Trabalho apresentado no CNMAC, Gramado - RS, 2016.

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Extração de Conhecimento Relacionado ao Aprendizado de Matemática de Alunos do Ensino Fundamental

Stella Oggioni da Fonseca[1]
Instituto Politécnico, UERJ, Nova Friburgo, RJ
Anderson Amendoeira Namen[2]
Instituto Politécnico, UERJ, Nova Friburgo, RJ

Resumo. O presente artigo apresenta um estudo relacionado à descoberta de conhecimento em bases de dados do INEP, instituto responsável pela avaliação do sistema educacional brasileiro. Por intermédio de um Classificador Bayesiano procura-se identificar fatores associados ao aprendizado de Matemática de alunos do 9º ano do ensino fundamental do Estado do Rio Grande do Sul. São apresentadas as etapas de limpeza e preparação dos dados para posterior aplicação do algoritmo de Mineração de Dados Naïve Bayes. Por fim, são analisados os resultados provenientes deste processo.

Palavras-chave. Mineração de Dados, Descoberta de Conhecimento, Classificador Bayesiano, Aprendizagem de Matemática

1 Introdução

Responsável por promover estudos e pesquisas sobre o sistema educacional brasileiro, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) tem como parte integrante de sua estrutura organizacional a Diretoria de Avaliação da Educação Básica (Daeb). Sob a responsabilidade desta diretoria, a Avaliação Nacional do Rendimento Escolar (Anresc) é aplicada com a finalidade de produzir informações acerca da qualidade do ensino fundamental público [1].

A Anresc, amplamente conhecida como Prova Brasil, é composta de testes que avaliam as habilidades em Matemática e Língua Portuguesa de alunos do 5º e 9º anos do ensino básico. Concomitante a estes testes, são aplicados questionários a alunos, professores e diretores objetivando coletar informações que possam subsidiar melhorias no processo de ensino-aprendizagem. Tais informações são armazenadas em bases de dados que são de domínio público e disponibilizadas no *site* do INEP.

Diante deste contexto, este estudo apresenta uma metodologia capaz de explorar os dados oriundos da Prova Brasil. Mais especificamente, por meio dos questionários preenchidos pelos professores e alunos, procuram-se fatores que possam influenciar positivamente ou negativamente no desempenho obtido em Matemática dos discentes. No presente trabalho, foram selecionados os alunos do 9º ano do Estado do Rio Grande do Sul.

[1]sfonseca@iprj.uerj.br
[2]aanamen@iprj.uerj.br

O artigo apresenta, inicialmente, as bases selecionadas e o processo de limpeza e transformação dos dados que as compõem. Posteriormente, são apresentados de forma sucinta os conceitos do algoritmo *Naïve Bayes*, utilizado para a extração de informações. Por fim, os resultados e conclusões extraídas a partir desse estudo são apresentados.

2 Bases de Dados da Prova Brasil 2011

Com o intuito de cumprir o objetivo apresentado no escopo deste trabalho, as seguintes bases de dados foram selecionadas:

- TS_QUEST_PROFESSOR: arquivo composto pelos dados das respostas ao questionário aplicado ao professor de cada disciplina de cada série;
- TS_QUEST_ALUNO: arquivo que contém os dados das respostas ao questionário aplicado ao aluno de cada série;
- TS_RESULTADO_ALUNO: arquivo que armazena informações da proficiência dos alunos nas provas de Língua Portuguesa e Matemática.

Os arquivos TS_QUEST_ALUNO e TS_RESULTADO_ALUNO contêm 5201730 registros, sendo que o primeiro possui 70 atributos (58 são perguntas sobre o perfil socioeconômico, 1 aloca sobre o preenchimento ou não do questionário e os restantes permitem a identificação de cada aluno) e o segundo é composto de 22 atributos (2 contêm a proficiência obtida em Matemática e Língua Portuguesa e os restantes são atributos identificadores). Já o arquivo TS_QUEST_PROFESSOR possui 304412 registros e 161 atributos (152 referentes às perguntas do questionário e 9 possibilitam a identificação de cada professor).

