

Desempenho das Instituições de Ensino Brasileiras no ENEM: uma Abordagem Usando Mineração de Dados

Raphael Hoed

Instituto Federal de Educação,
Ciência e Tecnologia do Norte de Minas Gerais
Brazil
raphael.hoed@gmail.com

Pedro Fábio Saraiva

Instituto Federal de Educação,
Ciência e Tecnologia do Norte de Minas Gerais
Brazil
rpedro.fabio@ifnmg.edu.br

ABSTRACT

In this paper we used the mining of association rules via the Apriori algorithm, using the High School National Exam (ENEM) microdata provided by the National Institute for Educational Studies and Research “Anísio Teixeira” (INEP). Factors influencing the performance of educational institutions in ENEM were verified, based on variables such as the percentage of teacher education, dropout rate, pass and fail rate, location of the institution (whether urban or rural), type of institution. (whether public or private), etc. From the use of the Apriori algorithm, it is expected to gain a better understanding of the factors that can contribute to the positive performance of schools, thus proposing ways to expand the good performance.

RESUMO

Neste artigo foi empregada a mineração de regras de associação via algoritmo *Apriori*, usando os microdados do Exame Nacional do Ensino Médio (ENEM) por escola, disponibilizados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). Foram verificados os fatores que influenciam o desempenho das instituições de ensino no ENEM, tomando como base variáveis como o percentual de formação docente, taxa de abandono, taxa de aprovação e de reprovação, localização da instituição (se urbana ou rural), tipo da instituição (se pública ou privada), etc. A partir do emprego do algoritmo *Apriori*, espera-se obter maior entendimento sobre os fatores que podem contribuir para o desempenho positivo das escolas, propondo assim formas de ampliar o bom desempenho.

Palavras-Chave

Mineração de Dados; *Apriori*; ENEM; Escolas

Classificação de palavras-chave ACM

Educação, Aprendizagem de Máquina.

INTRODUÇÃO

O Brasil passou, ao longo da última década, por alterações no processo de seleção para ingresso no ensino superior. Com essas alterações o tradicional exame vestibular foi substituído em muitas universidades pelo Exame Nacional do Ensino Médio (ENEM). O ENEM foi criado inicialmente como um instrumento para avaliar o desempenho dos estudantes ao concluir a educação básica, sendo que a partir de 2009 começou a ser utilizado como forma de acesso ao ensino superior.

O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) disponibiliza os dados individuais (microdados) das instituições ofertantes de ensino médio que dispõem de discentes que realizaram a prova do ENEM. Esses microdados estão disponíveis a partir do ano de 2005 até o ano de 2015 e são bastante ricos, podendo ser estudados por meio de técnicas de mineração de dados. A mineração de dados consiste em analisar dados e aplicar algoritmos que produzirão um conjunto de padrões de certos dados [6].

O presente trabalho tem por objetivo estudar os microdados do ENEM por instituição de ensino, disponibilizados gratuitamente pelo INEP, afim de verificar quais os fatores estão relacionados ao desempenho dos alunos, tendo como base a instituição de ensino onde estudaram. Para isso, foi adotada a mineração de regras de associação via algoritmo *Apriori*.

Este artigo está organizado da seguinte forma: A Seção “Fundamentação Teórica” explora os trabalhos correlatos sobre o desempenho dos alunos no ENEM e discorre sobre a técnica de mineração de dados empregada; A Seção “Metodologia” mostra detalhes do emprego da metodologia de mineração de dados *Cross Industry Standard Process for Data Mining* (CRISP-DM) ao estudo em questão; A Seção “Estudos Realizados” mostra os resultados obtidos com os estudos conduzidos. A Seção “Conclusões” exhibe as conclusões obtidas a partir dos estudos realizados. A Seção “Referências Bibliográficas” exhibe os autores e respectivas obras que embasaram esse estudo.

FUNDAMENTAÇÃO TEÓRICA

Esta seção está dividida em duas subseções: a subseção 2.1 trata de estudos similares feitos sobre o desempenho de alunos no ENEM usando os microdados disponibilizados pelo INEP e a subseção 2.2 discorre sobre a mineração de regras de associação empregada nesse estudo.

Dados do ENEM

O objetivo atual do ENEM vai além da avaliação de qualidade do ensino médio no Brasil, atuando também como substituto do vestibular tradicional [12]

Embora haja um discurso governamental no sentido da adoção de práticas que melhorem a qualidade do ensino básico (fundamental e de nível médio) brasileiro, de acordo com dados do *Programme for International Student Assessment* (PISA) referentes ao ano de 2015, o

desempenho dos estudantes do Brasil em áreas como ciências, leitura e matemática é inferior ao de vários países membros da Organização para Cooperação e Desenvolvimento Econômico (OCDE) [13]. Isso reitera a importância de se conhecer as causas que levam as instituições de ensino médio a formarem alunos com deficiências em diferentes campos do conhecimento.

