9ª Jornada Científica e Tecnológica do IFSULDEMINAS
6º Simpósio da Pós-Graduação
ISSN 2319-0124

TÉCNICAS DE MINERAÇÃO DE DADOS APLICADAS EM DADOS DO ENEM 2015

Juliete A. R. COSTA[1]; André L. REIS[2]; Daniel C. L. SOUZA[3]; Kaessa G. S. CRISTINO[4]; Marcelo M. AURELIANO[5]; Salles R. SOARES[6]; Thiago E. SANTOS[7]; Yasmin V. S. SILVA[8]

RESUMO

Este trabalho tem como objetivo relatar a experiência de um Projeto Integrador vivenciado por alunos do curso Técnico em Informática Integrado ao Ensino Médio. O projeto intitulado "Mineração de Dados Educacionais" objetiva integrar os alunos no âmbito da pesquisa acadêmica. No primeiro momento do projeto foram analisados dados do ENEM de 2015 e aplicadas técnicas de mineração de dados para extrair novas informações. Os resultados destacam que algumas variáveis, tais como, de nível socioeconômico e taxa de permanência dos alunos influenciam no resultado das escolas.

Palavras-chave:
Dados Educacionais; Classificação; Regras de Associação

1. INTRODUÇÃO

O grande volume de dados gerados por ambientes educacionais pode fornecer muito mais que simples relatórios. Segundo RODRIGUES (2014), usar técnicas de Mineração de Dados Educacionais (EDM) pode ser útil para descobrir novos conhecimentos. Em BACKER (2011) são apresentadas as principais subáreas de pesquisas em EDM e as tarefas relacionadas. O trabalho apresentado por GOTTARDO (2013) utiliza a técnica de classificação para investigar as inferências com relação ao desempenho dos estudantes no cenário de atividades em grupos.

Em regras de associações, busca-se por regras do tipo "SE...ENTÃO" para avaliar se o valor de uma variável influencia em outra (GARCIA, 2011). No trabalho de BEZERRA (2016), os autores utilizam essa técnica para identificar variáveis que propiciam a evasão de alunos do ensino fundamental. SILVA (2014) utiliza a indução de regras em dados do Exame Nacional do Ensino Médio (ENEM) para avaliar se o fator socioeconômico influencia no desempenho dos alunos.

Atualmente, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) disponibiliza dados para pesquisadores[9]. Este trabalho tem como objetivo proporcionar aos

[1] IFSULDEMINAS - juliete.costa@ifsuldeminas.edu.br
[2] IFSULDEMINAS - andreluiz.dosreis.mg@gmail.com
[3] IFSULDEMINAS - daniellschristian91533444@gmail.com
[4] IFSULDEMINAS - kaessa.gabrieli06@gmail.com
[5] IFSULDEMINAS - mandre.ma648@gmail.com
[6] IFSULDEMINAS - salles.ribeiro.2017@gmail.com
[7] IFSULDEMINAS - thiagoeds2012@gmail.com
[8] IFSULDEMINAS - yasminsilva29.YS@gmail.com
[9] http://portal.inep.gov.br/microdados

ISSN 2319-0124

alunos do Ensino Médio um primeiro contato com pesquisa acadêmica e, consequentemente, descobrir informações interessantes nos dados ENEM por Escola 2015.

3. MATERIAL E MÉTODOS

O banco de dados utilizado é disponibilizado pelo INEP em formato .xlsx com cinco planilhas que destacam o desempenho de 15.597 escolas nas áreas de linguagens, ciências humanas, ciências da natureza, matemática e redação.

A primeira etapa do processo de EDM é realizar a normalização dos dados e neste processo, foram considerados alguns atributos e calculada a média das cinco modalidades de prova para cada escola. Após a normalização, notou-se que as médias variam de 406.49 a 763.71 e, baseado nessa média, foi atribuído um conceito para cada escola (veja Figura 1(a)).

