

## **Mineração de Dados da Educação Básica Brasileira Usando as Bases do INEP:**

### **Uma Revisão Sistemática da Literatura**

Raimundo de Castro Soares, UFMA, soares.raimundo@ufma.br, 0000-0002-0904-0370

Nelson Weber Neto, UFMA, nelsonweberneto@gmail.com, 0000-0002-1136-768X

Luciano Reis Coutinho, UFMA, luciano.rc@ufma.br, 0000-0001-7996-7334

Francisco José da Silva e Silva, UFMA, fssilva@lsdi.ufma.br, 0000-0001-8339-3679

Davi Viana dos Santos, UFMA, davi.viana@lsdi.ufma.br, 0000-0003-0470-549X

Ariel Soares Teles, IFMA/UFMA/UFDPar, ariel.teles@ifma.edu.br, 0000-0002-0840-3870

**Resumo:** A análise de problemas educacionais brasileiros através da mineração de dados tem se tornado cada vez mais presente em estudos científicos. Neste contexto, esta Revisão Sistemática da Literatura (RSL) tem o objetivo de identificar os estudos focados em minerar os dados da educação básica brasileira produzidos pelo Instituto Nacional de Estudos Pedagógicos Anísio Teixeira (INEP). Os artigos foram obtidos por meio de buscas em fontes de artigos nacionais e internacionais, retornando um total de 410 estudos, dos quais 19 atenderam aos critérios de seleção. Adicionalmente, buscou-se entender como ocorre o processo de mineração desses dados. Os resultados mostram que os pesquisadores se interessam por esses dados para estudar os problemas da educação básica brasileira, sugerindo melhorias e tomadas de decisão para abordá-los.

**Palavras-chave:** Mineração de Dados, INEP, IDEP, SAEB, Censo Escolar, Indicadores Educacionais.

### **Data Mining of Brazilian Basic Education Using the INEP Bases: A Systematic Literature Review**

**Abstract:** The analysis of Brazilian educational problems through data mining has become increasingly present in scientific studies. In this context, this Systematic Literature Review (SLR) aims to identify studies focused on mining the data of basic Brazilian education acquired by the Brazilian Institute of Studies and Educational Research Anísio Texeira (INEP). The articles were obtained by searching national and international article sources, so returning a total of 410 studies, of which 19 met the selection criteria. Additionally, we sought to understand how the data mining process occurs. The results shown that researchers are interested in such data to study the problems of Brazilian education, hence suggesting improvements and decision making to address them.

**Keywords:** Data Mining, INEP, IDEP, SAEB, School Census, Educational Indicators.

## **1. Introdução**

O Instituto Nacional de Estudos Pedagógicos Anísio Teixeira\*(INEP), um órgão do Ministério da Educação do Brasil, é responsável pela realização de avaliações e questionários para a coleta de dados educacionais relevantes. Estas coletas de dados são direcionadas ao corpo docente, corpo discente e gestores de todos os níveis de ensino. Ao serem analisados de forma minuciosa, esses dados podem mostrar o diagnóstico educacional sob vários aspectos, como a taxa de eficiência, infraestrutura escolar, formação dos professores, complexidade da gestão escolar, nível socioeconômicos dos alunos, dentre outros.

Todos os anos, o INEP realiza o Censo Escolar, sendo um importante instrumento de coleta de informações da educação básica e pesquisa estatística educacional brasileira.

---

\*(<<https://www.gov.br/inep/pt-br>>)

A educação básica compreende a educação infantil, ensino fundamental I e II, e ensino médio. Para a realização do Censo, o INEP conta com a colaboração das secretarias de educação municipais e estaduais, e devem participar escolas públicas e privadas de todas as etapas da educação básica e profissional. O Censo Escolar contribui para o INEP elaborar os Indicadores Educacionais.

O INEP também é responsável pela realização do Sistema de Avaliação da Educação Básica (SAEB) e do Índice de Desenvolvimento da Educação Básica (IDEB). As avaliações aplicadas pelo SAEB acontecem a cada dois anos. Provas de duas disciplinas são aplicadas, Língua Portuguesa e Matemática, para estudantes do 5º ano do ensino fundamental I, 9º ano do ensino fundamental II, e 3º ano do ensino médio.

O IDEB é um indicador que mede a qualidade da educação básica brasileira, calculado a partir da média da proficiência das provas de Língua Portuguesa e Matemática aplicadas pelo SAEB, padronizada entre 0 e 10, e o indicador de rendimento baseado na taxa de aprovação dos alunos. Além das avaliações, os questionários do SAEB também são aplicados aos alunos, professores e gestores da educação. Eles visam obter dados sobre a infraestrutura das escolas e aspectos administrativos.

