Proceeding Series of the Brazilian Society of Applied and Computational Mathematics, v. 6, n. 1, 2018.

Trabalho apresentado no XXXVII CNMAC, S.J. dos Campos - SP, 2017.

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

Fatores relacionados aos concluintes do curso de Licenciatura em Matemática - uma análise dos resultados do Enade

Stella Oggioni da Fonseca[1]
Instituto Politécnico, UERJ, Nova Friburgo, RJ
Adriana da Rocha Silva[2]
Instituto Politécnico, UERJ, Nova Friburgo, RJ
Anderson Amendoeira Namen[3]
Instituto Politécnico, UERJ, Nova Friburgo, RJ

Resumo. O artigo apresenta um estudo exploratório em bases de dados provenientes do Exame Nacional de Desempenho de Estudantes (Enade), que é um instrumento para avaliar e gerar informações acerca dos concluintes dos cursos de graduação do Brasil. Por intermédio dos dados coletados, este trabalho visa identificar fatores que possam influenciar no processo de aprendizagem dos discentes do curso de Licenciatura em Matemática. Inicialmente são apresentadas as tarefas de pré-processamento destes dados e os conceitos do algoritmo de Mineração *Naïve Bayes*, presente na abordagem utilizada. Por fim, são tecidas algumas conclusões a respeito dos resultados obtidos.

Palavras-chave. Mineração de Dados, Enade, Licenciatura em Matemática

1 Introdução

O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) realiza um diagnóstico do sistema educacional em diversos níveis e modalidades de ensino, incluindo a Educação Superior, cuja avaliação é feita pelo Sistema Nacional de Avaliação da Educação Superior (Sinaes). Com o objetivo de realizar um acompanhamento do processo de ensino-aprendizagem dos estudantes dos cursos de graduação, o Sinaes possui, dentre os seus instrumentos de avaliação, o Exame Nacional de Desempenho de Estudantes (Enade) [2].

O Enade, que teve a sua primeira edição em 2004 e tem periodicidade máxima trienal para cada área do conhecimento, é aplicado a todos os concluintes dos cursos de graduação, sendo um componente curricular obrigatório. O Exame é composto por uma prova, bem como um questionário aplicado aos alunos com o intuito de coletar informações quanto ao seu perfil socioeconômico e aspectos relacionados à sua formação. Portanto, o Enade gera um grande volume de dados que permite estudos visando à melhoria da qualidade do ensino.

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

No presente trabalho, procura-se identificar fatores que possam influenciar no processo de aprendizagem dos discentes. Mais especificamente, buscam-se relações entre as respostas dadas ao questionário e o resultado obtido na prova. O enfoque é dado aos alunos do curso de Licenciatura em Matemática, uma vez que, muitos destes, serão futuros professores de uma disciplina que apresenta um cenário crítico no país.

Diante desse contexto, para extrair esses fatores utiliza-se a Mineração de Dados, que consiste em aplicar métodos capazes de descobrir informações. Assim, o artigo apresenta, inicialmente, as etapas de pré-processamento para efetuar a limpeza e tratamento dos dados. Posteriormente, são apresentados os conceitos do algoritmo de Mineração *Naïve Bayes* e, por fim, os resultados e conclusões extraídas a partir desse estudo.

2 Dados do Enade 2014

Conforme mencionado, o Enade é composto de uma prova e um questionário aplicado aos alunos. A prova é constituída de 30 questões de formação específica, que explanam sobre os conteúdos programáticos de cada curso, e 10 questões de formação geral, que abordam temas diversos relacionados à realidade contemporânea. Já o questionário é composto de 81 perguntas, sendo que 13 destas são respondidas somente por alunos das licenciaturas.

Os dados gerados pelo Enade são de domínio público e encontram-se disponíveis no *site* do Inep. Neste trabalho, foram utilizados os dados referentes ao ano de 2014, resultado mais recente disponibilizado a respeito do curso de Licenciatura em Matemática.

Para cumprir o objetivo exposto no escopo deste artigo, foi efetuado o *download* do arquivo intitulado como:

- *microdados_enade_*2014: contém os dados referentes as respostas dadas ao questionário e o resultado obtido na prova.

O arquivo *microdados_enade_*2014 é composto de um conjunto de 481720 registros, que correspondem, em âmbito nacional, aos concluintes dos cursos avaliados no ano de 2014, e 155 atributos, sendo que 81 armazenam as respostas dadas ao questionário, 2 atributos alocam as notas obtidas na formação geral e específica, e os restantes são atributos relacionados a prova e a identificação de cada aluno (curso, código identificador da Instituição de Educação Superior, seu gênero, idade, dentre outras informações).