A Tabela 1 exibe alguns autores que estudaram os microdados do ENEM, algumas evidências obtidas e as técnicas de mineração de dados empregadas no estudo:

Autor	Técnica Empregada	Resultados Obtidos
Simon e Cazella [13]	Árvore de Decisão	Os melhores desempenhos de candidatos nas provas da áreas de ciências da Natureza e suas Tecnologias foram de alunos de escolas privadas e de nível socioeconômico elevado.
Silva, Morino E Sato [12]	Algoritmo <i>Apriori</i>	Baixa renda, baixa escolaridade dos pais e número elevado de pessoas residentes na mesma casa influenciam alunos do estado de São Paulo no mau desempenho no ENEM.
STEARN S. et al. [14]	Árvore de Decisão	Indicadores socioeconômicos podem ajudar a prever a nota do aluno no ENEM
Alves, Cechinel E Queiroga [1]	Árvore de Decisão	O tipo de escola (se público ou privada) e nível socioeconômico influenciam o desempenho do aluno na prova do ENEM

Tabela 1. Estudos envolvendo microdados do ENEM.

Nos estudos apresentados na Tabela 1, os microdados utilizados para análise foram os de desempenho individual de cada aluno no ENEM. Ao realizar a prova, o discente preenche um questionário socioeconômico, o que permite

verificar de que forma fatores deste tipo influenciam no rendimento do aluno neste exame. Contudo, neste artigo, pretende-se analisar o desempenho dos alunos levando-se em consideração a instituição ofertante do ensino médio onde estudaram. Nesse sentido, outras variáveis, não elencadas nos estudos mencionados na Tabela 1, foram levadas em consideração, como nível de formação dos docentes, taxa de evasão, localização da escola (se urbana ou rural), grupo socioeconômico da instituição de ensino etc.

Em relação à formação do professor, Carmo et al., ao analisarem a relação entre o desempenho escolar no ensino médio e a adequação na formação docente, estudando dados do INEP dos anos de 2013 e 2014, verificaram que “a política de adequação entre a formação docente na licenciatura e a disciplina ministrada produz resultados positivos sobre a proficiência dos alunos”[4]. Ainda de acordo com Carmo et al., “não há na legislação educacional a exigência da atuação docente exclusivamente na disciplina de sua formação” [4]. Assim, o professor pode acabar atuando em disciplinas fora da sua área de formação.

No que diz respeito à localização da escola (se urbana ou rural), Pieri e Santos [8] verificaram que a formação dos professores varia consideravelmente entre as escolas urbanas e rurais, sendo que nessas últimas o índice tende a ser menor, o que pode levar a um menor desempenho dos alunos da zona rural em algumas disciplinas.

Já em relação à condição socioeconômica dos alunos e grupo socioeconômico ao qual pertence a instituição de ensino, Sampaio et al. [11] enfatizam que a renda tem um papel decisivo no desempenho em exames vestibulares, uma vez que alunos mais ricos tem melhores condições de estudo ao cursar instituições privadas e cursinhos preparatórios, o que amplia as desigualdades em relação aos discentes com condições socioeconômicas deficientes.

No que diz respeito às taxas de evasão, Batista, Souza e Oliveira [3], estabelecem uma correlação entre a evasão no ensino médio e as condições socioeconômicas dos alunos, condições estas que, conforme mencionado no parágrafo anterior desta seção, influenciam o desempenho do aluno.

A melhoria da educação envolve uma série de medidas tais como: valorização docente, melhoria das condições salariais, qualificação constante, flexibilização da jornada de trabalho com espaço para atuação em pesquisa, melhoria das condições de trabalho (infraestrutura e número de discentes por turma) [4]. Percebe-se que o caminho para se consolidar uma educação de qualidade não é trivial, sendo que esse estudo não pretende limitar os indicadores de qualidade na educação básica às variáveis estudadas por meio dos microdados do ENEM. Contudo, a compreensão dos dados disponibilizados pelo INEP pode gerar importantes subsídios no sentido de mitigar os problemas que há muitos anos assolam a educação brasileira.

Mineração de Regras de Associação

O algoritmo *Apriori* será usado nesse trabalho para mineração das regras de associação. Procura-se descobrir associações importantes entre o desempenho do aluno na prova do ENEM e os informações referentes à instituição de ensino onde curso o ensino médio.

De acordo com Romão et al.: “Uma das técnicas mais atraentes é a Mineração de Regras de Associação, que tem como destaque o algoritmo *Apriori*. Ele pode trabalhar com um número grande de atributos, gerando várias alternativas combinatórias entre eles” [10].

