Mapeamento Sistemático da Mineração de Dados Educacionais no Combate à Evasão Escolar no Brasil*

Neila T. M. Bastos[1], Lara Gomes[1], Raquel Silveira[1], Carina Oliveira[1]

[1]Programa de Pós-Graduação em Ciência da Computação (PPGCC)
Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE)

{neila.temoteo.matos76,lara.beatriz.soares03}@aluno.ifce.edu.br,
{carina.oliveira,raquel_silveira}@ifce.edu.br

Abstract. *This article presents a Systematic Mapping Study focusing on the use of Educational Data Mining (EDM) and Machine Learning (ML) algorithms for identifying and preventing school dropout in Brazil. A total of 46 national studies (2020–2024) were analyzed, considering methodologies, tools, algorithms, datasets, and types of attributes used. The analysis revealed a predominance of academic and demographic data, with Random Forest and Decision Trees standing out for their performance. Since 2021, the field has shown progress through the use of larger datasets. This work provides a comprehensive overview of the state of the art and highlights gaps for future research.*

Resumo. *Este artigo apresenta um Mapeamento Sistemático da Literatura com foco no uso de Mineração de Dados Educacionais (EDM) e algoritmos de Aprendizado de Máquina (ML) na identificação e prevenção da evasão escolar no Brasil. Foram analisados 46 estudos nacionais (2020-2024), considerando metodologias, ferramentas, algoritmos, bases de dados e tipos de atributos utilizados. A análise revelou predomínio do uso de dados acadêmicos e demográficos, com destaque para o desempenho dos algoritmos Floresta Aleatória e Árvores de Decisão. Observa-se avanço na área a partir de 2021, com uso de bases maiores. Este trabalho oferece uma visão abrangente do estado da arte e aponta lacunas para pesquisas futuras.*

1. Introdução

A evolução tecnológica tem transformado profundamente o cenário educacional, proporcionando avanços significativos na forma como os dados são utilizados para aprimorar o ensino e enfrentar desafios persistentes, como a evasão e a retenção escolar. Ferramentas computacionais modernas permitem a análise de grandes volumes de dados educacionais, oferecendo oportunidades para o desenvolvimento de metodologias inovadoras voltadas à melhoria do desempenho acadêmico. Nesse contexto, a aplicação de tecnologias avançadas no combate à evasão escolar no Brasil tem se mostrado uma estratégia promissora, possibilitando intervenções mais precisas e eficazes para promover a permanência e o sucesso dos estudantes.

A Mineração de Dados Educacionais (*Educational Data Mining* - EDM) e o Aprendizado de Máquina (*Machine Learning* - ML) destacam-se como ferramentas essenciais nesse processo, permitindo a identificação de padrões complexos, a previsão de

*O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

comportamentos e a proposição de soluções personalizadas para o ambiente de aprendizagem [Cechinel and Camargo 2019]. Por meio dessas técnicas, é possível analisar dados acadêmicos, demográficos e socioeconômicos para prever quais estudantes estão em risco de evasão, fornecendo subsídios para ações preventivas. A capacidade de processar grandes bases de dados e extrair informações relevantes torna essas abordagens indispensáveis para otimizar a gestão educacional e melhorar os resultados de aprendizagem.

Apesar do crescente interesse em EDM e ML, há uma notável escassez de referências consolidadas que sintetizem o estado da arte dessas técnicas no contexto brasileiro, especialmente para pesquisadores que estão iniciando seus estudos na área. A ausência de Mapeamentos Sistemáticos da Literatura (do inglês, *Systematic Literature Mapping* - SLM) que abordem especificamente o uso de EDM no combate à evasão escolar no Brasil dificulta a compreensão das tendências, ferramentas e metodologias predominantes, bem como a identificação de lacunas de pesquisa. Essa lacuna representa um obstáculo para o avanço do campo, pois limita a disseminação de boas práticas e a orientação de novos estudos.

Diante desse cenário, este artigo propõe um Mapeamento Sistemático da Literatura para investigar o uso de EDM e ML na identificação e prevenção da evasão escolar no Brasil. A pesquisa analisa 46 estudos nacionais publicados entre 2020 e 2024, com foco nas metodologias, ferramentas, algoritmos, bases de dados e atributos utilizados. O objetivo é oferecer uma visão abrangente do estado atual da pesquisa na área, destacando as principais abordagens e resultados obtidos, além de mapear tendências e lacunas que possam orientar investigações futuras.

Os benefícios desta proposta são múltiplos, tanto para a área educacional quanto para pesquisadores e gestores. Ao consolidar o conhecimento produzido no Brasil, o estudo fornece uma base sólida para analistas de dados, educadores e tomadores de decisão, promovendo o desenvolvimento de estratégias mais eficazes contra a evasão escolar. Além disso, ao identificar lacunas de pesquisa, o trabalho oferece diretrizes para novos estudos, incentivando a inovação e o amadurecimento da aplicação de EDM e ML no contexto educacional brasileiro. Assim, esta pesquisa contribui para fortalecer a interface entre tecnologia e educação.