Intervalo de média	Conceito
Maior ou igual a 406 e menor que 500	4 = Ruim
Maior ou igual a 500 e menor que 600	3 = Regular
Maior ou igual a 600 e menor que 700	2 = Bom
Maior que 700	1 = Ótimo

(a) Conceitos das Escolas

UF	DA	LOC	TP	IPE	INS	IFD	C
RO	Estadual	Urbana	68	De_60_a_80	Médio_Alto	55	4
RO	Estadual	Urbana	58	80_ou_mais	Médio	48	4
RO	Estadual	Urbana	55	De_60_a_80	Médio	56	4
RO	Federal	Rural	79	80_ou_mais	Médio	71	3

(b) Instância de dados normalizados – ENEM por Escola

Figura 1 - Características dos dados

A Figura 1(b) ilustra uma instância da base de dados normalizada, onde: **UF** é estado da escola, DA é a dependência administrativa: municipal, estadual, federal ou privada, **LOC** é a sua localização: urbana ou rural, **TP** indica a taxa de participação dos alunos no ENEM, **IPE** é o índice de permanência dos alunos na escola, **INS** é o indicador de nível sócio econômico, **IFD** corresponde ao índice de formação docente e **C** é o conceito da escola (Figura 1(a)). A partir destes dados, foi criado um arquivo para experimentos com a Ferramenta Weka[10].

4. RESULTADOS E DISCUSSÕES

Foram escolhidas as técnicas Regras de Associação e Classificação para os experimentos e todos os testes foram realizados em uma máquina com processador Athlon X2 B26 3.2Ghz Dual Core com 8GB de memória RAM rodando sob o sistema operacional Linux Mint x64.

4.1. Resultados – Regras de Associação

[10] http://www.cs.waikato.ac.nz/ml/weka/index.html

ISSN 2319-0124

Ao analisar os dados notou-se uma discrepância, pois existem mais escolas privadas e estaduais que federais e municipais. Dessa forma, o algoritmo *Apriori* detectaria regras das instituições que correspondem à maioria dos dados. Portanto, decidiu-se analisá-los de forma isolada a partir da dependência administrativa.

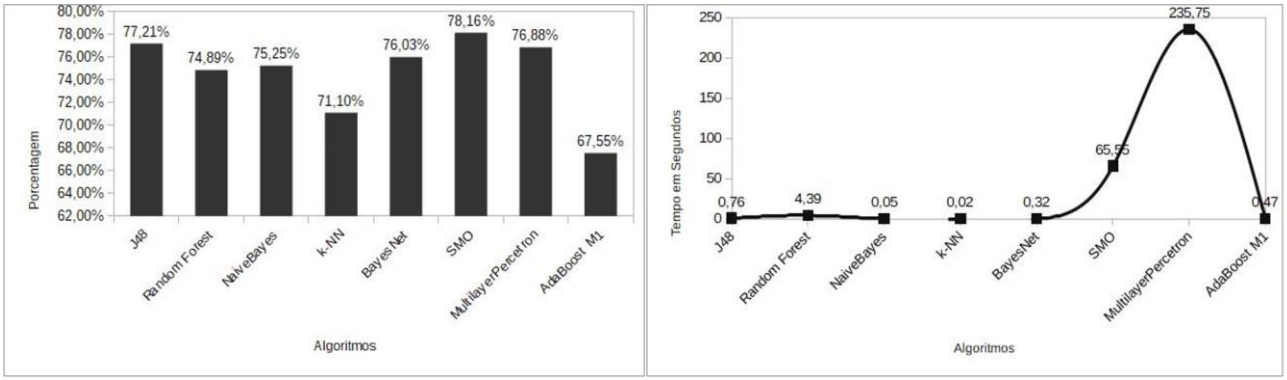
Diante do exposto e alicerçado nos testes efetuados, concluiu-se que ao elevar a medida de suporte e/ou confiança, diminuiu-se a quantidade de regras. Por cúmulo, para haver abundância de regras, fixou-se o suporte no menor índice possível e a confiança em um valor razoável, sendo esta igual a 0,6 e aquele igual a 0,1. Com esses valores de suporte e confiança, o algoritmo encontrou para escolas com DA estadual, federal, municipal e privada 38, 72, 90 e 64 regras, respectivamente.

