XXI ENMC
ENCONTRO NACIONAL DE MODELAGEM COMPUTACIONAL
IX ECTM
ENCONTRO DE CIÊNCIA E TECNOLOGIA DE MATERIAIS

08 a 11 de Outubro de 2018
Instituto Federal Fluminense
Búzios - RJ

UTILIZAÇÃO DE UM ALGORITMO DE ASSOCIAÇÃO PARA A DESCOBERTA DE ASPECTOS RELACIONADOS À REPETÊNCIA ESCOLAR

Stella Oggioni da Fonseca[1] - stella.oggioni@gmail.com
Adriana da Rocha Silva[2] - arsilva@iprj.uerj.br
Anderson Amendoeira Namen[3] - aanamen@iprj.uerj.br
[1,2]Universidade do Estado do Rio de Janeiro, Instituto Politécnico – Nova Friburgo, RJ, Brasil
[3]Universidade do Estado do Rio de Janeiro, Instituto Politécnico e Universidade Veiga de Almeida – Nova Friburgo, RJ, Brasil

Resumo. *O presente artigo efetua um estudo exploratório acerca de aspectos relacionados à repetência escolar dos alunos do 9º ano do ensino fundamental, residentes no Estado do Rio de Janeiro. Por intermédio dos dados coletados pela Prova Brasil, analisam-se características ligadas aos discentes que estão associadas à repetência. Para a extração destas informações, foram conduzidas etapas de limpeza e tratamento dos dados e a aplicação do algoritmo de associação Apriori, utilizado na etapa de mineração de dados. São apresentadas conclusões a respeito do modelo gerado e dos resultados obtidos.*

Palavras-chave: *Repetência Escolar, Dados Educacionais, Prova Brasil, Mineração de Dados, Algoritmo Apriori.*

1. INTRODUÇÃO

A Prova Brasil é uma avaliação desenvolvida pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) com a finalidade de efetuar um diagnóstico da qualidade do ensino básico ofertado nas escolas públicas. A primeira edição da Prova Brasil ocorreu em 2005 e, desde então, acontece bienalmente. Em 2013, ano em que os dados utilizados no presente trabalho foram coletados, esta avaliação foi composta de testes de Língua Portuguesa e Matemática, bem como de questionários visando à obtenção de informações a respeito dos discentes, professores, diretores e escolas (Inep, 2015).

Por intermédio dos dados provenientes da aplicação da Prova Brasil, pesquisas vêm sendo conduzidas com o intuito de obter possíveis explicações para os fenômenos educacionais e, consequentemente, nortear melhorias no processo ensino-aprendizagem.

Diante desta perspectiva, o presente trabalho apresenta resultados iniciais referentes à mineração de dados relacionados aos alunos do 9º ano do ensino fundamental, inseridos em escolas do Estado do Rio de Janeiro. Segundo Fayyad & Piatetsky-Shapiro & Smyth (1996), a mineração de dados consiste na aplicação de algoritmos capazes de extrair informações embutidas em

grandes volumes de dados. Para a descoberta de conhecimento, a mineração engloba conceitos advindos da Estatística e Aprendizagem de Máquina.

O objetivo do trabalho é aplicar o algoritmo *Apriori* para mineração das regras de associação entre as características dos discentes e a repetência escolar. Mais especificamente, por meio das respostas dadas às perguntas presentes no questionário, busca-se identificar as variáveis que caracterizam os estudantes que poderiam estar associadas à repetência.

O artigo apresenta, inicialmente, as tarefas de seleção, limpeza e tratamento dos dados que foram efetuadas. Posteriormente, são abordados, de forma sucinta, os conceitos do algoritmo *Apriori*, responsável pela descoberta de associações entre as variáveis. Finalmente, são expostos os resultados iniciais obtidos e as conclusões extraídas.

2. BASE DE DADOS E METODOLOGIA

Os dados da Prova Brasil são de acesso público e podem ser encontrados no *site* do Inep (vide www.inep.gov.br). Para cumprir o objetivo exposto no escopo deste artigo, foi efetuado o *download* do arquivo intitulado TS_ALUNO_9EF. Este arquivo contém 2720588 registros correspondentes, em âmbito nacional, aos alunos do 9º ano do ensino fundamental, com 92 campos que compreendem diversos atributos identificadores (informações como turma, escola e estado em que os alunos estão inseridos), as notas obtidas nos testes de Matemática e Língua Portuguesa, bem como as respostas dadas às 57 perguntas presentes no questionário.

O questionário preenchido pelos estudantes captura informações acerca do perfil socioeconômico, cotidiano, hábitos de leitura e estudo, trajetória escolar, incentivo familiar e suas percepções com relação à escola. Além disso, versa sobre a repetência, ou seja, o aluno indica se já foi reprovado. Conforme mencionado, esta última informação é central para o presente estudo.