2. Trabalhos Relacionados

Nesta seção, foram buscados por trabalhos que realizam mapeamento sistemático com foco em mineração de dados educacionais no contexto da evasão escolar.

O artigo [Colpo et al. 2020] realiza um mapeamento a partir das publicações do Congresso Brasileiro de Informática na Educação (CBIE), buscando responder a questões sobre a natureza da evasão (se em nível de curso, disciplina, instituição ou nível de ensino), os dados utilizados (tipos e volume de dados) e as técnicas aplicadas (algoritmos e ferramentas). Os autores destacam o uso expressivo de tarefas de classificação, com predominância de algoritmos baseados em árvores de decisão. Apesar da relevância e da proximidade temática com o presente estudo, a análise se limita às publicações do CBIE e a um período anterior (2006–2019), distinto do foco desta pesquisa (2020–2024).

No mesmo contexto, [Morais et al. 2021] conduz uma revisão sistemática com foco na aplicação de modelos de regressão para prever a evasão no ensino básico. O

estudo busca responder quais os métodos de regressão são utilizados, quais fatores influenciam na evasão escolar no ensino básico, as ações adotadas para mitigar a evasão e as técnicas de machine learning utilizadas para esse contexto. Em um dos resultados obtidos, é identificado 14 fatores influentes, mas restringe-se ao uso de modelos de regressão e não considera especificamente o cenário nacional.

De maneira similar, [Santos et al. 2021], [Silva and Roman 2021] e [Jesus and Gusmão 2024] também investigam aspectos como algoritmos, técnicas e tipos de dados utilizados. Dentre esses, apenas [Santos et al. 2021] realiza uma busca em bases nacionais. Em comum, os três apontam o uso predominante de algoritmos baseados em árvores de decisão.

Embora todos esses trabalhos sejam relevantes e contribuam para os estudos em EDM e evasão escolar, o presente estudo se diferencia por oferecer um mapeamento mais recente, abrangendo um período mais atual e aplicando a pesquisa no cenário nacional.

3. Proposta

Este estudo apresenta um Mapeamento Sistemático da Literatura (SLM) do uso de Mineração de Dados Educacionais (EDM) e algoritmos de Aprendizado de Máquina (ML) na identificação e prevenção da evasão escolar no Brasil. A pesquisa busca caracterizar o estado da arte da área, mapear tendências, destacar boas práticas e apontar lacunas que possam orientar futuras investigações, contribuindo para o avanço da aplicação de EDM e ML no enfrentamento da evasão escolar. O SLM foi estruturado com base em diretrizes consolidadas para revisões sistemáticas, adaptadas às especificidades da área de tecnologia educacional. O método compreende 4 etapas, detalhadas nesta seção. Essa abordagem garante rigor metodológico, transparência e reprodutibilidade na análise.

3.1. Etapa 1: Definição das Questões de Pesquisa (QP)

Nesta etapa, as QP foram formuladas para responder aos objetivos do mapeamento, abrangendo aspectos centrais da aplicação de EDM e ML no combate à evasão escolar. Um processo iterativo de revisão da literatura e análise de demandas práticas foi realizado para garantir a relevância e a abrangência das questões. Assim, as QP elaboradas foram:

- **QP1:** Quais trabalhos científicos utilizam EDM no contexto da evasão escolar?
- **QP2:** Quais as ferramentas/bibliotecas utilizadas pelos trabalhos?
- **QP3:** Quais os algoritmos de aprendizagem de máquina utilizados nos trabalhos?
- **QP4:** Quais os algoritmos com melhor desempenho?
- **QP5:** Quantos registros existem nas bases de dados analisadas?
- **QP6:** Quantos atributos estão presentes nas bases de dados analisadas?
- **QP7:** Quais os tipos de dados analisados nos trabalhos?

3.2. Etapa 2: Identificação de Estudos

Na segunda etapa, foi estruturada a *string* de busca com termos-chave das áreas de tecnologia da informação e educação: ((*Prediction* OR *Classification*) AND "*Student Dropout*" AND ("*Machine Learning*" OR "*Data Mining*" OR "*Data Science*")).

As buscas, realizadas em dezembro de 2024, abrangeram seis repositórios científicos de relevância nacional e internacional: *ACM Digital Library*, *IEEE Xplore*, *ScienceDirect*, *SpringerLink*, *SBC OpenLib* (SOL) e Revista Brasileira de Informática na

Educação (RBIE). Esses repositórios foram selecionados por sua representatividade em publicações de tecnologia educacional e informática, com ênfase em fontes que abrigam estudos brasileiros, como SOL e RBIE. Ao final dessa fase, foram identificados 67 artigos.

