Aplicações de Aprendizagem de Máquina com dados do INEP: Uma Revisão de Literatura

Thiago C. Feitosa[1], Antônio R. Braga[2]

[1] Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE)
Av. Treze de Maio, 2081 - Benfica, Fortaleza – CE

[2] Universidade Federal do Ceará, Campus Quixadá (UFC)
Av. José de Freitas Queiroz, 5003 - Quixadá, CE, 63902-580

thiago@ifce.edu.br, rafaelbraga@ufc.br

Abstract. *This study reviews 27 academic works on machine learning applied to INEP datasets, mainly ENADE and ENEM. It identifies key themes such as educational inequalities, performance, and dropout, with a focus on supervised learning—particularly classification algorithms like Random Forest, SVM, and neural networks. The findings highlight analytical advances and suggest integrating datasets and developing interactive tools to inform public policy.*

Resumo. *Este estudo revisa 27 artigos sobre o uso de aprendizado de máquina em dados do INEP, com ênfase no ENADE e no ENEM. Destacam-se temas como desigualdades, desempenho e evasão, com predominância de técnicas supervisionadas, como Random Forest, SVM e redes neurais. Os resultados apontam avanços analíticos e sugerem a integração de bases de dados e o desenvolvimento de ferramentas interativas para subsidiar políticas públicas.*

1. Introdução

No contexto da educação superior, a Lei nº 10.861, de 14 de abril de 2004, instituiu o Sistema Nacional de Avaliação da Educação Superior (SINAES), composto por diferentes instrumentos de avaliação. Entre eles, pode-se destacar o Exame Nacional de Desempenho dos Estudantes (ENADE), cuja finalidade, conforme estabelecido nos artigos 5º, §§ 1º a 11 da referida Lei, é divulgar a qualidade dos cursos superiores por meio do desempenho dos estudantes em relação às matrizes curriculares da Instituição de Ensino (IE), como também às habilidades desenvolvidas ao longo da graduação [Brasil, 2004].

Conforme a Portaria Normativa MEC nº 840, de 24 de agosto de 2018, cabe ao INEP, por meio da Diretoria de Avaliação da Educação Superior (DAES), realizar o ENADE [da Educação 2018] e divulgar, a partir de seus resultados, os Indicadores de Qualidade da Educação Superior, que incluem o Conceito ENADE, o Indicador de Diferença entre os Desempenhos Observado e Esperado (IDD), o Conceito Preliminar de Curso (CPC) e o Índice Geral de Cursos Avaliados da Instituição (IGC).

Desde 2004, o ENADE disponibiliza microdados detalhados sobre estudantes, cursos e instituições de ensino superior, constituindo uma base valiosa para análises sobre o desempenho discente [IBGE 2024; INEP 2025]. Com o avanço da ciência de dados e da aprendizagem de máquina, esses dados vêm sendo cada vez mais utilizados em estudos que buscam identificar padrões, prever resultados e caracterizar perfis, superando limitações das abordagens estatísticas tradicionais [Garcia et al. 2022; Neto et al. 2020].

O presente trabalho tem como objetivo realizar uma **Revisão Sistemática de Literatura** (RSL), limitando-se a pesquisas publicadas entre 2020 e 2025, sobre as aplicações de algoritmos de aprendizagem de máquina utilizando *datasets* fornecidos pelo INEP, dando mais ênfase às aplicadas em *datasets* do ENADE, mapeando os tipos de dados utilizados, os problemas da educação que foram pesquisados, as metodologias de aprendizagem abordadas e as técnicas de *machine learning (ML)* aplicadas. Optou-se por este período, por ser o período em que se observa maior maturidade no uso de *ML*, alinhando-se com o crescente interesse da comunidade científica após a pandemia de Covid-19.

2. Trabalhos Relacionados

Os trabalhos relacionados foram selecionados devido à similaridade entre suas questões de pesquisa e as do presente estudo, abordando revisões sistemáticas e mapeamentos do uso de dados do INEP e técnicas de machine learning na educação brasileira. Dessa forma, oferecem referência comparativa e contextualizam os avanços e lacunas na área.

O estudo de Dutra et al. (2023) realiza uma **Revisão Sistemática da Literatura** (RSL) com foco nos fatores que influenciam o desempenho no ENEM, a partir da análise de estudos publicados entre 2013 e 2022. Utilizando dados extraídos dos microdados do exame e técnicas de mineração de dados, a pesquisa destaca a relevância de variáveis socioeconômicas como determinantes do desempenho estudantil.

