025) avançam nessa perspectiva ao investigar a autorregulação da aprendizagem com base no questionário de hábitos de estudo do ENEM. Os resultados evidenciam a relação entre práticas de estudo, gestão do tempo e desempenho acadêmico, destacando a relevância de fatores comportamentais no contexto do ensino remoto. Ainda assim, o estudo concentra-se na identificação de padrões de autorregulação, sem explorar de forma integrada a formação de perfis de estudantes combinando múltiplas dimensões dos dados.

Estudos anteriores à pandemia também empregaram técnicas de mineração de dados e aprendizado de máquina para analisar o desempenho no ENEM. Sousa Lima et al. (2020) utilizaram o algoritmo K-means para agrupar estudantes em categorias de desempenho, evidenciando desigualdades regionais, diferenças associadas ao tipo de escola e a relevância da redação como fator discriminante. Silva et al. (2020) exploraram técnicas de agrupamento e regras de associação para identificar perfis de estudantes no ENEM 2019, evidenciando a centralidade de fatores como renda familiar, tipo de escola e escolaridade materna. Entretanto, esses estudos não consideram o contexto do ensino remoto nem incorporam variáveis relacionadas aos hábitos de estudo.

Diante dessas lacunas, este trabalho se diferencia por integrar variáveis tradicionais do ENEM com informações sobre hábitos de estudo em um contexto pandêmico, permitindo uma análise mais abrangente dos fatores associados ao desempenho estudantil. Além disso, combina técnicas de agrupamento e regras de

associação para identificar perfis de estudantes e padrões relevantes nos dados, explorando múltiplas dimensões de forma integrada.

Como contribuição adicional, disponibiliza o dataset ENEM_HE, favorecendo a reprodutibilidade e o avanço de pesquisas na área. Do ponto de vista prático, os resultados podem subsidiar a elaboração de estratégias educacionais mais direcionadas, especialmente no desenvolvimento de hábitos de estudo e no fortalecimento da autorregulação da aprendizagem em contextos de ensino remoto.

6. Considerações e Trabalhos Futuros

O presente estudo permitiu identificar perfis distintos de estudantes do ENEM a partir da integração de dados socioeconômicos e de hábitos de estudo, evidenciando o papel central da autorregulação da aprendizagem no desempenho acadêmico, especialmente no contexto do ensino remoto durante a pandemia de COVID-19.

Os resultados indicam que a frequência e a qualidade dos hábitos de estudo, particularmente aqueles relacionados à gestão do tempo, planejamento e práticas de revisão, apresentam maior capacidade de discriminação entre os perfis de desempenho quando comparados a variáveis socioeconômicas isoladas. Esse achado reforça a importância de considerar fatores comportamentais na análise do desempenho educacional, ampliando a abordagem tradicional centrada apenas em condições estruturais.

Do ponto de vista da Informática na Educação, o estudo contribui ao demonstrar o potencial de técnicas de mineração de dados educacionais, como agrupamento e regras de associação, para a identificação de padrões complexos em cenários reais e de larga escala. Além disso, a integração de diferentes fontes de dados possibilita uma análise mais abrangente do processo de aprendizagem, favorecendo a construção de modelos analíticos mais robustos e interpretáveis.

Em termos práticos, os resultados podem subsidiar a elaboração de estratégias pedagógicas

mais direcionadas, especialmente no desenvolvimento de hábitos de estudo e no fortalecimento da autorregulação da aprendizagem. Também podem apoiar a formulação de políticas públicas voltadas à redução das desigualdades educacionais, ao evidenciar a interação entre fatores comportamentais e condições de acesso e permanência escolar.

Apesar das contribuições, o estudo apresenta limitações. Destaca-se o uso de dados autodeclarados no questionário de hábitos de estudo, sujeitos a vieses de percepção e resposta. A taxa de participação de 37,6% também pode impactar a representatividade dos resultados, além de possíveis vieses associados à amostra analisada. Adicionalmente, as escolhas metodológicas, como a definição de parâmetros nos algoritmos e os limiares adotados nas regras de associação, podem influenciar os padrões identificados. Ressalta-se ainda que as técnicas utilizadas permitem identificar associações, mas não estabelecem relações de causalidade.

