Avaliação da Qualidade das Instituições de Ensino Superior utilizando os microdados do Enade e do Censo da Educação Superior

Felipe Navarro Balbino Alves
Centro de Informática - CIn
Universidade Federal de Pernambuco - UFPE
Recife, Brasil
fnba@cin.ufpe.br

Débora da Conceição Araújo
Centro de Informática - CIn
Universidade Federal de Pernambuco - UFPE
Recife, Brasil
dca2@cin.ufpe.br

Paulo de Assis Nascimento
Centro de Informática - CIn
Universidade Federal de Pernambuco - UFPE
Recife, Brasil
pan2@cin.ufpe.br

1 Introdução

A qualidade e a acessibilidade à Educação são aspectos-chave no desenvolvimento de uma sociedade mais justa [Adeodato, 2016]. É neste sentido que políticas educacionais como o Enade e o Censo da Educação Superior atuam ao avaliar o rendimento dos estudantes de cursos de graduação. Através dessas políticas, tornou-se possível reunir uma enorme quantidade de dados acerca do panorama da educação de nível superior brasileira, uma vez que tais bases contém informações relativas a milhares de instituições de ensino e professores, além de milhões de estudantes.

Enquanto o Enade avalia a proficiência dos estudantes por meio da realização de uma prova, o Censo da Educação Superior considera informações detalhadas acerca da situação de uma instituição de ensino, tais como o perfil dos seus alunos, professores e dos seus cursos. Estes dados são obtidos através de um questionário.

Nesse sentido, este trabalho pretende se utilizar das técnicas de mineração de dados para atuar na extração de regras relevantes acerca da qualidade das Instituições de Ensino Superior no Brasil. Para nortear as atividades, foi escolhida a metodologia CRISP-DM (CRoss Industry Standard Process for Data Mining) [8], por proporcionar uma visão sistêmica das etapas do processo, que são: entendimento do negócio, compreensão dos dados, preparação dos dados, modelagem, avaliação e desenvolvimento. Vale salientar que esse é um processo não linear.

2 Caracterização do Problema e definição da variável alvo.

Segundo o Mapa do Ensino Superior do Brasil (2015), durante os últimos 13 anos, a quantidade de universidades no país mais que dobrou [1], porém, o déficit educacional brasileiro se torna evidente quando se considera que o Brasil ainda é o sexagésimo colocado em educação básica em um ranking de setenta e dois países criado pela OCDE (Organização para a Cooperação e Desenvolvimento Econômico) [2]. Este resultado da educação básica tende a refletir na qualidade do ensino superior, visto que nenhuma das universidades brasileiras estão entre as 100 melhores do mundo, de acordo com um ranking publicado pelo *Center for World University Rankings* (CWUR) [4].

Dessa forma, é objetivo deste artigo investigar quais parâmetros são relevantes para a qualidade dos cursos de graduação brasileiros, a fim de inspirar diretrizes mais eficientes para atacar os problemas da educação no Brasil.

O primeiro passo para tal é a definição de uma métrica para se avaliar a qualidade de uma instituição de ensino. Optou-se por utilizar a nota média dos alunos no Enade, uma vez que este é um dos parâmetros já usados pelo INEP para averiguar a qualidade de um curso de graduação [9]. Definida a métrica de desempenho, procede-se à escolha do grão final de análise.

Uma vez que deseja-se avaliar as instituições de ensino respeitando as diferenças entre os cursos, definiu-se o grão final como sendo o par *instituição x área geral*. Por exemplo, UFPE - Educação. O par *instituição x curso* não foi utilizado como grão pois não existia um índice único que permitisse relacionar os cursos da base de dados do Enade com os cursos da base de dados do Censo Escolar. Tal fato não ocorre com as instituições de ensino, que possuem um código de identificação único nas duas bases.

Por fim, considerou-se um problema de decisão binária onde a variável alvo são os registros (pares instituição x área geral) que estão entre os 25% melhores de acordo com a nota do Enade 2014.

