Perfil das Bases de Dados Nacionais na Área de Mineração de Dados Educacionais*

Lara Gomes, João Linhares, Neila Bastos, Raquel Silveira, Carina Oliveira

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

{lara.beatriz.soares03,joao.linhares09}@aluno.ifce.edu.br,
{neila,carina.oliveira,raquel_silveira}@ifce.edu.br

Abstract. *Educational Data Mining (EDM) has emerged as an effective approach for investigating challenges within the educational context. School dropout, in particular, is a recurring phenomenon that affects many educational institutions and has been widely studied using EDM and Machine Learning techniques. The quality of such analyses largely depends on the selection of appropriate data sources — those that are comprehensive, accessible, rich in relevant attributes, and sufficiently large. In this context, this study aims to map and characterize the data sources used in EDM research focused on school dropout. The analysis considers aspects such as dataset size, number of attributes, availability (public or private), and types of data used.*

Resumo. *A Mineração de Dados Educacionais (MDE) tem se consolidado como uma abordagem eficaz para investigar desafios no contexto educacional. A evasão escolar, por sua vez, é um fenômeno recorrente que afeta diversas instituições de ensino e tem sido amplamente estudada com o apoio de técnicas de MDE e Machine Learning. A qualidade das análises realizadas depende, em grande parte, da escolha de bases de dados adequadas — que sejam abrangentes, acessíveis, ricas em atributos relevantes e com volume suficiente. Neste contexto, este trabalho propõe mapear e caracterizar as fontes de dados utilizadas em estudos de MDE voltados à evasão escolar. São considerados aspectos como o tamanho das bases, quantidade de atributos, disponibilidade (pública ou privada) e os tipos de dados analisados.*

1. Introdução

A Mineração de Dados (MD, do inglês *Data Mining*) tem se consolidado como uma ferramenta estratégica para explorar grandes volumes de dados e extrair conhecimentos relevantes em diferentes áreas do saber. No contexto educacional, essa abordagem dá origem à Mineração de Dados Educacionais (MDE), um campo de estudo voltado à análise de dados provenientes de ambientes educacionais [Baker et al. 2011], com o objetivo de apoiar decisões, identificar padrões e propor intervenções que melhorem a qualidade da educação, como demonstrado em [Santos et al. 2025]. Um dos temas mais recorrentes

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

nesse campo é a evasão escolar — um fenômeno complexo que representa a saída prematura de estudantes das instituições de ensino, sem a conclusão do curso, e que impacta diretamente nos indicadores de desempenho e na permanência estudantil, além de prejuízos econômicos e organizacionais para as instituições de ensino [Silva and Sampaio 2022].

Para que estudos em MDE possam ser realizados de maneira eficaz, a disponibilidade e a qualidade das bases de dados utilizadas são fatores determinantes. As fontes de dados variam entre bases institucionais privadas, como registros acadêmicos internos das instituições de ensino, e bases públicas, como os Censos da Educação Básica (principal instrumento de coleta de informações da educação básica [INEP 2025b]) e Superior (instrumento de pesquisa mais completo do Brasil sobre as instituições de educação superior [INEP 2025a]) disponibilizados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). Além disso, é possível construir bases próprias por meio do cruzamento de diferentes fontes ou da aplicação de instrumentos como questionários.

A aplicação de algoritmos de Aprendizado de Máquina (do inglês *Machine Learning*) na predição da evasão escolar tem ganhado destaque, como demonstram os estudos de [Kantorski et al. 2023a] e [Oliveira and Medeiros 2024]. De modo complementar, análises exploratórias realizadas em trabalhos como [Martins et al. 2023] têm contribuído para a compreensão dos fatores associados ao abandono escolar. Apesar das abordagens distintas, esses estudos compartilham um elemento essencial: a utilização de dados. Iniciar uma pesquisa nessa área exige não apenas conhecimento do contexto, mas também acesso a dados relevantes, consistentes e de qualidade. A coleta de boas bases de dados é uma etapa crítica do processo científico, pois define os limites e as possibilidades da análise. Bases bem estruturadas permitem a extração de evidências confiáveis, a construção de modelos mais precisos e a identificação de padrões com maior poder explicativo. Por outro lado, a ausência de dados adequados compromete a validade dos resultados e pode levar a conclusões equivocadas.

A etapa de busca por bases de dados, portanto, não é trivial. Muitas vezes, pesquisadores enfrentam dificuldades para encontrar fontes acessíveis, atualizadas e adequadas aos objetivos do estudo. Além disso, nem sempre há um guia claro sobre quais bases já foram utilizadas na área, quais suas características e limitações. Como demonstrado no mapeamento sistemático realizado por [Nascimento et al. 2024], dos 65 trabalhos analisados sobre mineração de dados e aprendizado de máquina aplicados à evasão estudantil, apenas 23% utilizavam bases abertas, 20% bases privadas (de universidades), e 57% sequer informavam a origem dos dados. Esse cenário evidencia não apenas a carência de dados abertos, mas também a necessidade de maior transparência e padronização na divulgação das fontes utilizadas.

Neste contexto, o presente trabalho propõe, por meio de uma revisão sistemática da literatura, analisar e caracterizar as bases de dados utilizadas em estudos que abordam a evasão escolar com o uso de MDE no contexto nacional. Para apoiar a análise dos resultados, o Power BI foi utilizado como ferramenta para construir visualizações que respondem de maneira objetiva às questões de pesquisa formuladas. A partir de informações como tamanho das bases, quantidade de atributos, disponibilidade (pública ou privada) e tipos de dados analisados (acadêmicos, socioeconômicos, demográficos etc.), pretende-se traçar um panorama das fontes de dados utilizadas no Brasil.