As etapas conduzidas para a descoberta de conhecimento são apresentadas na Fig. 1.

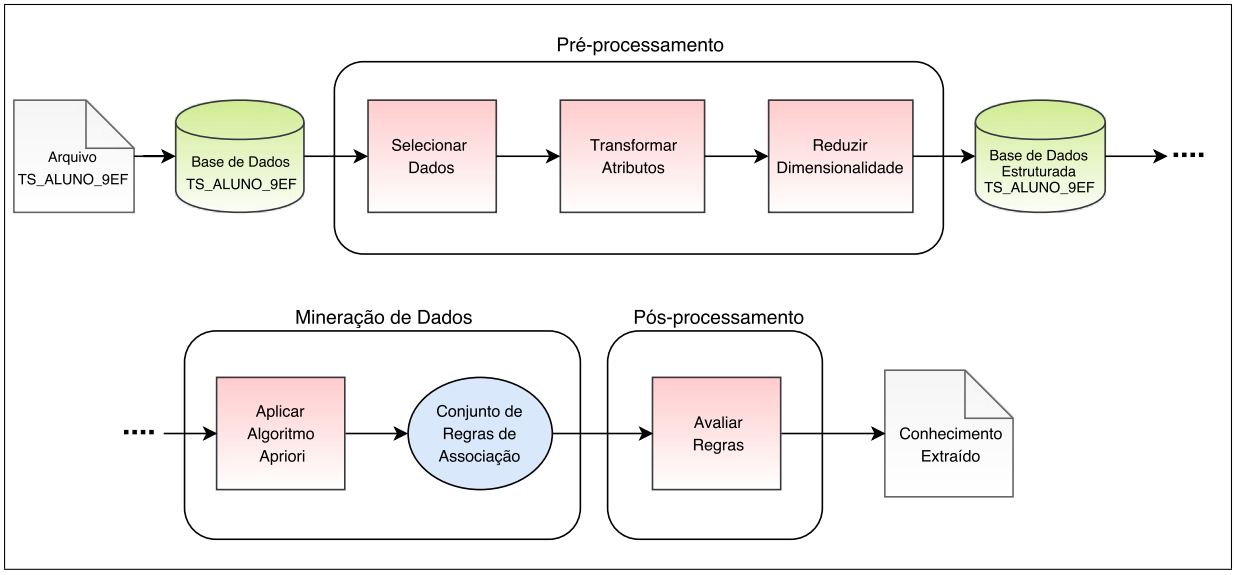

Figura 1- Metodologia utilizada para a descoberta de conhecimento

A metodologia apresentada na Fig. 1 consiste de três etapas centrais: pré-processamento, mineração de dados e pós-processamento. Tais etapas serão descritas nas seções posteriores.

3. PRÉ-PROCESSAMENTO

O propósito do pré-processamento é selecionar os dados e transformá-los em um formato apropriado para futuras análises (Tan & Steinbach & Kumar, 2009). Assim, nesta etapa são executadas tarefas como remoção de dados faltantes, transformação de atributos e seleção de variáveis com o intuito de reduzir a dimensionalidade.

3.1 Seleção dos dados

Para executar as tarefas de seleção e manipulação dos dados, o arquivo TS_ALUNO_9EF foi importado para o *software PostgreSQL* (The PostgreSQL Global Development Group, 1995), que é um sistema gerenciador de banco de dados *open source* e que suporta tabelas com extensão *.csv* (*Comma-Separated Values*), formato em que se encontra o arquivo mencionado. Por meio de códigos escritos em SQL (*Structured Query Language*) no *PostgreSQL*, foi possível realizar diversas operações relacionadas aos registros (que compõem as linhas de uma tabela) e aos atributos (que compõem as colunas de uma tabela).

Inicialmente foram selecionados os alunos residentes no Estado do Rio de Janeiro. Esta identificação foi efetuada por meio do atributo numérico ID_UF. Como o código correspondente ao Rio de Janeiro é o número 33, foram selecionados os registros de alunos cujo ID_UF alocava o referido valor. Após esta seleção, restaram para o estudo 157602 estudantes.

Como se objetivou analisar as respostas dadas pelos discentes, foram selecionados somente os alunos que efetuaram o preenchimento do questionário. Esta identificação ocorre por intermédio do atributo IN_PREENCHIMENTO_QUESTIONARIO. O conjunto de dados foi reduzido para 116441 registros. Finalmente, foi decidido manter na base de dados somente os discentes que responderam a, no mínimo, 70% das 57 questões. Portanto, restaram 113300 registros de alunos para as análises posteriores.

