Revista Latinoamericana de Metodología de la Investigación Social.
Nº28. Año 14. Octubre 2024 – Marzo 2025. Argentina. ISSN 1853-6190. Pp. 70-86.

# Uma proposta de *framework* para pesquisas com *educational data mining* no contexto da América Latina

A framework proposal for research with educational data mining in the context of Latin America

**Bruno Francisco Batista Dias e Deborah Moraes Zouain** †

**Resumen**

A oferta gratuita e ecumênica da educação básica desempenha papel fundamental para o progresso e desenvolvimento dos povos latino-americanos. Por isso, os governos dessa região vêm se empenhado em garantir que isso ocorra de forma satisfatória. Nesse sentido, nós pesquisadores, por meio de achados científicos, podemos contribuir significativamente para esse processo de melhoria continue ocorrendo. Para isso, o uso de métodos de pesquisa como a Mineração de Dados Educacionais (*Education Data Mining — EDM*) —que combina conhecimentos de computação, educação e estatística— pode ser um forte alinhado. Diante das possibilidades da aplicação da EDM, este estudo propõe um *framework* para auxiliar o desenvolvimentos de pesquisas que utilizem a EDM como estratégia metodológica. O *framework* foi elaborado seguindo as práticas de Design Science Research (DSR), resultando em um modelo objetivo e de fácil compreensão, organizado 4 etapas. Esperamos que o framework auxilie a formulação de uma boa questão de pesquisa até escolha dos caminhos adequados que o levem a respondê-la satisfatoriamente.

**Palabras clave**: Data Mining; EDM; efeito escola; Design Science Research (DSR); descoberta de conhecimento.

**Abstract**

The free and ecumenical provision of basic education plays a fundamental role in the progress and development of Latin American peoples. Therefore, governments in this region have been committed to ensuring that this occurs satisfactorily. In this sense, we researchers, through scientific findings, can significantly contribute to this improvement process to continue occurring. To this end, the use of research methods such as Education Data Mining (EDM) —which combines knowledge from computing, education, and statistics— can be a strong alternative. Given the possibilities of applying EDM, this study proposes a framework to assist in the development of research that uses EDM as a methodological strategy. The framework was developed following Design Science Research (DSR) practices, resulting in an objective and easy-to-understand model, organized in 4 stages. We hope that the framework will help you formulate a good research question and choose the appropriate paths that will lead you to answer it satisfactorily.

**Keywords**: Data Mining; EDM; school effect; Design Science Research (DSR); knowledge discovery.

## 1. Introdução

Nos 20 países da América Latina, os sistemas educacionais públicos ofertam educação gratuita para crianças e adolescentes em idade escolar, de modo a prepará-los com conhecimentos e habilidades necessários para a vida adulta. Além disso, nas últimas duas décadas, os governos latino-americanos têm se empenhado para universalizar o acesso à educação básica (Junior & Nunes, 2022). Porém, ainda há um longo caminho a ser percorrido para se alcançar esses ideais. Pois, embora de difícil comparação, nas diversas realidades dos sistemas educacionais desses países, se observa que a evasão e a baixa qualidade do ensino ofertado ainda são problemas endêmicos, evidenciados pelas baixas taxas de conclusão no tempo adequado dos estudantes e altos níveis de analfabetismo funcional dos egressos (Anônimo, 2017).

Em 2020, 94% das crianças e adolescentes latino-americanas de 4 a 15 anos estavam matriculadas nos primeiros anos da educação básica, porém somente 85% concluíram essa etapa do ensino (OECD, 2020). Esse número é ainda menor quando consideramos os anos finais da educação básica, em que as taxas de matrícula e conclusão estão no patamar de 85% e 80%, respectivamente (OECD, 2020). Essa situação é gravíssima, pois a evasão reflete "o próprio fracasso das relações sociais que se expressam na realidade desumana que se vivencia no cotidiano, no qual a distância formada pela teoria e a prática desafia a inteligência do indivíduo" (Ferreira, 2013: 12). Quanto a qualidade do aprendizado, o problema também é notório: 1 a cada 3 estudantes latino-americanos que concluem a educação básica não possuem níveis adequados de conhecimento em leitura e matemática (Schleicher, 2019).

Nesse sentido, é necessário um amplo entendimento contextualizado das causas desses problemas, assim como das suas possíveis soluções. Para isso, diversos estudiosos do campo das ciências sociais têm se empenhado em fazer grandes variedades de pesquisas, por meio do emprego de métodos e técnicas de coletas de dados *in loco* nas escolas. Porém, dado o elevado quantitativo de escolas, com contextos bastantes distintos uma das outras, boa parte desses estudos não podem ser generalizados, de modo a fomentar a criação de políticas públicas para enfrentar sistematicamente os problemas educacionais. Podemos citar, por exemplo, as pesquisas que utilizam como método os estudos de caso, que apesar de facilitarem o entendimento de uma realidade local e prover soluções a ela, são, praticamente, não generalizáveis (Yin, 2015).

Como solução, acreditamos que os recentes desenvolvimentos tecnológicos, especificamente as técnicas de mineração de dados, podem vir a contribuírem para analisar informações em larga escala, de modo a extrair conhecimento útil em bancos de dados educacionais, permitindo um entendimento holístico do campo educacional. Esse processo, também chamado de descoberta de conhecimento em base de dados, permite a identificação de padrões por meio do uso de métodos estatísticos combinados com ferramentas de tecnologia da informação (Romero *et al.*, 2010; Shin & Shim, 2021). Como informa Dhankhar *et al.* (2021), o uso dessas técnicas tem desempenhado um papel central no avanço da melhoria de qualidade de diversos tipos de organizações.

Por isso, esta pesquisa se propõe a fornecer um *framework* básico para o uso dessa metodologia, provendo um *workflow* de como aplicar a mineração de dados na investigação da realidade escolar e do efeito escola, isto é, "o quanto uma organização escolar, pelas suas políticas e práticas internas, acrescenta ao aprendizado do aluno" (Brooke *et al.*, 2011: 10). O *framework* proposto é direcionado, principalmente, a pesquisadores que já possuem conhecimento prévio em estatísticas e noções de banco de dados, mas ainda não possuem um saber sistematizado de como estruturar uma pesquisa de mineração de dados em bases educacionais.

## 2. Educational data mining

Sistemas de coletas de dados educacionais e avaliações periódicas da qualidade do ensino, disponibilizadas em repositórios públicos, são boas práticas de governança adotadas por alguns governos latino-americanos (Reia & Cruz, 2021). Nesses repositórios costumam ser possível extrair informações como o desempenho das escolas, nível de aprendizagem dos estudantes, disponibilidade de recursos, características socioeconômicas dos estudantes e perfil dos docentes. Além disso, essas bases de dados são, em geral, confiáveis e possuem níveis

adequados de acurácia, isto é, refletem fidedignamente a realidade em que os dados foram coletados (Hopfenbeck *et al.*, 2018).

Atualmente a Colômbia e o Brasil são os países da América latina que possuem os melhores sistemas de coleta e divulgação dos dados sobre os seus sistemas de ensino. A finalidade, de um modo geral, desses sistemas consiste em subsidiar a tomada de decisão e a elaboração de políticas educacionais desses países (Moreno-Gómez *et al.*, 2020; MEC, 2018).

