EMP Educação Matemática Pesquisa
ISSN 1983-3156
A2
http://dx.doi.org/10.23925/1983-3156.2018v20i2p257-282

Uma metodologia para a descoberta de conhecimento em bases de dados da Prova Brasil

A methodology for knowledge discovering in Prova Brasil databases

STELLA OGGIONI DA FONSECA [1]
ADRIANA DA ROCHA SILVA [2]
ANDERSON AMENDOEIRA NAMEN [3]

Resumo

A Prova Brasil é uma avaliação que, por intermédio da aplicação de testes e questionários, coleta informações sobre o ensino fundamental. O presente trabalho objetiva apresentar uma metodologia capaz de identificar aspectos, relacionados ao ambiente educacional, que possam ter influenciado positiva ou negativamente no resultado obtido pelos alunos nos testes de Matemática, aplicados em 2013. A abordagem proposta consiste, essencialmente, de um processo de redução de dimensionalidade com posterior aplicação de mineração de dados visando à descoberta de conhecimento nas bases. A partir das conclusões obtidas é possível fomentar a discussão que busquem o alcance de melhorias no processo de ensino-aprendizagem, bem como estimular pesquisas acerca dos dados disponibilizados pelo Governo Federal.

Palavras-chave: *Prova Brasil, Redução de Dimensionalidade, Mineração de Dados.*

Abstract

Prova Brasil is an evaluation that, through the application of tests and questionnaires, collects information about elementary education. The present work aims to present a methodology for the extraction of aspects related to the educational environment that may have influenced positively or negatively students' results in the Mathematics tests applied in 2013. The proposed approach consists of a dimensionality reduction process followed by data mining, aiming to get knowledge discovery in databases. Based on the conclusions obtained, discussions about actions for improvements in the teaching-learning process can be made, as well the fostering of researches on the data provided by the Federal Government.

Keywords: *Prova Brasil, Reduction of Dimensionality, Data Mining.*

[1] Doutora em Modelagem Computacional - Programa de Pós-Graduação em Modelagem Computacional do Instituto Politécnico da Universidade do Estado do Rio de Janeiro - e-mail: stella.oggioni@gmail.com.
[2] Mestranda em Modelagem Computacional - Programa de Pós-Graduação em Modelagem Computacional do Instituto Politécnico da Universidade do Estado do Rio de Janeiro - Professora do Centro de Educação a Distância do Estado do Rio de Janeiro (Cecierj) - e-mail: arsilva@iprj.uerj.br.
[3] Doutor em Engenharia de Sistemas e Computação pela COPPE/UFRJ - Professor do Programa de Pós-Graduação em Modelagem Computacional do Instituto Politécnico da Universidade do Estado do Rio de Janeiro e da Universidade Veiga de Almeida - e-mail: anamen@uva.br.

Introdução

O sistema educacional brasileiro tem se caracterizado como um dos divisores da sociedade em grupos econômicos distintos, mantendo e reforçando a desigualdade e os problemas sociais (GUZZO; EUZEBIOS FILHO, 2005). Para mudar este cenário, algumas iniciativas vêm buscando ampliar os mecanismos de avaliação, de modo a alicerçar medidas que possam melhorar a qualidade e a equidade do ensino.

Segundo Azevedo (2001), no contexto da educação básica, constata-se que desde 1930 já havia o interesse, por parte do Estado, em avaliar o setor educacional. No entanto, somente no final dos anos 80 é que, de fato, iniciou-se um planejamento para que a avaliação externa, em larga escala e dando ênfase no rendimento do aluno, se tornasse uma prática sistêmica, passando a integrar ações governamentais (FREITAS, 2007).

No ano de 1990 foi criado o Sistema Nacional de Avaliação da Educação Básica (Saeb). O Saeb é composto por um conjunto de instrumentos que têm como intuito realizar um diagnóstico da educação básica e fornecer indicadores sobre a qualidade do ensino oferecido aos estudantes. De modo geral, as avaliações realizadas pelo Saeb são compostas por testes, que visam mensurar o aprendizado dos conteúdos propostos, e por questionários, aplicados junto à comunidade escolar, que procuram avaliar os fatores que afetam a qualidade do ensino dos discentes.

Conforme amplamente discutido em Coelho (2008), as informações coletadas pelo Saeb subsidiaram diversas pesquisas na área de educação, tornando possível a investigação das características escolares, fatores socioeconômicos e dos níveis de aprendizagem dos alunos. Nesse sentido, inúmeros estudos têm evidenciado que as escolas são distintas, não somente quanto ao perfil dos estudantes, mas também quanto aos seus aspectos internos (SOARES, 2004; FRANCO; SZTAJN; ORTIGÃO, 2007; RODRIGUES; GUIMARÃES; RIOS-NETO, 2011).

Diante da possibilidade de extração de conhecimento acerca das relações entre ambiente educacional e desempenho dos estudantes, acredita-se que a contribuição desta pesquisa se dá por intermédio da proposição de uma metodologia para descoberta de conhecimento relevante nas bases de dados que armazenam informações sobre a educação básica. Foram analisados os dados oriundos da avaliação denominada Prova Brasil, aplicada ao ensino fundamental nas escolas públicas. Objetivou-se identificar os aspectos que pudessem ter influenciado positiva ou negativamente os resultados dos alunos nos testes de matemática.

Dentre as diversas tarefas desenvolvidas no trabalho, destaca-se a aplicação de um processo de redução de dimensionalidade, motivado pelo alto número de perguntas presentes nos questionários da Prova Brasil. A abordagem proposta no presente trabalho, combinou esta redução de dimensionalidade com o processo de mineração de dados (*Data Mining*). Este último visou identificar relações e padrões relevantes relacionados à aprendizagem dos estudantes e consistiu na aplicação de algoritmos baseados em conceitos estatísticos e inteligência computacional (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996). Salienta-se que inúmeros pesquisadores têm mostrado interesse em utilizar mineração de dados para investigar perguntas científicas na área educacional (BAKER; ISOTANI; CARVALHO, 2011).

Prova Brasil

A Avaliação Nacional do Rendimento Escolar (Anresc), amplamente conhecida como Prova Brasil, avalia alunos de Ensino Fundamental dos 5º e 9º anos de todas as escolas públicas que possuem, no mínimo, 20 alunos matriculados nos anos mencionados. Esta avaliação foi efetuada pela primeira vez em 2005 e vem ocorrendo a cada dois anos, sendo a quinta edição, realizada no ano de 2013, a base utilizada neste trabalho.

Em 2013, a Prova Brasil foi composta da aplicação de testes de Matemática e Língua Portuguesa, além de quatro questionários contextuais que visavam coletar informações a respeito dos alunos, professores, diretores e escolas. Estes questionários englobavam questões sobre gestão escolar, práticas pedagógicas, qualificações profissionais, infraestrutura escolar e perfil socioeconômico. Cabe ressaltar que as bases de dados que armazenam essas informações constituem um importante acervo que pode ser usado para descoberta de conhecimento por educadores, pesquisadores e elaboradores de políticas públicas.

Muitos estudos têm explorado as bases de dados da Prova Brasil. Para exemplificar, Pereira e Mori (2011), ao analisarem a proposta curricular e o desempenho dos alunos do 9º ano das escolas paranaenses na Prova Brasil, do ano de 2007, dando enfoque às respostas dadas ao questionário dos professores e dos alunos, puderam concluir que a qualidade do ensino não está somente vinculada ao currículo. É fundamental a mediação do professor na organização de atividades que despertem o interesse dos discentes. Ademais, as autoras reconhecem que os docentes não devem ser responsabilizados pelo

fracasso escolar, uma vez que existem outros fatores associados, como a distribuição desigual e injusta de recursos materiais e simbólicos.

Ortigão e Aguiar (2013) utilizaram os dados da Prova Brasil 2009 referentes aos alunos de 5º ano do ensino fundamental para discutir a reprovação no âmbito da escola pública brasileira. Por intermédio da programação de um modelo de regressão logística, puderam investigar aspectos relacionados aos alunos e familiares que exerceriam influência na repetência escolar. Neste estudo, concluíram que os meninos, quando comparados com as meninas, são mais propensos à repetição em Matemática. Além disso, cursar a pré-escola, fazer os deveres de casa e ter apoio da família nos estudos são importantes fatores associados à diminuição do risco de repetência.