3.2 Transformação de atributos

Após a etapa de seleção, todos os 57 atributos do arquivo TS_ALUNO_9EF, que correspondem às perguntas do questionário, tiveram seus valores alterados, pois em seu conteúdo havia somente um caractere que representava a resposta escolhida (letra “A”, “B”, “C”, ...). Esses caracteres foram alterados para um *string*, contendo a resposta correspondente à letra, de modo que os resultados obtidos pela mineração pudessem ter mais fácil visualização, propiciando melhor entendimento por parte do usuário.

Foi necessário, ainda, analisar a questão a respeito da repetência. Esta pergunta tem o seguinte enunciado:

> Questão 48) Você já foi reprovado?
> A - Não;
> B - Sim, uma vez;
> C - Sim, duas vezes ou mais.

A partir desta questão, foi criado o atributo categórico *Repetência*. Logo, caso o estudante tivesse respondido a alternativa “A”, seria indicado como não repetente. As alternativas “B” e “C” foram agregadas, uma vez que ambas expressam a situação de repetência do discente. Ademais, 848 alunos não responderam à esta questão e, portanto, foram removidos da base de dados. Após esta remoção, permaneceram 112452 registros na tabela TS_ALUNO_9EF.

Informações sobre as categorias da variável *Repetência* são apresentadas na Tabela 1.

Tabela 1- Categorias da variável alvo *Repetência*

Questão 48: Você já foi reprovado?	**Repetência**	**Nº de Alunos**	**Média no teste de Matemática**	**Média no teste de Língua Portuguesa**
"A" - Não	Não	72777	254,98	250,59
"B" - Sim, uma vez "C" - Sim, duas vezes ou mais	Sim	39675	229,83	223,40

A Tabela 1 mostra que cerca de 35% do total de alunos presentes na base já foram reprovados. Ela apresenta, ainda, a média obtida nos testes das disciplinas de Matemática e Língua Portuguesa aplicados pela Prova Brasil 2013. É importante mencionar que a escala de proficiência varia de 0 a 500. Nota-se uma diferença de aproximadamente 25 e 27 pontos, respectivamente, em Matemática e Língua Portuguesa, entre alunos não repetentes e repetentes.

Após estas análises, se objetivou buscar relações entre as variáveis que alocavam as respostas dadas às perguntas presentes no questionário e a variável alvo criada, denominada *Repetência*, com as categorias "Sim" e "Não".

3.3 Redução de dimensionalidade

Conforme mencionado, o questionário preenchido pelos estudantes é composto por 57 perguntas. Para reduzir a dimensionalidade, foram selecionadas algumas variáveis para o processo de mineração de dados. A seleção destas variáveis foi efetuada por uma medida denominada Incerteza Simétrica, que baseia-se no conceito de entropia (Quinlan, 1986).

Seja $\mathcal{D} = \{X_1, X_2, \ldots, X_n, C\}$, com $n \geq 1$, o conjunto com $n+1$ atributos de uma base de dados, onde C é o atributo alvo composto de $\{c_1, c_2, \ldots, c_m\}$, com $m \geq 2$, categorias. Seja ainda um atributo arbitrário $X_l \in \mathcal{D}$, com $l = 1, \ldots, n$, cujos valores são $\{x_1^l, x_2^l, \ldots, x_q^l\}$, sendo $q \geq 1$. Define-se a entropia, denotada por H, do atributo X_l como:

$$H(X_l) = -\sum_{i=1}^{q} p_i \log_2(p_i), \tag{1}$$

onde p_i é a razão entre o número de registros da base em que ocorre o valor x_i^l e o número total de registros.

Pode-se estender os conceitos e definir a entropia condicional da variável alvo C, dado o atributo X_l, uma vez que precisa-se avaliar o grau de associação entre o atributo e a variável alvo. Então, sabendo-se que $p_{j|i}$ é a razão entre o número de registros que pertencem à categoria c_j em que ocorre o valor x_i^l do atributo X_l, e o número total de registros da base, a entropia condicional é dada por:

$$H(C|X_l) = -\sum_{i=1}^{q}\sum_{j=1}^{m} p_{j|i} \log_2\left(\frac{p_{j|i}}{p_i}\right). \tag{2}$$

Quanto menor for o valor de $H(C|X_l)$, mais informativo um atributo será em relação ao alvo.

Por meio destas definições, formulou-se uma expressão que avalia a qualidade do atributo, expressão esta denominada Ganho de Informação (*Information Gain*) (Quinlan, 1986). O ganho é obtido subtraindo-se a entropia do atributo alvo pela entropia condicional. Em termos matemáticos, tem-se:

$$Ganho(C|X_l) = H(C) - H(C|X_l). \tag{3}$$

Quanto maior o valor do ganho ou, equivalentemente, quanto menor o valor da entropia condicional, uma vez que a informação do atributo alvo $H(C)$ é fixa, melhor será a qualidade do atributo. Portanto, a relação $H(C) - H(C|X_l)$ indica a quantidade de informação que o atributo X_l traz para a discriminação do atributo alvo C. No entanto, segundo Quinlan (1986), este critério tende a selecionar atributos com um número maior de valores. Para contornar este problema, pode-se normalizar a Eq. (3), o que resulta na medida Incerteza Simétrica.

