http://dx.doi.org/10.18222/eae.v30i74.5696

MINERAÇÃO DE DADOS ORIENTADA PELO DOMÍNIO EDUCACIONAL: UMA PROVA DE CONCEITO

STELLA OGGIONI DA FONSECA[I]
ANDERSON AMENDOEIRA NAMEN[II]
FRANCISCO DUARTE MOURA NETO[III]
ADRIANA DA ROCHA SILVA[IV]
MARIA ISABEL RAMALHO ORTIGÃO[V]
URSULA ANDREA BARBARA VERDUGO ROHRER[VI]

RESUMO

Este trabalho propõe uma metodologia para identificação de padrões relacionados ao aprendizado de matemática e às características do ambiente escolar, a qual foi aplicada aos dados da Prova Brasil 2013, com ênfase nos estudantes do 9º ano do ensino fundamental do estado do Rio de Janeiro. A abordagem, apoiada pelo conhecimento de especialistas em educação, consistiu na proposição de um processo de redução de dimensionalidade integrado à mineração de dados, sendo sua avaliação efetuada por intermédio de medidas técnicas em conjunto com medidas de interesse do domínio educacional. Foi possível identificar ações, bem como analisar sua viabilidade para solução das questões educacionais. A metodologia, orientada pela área de aplicação e cuja avaliação não se restringiu ao uso de métricas técnicas da mineração de dados, pode servir como referência – uma prova de conceito – a outras pesquisas em ações e políticas educacionais.

PALAVRAS-CHAVE PROVA BRASIL • MINERAÇÃO DE DADOS • POLÍTICAS EDUCACIONAIS • RENDIMENTO ESCOLAR.

I Universidade do Estado do Rio de Janeiro (UERJ), Rio de Janeiro-RJ, Brasil; https://orcid.org/0000-0002-7697-2151; *stella.oggioni@gmail.com*

II Universidade do Estado do Rio de Janeiro (UERJ) e Universidade Veiga de Almeida (UVA), Rio de Janeiro-RJ, Brasil; https://orcid.org/0000-0002-5379-4495; *aanamen@iprj.uerj.br*

III Universidade do Estado do Rio de Janeiro (UERJ), Rio de Janeiro-RJ, Brasil; https://orcid.org/0000-0003-4944-9450; *fmoura@iprj.uerj.br*

IV Universidade do Estado do Rio de Janeiro (UERJ), Rio de Janeiro-RJ, Brasil; https://orcid.org/0000-0002-4435-2986; *arsilva@iprj.uerj.br*

V Universidade do Estado do Rio de Janeiro (UERJ), Rio de Janeiro-RJ, Brasil; https://orcid.org/0000-0001-7269-592X; *isabelortigao@terra.com.br*

VI Universidade do Estado do Rio de Janeiro (UERJ), Rio de Janeiro-RJ, Brasil; https://orcid.org/0000-0001-5743-5206; *ursula@iprj.uerj.br*

MINERÍA DE DATOS ORIENTADA POR EL DOMINIO EDUCATIVO: UNA PRUEBA DE CONCEPTO

RESUMEN

Este trabajo tiene el objetivo de proponer una metodología para identificar patrones relacionados con el aprendizaje de matemáticas y con las características del ambiente escolar, que fue aplicada a los datos de la Prova Brasil 2013, específicamente a los alumnos del 9º año de la educación fundamental del estado de Rio de Janeiro. El enfoque, apoyado por el conocimiento de especialistas en el área de educación, consistió en proponer un proceso de reducción de dimensionalidad integrado a la minería de datos, y su evaluación fue efectuada por medio de medidas técnicas en conjunto con medidas de interés del dominio educativo. Este abordaje permitió identificar acciones y analizar su viabilidad para solucionar algunos problemas de la educación básica. La metodología, orientada por el área de aplicación y cuya evaluación no se restringió al uso de métricas técnicas de la mineria de datos, puede servir de referencia – una prueba de concepto – a otras investigaciones en acciones y políticas educativas.

PALABRAS CLAVE PROVA BRASIL • MINERÍA DE DATOS • POLÍTICAS EDUCATIVAS • RENDIMIENTO ESCOLAR.

DATA MINING GUIDED BY THE EDUCATIONAL DOMAIN: A PROOF OF CONCEPT

ABSTRACT

This paper aims to propose a methodology for identifying patterns related to learning mathematics and the characteristics of the school environment. It was applied to the data of the Prova Brasil 2013, focusing on students in their 9th year of education, in the state of Rio de Janeiro. The approach, supported by education experts, consists of proposing a dimensionality reduction process integrated to data mining. Its evaluation was conducted using technical measures together with measures of interest to the educational domain. This approach made it possible to identify actions, as well as to analyze their viability, to solve educational issues. The methodology, guided by the area of application and the evaluation of which is not restricted to the use of technical metrics of data mining, can serve as a reference – a proof of concept – for other studies of educational actions and policies.

KEYWORDS PROVA BRASIL • DATA MINING • EDUCATIONAL POLICIES • SCHOOL PERFORMANCE.

INTRODUÇÃO

Avaliações visando à elaboração de diagnósticos para melhoria do ensino público passaram a ocupar papel de destaque na agenda política educacional (BAUER, 2012; BAUER; GATTI, 2013; BAUER; GATTI; TAVARES, 2013). Tais avaliações, geralmente compostas por testes e questionários, coletam uma série de informações que possibilitam mensurar o aprendizado dos estudantes e identificar as condições escolares a que esses estão submetidos, possibilitando nortear ações e, quiçá, garantir melhor gerenciamento dos recursos disponíveis.

O emprego de avançadas metodologias para análise profunda de dados, alicerçadas em conceitos estatísticos, torna-se essencial para o processamento de um grande volume de dados (SOARES, 2007; TCHIBOZO, 2009; VELOSO *et al.*, 2009). Busca-se, dessa forma, otimizar os benefícios advindos da aplicação de avaliações, configurando-se como a nova fronteira educacional (PIETY, 2013).

Nesse sentido, a mineração de dados, que promove a junção de conceitos de banco de dados, estatística e aprendizado de máquina, vem se estabelecendo como método de pesquisa com potencial para explorar conjuntos de dados educacionais. Segundo Romero e Ventura (2013), a crescente utilização

de técnicas de mineração de dados no contexto da educação decorre da possibilidade de fornecerem melhor entendimento de como os estudantes aprendem; por conseguinte, obtêm-se *insights* para explicar os fenômenos educacionais, viabilizando a identificação de ações para a melhoria do processo ensino-aprendizagem.

A difusão da aplicação de mineração de dados nas mais distintas áreas suscitou a discussão, de fato, os modelos computacionais gerados e os resultados obtidos estão sendo capazes de alicerçar decisões em problemas práticos. Essa contestação surge por se perceber uma lacuna entre os objetivos dos pesquisadores que utilizam os processos da mineração de dados e as expectativas dos tomadores de decisões. Para exemplificar, Cao *et al.* (2010) mencionam que, no vocabulário de pesquisadores, formulam-se frases como "muitos padrões foram encontrados" ou "os padrões satisfazem as medidas técnicas que foram estipuladas". No entanto, pessoas que não conhecem detalhes do processo e precisam dos resultados estão interessadas em respostas a questões do tipo: "como posso usar os padrões encontrados?", "eles satisfazem medidas técnicas, mas apoiam a formulação de mudanças?" ou "é possível compreender as saídas do modelo?".

Para enfrentar os desafios de conciliar as visões diferentes de pesquisadores e tomadores de decisões, o presente trabalho tem como principal objetivo o desenvolvimento de uma metodologia de mineração de dados que integre o interesse técnico à temática educacional. Em outras palavras, busca-se orientar o processo de mineração de dados por meio do conhecimento de especialistas em educação. Logo, esse estudo tem caráter interdisciplinar, envolvendo um grupo de pesquisadores especialistas em diferentes áreas nos campos da educação, educação matemática, matemática e computação.

Os dados analisados são provenientes da Prova Brasil, avaliação desenvolvida pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) com o intuito de mensurar a qualidade do ensino fundamental brasileiro (BRASIL, 2015). Por intermédio da metodologia aqui desenvolvida, buscaram-se relações entre as respostas fornecidas por professores, diretores e alunos a questionários contextuais e a proficiência obtida no teste de matemática pelos estudantes do 9º ano do ensino fundamental, tendo sido descobertos padrões/aspectos acerca do desempenho em matemática. Os padrões identificados foram avaliados por medidas técnicas, baseadas em modelos probabilísticos, e medidas qualitativas, aqui formuladas, relacionadas à temática educacional.

O presente trabalho apresenta, inicialmente, a revisão bibliográfica das pesquisas que vêm sendo conduzidas na linha de mineração de dados educacionais.

Posteriormente, apresentam-se os conceitos básicos de um processo de descoberta de conhecimento em bases de dados. Em seguida, as tarefas efetuadas para extração e avaliação dos padrões são descritas e, finalmente, tecem-se conclusões a respeito dos resultados obtidos.

PESQUISAS COM MINERAÇÃO DE DADOS EDUCACIONAIS

Inúmeros pesquisadores têm mostrado interesse em utilizar mineração de dados para investigar questões científicas na área educacional. Segundo Rodrigues *et al*. (2014), por exemplo, os estudos no Brasil minerando dados educacionais focam principalmente na estimação ou modelagem do desempenho de estudantes. Além disso, buscam abordar a aprendizagem colaborativa, dar suporte e recomendações pedagógicas aos professores em cursos de educação a distância, fornecer *feedback* aos alunos sobre as condições de aprendizagem, bem como detectar ou prever problemas frequentes, como evasão escolar.

Um dos estudos pioneiros no Brasil, nessa linha, foi o efetuado por Gomes, Levy e Lachtermacher (2004), que fizeram uso de dados relacionados aos indicadores do censo demográfico e do censo escolar de 2000 publicados no *site* do Inep. O objetivo do trabalho era responder às seguintes questões:

- Pode-se agrupar os municípios visando a encontrar características comuns em termos de indicadores de desempenho e investimentos em educação?
- Pode-se encontrar alguma relação entre o Índice de Desenvolvimento Humano (IDH) e os indicadores de desempenho?

Para responder essas perguntas, foram minerados dados, por meio da técnica de agrupamento, de 5.507 municípios brasileiros. Dentre os resultados do estudo, destaca-se que a região geográfica e a unidade da federação não são critérios para uniformizar os municípios em relação à educação.