Os arquivos foram importados para um sistema gerenciador de banco de dados, a saber, o PostgreSQL, que possibilita a seleção, remoção e diversas outras operações sobre registros e atributos de tabelas.

Nestes arquivos foram selecionados somente alunos do 9º ano do ensino fundamental, bem como somente professores que lecionam Matemática para esta série. Além disso, somente alunos e professores do Rio Grande do Sul foram selecionados. Estas seleções foram feitas por meio de atributos identificadores e critérios baseados nas perguntas dos questionários. Mais detalhes desse processo podem ser vistos em [2].

Ao final destas seleções, a tabela TS_QUEST_PROFESSOR passou a conter 3180 registros e as tabelas TS_QUEST_ALUNO e TS_RESULTADO_ALUNO passaram a ter 116061 registros. Ademais, foram removidos da base registros correspondentes a alunos que não preencheram o questionário ou não possuíam proficiência em Matemática, uma vez que não acrescentariam informações para o estudo exposto.

Com a finalidade de manter as bases consistentes e permitir um estudo entre docentes e seus pupilos, foram selecionados, ainda, somente alunos que possuíam professor na base de dados TS_QUEST_PROFESSOR. Portanto, após estas remoções e restrições, restaram nas tabelas TS_QUEST_ALUNO e TS_RESULTADO_ALUNO 63987 registros de alunos e na tabela TS_QUEST_PROFESSOR restaram 3180 registros de professores.

DOI: 10.5540/03.2017.005.01.0570

2.1 Criação de novos atributos

Para identificar fatores relacionados a professores que possam ter influenciado positivamente ou negativamente o desempenho dos discentes em Matemática, foi necessário criar um atributo na tabela TS_QUEST_PROFESSOR que quantificasse, para cada professor, se sua turma obteve ou não um bom resultado.

Determinou-se que este atributo criado deveria armazenar o percentual de alunos, para cada professor, que estivessem acima da média geral de Matemática. Para isso, foi considerada a proficiência de cada aluno presente na tabela TS_RESULTADO_ALUNO. Mais detalhes do processo de criação deste atributo podem ser vistos em [2].

A etapa seguinte foi transformar este atributo numérico em categórico, uma vez que alguns algoritmos de mineração de dados, caso do presente trabalho, não podem ter como alvo um atributo contínuo [4]. Duas formas de mapear este atributo foram feitas, uma que permitiria a análise da influência positiva e outra que possibilitaria a análise da influência negativa do perfil do professor no desempenho dos alunos. A discretização em classes (ou conjuntos) deste atributo presente na tabela TS_QUEST_PROFESSOR pode ser vista na Tabela 1.

Tabela 1: Distribuição das classes para professores.

Atributo	**Classes**	**Descrição das Classes**	**Nº de Registros**
$PERC\_CLASSE\_1$ (influência positiva)	Até 65%	Percentual de até 65% dos alunos acima da média	2370
	Maior que 65%	Percentual maior que 65% dos alunos acima da média	810
$PERC\_CLASSE\_2$ (influência negativa)	Até 35%	Percentual de até 35% dos alunos acima da média	797
	Maior que 35%	Percentual maior que 35% dos alunos acima da média	2383

Com esta divisão apresentada na Tabela 1, esperava-se descobrir quais fatores influenciariam um professor ter um percentual maior do que 65 de alunos acima da média. Acredita-se que tal percentual seja relevante para a identificação de fatores determinantes para a obtenção de resultados mais positivos no processo de ensino-aprendizagem. De modo análogo, por meio da classe "Até 35%", esperava-se encontrar fatores relacionados ao professor que influenciariam negativamente o desempenho dos alunos. Portanto, $PERC\_CLASSE\_1$ e $PERC\_CLASSE\_2$ foram os atributos alvos, sendo que a escolha de um deles dependeria da análise que estivesse sendo feita.

Baseando-se na distribuição dos registros dos professores nas classes apresentadas na Tabela 1, os alunos presentes na tabela TS_QUEST_ALUNO também foram divididos, conforme descrito na Tabela 2.