Essa caracterização visa subsidiar futuras pesquisas na etapa de identificação de bases de dados adequadas, além de incentivar a transparência, o reuso e a documentação de dados na área educacional. Ao fornecer um mapeamento claro das fontes mais utilizadas e suas características, este trabalho pode servir como um guia para novos pesquisadores, reduzir barreiras de acesso à informação e fortalecer a produção científica na área, promovendo estudos mais robustos, replicáveis e com maior potencial de impacto nas políticas públicas de combate à evasão escolar.

2. Trabalhos Relacionados

Nesta seção foram buscados estudos relacionados à caracterização das bases de dados utilizadas em pesquisas de MDE, com foco específico na evasão escolar. Além disso, também foram consideradas revisões sistemáticas que abordam esse mesmo tema.

Batista e Fagundes [Batista and Fagundes 2023] realizaram uma revisão sistemática com 35 estudos entre 2010 e 2022, destacando algoritmos como *Decision Tree*, *Naive Bayes* e *Random Forest*, e a importância de atributos educacionais, demográficos e comportamentais. Contudo, o estudo não apresenta uma categorização sistemática das bases de dados nem detalha sua origem, estrutura ou disponibilidade pública. Por outro lado, o presente trabalho, visa mapear e caracterizar de forma aprofundada bases nacionais aplicadas à evasão escolar, contribuindo para uma abordagem mais prática e localizada.

Colpo et al. [Colpo et al. 2024] também realizaram revisão sistemática focada na predição da evasão escolar, destacando o uso predominante de dados acadêmicos, demográficos e econômicos de estudantes de graduação em instituições públicas, com destaque para algoritmos *ensemble*, como *Random Forest*. O estudo aponta a escassez de bases de dados oriundas de países em desenvolvimento, incluindo o Brasil, onde a MDE teria grande impacto. Apesar da análise sólida das técnicas, a caracterização das bases permanece superficial, sem detalhar estrutura, acessibilidade ou abrangência.

Santos et al. [Santos et al. 2021c] analisaram 50 trabalhos nacionais e internacionais sobre MDE no contexto da evasão escolar, incluindo publicações em português e diversos níveis de ensino. O estudo destaca ferramentas, algoritmos e bases de dados, utilizando o software Tableau para visualizações. Contudo, não aprofunda a análise das bases, deixando de discutir sua disponibilidade pública, tipos de dados e estrutura, além de não sistematizar as bases para facilitar sua reutilização. Tal limitação compromete a replicação e expansão dos estudos.

Complementando essa discussão, [Marques et al. 2024] apresentam um estudo de caso aplicado a dados do Instituto Federal do Piauí, utilizando algoritmos de aprendizado de máquina para identificar padrões associados à evasão escolar. Esse trabalho exemplifica o uso efetivo de bases institucionais brasileiras, reforçando a importância da curadoria de dados e da seleção criteriosa de atributos para a construção de modelos preditivos mais precisos. Os autores também destacam o potencial da MDE como ferramenta estratégica para apoiar políticas de permanência e melhoria do desempenho acadêmico.

Apesar das contribuições relevantes dos estudos mencionados, observa-se que poucos se dedicam a uma análise detalhada das bases de dados em si: suas origens, disponibilidade, tipos de dados e limitações. Nesse sentido, o presente trabalho propõe um avanço ao preencher essa lacuna por meio da caracterização sistemática das bases utilizadas em pesquisas de MDE no contexto nacional.

3. Metodologia

Este trabalho apresenta uma análise do perfil das bases de dados utilizadas em estudos de Mineração de Dados Educacionais com foco na evasão escolar, no período de 2020 a 2024. A metodologia adotada baseou-se, inicialmente, em uma Revisão Sistemática da Literatura (RSL), cujas etapas são descritas na Seção 3.1. As etapas subsequentes à RSL correspondem à Estruturação e Análise dos dados.

3.1. Etapa 1: Revisão Sistemática da Literatura

Como etapa inicial do trabalho, foi realizada uma RSL, abrangendo artigos publicados em eventos e periódicos nacionais e internacionais[1]. A escolha pela RSL se justifica por sua capacidade de identificar trabalhos já desenvolvidos, delimitar o problema de pesquisa e apontar novas possibilidades de investigação [Brizola and Fatin 2017].

O objetivo principal desta revisão é selecionar e reunir informações relevantes sobre o tema proposto, com foco no contexto educacional brasileiro. As etapas de planejamento, a estratégia de busca adotada, bem como os critérios de inclusão e exclusão, são detalhados ao longo desta subseção.

3.1.1. Planejamento

Esta etapa consistiu no planejamento e elaboração das Questões de Pesquisa (QP) norteadoras do estudo do trabalho. Posteriormente, foram definidos os eventos e periódicos para coleta dos artigos no foco da mineração de dados educacionais no contexto de evasão escolar. As QP definidas fazem referência exclusivamente a atributos das bases de dados utilizadas nos trabalhos coletados. São elas:

- **QP1)** Quais artigos foram publicados por ano?
- **QP2)** Qual a disponibilidade das bases de dados?
- **QP3)** Quais as bases de dados?
- **QP4)** Qual o tamanho das bases de dados?
- **QP5)** Quais os tipos de dados analisados?
- **QP6)** Qual a quantidade de atributos analisados?

3.1.2. Estratégia de Busca

Nesta etapa, foi definida a estratégia de busca para a identificação dos estudos de MDE no contexto da evasão escolar. A busca foi realizada em bases internacionais amplamente reconhecidas, como *ACM Digital Library*, *SpringerLink*, *ScienceDirect* e *IEEE Xplore*, bem como em publicações nacionais vinculadas à Sociedade Brasileira de Computação (SBC). A busca foi realizada no período de novembro de 2024.