3.3. Etapa 3: Seleção e Avaliação dos Estudos

Com o objetivo de selecionar os artigos mais adequados para responder às QP, foram definidos critérios de inclusão e exclusão. Os critérios de inclusão exigiam que os trabalhos fossem nacionais, publicados entre 2020 e 2024, utilizassem EDM e abordassem a evasão escolar em qualquer nível de ensino. Foram excluídos estudos não escritos em português ou inglês, trabalhos em andamento, duplicados ou que estivessem fora do escopo da pesquisa. Após a aplicação desses critérios, 46 artigos foram incluídos na análise.

3.4. Etapa 4: Síntese dos Dados e Apresentação dos Resultados

Nesta etapa, foram desenvolvidas visualizações para responder a cada QP da Seção 3.1, a fim de analisar os atributos de forma mais específica e direcionada. Essas visualizações foram construídas utilizando a ferramenta *Business Intelligence* Power BI, versão Desktop de maio de 2025. Os resultados são apresentados e discutidos na próxima seção.

4. Resultados e Discussões

4.1. QP1: Quais trabalhos científicos utilizam EDM no contexto da evasão escolar?

A Tabela 1 apresenta os 46 trabalhos selecionados. Destaca-se 2021 com mais publicações (12 trabalhos) e 2020 com menos (7 trabalhos).

Tabela 1. Trabalhos nacionais selecionados

ID	Referências	ID	Referências
E1	Sonnenstrahl, T. S., Bernardi, G., and Pertile, S. (2021). **Análise de interações do ambiente virtual de aprendizagem para predição de evasão em cursos no ensino a distância.** EaD em Foco, 11(1).	E7	Silva, D., Tamayo, S., Pessoa, M., Pires, F., Oliveira, D., Oliveira, E., and Carvalho, L. (2020). **Minerando dados de um juiz on-line para prever a evasão de estudantes em disciplinas introdutórias de programação.** XXXI Simpósio Brasileiro de Informática na Educação
E2	Noetzold, E. and de L. Pertile, S. (2021). **Análise e predição de evasão dos alunos de um curso de graduação em sistemas de Informação por meio da mineração de dados educacionais**. RENOTE.	E8	Kantorski, G., Martins, R., Balejo, A., and Frick, M. (2023). **Mineração de Dados Educacionais para Predição da Evasão em Cursos de Graduação Presenciais no Ensino Superior**. XXXIV Simpósio Brasileiro de Informática na Educação
E3	Fonseca Silveira, R., Holanda, M., Ramos, G. N., Victorino, M., and Da Silva, D. (2022). **Analysis of Student Performance and Social-economic Data in Introductory Computer Science Courses at the University of Brasília** 2022 IEEE Frontiers in Education Conference (FIE)	E9	Carvalho, C., Mattos, J., and Aguiar, M. (2023).**Avaliação da interpretabilidade de modelos por meio da clusterização de explicações no contexto da predição de evasão no ensino superior**. XXXIV Simpósio Brasileiro de Informática na Educação.
E4	Silva, J. C. L. (2023). **Definição de modelos de aprendizado de máquina para predição de evasão de alunos do curso técnico** Refas - Revista Fatec Zona Sul.	E10	Falcão, A., Villwock, R., and Miloca, S. (2023). **Análise de dados pré-universidade para predizer a evasão de alunos ingressantes em uma instituição de ensino superior**. XXXIV Simpósio Brasileiro de Informática na Educação
E5	Bitencourt, W. A., Silva, D. M., and Xavier, G. d. C. (2022). **Pode a inteligência artificial apoiar ações contra evasão escolar universitária?**. Ensaio: Avaliação e Políticas Públicas em Educação	E11	Erica Carmo, Gasparini, I., and Oliveira, E. (2022). **Identificação de Trajetórias de Aprendizagem em um Curso de Graduação e sua relação com a Evasão Escolar**. XXXIII Simpósio Brasileiro de Informática na Educação

Tabela 1. Trabalhos nacionais selecionados (continuação)