O estudo proposto por Kampff, Reategui e Lima (2008) abordou dados coletados do ensino a distância por meio dos ambientes virtuais de aprendizagem (AVA). O objetivo foi descobrir a relação entre comportamentos e características dos alunos e suas possíveis repercussões no desempenho. Com esse mapeamento, seria possível gerar alertas para o professor a partir de agrupamentos de estudantes com necessidades similares, tornando sua mediação mais efetiva. Fez-se a mineração de dados por meio da tarefa de classificação de 161 estudantes de cursos de graduação. Observou-se que dificilmente os

alunos que não realizaram todas as atividades avaliativas propostas obtiveram nota acima da média.

Outro trabalho a ser ressaltado é o de Bezerra *et al*. (2016), que analisaram a evasão escolar dos alunos do 9º ano do ensino fundamental das escolas do estado de Pernambuco, baseados nos censos escolares 2011 e 2012. A análise foi efetuada por mais de uma técnica de mineração de dados, a saber, Árvore de Decisão, Indução de Regras e Regressão Logística, com o intuito de detectar o perfil do aluno evadido. Os resultados evidenciaram que aspectos como idade, turno das aulas e região geográfica das escolas influenciam fortemente a evasão.

Já os pesquisadores Fernandes *et al*. (2018) apresentaram uma análise preditiva do desempenho acadêmico de estudantes de escolas públicas do Distrito Federal durante os anos letivos de 2015 e 2016. Por intermédio de dados coletados pelos pesquisadores e da utilização de modelos de classificação baseados no algoritmo *gradient boosting machine* (GBM), foi possível identificar que atributos relacionados a notas, características escolares e idade dos discentes são indicadores potenciais do sucesso ou fracasso acadêmico de um aluno.

Fonseca e Namen (2016), por intermédio dos dados coletados pela Prova Brasil aplicada em 2011 e da utilização de mineração de dados, puderam identificar fatores que relacionam o perfil de professores que lecionam matemática com a proficiência obtida por seus alunos. Nesse estudo, deu-se enfoque aos estudantes do 9º ano do ensino fundamental residentes no estado do Rio de Janeiro.

Os dados utilizados na presente pesquisa contemplam não somente o perfil do professor, mas também as características dos alunos, diretores e escolas, envolvendo um volume de dados muito maior do que o utilizado em Fonseca e Namen (2016). Dessa forma, possibilitou-se expandir a descoberta de fatores a respeito de diferentes agentes (*stakeholders*) que constituem o ambiente educacional e mensurar a influência, positiva ou negativa, no desempenho dos discentes. Outro fato é que aqui se abordam os dados referentes a 2013, resultado mais recente disponibilizado pelo Inep, quando esse estudo teve início. Finalmente, como ponto central, a presente pesquisa transcende a efetuada em Fonseca e Namen (2016), por analisar os padrões descobertos por intermédio de uma perspectiva mais ampla. Os resultados obtidos foram analisados não somente por medidas técnicas (provenientes de um processo automatizado), mas também por medidas que abordam questões específicas da temática educacional.

Conforme discutido no decorrer deste artigo, os padrões descobertos, baseados na análise de medidas técnicas, permitiram a proposição de ações

visando à melhoria dos resultados em matemática. Ademais, as ações propostas foram avaliadas por intermédio da comparação de métricas relacionadas à viabilidade de sua implementação, aqui denominadas medidas de interesse do domínio educacional. Acredita-se que essa abordagem de caráter inovador contribui para uma discussão mais ampla entre os pesquisadores, especialistas em distintas áreas, sobre as questões educacionais.

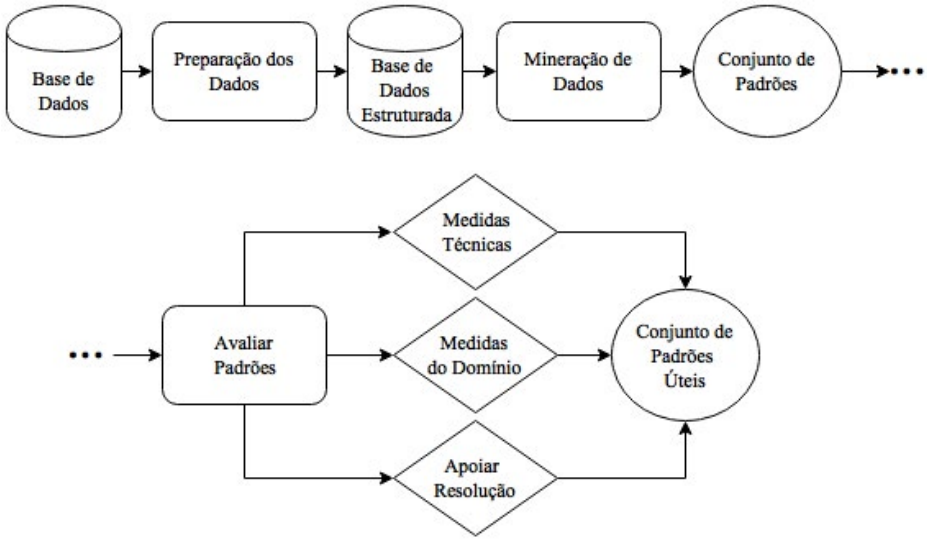
PROCESSO DE DESCOBERTA DE CONHECIMENTO

A descoberta de conhecimento, frequentemente conduzida em diversas aplicações por meio de processos tradicionais de mineração de dados, apresenta dificuldades para apoiar decisões referentes à resolução de problemas reais. Se, por um lado, as técnicas de mineração são eficientes na descoberta de padrões em grandes volumes de dados, por outro, tais técnicas são fracas na interpretação dos resultados, etapa fundamental para identificação e obtenção de conhecimento efetivamente útil (ADEJUWON; MOSAVI, 2010). Algumas razões podem ser listadas: aplicação de um enfoque reducionista, em vez de um enfoque holístico; preferência por paradigmas automatizados de análise de dados, em vez de considerar o julgamento humano como central; e maior foco na adaptação automática, desconsiderando-se o apoio da intervenção humana (CAO, 2010).

Para contornar tais limitações, Cao e Zhang (2005) propuseram a metodologia de mineração de dados orientada pelo domínio (*domain-driven data mining* – D³M), que tem como objetivo o uso de métodos, técnicas, ferramentas e aplicações eficientes que possam prover a descoberta de conhecimento orientada pelo domínio do problema envolvido. Nesse sentido, o conhecimento do domínio é crucial para garantir o uso efetivo dos resultados da mineração de dados (CAO *et al.*, 2010).

A Figura 1 apresenta a visão geral do processo de descoberta de conhecimento seguindo os conceitos da D³M.

FIGURA 1 – Processo de descoberta de conhecimento



Fonte: Elaboração dos autores.

Conforme apresentado na Figura 1, inicialmente a base de dados é submetida às tarefas de preparação dos dados, essenciais para garantir a confiabilidade dos resultados posteriores. Dentre as tarefas, podem-se citar o entendimento do domínio da aplicação para seleção dos dados necessários, remoção de dados faltantes, tratamento de inconsistências, redução de dimensionalidade, normalização e discretização de atributos.

Com os dados estruturados, a etapa seguinte consiste na mineração de dados. Segundo Fayyad, Piatetsky-Shapiro e Smyth (1996), nessa fase, aplicam-se algoritmos de descoberta que, sob certas limitações computacionais, identificam um conjunto de padrões nos dados. Finalmente, os padrões obtidos precisam ser avaliados. Em D³M, diferentemente das abordagens tradicionais, os padrões são analisados por medidas técnicas e medidas de interesse do domínio de aplicação, auxiliando a escolha das ações a serem implementadas para resolução do problema, de tal modo a otimizar as melhorias no cenário em questão.

Dito de outra forma, um padrão extraído por meio de um algoritmo de mineração será considerado útil se satisfizer as medidas técnicas e do domínio e, ainda, se for capaz de apoiar, por intermédio da seleção das ações a serem implementadas, melhorias na correção dos problemas detectados pela consulta à base de dados. Ressalta-se que tais medidas e ações devem ser formuladas

levando-se em conta fortemente a opinião dos pesquisadores especialistas do domínio da aplicação.

Ocorre, portanto, uma contínua intervenção humana no processo, de tal modo a inserir o conhecimento do domínio e garantir que os resultados fomentem e enriqueçam discussões na prática. Diante do exposto, segue-se a apresentação do domínio de aplicação e do problema aqui abordado; posteriormente são apresentadas as etapas ilustradas na Figura 1.

DOMÍNIO: DADOS DA PROVA BRASIL

Conforme dito, faz-se uso dos dados da Prova Brasil. Essa avaliação teve sua primeira edição em 2005 e, desde então, ocorre bienalmente. Em 2013, aplicaram-se testes nas áreas de língua portuguesa e matemática aos estudantes dos 5º e 9º anos do ensino fundamental, bem como questionários que coletavam informações a respeito desses alunos, professores, diretores e escolas. Os dados, de acesso público, são disponibilizados no *site* do Inep (www.inep.gov.br).

No presente trabalho, buscou-se relacionar as respostas dadas aos questionários ao desempenho obtido no teste de matemática aplicado a estudantes do 9º ano do ensino fundamental. A medida de desempenho é obtida por meio da aplicação da Teoria da Resposta ao Item, abordagem metodológica que possibilita a comparabilidade dos resultados e tem como referência as respostas dadas em cada item (KLEIN, 2003). A Tabela 1 apresenta informações sobre os arquivos utilizados no estudo, incluindo número de registros e de atributos de cada instrumento contextual.

TABELA 1 – Informações sobre os arquivos da Prova Brasil de 2013 utilizados no estudo

ARQUIVO	DESCRIÇÃO	Nº. DE REGISTROS	Nº. DE ATRIBUTOS
TS_ALUNO_9EF	Respostas ao *Questionário do aluno* e proficiência dos alunos do 9º ano do ensino fundamental	2.720.588	92
TS_PROFESSOR	Respostas ao *Questionário do professor* de cada disciplina de cada série	237.186	134
TS_DIRETOR	Respostas ao *Questionário do diretor* de cada escola	56.737	118
TS_ESCOLA	Média da proficiência dos alunos por disciplina e respostas ao *Questionário da escola*	59.251	127

Fonte: Adaptado do Inep (BRASIL, 2015).

O número de registros expressa, em âmbito nacional, o número de alunos, professores, diretores e escolas com informações nos respectivos arquivos. Os atributos, por sua vez, armazenam as respostas dadas às perguntas presentes nos questionários, notas obtidas, além de identificarem e permitirem a interligação das informações nos diferentes arquivos como, por exemplo, conectar aluno e escola. Com relação aos questionários, é importante mencionar que o respondido pelo aluno é composto por 57 questões, o aplicado ao professor contém 125, o preenchido pelo diretor tem 111 e o que coleta informações acerca da escola contém 68, perfazendo o total de 361 variáveis associadas a cada aluno.