Para efetuar o pré-processamento dos dados, este arquivo foi importado para o PostgreSQL, que é um sistema gerenciador de banco de dados que permite, por meio de códigos implementados em Linguagem de Consulta Estruturada (SQL - *Structured Query Language*), realizar diversas operações sobre registros e atributos.

Como o presente estudo é relacionado somente aos alunos que cursam licenciatura em Matemática, o primeiro passo realizado foi selecioná-los. A identificação desses discentes é feita por meio do atributo numérico *co_grupo*, que aloca o código da área de enquadramento do curso no Enade. Portanto, foram selecionados os registros dos alunos que possuem *co_grupo* igual a 702 (código referente ao curso de Matemática). O conjunto de dados passou a conter 16825 registros.

DOI: 10.5540/03.2018.006.01.0367

Outro passo efetuado foi embasado no fato da pesquisa consistir em extrair fatores que possam afetar o processo de aprendizagem durante o curso. Desse modo, o estudo deu ênfase no resultado obtido na formação específica. Foi necessário, então, selecionar somente os estudantes que fizeram esta parte da prova, ou seja, cujo campo numérico nt_ce, que aloca a nota obtida no componente específico, não fosse vazio.

Após este processo, foram selecionados somente alunos que responderam, no mínimo, uma pergunta do questionário. Portanto, ao final dessas seleções, tem-se 13382 registros de alunos para o estudo.

2.1 Atributo Alvo

No presente trabalho, o objetivo é identificar variáveis que possam influenciar no desempenho dos estudantes. Desse modo, o atributo chave, também denominado como alvo, é o nt_ce. Esse atributo numérico, responsável por armazenar uma pontuação que varia de 0 a 100, foi transformado em categórico, uma vez que alguns algoritmos de mineração de dados não podem ter como alvo um atributo contínuo.

O atributo nt_ce foi mapeado em categorias (ou classes) de forma a permitir a identificação de aspectos que possam afetar positivamente ou negativamente o resultado dos alunos. Para isso, os estudantes que obtiveram nota inferior ou igual a 9,7 foram categorizados como "Nota Baixa"; os alunos com nota superior ou igual a 43,5 foram mapeados para a categoria "Nota Alta"; e os restantes, que possuem notas entre os limites mencionados, foram categorizados para "Nota Mediana".

Os limites das notas, 9,7 e 43,5, foram determinados de modo que as classes "Nota Baixa" e "Nota Alta" possuíssem um conjunto de aproximadamente 1000 alunos. Observa-se que cada uma dessas categorias possui cerca de 7,5% de 13382, total de alunos presentes na base em estudo. Acredita-se que esse percentual seja relevante para levantar aspectos e ressaltar as diferenças entre as respostas dadas ao questionário pelos alunos que obtiveram desempenho inferior e superior. A Tabela 1 sumariza o exposto em relação a forma com que o atributo nt_ce foi categorizado.

Tabela 1: Categorias do atributo nt_ce.

Condição	**Categorias**	**Número de registros**
$0 \leq nt\_ce \leq 9,7$	Nota Baixa	1008
$9,7 < nt\_ce < 43,5$	Nota Mediana	11363
$nt\_ce \geq 43,5$	Nota Alta	1011

É importante salientar que a média das notas obtidas pelos estudantes presentes na base é de 25,7. Esse número revela que os discentes dos cursos de graduação em Matemática não apresentam domínio dos conteúdos abordados ao longo do curso.

2.2 Seleção de Atributos

Como deseja-se relacionar as respostas dadas as 81 questões com o atributo alvo, tem-se um problema de alta dimensionalidade (alto número de atributos envolvidos no estudo) e, portanto, a etapa de selecionar algumas dessas variáveis é essencial.

DOI: 10.5540/03.2018.006.01.0367

Neste trabalho, a seleção dos atributos foi efetuada pelo critério Incerteza Simétrica (*Symmetrical Uncertainty*), que gera um ranqueamento. Para ordenar cada variável, são utilizadas medidas que avaliam a sua qualidade, ou seja, o quanto ela está correlacionada com o alvo (alta correlação implica em melhor qualidade). No critério Incerteza Simétrica, a medida utilizada baseia-se no conceito de entropia, que permite identificar os atributos que mais realçam as diferenças entre as classes (ver detalhes em [3]). A implementação deste critério está disponibilizada dentro do *software* Weka, que é uma ferramenta, de código aberto, que possui algoritmos de seleção de atributos e mineração de dados [4].