Todos os dados produzidos por meio desses mecanismos são disponibilizados pelo INEP através de uma página para acesso a dados abertos<sup>†</sup>. O volume de dados disponível é grande e tem sido analisado para gerar conhecimento para gestores, permitindo a realização de tomadas de decisão baseadas em evidências. Considerando isso, diversas pesquisas em mineração de dados educacionais têm sido conduzidas.

Devido aos trabalhos nesta área, é necessário agregar o conhecimento já publicado. A realização de Revisões Sistemáticas de Literatura (RSLs) pode auxiliar nesta agregação. A RSL realizada por (Coelho e Silveira, 2017) analisou a aplicação de *Deep Learning* na mineração de dados educacionais e análise da aprendizagem. (Santos; Ferreira e Miranda, 2017) conduziram uma RSL sobre dados educacionais abertos no Brasil. Já o mapeamento sistemático realizado por (Maschio *et al.*, 2018) abordou os estudos focados em mineração de dados educacionais do Brasil publicados em eventos nacionais. Diferente dessas revisões, a RSL reportada nesta pesquisa objetiva identificar os estudos focados em minerar os dados produzidos pelo INEP (i.e., os dados do SAEB, IDEB, Censo Escolar e Indicadores Educacionais), visando entender como este processo é realizado para estudar e sugerir melhorias aos problemas da educação básica brasileira.

O restante deste artigo está organizado como segue. A Seção 2 descreve a metodologia utilizada na condução desta RSL. A Seção 3 descreve os resultados da revisão, enquanto a Seção 4 os discute. Por fim, a Seção 5.

## 2. Metodologia

A produção dessa RSL foi baseada no protocolo desenvolvido por Kitchenham e Charters (Kitchenham e Charters, 2007). O guia estabelece três etapas de uma RSL: planejamento, condução e produção do relatório. Essas etapas foram suportadas por uma ferramenta online de condução colaborativa de RSLs, o *Parsif.al*<sup>‡</sup>.

### 2.1. Questões de Pesquisa

Para alcançar o objetivo dessa RSL, as seguintes questões de pesquisa (QP) foram formuladas.

- (QP1) Quais dados do INEP (IDEB, SAEB, Censo Escolar ou Indicadores Educacionais) têm sido minerados?

<sup>†</sup> <https://www.gov.br/inep/pt-br/aceso-a-informacao/dados-abertos>

<sup>‡</sup> <https://parsif.al/>

- (QP2) Quais problemas da educação básica brasileira estão sendo abordados?
- (QP3) Quais metodologias de mineração de dados estão sendo usadas nos estudos?
- (QP4) Quais as técnicas estão sendo empregadas na mineração de dados?

## 2.2. Estratégia de Busca

Duas strings de busca foram elaboradas e aplicadas, conforme a Tabela 1. Os termos utilizados nas strings foram escolhidos visando maximizar a quantidade de artigos sobre o assunto. No processo de seleção das fontes de artigos, foi dada prioridade para as bibliotecas digitais, revistas e eventos científicos com escopo da área de informática na educação. A pesquisa foi realizada no dia 17 de abril de 2021 em todas as fontes. Após os artigos terem sido obtidos, a etapa de remoção de duplicados foi feita com o suporte da ferramenta *Parsif.al*.

Tabela 1. Strings de busca utilizadas.

String de Busca	Fontes de Artigos
("data analytics" OR "data mining" OR "data analysis") AND (SAEB OR INEP OR IDEB OR "basic education" OR "school census" OR "educational indicators") AND (Brazil OR Brazilian)	ACM Digital Library IEEE Digital Library ScienceDirect Web of Science
(dados OR data) AND (SAEB OR INEP OR IDEB OR "educação básica" OR "basic education" OR "censo escolar" OR "indicadores educacionais")	Workshops do Congresso Brasileiro de Informática na Educação (CBIE) Revista Brasileira de Informática na Educação (RBIE) Revista Novas Tecnologias na Educação (RENOTE) Simpósio Brasileiro de Informática na Educação (SBIE)

## 2.3. Critérios de Seleção

Os critérios de seleção dos artigos são especificados na Tabela 2. De maneira independente, dois pesquisadores foram responsáveis por aplicar os critérios para selecionar os artigos. Inicialmente, eles realizaram a leitura dos títulos e resumos para selecionar os artigos. Ao final deste primeiro passo, o coeficiente de Kappa de Cohen (Cohen, 1968) foi calculado para identificar o nível de concordância entre os pesquisadores em relação aos resultados obtidos. Os artigos conflitantes foram então reavaliados de forma a buscar um consenso entre os pesquisadores. Quando necessário, um terceiro pesquisador contribuiu com discussões para sanar conflitos.