Para a busca nas bases internacionais, uma *string* foi aplicada diretamente nas ferramentas de busca de cada base, com filtros por título, resumo e palavras-chave, quando disponíveis. Foram definidos termos-chave que compuseram a seguinte *string* de busca:

[1]Consideram-se as publicações internacionais que tratem de estudos realizados no contexto nacional.

```
(("Full Text & Metadata":prediction of students OR school
dropout OR school retention or school failure) AND ("Full Text &
Metadata":educational data mining OR knowledge discovery OR machine
learning) AND ("Full Text & Metadata":  institution or university))
```

No contexto nacional, foram selecionados eventos e periódicos de relevância na área de Informática na Educação, sendo eles: o Congresso Brasileiro de Informática na Educação (CBIE), a Revista Brasileira de Informática na Educação (RBIE), o Simpósio Brasileiro de Informática na Educação (SBIE), Workshop de Informática na Escola (WIE), Workshop de Aplicações Práticas de *Learning Analytics* em Instituições de Ensino no Brasil (WAPLA) e o Workshop sobre Educação em Computação (WEI). A busca dos trabalhos nessas fontes foi realizada manualmente, por meio da análise dos anais publicados em cada edição dos eventos e periódicos.

3.1.3. Critérios de Inclusão e Exclusão

Durante a etapa de seleção dos estudos na revisão sistemática, foram definidos critérios de inclusão e exclusão a fim de delimitar de forma clara o escopo da pesquisa, além de garantir a relevância temática e a atualidade das publicações analisadas.

Foram **incluídos** os trabalhos que:

- Utilizam técnicas de Mineração de Dados Educacionais;
- Possuem como foco principal o estudo da evasão escolar;
- Apresentam informações sobre as bases de dados utilizadas, seja em termos de origem, estrutura ou tipo de dados analisados (acadêmicos, socioeconômicos, etc.);
- Estão disponíveis integralmente para leitura e análise.

Foram **excluídos** os trabalhos:

- Publicados fora do período de 2020 a 2024;
- Fora do escopo da pesquisa (ex.: trabalhos que não abordam evasão escolar ou não utilizam MDE);
- Não analisam o contexto nacional;
- Duplicados.

Esses critérios foram aplicados ao longo da triagem dos títulos, resumos e textos completos dos trabalhos encontrados. A Figura 1 apresenta o diagrama PRISMA, utilizado para resumir o processo de identificação, triagem, elegibilidade e inclusão dos trabalhos na RSL. O processo iniciou com 2.557 registros encontrados pelos repositórios selecionados e ao final do processo de seleção 35 estudos foram considerados relevantes e incluídos na análise.

3.2. Etapa 2: Estruturação dos Dados

Após a seleção dos trabalhos, foi elaborada uma planilha para estruturar a coleta das informações extraídas dos artigos. As colunas da planilha contemplaram os seguintes campos: Portal científico de publicação do artigo, Ano de Publicação, Título, Autores, Origem da Base de Dados (instituição ou fonte responsável), Quantidade de Amostras (tamanho da base utilizada), Quantidade de Atributos, Disponibilidade da Base (pública

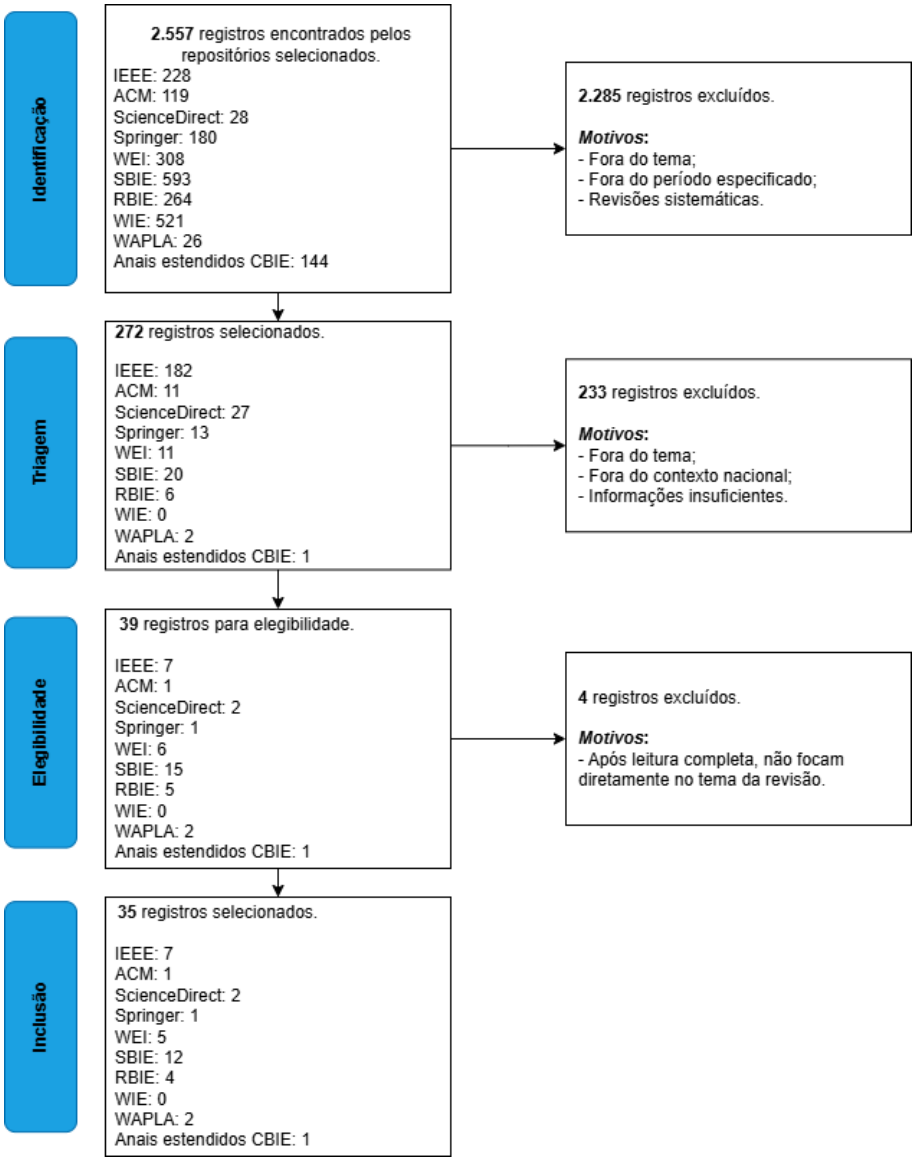

Figura 1. Diagrama PRISMA

ou privada) e Tipos de Dados analisados. Com essa estrutura definida, os dados foram inseridos à medida que as informações eram identificadas nos textos originais.

