

Identificação dos perfis de estudantes com nota elevada no ENADE

Identifying students profiles with high marks in ENADE

Eduardo Barbosa Rigueira Cavalcanti^a, Joel Mateus Lopes Ferreira^b

^a edurigueira@gmail.com

^b joel.jfx@gmail.com

Resumo: Entender as razões do desempenho superior dos alunos no Exame Nacional de Desempenho do Estudante (ENADE) é muito importante em vários aspectos. Esse conhecimento, além de contribuir para a formulação de políticas públicas no contexto da educação superior, permite que as instituições de ensino promovam estratégias para melhorar sua posição no ranking da educação de qualidade no Brasil. Este artigo aplica a metodologia CRISP-DM para construir modelos que expliquem os fatores que podem explicar o sucesso do aluno no exame, bem como antecipar suas chances de se classificar no grupo dos mais bem colocados. Foi utilizada a técnica de Regressão Logística Binária, tomando-se variáveis independentes extraídas dos microdados fornecidos pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), referentes ao exame de 2017.

Palavras chave: ENADE; Desempenho; Mineração de Dados.

Abstract: Understanding the reasons for superior performance of students in the National Exam of Student Performance (ENADE) is very important in many aspects. This knowledge, beyond contributing to the formulation of public policies in the context of higher education, enables educational institutions to promote strategies to improve their position regarding the quality education ranking in Brazil. This paper applies the CRISP-DM methodology to build models for explaining the factors that can explain the student's success in the exam, as well as anticipating their chances of ranking in the group of the best placed. The Binary Logistic Regression technique was used, taking independent variables extracted from the microdata provided by the National Institute of Educational Studies and Research Anísio Teixeira (INEP), referring to the 2017 exam.

Keywords: ENADE; Performance; Data Mining Dados.

1. Introdução

O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), autarquia federal vinculada ao Ministério da Educação (MEC), tem como missão promover estudos, pesquisa e avaliações sobre o Sistema Educacional Brasileiro.

Adicionalmente, o INEP tem por objetivo subsidiar a formulação e implementação de políticas públicas para a área educacional a partir de parâmetros de qualidade e equidade.

Por meio da Lei 10.861, de 14 de abril de 2004, foi criado o SINAES - Sistema Nacional de Avaliação da Educação Superior, no qual está contemplado o ENADE - Exame Nacional de Desempenho de Estudantes, como instrumento de avaliação de qualidade dos cursos

A responsabilidade de realizar o ENADE é do INEP, a sua concepção e operacionalização está

a cargo da Diretoria de Avaliação de Educação Superior.

Para melhor atender às demandas sobre o ENADE, além de garantir a transparência requerida pela sociedade e promover a participação dos diferentes agentes sociais envolvidos nas ações educativas, o INEP disponibiliza em seu site os microdados do ENADE.

O objetivo do ENADE é o acompanhamento do processo de aprendizagem e do desempenho acadêmico dos estudantes em relação aos conteúdos programáticos previstos nas diretrizes curriculares do respectivo curso de graduação, além de avaliar o desenvolvimento de competências e habilidades necessárias para formação do profissional no mercado de trabalho.

Em 2017, o ENADE foi realizado no dia 26 de novembro, quando se avaliou os estudantes dos cursos que conferem

- diploma de bacharel em: Arquitetura e

Urbanismo, Engenharia Ambiental, Engenharia Civil, Engenharia de Alimentos, Engenharia de Computação, Engenharia de Controle e Automação, Engenharia de Produção, Engenharia Elétrica, Engenharia Florestal, Engenharia Mecânica, Engenharia Química, Engenharia e Sistema de Informação;

- diploma de bacharel ou licenciatura em: Ciência da Computação, Ciências Biológicas, Ciências Sociais, Filosofia, Física, Geografia, História, Letras - Português, Matemática e Química,
- diploma de licenciatura em: Artes Visuais, Educação Física, Letras - Português e Espanhol, Letras - Português e Inglês, Letras - Inglês, Música e Pedagogia;
- diploma de tecnólogo em: Análise e Desenvolvimento de Sistemas, Gestão da Produção Industrial, Gestão da Tecnologia da Informação e Redes de Computadores.

Diante desse contexto, o objetivo do presente trabalho é identificar o perfil dos estudantes, a partir de estudos de mineração de dados, que obtiveram as notas mais elevadas no exame do ENADE no ano de 2017.

Como problema de pesquisa tentamos resolver a seguinte pergunta: Quais as variáveis relevantes para explicar a performance dos alunos no ENADE?

Nesse contexto, a abordagem da pesquisa é quantitativa, por meio da análise e tratamento dos microdados disponibilizados pelo INEP e coletados no ENADE 2017. Os sujeitos da pesquisa são os alunos de graduação presentes ao Exame Nacional de Desempenho de Estudantes - ENADE 2107. É utilizada a metodologia CRISP-DM (Chapman et al., 2000) no tratamento e modelagem dos dados.