ID	Referências	ID	Referências
E6	Nóbrega, B. S. d., Maia, J. d. S., Filho, M. A. S., Júnior, E. E. A., and Alves, M. B. (2022). **Sistema para identificação e monitoramento de estudantes em risco de evasão: System for identifying and monitoring students at risk of circumventin**. Brazilian Journal of Development.	E12	Viana, F., Santana, A., and Rabêlo, R. (2022). **Avaliação de Classificadores para Predição de Evasão no Ensino Superior Utilizando Janela Semestral**. XXXIII Simpósio Brasileiro de Informática na Educação
E13	Mathews de, N. S. L., Fachini Gomes, J. B., Holanda, M., Koike, C. C., and Leao Costa, M. T. (2023). **Study on Computer Science Undergraduate Students Dropout at the University of Brasilia** 2023 IEEE Frontiers in Education Conference (FIE)	E25	da Silva Garcia, L. L., Lara, D., Gomes, R., and Cazella, S. (2022). **Mineração de Dados Educacionais na Predição do Desempenho Acadêmico: um prognóstico a partir do percurso curricular realizado**. XXXIII Simpósio Brasileiro de Informática na Educação
E14	Oliveira, J. L., Paula Ambrósio, A., Silva, U., Brancher, J., and Franco, J. J. (2020). **Undergraduate Students' Effectiveness in an Institution With High Dropout Index**. 2020 IEEE Frontiers in Education Conference (FIE)	E26	Nascimento, P., Junior, A. S., Schulz, C., Santos, M., Maciel, A., Rodrigues, R., Nascimento, R., and Alencar, F. (2021). **Análise dos Impactos da Gestão do Tempo no Desempenho Acadêmico Através da Mineração de Dados Educacionais**. XXXII Simpósio Brasileiro de Informática na Educação
E15	Santos, G., Souza, A., Mantovani, R., Cruz, R., Cordeiro, T., and Souza, F. (2024). **An Exploratory Analysis on Gender-Related Dropout Students in Distance Learning Higher Education using Machine Learning**. Association for Computing Machinery	E27	Santos, C. H., Martins, S., and Plastino, A. (2021a). **É Possível Prever Evasão com Base Apenas no Desempenho Acadêmico?**. XXXII Simpósio Brasileiro de Informática na Educação
E16	Villar, A. and de Andrade, C. R. V. (2024). **Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study**. Discover Artificial Intelligence	E28	Santos, J., Sousa, J. D., Mello, R., Cristino, C., and Alves, G. (2021a). **Um Modelo para Análise do Impacto da Retenção e Evasão no Ensino Superior Utilizando Cadeias de Markov Absorventes**. XXXII Simpósio Brasileiro de Informática na Educação
E17	Rabelo, A. and Zárate, L. (2024). **A Model for Predicting Dropout of Higher Education Students**. Data Science and Management	E29	Colpo, M., Primo, T., and Aguiar, M. (2021). **Predição da evasão estudantil: uma análise comparativa de diferentes representações de treino na aprendizagem de modelos genéricos**. SBC
E18	Krüger, J. G. C., de Souza Britto, A., and Barddal, J. P. (2023). **An explainable machine learning approach for student dropout prediction**. Expert Systems with Applications	E30	Souza, A. L. and Braga, A. (2021). **Uma análise dos algoritmos de classificação com base na evasão dos estudantes dos cursos técnicos integrados ao Ensino Médio do Campus Ceres do IF Goiano**. XXXII Simpósio Brasileiro de Informática na Educação
E19	de Jesus, H. O., Rodriguez, L. C., and Costa Junior, A. d. O. (2021). **Predição de Evasão Escolar na Licenciatura em Computação**. Revista Brasileira de Informática na Educação	E31	Brito, B., Mello, R., and Alves, G. (2020). **Identificação de Atributos Relevantes na Evasão no Ensino Superior Público Brasileiro**. XXXI Simpósio Brasileiro de Informática na Educação
E20	Oliveira, R. d. S. and Medeiros, F. P. A. d. (2024). **Modelo de Predição de Evasão Escolar com Base em Dados de Autoavaliação de Cursos de Graduação**. Revista Brasileira de Informática na Educação	E32	Marques, L., Marques, B., Rocha, R., e Silva, L., de Castro, A., and Queiroz, P. G. (2020). **Evasão Acadêmica e suas Causas em Cursos de Bacharelado em Ciência da Computação: Um Estudo de Caso na UFERSA**. XXXI Simpósio Brasileiro de Informática na Educação
E21	Teodoro, L. d. A. and Kappel, M. A. A. (2020). **Aplicação de Técnicas de Aprendizado de Máquina Para Predição de Risco de Evasão Escolar em Instituições Públicas de Ensino Superior no Brasil**. Revista Brasileira de Informática na Educação	E33	Filho, F., Vinuto, T., and Leal, B. (2020). **Análise de Classificadores para Predição de Evasão dos Campi de uma Instituição de Ensino Federal**. XXXI Simpósio Brasileiro de Informática na Educação
E22	Souza, V. F. d. and Santos, T. C. B. d. (2021). **Processo de Mineração de Dados Educacionais aplicado na Previsão do Desempenho de Alunos: Uma comparação entre as Técnicas de Aprendizagem de Máquina e Aprendizagem Profunda**. Revista Brasileira de Informática na Educação	E34	Barbosa, D., Cabral, L., Dwan, F., Feitas, E., and Mello, R. (2023). **Previsão da Evasão Escolar através da Análise de Dados e Aprendizagem de Máquina: Um estudo de caso**. II Workshop de Aplicações Práticas de Learning Analytics em Instituições de Ensino no Brasil
E23	Carneiro, M. G., Dutra, B. L., Paiva, J. G. S., Gabriel, P. H. R., and Araújo, R. D. (2022). **Educational data mining to support identification and prevention of academic retention and dropout: a case study in introductory programming**. Revista Brasileira de Informática na Educação	E35	Oliveira, R., Medeiros, F., and Alves, K. (2023). **Predição de Evasão por meio de um Instrumento Sistemático de Avaliação Institucional**. II Workshop de Aplicações Práticas de Learning Analytics em Instituições de Ensino no Brasil