Os arquivos descritos na Tabela 1 foram importados, em forma de tabelas, para o *software* PostgreSQL (POSTGRESQL GLOBAL DEVELOPMENT GROUP, 1995), sistema gerenciador de banco de dados que permite manipulações em registros e atributos presentes em tabelas.

PREPARAÇÃO DOS DADOS

Seleção dos dados

Inicialmente foram selecionados os discentes, professores, diretores e escolas do estado do Rio de Janeiro. Tal seleção foi motivada por entender que cada estado brasileiro tem uma rede pública com características específicas e, portanto, os fatores relacionados ao desempenho devem ser estudados separadamente.

Outra tarefa executada foi a exclusão dos alunos que não fizeram ou que responderam somente a três ou menos itens do teste de matemática, uma vez que não se adequavam ao cálculo da proficiência. Foram removidos, ainda, todos aqueles que não responderam a, no mínimo, 70% das perguntas dos respectivos questionários. Ademais, para garantir a consistência entre as bases de dados, foram mantidos somente os docentes, diretores e escolas que se relacionavam a algum aluno na base TS_ALUNO_9EF. Portanto as tabelas foram interligadas com o intuito de identificar, para cada aluno, o seu professor de matemática, a sua escola e o seu diretor.

Após essas seleções e remoções, cada tabela passou a conter a seguinte quantidade de registros: 113.021 alunos em TS_ALUNO_9EF; 2.388 professores em TS_PROFESSOR; 1.771 diretores em TS_DIRETOR; e 1.764 escolas em TS_ESCOLA.

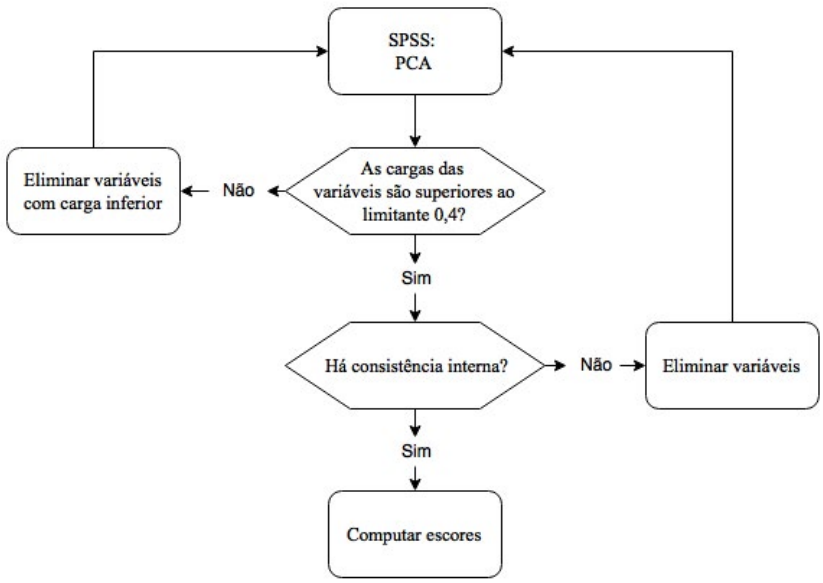
Redução de dimensionalidade

Conforme mencionado, os quatro questionários englobam 361 questões. Em virtude desse alto número de variáveis, a redução de dimensionalidade foi essencial na etapa de preparação dos dados. Etapa essa que consiste em reduzir o número de questões a serem analisadas, a partir de sua troca por atributos agregados, os construtos, mantendo a relevância empírica.

O processo de redução de dimensionalidade foi efetuado por intermédio do conhecimento dos especialistas envolvidos nesta pesquisa e, portanto, em conformidade com os conceitos preconizados pela D³M e tendo como referência uma revisão da literatura específica. Os pesquisadores, a partir da análise de cada questionário, conjecturaram quais questões estariam inter-relacionadas. Os grupos de variáveis abordando um tema comum seriam os possíveis construtos para o estudo. É importante salientar que, devido ao grande número de variáveis, muitos construtos foram identificados. Assim, alguns foram escolhidos para serem submetidos a análises posteriores.

Os construtos conjecturados foram analisados individualmente, seguindo o processo descrito na Figura 2.

FIGURA 2 – Processo para cada construto sugerido



Fonte: Elaboração dos autores.

Conforme exposto na Figura 2, por meio da utilização do *software* IBM SPSS Statistics 23 (IBM CORPORATION, 1989), a análise de componentes principais (*principal components analysis* – PCA) (SMITH, 2002) foi aplicada às questões escolhidas relacionadas a cada construto, com a especificação de obtenção de um único componente. Posteriormente, verificou-se a carga de cada variável, ou seja, sua correlação com cada um dos componentes principais. Caso a carga fosse inferior a 0,4 (em valor absoluto), a variável seria eliminada do construto e a análise refeita. Cabe ressaltar que esse valor mínimo especificado é recomendado por Stevens (1992), baseando-se na justificativa de que esse ponto de corte é apropriado para facilitar a interpretação. Além disso, o autor justifica que, elevando-se a carga de uma variável ao quadrado, obtém-se uma estimativa do montante da variância em um componente por ela explicada. Assim, o valor de 0,4 explica por volta de 16% da variância da variável.

A escolha das variáveis para compor cada construto seguiu também critérios qualitativos, como teoria e parcimônia. O primeiro relaciona-se com a teoria dos estudos sociológicos e o segundo, com o cuidado que pesquisadores devem ter em relação ao número de fatores selecionados para entrar no modelo. Não existe um número ideal, mas o melhor modelo, do ponto de vista desse critério, é um modelo tão simples quanto possível, que seja capaz de fornecer explicações plausíveis e consistentes.

Após verificar se todas as variáveis satisfaziam o limite mínimo definido a respeito da carga, analisou-se a consistência interna do construto. Essa etapa foi efetuada por meio da medida α de Cronbach (CRONBACH, 1951) e da média das correlações entre as variáveis. Em termos práticos, o limite inferior para α de Cronbach geralmente aceito é 0,7. Contudo, em pesquisas exploratórias, esse valor pode ser diminuído para 0,6 (HAIR *et al.*, 2005; KLINE, 1999; ROBINSON; SHAVER; WRIGHTSMAN, 1991). Ademais, segundo Clark e Watson (1995), a média das correlações entre as variáveis deve ficar entre 0,15 e 0,5. Portanto, caso uma variável comprometesse o valor da medida α de Cronbach ou tivesse baixa correlação, deveria ser eliminada e a análise refeita.

Os construtos submetidos a esse processo e as respectivas medidas obtidas são apresentados na Tabela 2.

TABELA 2 – Características dos construtos: α de Cronbach, média das correlações e questões utilizadas

QUESTIONÁRIO / CONSTRUTO	α DE CRONBACH	MÉDIA DAS CORRELAÇÕES	QUESTÕES UTILIZADAS
ALUNO			
Posse de bens	0,805	0,280	5 a 15
Escolaridade dos pais	0,653	0,487	19 e 23
Hábito de leitura	0,721	0,248	32 a 37, 39 e 56
PROFESSOR			
Experiência	0,826	0,612	13 a 15
Necessidade de aperfeiçoamento profissional	0,891	0,576	26 a 31
Impedimentos ao desenvolvimento profissional	0,716	0,386	34 a 37
Hábitos de leitura	0,656	0,241	38 a 43
Uso de recursos audiovisuais e didáticos	0,777	0,367	44 a 47,49 e 50
Integração da equipe escolar	0,925	0,406	51, 53 a 69
Problemas de aprendizagem dos alunos	0,647	0,169	70, 72 a 79
Expectativa sobre formação dos alunos	0,661	0,493	95 e 96
Livro didático	0,687	0,422	99 a 101
Práticas pedagógicas	0,705	0,285	107 a 110, 112 e 113
Práticas pedagógicas em matemática	0,783	0,376	120 a 125
DIRETOR			
Experiência	0,751	0,501	16 a 18
Frequência e fluxo escolar	0,739	0,361	44 a 48
Relação com a comunidade externa	0,635	0,367	53 a 55
Merenda escolar	0,683	0,301	62 a 66
Funcionamento da escola	0,788	0,292	67 a 71, 73 a 76
ESCOLA			
Segurança da escola	0,829	0,410	24 a 29, 35
Espaços da escola	0,716	0,239	57 a 64

Fonte: Elaboração dos autores.

Observa-se que 21 construtos foram analisados. Sete, dentre os 21, tinham α de Cronbach inferior a 0,7. No entanto, todos tinham valor superior a 0,6 (o menor valor, 0,635, refere-se ao construto "Relação com a comunidade externa"). Para corroborar essa confiabilidade, nota-se, ainda, que os sete construtos têm média das correlações entre suas variáveis com valor entre 0,15 e 0,5, satisfazendo o critério de Clark e Watson (1995).

Ao final do processo, computaram-se os escores. Assim, cada aluno presente na base de dados passou a ter uma pontuação nos construtos "Posse de bens", "Escolaridade dos pais" e "Hábito de leitura". Analogamente, cada professor, diretor e escola, presentes nas respectivas bases de dados, passaram a

ter pontuações referentes aos construtos extraídos, apresentados na Tabela 2. Mais detalhes do processo de redução de dimensionalidade descrito neste artigo podem ser vistos em Fonseca (2018).

É importante mencionar que, durante a condução do processo de redução de dimensionalidade, foram consultados trabalhos que abordavam os construtos presentes nos questionários das avaliações do Sistema de Avaliação da Educação Básica (Saeb). Podem-se citar o documento *Saeb 2001: Novas perspectivas* (BRASIL, 2001), o artigo de Franco *et al*. (2003) e a pesquisa efetuada por Karino, Vinha e Laros (2014). Tais trabalhos serviram como arcabouço teórico e direcionaram o estudo aqui desenvolvido.

Integração, normalização e discretização

Após o processo de redução de dimensionalidade, as bases de dados foram integradas com o intuito de agregar a cada aluno as pontuações correspondentes ao seu professor, diretor e escola. Portanto a tabela TS_ALUNO_9EF passou a conter os 21 novos atributos (construtos) apresentados na Tabela 2. Além disso, essas 21 variáveis, que alocam os escores, são contínuas com diferentes amplitudes, isto é, têm distintos valores de máximo e mínimo. Para melhor entendimento desses dados, os valores dessas variáveis foram normalizados entre 0 e 1, dividindo-se todos os valores da variável pelo máximo valor.