3 As Bases de Dados

Este trabalho foi desenvolvido sobre duas bases de dados. Uma delas é o Censo Escolar, que contém informações sobre alunos, professores, cursos, instituições de ensino superior [5]. A outra armazena os microdados do Enade, que contém o detalhamento do desempenho dos alunos na referida prova [6]. Diante da grande quantidade de registros e variáveis (atributos) nessas fontes, cabe aqui a descrição das bases e seus arquivos.

O Censo Escolar contém dados em diversas granularidades (aluno, curso, instituição e docente). Na Tabela I, podemos observar a quantidade de registros para cada uma das tabelas de dados do Censo Escolar. Na Base de Dados do Enade, os registros estão na granularidade aluno e temos 481.721 registros. As duas bases precisam ser relacionadas e por isso observaram-se os seguintes fatos:

- Há um índice comum entre as duas bases de dados para identificar a instituição de ensino (IES);
- Não há um índice comum para a identificação do curso;
- Nas duas bases de dados, há campos que identificam a área do curso;

Tabela I. Arquivos da base de dados do Censo da Educação Superior 2014

Arquivo	**Registros**
dm_aluno.csv	10.793.933
dm_instituicao.csv	2.368
dm_curso.csv	33.274
dm_professor.csv	396.596

As bases foram relacionadas por meio do código da área geral do conhecimento presentes na base do Censo da Educação Superior conforme a Tabela II. Esses códigos foram aplicados na base de dados do Enade para que os cursos das grandes áreas possam ter uma relação direta com a base do Censo da Educação Superior.

Tabela II. Código das Grandes Áreas de Conhecimento

Código OCDE	Área de Conhecimento
1	Educação
2	Humanidades e Artes
3	Ciências Sociais, Negócios e Direito
4	Ciências, Matemáticas e Computação
5	Engenharia, Produção e Construção
6	Agricultura e Veterinária
7	Saúde e bem estar social
8	Serviços

4 Pré Processamento dos dados

Durante o processo de conversão dos dados para o grão de análise final, foram necessárias transformações em alguns atributos. Assim, visando o agrupamento de informações provenientes de variáveis categóricas, tais atributos foram convertidos em variáveis *dummy*[1] e foi utilizada a aglutinação pela média. Dessa forma, no grão final, temos a proporção de elementos em cada uma das classes. Para as variáveis numéricas, foi utilizada a aglutinação pela média. Já nas variáveis referentes à datas (Ex: data de abertura do curso), considerou-se apenas o ano e tomou-se a média.

A fim de garantir que o classificador construído não aprenda aspectos regionais da educação, removemos as colunas relacionadas à localização das Instituições de Ensino Superior. Além disso, para alunos estrangeiros, considerou-se apenas a booleana: é brasileiro ou não, ou seja, ignorou-se o país de origem por possuir muitas categorias.

Para cada arquivo da base de dados do Censo da Educação Superior foram realizados os seguintes processamentos: No arquivo dm_aluno.csv, eliminamos os alunos que não estavam cursando aulas no ano de referência e também aqueles cujo curso não tem área geral definida (este fato ocorre quando se pode escolher entre licenciatura ou bacharelado em uma etapa posterior do curso). Estes alunos eliminados compreendiam 37,12% dos dados originais.

Em dm_curso.csv foram utilizados apenas os cursos ativos em que área geral de conhecimento estava presente, e que continham o ano de início devidamente preenchido. Perdemos dessa forma 7,22% dos registros originais.

No arquivo dm_docente.csv utilizamos apenas os professores em atividade, ocasionando uma perda de 3,33 % dos registros.

[1] Quando se transforma uma variável categórica em n variáveis numéricas, onde n é o número de categorias dessa variável. Este processo está descrito em [7].

Por fim, no arquivo dm_enade.csv considerou-se apenas os alunos presentes no dia da avaliação e cuja nota geral constava na base, o que resultou numa perda de 17,90 % dos registros.

Após o agrupamento por Instituição de Ensino Superior x Área de Conhecimento, ficamos com 2.626 registros no grão final de análise.

5 Modelagem

Para a avaliação dos parâmetros mais relevantes para a qualidade dos cursos, foi construído um modelo de Regressão Logística, que permite avaliar a influencia positiva ou negativa de cada atributo através dos coeficientes obtidos. Também é possível avaliar quais parâmetros são relevantes no processo decisório a partir da variância de tais parâmetros. Basta observar que se o zero está contido no intervalo de confiança do coeficiente de um parâmetro, não podemos decidir se esta variável é favorável ou desfavorável à classe alvo.