Na Colômbia, o sistema de avalições e coletas de informação visa acompanhar o cumprimento das metas estabelecidas no documento *Metas Educativas 2021: la educación que queremos para los Bicentenarios*. Segundo o Ministério da Educação da Colômbia essas avaliações:

> pressupõem diferentes perspectivas, estas podem ser: política, em que as relações de poder presente entre o avaliador (Ministério da Educação Nacional, ICFES, Secretários, Diretores, Professores) e os avaliados (Instituições de Ensino, Diretores, Professores, Alunos); pedagógico, pergunta sobre os pressupostos pedagógicos (concepções sobre Educação, Ensino, Avaliação, Aprendizagem, Didática, Currículo, Aluno, Professor, etc.), epistemológico (concepções sobre o Conhecimento, o modo de Produção do Conhecimento, Verdade, Validade, o Sujeito Cognoscente, os Objetos do Conhecimento, etc.), ontológico (as concepções sobre Mundo, Ser, Ser Fundador, etc.) e axiológica (as concepções sobre os Valores Ética e Estética) presentes nos discursos e normas sobre avaliação (ICFES, 2015: 1).

No Brasil, por sua vez, o Sistema de Avaliação da Educação Básica (SAEB) permite

> avaliar a qualidade, a equidade e a eficiência da educação praticada no país em seus diversos níveis governamentais; produzir indicadores educacionais para o Brasil, suas regiões e Unidades da Federação e, quando possível, para os municípios e as instituições escolares, tendo em vista a manutenção da comparabilidade dos dados, permitindo, assim, o incremento das séries históricas; subsidiar a elaboração, o monitoramento e o aprimoramento de políticas públicas baseadas em evidências, com vistas ao desenvolvimento social e econômico do Brasil; e desenvolver competência técnica e científica na área de avaliação educacional, ativando o intercâmbio entre instituições educacionais de ensino e pesquisa (INEP, 2019: 19).

Da mesma maneira, outros países da região —como a Argentina, através da Direção Nacional de Informação e Avaliação da Qualidade Educativa (DiNIECE)— possuem seus próprios sistemas de geração de dados educacionais. Há ainda a possibilidade de os dados desses sistemas serem combinados com os resultados de iniciativas internacionais de coletas de avaliação e dados, como o Programa Internacional de Avaliação de Estudantes (PISA), "uma avaliação internacional que mede o nível educacional de jovens de 15 anos por meio de provas de Leitura, Matemática e Ciências" (INEP, 2018: 1), da Organização para a Cooperação e Desenvolvimento Econômico (OCDE). O PISA contém informações de sete países latino-americanos: Argentina, Brasil, Chile, Colômbia, México, Peru e Uruguai.

A disponibilização desses repositórios tem permitido a realização de pesquisas científicas em profundidade e em larga escala na América Latina. Esse tipo de pesquisa combinando ferramentas de tecnologia da informação e estatística é denominado *Educational Data Mining* (EDM) ou, simplesmente, mineração de dados educacionais. Observa-se que:

> historicamente, tem sido difícil estudar o quanto as diferenças entre professores e turmas influenciam aspectos específicos da aprendizagem; esse tipo de análise se torna muito mais fácil com a mineração de dados educacionais. Da mesma forma, os impactos de diferenças individuais são difíceis de estudar estatisticamente com métodos tradicionais – A mineração de dados educacionais tem o potencial de estender um conjunto de ferramentas muito mais amplo para a análise de questões importantes, como essas, das diferenças individuais (Romero *et al*., 2010: 3).

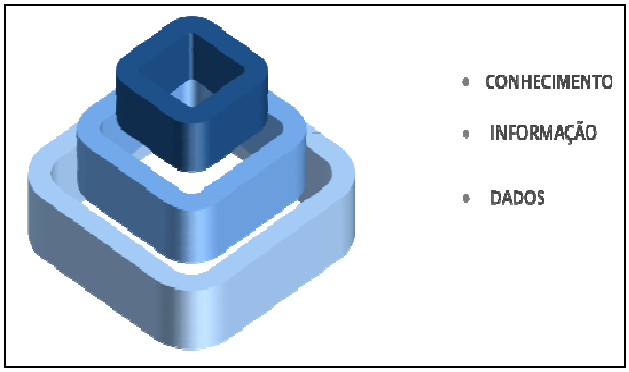
O objetivo do uso da EDM é obter conhecimento útil sobre aspectos educacionais. Em outras palavras, a EDM contempla *frameworks*, métodos e procedimentos modernos de investigação para identificar problemas e suas possíveis soluções na área da educação (Baker,

2015; Mohamad & Tasir, 2013; Dutt *et al.*, 2017). Entre as vantagens de seu uso, está sua capacidade de facilitar investigações sobre diversos aspectos da educação pública que seriam inimagináveis a poucas décadas atrás. Isso se dá, pois, a EDM utiliza dados, que já se encontram previamente coletados, gerando economia de tempo, redução de custos e escalabilidade das investigações (Romero *et al*., 2010; Baker & Inventado, 2014; Ahuja *et al.*, 2019).

As pesquisas que utilizam EDM partem de um problema (dúvida específica) e tentam solucioná-lo através de análise de dados decorrentes da iteração entre estudantes e ambiente de aprendizagem (escola, comunidade, rede de ensino, entre outros) (Baker, 2015). Os problemas costumam envolver fatores que afetam a qualidade do ensino ofertado. As análises são feitas por meio de testes estatísticos entre as variáveis educacionais. Já as soluções, apresentam os achados das análises, como respostas para orientar a aplicação melhores práticas organizacionais e pedagógicas, fornecer melhores arranjos organizacionais, permitir melhoria nos níveis de eficiência e eficácia dos gastos em educação, entre outras.

O processo de EDM é feito por meio da análise de dados de modo a identificar padrões, que quando contextualizados, se transformam em informações. Essas informações ao serem aplicadas na solução de problemas concretos são chamadas de conhecimento (Hand, 2001).