Durante esse processo, observou-se que muitas informações apresentavam variações de nomenclatura ou estavam descritas de forma não padronizada, o que exigiu um esforço de padronização. Para garantir consistência na análise, foram definidos intervalos numéricos para categorizar o tamanho das bases e o número de atributos, além da padronização dos tipos de dados em categorias como: acadêmicos, demográficos, socioeconômicos, sociais e outros.

3.3. Etapa 3: Análise dos Dados

A análise dos dados foi realizada por meio da ferramenta de visualização Power BI[2]. O uso de ferramentas como o Power BI facilita a visualização e interpretação das informações,

[2]https://www.microsoft.com/pt-br/power-platform/products/power-bi

permitindo a extração de *insights* relevantes para o contexto estudado. A planilha previamente preenchida e padronizada foi carregada na ferramenta e serviu como base para a construção de visualizações que auxiliam na resposta às QP definidas na Seção 3.1.1. A próxima seção apresenta os resultados obtidos, seguidos de suas respectivas discussões.

4. Resultados e Discussões

Nesta seção são apresentadas as respostas de cada questão de pesquisa da Seção 3.1.1. Para cada QP, é apresentada uma visualização que ilustra a resposta da questão.

4.1. QP1) Quais artigos foram publicados por ano?

A Figura 2 mostra a distribuição por ano dos trabalhos publicados na temática. A Tabela 1 apresenta os artigos que passaram pelo processo de triagem e foram selecionados para o estudo deste trabalho. Observa-se um número maior de publicações no ano de 2023 (9 artigos), seguido pelo ano de 2021 e 2024 (8 artigos). A produção científica na área se manteve relativamente constante e com um certo crescimento ao longo dos anos, demonstrando que a temática de MDE e evasão escolar permanecem atuais e relevantes.

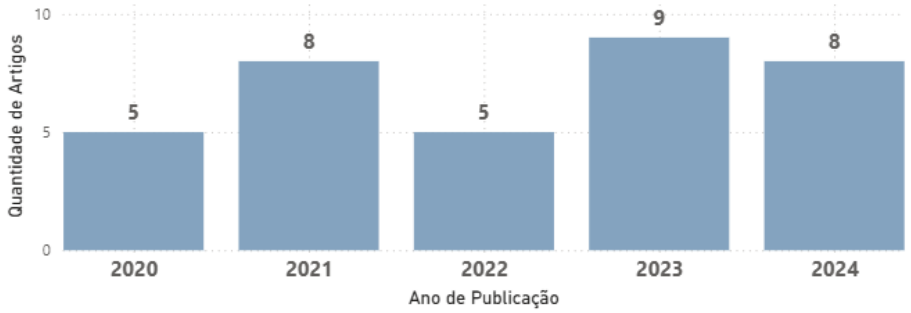

Figura 2. Quantidade de artigos publicados por ano.

Ano	Artigo
2020	[Brito et al. 2020] [Filho et al. 2020] [Marques et al. 2020] [Oliveira et al. 2020] [Teodoro and Kappel 2020]
2021	[Colpo et al. 2021] [de Jesus et al. 2021] [De Lima and Krohling 2021] [Leite et al. 2021] [Saraiva et al. 2021] [Santos et al. 2021a] [Santos et al. 2021b] [Souza and Braga 2021]
2022	[Carneiro et al. 2022] [Êrica Carmo et al. 2022] [Fonseca Silveira et al. 2022] [Souza et al. 2022] [Viana et al. 2022]
2023	[Barbosa et al. 2023] [Carvalho et al. 2023] [Dias et al. 2023] [Falcão et al. 2023] [Kantorski et al. 2023b] [Magalhães dos Santos et al. 2023] [Mathews de et al. 2023] [Oliveira et al. 2023] [Teodoro et al. 2023]
2024	[Correia et al. 2024] [Krüger et al. 2023] [Oliveira and Medeiros 2024] [Rabelo and Zárate 2024] [Rodrigues et al. 2024] [Santos et al. 2024] [Sousa et al. 2024] [Villar and de Andrade 2024]

Tabela 1. Trabalhos coletados.

4.2. QP2) Qual a disponibilidade das bases de dados?

A Figura 3 e a Tabela 2 ilustram a distribuição dos artigos segundo a disponibilidade das bases de dados utilizadas: privadas, públicas, NI (Não Informadas – artigos que não especificaram a disponibilidade da base) e ambas (artigos que utilizaram simultaneamente bases públicas e privadas). Observa-se um número expressivo de estudos que utilizaram bases privadas, o que pode estar associado ao acesso direto de pesquisadores a registros acadêmicos internos, frequentemente armazenados em bancos de dados institucionais ou a questionários aplicados pelos pesquisadores.

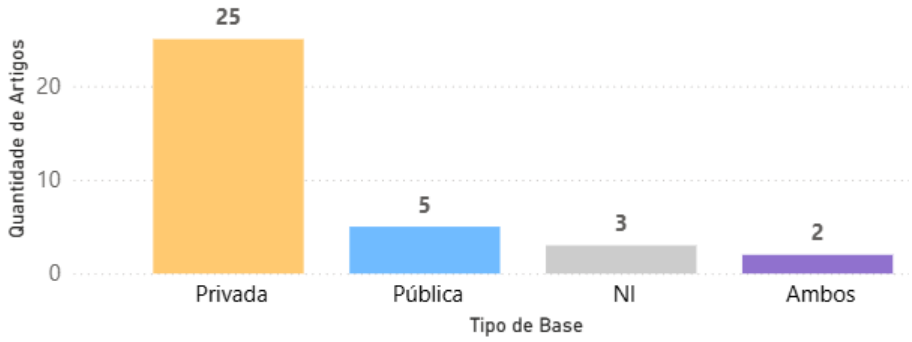

Figura 3. Disponibilidade das Bases.