Outra tarefa de preparação foi efetuada com a finalidade de transformar esses 21 atributos do tipo contínuo para discreto. Tal transformação recebe o nome de "discretização" e consiste, basicamente, de duas subtarefas: escolher o número de categorias que o atributo terá e ainda estabelecer como os valores contínuos serão mapeados para essas diferentes categorias (HAN; KAMBER; PEI, 2012).

Assim, primeiramente foi assumido que todos os 21 atributos teriam quatro categorias. Essas categorias foram intituladas 1, 2, 3 e 4. O mapeamento de uma variável arbitrária X_k, com k = 1, ..., 21 ocorreu de tal modo que a categoria 1 se referiu ao percentual de 10% dos registros com valores mais baixos do atributo X_k (considerando-se apenas os valores não nulos de X_k), ou seja, 10% do total de registros que obtiveram as menores pontuações; a categoria 4 referiu-se aos 10% dos registros com os valores mais altos do atributo X_k ou, equivalentemente, os 10% de registros, considerando-se apenas os com valores não nulos que obtiveram as maiores pontuações; os registros restantes foram alocados nas categorias 2 e 3, de modo que a primeira metade pertenceria à categoria 2 e a segunda metade pertenceria à categoria 3, cada uma representando 40% dos registros com valores de X_k não nulos.

Para exemplificar a discretização dos atributos, considere a variável relacionada à posse de bens. Dos 113.021 alunos, 11.033 não tiveram escore calculado (tal situação ocorre, pois o aluno deveria responder a todas as questões presentes no construto, mas há situações em que isso não ocorre). Assim, esse atributo tem 101.988 valores não nulos. A categoria 1 refere-se aos 10% dos 101.988 registros que obtiveram as menores pontuações (intervalo de pontuação da categoria 1: [0, 0,1848]); a categoria 4 refere-se aos 10% dos 101.988 registros que obtiveram as maiores pontuações (intervalo de pontuação da categoria 4: [0,42825, 1]); os registros restantes foram alocados nas categorias 2 (intervalo de pontuação da categoria 2: (0,1848, 0,28566)) e 3 (intervalo de pontuação da categoria 3: (0,28566, 0,42825)). Seguindo a lógica dessa divisão, os alunos alocados à categoria 1 do construto Posse de bens têm muito menos bens do que os localizados na categoria 4.

Processo análogo foi feito para os outros atributos; por exemplo, para o atributo "Problemas de aprendizagem dos alunos", os alunos na categoria 1 têm professores que estão vivenciando muito menos problemas de aprendizagem de seus alunos do que os professores que lecionam para os alunos presentes na categoria 4. Interpretação nesse sentido também é dada ao atributo "Necessidade de aperfeiçoamento profissional", ou seja, quanto maior o valor do escore, menos favorável é o cenário relacionado ao seu aperfeiçoamento.

Como se desejou descobrir relações entre os 21 atributos e o desempenho obtido em matemática, também foi efetuada a discretização da variável-alvo que armazena a nota obtida pelos discentes no teste. É importante mencionar que a escala de proficiência em matemática varia de 0 a 500.

A distinção em categorias ou classes do atributo-alvo foi realizada de duas formas: uma visando à descoberta de fatores que pudessem influenciar de forma positiva e outra à descoberta de fatores que pudessem influenciar negativamente o desempenho dos alunos.

Para a influência positiva, os registros de alunos foram mapeados de tal modo que aqueles com notas superiores ou iguais a 308,43 fossem alocados na classe Nota Alta. Os registros restantes, com proficiências inferiores a esse limite, ficaram alocados em uma classe denominada Outra. Já para a análise da influência negativa, a discretização ocorreu de forma a alocar os alunos com notas inferiores ou iguais a 180,44 na classe Nota Baixa, enquanto aqueles com notas superiores fossem alocados na categoria Outra.

Cabe salientar que os limites, 308,43 e 180,44, foram especificados para que, respectivamente, as classes Nota Alta e Nota Baixa tivessem 10% do total de alunos presentes na base, ou seja, 11.302 discentes. Nota-se que esse

percentual mantém a coerência em relação à discretização efetuada para os outros 21 atributos.

Após as tarefas de preparação, a base de dados contendo os 113.021 registros de alunos, as 21 variáveis explicativas (construtos) e a variável-alvo (discretizada de duas formas) encontrava-se estruturada para a aplicação do algoritmo de mineração de dados.

MINERAÇÃO DE DADOS

A etapa de mineração consiste na aplicação de algoritmos para extração de padrões embutidos nos dados. No presente trabalho, utilizou-se o algoritmo Naïve Bayes (JOHN; LANGLEY, 1995), que consiste em classificar um registro em determinada classe, alicerçando-se na probabilidade de esse registro pertencer a essa classe (HAN; KAMBER; PEI, 2012). Especificamente, neste estudo, utilizou-se a implementação do algoritmo Naïve Bayes disponibilizada pelo *software* Weka (UNIVERSITY OF WAIKATO, 1999), ferramenta de código aberto que contém uma série de algoritmos de mineração de dados. Mais detalhes do algoritmo e da ferramenta podem ser vistos em Witten, Frank e Hall (2011).

Por intermédio das variáveis explicativas, pretendeu-se identificar em qual classe do atributo-alvo os estudantes se enquadravam. Em outras palavras, buscou-se identificar quais os construtos que permitiam classificá-los como Nota Alta (influência positiva) ou Nota Baixa (influência negativa).

RESULTADOS E DISCUSSÕES

Nesta seção, discutem-se os padrões extraídos, por meio do algoritmo Naïve Bayes, sobre os fatores que poderiam influenciar o desempenho em matemática dos discentes do 9º ano do ensino fundamental residentes no estado do Rio de Janeiro. A relevância dos padrões é avaliada por meio de medidas técnicas em conjunto com medidas de interesse do domínio educacional.

O modelo gerado por Naïve Bayes apresenta o percentual de registros de alunos da classe Nota Alta que se encontram em cada categoria dos atributos. Quanto maior a porcentagem de uma categoria, maior será a sua influência para a classe Nota Alta. Por exemplo, para o atributo Posse de bens, primeira linha da Tabela 3, 44,54% dos registros da base classificados como Nota Alta têm valor do atributo na categoria 3. Além disso, pode-se observar que 56,62% concentram-se nas categorias 3 e 4, correspondentes às maiores pontuações do construto Posse de bens. Isso quer dizer que a maioria dos alunos com

Nota Alta tem melhores condições econômicas do que seus colegas com igual desempenho, confirmando resultados apontados em diversos estudos, desde a divulgação do Relatório Coleman, na década de 1960 (COLLEMAN, 1966).

TABELA 3 – Distribuição percentual dos alunos Nota Alta por categoria do atributo, segundo o construto

QUESTIONÁRIO	ATRIBUTO REFERENTE AO CONSTRUTO	CATEGORIAS DO ATRIBUTO (%)				SOMA[1] DE 3 E 4
		1	2	3	4	
Aluno	Posse de bens	6,50	36,88	44,54	12,08	56,62
	Escolaridade dos pais	10,37	23,91	31,62	34,10	65,72
	Hábito de leitura	9,30	39,34	39,59	11,77	51,36
Professor	Experiência	8,94	32,67	43,47	14,92	58,39
	Necessidade de aperfeiçoamento profissional	11,20	40,07	39,13	9,6	48,73
	Impedimentos ao desenvolvimento profissional	10,56	39,91	40,84	8,69	49,53
	Hábitos de leitura	10,32	42,08	39,50	8,10	47,60
	Uso de recursos audiovisuais e didáticos	9,16	39,59	39,52	11,73	51,25
	Integração da equipe escolar	10,75	36,91	38,99	13,35	52,34
	Problemas de aprendizagem dos alunos	16,98	37,93	35,34	9,75	45,09
	Expectativa sobre formação dos alunos	7,47	29,62	26,95	35,96	62,91
	Livro didático	6,88	36,73	43,43	12,96	56,39
	Práticas pedagógicas	9,51	37,4	41,67	11,42	53,09
	Práticas pedagógicas de matemática	10,03	38,35	37,64	13,98	51,62
Diretor	Experiência	10,86	42,46	34,65	12,03	46,68
	Frequência e fluxo escolar	9,52	32,99	42,01	15,48	57,49
	Relação com a comunidade externa	10,25	42,42	35,30	12,03	47,33
	Merenda escolar	10,81	38,30	38,4	12,49	50,89
	Funcionamento da escola	6,57	31,89	44,92	16,62	61,54
Escola	Segurança da escola	13,19	35,14	41,62	10,05	51,67
	Espaços da escola	8,29	35,74	35,69	20,28	55,97

Fonte: Fonseca (2018).

Para identificação dos padrões, a partir dos resultados advindos da mineração, medidas técnicas e da temática educacional foram formuladas e analisadas.

1 Percentuais maiores relacionados às categorias 3 e 4 indicam melhores condições associadas ao construto, exceto para os construtos "Necessidade de aperfeiçoamento profissional', 'Impedimentos ao desenvolvimento profissional' e 'Problemas de aprendizagem dos alunos", nos quais menores percentuais indicam melhores condições.

Medidas técnicas

Para avaliar os padrões, duas medidas técnicas foram estabelecidas. A primeira objetivou mensurar a importância dos padrões, de modo a ordená-los. Já a segunda procurou complementar a análise, verificando se a situação inversa dos padrões, analisando o aspecto da nota baixa (no lugar da nota alta), tinha efeito adverso.

A primeira medida a ser fixada foi "porcentagem de alunos da classe Nota Alta que se encontram nas duas categorias que representam o melhor cenário", nomeada *MT*. Considerou-se que seu limite fosse 57%, isto é, exigiu-se que, no mínimo, 57% dos alunos da classe Nota Alta estivessem nas categorias consideradas favoráveis ao desempenho. Ressalta-se que, para definir esse limite, buscou-se um valor que possibilitasse a seleção de um número viável de construtos para posterior análise. Caso esse limite fosse cumprido, acreditou-se que os alunos com êxito em matemática apresentariam uma característica semelhante (ou seja, um padrão) e, consequentemente, permitiriam apontar a relevância do construto para um melhor desempenho em matemática.

Conforme pode ser visto na Tabela 3, cinco construtos, dentre os 21, satisfazem a condição $MT \geq 57\%$, a saber: "Escolaridade dos pais"; "Expectativa dos professores sobre formação dos alunos"; "Funcionamento da escola"; "Experiência do professor"; e "Frequência e fluxo escolar".