A relevância do trabalho está na diversidade de aplicações dos modelos obtidos e seus resultados, bem como na necessidade de aplicação de técnicas quantitativas de análise exploratória para transformações dos dados do ENADE em informação útil para o entendimento do problema de pesquisa.

2. Referencial Teórico

2.1. ENADE

Para revisão de literatura sobre o Exame Nacional de Desempenho de Estudantes (ENADE), procurou-se o acrônimo ENADE nos campos título ou assunto da ferramenta de busca do

Portal Capes, obtendo-se 61 artigos publicados em periódicos revisados por pares, entre setembro de 2006 a agosto de 2019.

De forma a complementar a pesquisa, igual procedimento foi adotado na base Scopus (Elsevier), com retorno de 39 artigos publicados em revistas especializadas e 7 apresentados em conferências, de julho de 2006 a setembro de 2019. Do total de artigos selecionados, um foi descartado por não possuir relação com o ENADE. O apêndice 1 relaciona os artigos encontrados.

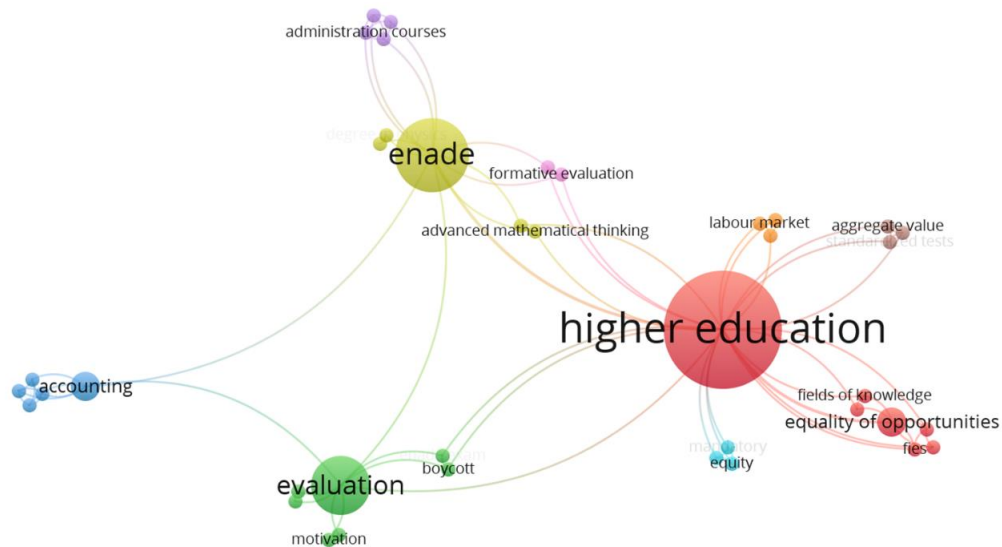
A disponibilidade dos microdados do ENADE, sua completude e extensa série histórica estimula a realização de pesquisas quantitativas. Nesse sentido, Leão et al. (2018) afirmam que os microdados do ENADE, disponibilizados pelo INEP, são uma rica fonte de informações para a melhoria da qualidade da educação oferecida pelas instituições de ensino superior (IES), mas que requerem a aplicação de técnicas de *data mining* para alcançar os padrões do processo de aprendizado e performance acadêmica de estudantes de diferentes cursos.

Essa percepção é constatada ao se analisar os 45 artigos selecionados da base Scopus (Elsevier), onde observa-se que predominam pesquisas centradas na avaliação de performance de estudantes de cursos específicos, de IES ou comparativos de performance. Dentre tais comparativos destaca-se a confrontação de desempenho entre estudantes de IES públicas e privadas, EAD ou de beneficiários de políticas de acesso à educação de nível superior, em grande parte com a utilização de métodos quantitativos para a obtenção de resultados.

Das 45 pesquisas, 34 são focadas na avaliação de áreas específicas do conhecimento, seja pela captura de performance dos estudantes a elas pertencentes ou pela pontuação dos respectivos cursos, com abrangência regional ou nacional. Elas estão distribuídas da seguinte maneira: 35% relativas a áreas das Ciências Exatas, 47% nas áreas de Saúde e Ciências Biológicas e 18% nos domínios das Ciências Sociais e Humanas.

A Figura 1 apresenta uma rede de coocorrência de termos, formada a partir da busca efetuada na base Scopus.

Em sua grande maioria, as pesquisas geram conhecimento passível de ser usado na formulação de políticas públicas e muitas outras permitem traçar estratégias para a melhoria da qualidade de ensino nas IES. Alguns dos artigos

Figura 1. Rede de coocorrência de termos.

destinam-se a avaliações mais específicas de políticas públicas, tais como equidade e incentivos de acesso à educação superior como mecanismos de cotas, ProUni e Fies.