Santos e Tolentino-Neto (2015), por sua vez, analisaram as notas médias em Matemática obtidas pelos alunos dos 5º e 9º anos do Ensino Fundamental e 3ª série do Ensino Médio nas edições de 2005 a 2013 do Saeb. O estudo deu enfoque aos alunos residentes no Estado do Rio Grande do Sul. Por meio desta análise, os autores puderam concluir que os resultados acerca dos discentes do 5º ano revelam que algumas habilidades notadamente não estão bem desenvolvidas, como é o caso das operações de multiplicação e divisão. Já sobre os alunos do 9º ano, constatou-se que a maioria dos estudantes não resolve equações de 1º e 2º grau com uma incógnita e não soluciona problemas envolvendo relações métricas do triângulo retângulo.

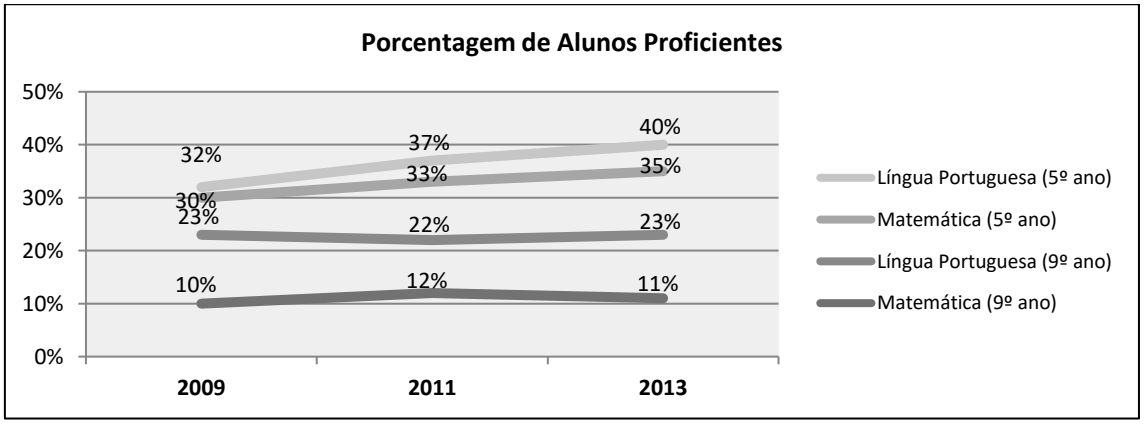
Fonseca e Namen (2016) aplicaram o processo de Descoberta de Conhecimento em Bancos de Dados (*Knowledge Discovery in Databases* – KDD) com o intuito de identificar fatores que relacionavam o perfil de professores que lecionam Matemática com a proficiência obtida por seus alunos. Nesta pesquisa, utilizaram as respostas dos docentes aos questionários, bem como o resultado obtido no teste de Matemática pelos estudantes do 9º ano do ensino fundamental. Os dados utilizados foram provenientes da Prova Brasil do ano de 2011. Por meio da mineração de dados, puderam ser observados alguns fatores favoráveis ao aprendizado de matemática, como alto percentual de conteúdo previsto desenvolvido, assiduidade dos discentes na escola e a expectativa, por parte dos docentes, de que muitos dos seus alunos teriam uma boa formação educacional. Também foram analisados aspectos que tendiam a influenciar negativamente o desempenho dos estudantes, como a desvalorização salarial da profissão de educador e a crença do professor de que poucos alunos entrariam em uma universidade.

A ampliação dos objetivos deste último trabalho citado culminou na realização do presente artigo. O estudo passou então a considerar não somente os questionários

respondidos pelos docentes, mas também os preenchidos pelos alunos, diretores e pelos aplicadores da Prova Brasil (ano de 2013), agentes externos às escolas, responsáveis por coletar informações acerca da infraestrutura escolar. Logo, o volume de dados tornou-se muito maior do que o analisado em Fonseca e Namen (2016), gerando desafios relacionados à capacidade de análise. Além disso, o presente trabalho, de caráter interdisciplinar, foi feito com o auxílio de especialistas da área educacional, o que possibilitou melhor interpretação dos resultados obtidos.

Salienta-se que o presente estudo foi desenvolvido analisando-se as relações entre os dados dos questionários mencionados e a proficiência obtida no teste de Matemática pelos estudantes do 9º ano do ensino fundamental. A escolha da disciplina e série mencionadas ocorreu motivada pela análise dos resultados apresentados na Figura 1.

Figura 1 - Porcentagem de alunos que aprenderam o adequado por disciplina e série avaliada



Nota: Dados da Fundação Lemann e Meritt (2015).
Fonte: Elaboração dos autores.

A Figura 1 apresenta a porcentagem de alunos, em âmbito nacional, que aprenderam o adequado em cada disciplina e série avaliada pela Prova Brasil nos anos de 2009, 2011 e 2013. Observe que o cenário mais crítico relaciona-se ao desempenho em Matemática dos alunos do 9º ano. Segundo os dados apresentados pela Fundação Lemann e Meritt (2015), em 2009 somente 10% dos alunos do 9º ano que fizeram o teste de Matemática obtiveram uma pontuação[4] que indicava que estavam preparados para continuar os estudos. Em 2011 a porcentagem passou a ser 12% e em 2013 decaiu para 11%. Logo, é notório que ao longo dos anos os resultados estão estagnados em uma situação muito aquém da esperada.

[4] Esta pontuação foi definida pelo pesquisador Francisco Soares (Fundação Lemann e Meritt (2015)) com base na escala do Saeb.

Visão geral do processo

Para cumprir o objetivo apresentado no escopo deste trabalho, foi feito o *download* das bases de dados da Prova Brasil do ano de 2013. Tais informações, de acesso público, são armazenadas em diferentes arquivos que podem ser encontrados no *site* do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep)[5]. Os arquivos, referentes a este ano, estão intitulados como apresentado no Quadro 1, onde encontra-se também uma breve descrição a respeito de cada um.

Quadro 1 - Arquivos da Prova Brasil 2013

Arquivo	Descrição	Nº. de registros	Nº. de atributos
TS_ALUNO_9EF	Dados das respostas ao questionário do aluno e das proficiências dos alunos do 9º ano do ensino fundamental	2720588	92
TS_PROFESSOR	Dados das respostas ao questionário aplicado ao Professor de cada disciplina de cada série	237186	134
TS_DIRETOR	Dados das respostas ao questionário aplicado ao Diretor de cada Escola	56737	118
TS_ESCOLA	Média da proficiência dos alunos por disciplina e dados das respostas ao questionário aplicado à Escola	59251	127

Fonte: Adaptado de Inep (2015).

O arquivo TS_ALUNO_9EF contém 2720588 registros correspondentes, em âmbito nacional, aos alunos do 9º ano do ensino fundamental, com 92 campos que compreendem diversos atributos identificadores, a proficiência obtida nas disciplinas de Matemática e Língua Portuguesa, bem como as respostas dadas às 57 perguntas presentes no questionário.

O segundo arquivo, TS_PROFESSOR, armazena informações de professores que lecionam Língua Portuguesa e Matemática para alunos dos 5º e 9º anos do ensino fundamental e 3ª série do ensino médio. Possui 237186 registros e 134 atributos, incluindo 125 referentes às respostas das perguntas do questionário e os restantes, que possibilitam a identificação de cada professor.

O arquivo TS_DIRETOR contém 56737 registros e 118 atributos, sendo que, destes, seis são variáveis identificadoras, um informa se houve ou não o preenchimento do questionário e 111 armazenam as respostas dadas às perguntas.

[5] *Site* do Inep: www.inep.gov.br.

O último arquivo acessado denomina-se TS_ESCOLA. Composto por 59251 registros e 127 atributos, este arquivo armazena dados sobre a identificação da escola, índices calculados sobre os alunos, professores e o resultado obtido nos testes, além de alocar as respostas dadas às 68 questões presentes no questionário.

As bases de dados descritas no Quadro 1 passaram por diversas tarefas de manipulação, como seleção dos dados, transformação e discretização de alguns atributos. Convém destacar a tarefa de redução de dimensionalidade dos dados, com a redução significativa do número original de atributos. Esse processo foi essencial para viabilização das análises, dado o grande volume de dados envolvido. Posteriormente, foi conduzida a etapa de mineração de dados, que aplicou o algoritmo *Naïve Bayes* para a identificação de padrões embutidos nos dados. Finalmente, os resultados gerados foram interpretados. Cada um desses passos será descrito nas seções seguintes.

Seleção dos dados

Para executar as tarefas de seleção dos dados, os arquivos foram importados para o *software* PostgreSQL (POSTGRESQL, 2007), que é um sistema gerenciador de banco de dados. Por meio de Linguagem de Consulta Estruturada (SQL - *Structured Query Language*), foi possível realizar diferentes operações sobre os registros e atributos selecionados.