A segunda medida técnica adotada, denominada MT_2, analisou se os construtos, selecionados a partir de *MT*, apresentavam situação inversa quando os alunos da classe Nota Baixa fossem analisados. O objetivo buscado, simplesmente, foi o de confirmar se condições piores relacionadas aos construtos implicavam também pior desempenho dos alunos. Ressalta-se que, nesse caso, para obtenção dessa confirmação, bastaria que, pelo menos, mais de 50% dos alunos da classe Nota Baixa estivessem nas categorias que representassem o pior cenário (categorias 1 e 2). Assim, a escolha da segunda medida técnica (MT_2: = porcentagem de alunos da classe Nota Baixa que se encontram nas duas categorias que representam o pior cenário) considerou um limite igual a 50%. Observou-se que todos os cinco construtos satisfizeram a restrição sobre a medida técnica MT_2, ou seja, $MT_2 > 50\%$ (Tabela 4).

TABELA 4 – Distribuição percentual dos alunos Nota Baixa por categoria do atributo, segundo o construto

QUESTIONÁRIO	ATRIBUTO REFERENTE AO CONSTRUTO	CATEGORIAS DO ATRIBUTO (%)				SOMA DE 1 E 2
		1	2	3	4	
Aluno	Escolaridade dos pais	21,15	32,98	27,49	18,38	54,13
Professor	Experiência	11,50	44,72	35,07	8,71	56,22
	Expectativa sobre formação dos alunos	13,76	44,49	26,34	15,41	58,25
Diretor	Frequência e fluxo escolar	12,25	40,37	37,61	9,77	52,62
	Funcionamento da escola	12,69	45,44	34,72	7,15	58,13

Fonte: Fonseca (2018).
Nota: 'Soma de 1 e 2' > 50%.

Portanto, a partir da análise das medidas técnicas, cinco padrões puderam ser identificados como relevantes para o desempenho dos estudantes em matemática:

- P_1: quanto maior o nível escolar dos pais, melhores são os resultados alcançados pelos seus filhos;
- P_2: quanto maior a expectativa dos professores em relação à formação futura de seus alunos, melhores são os resultados dos discentes;
- P_3: quanto melhor o funcionamento da escola, melhores são os resultados do corpo discente;
- P_4: quanto maior a experiência do professor, melhores são os resultados dos seus alunos;
- P_5: quanto maior a frequência dos alunos na escola, melhores são os resultados alcançados por esses.

Analisando os valores da medida técnica *MT*, presentes na Tabela 3, nota-se que o padrão P_1 apresentou o resultado mais expressivo (65,72%), seguido por P_2, e assim sucessivamente. Nessa ótica, caso estivesse sendo executada a mineração tradicional, o processo de descoberta de padrões deveria ser encerrado e o resultado obtido seria entregue aos tomadores de decisão. Esses, por sua vez, priorizariam a elaboração de políticas públicas, de acordo com a ordem de importância desses construtos. No entanto, conforme mencionado anteriormente, o processo de caráter inovador elaborado nesta pesquisa propõe a análise dos padrões não somente considerando medidas técnicas, provenientes de um processo automatizado, mas utilizando medidas de interesse da temática em questão.

Medidas do domínio

Para que os padrões possam apoiar soluções para o problema do baixo desempenho obtido no teste de matemática, é necessário considerar as preocupações acerca do domínio da questão abordada. O objetivo é formular medidas que captem informações e insiram o conhecimento de especialistas em educação.

Sob essa perspectiva, foram formuladas três medidas do domínio educacional. Assim, cada padrão obtido anteriormente será analisado com base nas seguintes medidas:

- Custo: C : = "custo requerido para resolver os problemas relacionados com o padrão";
- Tempo: T : = "tempo indispensável para resolver as questões relacionadas com o padrão";
- Rejeição: R : = "nível de rejeição dos professores com respeito às decisões tomadas baseadas no padrão".

A abordagem consiste em indicar ações a serem desenvolvidas com base em cada padrão descoberto. A partir dessas ações, deve-se avaliar qual o custo para a sua implementação (C), qual o tempo necessário para que as ações possam resolver o problema envolvido (T) e, por fim, qual o nível de resistência do corpo docente para que as ações formuladas sejam colocadas em prática (R).

Nesses moldes de avaliação dos padrões, o julgamento humano é central, substituindo-se paradigmas automatizados de análises de dados. É importante ressaltar que, obviamente, outras medidas de interesse da temática educacional poderiam ser formuladas. No entanto, as aqui mencionadas permitem ilustrar o processo – uma prova de conceito – e fomentar uma discussão mais ampla relacionada à elaboração de políticas públicas na área da educação.

Inicialmente foram verificados quais padrões, entre os cinco identificados, permitiriam a elaboração de ações e, consequentemente, a avaliação das medidas C, T e R. Os autores do presente artigo consideraram, em um primeiro momento, a exclusão dos padrões P_3 e P_4 relacionados, respectivamente, ao funcionamento da escola e à experiência do professor. O primeiro, por apresentar um alto grau de complexidade no que tange à definição de ações, uma vez que abrange temas diversos, como insuficiência de recursos financeiros e pedagógicos, carência de apoio administrativo e de professores, alto índice de falta por parte de professores e alunos, alta rotatividade do corpo docente e indisciplina dos discentes. Já o construto referente à experiência do professor não foi considerado nas análises posteriores, pois avaliou-se que ações relacio-

nadas a essa questão seriam restritas, uma vez que essa experiência está ligada diretamente ao tempo de atuação do profissional. Assim, haveria limitações relacionadas a ações sobre essa variável, visto que há grande dificuldade em agir sobre o tempo.

Portanto foi decidido priorizar intervenções visando à elevação da escolaridade dos pais, ao aumento da expectativa dos professores em relação à formação futura de seus alunos e ao incremento da frequência dos estudantes na escola. Cabe ressaltar que se pretende efetuar uma análise futura do padrão P_3, sendo a escolha inicial de aprofundamento do estudo referente aos outros três padrões uma decisão puramente pragmática, baseada em um plano de desenvolvimento de pesquisa, visando à obtenção de diferentes resultados em etapas distintas.

Nessa prova de conceito, a formulação das intervenções partiu da ideia-base de contratar professores, com a mesma formação e carga horária de 40 horas semanais, para solucionar os três problemas, evidentemente atuando de modos distintos de acordo com cada situação. Dessa forma, fez-se a "normalização" da solução, buscando-se um recurso comum (o professor), com alocações diferenciadas em conformidade com cada caso. Assumiu-se, também, a hipótese de que as ações formuladas poderiam ser implementadas em escolas com a característica de ter até 60 alunos matriculados no 9º ano do ensino fundamental, situação encontrada em 58% das escolas públicas no estado do Rio de Janeiro, de acordo com análise das bases de dados do Inep relativas a 2013. Essa hipótese viabilizaria as ações recomendadas, uma vez que a alocação dos professores previstos seria suficiente para atendimento às demandas previstas. A seguir são descritas as ações para cada padrão.

Ações definidas

Escolaridade dos pais (P_1): para elevar o nível de escolaridade, sugere-se a criação de cursos na modalidade Educação de Jovens e Adultos (EJA) seguindo-se o formato do já existente Programa Brasil Alfabetizado (BRASIL, 2011), do Governo Federal. Para isso, a escola deveria contratar três professores, sendo um para atuar na alfabetização, outro nos anos iniciais e um terceiro para os anos finais do ensino fundamental. O curso de alfabetização teria oito meses de duração com, no mínimo, 320 horas/aula; os segmentos dos anos iniciais e finais do ensino fundamental teriam a duração dos cursos de EJA, ou seja, quatro semestres (total de 24 meses), com carga horária de 1.600 horas.

Expectativa do professor quanto à formação futura dos alunos (P_2): para aumentar a expectativa dos professores em relação ao sucesso na formação futura

dos seus alunos, uma ação seria oferecer um curso (ou grupo de discussão) sobre o Efeito Rosenthal (ROSENTHAL; JACOBSON, 1966), de modo a levar os participantes a refletir sobre a ideia de que, quanto maior a expectativa em relação a uma pessoa, melhor acaba sendo seu desempenho. Esse curso seria ministrado/conduzido por um professor, com a mesma carga horária semanal e base salarial da ação anterior, com o foco de sensibilizar os professores da escola a mudar sua postura com respeito à expectativa de formação futura dos seus alunos (não somente os professores de matemática, de forma a otimizar a alocação do professor contratado). Detalhes sobre essa questão podem ser encontrados em Barbosa e Randall (2004) e Fonseca e Namen (2016). Cabe ressaltar que, de modo empírico, ficou estabelecido, pelos especialistas que formularam essa proposta, que o curso tivesse duração de seis meses.

Frequência e fluxo escolar (P_5): para ter compatibilidade com as duas primeiras ações e alcançar o objetivo de ampliar a frequência dos discentes na escola, propõe-se a contratação de dois professores para se dedicarem à elaboração de atividades estimulantes, com a organização de tarefas extracurriculares envolvendo atividades esportivas, culturais, entre outras. Consequentemente, a escola se tornaria mais atrativa. Além disso, essas atividades seriam oferecidas somente aos alunos com boas frequências, com o intuito de motivar a assiduidade dos estudantes faltosos. Diferentemente das ações anteriores, a contratação desses docentes seria por tempo indeterminado, ou seja, tratar-se-ia de atividade contínua. No entanto, para quantificar o custo, será considerado um período de 24 meses (tempo máximo dos cursos ofertados).

Assume-se que as ações definidas implicarão um nível igual de impacto positivo nas diferentes variáveis consideradas. Em outras palavras, parte-se da hipótese de que o grau de elevação no nível de escolaridade dos pais dos estudantes, na expectativa dos professores em seus alunos e na frequência dos discentes na escola, dar-se-á de forma idêntica, a partir das ações propostas relacionadas a essas três dimensões do problema.

As ações que envolvem os padrões P_z, com $z = 1, 2$ e 5, devem ser avaliadas pelas medidas C, T e R.

Cálculo das medidas do domínio

Inicialmente, é importante mencionar a notória dificuldade em quantificar o custo de cada ação proposta, uma vez que envolve várias variáveis inerentes a situações específicas. Nesta pesquisa, efetuou-se a normalização da solução por intermédio da contratação de professores. Desse modo, o

custo foi calculado considerando somente uma variável: o salário do professor. Além disso, os docentes contratados têm a mesma carga horária semanal e base salarial. Por conseguinte, para comparar os custos relacionados às ações, basta observar o montante de meses envolvidos em cada uma. No entanto, para ilustrar mais concretamente a abordagem, foi considerado que o docente receberia o salário mensal de R$ 4.692,77 (vencimento básico de R$ 4.234,77, acrescido do auxílio alimentação no valor de R$ 458,00). Esse valor corresponde à remuneração de um professor do ensino básico, técnico e tecnológico do quadro permanente do Instituto Federal de Educação, Ciência e Tecnologia do Rio de Janeiro (IFRJ, 2016). Considerou-se um docente graduado com carga horária de 40 horas semanais e dedicação exclusiva.