Tabela 2. Critérios de inclusão e exclusão de artigos.

Critérios de Inclusão (CI)	Critérios de Exclusão (CE)
(CI1): Artigos que realizam mineração dos dados do SAEB, IDEB, Censo Escolar e Indicadores Educacionais	(CE1): Artigos que realizam mineração de dados que não sejam do SAEB, IDEB, Censo Escolar e Indicadores Educacionais
(CI2): Artigos em inglês ou português	(CE2): Artigos em idioma diferente do inglês e português
(CI3): Artigos completos	(CE3): Literatura cinza (e.g., teses, dissertações, artigos curtos, capítulos de livro, relatórios técnicos)
	(CE4): Texto completo não disponível online

## 2.4. Avaliação de Qualidade

A qualidade dos artigos selecionados foi avaliada respondendo as cinco questões de qualidade (QQ) listadas abaixo. A pontuação dada a cada uma, conforme as respostas, foi a seguinte: sim = 1 ponto, parcialmente = 0,5 ponto, e não = pontuação zerada. A análise de qualidade foi realizada por dois pesquisadores.

- (QQ1): O estudo apresenta explicitamente o problema de pesquisa?

- (QQ2): O estudo apresenta claramente o seu objetivo?
- (QQ3): Existe uma descrição adequada do contexto em que o estudo foi realizado?
- (QQ4): A mineração dos dados foi feita de maneira rigorosa?
- (QQ5): O estudo apresenta novos achados?

### 2.5. Extração de Dados

Com o objetivo de responder às questões de pesquisa, os dados apresentados abaixo foram extraídos dos artigos selecionados. Esta fase de extração de dados foi realizada por um pesquisador e verificada por outros pesquisadores com experiência na condução de RSLs, informática na educação e mineração de dados.

1. Título, ano e fonte do artigo.
2. Dados do INEP (i.e., se SAEB, IDEB, Censo Escolar ou Indicadores Educacionais relacionados a alunos, escola, professor e notas do IDEB).
3. Problema abordado.
4. Metodologia de mineração de dados.
5. Técnica empregada na mineração de dados.

### 3. Resultados

A Figura 1 ilustra o fluxo do processo de revisão. A pesquisa nas bases de dados resultou um total de 410 publicações, com apenas 3 artigos removidos por serem duplicados. Houve então a aplicação dos critérios de inclusão e exclusão aos 407 artigos. Após a seleção realizada independentemente por cada pesquisador, o coeficiente de Kappa de Cohen foi calculado, buscando verificar o grau de concordância. O resultado do Kappa foi de aproximadamente 0,5, o que representa uma concordância moderada (Viera e Garrett, 2005). Portanto, os artigos foram reavaliados pelos pesquisadores envolvidos na fase de seleção até chegarem a um consenso. Ao final, um total de 19 artigos foram selecionados.

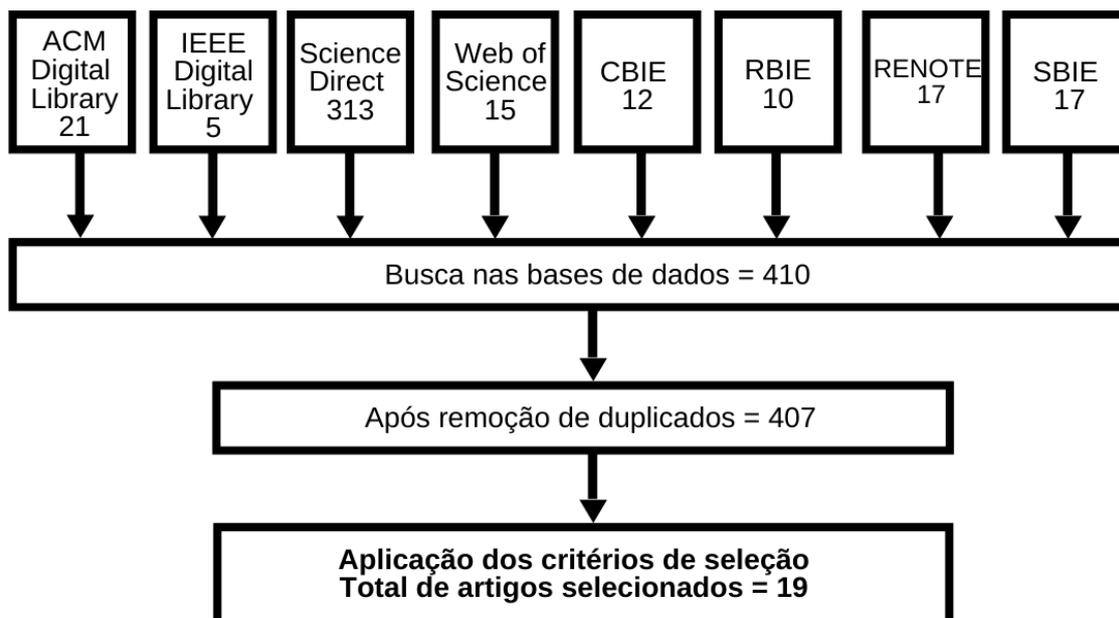


Figura 1. Fluxo do processo de condução desta revisão.