Tabela 1. Trabalhos nacionais selecionados (continuação)

ID	Referências	ID	Referências
E24	Rodrigues, H., Moraes, L., Santiago, E., Campos, J., Júnior, E. G., Wanderley, G., Garcia, A., Mello, C., Alvares, R., and Santos, R. (2024). **Predicting Student Dropout on the Information Systems Undergraduate Program of UNIRIO Using Decision Trees**. XXXII Workshop sobre Educação em Computação	E36	Teodoro, L., Ferreira, A., and Kappel, M. (2023). **GraduAI – Sistema com Aprendizagem de Máquina para Avaliação de Risco de Evasão**. Anais Estendidos do XII Congresso Brasileiro de Informática na Educação
E37	Freire, J., Landim, F., Moraes, L., Delgado, C., and Pedreira, C. (2024). **Modelo para previsão precoce de abandono de uma disciplina de introdução à programação**. XXXII Workshop sobre Educação em Computação	E42	Schoeffel, P., Ramos, V., and Wazlawick, R. (2020). **A Method to Predict At-risk Students in Introductory Computing Courses Based on Motivation**. Anais Estendidos do IX Congresso Brasileiro de Informática na Educação
E38	Sousa, R., Fachini-Gomes, J., Holanda, M., and Leão, M. (2024). **Um Estudo da Evasão no Curso de Licenciatura em Computação da Universidade de Brasília**. XXXII Workshop sobre Educação em Computação	E43	Magalhães dos Santos, J. K., da Rocha, H. O., Rodrigues Okamura, E. M., Araujo Dias, V. A., Pessoa de Melo, L. H., Oliveira Viana, G. B., Rodrigues, V. C., and da Silva, D. A. (2023). **A Review of IA Use in Education Analysis**. 2023 Workshop on Communication Networks and Power Systems (WCNPS)
E39	Correia, R., Mendonça, H., Silva, C., and Toledo, D. (2024). **Análise dos principais fatores que influenciam a evasão no ensino superior utilizando técnicas de mineração de dados educacionais**. XXXII Workshop sobre Educação em Computação	E44	De Lima, L. M. and Krohling, R. A. (2021). **Discovering an Aid Policy to Minimize Student Evasion Using Offline Reinforcement Learning**. 2021 International Joint Conference on Neural Networks (IJCNN)
E40	Souza, J., Komati, K., and Andrade, J. (2022). **Análise de Sobrevivência: um estudo de caso em um Curso de Sistemas de Informação**. XXX Workshop sobre Educação em Computação	E45	Leite, D., Filho, E., de Oliveira, J. F. L., Carneiro, R. E., and Maciel, A. (2021). **Early detection of students at risk of failure from a small dataset**. 2021 International Conference on Advanced Learning Technologies (ICALT)
E41	Saraiva, D., Pereira, S., Braga, R., and Oliveira, C. (2021). **Análise de Agrupamentos para Caracterização de Indicadores de Evasão**. XXIX Workshop sobre Educação em Computação	E46	Dias, J. C., Da Silva, T. L., Juliatto, M. A., Da Paix˜ao, A. N., and Prata, D. N. (2023). **School dropout in the Federal Network Education of Brazil: is it an inherent individual attribute or it lies on setting conditions?**. 2023 International Symposium on Computers in Education (SIIE)

4.2. QP2: Quais as ferramentas/bibliotecas utilizadas pelos trabalhos?

A Figura 1 apresenta 55 registros de uso de ferramentas e bibliotecas de Mineração de Dados ao longo dos anos, com destaque para 2021, que teve o maior número de ocorrências (20). Os anos de 2020 e 2024 registraram menos usos (8 e 6, respectivamente).

A biblioteca *Scikit-learn*, amplamente utilizada em ML com Python, apareceu em 9 estudos, mantendo presença constante entre 2021 e 2024. A metodologia CRISP-DM foi citada em 6 trabalhos, com destaque para 2021 e o biênio 2023-2024, mostrando sua relevância como estrutura de projeto. *Pandas* (5 usos) teve maior destaque em 2020 e 2021, enquanto *Weka* também foi utilizada 5 vezes, principalmente em 2021 e 2024.