De acordo com Hoed [7], a mineração de regras de associação tem muitas aplicações comerciais em se tratando, por exemplo, de supermercados, quando se pode averiguar, a partir de um banco de dados, se a venda de um determinado produto também está associada à venda de outro produto. A descoberta de regras de associação desse tipo podem subsidiar decisões como melhor disposição das mercadorias no supermercado, colocando estrategicamente os produtos correlacionados uns próximos aos outros. Romão et al. afirmam que: “O objetivo, então, é encontrar todas as regras de associação relevantes entre os itens, do tipo X(antecedente) \Rightarrow Y(consequente)”[10].

A mineração de regras de associação não é útil apenas no contexto de transações comerciais, mas pode ser empregada também para análise de dados do ensino [7]. No que diz respeito ao contexto desse artigo, pode-se verificar, por exemplo, quais fatores institucionais se correlacionam ao bom ou ao mau desempenho dos alunos no ENEM.

A descoberta de regras de associação pode ser decomposta em duas etapas, de acordo com AGRawal et al. (1993 apud ROMÃO et al., 1999) [10]: localizar os conjuntos de itens (*itemsets*) que apresentam suporte superior ao mínimo definido à partida; utilizar os *itemsets* obtidos na etapa 1 para gerar as regras de associação do banco de dados. Algumas definições importantes sobre mineração de regras de associação: “A toda regra de associação $A \rightarrow B$ associamos um grau de confiança, denotado por $conf(A \rightarrow B)$ ” [2]. O grau de confiança seria a probabilidade de que uma transação que tenha um item, também contenha o outro item. A Equação (1) a seguir, formaliza essa definição [2]:

$$conf(A \rightarrow B) = \frac{\text{número de transações que suportam } (A \cup B)}{\text{número de transações que suportam } A} \quad (1)$$

Uma outra definição importante seria que: “A toda regra de associação $A \rightarrow B$ associamos um suporte, denotado por $sup(A \rightarrow B)$ definido como sendo o suporte do *itemset* $A \cup B$ ” [2].

O suporte seria a proporção de transações que contém os itens. Ao definir um grau mínimo de confiança e um grau mínimo de suporte, uma regra de associação interessante seria então aquela que possui um suporte igual ou superior ao mínimo definido e aquela que possui uma confiança igual ou superior ao mínimo definido.

De acordo com Ribeiro [9], para encontrar regras consideradas fortes, além do suporte e da confiança é

também utilizada a medida *Lift*, que é definida pela Equação 2:

$$Lift(A,B) = \frac{P(A \cup B)}{P(A)P(B)} \quad (2)$$

Ainda de acordo com Ribeiro: “A ocorrência de um item A é independente de um item B se $P(A \cup B) = P(A)P(B)$. Se não, existe uma correlação entre os itens” [9].

Desta forma, se o valor da Equação 2 for menor que 1, então a ocorrência de A correlaciona-se negativamente com a ocorrência de B. Se o resultado for superior a 1, A e B se correlacionam positivamente, evidenciando que a ocorrência de A implica na ocorrência de B. Sendo assim, no âmbito desse estudo, só serão consideradas como válidas as regras obtidas cujo *Lift* seja superior a 1.

As fases de execução do algoritmo *Apriori* compreendem geração, poda e validação [2]. Resumidamente, sem entrar em detalhes sobre cada fase, na fase de geração são gerados os *itemsets* que tenham alguma chance de serem frequentes, na fase de poda são descartados os *itemsets* sem chances de serem frequentes, e na última é calculado o suporte de cada um dos *itemsets* do conjunto [2]. O funcionamento do algoritmo *Apriori* é descrito da seguinte forma: “Na primeira passagem, o suporte para cada item individual (conjuntos-de-1-item) é contado e todos aqueles que satisfazem o suporte_mínimo são selecionados, constituindo-se os conjuntos-de-1-item frequentes (F1). Na segunda iteração, conjuntos-de-2-itens candidatos são gerados pela junção dos conjuntos-de-1-item (a junção é feita através da função *apriori-gen*) e seus suportes são determinados pela pesquisa no banco de dados, sendo, assim, encontrados os conjuntos-de-2-itens frequentes. O algoritmo *Apriori* prossegue iterativamente, até que o conjunto-de-k-itens encontrado seja um conjunto vazio” [15].

METODOLOGIA

A metodologia de estudo a ser aplicada para analisar os microdados do ENEM segue as etapas previstas na metodologia de mineração de dados CRISP-DM compreendendo as seguintes etapas: compreensão do negócio, compreensão dos dados, preparação dos dados, modelação (aplicação das técnicas de mineração de dados), avaliação dos resultados e desenvolvimento [5].

A etapa de compreensão do negócio envolve a compreensão do objetivo da pesquisa, conforme foi apresentado anteriormente na Seção 1 (Introdução) deste trabalho.