Inicialmente, apesar da avaliação Prova Brasil ser em âmbito nacional, considerou-se nesta pesquisa a seleção dos alunos, professores, diretores e escolas somente do Estado do Rio de Janeiro. Tal seleção foi motivada por entender-se que cada estado possui peculiaridades inerentes a sua rede pública. Portanto, os resultados que serão obtidos não necessariamente retratam a realidade de outras regiões do país.

Foi necessário, ainda, remover os alunos que não preencheram o teste e os que não se adequavam ao cálculo da proficiência (mínimo de 3 itens respondidos no teste), uma vez que a informação chave é o resultado obtido em Matemática. Por fim, como o intuito era identificar características provenientes do questionário, foram removidos os alunos que não responderam a, no mínimo, 70% das 57 perguntas.

Além da seleção dos estudantes, os docentes, diretores e escolas também foram analisados. Assim como para os alunos, foram removidos todos aqueles que não haviam respondido a, no mínimo, 70% das perguntas presentes nos questionários. Ademais, foram desconsiderados os casos de docentes, diretores ou escolas que não possuíam

alunos registrados na base de dados. Para tal critério, foi necessário interligar as tabelas por meio de atributos identificadores que possibilitam detectar, para cada aluno, a sua escola, diretor e professor. Finalmente, foram selecionados somente os professores que lecionavam Matemática, que é a disciplina de interesse neste estudo.

Ao final do processo de seleção dos dados, cada tabela passou a conter a seguinte quantidade de registros: 113021 alunos em TS_ALUNO_9EF; 2388 professores em TS_PROFESSOR; 1771 diretores em TS_DIRETOR; e 1764 escolas em TS_ESCOLA.

Redução de dimensionalidade

O próximo passo efetuado foi reduzir a dimensionalidade, uma vez que havia um grande número de variáveis (361 atributos) referentes às perguntas presentes em cada questionário.

Procurou-se identificar grupos de questões inter-relacionadas que poderiam estar medindo um tema comum ou, equivalentemente, uma dimensão subjacente. Essas dimensões são chamadas de construtos ou variáveis latentes (FIELD, 2009). Os grupos foram identificados por meio de análise de componentes principais (PCA - *Principal Components Analysis*). Segundo Hair et al. (2005), PCA é uma técnica multivariada que consiste em condensar a informação presente nas variáveis originais em um conjunto menor de variáveis (resultantes de combinações lineares d0s originais), perdendo o mínimo de informação. Sabendo-se que a informação e variabilidade dos dados são capturadas pela matriz de correlações de um conjunto de variáveis, a PCA consiste em explicitar a estrutura desta matriz.

Todo este processo e resultados extraídos foram efetuados por meio do *software* estatístico *Statistical Package for the Social Sciences* (SPSS). Este *software* possui uma série de técnicas de análise de dados, incluindo a PCA.

A Tabela 1 apresenta os construtos extraídos de cada questionário, os números das questões agrupadas, e métricas de avaliação, como o valor do coeficiente α de Cronbach e a média das correlações entre as variáveis (detalhes sobre essas medidas podem ser encontrados em HAIR et al., 2005 e CLARK; WATSON, 1995). Foram dados nomes aos construtos, pelos autores do presente artigo, de modo a permitir o entendimento das temáticas referentes às questões inter-relacionadas.

Tabela 1 - Construtos extraídos dos questionários

Questionário	Construto	Questões agrupadas[6]	α de Cronbach	Média das correlações
Aluno	Posse de bens	5 a 15	0,805	0,280
	Utilização da biblioteca	33, 34, 39, 56	0,741	0,422
	Dever de casa e correções	51, 52, 54 e 55	0,747	0,425
	Incentivo dos pais	27 a 31	0,547	0,247
	Hábito de leitura geral	32, 35 a 37	0,603	0,276
	Escolaridade dos pais	19 e 23	0,653	0,487
Professor	Direção da escola, gestão e liderança	58 a 69	0,935	0,546
	Recursos pedagógicos	41,44 a 47, 49, 50, 110, 123 e 125	0,789	0,293
	Necessidade de aperfeiçoamento profissional	26 a 31	0,891	0,576
	Experiência e salário	5, 10, 12 a 15 e 97	0,783	0,376
	Práticas pedagógicas	112, 113, 120, 121, 122 e 124	0,754	0,403
	Violência na escola	85 a 92	0,706	0,288
	Desenvolvimento profissional	8, 9, 21 a 24	0,773	0,362
Diretor	Projeto na escola	100 a 108	0,797	0,310
	Funcionamento da escola	67 a 71, 73 a 76	0,783	0,292
	Políticas, ações e programas escolares	44 a 48, 50 a 52	0,742	0,271
	Experiência e salário	5, 10, 12, 16 a 18	0,749	0,342
	Nível de escolaridade e atividade de desenvolvimento profissional	4, 8, 9, 19 e 20	0,640	0,307
	Conselho escolar	29 e 30	0,905	0,865
Escola	Infraestrutura	7 a 21, 31 e 36	0,917	0,402
	Recursos de apoio	37 a 45, 47 a 49, 51, 53, 56 e 60	0,857	0,292
	Situação da biblioteca	65, 67 a 74	0,829	0,404
	Segurança da escola	24 a 29	0,847	0,482

Fonte: Elaboração dos autores.

Em termos práticos, um componente apresenta confiabilidade quando α de Cronbach tem um valor superior a 0,7 (KLINE, 1999; HAIR et al., 2005). Além disso, segundo Clark e Watson (1995), a média das correlações entre as variáveis deve ficar entre 0,15 e 0,5.

Observa-se que foram extraídos 23 construtos dos quatro questionários. Dentre eles, nota-se que quatro não apresentaram α de Cronbach superior a 0,7. Contudo, concomitantemente, foi analisada a média das correlações entre as variáveis e, em todos os quatro casos, esta ficou entre a faixa sugerida por Clark e Watson (1995). Portanto, foi decidido manter os 23 construtos para as análises posteriores. Lembrando que, originalmente, a base era composta por um total de 361 atributos. Mais detalhes do

[6] Os enunciados das questões podem ser vistos no *site* do Inep (www.inep.gov.br), onde os questionários são apresentados na íntegra e disponibilizados para *download*.

processo de redução de dimensionalidade aqui conduzido podem ser vistos em Fonseca (2018).

É importante ressaltar que existem na literatura outros trabalhos que buscam determinar os construtos presentes nos questionários das avaliações do Saeb. Podem-se citar o documento *Saeb 2001: Novas perspectivas* (BRASIL, 2001) e o artigo de Franco et al. (2003). Em estudo mais recente, Karino, Vinha e Laros (2014) analisaram os questionários da Prova Brasil do ano de 2009 por meio de análise fatorial dos eixos principais. Este último trabalho citado direcionou o processo de redução de dimensionalidade desenvolvido, diferenciando-se apenas a técnica aqui utilizada - PCA. Conforme pode ser observado na Tabela 1, ao final deste processo, cada registro de aluno passou a ter seis novos atributos: um novo atributo contendo uma pontuação que mensurava a posse de bens, outro atributo com uma pontuação que quantificava a sua frequência de utilização da biblioteca, e assim sucessivamente. Além disso, cada professor passou a ter sete novos atributos com pontuações relacionadas aos construtos identificados; cada diretor passou a ter seis novos atributos com pontuações para seus diferentes construtos; e cada escola quatro novos atributos, representando os construtos extraídos.

Integração, transformação e discretização dos dados

Para cumprir o objetivo de relacionar o desempenho dos alunos com as características associadas aos docentes, diretores e escolas, foi necessário transportar para cada aluno os novos atributos com as pontuações obtidas por seus professores, diretores e escolas, relacionadas aos construtos apresentados na Tabela 1. Para isso, foi preciso integrar os dados presentes em todas as quatro tabelas.

É importante mencionar, ainda, que após o transporte dessas informações para a tabela TS_ALUNO_9EF, foi detectado que 1159 alunos não possuíam escores em relação a nenhum dos 23 componentes. Tal fato ocorreu, pois a pontuação era calculada se todas as variáveis possuíssem resposta. No entanto, alguns alunos, diretores e professores não preencheram totalmente o questionário. Assim, uma vez que esses 1159 alunos não acrescentariam informações em análises posteriores, decidiu-se removê-los da base, restando 111862 registros de alunos.

Outra tarefa consistiu em transformar esses 23 novos atributos presentes na tabela TS_ALUNO_9EF. Essas variáveis eram contínuas com diferentes amplitudes, isto é,

possuíam distintos valores de máximo e mínimo. Visando um melhor entendimento desses dados, os valores desses atributos foram normalizados e transformados para o intervalo [0, 1].