Bibliotecas auxiliares como *NumPy* (2 usos), *Excel*, *Matplotlib*, *Scipy* e *Tableau* (1 uso cada) foram pouco frequentes. A categoria "Outros", que inclui diversas ferramentas não especificadas (ex.: KDD, Shap), apareceu em 13 trabalhos, especialmente em 2021. Além disso, 11 estudos não especificaram as ferramentas usadas, o que compromete a reprodutibilidade dos resultados.

4.3. QP3: Quais os algoritmos de aprendizagem de máquina utilizados nos trabalhos?

A Figura 2 evidencia que o algoritmo mais utilizado foi a Árvore de Decisão (DT), com 27 usos, destacando-se por sua simplicidade e boa performance, especialmente em 2021

Ano	CRISP-DM	Excel	Matplotlib	Não Informado	Numpy	Outros	Pandas	Scikit-learn	Scipy	Tableau	Weka	Total
2020			1	2		1	2	1	1			8
2021	2	1		3	1	5	2	2		1	3	20
2022				3		3		2			1	9
2023	3			3	1	1	1	2			1	12
2024	1					3		2				6
Total	**6**	**1**	**1**	**11**	**2**	**13**	**5**	**9**	**1**	**1**	**5**	**55**

Figura 1. Ferramentas e bibliotecas utilizadas pelos trabalhos.

e 2024 (7 usos cada). A seguir, a Floresta Aleatória (*Random Forest* - RF) apareceu 23 vezes, com maior incidência em 2022 (6). O *Naive Bayes* (NB), eficiente para dados categóricos e textos, teve 14 usos, principalmente entre 2020 e 2022. O *Support Vector Machine* (SVM) foi moderadamente usado (12 vezes), possivelmente limitado pelo alto custo computacional. Técnicas de *Boosting* cresceram, somando 12 usos, com destaque em 2024 (4 usos), indicando maior adoção de métodos *ensemble*. Redes Neurais foram menos frequentes (7 usos), talvez por sua complexidade. Algoritmos tradicionais como Regressão Logística (LR), *K-Nearest Neighbors* (KNN) e *MultiLayer Perceptron* (MLP) tiveram uso moderado (7 a 10 vezes). A categoria "Outros" agrupou 14 algoritmos não especificados. Por ano, observa-se um surgimento do uso de MLP a partir de 2021 (5 trabalhos) que decai depois (2 trabalhos em 2022 e 1 em 2023). Em 2022, o uso foi mais equilibrado, e em 2023-2024 houve diversificação, com maior foco em *Boosting* e DT.

Ano	Boost	DT	GB	KNN	LR	MLP	Não Informado	NB	Outros	Redes Neurais	RF	SVM	Total
2020	2	2	1	2			1	3	2	1	3	2	**19**
2021	2	7	1		2	5	1	5	6	2	5	4	**40**
2022	2	5	2	3	1	4	2	4	4		6	2	**35**
2023	2	6		2	2	1		2	1	3	4	1	**24**
2024	4	7	1		4				1	1	5	3	**26**
Total	**12**	**27**	**5**	**7**	**9**	**10**	**4**	**14**	**14**	**7**	**23**	**12**	**144**

Figura 2. Algoritmos utilizados pelos trabalhos.

4.4. QP4: Quais os algoritmos com melhor desempenho?

A Figura 3 mostra os algoritmos de ML com melhor desempenho nos estudos. O algoritmo RF lidera, indicado por 9 artigos entre 2021 e 2024, refletindo sua eficácia e popularidade. Algoritmos baseados em árvores (DT, RF, *Boosting*) predominam, somando 17 usos, evidenciando preferência por métodos interpretáveis. *Boosting* ganha destaque em 2024 (3 artigos), sugerindo tendência de crescimento. SVM tem baixa frequência (2 artigos), possivelmente pela complexidade e necessidade de ajuste. *Extra Trees Classifier* (ETC), *Generalized Linear Model* (GLM) e KNN aparecem pouco representativos.

Ano	Boost	DT	ETC	GLM	KNN	Não informado	RF	SVM	Total
2020			1						**1**
2021		2				1	2	1	**6**
2022		1		1	1		2	1	**6**
2023	1	1				1	4		**7**
2024	3						1		**4**
Total	**4**	**4**	**1**	**1**	**1**	**2**	**9**	**2**	**24**

Figura 3. Melhores algoritmos apontados pelos trabalhos.