Por meio da criação dos atributos $ALUNOS\_CLASSE\_1$ e $ALUNOS\_CLASSE\_2$ na tabela TS_QUEST_ALUNO, buscava-se analisar as respostas dadas ao questionário dos alunos que os levariam a ter um professor com resultado positivo ou negativo. Em outras palavras, permitiriam identificar qual o perfil dos alunos que têm professores com alto percentual de pupilos acima da média (Maior que 65%) e com baixo percentual de

DOI: 10.5540/03.2017.005.01.0570

discentes acima da média (Até 35%).

Tabela 2: Distribuição das classes para alunos.

Atributo	Classes	Nº de Registros
$ALUNOS\_CLASSE\_1$ (influência positiva)	Alunos de professores pertencentes à classe Até 65%	46967
	Alunos de professores pertencentes à classe Maior que 65%	17020
$ALUNOS\_CLASSE\_2$ (influência negativa)	Alunos de professores pertencentes à classe Até 35%	14443
	Alunos de professores pertencentes à classe Maior que 35%	49544

3 Classificador Bayesiano

Após a preparação dos dados, deve-se utilizar um algoritmo que efetue a extração de informações embutidas nos dados. Neste trabalho foi utilizado um classificador bayesiano denominado *Naïve Bayes*. De forma sucinta, este algoritmo classifica um registro como sendo de uma determinada classe, com base na probabilidade deste registro pertencer a esta classe. Esta abordagem tem como característica principal o fato de assumir independência condicional, ou seja, os atributos não-alvos não se correlacionam uns com os outros. Apesar dessa hipótese ter limitações, este algoritmo tem se mostrado eficiente [4].

Considere um registro arbitrário X que seja descrito por um conjunto de atributos $\{X_1, X_2, \ldots, X_d\}$. Suponha que pretende-se classificar este registro para uma das classes $C_1, C_2, \ldots, C_k$ do atributo alvo. O algoritmo *Naïve Bayes* efetua esta classificação analisando qual classe torna máxima a probabilidade à posteriori $P(C_j|X)$. Esta probabilidade é calculada para cada classe aplicado-se o Teorema de Bayes:

$$P(C_j|X) = \frac{P(C_j)P(X|C_j)}{P(X)}. \tag{1}$$

Considerando a hipótese de que *Naïve Bayes* assume independência condicional, a probabilidade $P(X|C_j)$ é calculada multiplicando-se a contribuição de cada X_i. Logo, por meio da equação 1 e da hipótese anterior, a probabilidade para cada classe pode ser encontrada. O algoritmo então classifica o registro para a classe que maximar este valor.

Neste trabalho, foi utilizado uma implementação do *Naïve Bayes* disponibilizada dentro do *software* Weka, que é uma ferramenta de mineração de dados de código aberto [5].

4 Resultados e Conclusões

É importante mencionar que, antes de aplicar o algoritmo de mineração, foi efetuado um processo de seleção de atributos por meio de um algoritmo denominado *Correlation-based Feature Selection* (CFS) (ver detalhes em [2]). Este algoritmo, cuja implementação também se encontra no *software* Weka, analisa subconjuntos de atributos que apresentam

DOI: 10.5540/03.2017.005.01.0570

maior correlação com o atributo alvo. Os atributos selecionados podem ser vistos nas Tabelas 3 e 4.

O algoritmo *Naïve Bayes* foi executado quatro vezes neste trabalho. As duas primeiras vezes almejaram analisar questões relacionadas ao professor. De posse da tabela TS_QUEST_PROFESSOR (que contém as questões selecionadas dentre as 152 presentes no questionário), primeiramente foi considerado o atributo alvo $PERC\_CLASSE\_1$, visando à descoberta de fatores que poderiam afetar positivamente o desempenho dos estudantes. Logo, os resultados apresentam atributos, com seus respectivos valores, que descrevam a classe "Maior que 65%". A segunda execução do algoritmo analisou o atributo alvo $PERC\_CLASSE\_2$, ou seja, descobriu fatores que poderiam influenciar negativamente o desempenho em Matemática. Neste caso, a classe de interesse foi "Até 35%".