Disponibilidade	Artigo
Privada	[Oliveira et al. 2020] [Marques et al. 2020] [de Jesus et al. 2021] [Saraiva et al. 2021] [Santos et al. 2021a] [Colpo et al. 2021] [De Lima and Krohling 2021] [Leite et al. 2021] [Fonseca Silveira et al. 2022] [Carneiro et al. 2022] [Souza et al. 2022] [Êrica Carmo et al. 2022] [Viana et al. 2022] [Mathews de et al. 2023] [Carvalho et al. 2023] [Falcão et al. 2023] [Barbosa et al. 2023] [Oliveira et al. 2023] [Villar and de Andrade 2024] [Rabelo and Zárate 2024] [Krüger et al. 2023] [Oliveira and Medeiros 2024] [Rodrigues et al. 2024] [Sousa et al. 2024] [Correia et al. 2024]
Pública	[Teodoro and Kappel 2020] [Filho et al. 2020] [Teodoro et al. 2023] [Magalhães dos Santos et al. 2023] [Dias et al. 2023]
Não informado	[Santos et al. 2021b] [Kantorski et al. 2023b] [Santos et al. 2024]
Ambos	[Brito et al. 2020] [Souza and Braga 2021]

Tabela 2. Trabalhos coletados por disponibilidade da base.

4.3. QP3) Quais as bases de dados?

Conforme ilustrado na Figura 4, foram identificadas 19 instituições ou entidades como origem das bases de dados utilizadas nos estudos. A figura revela que a maioria dos tra-

balhos (57,1%) utilizou bases privadas provenientes de instituições de ensino brasileiras (institutos e universidades federais).

Já as bases públicas, representadas na figura pelo MEC e IFCE, incluem fontes como o INEP[3], a Plataforma Nilo Peçanha (PNP)[4] e o IFCE em Números[5], cujos dados são disponibilizados abertamente por meio de painéis interativos ou conjuntos de microdados. A ampliação do acesso a bases públicas de qualidade pode contribuir significativamente para o avanço da área e para a democratização das investigações em MDE, especialmente entre pesquisadores com menos acesso a sistemas institucionais.

O INEP, através do Plano de Dados Abertos, sistematiza informações sobre o sistema educacional brasileiro em diferentes níveis de ensino, da educação básica e superior [INEP 2020]. Diversas bases de dados são disponibilizadas pelo INEP: indicadores educacionais (ex.: formação docente, gestão escolar, taxas de rendimento); painéis interativos; microdados (censos da educação superior e escolar, Enade, Enem); entre outras fontes estatísticas da educação brasileira.

A Plataforma Nilo Peçanha é um ambiente virtual de coleta, validação e disseminação das estatísticas oficiais da Rede Federal de Educação Profissional, Científica e Tecnológica (Rede Federal) [MEC 2025]. A plataforma reúne dados sobre o corpo docente, discente, técnico-administrativo e gastos financeiros de cada unidade da Rede Federal através de painéis interativos e em formato *.csv* disponíveis para análise. Ambas as instituições fornecem documentação clara que orienta a interpretação dos atributos.

O IFCE Em Números corresponde a um conjunto de painéis que utiliza registros acadêmicos do Instituto Federal do Ceará para apresentar informações sobre o ensino, evasão, origem, ingresso e trajetória de desempenho dos estudantes. A plataforma, criada utilizando o Tableau[6], possui dados de 2013 até o momento atual e permite a aplicação de diversos filtros para especificar as análises.

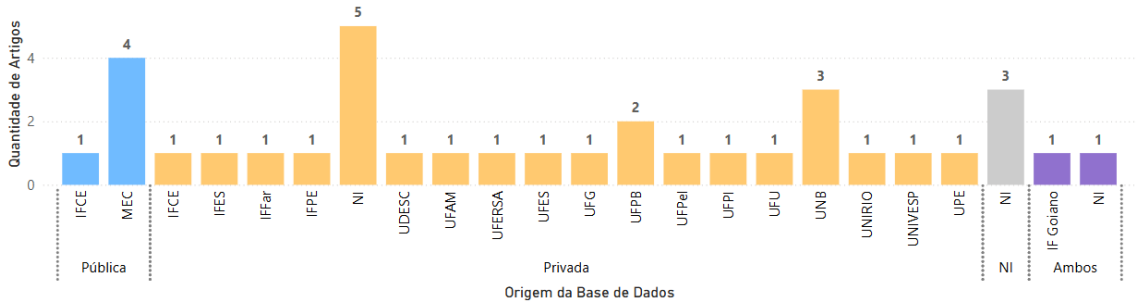

Figura 4. Origem das Bases de Dados Utilizadas.

4.4. QP4) Qual o tamanho das bases de dados?

A Figura 5 apresenta uma visão do tamanho das bases de dados, associando-as à sua disponibilidade (pública ou privada) e à base utilizada. O tamanho das bases foi agrupado em seis categorias: até 100 registros; de 101 a 500; de 501 a 1.000; de 1.001 a 5.000;

[3]https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos
[4]https://www.gov.br/mec/pt-br/pnp
[5]https://emnumeros.ifce.edu.br/
[6]https://www.tableau.com/pt-br/products/tableau

acima de 5.000; e NI (Não Informado – trabalhos que não especificaram o tamanho da base utilizada).