A medida Incerteza Simétrica divide o ganho pela soma entre a entropia do atributo e a entropia do atributo alvo. O valor resultante é ainda multiplicado pelo fator 2 para que os valores desta métrica recaiam no intervalo $[0, 1]$. Portanto, o critério Incerteza Simétrica, denotado por U, é dado por:

$$U(C|X_l) = 2\left[\frac{Ganho(C|X_l)}{H(C) + H(X_l)}\right] = 2\left[\frac{H(C) - H(C|X_l)}{H(C) + H(X_l)}\right]. \quad (4)$$

Quando esta expressão assume o valor máximo 1, indica que o atributo X_l está completamente correlacionado com o atributo alvo C. Por outro lado, se a expressão resulta no valor mínimo 0, indica que estes atributos são independentes, ou seja, não há qualquer correlação.

No presente trabalho foi utilizado o *software* Weka (The University of Waikato, 1999) para computar os valores da medida Incerteza Simétrica. Este *software* é de código aberto e possui uma série de critérios de seleção de atributos e algoritmos de mineração de dados. Mais detalhes desta ferramenta podem ser vistos em Witten & Frank & Hall (2011).

Ao computar a medida Incerteza Simétrica para cada variável presente na base de dados TS_ALUNO_9EF, identificou-se as mais correlacionadas com o atributo alvo *Repetência*. A Tabela 2 apresenta as oito questões ranqueadas de acordo com a medida. As variáveis foram intituladas pelos autores do presente artigo de modo a facilitar a interpretação dos resultados posteriores.

Tabela 2- Variáveis ranqueadas pela medida Incerteza Simétrica

#	Questão	Enunciado	Valores	Variável
1	49	Você já abandonou a escola durante o período de aulas e ficou fora da escola o resto do ano?	A - Não; B - Sim, uma vez; C - Sim, duas vezes ou mais.	Q49_AbandonouEscola
2	45	Atualmente você trabalha fora de casa (recebendo ou não salário)?	A - Sim; B - Não.	Q45_TrabalhaFora
3	57	Sua pretensão após terminar o 9º ano.	A - Somente continuar estudando; B - Somente trabalhar; C - Continuar estudando e trabalhar; D - Ainda não sei.	Q57_Pretensao
4	01	Qual é o seu sexo?	A - Masculino; B - Feminino.	Q01_Sexo
5	26	Com qual frequência seus pais, ou responsáveis por você, vão à reunião de pais?	A - Sempre ou quase sempre; B - De vez em quando; C - Nunca ou quase nunca.	Q26_PaisVaoAReuniao
6	33	Com qual frequência você lê livros em geral?	A - Sempre ou quase sempre; B - De vez em quando; C - Nunca ou quase nunca.	Q33_LeLivros
7	43	Em dia de aula, quanto tempo você gasta assistindo à TV, navegando na internet ou jogando jogos eletrônicos?	A - Menos de 1 hora; B - Entre 1 e 2 horas; C - Mais de 2 horas, até 3 horas; D - Mais de 3 horas; E - Não vejo.	Q43_TempoVendoTVInternet
8	54	Você faz o dever de casa de Matemática?	A - Sempre ou quase sempre; B - De vez em quando; C - Nunca ou quase nunca; D - Não passam dever.	Q54_FazDeverDeMatematica

Foi possível identificar que a variável nomeada como Q49_AbandonouEscola, correspondente à pergunta de número 49 no questionário, é a que melhor permite a discriminação entre as categorias presentes no atributo alvo. A segunda melhor é a variável Q45_TrabalhaFora, e assim sucessivamente. Portanto, no processo de mineração de dados, buscaram-se relações entre as oito variáveis apresentadas na Tabela 2 e a variável alvo *Repetência*.

4. DESCOBERTA DE ASSOCIAÇÕES: MINERAÇÃO DE DADOS

No processo de mineração de dados foi aplicada uma metodologia conhecida como análise de associação, útil para descobrir relacionamentos embutidos em grandes conjuntos de dados. Os relacionamentos encontrados podem ser representados na forma de regras de associação.

Uma regra de associação tem o formato $\mathcal{A} \longrightarrow \mathcal{B}$, onde $\mathcal{A}$ e $\mathcal{B}$ são conjuntos disjuntos de itens (atributos, com seus respectivos valores), isto é, $\mathcal{A} \cap \mathcal{B} = \emptyset$. $\mathcal{A}$ é o antecedente e $\mathcal{B}$ é o consequente da regra. Além disso, a força de uma regra pode ser avaliada em termos de duas medidas: o suporte e a confiança.