Conforme mencionado, dentre as 81 perguntas, 13 são específicas para alunos de licenciatura. Assim, decidiu-se abordar separadamente essas questões. O processo de seleção foi efetuado duas vezes: dentre as 68 primeiras questões, as mais relacionadas com o atributo alvo foram as de número 13, 17, 8, 2 e 4; já dentre as 13 questões, foram selecionadas as de número 72 e 77 (os enunciados dessas questões são apresentados na Seção 4). A posterior etapa de mineração foi efetuada com os atributos selecionados e o alvo nt_ce.

3 Mineração de Dados: *Naïve Bayes*

Após as tarefas de pré-processamento, deve-se aplicar um algoritmo de mineração responsável pela extração de informações que encontram-se embutidas nos dados. Neste artigo, 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.

Para compreender o algoritmo, considere um registro arbitrário X que seja descrito por um conjunto de atributos $\{X_1, X_2, \ldots, X_d\}$, onde d é o número máximo de atributos. Suponha que pretende-se classificar este registro para uma das k classes $C_1, C_2, \ldots, C_k$ presentes no atributo alvo. O algoritmo *Naïve Bayes* efetua esta classificação analisando qual classe torna máxima a probabilidade à posteriori $P(C_j|X)$, com $j = \{1, 2, \ldots, k\}$. A probabilidade, para cada classe, é calculada 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 (os atributos não-alvos não se correlacionam uns com os outros), o termo $P(X|C_j)$ presente na equação 1 é calculado multiplicando-se a contribuição de cada X_i, com $i = \{1, 2, \ldots, d\}$. 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 utilizada uma implementação do *Naïve Bayes* disponibilizada dentro do *software* Weka [4].

4 Resultados e Conclusões

O algoritmo *Naïve Bayes* foi aplicado para identificar as relações entre os atributos selecionados e o atributo alvo. Conforme mencionado na seção anterior, o modelo gerado

DOI: 10.5540/03.2018.006.01.0367

pelo *Naïve Bayes* apresenta como resultado a probabilidade correspondente a cada atributo, com seu respectivo valor, dado que uma classe ocorra, ou seja, $P(X_i|C_j)$, sendo que as classes C_j são "Nota Baixa", "Nota Mediana" e "Nota Alta".

O modelo resultante é apresentado na Tabela 2. Para facilitar a visualização dos resultados, as probabilidades foram convertidas para o formato de porcentagem.

Tabela 2: Modelo gerado por *Naïve Bayes*.

Atributo	Valores	Probabilidade (Nota Baixa)	Probabilidade (Nota Mediana)	Probabilidade (Nota Alta)
Questão 13) Ao longo de sua trajetória acadêmica, você recebeu algum tipo de bolsa?	Nenhum	83,8	74,3	49,2
	Bolsa de iniciação científica	2,4	4,8	14,3
	Bolsa de extensão	1,2	3,5	7,3
	Bolsa de monitoria/tutoria	1,8	3,7	9,4
	Bolsa PET	0,3	0,5	2,2
	Outro tipo de bolsa acadêmica	10,4	13,2	17,5
	Nulo	0,1	0,0	0,1
Questão 17) Em que tipo de escola você cursou o ensino médio?	Todo em escola pública	87,7	83,9	69,7
	Todo em escola privada	5,9	10,2	22,7
	Todo no exterior	0,2	0,1	0,1
	A maior parte em escola pública	3,7	3,5	3,9
	A maior parte em escola privada	2,3	2,3	3,2
	Parte no Brasil e parte no exterior	0,1	0,0	0,3
	Nulo	0,1	0,0	0,1
Questão 8) Qual a renda total de sua família, incluindo seus rendimentos (número de salários mínimos)?	Até 1,5	34,7	25,6	14,6
	De 1,5 a 3	34,6	34,2	28,3
	De 3 a 4,5	18,3	19,5	21,3
	De 4,5 a 6	6,2	9,9	13,4
	De 6 a 10	5,5	8,1	13,6
	De 10 a 30	0,6	2,6	7,8
	Acima de 30	0,1	0,1	1,0
Questão 2) Como você se considera?	Branco(a)	30,7	41,2	57,1
	Negro(a)	14,8	11,6	6,4
	Pardo(a)/mulato(a)	51,7	44,8	34,0
	Amarelo(a) (de origem oriental)	1,3	1,3	1,7
	Indígena ou de origem indígena	1,5	1,1	0,8
Questão 4) Até que etapa de escolarização seu pai concluiu?	Nenhuma	20,9	15,5	9,0
	1º ao 5º ano	45,8	42,7	31,7
	6º ao 9º ano	11,1	15,8	14,8
	Ensino Médio	16,4	19,2	28,7
	Graduação	4,9	5,1	11,6
	Pós-graduação	0,9	1,7	4,2
Questão 72) Se você tem experiência no magistério, em qual etapa/modalidade atuou? Assinale a alternativa mais relevante para você	Educação Infantil	4,5	3,1	1,5
	Ensino Fundamental - anos iniciais	15,7	11,4	5,2
	Ensino Fundamental - anos finais	27,9	30,2	23,7
	Ensino Médio	12,8	18,5	28,5
	Educação Profissional Técnica	1,3	1,4	1,8
	Educação de Jovens e Adultos	1,9	2,2	1,8
	Ensino Superior	0,4	0,5	2,6
	Outra modalidade de ensino	0,6	0,8	1,6
	Não tenho experiência no magistério	34,8	31,9	33,2
	Nulo	0,1	0,0	0,1
Questão 77) Onde você pretende atuar daqui a 5 anos?	Escola pública, como professor	72,4	66,2	53,1
	Escola privada, como professor	5,8	8,7	18,5
	Escola/instituição pública, na gestão	7,3	8,9	9,7
	Escola/instituição privada, na gestão	1,9	2,2	3,0
	Outro não vinculado à educação	12,5	14,0	15,6
	Nulo	0,1	0,0	0,1