Também se destaca a ocorrência de pesquisas descritivas, muitas longitudinais, que visam traçar o perfil sociodemográfico dos estudantes de graduação, qualificar a oferta e colocação de profissionais no mercado de trabalho e avaliar tendências de privatização do ensino superior na área de saúde. Outros estudos buscam avaliar o conteúdo das questões do exame, capturar a percepção e satisfação dos estudantes com as IES e medir o ganho de conhecimento entre estudantes novatos e veteranos.

Dentre as pesquisas quantitativas, destacam-se por sua abordagem as que seguem.

Knop e Collares (2019), que avaliam influência da origem social na probabilidade de concluir diferentes cursos de ensino superior, por meio do uso da regressão logística para análise do impacto da escolaridade dos pais e da renda familiar, levando em consideração os diferentes retornos econômicos da graduação.

Os resultados mostram que alunos com pais mais instruídos geralmente têm maior probabilidade de se formar e obter maiores recompensas econômicas no mercado de trabalho, embora essa relação não tenha sido determinante nas instituições públicas. Já a renda familiar aumenta as chances de obtenção de um diploma em instituições públicas e privadas, com o mesmo efeito, independentemente do retorno econômico do aluno.

Bezerra e Tassigny (2018), procuram medir por meio de modelos Probit de escolha binária

a probabilidade dos estudantes de administração de empresas participantes dos programas de financiamento Fies e ProUni venham a alcançar uma boa performance acadêmica no ENADE 2015.

Já o objetivo do estudo de Anbar Neto et al. (2018) foi avaliar os fatores que afetam as notas do Exame Nacional de Desempenho dos Cursos de Medicina no Brasil. Para isso foram examinados os fatores que influenciaram as pontuações do ENADE 2010 em 100 programas médicos. As variáveis de infraestrutura acadêmica e de saúde foram investigadas. A análise dos dados foi baseada no desempenho dos programas médicos no ENADE 2010, com os programas divididos em dois grupos: ENADE 1-2 (pontuação insatisfatória) versus ENADE 4-5 (pontuação satisfatória).

Na análise univariada, foram associadas ao melhor desempenho no ENADE as variáveis de universidade, instituição pública, o menor número de vagas por curso, menor número de professores especialistas, maior número de doutores, programa de pós-graduação strictu sensu, maior duração do curso. Na análise multivariada da Regressão Logística Binária, quatro variáveis foram associadas de forma independente ao melhor desempenho do ENADE: instituição pública, menor número de vagas por curso, programa de pós-graduação strictu sensu e maior duração do curso.

Santos Neto, Dantas e Machado (2017) buscam modelar a satisfação dos estudantes de engenharia industrial de instituições privadas de ensino superior, a partir dos microdados do ENADE 2011. As variáveis utilizadas foram

envolvimento do professor, interesse do estudante, demandas do curso, organização do curso, satisfação geral e infraestrutura. Foram empregados os métodos de análise fatorial confirmatória e modelagem de equações estruturais. A hipótese proposta no modelo indica que o envolvimento do professor é chave para a satisfação geral do estudante.

Rossetto e Gonçalves (2015), por meio de estatísticas descritivas e de um modelo logístico multinomial, usando os microdados de 2008 a 2010, provêm evidências de que o sistema de equidade de acesso à educação de nível superior, compreendendo políticas de cotas, Fies e ProUni, não oferece os mesmos benefícios para pessoas em igual situação e conclui com uma discussão sobre possíveis reformas.

Uma análise de performance dos estudantes de matemática e Ciências das Universidades Brasileiras, baseadas nas notas do ENADE 2014, foi conduzida por Silva et al. (2017). Os autores utilizaram técnicas de regressão linear múltipla pelo método *stepwise* e perceberam que variáveis como curso, tempo do curso, instituição, gênero, deficiência e cidade das instituições são relevantes e influentes na performance dos alunos que fizeram parte do exame em 2014.

A pesquisa anterior possui similitudes com o presente estudo, permitindo avaliar se os fatores então encontrados como relevantes à época do ENADE 2014 permanecem discriminantes da performance dos estudantes do ENADE 2017.

2.2. Mineração de Dados

A mineração de dados pode ser entendida como um processo analítico que comporte o uso de uma grande quantidade de dados, na busca por relacionamentos sistemáticos e padrões consistentes entre variáveis, com objetivo de aplicá-los a um novo conjunto de dados para obtenção de resultados.

Com a evolução da mineração de dados e o aumento da capacidade de processamento, houve a necessidade de padronizar a necessidade de sistematizar as melhores práticas aprendidas em uma metodologia de aplicação comum.