Inicialmente, defina-se suporte de um conjunto arbitrário $\mathcal{G}$ como:

$$sup(\mathcal{G}) = \frac{\sigma(\mathcal{G})}{N}, \tag{5}$$

onde $\sigma(\cdot)$ é o número total de ocorrências de registros contendo o conjunto de itens, no caso, $\mathcal{G}$, e N é o número total de registros na base de dados.

Assim, o suporte de uma regra $\mathcal{A} \longrightarrow \mathcal{B}$ determina a frequência com que um conjunto de itens $\mathcal{A} \cup \mathcal{B}$ ocorre, ou seja, é o percentual de ocorrências de registros que contêm todos os itens para os quais uma regra é aplicável. A confiança determina a frequência na qual os itens de $\mathcal{B}$ aparecem em ocorrências que contenham $\mathcal{A}$. Em outras palavras, a confiança não trabalha com todas as ocorrências, apenas com as que possuem o antecedente da regra. Para uma determinada regra $\mathcal{A} \longrightarrow \mathcal{B}$, quanto maior a confiança, maior a probabilidade de que $\mathcal{B}$ esteja presente em ocorrências que contenham $\mathcal{A}$.

Em termos matemáticos, o suporte e confiança são definidos, respectivamente, como:

$$sup(\mathcal{A} \longrightarrow \mathcal{B}) = sup(\mathcal{A} \cup \mathcal{B}) \qquad \text{e} \qquad conf(\mathcal{A} \longrightarrow \mathcal{B}) = \frac{sup(\mathcal{A} \longrightarrow \mathcal{B})}{sup(\mathcal{A})}.$$

O objetivo da mineração de regras de associação é gerar todas as regras possíveis que satisfaçam os valores mínimos de suporte e de confiança determinados pelo usuário. O problema, portanto, é decomposto em dois subproblemas:

1. Gerar todos os conjuntos de itens que possuem suporte maior do que um limite mínimo especificado. Esses conjuntos são chamados de conjuntos de itens frequentes;

2. Para cada conjunto de itens frequentes, gerar todas as regras que possuem confiança maior que um valor de confiança mínimo.

Para solucionar estes subproblemas, foi utilizado o algoritmo *Apriori* (Agrawal & Srikant, 1994). Este algoritmo realiza a etapa de mineração de dados em dois passos. No primeiro, é feita uma varredura sobre a base de dados de entrada, a fim de gerar todos os conjuntos de combinações de itens que satisfaçam um valor maior do que o suporte mínimo definido pelo usuário. No segundo passo, são extraídas todas as regras de alta confiança dos conjuntos gerados. Estas regras, que satisfazem o limitante especificado da medida confiança, são chamadas

de regras fortes. Detalhes do algoritmo *Apriori* podem ser encontrados em Han & Kamber & Pei (2012) e Tan & Steinbach & Kumar (2009).

No presente trabalho, foi utilizada uma implementação do algoritmo *Apriori* disponibilizada dentro do *software* Weka.

4.1 Outras medidas para avaliar as regras de associação

Outras medidas para avaliar as regras devem ser definidas, uma vez que a confiança apresenta algumas limitações e pode originar uma situação conhecida como armadilha de confiança. Para exemplificar essa questão, é apresentado um exemplo clássico, no contexto da análise de associação, retirado de Tan & Steinbach & Kumar (2009). Os autores abordam que mesmo valores significativos de confiança podem não identificar uma regra relevante. A Tabela 3 mostra uma situação onde se busca analisar o relacionamento entre pessoas que bebam chá e café.

Tabela 3- Preferências de bebida em um grupo de 1000 pessoas

	Bebe café	**Não bebe café**	**Total**
Bebe chá	150	50	200
Não bebe chá	650	150	800
Total	800	200	1000

Fonte: Tan & Steinbach & Kumar (2009, p. 445).

Nota-se que, a princípio, a regra $\{\text{Bebe chá}\} \longrightarrow \{\text{Bebe café}\}$ poderia ser considerada relevante, pois seus valores de suporte (15%) e confiança (75%) são relativamente altos. No entanto, pode-se observar que 80 por cento do número total de pessoas bebe café independente de beberem chá, enquanto o percentual de pessoas que bebem chá e café é 75 por cento. Assim, a regra de associação $\{\text{Bebe chá}\} \longrightarrow \{\text{Bebe café}\}$ é ilusória, já que o fato de uma pessoa beber chá, na realidade, diminui a possibilidade de que beba café de 80 para 75 por cento.

Na Tabela 4 são apresentadas outras medidas que possuem informação adicional, evitando situações enganosas, como a mencionada acima. Mais detalhes acerca das medidas apresentadas podem ser vistas em Tan & Steinbach & Kumar (2009) e Hahsler (2015).