Para exemplificar a interpretação do modelo, considere o atributo referente à Questão 13. Observa-se que, dentre os 1008 alunos pertencentes à classe "Nota Baixa", 83,8% não receberam bolsa ao longo de sua trajetória acadêmica; 2,4% possuíram bolsa de iniciação científica; e assim sucessivamente. Nota-se que 0,1% desses alunos não responderam a essa questão, sendo representado pelo valor "Nulo". Interpretação análoga pode ser feita em relação às classes "Nota Mediana" e "Nota Alta".

Ao observar a Tabela 2, alguns resultados relevantes puderam ser extraídos. Nota-se que mais de 50% dos alunos da classe "Nota Alta" possuíram bolsa em sua trajetória acadêmica, enquanto que somente 16,1% dos alunos pertencentes à classe "Nota Baixa"

DOI: 10.5540/03.2018.006.01.0367

estiveram na mesma situação. Tal fato pode apontar a influência positiva de ingressar os alunos, ao longo do curso de graduação, em atividades que enriquecem o conhecimento e são remuneradas.

Outro fator que pode ser notado é que, independente da classe, a grande maioria dos alunos dos cursos de licenciatura em Matemática estudaram o ensino médio em escola pública. Mesmo com essa constatação, ainda, percebe-se que os alunos provenientes de escola privada configuram em maior número no conjunto "Nota Alta". Nota-se, ainda, a influência de outros aspectos amplamente discutidos na literatura, como a relação entre o desempenho e o perfil econômico, bem como a corroboração das diferenças histórias quanto à cor/etnia declarada. Além disso, conclui-se que com o aumento do nível escolar dos pais, melhores são os resultados dos seus filhos.

É importante salientar que este último fator mencionado está presente nos mais diversos níveis de ensino. Alves, Ortigão e Franco (2007, p.176), ao analisarem os dados do Sistema de Avaliação da Educação Básica do ano de 2001, concluíram que a "instrução dos pais é um dos fatores que mais se relaciona com o desempenho escolar dos estudantes e, no caso da repetência, quanto maior a instrução, menor é o risco de ocorrência desse fenômeno".

Fatores relevantes também puderam ser observados em relação às questões 72 e 77. Observa-se que os alunos pertencentes à classe "Nota Alta" possuem mais experiência em etapas superiores ao Ensino Fundamental do que os alunos pertencentes à classe "Nota Baixa". Ademais, em ambas as classes, muitos alunos pretendem atuar como professor em escolas públicas. No entanto, um número maior de alunos pertencentes à classe "Nota Alta" almeja ser professor em escolas privadas.

Após a etapa de mineração, para fomentar a discussão de que o ensino é também influenciado por vários fatores acerca do ambiente escolar, foram analisados dois indicadores de qualidade da Educação Superior, a saber: o Conceito Enade e o Conceito Preliminar de Curso (CPC). Ambos construídos a partir dos resultados obtidos no Enade e apresentados em uma escala de cinco níveis, os indicadores são calculados para cada curso de cada instituição. O Conceito Enade avalia somente o desempenho dos discentes no exame, já o CPC é mais abrangente, considerando também em seu cálculo as respostas dadas ao questionário do estudante, infraestrutura das instituições e as informações sobre o corpo docente. Mais detalhes sobre esses indicadores podem ser vistos em [2].