A Tabela 3 apresenta, em ordem cronológica de publicação, uma caracterização

dos 19 estudos selecionados nesta RSL, com as seguintes informações: referências, dados analisados (i.e., SAEB, IDEB, Censo Escolar ou Indicadores Educacionais), os problemas estudados, as metodologias e as técnicas utilizadas.

Tabela 3. Caracterização dos estudos selecionados.

Referência	Dados	Problema	Metodologia	Técnica
A1 (Soares, 2006)	SAEB	Baixo desempenho escolar	Análise exploratória	Análise estatística
A2 (Waltenberg e Vandenberghe, 2007)	SAEB	Baixo desempenho escolar	Análise exploratória	Algoritmo EOP e Modelo de Regressão
A3 (Ramos; Machado e Cordeiro, 2015)	IDEB e SAEB	Fragilidade do sistema educacional	Análise exploratória	Análise de agrupamento
A4 (Penteado, 2016a)	SAEB	Melhoria de desempenho de alunos	Análise exploratória	Algoritmo POKS
A5 (Penteado, 2016b)	IDEB	Baixo desempenho escolar	Análise exploratória	Análise de correlação
A6 (Júnior <i>et al.</i> , 2017)	Censo	Baixo desempenho escolar	KDD	Análise de correlação e identificação de <i>outliers</i>
A7 (Carvalho; Cruz e Gouveia, 2017)	Censo	Evasão de alunos	KDD	Árvore de decisão e Classificação Bayesiana
A8 (Bem; Pereira e Souza, 2017)	IDEB	Falta de padronização entre bases educacionais	Proposta de solução	Ferramentas OLAP
A9 (Avila <i>et al.</i> , 2018)	Indicadores	Falta de padronização entre bases educacionais	Framework LDIF	Linked data mashup
A10 (Nascimento e Júnior, 2018)	Indicadores	Formação docente de baixa qualidade	CRISP-DM	Modelos de regressão
A11 (Nascimento; Junior e Fagundes, 2018)	Indicadores	Evasão e reprovação de alunos	CRISP-DM	Análise de correlação e Modelos de regressão
A12 (Nascimento; Fagundes e Maciel, 2019)	Indicadores	Aproximação do desempenho previsto das taxas de eficiência	CRISP-DM	Modelos de Regressão
A13 (Pinto; Júnior e Costa, 2019)	SAEB	Baixo Desempenho escolar	CRISP-DM	Algoritmos de classificação
A14 (Junior <i>et al.</i> , 2019)	Indicadores	Reprovação de alunos	Análise preditiva	Modelos de regressão
A15 (Júnior <i>et al.</i> , 2019)	IDEB e SAEB	Gestão de baixa qualidade e baixo desempenho escolar	CRISP-DM	Modelos de regressão e Árvore de decisão
A16 (Pinto <i>et al.</i> , 2019)	SAEB	Baixo desempenho escolar	CRISP-DM	Algoritmos de classificação
A17 (Pacini, 2020)	SAEB	Baixo desempenho escolar	Análise exploratória	Análise de correlação
A18 (Santos e Medeiros, 2020)	IDEB	Baixo investimento na educação	KDD	Análise de correlação e Modelo de regressão
A19 (Silva; Souza e Cysneiros, 2021)	SAEB e Censo	Qualidade do processo de mineração de dados	Proposta de solução	Modelo de regressão

Para o preenchimento da Tabela 3, algumas informações extraídas dos artigos tiveram que ser categorizadas e padronizadas. Ao relatar sobre os problemas educacionais abordados pelos estudos, o baixo desempenho escolar é uma categorização usada para os estudos que não somente buscaram explicações para o baixo rendimento das escolas nos resultados do SAEB e IDEB, mas também sugerir melhorias.