Por fim, foi efetuada uma tarefa com o intuito de transformar esses atributos do tipo contínuo para discreto. Tal transformação recebe o nome de discretização. Foi definido que todos os 23 construtos possuiriam quatro categorias, contendo os valores 1, 2, 3 e 4. Considere X_j um atributo arbitrário dentre esses 23 construtos. Para cada X_j, j variando de 1 a 23, a categoria 1 passou a se referir aos 10% primeiros valores não nulos do atributo X_j, ou seja, os 10% do total de registros que obtiveram as menores pontuações; a categoria 4 relacionava-se aos 10% últimos valores não nulos do atributo X_j ou, equivalentemente, os 10% de registros 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.

Para exemplificar o processo descrito, considere o atributo referente à posse de bens. Conforme visto, este atributo armazena, para cada aluno, uma pontuação em relação aos bens presentes em sua residência. Dos 111862 alunos, 29154 não possuem escore calculado, ou seja, este atributo possui 82708 valores não nulos. Assim, a categoria 1 refere-se aos 10% dos 82708 registros que obtiveram as menores pontuações (intervalo de pontuação da categoria 1: [0, 0,19035]); a categoria 4 refere-se aos 10% dos 82708 registros que obtiveram as maiores pontuações (intervalo de pontuação da categoria 4: [0,42635, 1]); os registros restantes foram alocados nas categorias 2 (intervalo de pontuação da categoria 2: (0,19035, 0,28656]) e 3 (intervalo de pontuação da categoria 3: (0,28656, 0,42635)). Por meio dessa divisão, torna-se possível categorizar os alunos localizados na categoria 1 como possuidores de muito menos bens do que os localizados na categoria 4.

Processo análogo foi feito para os outros atributos; por exemplo, para o atributo ***Violência na escola***, do questionário do professor, pode-se inferir que os alunos na categoria 1 possuem professores que estão vivenciando muito menos violência que os professores que lecionam para os alunos presentes na categoria 4. Interpretação nesse sentido também é dada ao atributo, também oriundo do questionário preenchido pelo professor, denominado ***Necessidade de aperfeiçoamento profissional***. Quanto maior o valor do escore, maior a necessidade de aperfeiçoamento.

O atributo contínuo que aloca a nota do estudante, que contém uma pontuação variando de 0 a 500, foi também discretizado, uma vez que se buscava descobrir relações entre os

23 construtos mencionados e o desempenho obtido pelos alunos no teste de Matemática. Este atributo é denominado atributo alvo e suas categorias recebem o nome de classes. Essa distinção em classes foi realizada de duas formas: uma visando à análise da descoberta de fatores que pudessem influenciar de forma positiva e a outra à descoberta de fatores que pudessem influenciar negativamente o desempenho dos alunos.

Para a análise da influência positiva, os 10 por cento dos estudantes que obtiveram as melhores notas nos testes de matemática (proficiência superior ou igual a 308,565083) foram inseridos na classe nomeada "Nota Alta", totalizando 11186 registros. Os restantes, ou seja, 90 por cento da base de alunos, com nota inferior a 308,565083, foram inseridos em uma classe denominada "Outra".

De modo análogo, para a análise da influência negativa, os alunos foram divididos em outras duas classes: "Nota Baixa", que considerava os 10 por cento dos alunos com os piores resultados, ou seja, nota inferior ou igual a 180,527945, e "Outra", contendo alunos com nota superior a este valor.

Assim, por meio das tarefas anteriormente descritas, os dados foram estruturados de tal modo que possibilitassem a aplicação de um algoritmo de mineração de dados.

Mineração de dados

A etapa de mineração de dados, responsável pela extração de padrões, foi efetuada aplicando-se um classificador bayesiano denominado *Naïve Bayes*. Um classificador bayesiano consiste em classificar um registro em uma determinada classe, alicerçando-se na probabilidade deste registro pertencer a esta classe (HAN; KAMBER; PEI, 2012).

No caso do presente trabalho, baseando-se nos construtos e em suas categorias 1, 2, 3 ou 4 em que os alunos estão inseridos, pretendeu-se identificar a qual classe 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).

Neste trabalho foi utilizada uma implementação do algoritmo *Naïve Bayes*, disponibilizada dentro do *software* Weka (HALL et al., 2009), que é uma ferramenta de código livre que contém algoritmos de mineração (WITTEN; FRANK; HALL, 2011).

Resultados e discussões

O algoritmo *Naïve Bayes* foi executado com o intuito de detectar quais construtos, dentre os 23, foram mais relevantes para a descrição da classe "Nota Alta". Análise análoga foi feita para a classe "Nota Baixa".

Influência positiva

O modelo gerado por *Naïve Bayes* apresenta a probabilidade de o construto ocorrer dado que uma classe ocorra. Essas probabilidades podem ser expressas em porcentagem. Assim, para esta primeira análise, o modelo gerado pelo algoritmo apresenta o percentual de registros da classe "Nota Alta" que se encontram em cada categoria dos atributos. Quanto maior a porcentagem de uma categoria, maior é a influência para a classe "Nota Alta". Por exemplo, para o construto ***Posse de bens***, primeira linha da Tabela 2, 43,77% dos registros da base classificados como "Nota Alta", possuem o valor do construto igual a 3. A tendência é que alunos com bom desempenho estejam presentes nas categorias que representem o melhor cenário relacionado ao construto. Novamente, para exemplificar, nota-se que 55,49% dos alunos com Nota Alta encontram-se nas categorias 3 e 4, referentes as maiores pontuações, do construto ***Posse de bens***. Isso implica que a maioria dos alunos com Nota Alta possuem melhores condições acerca deste tema, ou seja, suas famílias possuem automóveis e suas residências são melhores equipadas com televisão, geladeira, banheiros, computadores, entre outros aspectos.

Tabela 2 - Resultado de mineração de dados gerado pelo algoritmo *Naïve Bayes* (influência positiva)

Questionário	Construto	Categorias (valores em %)			
		1	2	3	4
Aluno	Posse de bens	6,79	37,72	43,77	11,72
	Utilização da biblioteca	8,07	34,48	40,72	16,73
	Dever de casa e correções	7,02	32,12	47,83	13,03
	Incentivo dos pais	10,18	34,24	41,72	13,86
	Hábito de leitura geral	14,47	47,66	32,05	5,82
	Escolaridade dos pais	4,87	28,87	45,87	20,39
Professor	Direção da escola, gestão e liderança	10,56	37,42	39,28	12,74
	Recursos pedagógicos	10,83	38,44	39,39	11,34
	Necessidade de aperfeiçoamento profissional	10,87	40,04	39,29	9,80
	Experiência e salário	8,33	33,40	43,54	14,73
	Práticas pedagógicas	6,59	42,58	43,00	7,83
	Violência na escola	12,39	43,21	36,38	8,02
	Desenvolvimento profissional	8,87	41,04	40,11	9,98
Diretor	Projeto na escola	13,37	40,18	38,03	8,42

	Funcionamento da escola	5,54	33,03	47,21	14,22
	Políticas, ações e programas escolares	9,26	35,47	41,20	14,07
	Experiência e salário	7,36	38,18	43,83	10,63
	Nível de escolaridade e atividade de desenvolvimento profissional	9,49	39,09	37,37	14,05
	Conselho escolar	11,55	37,12	40,42	10,91
Escola	Infraestrutura	8,81	35,35	44,06	11,78
	Recursos de apoio	8,03	37,19	44,06	10,72
	Situação da biblioteca	8,30	39,28	43,68	8,74
	Segurança da escola	11,10	38,33	41,90	8,67

Fonte: Elaboração dos autores.