Tendo em vista tais fatos, procedeu-se à construção do modelo de regressão logística, inicialmente normalizando variáveis que não representam proporções, como por exemplo o valor recebido por transferência por uma instituição de ensino.

Isto é necessário pois grande parte das variáveis da nossa base de dados aglutinada é composta por médias de variáveis *dummy*, que representam proporções e estão naturalmente entre zero e um. As demais variáveis assumem intervalos bem diferentes e podem interferir na convergência da Regressão Logística.

Uma vez normalizados os dados, iniciou-se o treinamento dos modelos, onde eram realizadas as retiradas das variáveis que não eram relevantes. Para tal, utilizou-se como métrica a probabilidade de um coeficiente possuir um *z-score* (número de desvios pelo qual um valor está afastado da média) maior do que o valor amostral obtido. No conjunto final de atributos, este parâmetro era inferior a 10% para todas as variáveis.

Após esta etapa, buscou-se validar o conhecimento fornecido pelo modelo de regressão através da extração de regras.

Uma vez que o output de um algoritmo de extração de regras é mais fácil de ser interpretado quando as variáveis de entrada são categóricas, realizamos a quantização das variáveis contínuas em três conjuntos, denominados Bin 1, Bin 2 e Bin 3. A quantização foi realizada de forma que cada nível (Bin) tivesse aproximadamente o mesmo número de registros, de forma que os intervalos da quantização não possuíam comprimentos iguais. Podemos interpretar esta quantização como uma divisão em *tercis*.

Assim, tornou-se possível interpretar as saídas da indução de regras da seguinte forma: Quando uma variável é associada ao Bin 1, interpretamos como uma baixa intensidade da variável, por exemplo, "A média da idade dos alunos = Bin 1" significa que temos instituições onde a média da idade dos alunos é baixa. Da mesma forma, quando uma variável é atribuída ao Bin 2, interpretamos como uma intensidade intermediária e, analogamente, quando a variável é atribuída ao Bin 3, temos uma intensidade forte. Assim, "A média da idade dos alunos = Bin 3" significa que temos instituições onde os alunos são mais velhos.

6 Regressão logística

Após a eliminação das variáveis não relevantes, chegamos aos coeficientes mostrados na Tabela III.

Tabela III. Modelo de regressão logística obtido

Atributo	**Coeficiente**	**Desvio padrão**	**P > |z|**
Porcentagem de mulheres alunas	2.257	0.275	0.000
Porcentagem de alunos não ingressantes por seleção simplificada	1.564	0.502	0.002
Porcentagem de alunos não ingressantes por vestibular próprio	1.162	0.280	0.000
Valor recebido pela instituição através de transferência	0.300	0.070	0.000
Número de horas integral	0.280	0.068	0.000
Número de horas EAD	-0.138	0.052	0.007
Idade média dos alunos	-0.217	0.069	0.002
Porcentagem de professores em regime parcial	-0.523	0.279	0.061
Constante	-0.832	0.662	0.209
Porcentagem de docentes pardos	-0.855	0.418	0.041
Porcentagem de	-1.121	0.396	0.005

alunos pardos			
Porcentagem de alunos que não ocupam vagas novas	-1.270	0.452	0.005
Porcentagem de professores com especialização	-2.514	0.375	0.000
Porcentagem de mulheres docentes	-2.582	0.628	0.000

Na figura I, podemos observar a curva ROC obtida para este modelo final, que possui uma área sob a curva de cerca de 77,25%.