As metodologias utilizadas pelos artigos foram: *Cross Industry Standard Process for Data Mining* (CRISP-DM), *Knowledge Discovery in Databases* (KDD), e a metodologia *Linked Data Integration Framework* (LDIF), uma ferramenta para a construção de aplicações que utilizam dados relacionados (LDIF, 2021). Além disso,

alguns estudos não definiram explicitamente uma denominação para a metodologia utilizada. Para esses estudos, categorizamos as metodologias de três formas: análise exploratória, usada para consultar a estrutura implícita dos dados e aprender sobre os relacionamentos entre as diversas variáveis; análise preditiva, cujo objetivo é analisar dados e classificá-los ou prever comportamentos futuros (e.g., notas futuras do IDEB); e dois artigos não especificaram exatamente uma metodologia, mas propuseram soluções para serem utilizadas na mineração de dados educacionais.

Na coluna que lista as técnicas empregadas na mineração dos dados educacionais, as ferramentas *Online Analytical Processing* (OLAP) se referem a um conjunto de técnicas e aplicações direcionadas ao acesso e análise *ad-hoc* de dados, tendo como objetivo transformar dados em informações para auxiliar no processo de decisão.

A Figura 2 apresenta o resultado da avaliação de qualidade dos artigos selecionados. Como pode ser visto, apenas 5 artigos obtiveram a pontuação máxima, e a pontuação mínima foi 3,5, atingida por 3 estudos. Esses resultados mostram que os estudos que utilizam os dados do INEP, em sua grande maioria, têm boa qualidade por, principalmente, cumprir seus objetivos.

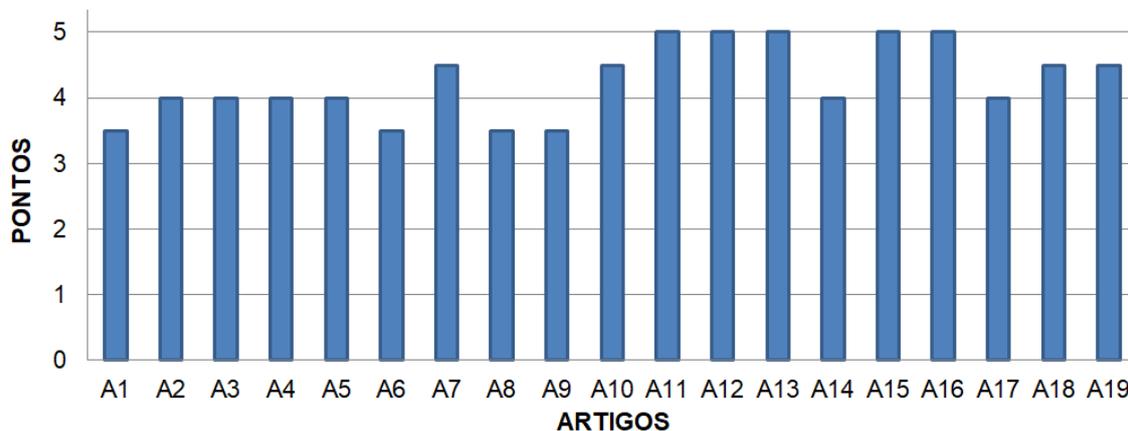


Figura 2. Pontuações dos estudos obtidas na avaliação de qualidade.

## 4. Discussão

### 4.1. Respondendo as Questões de Pesquisa

A questão de pesquisa **QP1** indagou quais os dados do INEP (IDEB, SAEB, Censo Escolar e Indicadores Educacionais) são analisados com técnicas de mineração. Constatou-se que, do total de artigos analisados, nove utilizam os dados do SAEB, sendo portanto o conjunto de dados mais utilizado. Cinco estudos utilizam dados dos Indicadores Educacionais, cinco utilizam os dados do IDEB e os dados menos utilizados são os do Censo Escolar, com apenas 3 artigos. Ressalta-se que, se somarmos a quantidade de vezes que os dados são utilizados, será maior que o total de 19 estudos selecionados, pois 3 desses trabalhos utilizam mais de um conjunto de dados do INEP (ver Tabela de resultados preliminares 3).

A questão **QP2** teve como objetivo identificar quais os problemas da educação básica brasileira estão sendo abordados pelas técnicas de mineração de dados. O principal problema estudado é o baixo desempenho escolar, o qual refere-se, por exemplo, ao baixo desempenho dos estudantes no SAEB e IDEB ou ENEM (Exame Nacional do Ensino Médio). O INEP fornece todos os anos uma estimativa sobre as notas do IDEB de acordo com os dados do ano anterior, desejando que as escolas alcancem aquelas notas para

melhorar o nível da educação básica brasileira. É tácito que algumas escolas não alcançam essa estimativa. Dessa forma, o baixo desempenho escolar foi um problema frequente tratado nos artigos revisados. Também foram encontrados diversos outros problemas como a fragilidade do sistema educacional, evasão e reprovação de alunos, problemas com a formação dos docentes, baixo investimento na educação e dois artigos trataram da falta de padronização das bases de dados educacionais.