Barbosa et al. (2023), por sua vez, conduzem um **Mapeamento Sistemático da Literatura** (MSL) voltado à análise de técnicas estatísticas, algoritmos de ML e sistemas computacionais aplicados aos dados do ENADE, com ênfase nos cursos da área de Computação. Com base em 32 estudos selecionados, o trabalho evidencia o uso recorrente de estatística descritiva e o desenvolvimento de sistemas de apoio ao estudante, incluindo gamificação e *dashboards.*

Já o artigo de [De Castro Soares et al. (2021)] apresenta uma RSL sobre a aplicação de técnicas de mineração de dados à educação básica brasileira, utilizando bases do INEP como SAEB, IDEB e Censo Escolar. A partir da análise de 410 estudos, dos quais 19 foram selecionados, os autores identificam problemas recorrentes como baixo desempenho e evasão, com destaque para metodologias como CRISP-DM e KDD, e técnicas como regressão e análise de correlação.

3. Quadro de Siglas

Quadro 1 - Siglas de técnicas e algoritmos de aprendizado

Sigla	Significado
1C-SVN	One-Class Support Vector Machine
AdaBoost	Adaptive Boosting
ANOVA	Analysis of Variance
CART	Classification and Regression Tree
CV	Cross-Validation
CWR	Class-Weighted Regressor
DT	Decision Tree (Árvore de Decisão)
DTR	Decision Tree Regressor
ElasticN	Elastic Net (combinação de Lasso e Ridge)
ETL	Extract, Transform, Load
FE	Feature Engineering
FI	Feature Importance
GBR	Gradient Boosting Regressor
GR	Gain Ratio
GSCV	Grid Search with Cross-Validation
IF	Isolation Forest
IG	Ou InfoGain, Information Gain
JRIP	Implementação de regras RIPPER no Weka
RM	Regressão Multipla
SFS	Sequential Forward Selection
SU	Symetrical Uncertainty
TF-IDF	Term Frequency-Inverse Document Frequency

Sigla	Significado
KDD	Knowledge Discovery in Databases
K-Means	Algoritmo de Agrupamento K-Médias
K-NN	k-Nearest Neighbors
K-NN-R	k-Nearest Neighbors Regressor
Lasso	Least Absolute Shrinkage and Selection Operator
LGBM	Ligth Gradient Boosting Machine
LR	Logistic Regression
MLP	Multi Layer Perception
ML-KNN	Multi-Label k-Nearest Neighbors
NB	Naive Bayes
PART	Partial Decision Trees
PCA	Principal Component Analysis
RFE	Recursive Feature Elimination
RF	Random Forest
RFR	Random Forest Regressor
Ridge	Regressão Ridge
RL	Regressão Linear
SKB	SelectKBest (seleção de atributo)
SMOTE	Synthetic Minority Over-sampling Technique
SVM	Support Vector Machine
XGB	eXtreme Gradient Boosting

4. Metodologia

Esta pesquisa aborda identificar, analisar e sintetizar estudos acadêmicos que empregam algoritmos de aprendizagem de máquina aplicados aos dados do ENADE. A RSL foi conduzida conforme as diretrizes propostas por [Kitchenham, 2007], ao qual define três etapas para realizar uma RSL, o planejamento, a condução e a produção do relatório.

4.1. Questões de Pesquisa

Para alcançar os resultados desta RSL, foram formuladas as seguintes questões de pesquisa (QP) :

- QP1. Quais dados do INEP estão sendo estudados?
- QP2. Quais problemas da educação estão sendo abordados?
- QP3. Quais tipos de aprendizagem de máquina estão sendo utilizadas?
- QP4. Quais técnicas estão sendo empregadas na aprendizagem de máquina?