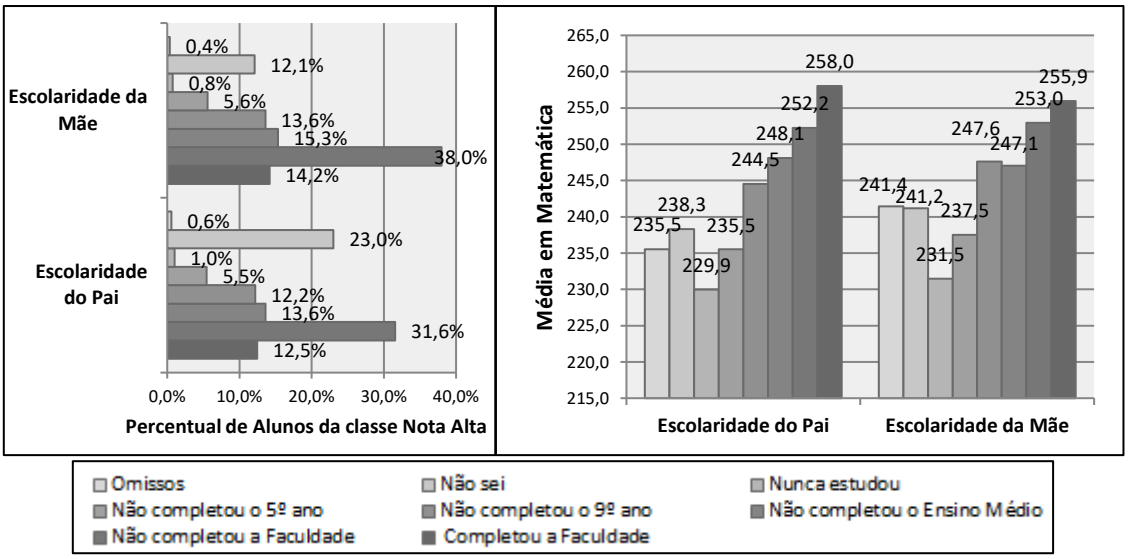
Na sequência, é feita uma discussão dos resultados evidenciados por meio da Tabela 2, apresentando-se uma interpretação dos construtos que possuíram porcentagem superior a 58% ao se analisar a soma das categorias 1 e 2 ou 3 e 4. Esse valor, 58%, foi escolhido, efetivamente, para permitir a seleção de um número viável de construtos para análise, considerando-se as limitações de espaço do artigo.

Escolaridade dos pais

Ao observar a Tabela 2, nota-se que as porcentagens mais expressivas correspondem ao construto ***Escolaridade dos pais***. Cerca de 66% dos alunos pertencentes à classe "Nota Alta" e que possuem escore calculado para este construto encontram-se nas categorias 3 e 4. Em outras palavras, dos 9392 alunos que possuem Nota Alta e escore computado, 6223 estão nas categorias correspondentes às maiores pontuações em relação à escolaridade dos pais. Portanto, pode-se interpretar que alunos que possuem pais com alto nível escolar tendem a obter melhor resultado no teste de Matemática.

Para corroborar esse resultado, o gráfico à esquerda, apresentado na Figura 3, expressa a escolaridade separadamente do pai e da mãe. Assim, permite analisar o número de alunos, dentre os que pertencem à classe "Nota Alta", que preencheram cada alternativa destas questões. Já o gráfico à direita realça a importância da escolaridade dos pais no desempenho escolar, ou seja, apresenta a média da proficiência obtida no teste de Matemática agrupando os alunos de acordo com a alternativa de cada questão. Para computar estas médias, considerou-se não somente os alunos com Nota Alta, mas todos os 111862 estudantes presentes na base.

Figura 3 - Escolaridade dos pais



Fonte: Elaboração dos autores.

Nota-se que 52,2% dos alunos com nota alta possuem mães com, pelo menos, ensino médio completo e 44,1% dos estudantes possuem pais nesta condição. Ao se analisar as médias, percebe-se que alunos que possuem pais que nunca estudaram obtiveram as menores proficiências, uma vez que a média é próxima a 230 pontos. Ao se comparar com os alunos que assinalaram a alternativa correspondente aos pais com ensino superior completo, pode-se ver uma diferença de aproximadamente 30 pontos. Desse modo, conclui-se que com o aumento do nível escolar dos pais, melhores são os resultados dos seus filhos.

Este resultado foi encontrado em diversas outras pesquisas, como a efetuada por Jesus e Laros (2004), ao analisarem o desempenho em Língua Portuguesa dos alunos do 9º ano, por intermédio dos dados do Saeb do ano de 2001. Além deste estudo, com base nos dados da Prova Brasil 2009, Ortigão e Aguiar (2013) verificaram que quanto maior o nível escolar dos pais, menores são as chances de repetência dos alunos do 5º ano.

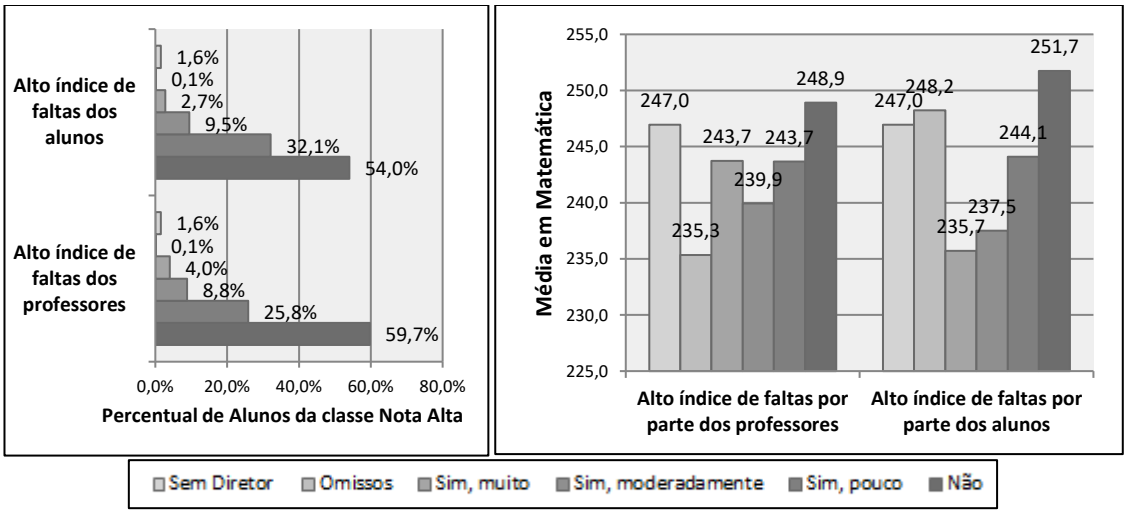
Funcionamento da escola

O construto ***Funcionamento da escola***, oriundo do questionário do diretor, engloba questões sobre recursos financeiros, quantidade insuficiente de professores, carência de apoio pedagógico e assiduidade por parte dos alunos e docentes. Observa-se na Tabela 2 que 61,43% dos alunos com nota alta estão nas categorias 3 e 4. Uma alta pontuação neste construto indica que a escola sofre poucos problemas relacionados a estes temas e, consequentemente, os alunos tendem a obter melhores resultados.

As questões referentes ao alto índice de faltas por parte dos alunos e por parte dos professores apresentaram maiores cargas fatoriais e, consequentemente, maior relevância em comparação às outras questões englobadas no construto (mais detalhes sobre esse aspecto podem ser encontrados em Fonseca (2018)). Por isso, decidiu-se analisá-las separadamente.

Observa-se na Figura 4 que, para ambas as questões, a maioria dos discentes da classe "Nota Alta" estão sob a gestão de diretores que responderam a alternativa "Não", ou seja, não ocorreram problemas de alto índice de faltas por parte de alunos e professores. É importante mencionar que a opção "Sem Diretor" corresponde aos registros de alunos que não possuem um diretor na tabela TS_DIRETOR.

Figura 4 - Assiduidade de alunos e professores



Fonte: Elaboração dos autores.

Em relação à média das proficiências obtidas por todos os alunos, nota-se que os discentes que estão sob a gestão de diretores que responderam a alternativa "Não", acerca do alto índice de faltas dos professores e alunos, possuíram, respectivamente, média de, aproximadamente, 249 pontos e 252 pontos. É notório que as médias dos alunos relacionadas às demais alternativas são inferiores. Portanto, o baixo índice de absenteísmo dos docentes e discentes indica influência positiva no desempenho.