Foram encontradas interessantes regras que influenciam no conceito, podendo destacar como exemplo as regras 1, 2, 3 e 4. Analisando-as, se percebe que escolas urbanas onde alunos têm IPE a partir de 80% obtêm conceito 3 ou 4, dependendo do INS. É interessante notar que em escolas federais, embora o INS tenha valor médio (Regra 1), o conceito obtido é igual a 3, ou seja, o mesmo conceito obtido em escolas particulares e municipais quando INS tem valor Alto (Regras 3 e 4).

- Regra 1 - Federal: LOC=Urbana, IPE=80_ou_mais, INS=Médio → C=3 (conf: 0.94)
- Regra 2 - Estadual: LOC=Urbana, IPE=80_ou_mais, INS=Médio → C=4 (conf: 0.7)
- Regra 3 - Municipal: LOC=Urbana, IPE=80_ou_mais, INS=Alto → C=3 (conf: 0.96)
- Regra 4 - Privada: LOC=Urbana, IPE=80_ou_mais, INS=Alto → C=3 (conf:0.82)

4.2. Resultados - Algoritmos de Classificação

Na tarefa de classificação deseja-se classificar um determinado indivíduo em uma classe de acordo com suas características. No contexto deste trabalho, deseja-se classificar uma escola com um conceito C (veja Figura 1(a)) de acordo com as características da escola (veja Figura 1(b)).



a. Taxa de acerto dos Classificadores **b. Tempo de criação dos modelos**

Figura 2 - Resultado da aplicação dos classificadores

ISSN 2319-0124

Foram aplicados vários algoritmos de classificação e utilizada a metodologia de teste *Cross Validation*. Percebe-se que o *SMO (Support Vector Machine)* possui melhor taxa de acerto (Figura 2a), no entanto, tem o segundo pior tempo de criação do modelo, demorando 65,55 segundos (Figura 2b). Um algoritmo com resultado interessante é o J48 (Árvore de Decisão), pois alcança a segunda melhor taxa de acerto (77,21%) e o tempo de criação do modelo é de 0,76 segundos.

5. CONSIDERAÇÕES FINAIS

Utilizar técnicas de DM no ambiente educacional é uma tarefa que está se tornando muito comum no ambiente de pesquisa. Com este trabalho foi possível notar a importância deste estudo e, consequentemente, proporcionar aos alunos o primeiro contato com a pesquisa acadêmica. Como trabalho futuro, pretende-se buscar os resultados individuais de alunos a fim de investigar as variáveis que têm influência sobre seus resultados. Além disso, pretende-se realizar testes estatísticos para melhorar a comparação dos resultados obtidos pelos algoritmos de classificação.

REFERÊNCIAS

BAKER, R.S.J.; CARVALHO, A.M.J.B. **Mineração de Dados Educacionais: Oportunidades para o Brasil. Revista Brasileira de Informática na Educação**. *Revista Brasileira de Informática na Educação*. Volume 19, Número 2, p. 1-11, 2011.

BEZERRA, C; SCHOLZ, R; ADEODATO, P; PONTES, R; Silva, I. **Evasão Escolar: Aplicando Mineração de Dados para Identificar Variáveis Relevantes**. *In: Simpósio Brasileiro de Informática na Educação.* p. 1096-1105. 2016.

GARCIA, E; ROMERO, C; VENTURA, S; CASTRO, C. **A collaborative educational association rule mining tool**. *The Internet and Higher Education*. Volume 14, p. 77-88. 2011.

GOTTARDO, E.; KAESTNER, C.; NORONHA, R.V. **Aplicação de técnicas de mineração de dados para estimativa de desempenho acadêmico de estudantes de um AVA utilizando dados com classes desbalanceadas**. In: *International Conference on Interactive Computer aided Blended Learning*. 2013.

RODRIGUES, Rodrigo Lins; RAMOS, Jorge Luis Calvacanti; SILVA, João Carlos Sedraz; GOMES, Alex Sandro. **A literatura brasileira sobre mineração de dados educacionais.** In: Anais do Congresso Brasileiro de Informática na Educação. 2014.

SILVA, Leandro A.; MORINO, Anderson Hideki; SATO, Thiago Massahiro Conti. **Prática de mineração de dados no Exame Nacional do Ensino Médio**. In: Anais do Congresso Brasileiro de Informática na Educação. 2014.