As duas outras execuções do algoritmo *Naïve Bayes* ocorreram com o intuito de descrever a relação entre o perfil do aluno e o seu professor. A tabela TS_QUEST_ALUNO (que contém os atributos selecionados dentre os 58) foi analisada com o atributo alvo $ALUNOS\_CLASSE\_1$, onde a classe de interesse foi "Alunos de professores pertencentes à classe Maior que 65%". Por fim, o alvo foi alterado para $ALUNOS\_CLASSE\_2$ na busca de identificar fatores de influência negativa. A classe de interesse nos resultados então foi "Alunos de professores pertencentes à classe Até 35%".

Tabela 3: Modelo gerado por *Naïve Bayes* (questionário do professor).

Atributo	**Valores**	**Probabilidade (Maior que 65%)**	**Probabilidade (Até 35%)**
Questão 81) Ocorreu alto índice de faltas dos alunos?	Não	0.5089	0.3939
	Sim, e foi grave	0.0718	0.2092
Questão 116) Você utiliza nesta turma: internet	Sim, utilizo	0.7070	0.6534
	Não utilizo porque a escola não tem	0.0606	0.1629
Questão 121) Quanto dos conteúdos você desenvolveu?	Entre 60% e 80%	0.3763	0.4188
	Mais de 80%	0.5877	0.4746
Questão 123) Quantos dos alunos desta turma você acha que: concluirão o ensino fundamental?	Quase todos	0.9355	0.8322
	Pouco mais da metade	0.0454	0.1432
	Pouco menos da metade	0.0039	0.0089
Questão 124) Quantos dos alunos desta turma você acha que: concluirão o ensino médio?	Quase todos	0.6412	0.3400
	Pouco mais da metade	0.2365	0.4052
	Pouco menos da metade	0.0190	0.0824
	Poucos	0.0145	0.0496
	Não sei	0.0588	0.0836
Questão 125) Quantos dos alunos desta turma você acha que: entrarão para a universidade?	Quase todos	0.1199	0.0200
	Pouco mais da metade	0.2825	0.1465
	Pouco menos da metade	0.2343	0.1304
	Poucos	0.2556	0.5674
	Não sei	0.0908	0.1148
Questão 126) Os alunos desta turma têm livros didáticos?	Sim, todos têm	0.9108	0.8283
	Sim, a maioria tem	0.0622	0.1086
	Não, esta turma não recebeu o livro didático	0.0145	0.0512

O modelo gerado pelo algoritmo *Naïve Bayes* apresenta como resultado a probabilidade correspondente a cada atributo, com seu respectivo valor, dado que uma classe ocorra.

Devido às limitações de espaço inerentes a um artigo, foram apresentados somente os valores dos atributos com maior probabilidade, ou seja, que colaboram para predizer as classes e que realçam as diferenças entre elas.

Ao observar a Tabela 3, alguns fatores significativos a respeito da percepção do professor podem ser vistos. Dentre eles, pode-se destacar a importância da assiduidade por parte dos alunos, da estrutura escolar em disponibilizar recursos como internet, o cumprimento de um alto percentual do conteúdo previsto e a distribuição do livro didático para os estudantes. Além disso, um fato amplamente discutido na literatura refere-se à expectativa que o professor deposita em seus alunos. As questões 123, 124 e 125 explicitam que professores que acreditam em uma boa formação futura para seus alunos, estes obtiveram um melhor resultado em Matemática. Situação inversa é válida ao notar que uma baixa expectativa por parte do professor acarreta em resultados inferiores de seus alunos.

O último fator mencionado afirma que professores com uma visão positiva dos alunos tendem a estimulá-los e estes, por conseguinte, obtêm um melhor desempenho. Ademais, docentes que tem uma visão negativa tendem a ter posturas que comprometem o desempenho dos estudantes. Este fenômeno é conhecido como Efeito Pigmalião (também chamado de Efeito Rosenthal) [3].

A Tabela 4 apresenta o resultado da aplicação do algoritmo *Naive Bayes* nos dados do questionário do aluno.

Tabela 4: Modelo gerado por *Naïve Bayes* (questionário do aluno).