Tabela 4- Medidas para avaliar uma regra $\mathcal{A} \longrightarrow \mathcal{B}$

Medida	**Definição**	**Intervalo da Medida**	**Descrição**
Lift	$\frac{conf(\mathcal{A} \longrightarrow \mathcal{B})}{sup(\mathcal{B})}$	$[0, \infty)$	< 1 ($\mathcal{A}$ e $\mathcal{B}$ são relacionados negativamente) $= 1$ ($\mathcal{A}$ e $\mathcal{B}$ são independentes) > 1 ($\mathcal{A}$ e $\mathcal{B}$ são relacionados positivamente)
Leverage	$sup(\mathcal{A} \longrightarrow \mathcal{B}) - (sup(\mathcal{A})sup(\mathcal{B}))$	$[-0,25, 0,25]$	< 0 ($\mathcal{A}$ e $\mathcal{B}$ são relacionados negativamente) $= 0$ ($\mathcal{A}$ e $\mathcal{B}$ são independentes) > 0 ($\mathcal{A}$ e $\mathcal{B}$ são relacionados positivamente)
Conviction	$\frac{1-sup(\mathcal{B})}{1-conf(\mathcal{A} \longrightarrow \mathcal{B})}$	$[0, \infty)$	< 1 ($\mathcal{A}$ e $\mathcal{B}$ são relacionados negativamente) $= 1$ ($\mathcal{A}$ e $\mathcal{B}$ são independentes) > 1 ($\mathcal{A}$ e $\mathcal{B}$ são relacionados positivamente)

Os valores das medidas suporte e confiança são calculados durante a execução do algoritmo *Apriori* com o intuito de gerar as regras. Já as medidas definidas na Tabela 4 foram computadas pelos autores do presente artigo. O cálculo foi implementado em linguagem SQL, ao importar as regras para o *PostgreSQL*.

5. MODELO GERADO E PÓS-PROCESSAMENTO

Por meio da base de dados estruturada, composta por 112452 registros de alunos, as oito variáveis apresentadas na Tabela 2, com seus respectivos valores, e o atributo alvo *Repetência*, com as categorias "Sim" e "Não", procurou-se identificar o perfil dos alunos do 9º ano do ensino fundamental que eram repetentes.

Para a execução do algoritmo *Apriori*, foram definidos como valores mínimos o suporte de 0,5% e confiança de 70%. Cabe salientar que o pequeno valor de suporte mínimo especificado objetivou a identificação de combinação de itens que, apesar de não tão frequentes, pudessem ter grande relevância (as medidas computadas revelam a relevância das regras encontradas). Ademais, como tem-se um grande volume de dados (112452 registros), a utilização de um suporte de 0,5% ainda possibilitou a extração de regras relacionadas a um grupo de estudantes significativo. Mais especificamente, com este suporte, as regras que foram obtidas envolveram conjuntos de pelo menos 562 registros de alunos com características idênticas.

O modelo gerado pelo algoritmo *Apriori* ficou composto por 27 regras com o consequente "Repetência = Sim". As primeiras 10 regras, ordenadas decrescentemente pela confiança, são apresentadas na Tabela 5.

Tabela 5- Regras com o consequente indicando a repetência e os valores das medidas

Regra	Sup	Conf	Lift	Leverage	Conviction
{Q45_TrabalhaFora=Sim, Q49_AbandonouEscola=Sim uma vez, Q57_Pretensao=Continuar estudando e trabalhar} 799 ⟶ {Repetencia=Sim} 616	0,005478	0,77	2,182433	0,002971	2,813838
{Q01_Sexo=Masculino, Q45_TrabalhaFora=Sim, Q49_AbandonouEscola=Sim uma vez} 769 ⟶ {Repetencia=Sim} 591	0,005256	0,77	2,182433	0,002843	2,813838
{Q45_TrabalhaFora=Sim, Q49_AbandonouEscola=Sim uma vez} 1171 ⟶ {Repetencia=Sim} 884	0,007861	0,75	2,125747	0,004187	2,588731
{Q01_Sexo=Masculino, Q49_AbandonouEscola=Sim uma vez, Q57_Pretensao=Continuar estudando e trabalhar} 1357 ⟶ {Repetencia=Sim} 1007	0,008955	0,74	2,097403	0,004697	2,489165
{Q26_PaisVaoAReuniao=De vez em quando, Q49_AbandonouEscola=Sim uma vez, Q57_Pretensao=Continuar estudando e trabalhar} 1106 ⟶ {Repetencia=Sim} 820	0,007292	0,74	2,097403	0,003822	2,489165
{Q49_AbandonouEscola=Sim uma vez, Q54_FazDeverDeMatematica=Sempre ou quase sempre, Q57_Pretensao=Continuar estudando e trabalhar} 1091 ⟶ {Repetencia=Sim} 800	0,007114	0,73	2,06906	0,003691	2,396973
{Q33_LeLivros=De vez em quando, Q49_AbandonouEscola=Sim uma vez, Q57_Pretensao=Continuar estudando e trabalhar} 1413 ⟶ {Repetencia=Sim} 1023	0,009097	0,72	2,040717	0,004664	2,311367
{Q49_AbandonouEscola=Sim uma vez, Q57_Pretensao=Continuar estudando e trabalhar} 2803 ⟶ {Repetencia=Sim} 2027	0,018025	0,72	2,040717	0,009231	2,311367
{Q01_Sexo=Masculino, Q49_AbandonouEscola=Sim uma vez, Q54_FazDeverDeMatematica=Sempre ou quase sempre} 937 ⟶ {Repetencia=Sim} 677	0,00602	0,72	2,040717	0,003081	2,311367
{Q26_PaisVaoAReuniao=De vez em quando, Q45_TrabalhaFora=Não, Q49_AbandonouEscola=Sim uma vez, Q57_Pretensao=Continuar estudando e trabalhar} 785 ⟶ {Repetencia=Sim} 567	0,005042	0,72	2,040717	0,002579	2,311367