4.2. Bases de dados consultadas e estratégia de pesquisa

A busca foi realizada nas seguintes bases de dados: **Google Scholar**, **SciELO** e **Portal de Periódicos CAPES**. Como o objetivo é selecionar as publicações que envolvam os dados do INEP ou ENADE e que também sejam relacionados ao aprendizado de máquina, foi criada as *strings* de pesquisa em inglês e português e incluído os sinônimos: `"ENADE" or "INEP") and ("Data Mining" or "Data Analysis" or "Machine Learning") e ("ENADE" or "INEP") and ("Mineração de Dados" or "Análise de Dados")`

As strings foram aplicadas em qualquer campo de pesquisa ("*All Fields*", "*Anywhere*" ou "*Full Text*"), os resultados retornados foram filtrados conforme ano de publicação, entre 2020 e 2025. Ao todo foram encontrados 66 resultados, sendo 44 no *Google Scholar*, através do Portal da CAPES foram 12 documentos por meio da coleção "*Web of Science*" e no portal SciELO foram encontrados 6 artigos.

4.3. Processo de seleção dos estudos

Em relação ao total de resultados obtidos na pesquisa, foram definidos critérios de seleção (**Quadro 2**) com o objetivo de excluir da análise os trabalhos que não se enquadram no escopo desta Revisão de Literatura.

Quadro 2 - Critérios de Inclusão / Exclusão

Aspecto Observado	Critério Aplicado
Trabalho de conclusão de curso, teses e dissertações	Exclusão
Capítulos de livros, páginas de revistas	Exclusão
Realiza mineração de dados mas não utiliza nenhuma base de dados do INEP	Exclusão
Não implementa técnicas de aprendizado de máquina	Exclusão
Publicado antes do ano 2020	Exclusão
Artigos em Inglês ou Português	Inclusão
Artigos Completos	Inclusão
Realiza mineração de dados utilizando qualquer base de dados do INEP	Inclusão

A seleção dos artigos seguiu três etapas: 1) Triagem inicial por título e resumo, 2) Aplicação dos critérios de inclusão e exclusão e 3) Leitura completa dos artigos potencialmente relevantes.

Ao final do processo, dos 66 documentos encontrados, 27 foram considerados elegíveis para análise e síntese dos resultados após a aplicação dos critérios de exclusão.

5. Resultados

Ao final das três etapas de seleção dos artigos o **Quadro 3** foi preenchido com informações retiradas dos respectivos artigos.

Quadro 3. Resumo dos principais pontos dos artigos selecionados

Referência	Fonte de Dados	Problema	Tipo de Aprendizado	Técnicas Aplicadas
[Alberto V. 2020]	ENEM	Desigualdade	Supervisionado	CART
[Lima et al. 2020]	Censo e Indicadores	Evasão	Supervisionado	RL, Ridge, Lasso, CWR

[Rodrigues 2021]	ENADE	Qualidade dos cursos	Supervisionado	DT, RF, SVM, k-NN, MLP
[Souza et al. 2025]	ENADE	Alinhamento Curricular	Supervisionado	RF, k-NN, DT, LR, SVM, NB
[Vieira et al. 2022]	ENADE	Desigualdade	Supervisionado	K-Means
[da Silva M 2024]	ENADE, CENSUP	Atraso social	Não supervisionado	DT, RF, AdaBoost, XGB, SMOTE
[Rosa et al. 2021]	ENADE	Desigualdade	Supervisionado	DT (J48), RF, SVM
[Marques et al. 2023]	ENADE, IGC e Censo	Impacto da escola	Supervisionado	K-Means, Apriori
[Freitas et al. 2023]	Censo	Gestão ineficiente	Não supervisionado	K-Means, LOF, IF, 1C SVM, k-NN
[Oliveira 2021]	ENADE, Relatórios	Classificação de questões	Não supervisionado	ML-KNN, TF-IDF
[Magalhães 2025]	ENADE (provas)	Dificuldade com questões	Supervisionado	Prompt-based GenAI
[Azevedo S 2024]	ENADE, CPC, IDD, IGC	Desigualdade	Não supervisionado	RM, K-Medoids
[Lima 2021]	ENADE, Relatórios, DCNs	Análise curricular	Não supervisionado	ANOVA, Pearson, Spear man
[da Silva R. 2022]	ENADE	Impacto socioeconômico	Não supervisionado	DT
[Maretti 2023]	ENEM	Desigualdade	Supervisionado	FE, FI
[Cunha 2021]	ENADE	Análise Curricular	Supervisionado	ETL + KDD
[Estivalete 2021]	ENADE	Formação curricular	Supervisionado	DT (J48)
[Dos Santos 2023]	ENEM	Desigualdade	Weka	MLP, RF, J48, JRIP, PART, InfoGain
[Barros 2023]	Censo, TDI	Distorção idade-série	Supervisionado	DTR, GSCV, FI
[Sakashita 2023]	ENADE	Predição de conceito	Supervisionado	DT, RF, SVM, k-NN, MLP
[Souza et al. 2024]	ENADE	Classificação curricular	Supervisionado	DT, RF, SVM, NB, LR, k NN, TF-IDF
[Gondran 2022]	ENADE	Desigualdade	Supervisionado	IG, GR, SU, Relief F
[Franco 2020]	ENEM	Fatores de desempenho	Supervisionado	IG, PCA, SKB, RFE, SFS, XGB, LGBM, MLP
[Silva S F 2023]	ENADE	Práticas alternativas	Supervisionado	Elastic N, CV
[Teixeira de J 2020]	ENEM	Previsão de notas	Supervisionado	RL, k-NN-R, GBR, RFR
[Silva 2020]	ENADE	Baixo desempenho	Supervisionado	K-Means
[Gomes et al. 2020]	ENEM	Baixo desempenho	Supervisionado	CART