Figura 1. Estrutura básica do produto das etapas do KDD

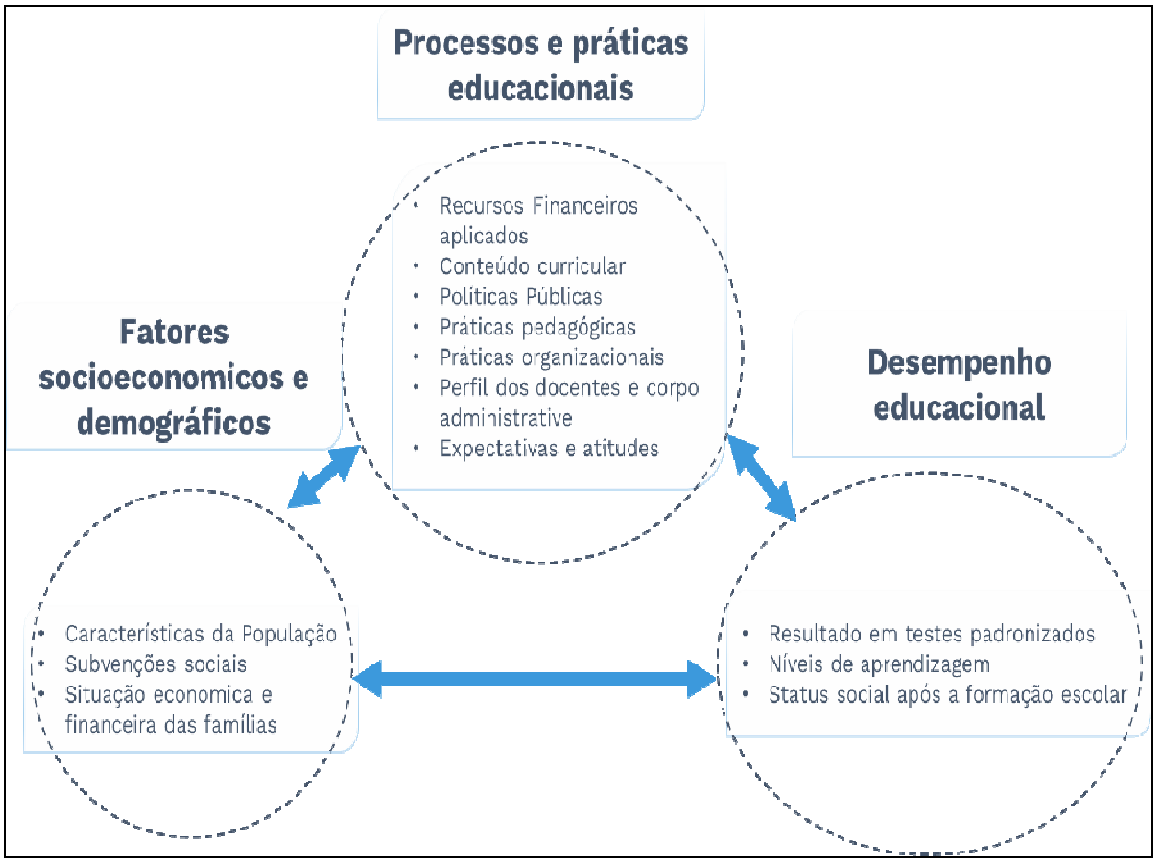



Fonte: desenvolvido pelo autor, a partir de Hand (2001).

Para isso, o campo da EDM considera que ao se trabalhar com dados educacionais deve-se considerar certas singularidades, tais como a alta comunalidade e a hierarquia dessas estruturadas de dados (Silva & Fonseca, 2017). Há também aspectos do campo educacional que podem ser importantes na EDM, como a necessidade de se distinguir as práticas organizacionais das práticas pedagógicas (Misoczky & Moraes, 2017), de compreender o contexto socioeconômico dos alunos (Menezes, 2018) e de entender as melhores configurações dos recursos dispostos na escola (Hanushek, 2005).

A Figura 2 ilustra alguns tipos de dados educacionais que são utilizados na EDM e suas relações.

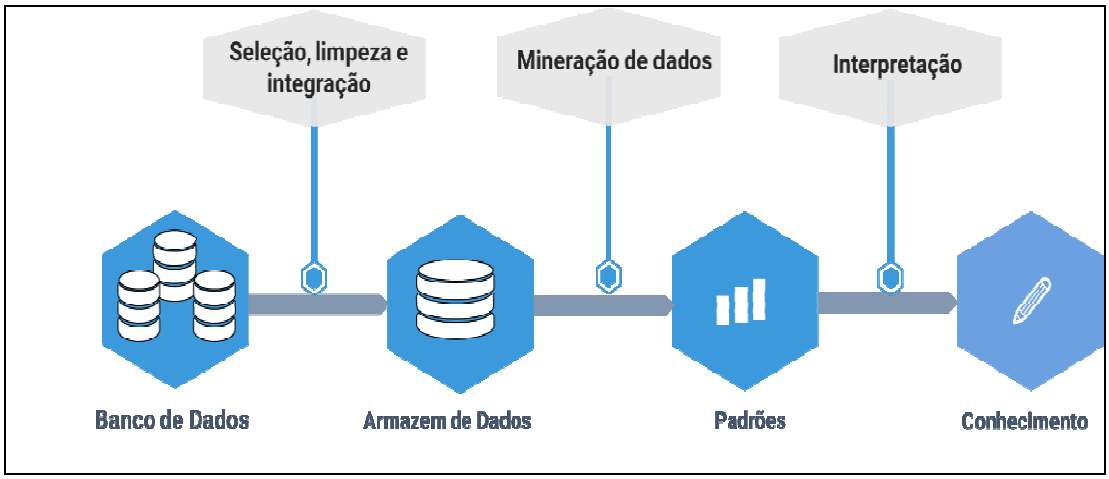
Figura 2. Proposta de indicadores OECD




Fonte: desenvolvido pelo autor, adaptado de OECD (2020).

Geralmente, a aplicação da EDM é realizada em 3 etapas ordenadas, conforme ilustrado na Figura 3.

Figura 3. Etapas do EDM




Fonte: desenvolvido pelo autor, a partir de Han & Kamber (2006).

Na primeira etapa, seleção, limpeza e integração são escolhidos os dados a serem analisados, que, passam pelo pré-processamento (limpeza) e são tratados (integração), de modo a permitir a sua mineração. A segunda etapa, mineração de dados, é a etapa mais complexa desse processo, nela os dados formatados são transformados em informação por meio de técnicas estatísticas avançadas que permitem a identificação de padrões de referências sobre um

determinado acontecimento ou grupo de sujeitos. Na última, interpretação, a informação obtida na etapa anterior é entendida por meio das lentes teóricas adotadas, gerando assim, o conhecimento científico válido (Han & Kamber, 2006).

Apesar dessas etapas serem, em regra, padronizadas, elas permitem diversos tipos de abordagens para se obter os mais variados tipos de conhecimento dentro do campo da educação. Sendo, na América Latina, predominante as abordagens que investigam questões sobre o desempenho estudantil, a evasão e os recursos escolares. Há, ainda, porém menos explorada, o uso da EDM para identificar padrões de comportamento entre alunos, que permite criar novas metodologias de aprendizado personalizadas de acordo com as necessidades individuais ou coletivas de um certo tipo de aluno (Jiménez *et al*., 2020).

Entretanto, como afirma Dwivedi & Singh (2016), a maior parte da produção acadêmica que utiliza a EDM na América Latina dedica-se apenas em realizar cruzamento de dados (*data crunching*) ao invés de obter conhecimento capaz de orientar soluções reais aos problemas educacionais. Isto é, ainda se faz necessário um maior engajamento das pesquisas latinas com EDM para auxiliar na melhoria dos processos educacionais.