Quanto à **QP3**, a metodologia categorizada como análise exploratória de dados foi a mais utilizada. Quanto aos estudos que descrevem minuciosamente a metodologia utilizada, seguindo rigorosamente suas etapas, a mais utilizada foi a CRISP-DM, com 6 estudos deixando bem claro em seus textos que a utilizaram. Portanto, no que diz respeito a seguir rigorosamente as etapas definidas por uma metodologia, ela é a predominante e preferida para a realização de mineração de dados. A outra metodologia mais usada, que possui etapas definidas, foi a KDD, com 3 artigos utilizando-a. As metodologias CRISP-DM e KDD são bem parecidas no que diz respeito às etapas, então acreditamos que a escolha de uso pode ser devido às preferências dos pesquisadores.

Quanto à **QP4**, dentre as técnicas utilizadas, houve uma predominância de modelos de regressão e análise de correlação. Através de análise de coeficientes de correlação se define o grau de um relacionamento entre variáveis. Esta técnica é muito utilizada, por exemplo, na análise de dados de desempenho e infraestrutura escolar. Por outro lado, o desenvolvimento de modelos de regressão está também muito presente em diversos artigos. Esses modelos permitem aos pesquisadores fazerem uma predição do resultado (uma variável numérica dependente) e comparar com o resultado real.

A Figura 3 mostra a coocorrência das técnicas de mineração de dados do INEP utilizadas nos estudos revisados nesta RSL. Ela evidencia a resposta para a QP1, com os dados do SAEB sendo utilizados em maior número ( $n = 9$ ) e a resposta da QP4, a qual identificou que as técnicas mais utilizadas são os modelos de regressão ( $n = 10$ ) e análises de correlação ( $n = 5$ ). Com a figura, é possível identificar ainda que os estudos têm focado mais frequentemente na elaboração de modelos de regressão com os dados dos Indicadores Escolares ( $n = 4$ ) e do SAEB ( $n = 3$ ).

SAEB	1	1	3	1	1	1	0	1	0	0	0	2
IDEB	0	0	2	1	2	0	0	1	0	1	0	0
Indicadores	0	0	4	0	1	0	0	0	0	0	1	0
Censo	0	0	1	0	1	0	1	1	1	0	0	0
	Análise estatística	Algoritmo EOp	Modelo de Regressão	Análise de Agrupamento	Análise de Correlação	Algoritmo POKS	Identificação de Outliers	Árvore de Decisão	Classificação Bayesiana	Ferramentas OLAP	Linked Data Mashup	Algoritmos de Classificação

Figura 3. Coocorrência entre os dados do INEP (linhas) e as técnicas (colunas).

## 4.2. Tendências

Existe uma tendência em grande parte dos artigos selecionados para tratar de problemas relacionados ao desempenho escolar de alunos. Minerando as bases de dados abertos do INEP, esses trabalhos visam identificar os fatores que influenciam no desempenho, criar modelos de predição, bem como sugerir possíveis soluções que podem ser feitas para melhorar as taxas de desempenho.

Outra tendência identificada foi a realização de análises exploratórias dos dados, sem um método bem definido nos artigos, e o uso da metodologia CRISP-DM. Quando utilizado aprendizado de máquina nos estudos, os modelos supervisionados de regressão foram os mais utilizados.

A Figura 4 faz uma análise entre o número de publicações por ano e os respectivos métodos de mineração de dados utilizados. Em relação ao quantitativo de publicações, fica evidente que no ano de 2019 houve um avanço significativo nas publicações. Em particular, muitos pesquisadores fizeram uso da metodologia CRISP-DM nos anos de 2018 e 2019. A figura ainda demonstra que, mesmo com o avanço nas tecnologias de mineração de dados, o ano de 2020 obteve um número baixo de publicações.

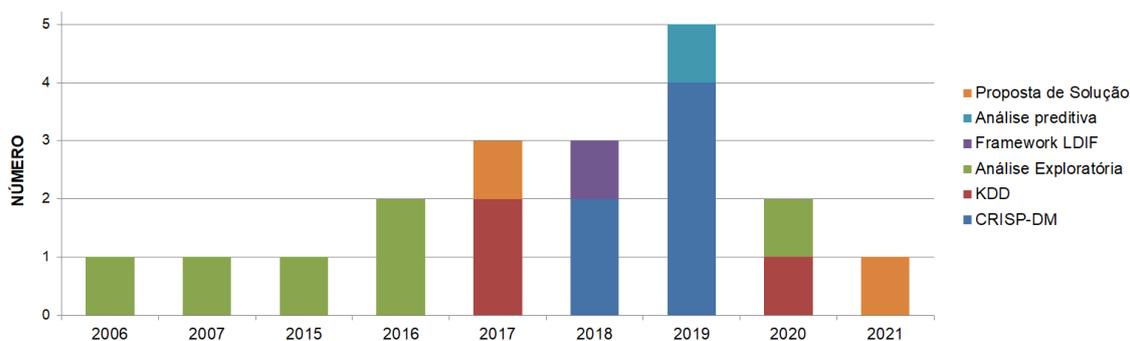


Figura 4. Quantidade de publicações por ano considerando as metodologias utilizadas.