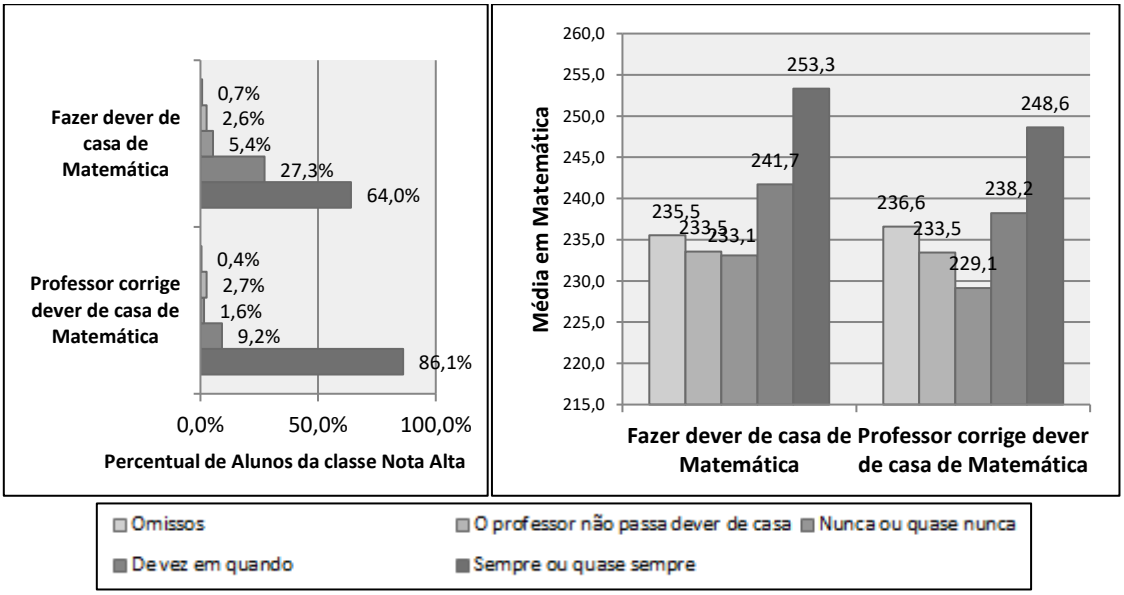
Este resultado reforça o que foi obtido em outras pesquisas, como a de Carvalho et al. (2012), que analisaram escolas da região metropolitana do Rio de Janeiro a partir dos dados dos questionários da Prova Brasil 2009, objetivando estudar fatores intraescolares que favoreciam o desempenho dos alunos do 5º ano em Matemática.

Dever de casa e correções

O construto ***Dever de casa e correções*** reafirma a importância do comprometimento na realização das tarefas extraescolares e a necessidade da discussão e correção destas atividades em sala de aula. 60,86% dos alunos pertencentes à classe "Nota Alta" e que possuem escore calculado para este construto encontram-se nas categorias 3 e 4.

Conforme exposto na Tabela 1, quatro questões compõem o ***construto Dever de casa e correções***. Por evidenciar o ensino de Matemática, gráficos são apresentados na Figura 5 a respeito do dever de casa desta disciplina e a sua correção.

Figura 5 - Dever de casa e correções



Fonte: Elaboração dos autores.

Constata-se que a maioria dos alunos pertencentes à classe "Nota Alta" fazem as tarefas escolares e os seus professores as corrigem. Nota-se, ainda, que, ao se analisar todos os alunos presentes na base, a média decai em 20 pontos quando se comparam os que sempre fazem e seus professores corrigem com os que nunca efetuam as atividades e estas nunca são corrigidas.

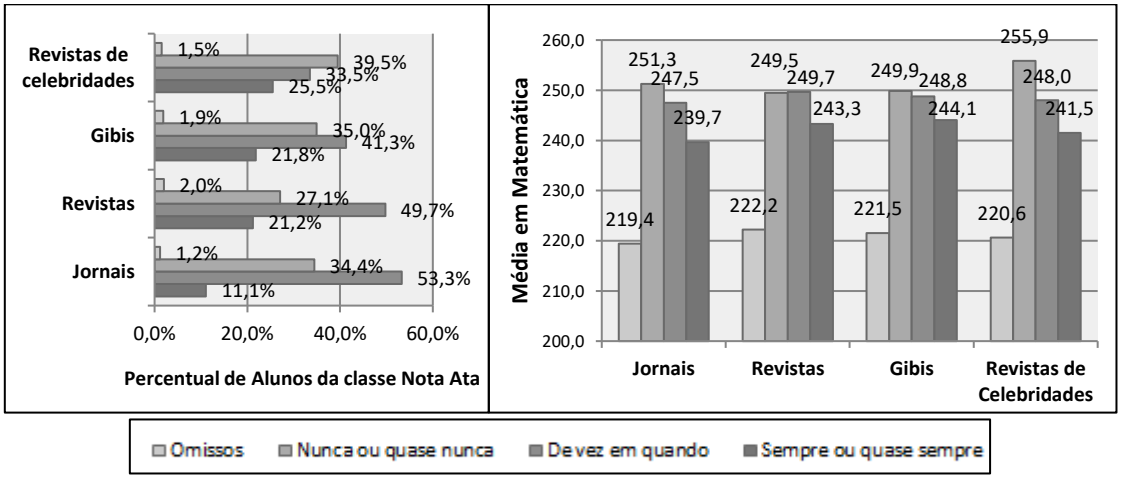
Este resultado corrobora outras pesquisas que dão enfoque não somente ao ensino fundamental, como a de Andrade e Laros (2007) que, ao analisarem os dados do Saeb de 2001, referentes aos alunos da 3ª série do ensino médio, concluíram que o cumprimento dos deveres de casa afeta o desempenho escolar em sentido positivo.

Hábito de leitura geral

Outro construto importante a ser destacado é ***Hábito de leitura geral***, proveniente do questionário preenchido pelo aluno. Observa-se, na Tabela 2, que cerca de 62% dos alunos pertencentes à classe "Nota Alta" encontram-se nas categorias 1 e 2. Estas categorias expressam as menores pontuações no construto. Portanto, este resultado leva à conclusão de que uma menor pontuação neste atributo foi favorável ao desempenho dos alunos em matemática.

Conforme visto na Tabela 1, a variável ***Hábito de leitura geral*** agrupou as questões 32, 35, 36 e 37. Estas questões referem-se, respectivamente, à frequência de leitura de jornais, revistas gerais, gibis e revistas sobre celebridades/esportes/televisão. O gráfico a esquerda da Figura 6 apresenta o número de alunos, pertencentes à classe "Nota Alta", em cada alternativa destas questões.

Figura 6 – Hábito de leitura geral



Fonte: Elaboração dos autores.

Para realçar a relação entre essas quatro questões e o desempenho, médias considerando todos os alunos da base foram computadas agrupando-os de acordo com cada alternativa. Os resultados são apresentados no gráfico à direita da Figura 6.

Observa-se que a média decai à medida que aumenta a frequência de leitura por meio destes veículos de informação. A relação que foi descoberta, a priori, vai contra o senso comum que, a princípio, consideraria essa prática como positiva. Buscando efetuar uma breve reflexão sobre a questão, essa relação poderia, por exemplo, sugerir que alunos que possuem maior interesse pela leitura de revistas, jornais ou gibis, podem ter um perfil menos focado em questões lógicas e matemáticas. Poderia indicar, portanto, a necessidade de um trabalho maior para motivação de alunos com essas características.

Uma possibilidade seria a utilização de artigos de jornais e de revistas para contextualizar problemas em que a matemática seria aplicada. Essa estratégia poderia motivar alunos com esse perfil. No entanto, essas são percepções que devem ser aprofundadas a partir do desenvolvimento de pesquisas adicionais. O trabalho de Ortigão, Franco e Carvalho (2007) aborda parte desse problema sob uma perspectiva distinta. Esses autores puderam identificar que o uso de jornais em aula é pouco eficaz por se tratar de uma mera fonte de pesquisa e recorte de números, gráficos e tabelas, sem relacionar, portanto, com os tópicos abordados em Matemática.

Outros construtos

Em relação aos outros construtos presentes na Tabela 2, observa-se que a maioria dos alunos pertencentes à classe "Nota Alta" encontra-se nas categorias 3 e 4. Já para o atributo intitulado ***Necessidade de aperfeiçoamento profissional***, do professor, e o atributo que mostra a visão do docente a respeito da ***Violência na escola***, que possuem sentido contrário aos demais, a maioria dos alunos está concentrada nas categorias 1 e 2. Desse modo, apesar das porcentagens das categorias que representam o pior e o melhor cenário não apresentarem uma diferença considerável, o resultado apresentado na Tabela 2 foi coerente.

Influência negativa

Após observar os fatores que afetam positivamente o desempenho, foi executado novamente o algoritmo *Naïve Bayes* a fim de identificar quais valores de construtos favoreciam a classificação do aluno em "Nota Baixa". Os resultados são apresentados na Tabela 3.