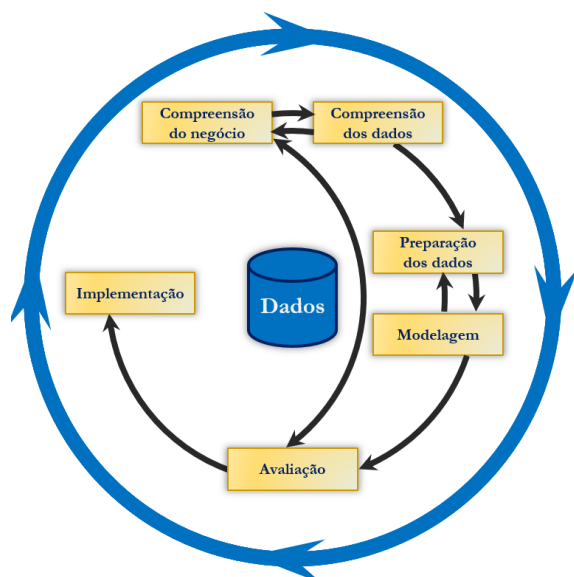
Chapman et al. (2000), descrevem a metodologia *Cross-Industry Standard Process for Data Mining* (CRISP-DM), que pode ser descrita em seis etapas, conforme mostrado na Figura 2:

- **Compreensão do negócio:** no qual busca-se entender as metas e os objetivos e requisitos do projeto de uma perspectiva comercial e

depois transformar esse conhecimento em uma definição de problema de mineração de dados e um plano preliminar.

- **Compreensão dos dados:** compreende a coleta inicial e familiarização com os dados de forma a identificar problemas de qualidade dos dados, desenvolver intuições preliminares sobre eles ou detectar subconjuntos interessantes de dados para formar hipóteses de informações ocultas.
- **Preparação de dados:** a fase de preparação de dados abrange todas as atividades que permitem a construção do conjunto de dados final a partir dos dados brutos iniciais.
- **Modelagem:** são selecionadas e aplicadas técnicas de modelagem. Como algumas técnicas têm requisitos específicos para a forma dos dados, pode ser necessário o retorno para a etapa de preparação dos dados.
- **Avaliação:** Uma vez que um ou mais modelos com boa qualidade sejam construídos, eles precisam ser testados para garantir sua generalização e que todos os principais problemas de negócios tenham sido tratados adequadamente. O resultado final é a seleção do modelo campeão.
- **Implantação:** geralmente significa implantar representações de código de modelo no sistema operacional para pontuar ou categorizar novos dados à medida que ocorrem e criar um mecanismo para usar essas novas informações para solução do problema de negócio original. A representação de código também deve incluir todas as etapas de preparação

Figura 2. Etapas metodologia CRISP-DM.



Fonte: Chapman et al. (2001).

de dados antes da modelagem, para que o modelo processe os novos dados brutos da mesma maneira que durante o desenvolvimento do modelo.

O presente estudo foi conduzido de acordo com as etapas do CRISP-DM, à exceção da etapa de implementação inexistente, e será mostrado de forma simplificada.

3. Etapas do CRISP-DM

3.1. Compreensão do problema

O Exame Nacional de Desempenho dos Estudantes (ENADE) avalia o rendimento dos estudantes que estão concluindo os cursos de graduação. A inscrição é obrigatória para estudantes concluintes habilitados aos cursos de bacharelado, licenciatura e superiores de tecnologia vinculados às áreas de avaliação da edição.

O objetivo do presente estudo é identificar as principais variáveis que expliquem o perfil dos estudantes que tiraram nota elevada no ENADE. Para isso, será utilizado modelagem estatística com técnica de regressão logística binária.

3.2. Compreensão dos dados

As informações relativas ao ENADE são disponibilizadas no portal do INEP (<http://portal.inep.gov.br/microdados>), o qual apresenta dados desde o ano de 2004 - data de início da aplicação do programa.

Para a elaboração do trabalho foram utilizados dados referentes ao ano de 2017, último período disponível. O arquivo referente aos microdados de 2017 possui 150 variáveis (colunas), as quais apresentam as informações, dividida por sessões: (i) Informações da Instituição de Ensino Superior e do Curso, (10 variáveis); (ii) Informações do Estudante (7 variáveis); (iii) Número de Itens da parte objetiva (8 variáveis); (iv) Vetores (8 variáveis); (v) Tipos de Presença (6 variáveis); (vi) Tipos de situação das questões da parte discursiva (5 variáveis); (vii) Notas na formação geral e componente específico (16 variáveis); (viii) Questionário de percepção da prova (9 variáveis); (ix) Questionário do Estudante (68 variáveis); (x) Específica para estudantes de Licenciatura (13 variáveis).

Adicionalmente, o arquivo de microdados totaliza 537.436 registros, referente à quantidade de estudantes inscritos no exame. Considerando que o objetivo é avaliar o perfil dos

estudantes que possuem maiores notas na avaliação, foram excluídos do arquivo os registros correspondentes aos estudantes que não compareceram na avaliação, bem como os que compareceram e entregaram a prova em branco. Nesse sentido, do arquivo original contendo 537.436, o arquivo foi reduzido para 449.206 registros. Foram também realizadas análises descritivas para conhecimento do perfil do público alvo e das distribuições das variáveis disponibilizadas.