Como trabalhos futuros, propõe-se a ampliação da análise para outras edições do ENEM, permitindo a investigação de mudanças nos perfis ao longo do tempo. Sugere-se também o uso de técnicas mais avançadas de imputação de dados, como métodos baseados em densidade ou modelos de aprendizado de máquina, para lidar com respostas incompletas. Além disso, análises com recortes geográficos mais específicos, como por região, estado ou tipo de escola, podem aprofundar a compreensão das desigualdades educacionais. Por fim, recomenda-se a aplicação das técnicas empregadas em outros níveis de ensino, ampliando o escopo e a aplicabilidade dos resultados.

7. Referências

- AGRAWAL, R.; MANNILA, H.; SRIKANT, R.; TOIVONEN, H.; VERKAMO, A. I. Fast discovery of association rules. In: FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. (ed.). *Advances in knowledge discovery and data mining*. Cambridge, MA: MIT Press, 1996. p. 307-328.
- ALPAYDIN, E. *Introduction to machine learning*. 3. ed. Cambridge, MA: MIT Press, 2014.

- ARGÔLO, R. F. Determinantes de desempenho dos estudantes do ensino superior: o caso do curso de Psicologia da UFBA. 2017. Dissertação (Mestrado em Psicologia) – Universidade Federal da Bahia, Salvador, 2017. Disponível em: <https://repositorio.ufba.br/handle/ri/23984>. Acesso em: 11 set. 2025.
- CUNHA, L. F. F.; SILVA, A. S.; SILVA, A. P. O ensino remoto no Brasil em tempos de pandemia: diálogos acerca da qualidade e do direito e acesso à educação. Revista Com Censo, Brasília, v. 7, n. 3, p. 27–37, 2020. Disponível em: <https://periodicos.se.df.gov.br/index.php/comcenso/article/view/924>. Acesso em: 9 set. 2025.
- 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, v. 31, p. 323–351, 2023. DOI: <https://doi.org/10.5753/rbie.2023.3087>.
- ELMASRI, R.; NAVATHE, S. B. Sistemas de banco de dados. 6. ed. São Paulo: Pearson, 2011.
- FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. C. P. L. F. de. Inteligência artificial: uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC, 2021.
- FRANCO, J. J.; MIRANDA, F. L. de A.; STIEGLER, D.; DANTAS, F. R.; BRANCHER, J. D.; NOGUEIRA, T. do C. Usando mineração de dados para identificar fatores mais importantes do Enem dos últimos 22 anos. In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 31., 2020, Porto Alegre. Anais [...]. Porto Alegre: SBC, 2020. p. 1112–1121. DOI: <https://doi.org/10.5753/cbie.sbie.2020.1112>.
- HAN, J.; KAMBER, M.; PEI, J. Data mining: concepts and techniques. 3. ed. Waltham: Morgan Kaufmann, 2012.
- HIPP, J.; GÜNTZER, U.; NAKHAEIZADEH, G. Algorithms for association rule mining - a general survey and comparison. ACM SIGKDD Explorations Newsletter, v. 2, n. 1, p. 58–64, 2000. DOI: <https://doi.org/10.1145/360402.360421>.
- INEP. Microdados do ENEM 2022. Brasília: INEP, 2022. Disponível em: <https://www.gov.br/inep/pt-br/aceso-a-informacao/dados-abertos/microdados/enem>. Acesso em: 28 jul. 2025.
- INEP. Painel apresenta pesquisa sobre estudo na pandemia. Brasília, DF: Ministério da Educação, 2023. Disponível em: <https://www.gov.br/inep/pt-br/assuntos/noticias/enem/painel-apresenta-pesquisa-sobre-estudo-na-pandemia>. Acesso em: 23 jul. 2025.
- KISSLER, S. M.; TEDIJANTO, C.; LIPSITCH, M.; GRAD, Y. H. Social distancing strategies for curbing the COVID-19 epidemic [preprint]. medRxiv, 2020. DOI: <https://doi.org/10.1101/2020.03.22.20041079>.
- NASCIMENTO, N. F. B. C.; RODRIGUES, R. L.; BARROS, M. A. M.; RODRIGUES, K. C. Identificação de autorregulação da aprendizagem em formulário de hábitos de estudo do ENEM. RENO-TE, Porto Alegre, v. 23, n. 1, p. 539–549, 2025. DOI: <https://doi.org/10.22456/1679-1916.149299>.