Essas bases apresentam grande variabilidade de tamanho, abrangendo desde menos de 100 até mais de 5.000 registros. Entre as bases públicas, destacam-se aquelas provenientes do INEP e da Plataforma Nilo Peçanha, ambas capazes de fornecer mais de 5.000 registros. Já o estudo com dados da plataforma pública IFCE em Números não informou o tamanho da base utilizada. No caso das bases construídas por meio de questionários, observa-se que um estudo alcançou entre 1.001 e 5.000 registros. Esse tipo de base pode variar bastante em tamanho, dependendo da metodologia adotada.

O uso de um volume adequado de registros é fundamental para que as técnicas e algoritmos aplicados nos estudos consigam aprender de forma eficaz com os dados de treinamento, possibilitando maior capacidade de generalização e melhor desempenho na fase de teste.

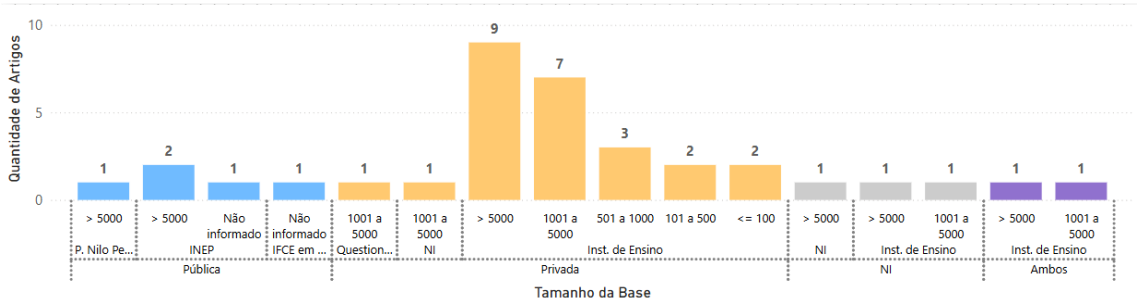

Figura 5. Tamanho das Bases de Dados.

4.5. QP5) Quais os tipos de dados analisados?

Os tipos de dados identificados, apresentados na Figura 6, incluem dados acadêmicos (ex.: disciplinas cursadas, período letivo, tipo de cota), demográficos (ex.: sexo, cor/raça, idade) socioeconômicos (ex.: renda familiar/per capita, se o estudante trabalha), de desempenho (ex.: notas, coeficiente de rendimento), sociais, outros tipos e os classificados como Não Informados (NI).

Conforme ilustrado na Figura 6, a maioria das bases permite a extração de dados acadêmicos e demográficos. Destacam-se as bases institucionais, que possibilitam o acesso a uma maior variedade de informações. No entanto, as bases públicas, como as do INEP e da PNP, também oferecem ampla diversidade de dados. A principal diferença observada é que, entre os estudos analisados, não foram identificados usos de dados de desempenho provenientes dessas bases públicas.

É importante destacar que muitos trabalhos não se restringiram a um único tipo de dado. A combinação de diferentes categorias, ou seja, uma abordagem multidimensional, permite uma análise mais completa dos fatores que influenciam a evasão escolar. Essa prática torna os estudos mais fundamentados, contribuindo para uma compreensão mais aprofundada do problema e para o desenvolvimento de estratégias de intervenção mais eficazes.

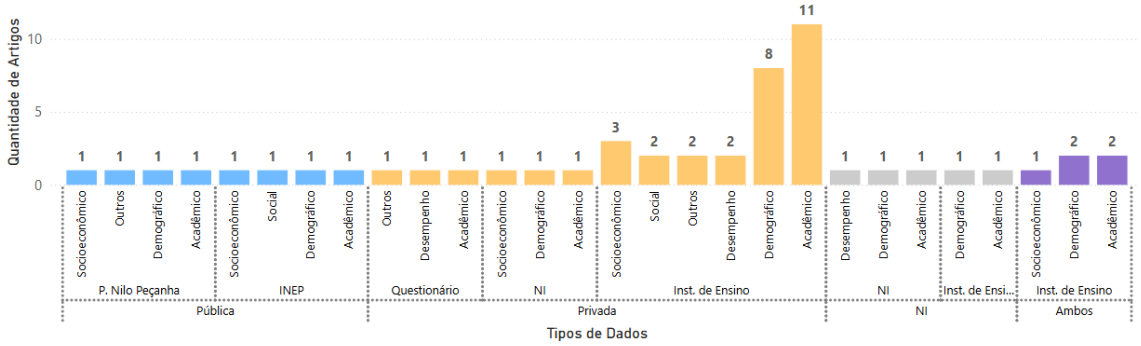

Figura 6. Tipos de Dados das Bases Utilizadas.

4.6. QP6) Qual a quantidade de atributos analisados?

A Figura 7, traz uma visão sobre a quantidade de atributos/características analisados nos trabalhos. A quantidade de atributos também foi agrupada em 6 categorias: Até 10 atributos; Entre 11 e 20; Entre 21 e 30; Entre 31 e 40; Acima de 40 atributos e os NI (Não informado – trabalhos que não especificaram a quantidade de atributos utilizados).

Nesta análise, observa-se que, entre as bases privadas, a maior diversidade de atributos está presente nas bases oriundas de instituições de ensino. Nessas bases, os trabalhos coletados utilizaram desde menos de 10 até mais de 40 atributos, sendo mais comum o uso entre 11 e 20 atributos. Entre as bases públicas, também se verificou certa diversidade, embora a maioria dos estudos tenha utilizado até 20 atributos. Destaca-se que, com a base do INEP, foi possível coletar mais de 40 atributos para análise.

A utilização de uma diversidade adequada de atributos é essencial para que os modelos de análise possam capturar com maior precisão os padrões presentes nos dados, contribuindo para a construção de modelos mais robustos e para a caracterização de indicadores que influenciam a evasão escolar.