O arquivo disponibilizado pelo INEP fornece 16 variáveis relacionada à avaliação do estudante, composta de notas específicas e de formação, as quais também são divididas entre parte objetiva e discursiva. Para esse estudo, optou-se por utilizar a Nota Final (NT_GER) que corresponde à nota bruta da prova, resultante da média ponderada da formação geral (peso 25%) e componente específico (peso 75%). A nota varia de 0 a 100.

O gráfico da Figura 3 apresenta o histograma da nota final dos estudantes no ENADE de 2017. A média da nota final do referido ano foi de 43,59. As demais estatísticas de distribuição são apresentadas na Tabela 1.

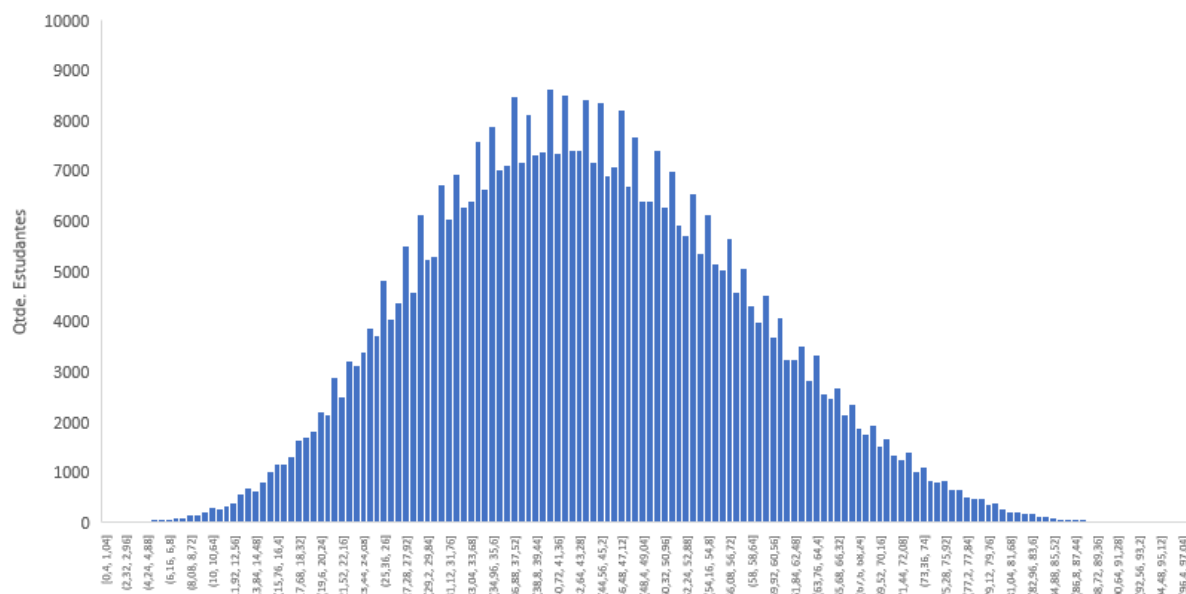
A partir da distribuição da nota final do ENADE, definiu-se que o objetivo seria estudar o perfil dos estudantes que tiraram nota acima de 60 pontos, os quais correspondem ao percentil P86 (P86,97%). Ou seja, 14% dos estudantes, que correspondem a 58.902 estudantes do total de 449 mil, tiraram nota acima de 60 pontos no exame.

Com objetivo de conhecer a distribuição das demais variáveis potenciais participantes do modelo em relação à nota final, foram realizadas nas análises univariadas apresentadas nos Tabelas de 2 a 11.

3.3. Preparação dos dados

Com o objetivo de preparar os dados para modelagem, foi selecionada uma amostra balanceada. O balanceamento da amostra é realizado a partir da variável target, que neste caso é a nota final do estudante no ENADE. A variável target ou variável de interesse foi dividida em duas categorias: (i) “Público 0”: estudantes que obtiveram nota final inferior a 60 pontos, e (ii) “Público 1”: estudantes que tiraram nota final igual ou superior a 60 pontos no ENADE.

A amostra balanceada é composta com a mesma quantidade de estudantes no “Público 0” e “Público 1”. Nesse sentido, considerando

Figura 3. Distribuição das notas dos estudantes do ENADE 2017.**Tabela 1:** Estatísticas da distribuição das notas.

N Válido	449.206	Média	43,592
Ausente	0	Erro da media padrão	0,0209
Percentis		Mediana	42,900
0		Modo	44,6
10	25,800	Desvio Padrão	12,9871
20	31,300	Variância	195,638
30	35,500	Assimetria	,198
40	39,300	Erro de assimetria padrão	,004
50	42,900	Curtose	-,308
60	45,700	Erro de curtose padrão	,007
70	50,900	Intervalo	97,7
80	55,700	Mínimo	,4
90	62,400	Máximo	98,1
95	67,700		
100	98,100		

que havia 449 mil estudantes no total, e que 58 mil (14%) possuíam a nota final igual ou superior a 60 pontos (“Público 1”), foi elaborada uma amostra aleatória de 30 mil estudantes associada ao “Público 0” e 30 mil estudantes associada ao “Público 1”, resultando em uma amostra final de 60 mil registros.