6. Resultados

A análise dos 27 artigos revela que a base de dados do **ENADE** é a mais utilizada nas pesquisas (74%), seguida pelo **ENEM** (22%) e pelo **Censo da Educação Superior** (15%), sendo comum o uso combinado de diferentes fontes do **INEP**. Os temas mais recorrentes são relacionados à **desigualdade educacional** – seja ela social, econômica, racial ou geográfica – presentes em 44% dos estudos. Os outros incluem a avaliação da qualidade e **desempenho acadêmico** (22%), **análises curriculares** e **alinhamento com diretrizes** (19%) e questões ligadas à **gestão institucional** e **evasão** (15%). As metodologias envolvem predominantemente **técnicas de aprendizado supervisionado** e **mineração de dados**, com ênfase em algoritmos como **árvores de decisão**, ***random forest*, SVM e k-NN**, refletindo o uso de métodos preditivos e classificatórios no contexto educacional brasileiro.

6.1. Resposta à Questão de Pesquisa 1 (QP1)

A questão de pesquisa 1 tem como objetivo identificar quais conjuntos de dados disponibilizados pelo INEP são utilizados nas pesquisas selecionadas

A análise mostrou que os dados mais utilizados em estudos que aplicam aprendizado de máquina e estatística sobre a educação superior brasileira são os microdados do **ENADE** e do **ENEM**, fornecidos pelo **INEP**. O **ENADE** aparece como a principal fonte, sendo empregado em 22 artigos para investigar o desempenho dos estudantes, características socioeconômicas, dados institucionais e até a análise textual das questões das provas. **O ENEM** é usado em outros 7 artigos, geralmente com foco em desempenho em matemática e fatores socioeconômicos. Além disso, o **Censo da**

Educação Superior (**CENSUP**) e o **Censo Escolar da Educação Básica** também são explorados em estudos voltados para evasão escolar, infraestrutura e análise financeira das instituições de ensino.

Os dados do INEP são utilizados de forma ampla, abrangendo desde atributos individuais dos estudantes (como renda, tipo de escola, trabalho, raça, moradia) até informações institucionais (tipo de IES, infraestrutura, indicadores como CPC, IDD, IGC). Vários estudos também criam atributos derivados (como tempo de graduação ou idade de ingresso) e analisam documentos como provas e relatórios do ENADE. Há um destaque para abordagens que combinam essas informações com técnicas de aprendizado de máquina, mineração de dados e estatística para resolver problemas como **predição de desempenho**, **agrupamento de perfis**, **identificação de desigualdades educacionais** e **classificação curricular**.

6.2. Resposta à Questão de Pesquisa 2 (QP2)

A segunda questão da pesquisa busca compreender os principais desafios da educação que têm sido objeto de análise nos artigos.

Os temas recorrentes observados nos artigos foram agrupados em 7 categorias principais e, em seguida os 27 documentos foram classificados conforme o assunto abordado, considerando múltiplas categorias para o mesmo artigo, gerando o **Quadro 4** a seguir.