Nas regras apresentadas, o número apresentado antes do símbolo "⟶" indica a quantidade de estudantes que assinalaram as mesmas alternativas nas questões expostas. Considerando a regra $\mathcal{A} \longrightarrow \mathcal{B}$, esta quantidade é o $\sigma(\mathcal{A})$. O número posterior à variável alvo corresponde à quantidade de discentes para o qual o consequente da regra se aplica. Este valor é o $\sigma(\mathcal{A} \cup \mathcal{B})$.

Todas as regras possuíram valores da medida *Lift* e *Conviction* superiores a um, bem como valores da medida *Leverage* maiores do que zero. Portanto, a relevância exposta pela confiança foi confirmada.

Ao analisar as regras geradas apresentadas na Tabela 5, alguns aspectos puderam ser identificados. Observa-se que foi possível corroborar a relação entre repetência e evasão. Diversas pesquisas apontam que a repetência é diretamente responsável pela evasão escolar e pela defasagem em relação à idade-série (Leon & Menezes-Filho, 2002).

Outro aspecto importante a ser citado consiste na situação econômica dos discentes. Nota-se que no antecedente das regras há a ocorrência de alunos que trabalham fora de casa e que pretendem continuar trabalhando após o término do 9º ano. Neste sentido, Ortigão & Aguiar (2013), ao analisarem os dados da Prova Brasil 2009, relacionados aos discentes do 5º ano do ensino fundamental, puderam concluir que os alunos provenientes de famílias de poder econômico médio e alto apresentam menor risco de reprovação quando comparados aos de famílias com baixas condições econômicas.

É importante destacar, ainda, a frequência de leitura e o acompanhamento dos pais em reuniões escolares. Observa-se que nas regras foram vistas regularidades medianas em ambos os fatores. Além disso, as diferenças em relação ao gênero têm sido apontadas pela literatura educacional, onde se enfatiza melhorias nos resultados obtidos por meninas em relação aos meninos (Madeira & Rodrigues, 1998). Ainda sobre esta questão, Baudelot & Establet (1991) abordam as possíveis causas para esta conclusão. Segundo os autores, as meninas têm melhor adaptação no ambiente escolar, uma vez que possuem um comportamento mais ordeiro decorrente de uma socialização primária e familiar. Ressaltam também que meninas possuem mais organização e disciplina, características essenciais para alcançar resultados positivos na trajetória escolar.

6. CONCLUSÃO

O presente trabalho extraiu aspectos relacionados à repetência de alunos do 9º ano do ensino fundamental, residentes no Rio de Janeiro. Para a extração destas informações, foram executadas tarefas de pré-processamento, essenciais para garantir a confiabilidade dos resultados posteriores.

Com o trabalho aqui desenvolvido, pode-se constatar o potencial da utilização de algoritmos de mineração para a descoberta de conhecimento em bases de dados educacionais. Mais especificamente, por intermédio do algoritmo *Apriori*, foram detectadas associações que podem fomentar a discussão sobre o perfil de estudantes com maiores dificuldades de aprendizagem. Na reportagem de Fajardo (2018), são apresentados importantes dados e argumentos a respeito da repetência escolar. No contexto econômico, o Brasil despendeu dos cofres públicos cerca de R$ 16 bilhões ao reprovar em 2016 aproximadamente 3 milhões de alunos da educação básica, o equivalente a 10,26% do total de estudantes da rede pública. Além dos aspectos financeiros, é amplamente discutido o custo social, uma vez que a repetência atinge principalmente os alunos com menor poder aquisitivo. Consequentemente, reforça a desigualdade social. Ao ser entrevistado por Fajardo (2018), Cesar Callegari, membro do Conselho Nacional da Educação (CNE), aponta que tais custos não podem implicar em aprovação automática, uma vez que deve existir um comprometimento com o efetivo aprendizado. Por outro lado, Callegari defende que a repetência deve estar aliada com estratégias complementares para não gerar a evasão.