Essa base corresponderá à Base de Treinamento ou Base de Modelagem. O restante dos registros, representados pela quantidade de 389 mil estudantes, referem-se à Base de Validação.

Tabela 2. Análise univariada da variável 1 - Categoria Administrativa da IES.

Categoria	Contagem	Média
Pública Federal	104.717	48,1
Pública Estadual	51.114	44,6
Pública Municipal	4.944	41,5
Privada com fins lucrativos	161.219	40,9
Privada sem fins lucrativos	125.414	43,1
Especial	1.798	37,1
Total	449.206	43,6

Tabela 3. Análise univariada da variável 2 - Modalidade de ensino do curso.

Categoria	Contagem	Média
EAD	90.992	40,6
Presencial	358.214	44,4
Total	449.206	43,6

Tabela 5. Análise univariada da variável 5 - Tipo de turno do curso de graduação.

Categoria	Contagem	Média
Matutino	63.512	45,1
Vespertino	19.241	45,0
Integral	85.292	47,8
Noturno	280.785	41,9
Total	448.830	43,6

Tabela 7. Análise univariada da variável 7 - Quantas pessoas da sua família moram com você?

Categoria	Contagem	Média
Vazio	6.993	37,9
Nenhuma	49.410	46,4
1 pessoa	68.055	44,4
2 pessoas	102.345	43,9
3 pessoas	111.817	43,8
4 pessoas	64.760	42,3
5 pessoas	27.441	41,3
6 pessoas	10.469	40,6
7 pessoas ou mais	7.916	40,3
Total	449.206	43,6

Tabela 9. Análise univariada da variável 9 - Alguém da sua família concluiu um curso superior?

Categoria	Contagem	Média
Vazio	6.993	37,9
Sim	290.780	44,6
Não	151.433	42,0
Total	449.206	43,6

Tabela 4. Análise univariada da variável 3 - Região geográfica da IES.

Categoria	Contagem	Média
Norte	30.128	40,0
Nordeste	79.928	43,1
Sudeste	208.115	44,2
Sul	97.109	44,2
Centro-Oeste	33.926	42,3
Total	449.206	43,6

Tabela 6. Análise univariada da variável 6 - Estado civil do estudante.

Categoria	Contagem	Média
Vazio	6.993	37,9
Solteiro	302.526	44,7
Casado	108.744	41,4
Divorciado	13.517	41,5
Viúvo	1.556	38,7
Outro	15.870	42,6
Total	449.206	43,6

Tabela 8. Análise univariada da variável 8 - Qual a renda total da sua família, incluindo seus rendimentos?

Categoria	Contagem	Média
Vazio	6.993	37,9
Até 1,5 SM	99.710	40,1
De 1,5 a 3 SM	128.182	42,4
De 3 a 4,5 SM	90.662	44,2
De 4,5 a 6 SM	46.471	45,7
De 6 a 10 SM	46.796	47,6
De 10 a 30 SM	26.571	50,3
Acima de 30 SM	3.821	50,9
Total	449.206	43,6

Tabela 10. Análise univariada da variável 10 - Foram oferecidas oportunidades para os estudantes participarem de projetos de iniciação científica e de atividades que estimularam a investigação acadêmica?

Categoria	Contagem	Média
Discordo Totalmente	20.971	42,5
Discordo Muito	19.623	43,5
Discordo Pouco	34.498	43,4
Concordo Pouco	58.782	43,7
Concordo Muito	89.693	43,9
Concordo Totalmente	190.290	44,2
Não sei responder	18.701	40,6
Não se aplica	9.655	40,2
Total	442.213	43,7

Tabela 11. Análise univariada da variável 11 - Os professores demonstraram domínio dos conteúdos abordados nas disciplinas?

Categoria	Contagem	Média
Discordo Totalmente	3.535	38,3
Discordo Muito	6.322	41,7
Discordo Pouco	20.214	42,7
Concordo Pouco	54.000	44,0
Concordo Muito	127.747	45,0
Concordo Totalmente	226.494	43,2
Não sei responder	2.736	39,1
Não se aplica	1.165	40,6
Total	442.213	43,7

Após a análise exploratória dos dados e o balanceamento da amostra de treinamento, foi realizada uma análise qualitativa para identificação das variáveis que melhor poderiam contribuir para explicação da variável de interesse.

Após a seleção das variáveis por meio de análise qualitativa, a etapa seguinte consistiu na “binagem das variáveis”. A criação de variáveis explicativas em categorias, de acordo com a variável de interesse (*target*), possibilita um maior percentual de acerto do modelo a ser desenvolvido.