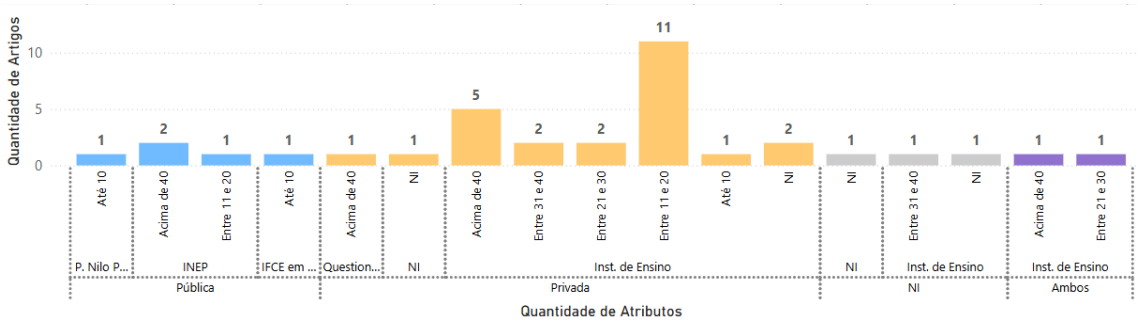

Figura 7. Quantidade de Atributos das Bases Utilizadas

5. Considerações Finais

Este trabalho teve como objetivo caracterizar as bases de dados utilizadas em estudos de Mineração de Dados Educacionais (MDE) no contexto da evasão escolar. Por meio de uma Revisão Sistemática da Literatura, foram encontrados 2.557 registros nos repositórios selecionados e, ao final, foram selecionados 35 trabalhos cujas informações foram

analisadas e organizadas com o propósito de compreender as características das bases utilizadas nas pesquisas.

As bases identificadas abrangem tanto fontes públicas quanto privadas, incluindo registros acadêmicos internos das instituições de ensino, bem como dados públicos disponibilizados pelo INEP e pela Plataforma Nilo Peçanha. Observou-se que a maioria dos trabalhos utiliza bases privadas (71,4%), o que, em geral, proporciona acesso a conjuntos de dados com maior volume, variedade de atributos e diversidade de tipos de dados. No entanto, as bases públicas também se mostraram relevantes e abrangentes. Estudos baseados nessas fontes conseguiram trabalhar com uma quantidade expressiva de registros e atributos, refletindo o esforço contínuo de coleta promovido por instituições como o INEP, por meio dos censos da educação básica e superior.

Os resultados reforçam a importância da disponibilidade e da qualidade das bases de dados para o desenvolvimento de estudos robustos em MDE. A caracterização apresentada neste trabalho não apenas resume informações relevantes sobre as fontes de dados utilizadas, mas também oferece um ponto de partida para futuros pesquisadores que enfrentam desafios para obter dados confiáveis e relevantes.

Referências

Baker, R., Isotani, S., and Carvalho, A. (2011). Mineração de dados educacionais: Oportunidades para o Brasil. *Revista Brasileira de informática na educação*, 19(02):03.

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. In *Anais do II Workshop de Aplicações Práticas de Learning Analytics em Instituições de Ensino no Brasil*, pages 42–50, Porto Alegre, RS, Brasil. SBC.

Batista, J. and Fagundes, M. (2023). Revisão sistemática sobre mineração de dados educacionais com foco em desempenho acadêmico. *Revista Brasileira de Informática na Educação*, 29(1):45–62.

Brito, B., Mello, R., and Alves, G. (2020). Identificação de atributos relevantes na evasão no ensino superior público brasileiro. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, pages 1032–1041, Porto Alegre, RS, Brasil. SBC.

Brizola, J. and Fatin, N. (2017). Revisão da lietratura e revisão sistemática da literatura. *Revista de Educação do Vale do Arinos - RELVA*, 3(2).

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*, 30:379–395.

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. In *Anais do XXXIV Simpósio Brasileiro de Informática na Educação*, pages 1191–1201, Porto Alegre, RS, Brasil. SBC.

Colpo, A., Rodrigues, L., and Teixeira, B. (2024). Revisão sistemática sobre a aplicação da MDE na predição de evasão escolar. *Journal of Educational Data Mining*, 16(1):15–34.

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. In *Anais do XXXII Simpósio Brasileiro de Informática na Educação*, pages 873–884, Porto Alegre, RS, Brasil. SBC.

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. In *Anais do XXXII Workshop sobre Educação em Computação*, pages 830–841, Porto Alegre, RS, Brasil. SBC.

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*, 29:255–272.

De Lima, L. M. and Krohling, R. A. (2021). Discovering an aid policy to minimize student evasion using offline reinforcement learning. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Dias, J. C., Da Silva, T. L., Juliatto, M. A., Da Paixão, 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? In *2023 International Symposium on Computers in Education (SIIE)*, pages 1–10.

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. In *Anais do XXXIV Simpósio Brasileiro de Informática na Educação*, pages 1293–1304, Porto Alegre, RS, Brasil. SBC.

Filho, F., Vinuto, T., and Leal, B. (2020). Análise de classificadores para predição de evasão dos campi de uma instituição de ensino federal. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, pages 1132–1141, Porto Alegre, RS, Brasil. SBC.

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. In *2022 IEEE Frontiers in Education Conference (FIE)*, pages 1–8.

INEP (2020). PolÍtica e plano de dados abertos do INEP. Acesso em: 06-06-2025.

INEP (2025a). Censo da educação superior. Disponível em: https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-da-educacao-superior. Acessado: 03-06-2025.

INEP (2025b). Censo escolar. Disponível em: https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-escolar. Acessado: 03-06-2025.

Kantorski, G., Martins, R., Balejo, A., and Frick, M. (2023a). Mineração de dados educacionais para predição da evasão em cursos de graduação presenciais no ensino superior. In *Anais do XXXIV Simpósio Brasileiro de Informática na Educação*, pages 1133–1142, Porto Alegre, RS, Brasil. SBC.