Os valores das três medidas, para cada ação envolvendo cada padrão, são apresentados na Tabela 5.

TABELA 5 – Valores das medidas do domínio

PADRÃO	CONSTRUTO	MEDIDAS DO DOMÍNIO		
		CUSTO (EM R$)	TEMPO (EM MESES)	REJEIÇÃO
P_1	Escolaridade dos pais	262.795,12	56	Baixa
P_2	Expectativa sobre formação dos alunos	28.156,62	12	Média
P_5	Frequência e fluxo escolar	225.252,96	6	Muito baixa

Fonte: Fonseca (2018).

Com relação ao custo, foi computado o salário dos professores ao longo dos meses de duração dos cursos propostos. Desse modo, para o padrão P_1, foram considerados 56 meses (8 meses para o curso de alfabetização, 24 meses para cada uma das modalidades do ensino fundamental – anos iniciais e anos finais); para o padrão P_2, foi considerado que o curso teria duração de 6 meses, conforme exposto na seção anterior; já para o padrão P_5 contabilizou-se o salário de dois professores durante 24 meses, relembrando que os cálculos consideram a hipótese da contratação de professores para o atendimento de escolas que contenham, no máximo, 60 alunos matriculados no 9º ano.

A segunda medida consistiu em mensurar o tempo indispensável para resolver as questões relacionadas com o respectivo padrão analisado. Antes de executar uma política pública, é possível conjecturar o prazo para que ocorram melhorias. No contexto desta pesquisa, o tempo foi estimado com base na duração dos cursos e atividades propostas, bem como na expectativa de geração dos resultados esperados. Assim, em relação ao padrão P_1, o período para

que um progenitor (e/ou responsável legal) de um aluno entre na alfabetização e conclua o 9º ano é de 56 meses. Logo, após esse tempo, é possível elevar consideravelmente o nível de escolaridade. Para o padrão P_2, considerou-se que, após 6 meses do curso de capacitação, seria possível visualizar o impacto da ação. Assim, o tempo necessário seria de 12 meses. Finalmente, para a ação, envolvendo o padrão P_5, ao ser comparada com as ações dos dois itens anteriores, percebe-se que atividades visando a estimular a frequência regular dos alunos na escola têm impacto em menor tempo. Isso ocorre pois se acredita que é mais simples diminuir a diferença de faltas dos alunos do que modificar a rotina dos pais e a mentalidade dos docentes. Desse modo, pode-se conjecturar que os resultados positivos da implementação dessas atividades poderiam começar a ocorrer após seis meses do seu início.

A terceira medida analisada foi o nível de rejeição dos professores às decisões tomadas baseadas nos padrões analisados. Salienta-se que na análise aqui desenvolvida concentrou-se o foco apenas no nível de rejeição dos professores com respeito às decisões tomadas. Nesse sentido, poderia ser imaginada uma nova medida de domínio que considerasse o nível de rejeição da comunidade com relação às ações desenvolvidas. Especialmente no que tange às ações relacionadas aos padrões P_1 e P_5, que envolvem a participação de pais de alunos e alunos, a voz desses sujeitos também poderia ser ouvida, uma vez que seriam esses os alvos da ação. Contudo, no presente trabalho, a análise limitou-se à avaliação da aceitação por parte apenas dos docentes envolvidos nas ações.

Considerando essas limitações, em relação ao padrão P_1, a ação foi classificada pelos autores do presente artigo como de baixa rejeição. Tal suposição pode ser justificada em virtude de o curso ser ofertado aos pais e não aos docentes. Além disso, os professores contratados serão devidamente remunerados conforme a carga horária de trabalho; portanto, acredita-se na baixa resistência dos docentes. Para o padrão P_2, a ação foi rotulada como de média rejeição, pois há maior resistência do professor quando se sugere sua participação em cursos de capacitação/formação. Essa afirmação pode ser interligada com suas intensificadas jornadas de trabalho. Conforme abordado por Barbosa (2011), a desvalorização do professor brasileiro faz com que os docentes trabalhem em várias escolas ou até mesmo tenham outras profissões para adquirir melhor remuneração. Em consequência, falta-lhes tempo para se comprometerem em outras atividades e investir no seu aprimoramento como educadores. Por fim, para o padrão P_5, a ação para aumentar a frequência regular dos alunos na escola foi classificada como de rejeição muito baixa. Observa-se que há coerência em se ter resistência inferior às demais, uma vez que as

atividades estimulantes têm impacto sobre a rotina dos discentes (e não dos docentes) e os professores contratados para sua elaboração/organização serão devidamente remunerados para exercer essa função.

Como comentários gerais, ao observar a Tabela 5, pode-se notar que a ação relacionada ao padrão P_1, elevar o nível de escolaridade dos pais, foi a que apresentou o maior custo para ser implementada e o maior tempo para gerar resultados; a política para o padrão P_5, atividades para estimular a assiduidade dos alunos na escola, requisitou o menor prazo para acarretar em melhorias e apresentou menor rejeição do que as demais; já a intervenção para o padrão P_2, sensibilizar para o aumento de expectativa dos professores, teve o menor custo, porém apontou maior rejeição por parte dos professores em aceitá-la.

Para padronizar as medidas Custo, Tempo e Rejeição, os valores foram transformados para números entre 0 e 1 e, em seguida, unificados, para se obter um valor que sintetizasse todas as três informações [mais detalhes em Fonseca (2018)]. Para essa unificação, poderiam ser dados pesos distintos para as diferentes medidas. Neste trabalho, como o objetivo é exemplificar a utilização de medidas orientadas pelo domínio, aplicou-se simplesmente uma média aritmética, isto é, a medida do domínio é

$$MD(P_z) = \frac{Custo(P_z) + Tempo(P_z) + Rejeição(P_z)}{3},$$

Após os cálculos, obteve-se $MD(P_1) = 0{,}75$, $MD(P_2) = 0{,}27$ e $MD(P_5) = 0{,}32$. Tais valores expressam que, em termos de custo, tempo e rejeição, o padrão tem ações mais viáveis de serem implementadas e promissoras de êxito, seguido, em termos de viabilidade, por ações relacionadas ao padrão P_5 e, por fim, por intervenções para os problemas relacionados a P_1.

Trade-off entre medidas técnicas e do domínio

Diferentemente da mineração de dados tradicional, em D³M busca-se efetuar o *trade-off* entre medidas técnicas e medidas do domínio. Conforme discutido, a medida técnica, responsável por ordenar os padrões, é a porcentagem mínima de alunos da classe Nota Alta que se encontram nas duas categorias que representam o melhor cenário. Por intermédio do algoritmo Naïve Bayes, foi obtido que $MT(P_1) = 0{,}66$, $MT(P_2) = 0{,}63$ e $MT(P_5) = 0{,}58$.

É importante ressaltar que as medidas têm sentido oposto, ou seja, quanto maior o valor da medida técnica, melhores são os resultados; em contrapartida,

a medida do domínio estabelece que a ação seja mais viável quanto menor for o valor associado ao padrão.

Para efetuar o *trade-off*, podem-se estabelecer pesos distintos para definir a contribuição da medida técnica e do domínio. No entanto, nesta pesquisa, foi efetuada a média aritmética. Em termos matemáticos:

$$Medida\ final(P_z) = \frac{MT(P_z) + (1 - MD(P_z))}{2}$$

O fato de as medidas técnica e de domínio terem sentido oposto justifica o termo 1 – $MD(P_z)$ presente na equação anterior. Os valores das medidas, referentes a cada padrão, podem ser visualizados na Tabela 6.

TABELA 6 – Valores das medidas dos padrões

PADRÃO	CONSTRUTO	MEDIDA TÉCNICA ($MD(P_z)$)	MEDIDA DO DOMÍNIO AJUSTADA (1 – $MD(P_z)$)	MEDIDA FINAL
P_1	Escolaridade dos pais	0,66	0,25	0,46
P_2	Expectativa sobre formação dos alunos	0,63	0,73	0,68
P_5	Frequência e fluxo escolar	0,58	0,68	0,63

Fonte: Fonseca (2018).

Nota-se que, ao se analisar somente a medida técnica, o padrão relacionado à escolaridade dos pais teria prioridade sobre os demais, seguido do padrão referente à expectativa dos professores e, por fim, do padrão sobre a frequência dos discentes. No entanto, conforme mencionado, ao considerar ações para resolução dos problemas sob a visão do custo, do tempo para obtenção de resultados e da resistência do professor, isto é, levando-se em consideração aspectos cruciais à temática e presentes no dia a dia do tomador de decisões, há uma nova ordem de relevância. Pensando em ações *versus* viabilidade para a solução de questões educacionais, buscou-se nessa prova de conceito considerar os dois aspectos, cujo resultado pode ser observado por meio da "Medida Final". Observa-se que a intervenção considerada mais promissora foi o oferecimento de uma capacitação aos professores, com o intuito de sensibilizá-los no sentido de aumentar suas expectativas em relação ao potencial de seus alunos e à possibilidade de continuidade nos estudos com sucesso. Ficou avaliado, em segundo lugar na viabilidade, elaborar atividades para os discentes de modo a gerar maior regularidade na frequência à escola e, por fim, restou implementar cursos para elevar o nível de escolaridade dos pais.

Aspectos críticos

Inúmeros estudos apontam que, quanto maior o nível de escolaridade dos pais, melhor o desempenho de seus filhos (JESUS; LAROS, 2004; ORTIGÃO; AGUIAR, 2013; ORTIGÃO; FRANCO; CARVALHO, 2007). Notoriamente, as consequências extrapolam a sala de aula e o êxito acadêmico, acarretando, futuramente, sucesso profissional. Logo, diferenças no nível educacional das famílias são determinantes para a persistência da desigualdade de rendimentos (REIS; RAMOS, 2011).

Diante do exposto, é evidente a necessidade de investimento na educação de jovens e adultos. A realidade é que 13 milhões de brasileiros com mais de 15 anos ainda não sabem ler ou escrever, conforme aponta a Pesquisa Nacional por Amostra de Domicílios (Pnad) mais recente, realizada em 2014 (BRASIL, 2016). Em contrapartida, há uma discussão sobre a complexidade em se mensurar o impacto econômico da alfabetização, o que torna questionável o alto investimento em programas destinados a jovens e adultos.