4.5. QP5: Quantos registros existem nas bases de dados analisadas?

A Figura 4 apresenta a quantidade de artigos por faixa de tamanho da amostra, distribuídos por ano (de 2020 a 2024). A faixa mais recorrente ao longo dos anos foi a de 1001 a 5000 amostras, com destaque em 2021 (6 artigos) e 2022 (4 artigos), isso sugere uma preferência por bases de dados de porte médio a grande, equilibrando disponibilidade de dados e viabilidade de análise. Em alta nos anos mais recentes, a faixa acima de 5000 amostras, utilizada em 5 artigos em 2023 e 6 artigos em 2024, indica uma tendência crescente de uso de grandes bases de dados, possivelmente relacionadas a sistemas educacionais completos ou bancos de dados integrados. Amostras menores ou iguais a 100 ou até 500 amostras, foram pouco frequentes e concentradas entre 2020 e 2022, o que pode refletir estudos de caso, pilotos ou pesquisas com menor acesso a dados. Aparecem em todos os anos, com pelo menos 1 artigo por ano omitindo o tamanho da amostra, o que pode comprometer a avaliação da robustez metodológica da parte dos estudos. Destaca-se, portanto, que há um aumento no uso de amostras maiores ao longo do tempo, especialmente a partir de 2021 o que pode indicar uma melhora no acesso a dados educacionais, como também uma maior preocupação com validade estatística dos resultados.

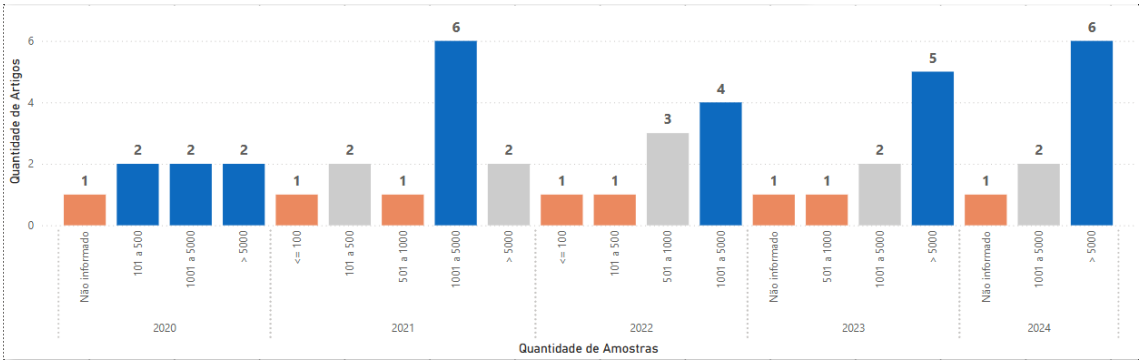

Figura 4. Tamanho da base

4.6. QP6: Quantos atributos estão presentes nas bases de dados analisadas?

A Figura 5 responde a QP6 mostrando informações sobre a quantidade de atributos presentes nas bases de dados utilizadas nos trabalhos selecionados. A faixa de quantidade de atributos mais frequente foi a entre 11 e 20 atributos (14 artigos), o que indica uma preferência dos pesquisados por um número moderado de variáveis, possivelmente equilibrando complexidade e interpretabilidade. A faixa acima de 40 atributos (9 artigos) também representa uma parcela significativa, mostra que alguns estudos optam por uma

abordagem mais robusta, com grande volume de dados. Estudos mais simples ou com foco restritos limitaram-se a utilizar até 10 atributos (6 artigos). Entre 31 e 40 atributos (5 artigos), foi pouco frequente, mas ainda representativo para estudos mais detalhados. A faixa menos utilizada foi a que utilizou entre 21 e 30 atributos (4 artigos), possivelmente evitada por ser uma zona intermediária que nem garante simplicidade nem profundidade analítica. Tivemos um número considerável de artigos (8) que não informaram o tamanho de suas amostras, o que pode indicar uma falta de detalhamento metodológico ou limitações na apresentação dos dados. A presença de artigos sem essa informação pode dificultar a avaliação da robustez de certos estudos. A maioria dos artigos utiliza entre 11 e 20 atributos, sugerindo uma preferência por modelos de análise com um número controlado de variáveis. Contudo, há também uma boa quantidade de estudos que trabalham com um número elevado de atributos (mais de 40), o que pode indicar o uso de técnicas mais avançadas, como algoritmos de aprendizado de máquina.