## 3. Procedimentos Metodológicos

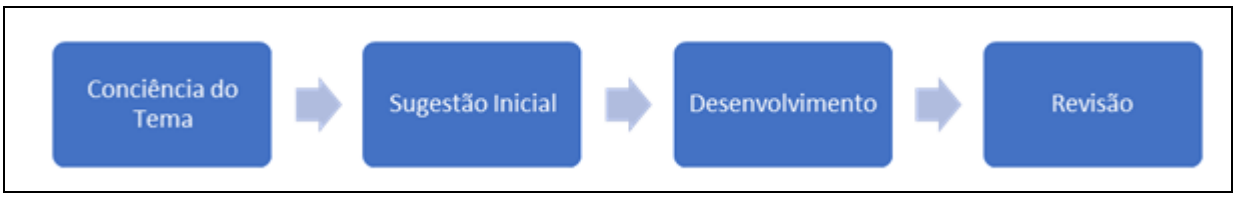
Os procedimentos metodológicos adotados nesse artigo tiveram como fim a construção de uma estrutura básica para o desenvolvimento de pesquisas científicas aplicando a mineração de dados educacionais. Objetivou-se também garantir que essa estrutura fosse de fácil entendimento e aplicação, de modo que um pesquisador ainda não familiarizado com o tema consiga obter orientações gerais sobre as possibilidades do uso da EDM. Para isso, a técnica metodológica adotada no estudo foi a Design Science Research (DSR).

Enquanto os métodos tradicionais de pesquisa fornecem soluções para o entendimento de uma determinada realidade, a DSR visa ampliar, justamente, essa capacidade de realizar investigações (Venable *et al*., 2016). A escolha pela aplicação dessa técnica deve-se a sua poderosa capacidade de orientar o desenvolvimento de *frameworks* no campo social para solução de problemas diversos (March & Smith, 1995). Em outras palavras, a DSR foi escolhida por permitir estruturar de formar fácil e amigável um caminho, cientificamente válido, para o desenvolvimento de soluções, como a que almejamos.

O produto final resultante da aplicação DSR são chamados de artefatos, que podem ser, entre outros, novos métodos de investigação, *frameworks* ou arquiteturas conceituais (Venable *et al*., 2016). Os artefatos, geralmente, são reutilizáveis e permitem encontrar soluções de problemas semelhantes, isto é, de um modo geral, são úteis para propósitos diversos. Para isso, a DSR deve ser aplicada de forma adequada e respeitar as singularidades do ambiente. O não atendimento dessas diretrizes pode implicar no desenvolvimento de artefatos inadequados ou que apresentem resultados que não reflitam adequadamente aos objetivos estabelecidos (Wieringa, 2014).

Para esse fim, adotamos nesse estudo as quatro etapas de construção de artefatos da DSR propostas por Ostrowski *et al.* (2014), conforme ilustrado na Figura 4.

Figura 4. Diagrama das etapas de construção de artefatos adotada no estudo




Fonte: desenvolvido pelo autor.

Na etapa inicial, consciência do tema, realizou-se um amplo entendimento do campo da mineração de dados e suas possibilidades de uso no contexto da educação da América Latina, por

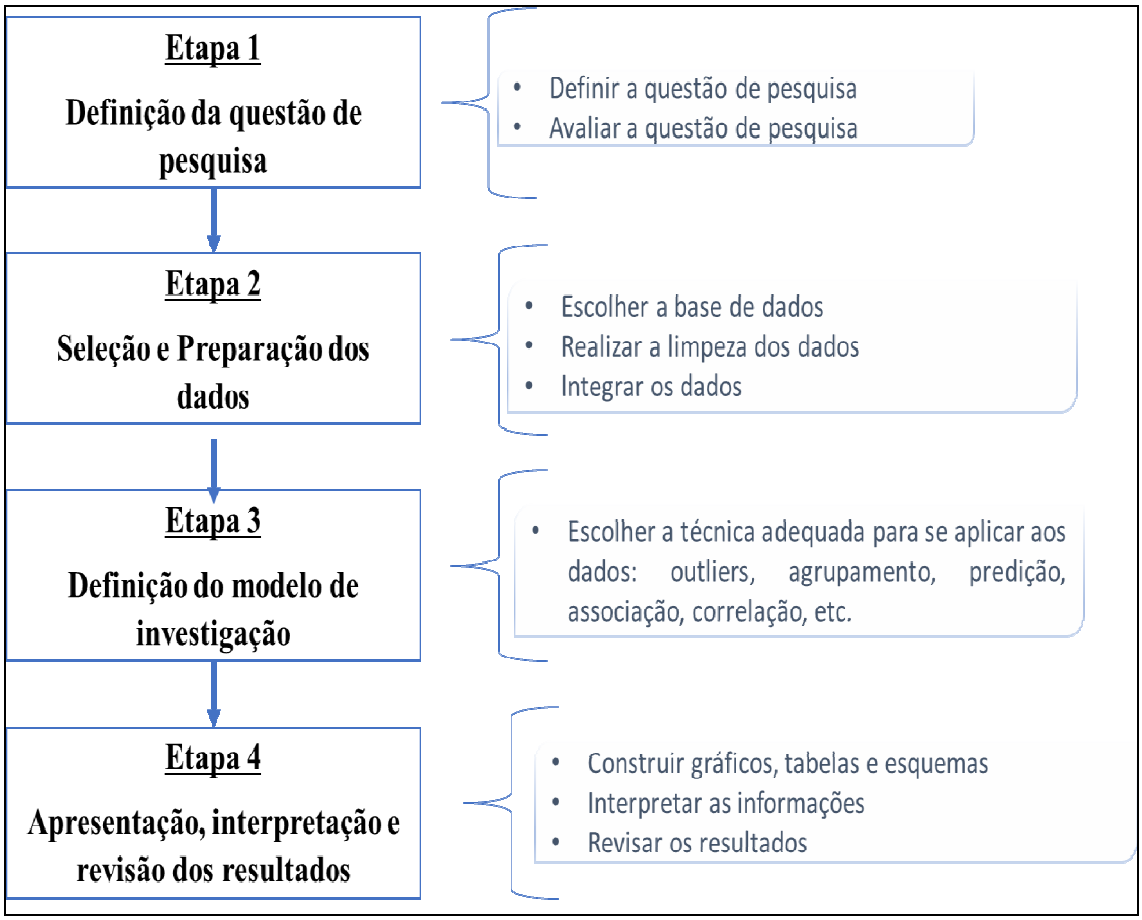
meio de uma sistemática revisão da literatura afim. Na segunda etapa, sugestão inicial, desenvolveu-se as ideias centrais do artefato, através de um processo criativo de ordenação dos conceitos levantados na etapa anterior. No desenvolvimento, estruturou-se o *framework* de forma a atender o objetivo desse estudo. Na revisão, fez-se uma minuciosa conferência para se apresentar um linguajar ubíquo e uma estrutura de fácil compreensão.

Ao final, foi possível apresentar um *framework* objetivo que seja útil aos diversos interesses de pesquisa no campo educacional na América Latina.

## 4. Proposta de *framework*

A Figura 5 contém a proposta de *framework* para pesquisas de mineração de dados educacionais que elaboramos nesse estudo como resultado do uso da DSR, contendo 04 etapas ordenadas de forma sequencial.

Figura 5. Proposta de Framework




Fonte: desenvolvido pelo autor.