- PEDREGOSA, F.; VAROQUAUX, G.; GRAMFORT, A.; MICHEL, V.; THIRION, B.; GRISEL, O.; BLONDEL, M.; PRETTENHOFER, P.; WEISS, R.; DUBOURG, V.; VANDERPLAS, J.; PASSOS, A.; COUNAPEAU, D.; BRUCHER, M.; PERROT, M.; DUCHESNAY, É. Scikit-learn: machine learning in Python. Journal of Machine Learning Research, v. 12, p. 2825–2830, 2011. Disponível em: <https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf>. Acesso em: 14 out. 2025.
- RAMOS, J. L. C.; RODRIGUES, R. L.; SILVA, J. C. S.; OLIVEIRA, P. L. S. de. CRISP-EDM: uma proposta de adaptação do Modelo CRISP-DM para mineração de dados educacionais. In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 31., 2020, Porto Alegre. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2020. p. 1092–1101. DOI: <https://doi.org/10.5753/cbie.sbie.2020.1092>.
- SILVA, V. A. A.; MORENO, L. L. O.; GONÇALVES, L. B.; SOARES, S. S. R. F.; SOUZA JÚNIOR, R. R. 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. DOI: <https://doi.org/10.5753/cbie.sbie.2020.72>.
- SILVA FILHO, R. L. C.; ADEODATO, P. J. L. Data mining solution for assessing the secondary school students of Brazilian federal institutes. In: BRAZILIAN CONFERENCE ON INTELLIGENT SYSTEMS, 2019, Salvador, Brazil. Proceedings. Salvador: IEEE, 2019. p. 574-579. DOI: <https://doi.org/10.1109/BRA-CIS.2019.00106>.
- SOARES, R. de A.; SILVA, G. A. Regulamentos da EaD no Brasil e o Impacto da Portaria Nº 343/2020 no Ensino Superior. EaD em Foco, v. 10, n.3, e1043, 2020. DOI: <https://doi.org/10.18264/eadf.v10i3.1043>.
- SOUZA LIMA, A. M.; CHOQUENAIRA FLOREZ, A. Y.; ASPAUZA LESCANO, A. I.; OLIVEIRA NOVAES, J. V. de; FATIMA MARTINS, N. de; TRAINA JUNIOR, C.; PARROS MACHADO DE SOUSA, E.; RODRIGUES JUNIOR, J. F.; FERREIRA CORDEIRO, R. L. Analysis of ENEM's attendants between 2012 and 2017 using a clustering approach. Journal of Information and Data Management, [S. l.], v. 11, n. 2, p. 15-130. 2020. DOI: <https://doi.org/10.5753/jidm.2020.2023>.
- SOUZA, M. de; KLUG, D. L. Desigualdades educacionais no Enem: uma perspectiva baseada em variáveis socioeconômicas e aprendizagem de máquina. Regae: Revista de Gestão e Avaliação Educacional, Santa Maria, v. 14, n. 23, e93251, 2025. DOI: <https://doi.org/10.5902/2318133893251>.
- WEBER NETO, N.; SOARES, R. C.; COUTINHO, L. R.; TELES, A. S. Análise Exploratória de Dados para Identificar o Impacto da Pandemia da COVID-19 no ENEM dos Estados do Ceará, Maranhão e Piauí. In: ESCOLA REGIONAL DE COMPUTAÇÃO DO CEARÁ, MARANHÃO E PIAUÍ, 2022, São Luís, MA. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2022a. p. 31-40. DOI: <https://doi.org/10.5753/ercemapi.2022.225842>.
- WEBER NETO, N.; SOARES, R. C.; COUTINHO, L. R.; TELES, A. S. A pandemia da COVID-19 impactou o ENEM? Uma análise comparativa de dados dos anos de 2019 e 2020. RE-NOTE, Porto Alegre, v. 20, n. 1, p. 223-232, 2022b. DOI: <https://doi.org/10.22456/1679-1916.126655>.
- WEBER NETO, N.; SOARES, R. C.; COUTINHO, L. R.; TELES, A. S. Data Analysis to Identify the Impact of the COVID-19 Pandemic on ENEM in 3 States of Northeast Brazil. Revista de Sistemas e Computação, v. 12, n. 3, p. 35-44. 2022c. DOI: <https://doi.org/10.36558/rsc.v12i3.7919>.
- ZAKI, M. J.; MEIRA JUNIOR, W. Data mining and machine learning: fundamental concepts and algorithms. 2. ed. Cambridge: Cambridge University Press, 2020.
- ZIMMERMAN, B. J. Attaining self-regulation: a social cognitive perspective. In: BOEKAERTS, M.; PINTRICH, P.; ZEIDNER, M. (org.). Handbook of self-regulation. San Diego: Academic Press, 2000. p. 13-39. DOI: <https://doi.org/10.1016/B978-012109890-2/50031-7>.