Quadro 4. Problemas da educação superior abordados nos artigos analisados

Aspecto Observado	Nº de Artigos
Problemas curriculares e de alinhamento pedagógico	7
Desempenho acadêmico insuficiente (especialmente em matemática)	6
Falta de ferramentas para análise pedagógica	5
Evasão escolar / abandono / tempo de graduação	5
Outros (infraestrutura, gestão institucional, concept drift, etc.)	4
Distorção idade-série / atraso escolar	1

A partir do **Quadro 4** o gráfico da **Figura 1** foi criado, com ele é possível observar que os principais problemas da educação básica estão fortemente concentrados nas desigualdades socioeconômicas e regionais, mencionadas em 17 deles, evidenciando o **impacto da renda**, tipo de escola e localização geográfica no desempenho dos estudantes. Além disso, são frequentes as discussões sobre **falhas curriculares** e **desalinhamento pedagógico** (7 artigos), **baixo desempenho acadêmico**, especialmente em **matemática** (6), e a **ausência de ferramentas adequadas para análise pedagógica** (5). Questões como **evasão escolar**, **abandono** e **prolongamento do tempo de graduação** também se destacam (5), junto de temas menos frequentes, como **infraestrutura deficiente**, **distorção idade-série e problemas institucionais de gestão**, totalizando 45 ocorrências temáticas distintas, já que muitos artigos abordam mais de um problema simultaneamente.

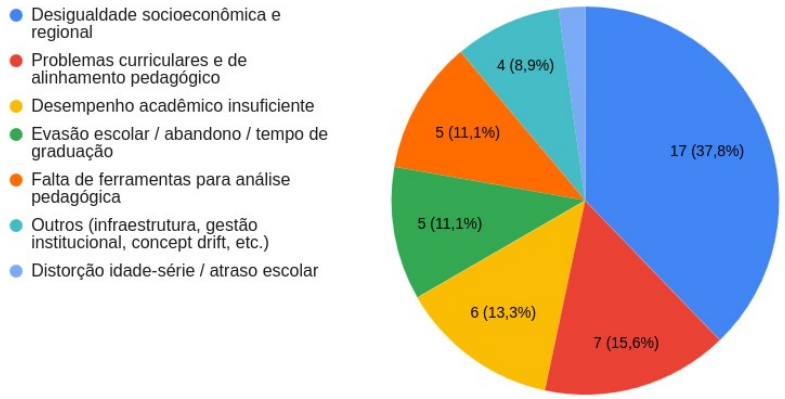

Figura 1 - Agrupamento dos problemas da educação abordados

6.3. Resposta à Questão de Pesquisa 3 (QP3)

A terceira questão de pesquisa busca identificar quais metodologias de aprendizagem de máquina têm sido empregadas nos estudos analisados.

As metodologias de aprendizagem de máquina utilizadas nos 27 artigos analisados revelam uma predominância do **aprendizado supervisionado**, presente em 20 estudos (74%). Dentre esses, destacam-se algoritmos como árvores de decisão (em pelo menos 6 artigos), regressão (linear, logística, ridge, lasso e *elastic net*, em cerca de 7 estudos), e classificadores como **KNN**, ***Random Forest***, **XGBoost** e **LightGBM**. Esses modelos foram frequentemente aplicados para tarefas de predição de desempenho acadêmico, evasão, tempo de graduação ou classificação de conteúdos do **ENADE**. As abordagens supervisionadas seguiram, em muitos casos, *frameworks* estruturados como **KDD** e **CRISP-DM**, além do uso de técnicas de validação cruzada e otimização de hiperparâmetros.

Já o aprendizado não supervisionado apareceu em 6 artigos (22%), sendo mais comum em análises de agrupamento (**K-means**, **K-medoids**) e descoberta de padrões por meio de regras de associação ou detecção de outliers. Um único estudo (4%) utilizou modelos generativos de linguagem natural para a criação automática de questões no estilo ENADE, representando uma abordagem inovadora com apoio em IA generativa. Alguns trabalhos, mesmo sem aplicar diretamente algoritmos clássicos de ML, seguiram processos estruturados de mineração de dados (como KDD), reforçando o uso da ciência de dados educacionais como base metodológica na área.