### *4.1. Etapa 1 – Definição da questão de pesquisa*

Em regra, os estudos científicos se dividem em duas fases, a de planejamento e a de execução. A primeira fase, planejamento, tem como objetivo nortear a investigação, definindo o que se pretende responder ou solucionar (escopo da pesquisa). Já a segunda, execução, é o momento em que se realizam os procedimentos como medições, análises, sumarizações, além, de apresentar os resultados condizentes com a proposta inicial. Nesse sentido, a fase de planejamento é tida como indispensável de qualquer trabalho científico, devendo ser realizado

com o máximo de precisão, para se evitar desperdícios ou adversidades na fase de execução (Yin, 2015).

No planejamento, muitos aspectos devem serem considerados para que a pesquisa consiga ser concluída com sucesso. Fontes de financiamento, o tempo disponível para sua realização, a capacidade técnica da equipe, entre outros elementos, terá grande impacto ao longo da execução (Marconi & Lakatos, 2012). Esses aspectos são limitadores da liberalidade da qual o pesquisador tem ao definir o escopo da pesquisa. Em outros termos, tudo isso deve ser minuciosamente considerado com o máximo de rigor, havendo, assim, apenas certa discricionaridade por parte do pesquisador e de sua equipe.

Considerando esses aspectos, este *framework* ocupa-se primordialmente em orientar como definir uma boa questão de pesquisa. Ela é quem garante o sucesso das etapas posteriores, por isso, é, sem dúvida, a tarefa mais desafiadora a qual o pesquisador se depara no processo científico (García, 2016). Isso se dá, pois uma má escolha, nessa fase, irá gerar repercussões em cascata ao longo do estudo, impossibilitando sua conclusão, gerando resultados imprecisos ou grande desperdícios de recursos. Além disso, um equívoco na escolha da questão de pesquisa pode dificultar significativamente o reconhecimento do trabalho pela comunidade científica.

Para aplicação da EDM, o processo de encontrar uma boa questão de pesquisa pode ser facilitado por meio do conhecimento da literatura prévia, das bases de dados disponíveis e da experiência do pesquisador (Gil, 2014). A literatura científica prévia, como os artigos já publicados sobre temas correlatos, costuma possuir sugestões para estudos futuros, sendo um ótimo ponto de partida. Já o conhecimento prévio das bases de dados existentes permite saber se a questão de pesquisa será exequível por meio da EDM, ou seja, se há dados disponíveis para respondê-la. Também é possível que novas questões de pesquisa surjam como efeito correlato da experiência do pesquisador com um determinado campo de investigação, ou seja, por meio da experiência e intuição.

A Tabela 1 apresenta uma lista de indexadores de pesquisas científicas que podem ser utilizados para conhecer a literatura prévia sobre o tema a que se deseja investigar.

Tabela 1. Indexadores que possuem pesquisas de origem latino-americanas

| Indexador | Endereço URL |
|---|---|
| Latindex | https://latindex.org/latindex/ |
| REDIB | https://www.redib.org/ |
| Web of Science | https://www.webofscience.com/ |
| Scielo | https://www.scielo.br/ |
| ResearchGate | https://www.researchgate.net/ |
| IEEE Xplore Digital Library | https:// https://ieeexplore.ieee.org |
| Scopus | https://www.scopus.com/ |
| Redalyc | https://www.redalyc.org/ |
| NB | https://www.caicyt-conicet.gov.ar/sitio/ |

Fonte: desenvolvido pelo autor.

Durante esse processo, uma boa estratégia criativa consiste no uso de questões orientadoras, como: que tipo de problemas existem no campo da educação básica? A causa desses problemas já foi bem definidas? As soluções existentes atualmente são satisfatórias para resolver estes problemas?

Quando finalmente encontrada uma possível questão de pesquisa, se faz necessário uma avaliação crítica de sua qualidade para saber se ela é adequada ou não. Para isso, pode se considerar como critério de validade, ao menos os seguintes aspectos: ser exequível, de modo que seja viável obter uma resposta considerando os recursos e os dados disponíveis; ser interessante, a partir dos aspectos de relevância e tempestividade da resposta para a sociedade; ser nova, no sentido de que acrescente conhecimento ainda não sabido cientificamente; ser ética, considerando a sua adequabilidade aos códigos de conduta e ordenamentos jurídicos aplicáveis (Marconi & Lakatos, 2012; Gil, 2014; Yin, 2015; García, 2016).

Como exemplo de uma pergunta que foi respondia por meio do uso da DME com dados educacionais latino-americanos, temos a pesquisas de dissertação de Menezes (2018), em que o autor suscitou a indagação: "Por que algumas escolas em contextos desafiadores conseguem obter desempenho superior a outras escolas, mesmo, quando comparadas com escolas inseridas em contextos mais favoráveis?" (Menezes, 2018: 18). Após a aplicação da EDM, Menezes (2018) respondeu o questionamento com a seguinte resposta: "Os resultados mostraram que o envolvimento e comprometimento dos administradores da escola é fundamental para obter bons resultados na aprendizagem e na melhoria contínua do desempenho escolar" (Menezes, 2018: 90).

Uma vez definida e validada a questão de pesquisa, pode-se passar para a próxima etapa.

### *4.2. Etapa 2 - Seleção e Preparação dos dados*

Na segunda etapa o pesquisador irá realizar o processo de seleção dos dados que serão utilizados no estudo. A seleção pode ser entendida como um procedimento que "busca identificar o conjunto de dados relevantes e seus subconjuntos de variáveis objetivando a criação de um conjunto de restrito de dados para a descoberta de conhecimento" (Coradine *et al*., 2011: 169). Em outros termos, é nesse momento em que se deve fazer a identificação e a prospecção dos dados necessários para fornecer as respostas para responder à pergunta elaborada na etapa anterior.

Para obter os dados necessários de forma satisfatória, pode ser que seja necessário selecionar mais de uma base de dados simultaneamente. Essa tarefa costuma ser complexa, pois existem diversos tipos de bases disponíveis que podem ser adequadas ou não para o estudo. Por isso, sugerimos dar preferência ao uso de base estruturadas, pois essas bases possuem um modelo conceitual (esquema) pré-documentado que fornece uma visão abstrata de alto nível, independente do banco de dados, indicando como os dados se relacionam entre si e refletem a realidade do ambiente em que foram coletados.

Isso irá facilitar o processo de entendimento dos dados, sua compatibilização com outras fontes e sua manipulação por meio de *softwares* estatísticos e de mineração de dados. A Tabela 2 apresenta alguns dos principais softwares de mineração de dados que podem ser utilizados na EDM.

**Tabela 2. Softwares para aplicação da EDM**

| Software | Uso | Características |
|---|---|---|
| Google Sheets | Gratuito | Uteis para análise de quantidades pequeñas de dados, tento como facilidade a apresentação das informações em uma interface visual. |
| EDM Workbench. | Gratuito | Adequado para criação de *labels* e definição de perfis de variáveis de forma automatizadas. |
| Python | Gratuito | Eficiente para procesar operações estatísticas em grande quantidade de dados (até 10 milhões de linhas) por meio de una linguagem de programação de fácil operacionalização. |
| Structured Query Language (SQL) | Gratuito | É uma linguagem de estruturação de banco de dados, adequada para criar, alterar, excluir e atualizar colunas e linhas de um banco de dados estruturado. |
| RapidMinder | Gratuito | É um pacote para realizar análises de mineração de dados e criar modelos, por meio de algoritmos e linguagem de programação. Permite análisis de alto nível pelo cruzamento multinivel de variáveis. |
| TraMineR | Gratuito | Agrega um pacote da linguagem de programação R que dá suporte à mineração de dados, ofrecendo um conjunto de funcionalidades estatísticas. |
| IMB SPSS Modeler Premium | Pago | Comporta um conjunto de ferramentas estatísticas de sua versão tradicional (SPSS) além de ferramentas avançãdas de análise e mieração de dados. |
| Tableau | Pago | Apresenta um conjunto de ferramentas para criar painéis interativos de visualizações dinâmicas em tempo real de dados. |

Fonte: desenvolvido pelo autor.