Atributo	**Valores**	**Probabilidade (Maior que 65%)**	**Probabilidade (Até 35%)**
Questão 13) Na sua casa tem computador?	Sim, com internet	0.6727	0.4883
	Não	0.1681	0.3570
Questão 17) Quantas pessoas moram com você?	Moro com mais 3	0.4106	0.3183
	Moro com mais 4 ou 5	0.2663	0.3626
Questão 19) Até que série sua mãe ou a mulher responsável por você estudou?	Não completou a 4ª	0.0689	0.1119
	Completou o Ensino Médio, mas não a faculdade	0.2542	0.1714
	Não sei	0.1433	0.2222
Questão 23) Até que série seu pai ou o homem responsável por você estudou?	Completou a 8ª, mas não o Ensino Médio	0.1266	0.1542
	Completou o Ensino Médio, mas não a faculdade	0.2086	0.1456
	Não sei	0.1839	0.2869
Questão 38) Você lê: sites da internet	Sempre ou quase sempre	0.5984	0.4930
	Nunca ou quase nunca	0.1033	0.1387
Questão 47) Quando você entrou na escola?	Na creche	0.2277	0.0985
	Na pré-escola	0.4731	0.4349
	No 1º ano	0.2807	0.4354

Ao analisar a Tabela 4 notam-se alguns aspectos sobre os alunos que colaboram para que eles tenham um professor cuja turma obteve bom resultado. Dentre eles, destaca-se o fato dos alunos possuírem computador com internet em casa, morar com no máximo 3 pessoas, ler *sites* frequentemente e inserir a criança em um ambiente escolar o quanto antes.

Outro fator relevante diz respeito ao nível escolar dos pais ou responsáveis pelos alu-

DOI: 10.5540/03.2017.005.01.0570

nos. Pais com maior nível escolar possuem filhos cujos professores obtiveram maior êxito. Situação inversa também pode ser constatada, alunos de professores que estão abaixo de 35% de alunos acima da média de Matemática possuem pais com baixo nível escolar.

Portanto, os resultados evidenciam que alunos com condições favoráveis extraescolares têm professores que possuem uma turma com bom desempenho. No entanto, alunos com condições adversas estão em escolas em que os professores obtêm resultados inferiores. Tal fato explicita o círculo vicioso de exclusão de alunos menos favorecidos no sistema educacional.

Conclui-se que o trabalho apresentou importantes fatores relacionados ao professor e ao aluno que podem influenciar tanto positivamente quanto negativamente o ensino-aprendizagem de Matemática. Além disso, apesar de ter sido discutido brevemente, a etapa de preparação de dados demanda grande parte do tempo do processo tendo em vista que é primordial garantir a consistência, qualidade e confiabilidade dos dados.

É importante mencionar, ainda, que estudo semelhante foi efetuado para o Estado do Rio de Janeiro em [2]. Portanto, a escolha do Estado do Rio Grande do Sul, com o entendimento da variabilidade de estruturas das redes públicas inerentes a cada estado, permitiu corroborar os resultados obtidos para o Rio de Janeiro.

Espera-se que este estudo sirva como base para estudos mais aprofundados, analisando em conjunto com profissionais da área da educação outras variáveis presentes nestes questionários e nos outros que não foram aqui abordados.

Agradecimentos

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

Referências

[1] Brasil. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). PDE/PROVA BRASIL Plano de Desenvolvimento da Educação 2011. Brasília, 2011. Disponível em: <http://portal.mec.gov.br/dmdocuments/prova%20brasil_matriz2.pdf>.

[2] S. O. da Fonseca, Utilização de modelos de classificação para mineração de dados relacionados à aprendizagem de matemática e ao perfil de professores do ensino fundamental, Dissertação de Mestrado, UERJ, 2014.

[3] R. Rosenthal, and L. Lenore. Teachers' expectancies: Determinates of pupils' I.Q. gains, *Psychological Reports*, 19:115–118, 1966.

[4] P. Tan, M. Steinbach, and V. Kumar. *Introdução ao Data Mining: Mineração de Dados*. Ciência Moderna Ltda, Rio de Janeiro, 2009.

[5] I. H. Witten, E. Frank, and M. Hall. *Data Mining: Practical Machine Learning Tools and Techniques*. USA: Morgan Kaufmann Publishers Inc., San Francisco, USA, 2011.