Durante a compreensão dos dados, foi necessário analisar os microdados do ENEM, disponibilizados no Portal do INEP (<http://portal.inep.gov.br/microdados>) para verificar quais as variáveis disponíveis no base de dados, identificando quais delas serão úteis no estudo. No âmbito

desse estudo foram utilizadas as seguintes variáveis referentes às escolas de ensino médio:

- UF: Unidade Federativa da escola.
- TP_DEPENDENCIA_ADM_ESCOLA: se a escola é de administração federal, estadual, municipal ou privada.
- TP_LOCALIZACAO_ESCOLA: se a escola é urbana ou rural.
- INSE: Indicador de Nível Socioeconômico da escola, variando de 1 (mais baixo) a 6 (mais alto).
- PORTE_ESCOLA: identifica o porte da escola a partir do número de alunos.
- *NU_MEDIA_CN: Média das notas de Ciências da Natureza do Ensino Médio Regular.
- *NU_MEDIA_CH: Média das notas de Ciências Humanas do Ensino Médio Regular.
- *NU_MEDIA_LP: Média das notas de Linguagens e Códigos do Ensino Médio Regular.
- *NU_MEDIA_MT: Média das notas de Matemática do Ensino Médio Regular.
- *NU_MEDIA_RED: Média das notas de Redação do Ensino Médio Regular.
- *PC_FORMACAO_DOCENTE: Indicador de Adequação da Formação Docente da escola para lecionar no Ensino Médio válido para os anos de 2013 a 2015. Essa variável indica o percentual de docentes na instituição que lecionam em áreas condizentes com a sua formação acadêmica.
- *NU_TAXA_PERMANENCIA: Indicador de Permanência na Escola para o Ensino Médio.
- *NU_TAXA_APROVACAO: Taxa de aprovação dos alunos no Ensino Médio.
- *NU_TAXA_REPROVACAO: Taxa de reprovação dos alunos no Ensino Médio
- *NU_TAXA_ABANDONO: Taxa de abandono dos alunos no ensino médio

Na fase de compreensão dos dados foi utilizado um *software* de planilha eletrônica (*Microsoft Excel*) pois a maioria dos microdados fornecidos encontra-se em formato *Comma Separated Values* (CSV). Foram utilizados os microdados correspondentes ao ano de 2017. O arquivo CSV analisado possui um total de 172305 registros referentes aos anos de 2005 a 2015.

Durante a preparação dos dados, foi feita a limpeza com remoção das variáveis que não serão úteis para o estudo e adequação para que as técnicas de mineração de dados sejam empregadas. Nessa fase o *software* de planilha eletrônica *Microsoft Excel* também foi usado. Variáveis como Código da escola, nome da escola, código do município, por não agregarem valor a essa pesquisa, foram removidas. Nessa etapa, foram removidos os registros referentes aos anos de 2005 a 2008, visto que, de acordo com o dicionário de variáveis que acompanha os microdados do ENEM, até 2008 a escala das notas

variavam de 0 a 100. Nos demais anos a escala varia de 0 a 1.000. Devido a essa diferença metodológica, decidiu-se trabalhar apenas com os dados compreendidos entre os anos de 2009 a 2015, totalizando assim, após a remoção dos dados descartados, 104687 registros.

Ainda em relação à fase de preparação, algumas variáveis foram discretizadas para facilitar a análise dos dados. As variáveis discretizadas são aquelas apresentadas anteriormente nesta seção que iniciam com “*”. Na discretização foram utilizados quartis estatísticos, obtendo-se os seguintes valores possíveis para as variáveis:

- ABAIXO DO QUARTIL 1
- ENTRE OS QUARTIS 1 E 2
- ENTRE OS QUARTIS 2 E 3
- ACIMA DO QUARTIL 3

A classificação “ABAIXO DO QUARTIL 1” corresponde aos menores percentuais enquanto “ACIMA DO QUARTIL 3” corresponde aos maiores percentuais.

Durante a fase de modelação foi aplicado o algoritmo *Apriori* usando o *Software R* versão 3.1.2. Foram filtradas as regras considerando os seguintes parâmetros:

- Suporte entre 0,01 e 0,02.
- Confiança mínima foi definida em 0,90.
- *Lift* maior ou igual a 5.
- Regras que não apresentaram alguma das variáveis de média de notas também foram excluídas, pois se pretende localizar regras associadas ao rendimento do discente.

Embora regras interessantes possam ser encontradas com parâmetros menos rígidos do que os apresentados, deve-se ressaltar que, nesse caso, obter-se-ia uma quantidade de regras muito grande, inviabilizando a apresentação de todas elas nesse trabalho.

Durante a fase de avaliação, os resultados obtidos na fase de modelação foram discutidos e analisados, o que será detalhado na Seção 4.