## 4.3. Limitações

Algumas limitações relacionadas à condução desta RSL devem ser reconhecidas. Primeiramente, apenas dois idiomas de escrita dos artigos foram considerados. Dessa forma, artigos publicados em idioma diferente (e.g., espanhol) poderiam ser considerados relevantes mas não foram inclusos na revisão. Além disso, esta RSL tentou maximizar os resultados na estratégia de busca adotada, mas trabalhos não publicados (i.e., literatura cinza) que podem ter obtido resultados interessantes não foram considerados. Este trabalho se limitou à quatro bases de dados relacionadas à educação básica, porém o INEP disponibiliza outras bases, tais como o ENADE, ENEM, Censo dos Profissionais do Magistério e ENCCEJA, dentre outras. Portanto, percebe-se que ainda há oportunidade para a expansão desta RSL para outras bases do INEP e outros níveis da educação.

## 5. Conclusão

Este artigo teve como objetivo realizar uma RSL sobre mineração de dados da educação básica brasileira através da utilização dos dados do INEP. Um total de 19 artigos publicados foram analisados, com seus dados extraídos para responder às questões de pesquisa. Os dados do SAEB foram os mais utilizados nos estudos, mas todos os outros três conjuntos de dados (IDEB, Censo Escolar e Indicadores Educacionais) têm sido minerados pelas pesquisas inclusas nesta RSL. Diversos problemas educacionais foram identificados, tendo como predominância os problemas relacionados ao baixo

desempenho escolar de alunos. Sobre as metodologias de mineração de dados utilizadas, não houve um método que predominou, mas a CRISP-DM foi a mais recorrente. Por fim, em relação às técnicas de mineração, pode-se concluir que, apesar de uma maior exploração do desenvolvimento de modelos de regressão, existe uma grande variedade das técnicas usadas.

### Agradecimentos

Os autores gostariam de agradecer a Fundação de Amparo à Pesquisa e ao Desenvolvimento Científico e Tecnológico do Maranhão (FAPEMA) pelo apoio dado a seus projetos de pesquisa.

### Referências

- Avila, C. V. *et al.* Um linked data mashup de dados de execuções financeiras e indicadores educacionais no ensino básico. In: **Brazilian Symposium on Computers in Education**. [S.l.: s.n.], 2018. v. 29, p. 1911–1915.
- Bem, L. do N.; Pereira, V. da S.; Souza, E. Data mart para análise comparativa de dados do ideb em municípios da microrregião do pajeú em pernambuco. In: **Anais dos Workshops do Congresso Brasileiro de Informática na Educação**. [S.l.: s.n.], 2017. v. 6, p. 704–713.
- Carvalho, J.; Cruz, L.; Gouveia, R. Descoberta de conhecimento com aprendizado de máquina supervisionado em dados abertos dos censos da educação básica e superior. In: **Anais dos Workshops do Congresso Brasileiro de Informática na Educação**. [S.l.: s.n.], 2017. v. 6, p. 674–683.
- Coelho, O. B.; Silveira, I. Deep learning applied to learning analytics and educational data mining: A systematic literature review. In: **Brazilian Symposium on Computers in Education**. [S.l.: s.n.], 2017. v. 28, p. 143–152.
- Cohen, J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. **Psychological bulletin**, 1968.
- Júnior, G. C.; Nascimento, R.; Alves, G.; Gouveia, R. Identificando correlações e outliers entre bases de dados educacionais. In: **Anais dos Workshops do Congresso Brasileiro de Informática na Educação**. [S.l.: s.n.], 2017. v. 6, p. 694–703.
- Júnior, O. d. G. F.; Rodrigues, W. R. M.; Barbirato, J. C. C.; Costa, E. de B. Melhoria da gestão escolar através do uso de técnicas de mineração de dados educacionais: um estudo de caso em escolas municipais de maceió. **RENOTE**, v. 17, n. 1, p. 296–305, 2019.
- Junior, R. N.; Nascimento, R. L. S. do; Fagundes, R. A. de A.; Neto, P. S. G. de M. Estimação de índices de aprovação e reprovação escolar do ensino médio. In: **Brazilian Symposium on Computers in Education**. [S.l.: s.n.], 2019. v. 30, p. 339–348.
- Kitchenham, B.; Charters, S. **Guidelines for performing systematic literature reviews in software engineering**. [S.l.], 2007.
- LDIF. **Linked Data Integration Framework**. 2021. <<http://ldif.wbbsg.de/>>. Online; Acessado em: 6 de maio de 2021.
- Maschio, P.; Vieira, M. A.; Costa, N.; Melo, S. de; Júnior, C. P. Um panorama acerca da mineração de dados educacionais no brasil. In: **Brazilian Symposium on Computers in Education**. [S.l.: s.n.], 2018. v. 29, p. 1936–1940.
- Nascimento, R.; Júnior, G. C. Estudo sobre docentes do ensino básico através de indicadores educacionais e modelos de regressão. In: **Anais dos Workshops do Congresso Brasileiro de Informática na Educação**. [S.l.: s.n.], 2018. v. 7, p. 379–388.
- Nascimento, R. L. do; Fagundes, R. A.; Maciel, A. M. Prediction of school efficiency rates through ensemble regression application. In: **IEEE 19th International Conference on Advanced Learning Technologies (ICALT)**. [S.l.: s.n.], 2019. v. 2161, p. 194–198.