6.4. Resposta à Questão de Pesquisa 4 (QP4)

A questão de pesquisa 4 trata da identificação das principais técnicas de *ML* empregadas nas aplicações de aprendizagem de máquina observadas na literatura:

Nos artigos analisados, As **árvores de decisão** (CART, J48) foram utilizadas em 6 artigos (22%), seguidas por ***Random Forest*** (RF) em 7 artigos (26%) e ***Support Vector*** *Machines* (SVM) em 6 artigos (22%). O ***k-Nearest Neighbors*** (k-NN) apareceu em 6 artigos (22%), enquanto as redes neurais ***Multilayer Perceptron*** (MLP) foram empregadas em 4 artigos (15%). Além disso, técnicas de regressão, como **Regressão Linear** e ***Elastic Net***, foram usadas em 4 artigos (15%), e métodos de aprendizado não supervisionado, como o **K-Means**, foram aplicados em 6 artigos (22%). O

balanceamento de classes com **SMOTE** e o pré-processamento de dados, como normalização e One-Hot Encoding, foram aplicados em vários estudos para otimizar os modelos.

Em relação à avaliação dos modelos, as métricas mais comuns foram F1-Score, usada em 6 artigos (22%) devido ao desbalanceamento de classes, acurácia em 12 artigos (44%), precisão e recall em 5 artigos (19%), e curvas ROC (AUC) em 4 artigos (15%). A validação cruzada foi utilizada em 16 artigos (59%), sendo a validação cruzada 5-fold a mais recorrente (30%). Além disso, 8 artigos (30%) aplicaram ajuste de hiperparâmetros, com o uso de **HalvingGridSearchCV** ou **GridSearchCV**. Métricas adicionais, como RMSE (Root Mean Squared Error), foram usadas em 3 artigos (11%), e a análise de importância de variáveis foi destacada em 5 artigos (19%).

A **Figura 2** abaixo apresenta o resultado da análise das pesquisas agrupadas por tipo de técnica que foi aplicada no estudo.

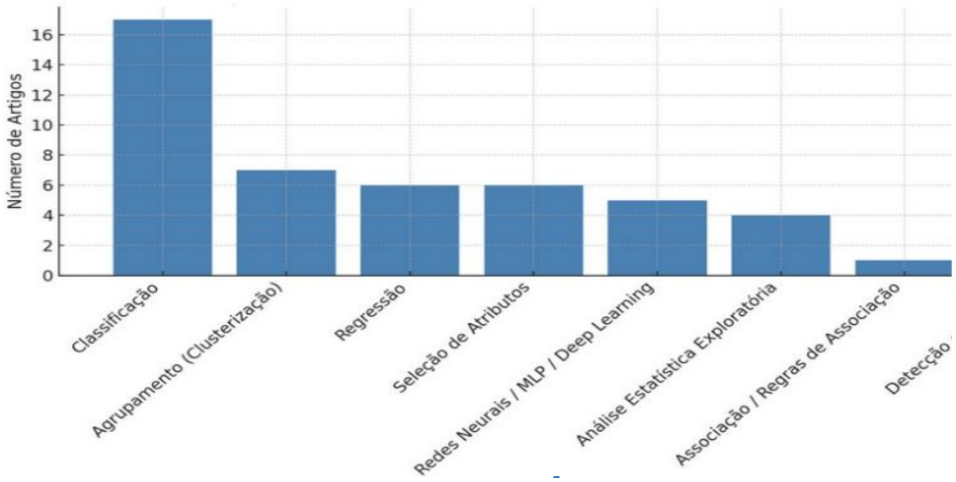

Figura 2 - Agrupamento dos problemas da educação abordados

7. Considerações Finais

A análise dos **27 artigos publicados entre 2020 e 2025** revela que, **nesse conjunto**, há uso intensivo dos microdados do ENADE e do ENEM em investigações que aplicam aprendizado de máquina e estatística à educação superior brasileira. Observa-se, nesses estudos, foco majoritário na permanência e no desempenho dos estudantes, com atenção a fatores socioeconômicos, institucionais e curriculares. Predominam metodologias supervisionadas, com destaque para algoritmos de classificação, como Árvores de Decisão, *Random Forest* e SVM.

Essas pesquisas demonstram um crescente rigor metodológico, com o uso de técnicas como validação cruzada e ajuste de hiperparâmetros, além da aplicação de métricas como acurácia e *F1-Score*. Os resultados evidenciam o potencial da ciência de dados educacionais para diagnosticar desigualdades, apoiar decisões pedagógicas e orientar políticas públicas voltadas à qualidade e equidade na educação superior.