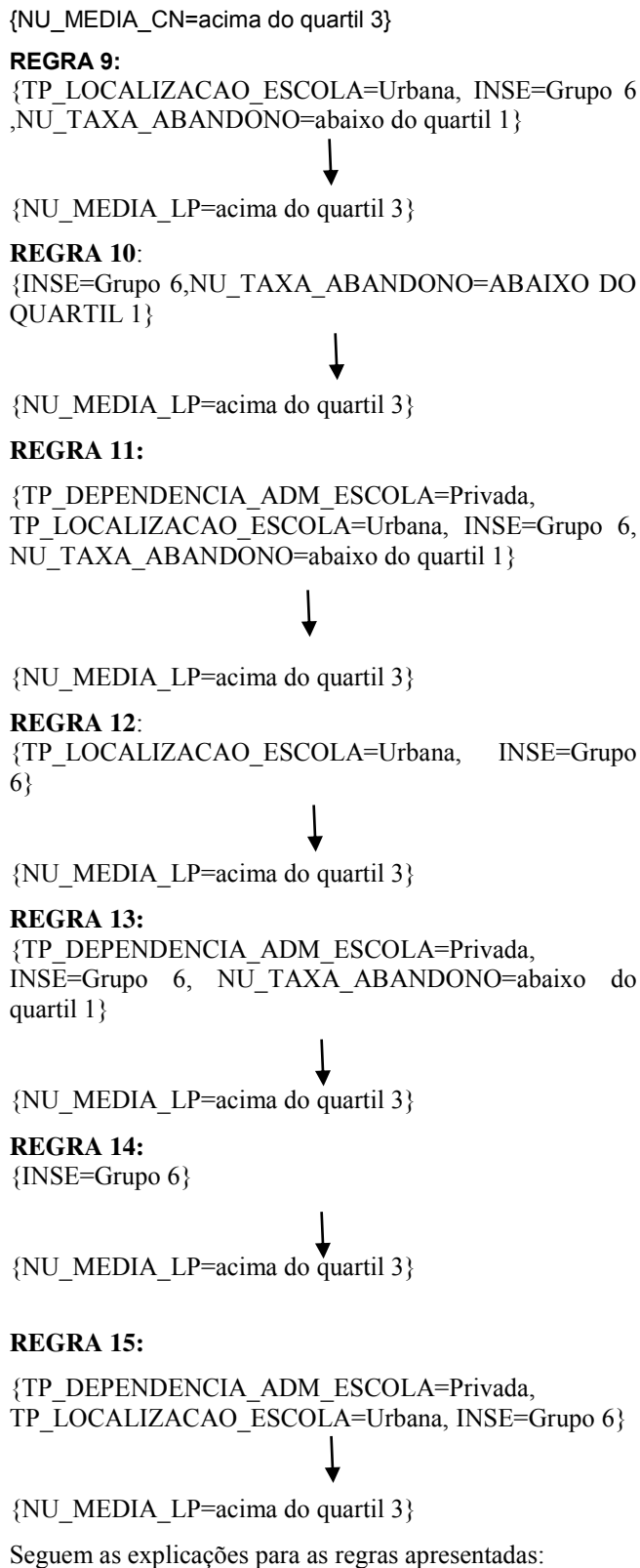
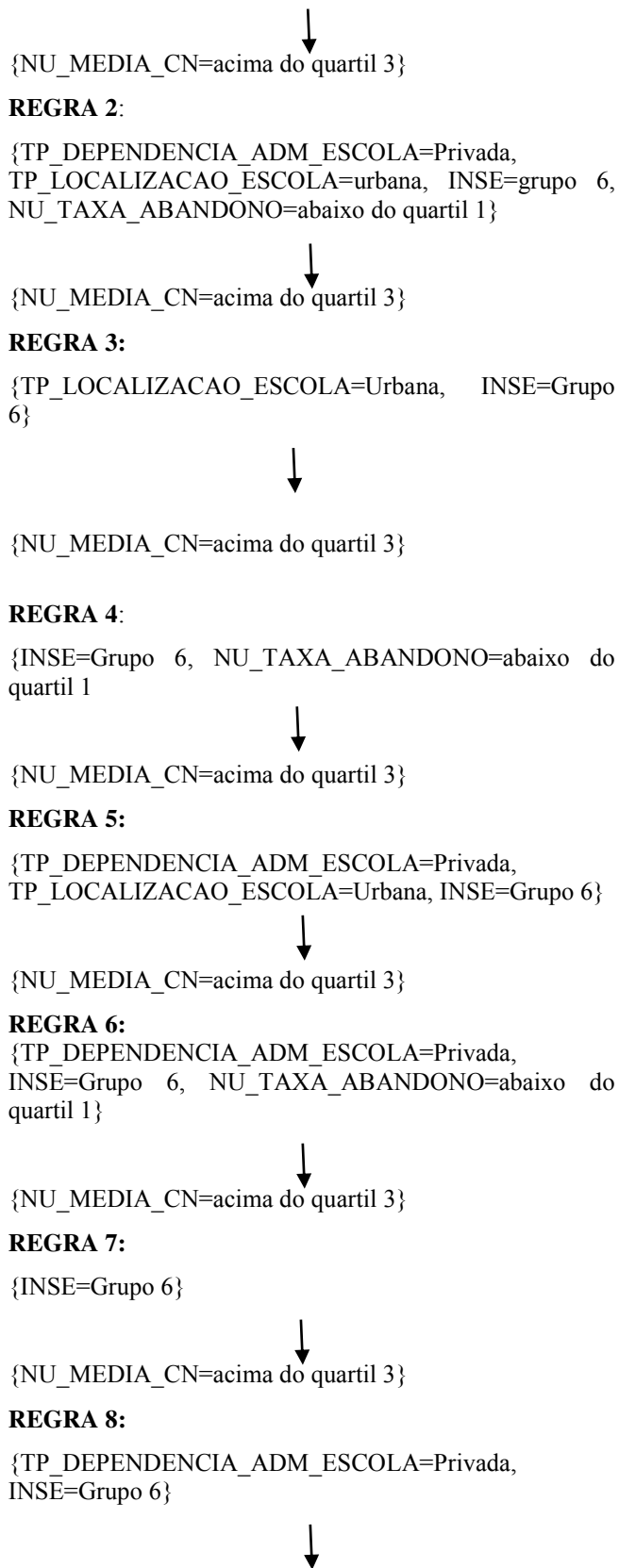
Na fase de desenvolvimento, avaliou-se os resultados obtidos, apresentando sugestões diante dos resultados encontrados, conforme será apresentado na Seção 5.