Tabela 3 - Resultado de mineração de dados gerado pelo algoritmo *Naïve Bayes* (influência negativa)

Questionário	Construto	Categorias (valores em %)			
		1	2	3	4
Aluno	Posse de bens	12,20	37,16	39,15	11,49
	Utilização da biblioteca	13,46	42,35	35,86	8,33
	Dever de casa e correções	15,18	44,08	32,56	8,18
	Incentivo dos pais	11,91	40,67	37,50	9,92
	Hábito de leitura geral	11,52	35,09	40,39	13,00
	Escolaridade dos pais	13,61	42,18	35,68	8,53
Professor	Direção da escola, gestão e liderança	10,60	42,82	38,32	8,26
	Recursos pedagógicos	9,95	38,82	41,40	9,83
	Necessidade de aperfeiçoamento profissional	10,14	41,23	37,80	10,83
	Experiência e salário	13,36	43,16	36,79	6,69

	Práticas pedagógicas	11,74	38,82	37,52	11,92
	Violência na escola	9,31	38,06	42,12	10,51
	Desenvolvimento profissional	11,22	41,23	38,05	9,50
Diretor	Projeto na escola	9,33	39,17	40,81	10,69
	Funcionamento da escola	13,41	45,10	34,63	6,86
	Políticas, ações e programas escolares	11,69	40,86	39,13	8,32
	Experiência e salário	12,25	40,63	37,51	9,61
	Nível de escolaridade e atividade de desenvolvimento profissional	10,85	40,89	39,28	8,98
	Conselho escolar	10,19	42,09	39,45	8,27
Escola	Infraestrutura	11,40	43,68	36,13	8,79
	Recursos de apoio	10,84	41,13	38,57	9,46
	Situação da biblioteca	12,07	41,79	35,86	10,28
	Segurança da escola	9,29	37,40	40,82	12,49

Fonte: Elaboração dos autores.

Para enfatizar a influência negativa no desempenho escolar, apresenta-se uma discussão dos resultados por meio de uma interpretação individual sobre os atributos que possuíram porcentagem superior a 58% (mesmo percentual utilizado para a análise da influência positiva) ao se observar a soma das categorias que representavam o melhor ou pior cenário.

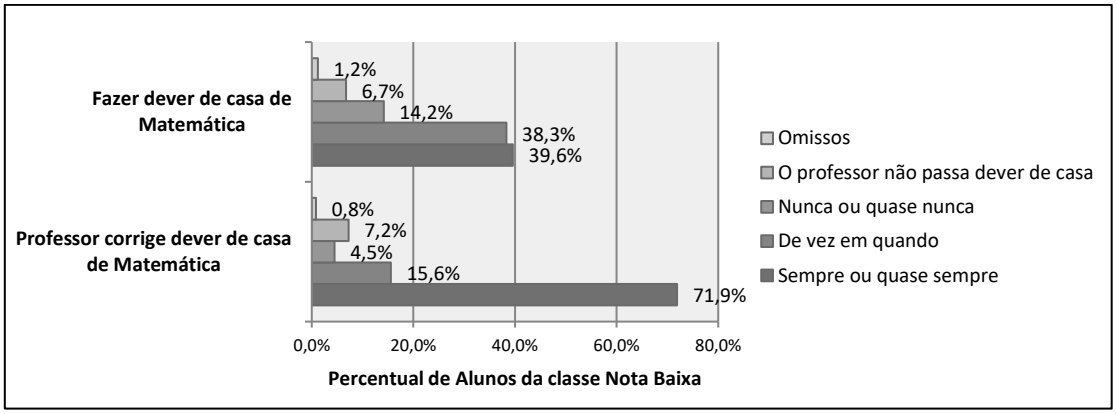
Dever de casa e correções

Analisando-se as porcentagens apresentadas na Tabela 3, pode-se observar que o construto ***Dever de casa e correções*** é o que apresenta diferenças mais expressivas entre as categorias que expõem o melhor e o pior cenário. Nota-se que 59,26% dos alunos pertencentes à classe "Nota Baixa" encontram-se nas categorias 1 e 2. Nestas categorias estão os alunos que obtiveram as menores pontuações em relação a fazer e corrigir o dever de casa.

Como verificado anteriormente, para situações em que o aluno faz o dever de casa e seu professor o corrige, havia influência positiva no desempenho em Matemática. Situação inversa agora foi constatada, ou seja, quando o aluno não faz o dever de casa e não há a correção do mesmo, existe um efeito adverso no processo de aprendizagem.

Na Figura 7 pode-se visualizar o número de alunos, pertencentes à classe "Nota Baixa", em cada alternativa dessas duas questões. Para facilitar a comparação entre a base de alunos com "Nota Baixa", "Nota Alta" e todos os alunos, a Tabela 4 sintetiza as informações.

Figura 7 – Dever de casa e correções referentes aos alunos da classe “Nota Baixa”



Fonte: Elaboração dos autores.

Tabela 4 - Percentuais de alunos de todas as bases sobre dever de casa e correções

	Fazer dever de casa			**Professor corrigir dever de casa**		
	Todos Alunos	Alunos “Nota Alta”	Alunos “Nota Baixa”	Todos Alunos	Alunos “Nota Alta”	Alunos “Nota Baixa”
Sempre ou quase sempre	48,2	64,0	39,6	80,0	86,1	71,9
De vez em quando	37,2	27,3	38,3	12,3	9,2	16,6
Nunca ou quase nunca	9,2	5,4	14,2	2,4	1,6	4,5
Professor não passa dever de casa	4,5	2,6	6,7	4,7	2,7	7,2
Omissos	0,9	0,7	1,2	0,6	0,4	0,8

Fonte: Elaboração dos autores.

Para exemplificar a interpretação da Tabela 4, considere a questão sobre fazer dever de casa. Constata-se que dos 111862 alunos de toda a base, 48,2% fazem sempre o dever de casa de Matemática. Ao se analisar somente os que pertencem à classe “Nota Alta”, ou seja, 11186 alunos, 64,0% destes cumprem as atividades. Já, ao se avaliar somente os 11186 alunos da classe “Nota Baixa”, pode-se ver que apenas 39,6% são comprometidos com as tarefas.

Portanto, os resultados evidenciam que alunos com “Nota Baixa” fazem menos o dever de casa de Matemática do que os que obtiveram “Nota Alta”. Além disso, atente-se ao fato de que 7,2% dos alunos com resultado inferior no desempenho em Matemática assinalaram que seus professores não passam dever de casa.

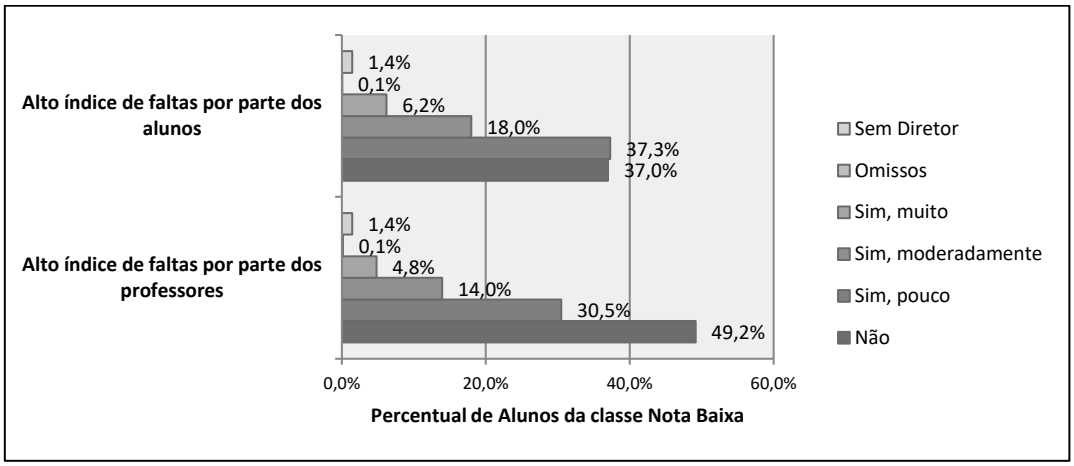
Funcionamento da escola

Ao se observar a Tabela 3, percebe-se que 58,51% dos alunos pertencentes à classe “Nota Baixa” encontram-se nas categorias 1 e 2 para esse construto. Logo, a maioria dos alunos

está concentrada nas menores pontuações, referentes à ocorrência de problemas no funcionamento escolar.

Na Figura 8 apresenta-se o percentual de alunos desta classe considerando as questões sobre assiduidade de alunos e professores.

Figura 8 - Assiduidade referente aos alunos da classe "Nota Baixa"



Fonte: Elaboração dos autores.

A Tabela 5 permite comparar as classes "Nota Alta" e "Nota Baixa". Observa-se nesta tabela que 59,7% dos alunos pertencentes à classe "Nota Alta" possuem diretores que responderam que não há alto índice de faltas por parte dos professores. Contudo, este valor decai para 49,2% ao se analisar os alunos da classe "Nota Baixa". Constata-se, ainda, que 54% com "Nota Alta" e 37% com "Nota Baixa" possuem diretores que responderam que não há alto índice de faltas por parte dos alunos. Logo, pode-se concluir que a assiduidade é um importante fator para o desempenho escolar.