Para a binagem, foram utilizadas as métricas WOE e IF, sendo que:

- $WOE = \logaritmo (\% \text{ estudantes "Público 0"} / \% \text{ estudantes "Público 1"})$
- $IF = WOE * (\% \text{ "Público 0"} - \% \text{ "Público 1"}) * 100$

O indicador IF pondera as porcentagens entre o “Público 0” e o “Público 1” em relação à cada categoria ou faixa. Quanto maior o IF, maior o poder de distinção da faixa. No processo de binagem, o objetivo é agrupar categorias que possuem IF semelhante, de forma a reduzir a quantidade de faixas da variável, sem a perda de valor explicativo para a variável.

Adicionalmente, $WOE > 0$ em determinada faixa ou categoria significa que a quantidade de estudantes no “Público 0” é maior que a quantidade de estudantes no “Público 1”, e $WOE < 0$, o inverso, ou seja, a quantidade do “Público 0” é menor que o “Público 1”. Portanto, com o objetivo de obter uma categoria explicativa, sugere-se que não se agrupe faixas de $WOE > 0$ com faixas de $WOE < 0$.

A título de ilustração, será apresentado o processo de binagem da Variável 03 - Região geográfica da IES. A Tabela 12 corresponde às categorias iniciais da região geográfica das Instituições de Ensino Superior. A partir da análise da tabela, verifica-se que para a Região Norte, 64% dos estudantes estão contemplados no “Público 0” e 36% no “Público 1”, ou seja, praticamente 2/3 dos estudantes que cursaram graduação em Instituições de Ensino Superior localizadas na região Norte obtiveram nota no ENADE inferior a 60 pontos.

Aplicando a mesma análise para a região Sudeste, verifica-se que 52% dos estudantes tiraram nota final acima de 60 pontos.

O poder explicativo de cada faixa ou categoria está representada pela métrica IF. Considerando que o IF da região Norte é muito superior às demais regiões (justificada pela proporção de estudantes entre “Público 0” e “Público 1”), e que o WOE das regiões Sul e Sudeste possuem inversão de sinal entre os WOE das regiões Nordeste e Centro-Oeste, optou-se por realizar as seguintes junções, conforme Tabela 13.

A categorização da informação da Região Geográfica em 3 faixas, possibilitou um valor de IF de 7,52. Em que pese a redução não significativa do IF da variável original de 8,02 para 7,52, a diminuição da quantidade de categorias possibilitará um maior ganho para a explicação do modelo.

De forma análoga ao processo acima, todas as demais variáveis selecionadas previamente

Tabela 12. Categorias de região geográfica da IES (variável 3).

Região geográfica da IES	Qtde Público 0	Qtde Público 1	Qtde Total	% Público 0	% Público 1	% Púb 0 / % Púb 1	WOE	IF
Norte	2082	1162	3244	64,2%	35,8%	1,79	0,25	7,18
Nordeste	5344	4945	10289	51,9%	48,1%	1,08	0,03	0,13
Sudoeste	13753	14902	28655	48,0%	52,0%	0,92	-0,03	0,14
Sul	6529	7008	13537	48,2%	51,8%	0,93	-0,03	0,11
Centro-Oeste	2292	1983	4275	53,6%	46,4%	1,16	0,06	0,45
Total	30000	30000	60000	50,0%	50,0%	1,00	0,00	8,02

Tabela 13. Junção de categorias de região geográfica da IES (variável 3)

Região geográfica da IES	Qtde Público 0	Qtde Público 1	Qtde Total	% Público 0	% Público 1	% Púb 0 / % Púb 1	WOE	IF
Norte	2082	1162	3244	64,2%	35,8%	1,79	0,25	7,18
Nordeste e Centro-Oeste	7636	6928	14564	52,4%	47,6%	1,10	0,04	0,21
Sudoeste e Sul	20282	21910	42192	48,1%	51,9%	0,93	-0,03	0,13
Total	30000	30000	60000	50%	50%	1,0	0,00	7,52

Tabela 14. Matriz de correlações das variáveis.

	Target_60	Var01	Var02	Var03	Var04	Var05	Var06	Var07	Var08	Var09	Var10	Var11
Target_60	1	,217**	,078**	,073**	,134**	,182**	,102**	,109**	,183**	,108**	,020**	,038**
Var01		1	,328**	-,193**	,215**	,502**	,222**	,120**	,102**	,094**	,078**	-,027**
Var02			1		,409**	,035**	,387**	,069**	,066**	,045**	,094**	,047**
Var03				1	-,027**	-,081**	-,023**	,115**	,135**	,009*	-,006	,000
Var04					1	,156**	,552**	,058**	,075**	,033**	,052**	,030**
Var05						1	,157**	,121**	,185**	,141**	,042**	-,027**
Var06							1	,115**	,059**	,060**	,037**	,062**
Var07								1	-,009*	,091**	,006	,035**
Var08									1	,279**	,021**	-,077**
Var09										1	,002	,148**
Var09											1	,146**
Var11												1

* Correlação significativa no nível 0,05 (2 extremidades) ** Correlação significativa no nível 0,01 (2 extremidades)

por meio de análise qualitativa foram recategorizadas ou “binadas”.