ESTUDOS REALIZADOS

Com o uso do algoritmo *Apriori*, aplicando-se os parâmetros já mencionados na Seção 3 (Metodologia) foram obtidas um total de 70 regras. Como essa quantidade de regras é inviável para ser apresentada neste artigo, serão exibidas apenas as 15 regras com maior valor obtido para o parâmetro *Lift*, filtrando-se assim apenas aquelas com maior correlação positiva entre as suas partes. Todas as regras a seguir apresentam confiança igual ou superior a 0,90, suporte abaixo de 0,02 e *Lift* superior a 3:

REGRA 1:

```
{TP_LOCALIZACAO_ESCOLA=urbana, INSE=Grupo 6, NU_TAXA_ABANDONO=abaixo do quartil 1}
```



- **REGRA 1:** 98% (valor correspondente à confiança da regra) das escolas de zona urbana, de nível socioeconômico 6 (mais elevado), onde a taxa de abandono

é pequena (abaixo do quartil 1), tiveram média elevada (acima do quartil 3) nas provas de Ciências da Natureza.

- **REGRA 2:** 98% das escolas privadas, localizadas na zona urbana, que pertencem ao grupo socioeconômico 6, onde a taxa de abandono é pequena (abaixo do quartil 1), tiveram média elevada (acima do quartil 3) nas provas de Ciências da Natureza.

- **REGRA 3:** 98% das escolas de zona urbana e que pertencem ao grupo socioeconômico 6 tiveram média elevada (acima do quartil 3) nas provas de Ciências da Natureza.

- **REGRA 4:** 98% das escolas que pertencem ao grupo socioeconômico 6 e onde a taxa de abandono é pequena (abaixo do quartil 1) tiveram média elevada (acima do quartil 3) nas provas de Ciências da Natureza.

- **REGRA 5:** 98% das escolas privadas, localizadas na zona urbana e pertencentes ao grupo socioeconômico 6 tiveram média elevada (acima do quartil 3) nas provas de Ciências da Natureza.

- **REGRA 6:** 98% das escolas privadas, que pertencem ao grupo socioeconômico 6 e onde a taxa de abandono é pequena (abaixo do quartil 1) tiveram média elevada (acima do quartil 3) nas provas de Ciências da Natureza.

- **REGRA 7:** 98% das escolas que pertencem ao grupo socioeconômico 6 tiveram média elevada (acima do quartil 3) nas provas de Ciências da Natureza.

- **REGRA 8:** 98% das escolas privadas que pertencem ao grupo socioeconômico 6 tiveram média elevada (acima do quartil 3) nas provas de Ciências da Natureza.

- **REGRA 9:** 93% das escolas de zona urbana, que pertencem ao grupo socioeconômico 6 e onde a taxa de abandono é pequena (abaixo do quartil 1) tiveram média elevada (acima do quartil 3) nas provas de Linguagens e Códigos.

- **REGRA 10:** 93% das escolas que pertencem ao grupo socioeconômico 6 e onde a taxa de abandono é pequena (abaixo do quartil 1) tiveram média elevada (acima do quartil 3) nas provas de Linguagens e Códigos.