Devido ao fato de os gestores perceberem tais dificuldades, é crescente a discussão de como a escola brasileira pode compensar a baixa escolaridade dos pais. Nesse sentido, soluções como capacitação do corpo docente e ampliação dos alunos em atividades de aprendizagem são frequentes nos debates acerca desse assunto. Apesar dessas compensações, os investimentos em educação de jovens e adultos não podem ser desencorajados. Os autores deste artigo compartilham a ideia de que devem ser considerados aspectos além do econômico, tratando-se, mais do que disso, de uma questão de justiça social.

Outro resultado constatado nesta pesquisa foi a importância de sensibilizar os professores para o chamado efeito Rosenthal. Diversas outras pesquisas também corroboram a relação entre expectativa dos docentes e desempenho dos alunos (BARBOSA; RANDALL, 2004; FONSECA; NAMEN, 2016; ROSENTHAL; JACOBSON, 1966). Conforme discutido nessas referências, muitas são as razões para a baixa crença dos professores em seus alunos, como a própria desmotivação com a profissão, decorrente dos baixos salários e altas jornadas de trabalho, da falta de infraestrutura e recursos da escola, de uma gestão escolar autoritária, da falta de interesse da turma em relação aos temas abordados, da indisciplina e alto índice de absenteísmo, bem como do contexto sociocultural em que os alunos estão inseridos.

Enfatizando esse último aspecto, um importante estudo efetuado por Damian (2006), ao entrevistar professores de escolas da periferia da cidade de Pelotas, Rio Grande do Sul, constatou que os docentes, atuantes em comunidades com problemas sociais, como a pobreza e as adversidades de ordem

emocional dela decorrentes, esperam desempenho inferior dos seus alunos. A autora relata que, nessas condições, a baixa expectativa dos professores era transmitida aos discentes por meio do discurso da educação compensatória (suprir as carências físicas, afetivas, intelectuais e escolares), no qual os conteúdos acadêmicos são colocados em segundo plano. Em contrapartida, professores que lecionam para alunos que não estão nessas condições esperam que esses sejam bem-sucedidos. Essa expectativa positiva era repassada aos discentes por intermédio do discurso instrucional. Consequentemente, os estudantes nesses moldes atingem melhores resultados e alcançam níveis mais altos no sistema educacional.

Portanto, apesar de se apontar um processo de capacitação/formação para os docentes como uma ação com melhor relação custo-benefício do que as demais, é importante salientar que a expectativa é decorrente de uma série de fatores extraescolares, configurando-se um grande desafio torná-la positiva. Ademais, ressalta-se que há limitações na assunção, feita pelos autores, de que haveria média resistência dos professores à capacitação. Dependendo do perfil da escola, de seu diretor e do próprio docente, essa resistência poderia ser bem mais acentuada, devido à questão da disponibilidade de tempo ou mesmo de motivação do professor para participar da capacitação.

Por fim, a ação visando à frequência regular dos alunos na escola foi superior à intervenção referente ao aumento da escolaridade dos pais, mas inferior à ação para elevar a crença dos professores em relação à formação futura de seus alunos. Discussões acerca do absenteísmo dos discentes são frequentes na literatura, como pode ser visto na pesquisa efetuada por Vasconcellos e Mattos (2011). Os autores relatam que mecanismos de controle não têm assegurado a presença dos alunos, sendo, portanto, um problema difícil de ser enfrentado.

O trabalho aqui desenvolvido propôs uma ação com o intuito de tornar o currículo mais atrativo, elaborando atividades que estimulassem a frequência dos estudantes. Contudo, é importante ressaltar que essa ação deveria ser realizada em conjunto com outras intervenções, tais como o estabelecimento de parceria família-escola na mediação de conflitos e no estímulo à presença na escola, bem como no incremento do diálogo entre professores e discentes para entender as possíveis causas das frequentes faltas.

Apesar desses aspectos críticos, a mineração de dados guiada pelo conhecimento do domínio educacional aqui conduzida permitiu a extração de três padrões, de fato, úteis, ou seja, capazes de apoiar a resolução dos problemas por meio de ações. As políticas sugeridas poderiam contribuir para a mudança do *status* crítico, no qual cerca de 90% dos estudantes do 9º ano não são

proficientes em matemática, em direção a um melhor *status*. Acredita-se que a abordagem aplicada, que agrega métricas estatísticas com métricas orientadas pelo domínio educacional, construídas com o apoio de especialistas nessa temática, possibilita a obtenção de melhores resultados.

CONSIDERAÇÕES FINAIS

Com auxílio de especialistas da área de educação, foi possível formular ações, com base nos padrões descobertos com a mineração de dados da Prova Brasil, as quais podem nortear melhorias e elevar a proficiência dos discentes. Tais ações foram avaliadas por intermédio da elaboração de três medidas: custo requerido, tempo indispensável para resolver as questões relacionadas e nível de rejeição dos professores com respeito às decisões tomadas. Conforme mencionado, outras medidas poderiam ser consideradas, porém, como prova de conceito, acredita-se que essas permitem fomentar a importância de se envolver o conhecimento do domínio no processo de mineração de dados.

Os resultados obtidos no estudo, contudo, devem ser analisados com parcimônia e cautela. Eles não permitem afirmar o que "faz diferença" nas escolas ou o que impulsiona sua qualidade. Não há dúvida de que qualidade é um conceito polissêmico e que a melhoria da aprendizagem escolar, em especial em relação à matemática, não pode ser obtida meramente pela criação de atividades extracurriculares ou por uma única ação de capacitação docente. Há muito mais fatores na escola a serem compreendidos e repensados, dentre os quais destacam-se as condições de infraestrutura escolar (física, pedagógica, humana, profissional, cultural), as condições de formação de professores e de gestores e os processos de responsabilização decorrentes das atuais avaliações externas.

Em um contexto mais geral, reduzir a melhoria da escola à atuação do professor ou ao estabelecimento de um currículo único, nacional, para todas as unidades escolares dificilmente resolve a questão. Ao contrário, alguns poderiam afirmar que são tentativas de produzir sentido às políticas educacionais, suprimindo do debate a responsabilidade do poder público em manter em condições dignas a infraestrutura das escolas.

O esforço para a aquisição das ferramentas empregadas na pesquisa e o rigor metodológico que as acompanha foram decisivos para permitir uma análise relevante dos dados apresentados. O estudo permitiu a identificação de padrões relacionados com o processo de aprendizagem ao abordar o uso combinado de métodos para redução de dimensionalidade e algoritmos de mi-

neração. Além disso, foram apresentadas as dificuldades em inserir medidas do domínio e a necessidade em superá-las para garantir resultados práticos.

No entanto, é desejável que haja um desdobramento deste trabalho no sentido de complementá-lo com outros estudos que possam conjugar diferentes abordagens de pesquisa, o que implica a inclusão de resultados de outros estados e a apropriação de aportes teóricos que possibilitam outras análises. Ademais, conforme mencionado, existe a possibilidade de inclusão e avaliação de novas medidas do domínio sobre os padrões estudados. Mais ainda, há a perspectiva de aprofundamento das análises das medidas do domínio relacionadas ao padrão referente ao funcionamento da escola e do *trade-off* com as respectivas medidas técnicas.

Sugere-se, também, a aplicação de outros algoritmos de mineração e, ainda, utilizar essa metodologia nas bases de dados de anos anteriores, já que o Inep disponibiliza dados da Prova Brasil desde 2005. Logo, poderia ser efetuada uma análise temporal para comparar as informações obtidas. Ademais, poderiam ser realizados estudos de sensibilidade para verificar a estabilidade dos resultados obtidos, novas medidas do domínio poderiam ser formuladas e a resistência às ações poderia ser avaliada não somente sob a perspectiva do professor.

Independentemente das possibilidades de futuros trabalhos, a metodologia desenvolvida nesta pesquisa e a prova de conceito, relacionadas ao processo de aprendizagem de matemática, podem apoiar outros estudos com dados educacionais e em áreas sociais. Espera-se, dessa forma, estimular a utilização da mineração de dados orientada pelo domínio, de modo a obter resultados relevantes e que fomentem e enriqueçam a discussão entre os tomadores de decisões para solução de problemas.

AGRADECIMENTOS

O presente trabalho foi realizado com o apoio financeiro da Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (Faperj).

REFERÊNCIAS

ADEJUWON, A.; MOSAVI, A. Domain driven data mining. *IJCSI International Journal of Computer Science Issues*, Mahebourg, Republic of Mauritius, v. 7, n. 2, p. 41-44, July 2010.

BARBOSA, A. *Os salários dos professores brasileiros:* implicações para o trabalho docente. 2011. 208 f. Tese (Doutorado em Educação Escolar) – Universidade Estadual Paulista "Júlio de Mesquita Filho", Faculdade de Ciências e Letras de Araraquara, Araraquara, SP, 2011.

BARBOSA, M.; RANDALL, L. Desigualdades sociais e a formação de expectativas familiares e de professores. *Caderno CRH*, Salvador, v. 17, n. 41, p. 299-308, maio/ago. 2004.

BAUER, A. É possível relacionar avaliação discente e formação de professores? A experiência de São Paulo. *Educação em Revista*, Belo Horizonte, v. 28, n. 2, p. 61-82, jun. 2012.

BAUER, A.; GATTI, B. A. (org.). *Vinte cinco anos de avaliação de sistemas educacionais no Brasil*: origens nas redes de ensino, no currículo e na formação de professores. Florianópolis: Insular, 2013. v. 2.

BAUER, A.; GATTI, B. A.; TAVARES, M. R. (org.). *Vinte e cinco anos de avaliação de sistemas educacionais no Brasil*: origens e pressupostos. Florianópolis: Insular, 2013. v. 1.

BEZERRA, C.; SCHOLZ, R; ADEODATO, P; PONTES, R; SILVA, I. Evasão escolar: aplicando mineração de dados para identificar variáveis relevantes. *In*: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 2016, Uberlândia, MG. *Anais* [...]. Uberlândia: Sociedade Brasileira de Computação, 2016. p. 1096-1105.

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. *SAEB 2001:* novas perspectivas. Brasília: Inep, 2001.

BRASIL. *Programa Brasil Alfabetizado*. Brasília: Ministério da Educação, 2011. Disponível em: http://portal.mec.gov.br/programa-brasil-alfabetizado. Acesso em: 1 ago. 2017.

BRASIL. Ministério da Educação. Secretaria de Educação Básica. Diretoria de Currículos e Educação Integral. *Diretrizes Curriculares Nacionais Gerais da Educação Básica*. Brasília: MEC, 2013. Disponível em: http://portal.mec.gov.br/docman/julho-2013-pdf/13677-diretrizes-educacao-basica-2013-pdf/file. Acesso em: 1 ago. 2017.

BRASIL. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira. *Microdados da Aneb e da Anresc 2013*. Brasília: Inep, 2015. Disponível em: http://portal.inep.gov.br/microdados#. Acesso em: 2 set. 2019.