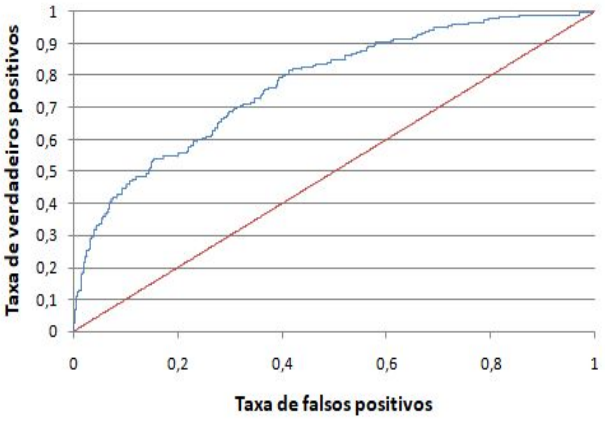

FIGURA I. CURVA ROC OBTIDA PARA O MODELO FINAL DE REGRESSÃO LOGÍSTICA

Em uma primeira observação dos coeficientes, chama a atenção a relevância que o modelo dá ao número de mulheres dentre os alunos e dentro dos professores. Podemos observar que, de acordo com o modelo, quanto maior o número de alunas, melhor o desempenho da escola. De fato, ao separarmos os registros onde a porcentagem de mulheres dentre os alunos é maior que 50%, a proporção da classe alvo era de cerca de 30%, contra 20% nos registros onde a porcentagem de mulheres era inferior a 50%.

No entanto, um fato que chama bastante atenção é que quando olhamos para o corpo docente, esta situação se inverte, ou seja, o modelo sugere que nas instituições onde existem mais professores homens tende-se a obter melhores desempenhos. Isto pode indicar uma mudança no perfil das mulheres nas universidades brasileiras entre as gerações passadas (docentes) e as gerações atuais (alunos).

Outro fator de interesse é o coeficiente relacionado à porcentagem de docentes com grau de escolaridade especialização, indicando que quanto maior é essa porcentagem, pior é o desempenho da escola. Tendo em vista que um professor que está registrado com este nível de escolaridade provavelmente não cursou o mestrado ou o doutorado, podemos interpretar que em universidades onde muitos professores não alcançaram estes graus superiores da educação tendem a terem baixo desempenho.

Observando os demais coeficientes, é possível perceber algumas características fortemente relacionadas a instituições de ensino particulares. Por exemplo, observamos que, de acordo com o modelo, quanto menos alunos ingressarem através de seleção simplificada, melhor. A seleção simplificada é um método de ingresso utilizado por alunos portadores de diploma ou transferência, e é usada primordialmente por universidades particulares. Outro parâmetro similar é a porcentagem de alunos que não ingressaram através de vestibular próprio. De fato, em 2014, a maioria das instituições públicas de ensino já haviam aderido ao ENEM, de forma que as universidades que utilizam vestibulares próprios são, em grande parte, particulares. Pelo modelo, pode-se observar que a predominância desses métodos de ingresso é prejudicial ao desempenho da instituição, o que levanta uma questão acerca da qualidade das IES particulares brasileiras, uma vez que os métodos de ingresso utilizados por elas são penalizados pelo modelo.

Uma variável com interpretação similar é a porcentagem de alunos que não ocupam vagas novas. Esses alunos, em geral, são aqueles que já estão no ensino superior, mas prestam exames como o ENEM com o objetivo de conseguir financiamento estudantil, e não ingressar em uma universidade. Assim, este aluno não ocupará uma nova vaga pois não mudará de curso, estando apenas a tentar conseguir financiamento. O modelo sugere que um número grande de alunos nessa situação é um demérito para a escola, o que novamente é uma situação típica de instituições particulares.

Outra análise interessante é que a variável associada ao número de horas em tempo integral de um curso possui sinal positivo, enquanto a variável associada ao número de horas em regime EAD de um curso possui coeficiente negativo, o que indica que cursos EAD ainda possuem baixa qualidade quando comparados com os cursos correspondentes em regime integral.

Com relação ao perfil dos alunos, podemos observar que alunos mais velhos tendem a ter um desempenho inferior, como podemos observar pelo coeficiente negativo dessa variável. Este fato levanta o questionamento acerca de quais motivos levaram esses alunos a ingressarem tardiamente no ensino superior e qual a influência desses fatores no desempenho do aluno ao longo do curso. Além disso, podemos observar que alunos de cor parda também tendem a ter um desempenho inferior, de fato, uma alta porcentagem de professores de cor parda também aparece como demérito no modelo. Dessa forma, também podem ser levantados questionamentos em relação às condições socioeconômicas da

população parda brasileira. Pode-se também questionar o acesso à educação do qual esse segmento da sociedade dispõe.