- **REGRA 11:** 93% das escolas privadas, localizadas na zona urbana, que pertencem ao grupo socioeconômico 6 e onde a taxa de abandono é pequena (abaixo do quartil 1) tiveram média elevada (acima do quartil 3) nas provas de Linguagens e Códigos.

- **REGRA 12:** 93% das escolas de zona urbana que , que pertencem ao grupo socioeconômico 6 tiveram média elevada (acima do quartil 3) nas provas de Linguagens e Códigos.

- **REGRA 13:** 93% das escolas privadas que pertencem ao grupo socioeconômico 6 e onde a taxa de abandono é pequena (abaixo do quartil 1) tiveram média elevada (acima do quartil 3) nas provas de Linguagens e Códigos.

- **REGRA 14:** 93% das escolas que pertencem ao grupo socioeconômico 6 tiveram média elevada (acima do quartil 3) nas provas de Linguagens e Códigos.

- **REGRA 15:** 93% das escolas privadas, localizadas na zona urbana e que pertencem ao grupo socioeconômico 6 tiveram média elevada (acima do quartil 3) nas provas de Linguagens e Códigos.

Como se pode verificar, as regras são similares entre si e apontam, de maneira geral, para o fato de que as instituições privadas, classificadas em um nível socioeconômico mais elevado, localizadas em zona urbana e com baixas taxas de evasão, formam alunos que apresentam bom desempenho nas provas do ENEM nas áreas de Ciências da Natureza e de Linguagens e Códigos. Regras similares foram encontradas para as provas das demais áreas de conhecimento.

Como as 15 primeiras regras classificadas só apresentaram resultados associados ao alto rendimento nas provas, convêm apresentar as regras encontradas que estão associadas ao desempenho mais baixo:

REGRA 16:

{TP_DEPENDENCIA_ADM_ESCOLA=Estadual, INSE=, PC_FORMACAO_DOCENTE=abaixo do quartil 1, NU_TAXA_ABANDONO=acima do quartil 3}



{NU_MEDIA_RED=abaixo do quartil 1}

REGRA 17:

{TP_DEPENDENCIA_ADM_ESCOLA=Estadual, TP_LOCALIZACAO_ESCOLA=Urbana, INSE=, PC_FORMACAO_DOCENTE=abaixo do quartil 1, NU_TAXA_ABANDONO=acima do quartil 3}



{NU_MEDIA_RED=abaixo do quartil 1}

REGRA 18:

{INSE=, PC_FORMACAO_DOCENTE=abaixo do quartil 1, NU_TAXA_ABANDONO=acima do quartil 3}



{NU_MEDIA_RED=abaixo do quartil 1}

Seguem as explicações para as regras 16, 17 e 18:

- **REGRA 16:** 90% das escolas estaduais, sem classificação de nível socioeconômico (apenas os dados de 2015 possuem esse classificador), cuja taxa de formação docente é pequena (abaixo do primeiro quartil) e a taxa de abandono é elevada (acima do terceiro quartil) tiveram média baixa (abaixo do quartil 1) nas provas de redação.

- **REGRA 17:** 90% das escolas estaduais, localizadas na zona urbana, sem classificação de nível

socioeconômico, cuja taxa de formação docente é pequena (abaixo do primeiro quartil) e a taxa de abandono é elevada (acima do terceiro quartil) tiveram média baixa (abaixo do quartil 1) nas provas de redação.

- REGRA 18: 90% das escolas sem classificação de nível socioeconômico, cuja taxa de formação docente é pequena (abaixo do primeiro quartil) e a taxa de abandono é elevada (acima do terceiro quartil) tiveram média baixa (abaixo do quartil 1) nas provas de redação

As regras 16, 17 e 18 também são similares entre si e indicam, de forma geral, que escolas de ensino médio mantidas por governos estaduais, com baixo índice de docentes atuando na sua área de formação e com taxas mais elevadas de evasão (se comparadas a outras instituições) formam alunos que apresentam desempenho ruim na prova de redação do ENEM.

CONCLUSÕES

Este artigo mostrou, via utilização do algoritmo *Apriori*, fatores institucionais correlacionados ao bom e ao mau desempenho dos alunos no ENEM. Foram encontradas evidências de que os discentes de instituições privadas apresentam melhores rendimentos neste exame. Isso reforça a necessidade de implementar políticas que promovam a elevação da qualidade de ensino nas instituições públicas, equiparando seu rendimento ao das instituições privadas. Uma dessas políticas deve focar no incentivo à qualificação dos professores, visando atrair para área da docência profissionais mais qualificados e com formação mais específica, visto que, conforme se verificou nas regras 16, 17 e 18, o baixo percentual de formação docente é um dos fatores que contribui para o baixo rendimento dos alunos nas provas de redação.

Verificou-se também, nas regras encontradas, que escolas com níveis mais elevados de evasão também apresentam resultados piores nas avaliações do ENEM. Embora fuja ao escopo desse trabalho determinar as causas que levam aos altos índices de evasão escolar, pode-se assumir que fatores institucionais podem contribuir para esse problema, o que remete novamente à necessidade de promoção de reformas no ensino público.