BRASIL. Ministério da Educação. *Educação de jovens e adultos é prioridade para o governo*. Brasília: MEC, 2016. Disponível em: http://portal.mec.gov.br/busca-geral/204-noticias/10899842/41601-educacao-de-jovens-e-adultos-e-prioridade-para-o-governo. Acesso em: 15 ago. 2017.

CAO, L. Domain-driven data mining: challenges and prospects. *IEEE Transactions on Knowledge and Data Engineering*, Piscataway, v. 22, n. 6, p. 755-769, June 2010.

CAO, L.; YU, P. S.; ZHANG, C.; ZHAO, Y. *Domain driven data mining*. New York: Springer, 2010.

CAO, L.; ZHANG, C. Domain-driven data mining: a practical methodology. *International Journal of Data Warehousing & Mining*, Sydney, v. 2, n. 4, p. 49-65, Jan. 2005.

CLARK, A. C.; WATSON, D. Constructing validity: basic issues in objective scale development. *Psychological Assessment*, Washington, v. 7, n. 3, p. 309-319, Sept. 1995.

COLLEMAN, J. S. *Equality of education opportunity*. Washington: National Center for Educational Statistics, 1966.

CRONBACH, L. Coefficient alpha and the internal structure of tests. *Psychometrika*, Madison, v. 16, n. 3, p. 297-334, Sept. 1951.

DAMIAN, M. F. Discurso pedagógico e fracasso escolar. *Ensaio*: Avaliação e Políticas Públicas em Educação, Rio de Janeiro, v. 14, n. 53, p. 457-478, out./dez. 2006.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *AI Magazine*, Anaheim, v. 17, n. 3, p. 37-54, Mar. 1996.

FERNANDES, E.; HOLANDA, M.; VICTORINO, M.; BORGES, V.; CARVALHO, R.; ERVEN, G.V. Educational data mining: predictive analysis of academic performance of public school students in the capital of Brazil. *Journal of Business Research*, v. 94, p. 335-343, Feb. 2018. https://doi.org/10.1016/j.jbusres.2018.02.012.

FONSECA, S. O. *Uma metodologia de mineração de dados orientada pelo domínio para a descoberta de conhecimento sobre o processo de aprendizagem no ensino básico.* 2018. 200 f. Tese (Doutorado em Modelagem Computacional) – Instituto Politécnico, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2018.

FONSECA, S. O.; NAMEN, A. A. Mineração em bases de dados do Inep: uma análise exploratória para nortear melhorias no sistema educacional brasileiro. *Educação em Revista*, Belo Horizonte, v. 32, n. 1, p. 133-157, mar. 2016.

FRANCO, C.; FERNANDES, C.; SOARES, J. F.; BELTRÃO, K.; BARBOSA, M. E.; ALVES, M. T. G. O referencial teórico na construção dos questionários contextuais do Saeb 2001. *Estudos em Avaliação Educacional*, São Paulo, n. 28, p. 39-74, jul./dez. 2003.

GOMES, J. C.; LEVY, A.; LACHTERMACHER, G. Segmentação do censo educacional 2000 utilizando técnicas de mineração de dados. O Impacto da pesquisa operacional nas novas tendências multidisciplinares. *In*: SIMPÓSIO BRASILEIRO DE PESQUISA OPERACIONAL, 36., 2004, São João del-Rei, MG. *Anais* [...]. São João del-Rei, MG: Sociedade Brasileira de Pesquisa Operacional, 2004. p. 820-831.

HAIR, J. F.; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C. *Análise multivariada de dados*. 5. ed. Porto Alegre: Bookman, 2005.

HAN, J.; KAMBER, M.; PEI, J. *Data mining:* concepts and techniques. 3. ed. Waltham: Morgan Kaufmann, 2012.

IBM CORPORATION. *IBM SPSS Statistics*. Versão 23. New York: IBM Corporation, 1989.

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO RIO DE JANEIRO – IFRJ. *Edital n. 44/2016* – Concurso Público – Professor de Ensino Básico, Técnico e Tecnológico. Rio de Janeiro: IFRJ, 2016. Disponível em: https://migra.ifrj.edu.br/webfm_send/10632. Acesso em: 2 set. 2019.

JESUS, G. R.; LAROS, J. A. Eficácia escolar: regressão multinível com dados de avaliação em larga escala. *Avaliação Psicológica*, v. 3, n. 2, p. 93-106, nov. 2004.

JOHN, G. H.; LANGLEY, P. Estimating continuous distributions in Bayesian classifiers. *In*: CONFERENCE ON UNCERTAINTY IN ARTIFICIAL INTELLIGENCE, 11., 1995, Montreal. *Anais* [...]. San Francisco: Morgan Kaufmann, 1995. p. 338-345.

KAMPFF, A. J. C.; REATEGUI, E. B.; LIMA, J. V. de. Mineração de dados educacionais para a construção de alertas em ambientes virtuais de aprendizagem como apoio à prática docente. *Novas Tecnologias na Educação*, Porto Alegre, v. 6, n. 2, dez. 2008.

KARINO, C. A.; VINHA, L. G. A.; LAROS, J. A. Os questionários do SAEB: o que eles realmente medem? *Estudos em Avaliação Educacional*, São Paulo, v. 25, n. 59, p. 270-297, set./dez. 2014.

KLEIN, R. Utilização da Teoria de Resposta ao Item no Sistema Nacional de Avaliação da Educação Básica (SAEB). *Ensaio:* Avaliação e Políticas Públicas em Educação, Rio de Janeiro, v. 11, n. 40, p. 283-296, 2003.

KLINE, P. *The handbook of psychological testing*. Londres: Routledge, 1999.

ORTIGÃO, M. I. R.; AGUIAR, G. S. Repetência escolar nos anos iniciais do ensino fundamental: evidências a partir dos dados da Prova Brasil 2009. *Revista Brasileira de Estudos Pedagógicos*, Brasília, v. 94, n. 237, p. 364-389, ago. 2013.

ORTIGÃO, M. I. R.; FRANCO, C.; CARVALHO, J. B. P. de. A distribuição social do currículo de matemática: quem tem acesso a tratamento da informação? *Educação Matemática Pesquisa*, São Paulo, v. 9, n. 2, p. 249-273, dez. 2007.

PIETY, P. J. *Assessing the educational data movement*. New York: Teachers College, 2013.

POSTGRESQL GLOBAL DEVELOPMENT GROUP. *PostgreSQL Database Management System.* Versão 9.2. Berkeley: PostgreSQL Global Development Group, c1995. Disponível em: https://www.postgresql.org. Acesso em: 1 jan. 2015.

REIS, M. C.; RAMOS, L. Escolaridade dos pais, desempenho no mercado de trabalho e desigualdade de rendimentos. *Revista Brasileira de Economia*, Rio de Janeiro, v. 65, n. 2, p. 177-205, abr./jun. 2011.

ROBINSON, J. P.; SHAVER, P. R.; WRIGHTSMAN, L. S. Criteria for scale selection and evaluation. *In*: ROBINSON, J. P.; SHAVER, P. R.; WRIGHTSMAN, L. S (ed.). *Measures of personality and social psychological attitudes*. San Diego: Academic Press, 1991. p. 1-16.

RODRIGUES, R. L.; RAMOS, J. L. C.; SILVA, J. C. S.; GOMES, A. S. A literatura brasileira sobre mineração de dados educacionais. *In*: CONGRESSO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 2014, Dourados, MS. *Anais* [...]. Dourados, MS: Sociedade Brasileira de Computação, 2014. p. 621-630.

ROMERO, C.; VENTURA, S. Data mining in education. *Wiley interdisciplinary reviews*: data mining and knowledge discovery, New York, v. 3, n. 1, p. 12-27, Dec. 2013.

ROSENTHAL, R.; JACOBSON, L. Teachers' expectancies: determinates of pupils' I. Q. gains. *Psychological Reports*, Boston, v. 19, n. 1, p. 115-118, Aug. 1966.

SMITH, L. *A tutorial on principal components analysis*. New York: Cornell University, 2002.

SOARES, J. F. Melhoria do desempenho cognitivo dos alunos do ensino fundamental. *Cadernos de Pesquisa*, São Paulo, v. 37, n. 130, p. 135-160, jan./abr. 2007. Disponível em: http://www.scielo.br/pdf/cp/v37n130/07.pdf . Acesso em: 10 dez. 2017.

STEVENS, J. P. *Applied multivariate statistics for the social sciences*. 2. ed. Hillsdale: Erlbaum, 1992.

TCHIBOZO, G. Applications in data analysis for educational research. *Policy Futures in Education*, Paris, v. 7, n. 4, p. 364-367, Jan. 2009.

UNIVERSITY OF WAIKATO. *Weka*. Versão 3.6.9. Waikato: University of Waikato, 1999. Disponível em: https://www.cs.waikato.ac.nz/ml/weka/. Acesso em: 2 set. 2019.

VASCONCELLOS, S. S. de; MATTOS, C. L. G. de. O absenteísmo escolar e sua regulamentação. *In*: MATTOS, C. L. G.; CASTRO, P. A. *Etnografia e educação*: conceitos e usos. Campina Grande: EDUEPB, 2011. p. 271-294.

VELOSO, F.; PESSOA, S.; HENRIQUES, R.; GIAMBIAGI, F. (org.). *Educação básica no Brasil*: construindo o país do futuro. Rio de Janeiro: Elsevier, 2009.

WITTEN, I. H.; FRANK, E.; HALL, M. A. *Data mining:* practical machine learning tools and techniques. 3. ed. San Francisco: Morgan Kaufmann, 2011.

NOTA: O artigo é resultante de pesquisa relacionada à tese de doutorado *Uma metodologia de mineração de dados orientada pelo domínio para a descoberta de conhecimento sobre o processo de aprendizagem no ensino básico*, de autoria de Stella Oggioni da Fonseca e orientada pelos professores Anderson Amendoeira Namen e Francisco Duarte Moura Neto.

Os autores Stella Oggioni da Fonseca e Anderson Amendoeira Namen participaram de todas as etapas do trabalho, compreendendo desde a elaboração do projeto até a redação final do presente artigo. O autor Francisco Duarte Moura Neto participou do processo de redução de dimensionalidade, análise dos dados e interpretação dos resultados obtidos. As autoras Adriana da Rocha Silva, Maria Isabel Ramalho Ortigão e Ursula Andrea Rohrer atuaram, principalmente, na análise das medidas relacionadas à temática educacional. Todos os autores participaram da elaboração e revisão deste artigo.

Recebido em: 25 JUNHO 2018
Aprovado para publicação em: 18 MARÇO 2019