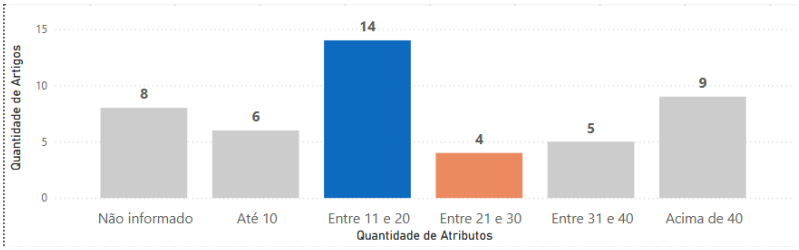

Figura 5. Atributos

4.7. QP7: Quais os tipos de dados analisados nos trabalhos?

A Figura 6 mostra a quantidade de artigos segundo os tipos de dados utilizados, totalizando mais que 46, pois alguns trabalhos analisaram múltiplos tipos. A maior parte (24 artigos) utiliza dados acadêmicos, como notas, frequência e histórico escolar, refletindo o foco principal das pesquisas. Dados demográficos, como idade, sexo, etnia e local de residência, aparecem em 18 estudos, indicando interesse no perfil dos alunos. Dados socioeconômicos, envolvendo renda familiar e ocupação dos pais, são considerados em menor escala (8 artigos). Métricas de desempenho externo, como avaliações padronizadas, foram usadas em 5 artigos, e dados sociais, como redes de relacionamento e convivência escolar, aparecem em 3. Um artigo não especificou os dados usados, e 6 se enquadram numa categoria genérica, que pode incluir variáveis institucionais ou ambientais.

Figura 6. Tipos de dados

5. Conclusões

O estudo analisou 46 trabalhos nacionais sobre Mineração de Dados Educacionais focados na evasão escolar, com visualização de resultados no Power BI. Observou-se diversidade de ferramentas, com destaque para bibliotecas Python (Scikit-learn, Pandas, Numpy) e uso frequente da metodologia CRISP-DM. Dos trabalhos apresentados, 23,91% não informam detalhadamente as ferramentas utilizadas.

A maioria dos artigos utiliza entre 11 e 20 atributos, sugerindo uma preferência por modelos de análise com um número controlado de variáveis. Contudo, há também uma boa quantidade de estudos que trabalham com um número elevado de atributos (mais de 40), o que pode indicar o uso de técnicas mais avançadas, como algoritmos de aprendizado de máquina. A presença de artigos sem essa informação pode dificultar a avaliação da robustez de certos estudos.

A diversidade de algoritmos observada ao longo dos anos indica um amadurecimento técnico nas pesquisas analisadas, especialmente a partir de 2021, com o aumento do uso de técnicas de *Machine Learning*. A Floresta Aleatória se destaca como o algoritmo mais utilizado, o que se justifica por sua eficácia em tarefas de classificação e sua interpretabilidade, especialmente em contextos educacionais. Algoritmos baseados em árvores (como Árvores de Decisão, Floresta Aleatória e *Boosting*) são predominantes, refletindo uma preferência por modelos compreensíveis. O crescimento do uso de técnicas ensemble nos últimos anos também reforça essa evolução metodológica. Quanto aos dados utilizados, há predominância de informações acadêmicas e demográficas, revelando um foco na análise do histórico escolar e perfil do estudante, enquanto dados sociais e de desempenho aparecem menos, indicando uma possível lacuna para investigações futuras.

Os estudos estão evoluindo em quantidade e qualidade, com uso crescente de métodos mais sofisticados, bases maiores e ferramentas mais modernas. Porém, ainda há lacunas importantes na descrição de metodologias, como linguagem usada e ferramentas, o que compromete a reprodutibilidade.

Referências

Cechinel, C. and Camargo, S. (2019). *Capítulo 12 Mineração de dados educacionais: avaliação e interpretação de modelos de classificação*.

Colpo, M., Primo, T., Pernas, A., and Cechinel, C. (2020). Mineração de dados educacionais na previsão de evasão: uma rsl sob a perspectiva do congresso brasileiro de informática na educação. In *XXXI Simpósio Brasileiro de Informática na Educação*, pages 1102–1111, Porto Alegre, RS, Brasil. SBC.

Jesus, J. A. d. and Gusmão, R. P. d. (2024). Investigação da evasão estudantil por meio da mineração de dados e aprendizagem de máquina: Um mapeamento sistemático. *Revista Brasileira de Informática na Educação*, 32:807–841.

Morais, F., Melo, A., Moutinho, M., and Fagundes, R. (2021). Modelos de regressão aplicados na previsão da evasão escolar do ensino básico: uma revisão sistemática da literatura. In *XXXII Simpósio Brasileiro de Informática na Educação*, pages 168–178, Porto Alegre, RS, Brasil. SBC.

Santos, V., Saraiva, D., and Oliveira, C. (2021). Uma análise de trabalhos de mineração de dados educacionais no contexto da evasão escolar. In *Anais do XXXII Simpósio*

Brasileiro de Informática na Educação, pages 1196–1210, Porto Alegre, RS, Brasil. SBC.

Silva, J. and Roman, N. (2021). Predicting dropout in higher education: a systematic review. In *Anais do XXXII Simpósio Brasileiro de Informática na Educação*, pages 1107–1117, Porto Alegre, RS, Brasil. SBC.