As instituições de ensino localizadas em áreas urbanas apresentaram melhores rendimentos nas provas do ENEM se relacionadas às localizadas na zona rural. Os fatores relacionados à essa disparidade requerem um estudo mais profundo, mas dentre as possíveis causas, pode-se elencar: infraestrutura mais precária, baixo índice de formação docente nessas instituições, altos níveis de evasão (os jovens podem ter que abandonar a escola precocemente para auxiliar os pais em trabalhos rurais).

Esse trabalho, pelo caráter multifatorial da questão, não esgota todas os fatores que estão relacionados ao bom e ao mau desempenho das instituições de ensino na preparação dos discentes para o ENEM. Contudo, foram apresentadas evidências importantes, que podem servir de subsídios para

trabalhos futuros e conscientização de dirigentes escolares e autoridades políticas, no sentido de desenvolver estratégias que coloquem a educação pública em um patamar de qualidade.

REFERÊNCIAS

1. Alves, Rafael Damiani; Cechinel, Cristian; Queiroga, Emanuel. Predição do desempenho de Matemática e Suas Tecnologias do ENEM utilizando técnicas de Mineração De Dados. Em: Anais dos Workshops do Congresso Brasileiro de Informática na Educação. 2018. p. 469.
2. Amo, Sandra de. Técnicas de mineração de dados, XXIV Congresso da Sociedade Brasileira de Computação, vol. 1–1, Jul/Ago 2004, pp. 43.
3. Batista, Santos Dias; Souza, Alesxsandra Matos; Oliveira, Júlia Mara da Silva. A evasão escolar no ensino médio: um estudo de caso. Revista Profissão Docente, UNIUBE. Uberaba/MG, v. 9, n. 19, 2009.
4. Carmo, Erinaldo Ferreira do; Rocha, Enivaldo Carvalho da; Figueiredo Filho, Dalson Brito; Silva, Lucas Emanuel de Oliveira; Ferreira, Giovana. A ampliação do Indicador de Formação Docente na melhoria do desempenho escolar. Cadernos de Estudos e Pesquisa na Educação Básica, v. 1, n.1, p. 11 -32, 2015.
5. Chapman, P.; Clinton, J.; Kerber, R.; Khabaza, T.; Reinartz, T.; Shearer, C.; Wirth, R. CRISP-DM 1.0: Step-by-step data mining guide. SPSS inc, v. 16, 2000.
6. Fayyad, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. AI magazine, v. 17, n. 3, p. 37-37, 1996.
7. Hoed, Raphael Magalhães. Análise da evasão em cursos superiores: o caso da evasão em cursos superiores da área de Computação. Brasília, DF: Universidade de Brasília, 2016.
8. Pieri, Renan Gomes; SANTOS, Alexandre André. Uma Proposta para o Índice de Infraestrutura Escolar e o Índice de Formação de Professores. Brasília: Inep, 2014. (Textos para Discussão 38) 40 p.
9. Ribeiro, Adriano Cesar. Correlação e visualização de alertas de segurança em redes de computadores. São José do Rio Preto, SP: Universidade Estadual Paulista - Campus de São José do Rio Preto, 2015.
10. Romão, Wesley; Niederauer, Carlos A. P.; Martins, Alejandro; Tcholakian, Aran; Tcholakian, Pacheco, Roberto C. S.; Barcia, Ricardo M. Extração de regras de associação em C&T: O algoritmo Apriori. XIX Encontro Nacional em Engenharia de Produção, v. 34, p. 37-39, 1999.
11. Sampaio, B.; Sampaio, Y.; Mello, E. P. de; Melo, A. S.. Desempenho no vestibular, background familiar e

evasão: evidências da UFPE. *Economia Aplicada*, v. 15, n. 2, p. 287-309, 2011.

12. Silva, Leandro A.; Morino, Anderson Hideki; Sato, Thiago Massahiro Conti. Prática de mineração de dados no exame nacional do ensino médio. Em: *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*. 2014. p. 651.
13. Simon, Augusto; Cazella, Sílvio. Mineração de Dados Educacionais nos Resultados do ENEM de 2015. Em: *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*. 2017. p. 754.
14. Stearns, B.; Rangel, F.; Firmino, F.; Rangel, F.; Oliveira, J.. Prevendo Desempenho dos Candidatos do ENEM Através de Dados Socioeconômicos. Em: *36º Concurso de Trabalhos de Iniciação Científica (CTIC 2017)*. SBC, 2017.
15. Vasconcelos, Lívia Maria Rocha de; Carvalho, Cedric Luiz de. Aplicação de regras de associação para mineração de dados na web. *Revista Telfract*, v. 1, n. 1, 2018.