Com essas informações, políticas públicas podem ser repensadas, de modo a oferecerem melhores condições de ensino a esses estudantes.

É notória também a importância do investimento financeiro nas instituições de ensino, uma vez que o coeficiente relacionado ao valor que uma instituição recebe por transferência possui um coeficiente positivo e, portanto, leva a desempenhos melhores.

O regime de trabalho dos docentes também aparece como um parâmetro relevante, mostrando que quanto mais professores estão trabalhando em regime parcial, pior o desempenho da instituição.

7 Extração de regras

Para a extração de regras, foi utilizado o algoritmo JRip. Além disso, conforme mencionado na seção 5, utilizamos as versões quantizadas das variáveis contínuas, onde foram construídos três níveis para cada variável, que podem ser interpretados como uma intensidade baixa, intermediária e alta de uma determinada variável. As regras obtidas podem ser visualizadas na Tabela IV.

Tabela IV. Regras extraídas utilizando o algoritmo JRip

(1)
Poucos professores com no máximo especialização;
Poucos professores com regime parcial de trabalho;
Alta integralização integral;
Baixa idade média dos alunos
=
A instituição é boa na área

Confiança	Suporte	Lift
76%	3.56%	3.04

(2)
Poucos professores com no máximo especialização;
Poucos professores com regime parcial de trabalho;
Alta idade média dos alunos;
Quantidade de alunos pardos intermediária;
=
A instituição é boa na área

Confiança	Suporte	Lift
63.71%	2.28%	2.508

(3)
Poucos professores com no máximo especialização;
Alto volume de recursos recebidos por transferência;
Poucos alunos pardos
=
A instituição é boa na área

Confiança	Suporte	Lift
54.80%	4.03%	2.1927

(4)
Poucos professores com no máximo especialização;
Poucos professores com regime parcial de trabalho;
Poucos alunos que não ocupam vagas novas
=
A instituição é boa na área

Confiança	Suporte	Lift
51.16%	1.66%	2.0465

(5)
Muitas mulheres dentre os alunos;
Quantidade intermediária de alunos pardos;
Idade média dos alunos intermediária;
Poucos recursos recebidos por transferência
=
A instituição é boa na área

Confiança	Suporte	Lift
59.09%	1.70%	2.3636

(6)
Poucos professores com no máximo especialização;
Quantidade intermediária de docentes pardos;
Poucos professores em regime parcial;
Poucas mulheres dentre os professores;
Muitos alunos não entraram por vestibular próprio
=
A instituição é boa na área

Confiança	Suporte	Lift
65.38%	1.00%	2.6153

(7)
Muitas mulheres dentre os alunos;
Poucos pardos dentre os docentes;

Poucos professores com no máximo especialização; Poucos alunos pardos = A instituição é boa na área		
Confiança	**Suporte**	**Lift**
66.66%	0.69%	2.666

(8) Poucos professores com no máximo especialização; Quantidade intermediária de alunos pardos; Muitas mulheres dentre os alunos; Quantidade intermediária de docentes pardos = A instituição é boa na área		
Confiança	**Suporte**	**Lift**
58.33%	0.93%	2.333

Podemos observar que foi possível a obtenção de algumas regras com lift alto (acima de 2.5). Obtivemos inclusive uma regra com lift 3.04, o que é bastante alto considerando que o lift máximo é 4, visto que a proporção da classe alvo no conjunto de dados final é 25%.

No entanto, é importante observar que as regras obtidas possuem um suporte baixo, inferior a 10%, o que põe em cheque a qualidade das estimativas de confiança e lift. Vale ressaltar porém que o objetivo principal da extração de regras é validar as conclusões fornecidas pelo modelo de regressão logística, o que foi possível, uma vez que observamos fenômenos como a preferência das regras por baixas porcentagens de professores com, no máximo, especialização. Também foi possível observar a influência da proporção de mulheres em regras como a (5), a (7) e a (8).