Em regra, as melhores fontes para se obter dados educacionais estruturados e confiáveis são os dados abertos governamentais. Sua classificação como abertos ocorre "quando qualquer pessoa pode livremente acessá-los, utilizá-los, modificá-los e compartilhá-los para qualquer finalidade, estando sujeito a, no máximo, a exigências que visem preservar sua proveniência e sua abertura" (Governo Federal do Brasil, 2021: 1). A escolha ou não do que usar depende de uma avaliação crítica das necessidades do pesquisador e do escopo da pesquisa.

Após a seleção das fontes de dados, o pesquisador deverá tratá-las para estarem aptas ao processo de mineração. Essa tarefa "envolve a limpeza dos dados, com operações de remoção dos ruídos, elaboração de esquemas e mapeamentos de valores desconhecidos" (Coradine *et al*., 2011: 169). Ainda, pode ser que durante esse processo seja necessário realizar a padronização entre as fontes, isso se faz necessário sempre que existirem duas ou mais fontes de dados distintas.

Geralmente os dados podem estar armazenados em arquivos de diversos tipos de formatos como: *.csv, .txt, docsql, .sql, .xml*. Por isso, o processo de preparação dos dados se inicia com a conversão dos dados para um mesmo formato em um único repositório, sempre que houver mais de uma fonte de dados com formatos distintos. Para integrá-los, a maior parte dos softwares de tratamento de dados possuem ferramentas de conversão chamadas de ETL (*extract, load transform*), que permitem a extração, carregamento e carga desses dados de diversos formatos e fontes para um único repositório padronizado.

Na sequência, é essencial que seja feito o processo de limpeza nesse repositório criado pelo ETL. Isso implica, num primeiro momento, apagar dados encontrados em duplicidade, isto é, deduplica-los. Na sequência, é conveniente a realização da adequação dos nomes das variáveis, por meio da remoção de espaços e acentos para facilitar a manipulação e pelo uso de nomes que permitam uma fácil compreensão do significado da variável. Ao final desse processo, pode ser que seja observado a existência de dados incompletos (*missing data*) que comprometa a qualidade das

análises a serem realizadas. Caso isso se verifique, pode ser que seja necessário repor esses dados ou complementá-los.

O produto dessa etapa será um armazém de dados, integrado a partir de diversas fontes, capaz de contemplar múltiplos dados sobre o sistema de ensino investigado, a ser utilizado na próxima etapa.

### *4.3. Etapa 3 - Definição do modelo de investigação*

Nessa etapa o pesquisador deve se encarregar de definir o modelo com a técnica computacional e estatística a ser aplicada aos dados, permitindo ao pesquisado obter as respostas necessárias para responder os questionamentos suscitados (Han & Kamber, 2006). É através da aplicação de uma ou mais dessas técnicas sobre os dados contidos no armazém de dados que se obterá informações hábeis, que ao serem interpretadas e validadas nas etapas posteriores, serão chamadas de conhecimento científico (Fayyad *et al*.,1996). A depender da técnica, pode incluir o retorno a etapa anterior, isto é, da preparação dos dados, sempre que os dados contidos no armazém de dados sejam insuficientes ou inadequados para prosseguir com as etapas posteriores.

De um modo geral, a maior parte dos *softwares* estatísticos e de mineração de dados possuem funcionalidades com as mais diversas técnicas de mineração, porém, a escolha da técnica certa a ser utilizada irá depender diretamente da questão de pesquisa e da perspectiva adotada pelo pesquisador para respondê-la. Em qualquer situação, por estarmos trabalhando com dados educacionais, deve-se sempre adequar a técnica de modo que ela forneça resultados contextualizados. Isso se faz mandatório na América Latina, pois como afirma Menezes (2018) "esta análise contextualizada é capital, pois os processos de ensino e de aprendizagem, em sociedades que apresentam desigualdades sociais relevantes, são condicionados, em parte, pelas posições dos públicos atendidos na hierarquia social, explicitadas por seu padrão de vida" (Menezes, 2018: 30).

A Tabela 3 contém as principais técnicas de mineração de dados que podem ser usadas na EDM, seguidas de sua definição, e ao menos um exemplo de real de sua aplicação com dados de escolas ou alunos latino-americanos.

Tabela 3. Técnicas de EDM

| Técnica | Objetivo | Possíveis aplicações |
|---|---|---|
| Análise de Outliers | Identificar valores discrepantes nos dados. | Análise do perfil de escolas de alto desempenho em contextos desafiadores. |
| Análise de agrupamento (*cluster*) | Criar subgrupos de objetos similares de acordo com algum critério pré-determinado. | Encontre semelhanças e diferenças entre escolas.<br>Criar perfis de estudantes de acordo com suas características familiares e socioeconômicas |
| Modelos de Predição | Prever o comportamento futuro de uma variável com base em outras variáveis ou no seu histórico. | Identificar alunos em risco.<br>Prever o alcance de metas com base nos resultados educacionais dos alunos. |
| Associação | Identificar padrões de comportamento das variáveis em que sempre que a situação X ocorrer existe grande probabilidade de Y também vir a se concretizar, em que X é chamado de antecedente e Y é chamado de consequente. | Combinação de conteúdos curriculares e desenvolvimento de competências em estudantes. |
| Análise de correlação entre variáveis | Identificar a relação entre duas ou mais variáveis. | Relação entre o nível de escolaridade dos pais e alunos que abandonam a escola.<br>Análise de quais práticas pedagógicas estão associadas a melhores níveis de aprendizagem. |
| Desenvolvimento de mapas conceituais | Definir as etapas de um processo não conhecido, estruturando-o de forma a entendê-lo. | Definição das melhores estratégias de planejamento e programação das atividades curriculares. |
| Extração de dados para inferências | Descrever os dados de uma maneira que permitem identificar características relevantes na população. | Identificação de padrões de aprendizagem do aluno.<br>Fornecimento de relatórios. |

Fonte: Bramer, 2007; Coradine *et al.*, 2011; Mohamad & Tasir, 2013; Khan & Ghosh, 2021; Sweta, 2021.

## *4.4. Etapa 4 - Apresentação, interpretação e revisão dos resultados*

Após os dados terem sidos analisados e aplicando o modelo definido na etapa anterior, o pesquisador irá apresentar, interpretar e revisar os resultados obtidos. Nesse momento, é que será possível entender, descrever e sistematizar o conhecimento resultante da aplicação da EDM para responder as inquietações definidas na etapa inicial.