Os resultados obtidos são disponibilizados para *download* no *site* do Inep, sendo que encontram-se em tabelas intituladas como *conceito_enade_*2014 e *cpc_*2014. A Figura 1 apresenta a porcentagem de alunos, referente a cada classe, nas faixas dos indicadores. Os estudantes na faixa "Nulo" do Conceito Enade estão inseridos em um curso que não reuniu condições para o cálculo do indicador, como por exemplo, menos de dois concluintes participaram da prova (das 443 unidades com curso de licenciatura em Matemática, 13 não possuem Conceito Enade). Já o valor "Nulo" no Conceito CPC engloba também os alunos pertencentes às unidades sem cursos reconhecidos pelo Ministério da Educação (das 443 unidades, 109 ainda não possuem curso de Matemática reconhecido).

Nota-se que, em ambos os conceitos, os alunos pertencentes à classe "Nota Baixa" concentram-se nas faixas inferiores, que revelam resultado insatisfatório, enquanto que os alunos da categoria "Nota Alta" estão nas faixas superiores. Por intermédio desse resultado é possível afirmar que os alunos com maiores notas estão inseridos em unidades que

DOI: 10.5540/03.2018.006.01.0367

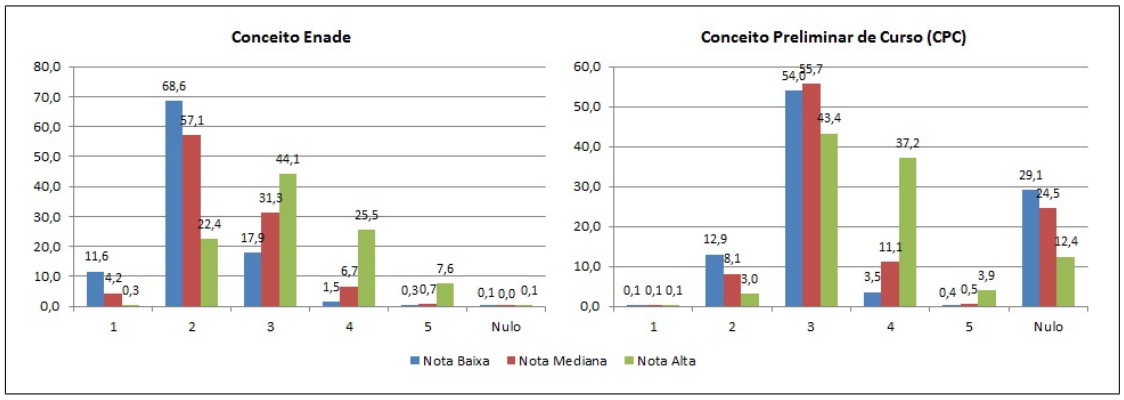

Figura 1: Percentual de alunos em cada faixa dos indicadores Conceito Enade e CPC.

apresentam melhores resultados e condições quanto à infraestrutura e nível do corpo docente. O inverso pode ser concluído em relação aos alunos que obtiveram notas inferiores. Tal fato reforça que vários fatores influenciam no processo de ensino-aprendizagem dos estudantes, isto é, nenhum fator isoladamente determina a qualidade da educação.

Portanto, conclui-se que o trabalho apresentou importantes aspectos que possam ter influenciado no desempenho obtido pelos concluintes do curso de Matemática. Expôs, ainda, como é executado o processo de descoberta de conhecimento em bases de dados, incluindo etapas essenciais como a seleção das variáveis mais relevantes e a Mineração de Dados, que realiza a junção entre a Estatística e a Inteligência Computacional. Ademais, espera-se que este trabalho sirva como base para estudos mais aprofundados e motive a utilização de algoritmos de Mineração em bases do Enade e de outras avaliações do Inep.

Agradecimentos

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

Referências

[1] F. Alves, I. Ortigão, C. Franco. Origem social e risco de repetência: interação entre raça-capital econômico. *Cadernos de Pesquisa*, v. 37, n. 130, p. 161-180, 2007.

[2] Brasil. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira - Inep. Microdados do Exame Nacional de Desempenho dos Estudantes. [online]. Brasília, 2014. Disponível em: <http://portal.inep.gov.br/enade>. Acesso em: 30 ago. 2016.

[3] 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.

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