Cabe destacar que as conclusões aqui apresentadas se referem exclusivamente ao corpus selecionado, limitado ao período de **2020** a **2025** e às bases ***Google Scholar***, **SciELO** e Portal **CAPES**, não sendo generalizáveis para toda a produção científica da

área**.** Entre as limitações, ressalta-se o tamanho da amostra e possíveis vieses introduzidos pelos critérios de seleção

Como perspectivas futuras, sugere-se ampliar o escopo para outros períodos, bases de dados e abordagens, bem como aprofundar a discussão sobre aspectos éticos e avaliação da qualidade dos estudos. A integração de diferentes bases do INEP e o desenvolvimento de métodos mais avançados, como *deep learning* e aprendizado por reforço, especialmente para análises preditivas complexas, podem contribuir para o avanço do campo. O desenvolvimento de ferramentas interativas com base nesses modelos também se apresenta como uma via promissora para aproximar pesquisa, gestão educacional e formulação de políticas públicas.

8. Referências

Alberto V., d. S. A. e. a. (2020). 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)* , pages 72–81. SBC.

Azevedo S, J. e. a. (2024). Desafios e oportunidades do ensino em computação: Uma análise exploratória e multivariada dos dados do enade. *Intersaberes*, 19.

Barbosa, P. L. S. et al. (2023). *O sucesso não é apenas uma questão de sorte: um mapeamento sistemático sobre técnicas de análise do Enade da área de Computação* . Sociedade Brasileira de Computação.

Barros, A. N. e. a. (2023). Aplicação de learning analytics para identificação de tomada de decisão sobre a distorção idade-série no Brasil. In *Workshop de Aplicações Práticas de Learning Analytics em Instituições de Ensino no Brasil (WAPLA)* , pages 21–31.

Brasil (2004). Lei nº 10.861, de 14 de abril de 2004. institui o sistema nacional de avaliação da educação superior (sinaes). http://www.planalto.gov.br/ccivil_ 03/_ato2004-2006/2004/lei/l10.861.htm. Acesso em: 30 abr. 2025.

Cunha, R. F. e. a. (2021). Análise automática dos microdados do ENADE para prover melhorias em cursos de ciência da computação.

da Educação, B. M. (2018). Portaria normativa mec nº 840, de 24 de agosto de 2018. https://www.in.gov.br/materia/asset_publisher/Kujrw0TZC2Mb/content/id/40288786/ do1-2018-08-27-portaria-n-840-de-24-de-agosto-de-2018-40288720. Acesso em: 30 abr. 2025.

da Silva M, e. a. (2024). Uso de técnicas de aprendizado de máquina para predição do tempo de graduação dos discentes de engenharia da computação na região sudeste do Brasil.

da Silva R., C. C. e. a. (2022). O impacto de aspectos socioeconômicos no desempenho de estudantes de sistemas de informação no Enade. Revista *Brasileira de Informática na Educação* , 30:157–181.

De Castro Soares, R., Weber Neto, N., Reis Coutinho, L., Da Silva e Silva, F. J., Viana dos Santos, D., and Soares Teles, A. (2021). Mineração de dados da educação básica brasileira usando as bases do INEP: Uma revisão sistemática da literatura.

Dos Santos, Vandeir Vioti, N. P. (2023). Analysis of the importance of social/racial quotas through enem 's microdata mining. *Informática na educação: teoria & prática ´* , 26(1):110– 117.

Dutra, J. F., Firmino Junior, J. B., and Fernandes, D. Y. S. (2023). Fatores que podem interferir no desempenho dos estudantes no enem: uma revisão sistemática da literatura. *Revista Brasileira de Informática na Educação*, 31:323–351. Acesso em: 29 abr. 2025.

Estivalete, P. e. a. (2021). Descoberta de conhecimentos sobre integração curricular nos estados da região sul do Brasil por meio do Enade 2012: um estudo utilizando mineração de dados educacionais.

Franco, J. J. e. a. (2020). Usando mineração de dados para identificar fatores mais importantes do enem dos últimos 22 anos. In *Simpósio Brasileiro de Informática na Educação (SBIE) ˜* , pages 1112–1121.