Existem diversas formas de se apresentar os resultados obtidos no processo de EDM, os mais comuns geralmente envolvem apresentações em forma de esquemas, tabelas e gráficos (Sweta, 2021). A escolha da melhor forma de apresentar os resultados depende diretamente da situação e do que se pretende elicitar. Nesse sentido, é adequado experimentar mais de uma forma de apresentação, até se escolher aquela que seja capaz de fornecer a melhor ideia do que se pretende explicar. Para isso, sugerimos alguns princípios orientadores: buscar o equilíbrio entre sumarizar e apresentar o maior número de detalhes que expliquem os resultados; preferir

apresentar gráficos à tabelas; preferir apresentações que permita ao leitor identificar por si mesmo as relações entre as variáveis; apresentar as frequências absolutas e relativas das variáveis quantitativas, assim como os máximos, mínimos e amplitude interquartil; apresentar a média e o desvio padrão de variáveis quantitativas normais e a mediana e a amplitude interquartil par as não normais.

A simples apresentação dessas informações obtidas pelas análises dos dados não configuram conhecimento. Por isso, elas precisarão ser integradas ao conhecimento já existente no campo e processadas para se obter inferências e conclusões. Para isso, é aconselhado que se faça uma adequada revisão bibliográfica, compostas principalmente por estudos científicos recentes (últimos 05 anos). O uso desses trabalhos irá facilitar a interpretação adequada dos resultados. Isso se dá, pois os estudos prévios possuem um conjunto de saberes pré-existentes do campo de estudo, permitindo translucidar as "lentes" pelas quais este pesquisador interpreta os resultados, colocando-o, previamente, em contato com as publicações científicas que versam sobre o tema (Marconi & Lakatos, 2012).

Por fim, bastante importante, deve-se realizar a revisão dos achados em termos de coerência e coesão. Considerando que os resultados da EDM são obtidos por meio de técnicas baseadas em métodos estatísticos, é comum que ocorram erros de interpretação por meio da construção de uma narrativa que extrapola o que se foi extraído pelas análises. Esses erros tendem a ser reduzidos com boas representações gráficas, respeito aso princípios do método científico e realização de análises baseadas em bons referencias teóricos.

Uma forma prática de se revisar é questionar a qualidade do trabalho nos seguintes aspectos: As informações foram validades e interpretadas adequadamente? Os resultados foram apresentados de forma clara e objetiva? Os resultados respondem à questão de pesquisa de forma adequada? A inferência estatística faz sentido?

Muitos erros podem ser corrigidos ao se realizar esses questionamentos, obtendo no final dessa etapa uma resposta cientificamente válida ao questionamento que orientou o estudo. Cabendo agora o pesquisador, a redação final de sua pesquisa conforme as normas de padronização adotadas pela comunidade científica.

## 5. Considerações Finais

Consideramos nesse estudo que o uso de *frameworks* que aplicam a EDM pode ser um forte aliado para realização de pesquisas capazes de extrair conhecimento científico em bases de dados educacionais.

Por isso, o objetivo deste artigo foi propor a sistematização de um *framework* metodológico para adoção da EDM na América Latina. Trata-se de uma proposta relevante, sobretudo por delinear um percurso metodológico factível para aqueles que possuem um certo conhecimento do tema, mas ainda estão dando os primeiros passos para trabalhar com essa abordagem. A proposta é predominantemente útil, considerando as diversas possibilidades de aplicação do *framework* para fazer pesquisas contextualizadas sobre questões que afetam o desempenho acadêmico e a permanecia dos estudantes nas escolas desses países.

De fato, para ter pleno domínio de como aplicar a EDM o pesquisador deve ter um profundo conhecimento de tecnologia da informação, estatística e do campo educacional. Essas competências não são desenvolvidas instantaneamente, pois requerem longo período de prática e muito esforço intelectual. Por isso, nesse trabalho, apresentamos um *framework* em forma de *workflow* em que o pesquisador que pertente utilizar essa ferramenta, saiba por onde começar, se atente a questões relevantes ao realizar sua pesquisa e perceba quais competências precisa aperfeiçoar-se para ser um cientista de dados educacionais.

As diversas realidades existentes na américa latina, assim como as variabilidades nos dados disponíveis são demasiadas complexas para serem investigadas de uma maneira rígida. Por isso, destacamos que a proposta apresentada não é uma regra a ser seguida rigidamente, mas sim um guia de boas práticas. Nesse sentido, além de delinear um caminho sistematizado, orientamos que o pesquisador deve-se nortear pelo respeito das regras estatísticas, o método

científico e, evidentemente, a ética. Esperamos, assim, que essa orientação faça sentido ao aprofundamento do pesquisador com o campo, desde a formulação de uma boa questão de pesquisa até escolha dos caminhos adequados que o levem a respondê-la satisfatoriamente, contribuindo para uma maior aceitabilidade de suas publicações científicas.

Por fim, por meio desse trabalho convidamos os pesquisadores, principalmente os que investigam o campo da educação na América Latina, para aplicar, complementar e aperfeiçoar o *framework* proposto.

## 6. Referências

AHUJA, R., JHA, A., MAURYA, R., SRIVASTAVA, R. (2019). Analysis of Educational Data Mining. In: Yadav, N., Yadav, A., Bansal, J., Deep, K., Kim, J. (Eds.), *Harmony Search and Nature Inspired Optimization Algorithms. Advances in Intelligent Systems and Computing*, vol 741. Springer: Singapore. https://doi.org/10.1007/978-981-13-0761-4_85

BAKER, R. S., & INVENTADO, P. S. (2014). Educational Data Mining and Learning Analytics. In: Larusson, J., White, B. (Eds.) *Learning Analytics*. Springer: New York. https://doi.org/10.1007/978-1-4614-3305-7_4

BAKER, R. S. (2015). *Big data and education.* New York: Teachers College, Columbia University.

BRAMER, M. (2007). *Principles of data mining.* Sant Lois: Springer.

GOVERNO FEDERAL DO BRASIL (2021). *O que são dados abertos.* Recuperado de: https://dados.gov.br/pagina/faq.

BROOKE, N., CUNHA, M. A. D. A., & FALEIROS, M. (2011). A avaliação externa como instrumento da gestão educacional nos estados. *Estudos & Pesquisas Educacionais, São Paulo*, 5(2), 17-79.

CORADINE, L. C., LOPES, R. V. V., & MACIEL, A. F. (2011). Mineração de Dados: Uma Introdução. *Journal of the Brazilian Neural Network Society*, 9 (3), 168-184.

DHANKHAR, A., SOLANKI, K., & DALAL, S. (2021). Predicting students performance using educational data mining and learning analytics: A systematic literature review. *Innovative Data Communication Technologies and Application*, 59(1), 127-140. https://doi.org/10.1007/978-981-15-9651-3_11

DUTT, A., ISMAIL, M. A., & HERAWAN, T. (2017). A systematic review on educational data mining. *IEEE Access*, *5*(1), 15991-16005.

DWIVEDI, T., & SINGH, D. (2016). Analyzing educational data through EDM process: A survey. *International Journal of Computer Applications*, *136*(5), 13-15.

FAYYAD, U., PIATETSKY-SHAPIRO, G., & SMYTH, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, *39*(11), 27-34.

FERREIRA, F. A. (2013). Fracasso e evasão escolar. Brasília: Brasil Escola.