- Nascimento, R. L. S. do; Junior, G. G. da C.; Fagundes, R. A. de A. Mineração de dados educacionais: Um estudo sobre indicadores da educação em bases de dados do inep. **RENOTE**, v. 16, n. 1, p. 1–11, 2018.
- Pacini, I. B. de A. Educational indicators: a study of the limits and potentialities of the brazilian proof of the state teaching network of. **Humanidades & Inovação**, v. 7, n. 18, p. 242–257, 2020.
- Penteado, B. Geração automática de modelo de relações de pré-requisitos a partir de avaliação de larga escala brasileiras: um estudo preliminar. In: **Anais dos Workshops do Congresso Brasileiro de Informática na Educação**. [S.l.: s.n.], 2016. v. 5, p. 990–998.
- Penteado, B. E. Correlational analysis between school performance and municipal indicators in brazil supported by linked open data. In: **Proceedings of the 25th International Conference Companion on World Wide Web**. [S.l.: s.n.], 2016. p. 507–512.
- Pinto, G. da S.; Júnior, O. d. G. F.; Costa, E. de B. Identificação dos fatores de melhorias no ideb pelo uso de mineração de dados: Um estudo de caso em escolas municipais de teotônio vilela-alagoas. **RENOTE**, v. 17, n. 3, p. 183–193, 2019.
- Pinto, G. da S.; Júnior, O. F.; Costa, E.; Barbirato, J. C. C.; Rodrigues, W. R. M. Identificação dos fatores de melhorias no ideb pelo uso de mineração de dados: Um estudo de caso em escolas municipais de maceió. In: **Brazilian Symposium on Computers in Education**. [S.l.: s.n.], 2019. v. 30, p. 1828–1837.
- Ramos, T. G.; Machado, J. C. F.; Cordeiro, B. P. V. Primary education evaluation in brazil using big data and cluster analysis. **Procedia Computer Science**, Elsevier, v. 55, p. 1031–1039, 2015.
- Santos, A.; Medeiros, F. P. A. de. Relationship of federal funding to ideb results in a state in brazil: an approach based on educational data mining. In: **IEEE. 15th Iberian Conference on Information Systems and Technologies (CISTI)**. [S.l.], 2020. p. 1–4.
- Santos, P.; Ferreira, R.; Miranda, P. Dados abertos educacionais: Uma revisao da literatura brasileira. In: **Brazilian Symposium on Computers in Education**. [S.l.: s.n.], 2017. v. 28, p. 11–20.
- Silva, W. J.; Souza, R. M.; Cysneiros, F. psda: A tool for extracting knowledge from symbolic data with an application in brazilian educational data. **Soft Computing**, Springer, v. 25, n. 3, p. 1803–1819, 2021.
- Soares, J. F. Measuring cognitive achievement gaps and inequalities: The case of brazil. **International Journal of Educational Research**, Elsevier, v. 45, n. 3, p. 176–187, 2006.
- Viera, A. J.; Garrett, J. M. Understanding interobserver agreement: the kappa statistic. **Family Medicine**, v. 37 5, p. 360–3, 2005.
- Waltenberg, F. D.; Vandenberghe, V. What does it take to achieve equality of opportunity in education?: An empirical investigation based on brazilian data. **Economics of Education Review**, Elsevier, v. 26, n. 6, p. 709–723, 2007.