Freitas, C. N., Gouveia, M. M. R., Rodrigues, M. E., Alves, G., Batista, M. C., and Rodrigues, R. L. (2023). Detecção de outliers em finanças de instituições de ensino superior brasileiras utilizando aprendizado de máquina não supervisionado. ˜

Garcia, D. H. A., Silva, A. C. R. d., Silva, M. G. d., and Santos, A. A. (2022). Mineração de dados educacionais na predição do desempenho acadêmico: um prognóstico a partir do percurso curricular realizado. In *Anais do Simpósio Brasileiro de Informática na Educação˜ (SBIE)*, páginas 1124–1134. Sociedade Brasileira de Computação. ˜

Gomes, C. M. A., Fleith, D. d. S., Marinho-Araujo, C. M., and Rabelo, M. L. (2020). Pre dictors of students' mathematics achievement in secondary education. *Psicologia: Teoria e Pesquisa*, 36 e 38.

Gondran, E. e. a. (2022). Analyzing the determinant characteristics for a good performance at ENADE brazilian exam stratified by teaching modality: Face-to-face versus online. In *ICEIS (1)*, pages 234–242.

INEP (2025). Microdados do enade. https://www.gov.br/inep/pt-br/ areas-de-atuacao/avaliacao-e-exames-educacionais/enade/microdados. Acesso em: 30 abr. 2025.

Instituto Brasileiro de Geografia e Estatística (IBGE) (2024). Exame nacional de desempenho dos estudantes (enade) - metadados. https: //ces.ibge.gov.br/pt/base-de-dados/metadados/inep/ exame-nacional-de-desempenho-dos-estudantes-enade. Acessado em: maio 2025.

Kitchenham, B. (2007). Guidelines for performing systematic literature reviews in software engineering. Acesso em: 23 abr. 2025.

Lima, N. C. A., Fagundes, A. d. A., and R. (2020). Educational data mining: A study of the factors that cause school dropout in higher education institutions in brazil.

Lima, P. d. S. N. e. a. (2021). Análise de conteúdo das provas do Enade para os alunos do curso de bacharelado em ciência da computação. Revista *Brasileira de Informática na Educação*, 29:385–413.

Magalhães, Leandro Cunha, S. T. F. M. (2025). Enadequest-um sistema de geração de questões no modelo Enade.*Caderno de Estudos em Sistemas de Informação*, 11(2).

Maretti, Leitao Lucas de Carvalho, V. B. P. R. (2023). Influências socioeconômicas e geográficas no desempenho do enem 2023: Um estudo estatístico e de aprendizado de máquina.

Marques, R., Gouveia, R., Junior, G., and Batista, M. (2023). Aprendizado de máquina para agrupamento e associação de dados do ensino superior público brasileiro.

Neto, F., Silva, R., Gouveia, R., Batista, M., and Oliveira, I. (2020). Computação em nuvem e aprendizado de máquina para análise de grandes volumes de dados educacionais. In *Anais do Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*. Sociedade Brasileira de Computação.

Rosa, E. R., Ferreira, D. J., Silva, N. F. F. d., and Assis, A. (2021). Estudo exploratório através de análises longitudinais aplicado à ciência da computação a partir da base de dados do Enade.

Sakashita, Renan GA, B. D. S. A. L. (2023). Predição do conceito enade dos cursos de computação no brasil. In *Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 1305–1316.

Silva, J. C. P. e. a. (2020). Data analysis of the performance of brazilian higher education medicine courses. In *Proceedings of the 10th Euro-American Conference on Telematics and Information Systems*, pages 1–8.

Silva S F, Carlos Enrique Carrasco Gutierrez, T. C. S. (2023). Percepções e práticas educativas no desempenho acadêmico: uma abordagem machine learning.

Souza, K. E., Neto, M. M., and Lana, C. A. (2024). Classificação curricular das questões do Enade em engenharia de computação: uma mineração de texto. In *Anais do Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 3009–3019, Porto Alegre. SBC.

Souza, K. E., Toledo, P. G. Neto, M. M., Villela, M. M., and Lana, A. C. (2025). C3 miner text: Uma ferramenta para classificação Ao de componentes curriculares de questões do Enade.

Teixeira de J, Alef Aparecido, G. L. A. F. (2020). Um estudo comparativo entre os frameworks de aprendizado de máquina scikit-learning vs. caret: Aplicado no desempenho de notas no enem.

Vieira, A. d. S., Bertolini, D., and Schwerz, A. L. (2022). Análise do desempenho no enredo dos concluintes de computação usando técnica de agrupamento.