Além disso, outros efeitos foram observados, como a influência da raça dos alunos, nas regras (2), (5), (6), (7) e (8), além de influências mais pontuais de variáveis, como a porcentagem de alunos que não entraram por vestibular próprio, que aparece na regra (6).

A partir disso, torna-se notório que as regras estão em adequação ao modelo de regressão logística.

8 Conclusões

Através dos resultados obtidos pelo modelo de regressão logística e pelas regras extraídas pelo algoritmo JRip, foi possível definir diversos fatores relevantes para a qualidade de uma instituição de ensino, como a influência do número de professores que param a sua formação no grau de especialização. Observou-se que esta situação tende a trazer resultados negativos da instituição.

Também foi perceptível que muitos dos fatores selecionados penalizam instituições privadas, visto que o alto número de alunos que ingressam no curso através de vestibular próprio ou seleção simplificada mostrou-se como um aspecto negativo. Estes métodos de ingresso são mais frequentemente observados em instituições particulares, levantando questionamentos acerca da qualidade dessas escolas.

Outros aspectos que se mostraram relevantes estão ligados a fatores sociais, como a influência do número de mulheres enquanto alunas e enquanto professoras. A alta proporção de mulheres no primeiro caso leva a bons resultados, no entanto, a situação se inverte para as mulheres no corpo docente. Acredita-se que esse fato deve-se a diferença de pensamento das gerações, mudanças no que se refere ao papel da mulher na sociedade, além de questões relacionadas ao incentivo e acesso desse grupo à educação.

Aspectos socioeconômicos foram considerados, ainda, no que se refere a raça, em que se analisa a influência da proporção de estudantes e professores pardos. Uma vez que esta proporção mostrou-se desfavorável à classe alvo, questionamentos podem ser levantados a respeito do acesso à educação que as pessoas pardas recebem no Brasil.

Outro aspecto interligado ao socioeconômico está relacionado à alta idade média dos alunos, pois acredita-se que esses estudantes não puderam ter acesso à universidade enquanto mais jovens, por necessidade de trabalhar, por exemplo.

Assim, por meio deste estudo tornou-se possível analisar não apenas os diferentes aspectos que influenciam na qualidade das Instituições de Ensino Superior do Brasil, mas também as dificuldades que diversos grupos sociais enfrentam para conquistar uma formação de qualidade.

Referências

[1] Adeodato, Paulo J.L. Data Mining Solution for Assessing Brazilian Secondary School Quality Based on ENEM and Census Data. In: 13th CONTECSI - International Conference on Information Systems and Technology Management, 2016, São Paulo-SP. Proc. of the 13th CONTECSI - International Conference on Information Systems and Technology Management. São Paulo-SP, 2016. p. 1112-1124.

[2] Scremin, Greice, and Daniela da Silva Aimi. "Qualidade na educação superior: conceitos e visões." Políticas Educativas 2.1 (2008).

[3] Mapa do Ensino Superior no Brasil. disponível em <http://convergenciacom.net/pdf/mapa-ensino-superior-brasil-2015.pdf> acessado em 17.07.2017

[4] As 100 melhores universidades do mundo, segundo o CWUR. disponível em<http://exame.abril.com.br/carreira/as-100-melhores-universidades-do-mundo/>

[5] Microdados Censo da Educação Superior 2014. Disponível em <http://dados.gov.br/dataset/microdados-do-censo-da-educacao-superior/resource/572ea703-014a-47eb-8fd6-cdc1d0e2d60c> acessado em 18.07.2017

[6] Microdados Enade 2014. Disnponível em <http://dados.gov.br/dataset/microdados-do-exame-nacional-de-desempenho-de-estudantes-enade/resource/452d1f80-4070-40b9-9b67-a3ad9ba07c8b> acessado em 18.07.2018

[7] Variável dummy?. Disponível em <http://webartigos.com/artigos/variavel-dummy/97922> acessado em 18.07.2017

[8] Shearer, C.: The CRISP-DM model: The new blueprint for data mining. Journal of Data Warehousing , 5 (4), pp. 13–22, 2000.

[9] Conceito preliminar do curso. Disponível em

<http://portal.inep.gov.br/conceito-preliminar-de-curso-cpc-> acessado em 20.07.2017