GARCÍA, C. S. R. (2016). Investigación científica. *Revista Científica Alas Peruanas*, 1(2), 1-15.

GIL, A. C. (2014). *Métodos e técnicas de pesquisa social*. São Paulo: Editora Atlas SA.

INSTITUTO COLOMBIANO PARA LA EVALUACIÓN DE LA EDUCACIÓN (ICFES). (2015). *Normograma*. Recuperado de: http://www2.icfes.gov.co/transparencia/normatividad/normogram

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (INPE). (2018). Programa Internacional de Avaliação de Estudantes. Recuperado de: http://portal.inep.gov.br/web/guest/educacao-basica/pisa.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (INEP). (2019). Sistema de Avaliação da Educação Básica. Recuperado de: http://portal.inep.gov.br/web/guest/educacao-basica/saeb.

JIMÉNEZ, C. V. E., GARCÍA, T. M., & VÁZQUEZ, N. J. L. (2020). Análisis de la Minería de Datos en el ámbito de la Educación. *Revista Paraguaya de Educación*, *8*(2), 75-85.

JUNIOR, A. M., & NUNES, A. F. (2022). 21st Century Education in Brazil. In *Globalization and Education: Teaching, Learning and Leading in the World Schoolhouse* (201-218), Charlote: Age Publishing.

HAN, J., & KAMBER, M. (2016). *Data mining concepts and techniques*. EUA: Elsevier Press.

HAND, J. (2001). *Principles of Data Mining.* Massachusetts: MIT Press.

HANUSHEK, E. A. (2005). *Economic outcomes and school quality*. International Academy of Education. International Institute for Educational Planning. Belgium/France: Stedi Média.

HOPFENBECK, T. N., LENKEIT, J., EL MASRI, Y., CANTRELL, K., RYAN, J., & BAIRD, J. A. (2018). Lessons learned from PISA: A systematic review of peer-reviewed articles on the programme for international student assessment. *Scandinavian Journal of Educational Research*, *62*(3), 333-353. Recuperado de: https://doi.org/10.1080/00313831.2016.1258726

KHAN, A., & GHOSH, S. K. (2021). Student performance analysis and prediction in classroom learning: A review of educational data mining studies. *Education and information technologies*, *26*(1), 205-240. Recuperado de: https://doi.org/10.1007/s10639-020-10230-3

MARCH, S. T., & SMITH, G. F. (1995). Design and natural science research on information technology. *Decision support systems*, *15*(4), 251-266. Recuperado de: https://doi.org/10.1016/0167-9236(94)00041-2

MARCONI, M. D. A., & LAKATOS, E. M. (2012). *Técnicas de pesquisa: planejamento e execução de pesquisa; amostragens e técnicas de pesquisa; elaboração, análise e interpretação de dados*. São Paulo: Atlas.

MENEZES, D. T. (2018). *Análise do efeito escola sob as lentes de Pierre Bourdieu.* Dissertação. Volta Redonda: Universidade Federal Fluminense.

MINISTÉRIO DA EDUCAÇÃO (MEC). (2018). *Educação Básica.* Recuperado de: http://portal.mec.gov.br/secretaria-de-educacao-basica

MISOCZKY, M. C., & MORAES, J. (2017). *Organisation and liberating praxis social movement schools.* Rio de Janeiro: EDUFF.

MOHAMAD, S. K., & TASIR, Z. (2013). Educational Data Mining: A Review. *Procedia-Social and Behavioral Sciences*, 97 (3), 320-324. Recuperado de: https://doi.org/10.1016/j.sbspro.2013.10.240

MORENO-GÓMEZ, J., CALLEJA-BLANCO, J., & MORENO-GÓMEZ, G. (2020). Measuring the efficiency of the Colombian higher education system: a two-stage approach. *International Journal of Educational Management*, 34(4), 794-804.

ORGANIZATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT. (2020). *The Impact of COVID-19 on Education: Insights from "Education at a Glance 2020"*. Paris: OECD Publishing.

OSTROWSKI, L., HELFERT, M., & GAMA, N. (2014). Ontology engineering step in design science research methodology: a technique to gather and reuse knowledge. *Behaviour & Information Technology*, 33(5), 443-451. Recuperado de: https://doi.org/10.1080/0144929X.2013.815276

REIA, J., & CRUZ, L. F. (2021). Seeing through the smart city narrative: Data governance, power relations, and regulatory challenges in Brazil. In *Power and Authority in Internet Governance,* (219-242). Londres: Routledge.

ROMERO, C., VENTURA, S., PECHENIZKIY, M., & BAKER, R. S. (2010). *Handbook of educational data mining*. Boca Raton: CRC Press.

SCHLEICHER, A. (2019). *PISA 2018: Insights and Interpretations*. Paris: OECD Publishing.

SILVA, C., & FONSECA, J. (2017). Educational Data Mining: a literature review. In Rocha, Á., Serrhini, M., Felgueiras, C. (Eds.), *Europe and MENA Cooperation Advances in Information and Communication Technologies. Advances in Intelligent Systems and Computing*. https://doi.org/10.1007/978-3-319-46568-5_9

SHIN, D., & SHIM, J. (2021). A systematic review on data mining for mathematics and science education. *International Journal of Science and Mathematics Education*, *19*(4), 639-659. Recuperado de: https://doi.org/10.1007/s10763-020-10085-7

SWETA, S. (2021). Educational Data Mining Techniques with Modern Approach. In *Modern Approach to Educational data mining and its Applications,* (25-38). Singapore: Springer.

VENABLE, J., PRIES-HEJE, J., & BASKERVILLE, R. (2016). FEDS: a Framework for Evaluation in Design Science research. *European journal of information systems*, *25*(1), 77-89. Recuperado de: https://doi.org/10.1057/ejis.2014.36

WIERINGA, R. J. (2014). *Design science methodology for information systems and software engineering.* Luxemburgo: Springer.

YIN, R. K. (2015). *Estudo de Caso: Planejamento e métodos*. Porto Alegre: Bookman Editora.

***Autor y autora.***

*Bruno Francisco Batista Dias*

Universidade do Grande Rio (UNIGRANRIO), Duque de Caxias-RJ, Brasil.

Doutorando pelo PPGA/Unigranrio (Duque de Caxias-RJ), Mestre pelo Programa de Pós-Graduação em Administração da Universidade Federal Fluminense e Especialista em Gestão Pública com ênfase em didática do ensino superior.

E-mail: brunofbd@id.uff.br

*Deborah Moraes Zouain* †

Universidade do Grande Rio (UNIGRANRIO), Duque de Caxias-RJ, Brasil.

Doutora em Engenharia de Produção (COPPE/UFRJ). Professora do PPGA da Universidade do Grande Rio (UNIGRANRIO).

**Citado**.

BATISTA DIAS, Bruno Francisco e MORAES ZOUAIN, Deborah (2024). Uma proposta de *framework* para pesquisas com *educational data mining* no contexto da América Latina. *Revista Latinoamericana de Metodología de la Investigación Social* – ReLMIS, Nº28, Año 14, pp. 70-86.

**Plazos.**

Recibido: 27/07/2022. Aceptado: 01/06/2023.