É notório, portanto, que pesquisas relacionadas à repetência escolar devem ser conduzidas para fomentar a discussão e nortear os tomadores de decisões. O presente trabalho apresentou resultados iniciais e terá continuidade com a execução de outras simulações. Pretende-se, como trabalhos futuros, enriquecer a discussão acerca dos resultados encontrados e explorar outras questões provenientes dos questionários que coletam informações a respeito dos professores,

diretores e escolas, uma vez que tais dados também são disponibilizados no *site* do Inep.

Agradecimentos

O presente trabalho foi realizado com o apoio financeiro da FAPERJ.

REFERÊNCIAS

Agrawal, R.; Srikant, R. (1994), "Fast Algorithms for Mining Association Rules". In: *20th INT'L CONFERENCE ON VERY LARGE DATABASES, 1994*. Proceedings. Santiago.

Baudelot, C.; Establet, R. (1991), *"Allez les filles"*. Paris: PUF.

Fajardo, V. (2018). *"Brasil gasta R$ 16 bilhões com reprovação de 3 milhões de alunos em 2016, aponta levantamento"*. Disponível em:< https://g1.globo.com/educacao/noticia/brasil-gasta-r-16-bilhoes-com-reprovacao-de-3-milhoes-de-alunos-em-2016-aponta-levantamento.ghtml>. Acesso em: 24 de mar. 2018.

Fayyad, U.; Piatetsky-Shapiro, G.; Smyth, P. (1996), "From data mining to knowledge discovery in databases", *AI Magazine, American Association for Artificial Intelligence*, California, USA, v. 17, n. 3, p. 37-54.

Hahsler, M. (2015), *"A Probabilistic Comparison of Commonly Used Interest Measures for Association Rules"*. Disponível em: <http://michael.hahsler.net/research/association_rules/measures.html>. Acesso em: 23 de mar. 2018.

Han, J.; Kamber, M.; Pei, J. (2012), *"Data Mining: Concepts and Techniques"*. 3ª ed., Waltham, USA: Morgan Kaufmann Publishers.

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

Leon, F. L. L. de; Menezes-Filho, N. A. (2002). "Reprovação, avanço e evasão escolar no Brasil". *Pesquisa e Planejamento Econômico*, Rio de Janeiro, v. 32, n. 3, p. 417-451.

Madeira, F.; Rodrigues, E. M. (1998). "Recado dos jovens: mais qualificação". In: COMISSÃO NACIONAL DE POPULAÇÃO E DESENVOLVIMENTO. *Jovens acontecendo na trilha das políticas públicas*. Brasília: CPND, 1998. v. 2, p. 427-498.

Ortigão, M. I. R.; Aguiar, G. S. (2013). "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.

Quinlan, J. R. (1986), "Induction of decision trees". *Machine Learning*, Kluwer Academic Publishers, Boston, USA, v. 1, n. 1, p. 81-106.

Tan, P.; Steinbach, M.; Kumar, V. (2009), *"Introdução ao Data Mining: Mineração de Dados"*. Rio de Janeiro: Editora Ciência Moderna Ltda.

The PostgreSQL Global Development Group. *"PostgreSQL Database Management System"* (Versão 9.2). 1995. Programa de Computador. Disponível em: <http://www.postgresql.org.br>.

The University of Waikato. *"Weka"* (Versão 3.8.2). 1999. Programa de Computador. Disponível em: <http://www.weka.org.br>.

Witten, I. H.; Frank, E.; Hall, M. (2011), *"Data Mining: Practical Machine Learning Tools and Techniques"*. USA: Morgan Kaufmann Publishers Inc., San Francisco.

USE OF AN ASSOCIATION ALGORITHM FOR DISCOVERING ASPECTS RELATED TO SCHOOL REPETITION

Abstract. *The present article shows an exploratory study about the repetition of the 9^{th} year of primary education students. The work focuses on the State of Rio de Janeiro. Through data collected by Prova Brasil, we analyze characteristics related to students that are associated with repetition. For the extraction of this information, steps are taken to clean and treat the data and the association algorithm Apriori is used in the data mining stage. Finally, conclusions are drawn about the generated model and the results obtained.*

Keywords*: School Failure, Educational Data, Prova Brasil, Data Mining, Apriori Algorithm.*