Tabela 5 - Percentuais de alunos de todas as bases sobre assiduidade de alunos e professores

	Alto índice de faltas (professores)			**Alto índice de faltas (alunos)**		
	Todos Alunos	Alunos "Nota Alta"	Alunos "Nota Baixa"	Todos Alunos	Alunos "Nota Alta"	Alunos "Nota Baixa"
Não	53,9	59,7	49,2	43,3	54,0	37,0
Sim, pouco	28,5	25,8	30,5	36,1	32,1	37,3
Sim, moderadamente	11,6	8,8	14,0	14,0	9,5	18,0
Sim, muito	4,4	4,0	4,8	5,0	2,7	6,2
Omissos	0,1	0,1	0,1	0,1	0,1	0,1
Sem Diretor	1,5	1,6	1,4	1,5	1,6	1,4

Fonte: Elaboração dos autores.

Outros construtos

É importante ressaltar que, para os construtos restantes presentes na Tabela 3, é notório que em grande parte deles a maioria dos alunos da classe "Nota Baixa" se concentra nas categorias que representam os piores cenários em relação a cada tema. Tal fato é esperado, pois o aprendizado está diretamente relacionado com o meio em que o aluno está inserido. No entanto, a diferença entre as categorias desses atributos não foi tão significativa como ocorreu no estudo da influência positiva.

Considerações finais

A partir dos resultados apresentados neste artigo, algumas questões puderam ser identificadas como relevantes no desempenho obtido em Matemática. Conforme demonstrado, quanto maior o nível escolar dos pais, melhores são os resultados alcançados por seus filhos. Além disso, a frequência regular dos alunos na escola, baixo índice de absenteísmo dos professores, o fato dos alunos gostarem de estudar Matemática e os docentes corrigirem as tarefas escolares, são quesitos importantes. Observou-se também uma relação inversa entre a frequência de leitura de revistas e jornais e o desempenho em matemática. Sugere-se que essa questão seja explorada com mais profundidade em pesquisas futuras.

Como ponto central, vale salientar a contribuição de se apresentar uma metodologia que permitiu a extração de conhecimento a partir da integração de diferentes dimensões de dados da Prova Brasil. A abordagem exposta, caracterizada pela redução da dimensionalidade dos dados combinada com uma mineração posterior, viabilizou, ainda, o estudo de um grande volume de dados, relacionando aspectos não somente intrínsecos aos alunos, mas também relacionados aos seus diretores, professores e escolas.

Cabe ressaltar que a análise foi efetuada apenas considerando a base de dados de 2013 e o Estado do Rio de Janeiro, podendo ser expandida para outros anos e regiões. Ademais, outros algoritmos de mineração de dados poderiam ser aplicados com o intuito de comparar os resultados obtidos. Ainda, como trabalho futuro, a avaliação dos construtos poderia ser feita não somente por porcentagens em cada categoria (que é uma medida técnica, proveniente de um processo automatizado), mas também por meio de outras medidas desenvolvidas por pesquisadores da área educacional. Tal campo de pesquisa é denominado mineração de dados orientada ao conhecimento do domínio (*Domain-Driven*

Data Mining – D³M) (CAO, 2010). Mesmo diante das perspectivas de estudos complementares, acredita-se que este artigo contribui como base e estímulo para a exploração das bases da Prova Brasil e de outras avaliações do Inep.

Agradecimentos

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

Referências

ANDRADE, J. M. d.; LAROS, J. A. Fatores associados ao desempenho escolar: estudo multinível com dados do Saeb/2001. *Psicologia: Teoria e Pesquisa*, Brasília, v. 23, n. 1, p. 33-42, 03 2007.

AZEVEDO, J. M. L. O Estado, a política e a regulação do setor educacional *no* Brasil*:* uma abordagem histórica. In: FERREIRA, N. S. C. e AGUIAR, M. A. S. *Gestão da educação: impasses, perspectivas e compromissos*. 2. ed. São Paulo: Cortez, p. 17-42, 2001.

BAKER, R. S. J. d.; ISOTANI, S.; CARVALHO, A. M. J. B. d. Mineração de dados educacionais: oportunidades para o brasil. *Revista Brasileira de Informática na Educação*, v. 19, n. 2, p. 3-13, ago. 2011.

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

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

CARVALHO, C. P. de; WALHELM, A. P. S.; ALVES, F.; KOSLINSKI, M. Gestão e desempenho escolar: um estudo nas redes municipais da Região Metropolitana do Rio de Janeiro a partir dos resultados da Prova Brasil 2009. In: CONGRESSO IBERO AMERICANO DE POLÍTICA E ADMINISTRAÇÃO DA EDUCAÇÃO. 3., 2012, Zaragoza. *Cadernos ANPAE*. Timbaúba: Biblioteca ANPAE - Cadernos ANPAE, 2012.

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

COELHO, M. I. A. d. M. Vinte anos de avaliação da educação básica no Brasil: aprendizagens e desafios. *Ensaio: Avaliação e Políticas Públicas em Educação*, v. 16, n. 59, p. 229-258, jun. 2008.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. *AI Magazine*, American Association for Artificial Intelligence, California, USA, v. 17, n. 3, p. 37-54, 1996.

FIELD, A. *Descobrindo a Estatística usando o SPSS*. 2. ed. Porto Alegre: Artmed, 2009.

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

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. Tese (Doutorado em Modelagem Computacional) - Instituto Politécnico, UERJ, 2018.

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*, n. 28, p. 39-74, 2003.

FRANCO, C.; SZTAJN, P.; ORTIGÃO, M. I. R. Mathematics teachers, reform and equity: results from the brazilian national assessment. *Journal for Research in Mathematics Education*, v. 38, n. 4, p. 393-419, 2007.

FREITAS, D. N. T. *A avaliação da educação básica no Brasil:* dimensão normativa, pedagógica e educativa. Campinas: Autores Associados, 2007.

FUNDACÃO LEMANN E MERITT. *QEdu: aprendizado em foco*. 2015. Sítio na internet. Disponível em: <http://qedu.org.br/>. Acesso em: 20 Jun. 2016.

GUZZO, R. S. L.; EUZEBIOS FILHO, A. Desigualdade social e sistema educacional brasileiro: a urgência da educação emancipadora. *Escritos educ.*, Ibirité, v. 4, n. 2, p. 39-48, dez. 2005.

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

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. The WEKA Data Mining Software: An Update; *SIGKDD Explorations*, New York, USA, v. 11, n. 1, p. 10-18, jun. 2009.

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

INEP. *Microdados da Aneb e da Anresc 2013*. Brasília: Inep, 2015. Acesso em: 2015-05-30. Disponível em: <http://portal.inep.gov.br/basicalevantamentos-acessar>.

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, 2004.

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.

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

ORTIGÃO, M. I. R.; FRANCO, C.; CARVALHO, J. B. P. 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, 2007.

ORTIGÃO, M. I. R.; AGUIAR, G. S. Repetência escolar nos anos iniciais do ensino fundamental: evidências a partir dos dados da prova Brasil 2009. *Rev. bras. Estud. pedagog.*, Brasília, v. 94, n. 237, p. 364-389, 2013.

PEREIRA, M. J.; MORI, N. N. R. Diretrizes curriculares e o desempenho de alunos paranaenses da 8ª série do ensino fundamental na Prova Brasil. *RBPG*, Brasília, supl. 1, v. 8, p. 121 - 143, dezembro 2011.

POSTGRESQL. *Documentação do PostgreSQL 8.0.0*. Rio de Janeiro, Brasil, 2007. 1310 p.

RODRIGUES, C. G.; GUIMARÃES, R. R. de M.; RIOS-NETO, E. L. G. O papel das origens sociais sobre a proficiência escolar e a probabilidade de progressão por série no Brasil: evidência de persistência. *RBPG*, Brasília, supl. 1, v. 8, p. 87-116, dezembro 2011.

SANTOS, J. B. P; TOLENTINO-NETO, L. C. B. O que os dados do SAEB nos dizem sobre o desempenho dos estudantes em Matemática? *Educ. Matem. Pesq*., São Paulo, v.17, n.2, p.309-333, 2015.

SOARES, J. F. Qualidade e equidade na Educação Básica brasileira: a evidência do Saeb-2001. *Arquivos Analíticos de Políticas Educativas*, Tempe, USA, v. 12, n. 38, 2004.

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

Texto recebido: 20/03/2018
Texto aprovado: 24/08/2018