Kantorski, G., Martins, R., Balejo, A., and Frick, M. (2023b). Mineração de dados educacionais para predição da evasão em cursos de graduação presenciais no ensino su-

perior. In *Anais do XXXIV Simpósio Brasileiro de Informática na Educação*, pages 1133–1142, Porto Alegre, RS, Brasil. SBC.

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*, 233:120933.

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. In *2021 International Conference on Advanced Learning Technologies (ICALT)*, pages 42–46.

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. In *2023 Workshop on Communication Networks and Power Systems (WCNPS)*, pages 1–7.

Marques, J., Carvalho, A., and Silva, R. (2024). Modelo preditivo aplicado à evasão no IFPI utilizando MDE. *Revista Brasileira de Tecnologia Educacional*, 11(1):88–105.

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. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, pages 1042–1051, Porto Alegre, RS, Brasil. SBC.

Martins, C., Lacerda, F., Carmo, I., Silva, E., Alves, T., and Campos, R. (2023). Análise exploratória sobre evasão tardia da graduação de uma universidade pública. In *Anais do VIII Congresso sobre Tecnologias na Educação*, pages 31–40, Porto Alegre, RS, Brasil. SBC.

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. In *2023 IEEE Frontiers in Education Conference (FIE)*, pages 1–7.

MEC (2025). Plataforma nilo peçanha. https://www.gov.br/mec/pt-br/pnp. Acesso em: 06-06-2025.

Nascimento, F. F. d., Dantas, L. C. d. O., Castro, A. F. d., and Queiroz, P. G. G. (2024). Técnicas de mineração de dados e aprendizado de máquina aplicados à evasão estudantil: um mapeamento sistemático da literatura. *Revista Brasileira de Informática na Educação*, 32:270–294.

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. In *2020 IEEE Frontiers in Education Conference (FIE)*, pages 1–7.

Oliveira, R., Medeiros, F., and Alves, K. (2023). Predição de evasão por meio de um instrumento sistemático de avaliação institucional. In *Anais do II Workshop de Aplicações Práticas de Learning Analytics em Instituições de Ensino no Brasil*, pages 118–127, Porto Alegre, RS, Brasil. SBC.

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*, 32:1–21.

Rabelo, A. M. and Zárate, L. E. (2024). A model for predicting dropout of higher education students. *Data Science and Management*, 8(1):72–85.

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. In *Anais do XXXII Workshop sobre Educação em Computação*, pages 588–598, Porto Alegre, RS, Brasil. SBC.

Santos, C. H., Martins, S., and Plastino, A. (2021a). É possível prever evasão com base apenas no desempenho acadêmico? In *Anais do XXXII Simpósio Brasileiro de Informática na Educação*, pages 792–802, Porto Alegre, RS, Brasil. SBC.

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. In *Proceedings of the 20th Brazilian Symposium on Information Systems*, SBSI '24, New York, NY, USA. Association for Computing Machinery.

Santos, J., Sousa, J. D., Mello, R., Cristino, C., and Alves, G. (2021b). Um modelo para análise do impacto da retenção e evasão no ensino superior utilizando cadeias de markov absorventes. In *Anais do XXXII Simpósio Brasileiro de Informática na Educação*, pages 813–823, Porto Alegre, RS, Brasil. SBC.

Santos, L. C. B., Schafer, A. G., Oliveira, E. S. d., Costa, V. G. d. J. S. D., Lima, M., and Souza, J. D. d. (2025). Mineração de dados educacionais com python: descobertas e aplicações para a melhoria da qualidade de ensino. *Cuadernos de Educación y Desarrollo*, 17(5):e8380.

Santos, V., Saraiva, D., and Oliveira, C. (2021c). 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.

Saraiva, D., Pereira, S., Braga, R., and Oliveira, C. (2021). Análise de agrupamentos para caracterização de indicadores de evasão. In *Anais do XXIX Workshop sobre Educação em Computação*, pages 238–247, Porto Alegre, RS, Brasil. SBC.

Silva, P. T. d. F. e. and Sampaio, L. M. B. (2022). Políticas de permanência estudantil na educação superior: reflexões de uma revisão da literatura para o contexto brasileiro. *Revista de Administração Pública*, 56(5):603–631.

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. In *Anais do XXXII Workshop sobre Educação em Computação*, pages 715–725, Porto Alegre, RS, Brasil. SBC.

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. In *Anais do XXXII Simpósio Brasileiro de Informática na Educação*, pages 1276–1285, Porto Alegre, RS, Brasil. SBC.

Souza, J., Komati, K., and Andrade, J. (2022). Análise de sobrevivência: um estudo de caso em um curso de sistemas de informação. In *Anais do XXX Workshop sobre Educação em Computação*, pages 392–403, Porto Alegre, RS, Brasil. SBC.

Teodoro, L., Ferreira, A., and Kappel, M. (2023). GraduAI – sistema com aprendizagem de máquina para avaliação de risco de evasão. In *Anais Estendidos do XII Congresso Brasileiro de Informática na Educação*, pages 84–95, Porto Alegre, RS, Brasil. SBC.

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*, 28:838–863.

Viana, F., Santana, A., and Rabêlo, R. (2022). Avaliação de classificadores para predição de evasão no ensino superior utilizando janela semestral. In *Anais do XXXIII Simpósio Brasileiro de Informática na Educação*, pages 908–919, Porto Alegre, RS, Brasil. SBC.

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*, 4(1):2.

Êrica Carmo, Gasparini, I., and Oliveira, E. (2022). Identificação de trajetórias de aprendizagem em um curso de graduação e sua relação com a evasão escolar. In *Anais do XXXIII Simpósio Brasileiro de Informática na Educação*, pages 323–333, Porto Alegre, RS, Brasil. SBC.