Na etapa seguinte foi realizada uma análise de correlação entre as variáveis explicativas, com objetivo de identificar uma correlação forte entre essas variáveis. A referida análise é relevante no processo de modelagem e tem por objetivo eliminar a multicolineariedade.

Por meio da análise de correlação, apresentado na Tabela 14, verifica-se que não foi identificada correlação forte, ou seja, correlação acima de |0,7|.

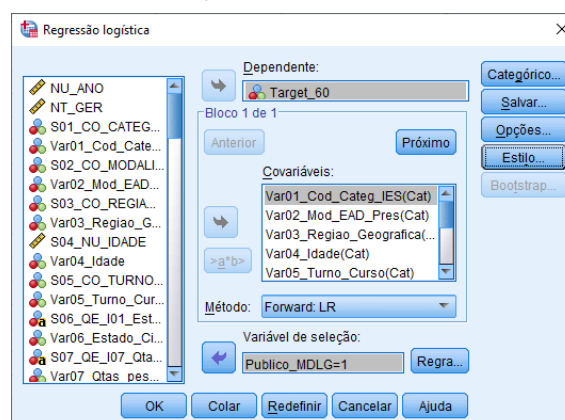
3.4. Modelagem

Realizada a etapa de preparação de dados, a qual contemplou a categorização das variáveis explicativas potenciais e a análise de correlação entre elas, foi desenvolvido o modelo de regressão logística binário.

A Figura 4 apresenta a tela de declaração de variáveis e parametrização de estatísticas para o desenvolvimento do modelo. No campo Variável Dependente é selecionada a variável de

interesse “Target_60”, que representa o valor de 1 para o “Público 1” e valor 0 para o “Público 0”. No campo Covariáveis foram selecionadas as variáveis explicativas potenciais, da variável de “Var01” até a variável “Var11”.

Figura 4. Declaração de variáveis.



Para a definição do método, foi selecionado o *Stepwise Forward LR*. O método *stepwise* realiza a modelagem passo a passo, incluindo ou excluindo as variáveis de acordo com o seu nível

de significância. A opção *Forward LR* parte de um modelo com nenhuma variável selecionada e inclui as variáveis mais relevantes até atingir o modelo mais adequado.

Portanto, a partir da parametrização mencionada foi gerado o modelo de regressão logística, cujos resultados serão apresentados no tópico a seguir.

3.5. Avaliação

A Tabela 15 apresenta a quantidade de registros utilizada na modelagem, sendo que foram utilizados dados de 60.000 estudantes para a modelagem e o restante, 389.206 registros, para a validação do modelo.

Tabela 15. Quantidade de registros.

Casos não ponderados ^a	N	%
Casos selecionados		
Incluído na análise	60000	13,4
Casos ausentes	0	,0
Total	60000	13,4
Casos não selecionados	389206	86,6
Total	449206	100,0

^a Se a ponderação estiver em vigor, veja a tabela de classificação para o número total de casos

O Bloco 0 ou o Bloco Inicial corresponde à elaboração do modelo sem nenhuma variável explicativa (Tabela 16). Partindo dessa premissa, o software definiu, de forma aleatória, que todos os registros do modelo seriam classificados como “Público 1”. A partir dessa definição, o modelo teria 50% de acerto na Base de Modelagem e 7,4% de acerto da Base de Validação.

Nas etapas seguintes, são adicionadas variáveis de acordo com o seu nível de significância (relevância) em relação à variável target. Por meio da Tabela 17 seguir é possível verificar que na Etapa 1 foi incluída a variável “Var01” Categoria administrativa da IES e na Etapa 2, a variável “Var 08” Renda familiar total.

Após a iteração do software e a geração de 10 etapas, chegou-se à versão final do modelo, cujas variáveis foram incluídas por ordem de acordo com o seu nível de importância. A ordem de inclusão das variáveis está apresentada na Figura 5. Por meio da figura, é possível afirmar que dentre as variáveis testadas, a variável mais relevante para explicar a nota dos estudantes que tiraram acima de 60 pontos no ENADE de 2017 é a Categoria administrativa da

IES. A segunda variável por ordem de importância é o Rendimento familiar total do qual o estudante está inserido.

Figura 5. Ordenamento das variáveis incluídas.

- a. Variável(is) inserida(s) na etapa 1: Var01_Cod_Categ_IES.
- b. Variável(is) inserida(s) na etapa 2: Var08_Salario_Minimo.
- c. Variável(is) inserida(s) na etapa 3: Var03_Regiao_Geografica.
- d. Variável(is) inserida(s) na etapa 4: Var04_Idade.
- e. Variável(is) inserida(s) na etapa 5: Var07_Qtas_pessoas_moram.
- f. Variável(is) inserida(s) na etapa 6: Var05_Turno_Curso.
- g. Variável(is) inserida(s) na etapa 7: Var11_Dominio_professores.
- h. Variável(is) inserida(s) na etapa 8: Var09_Familia_curso_superior.
- i. Variável(is) inserida(s) na etapa 9: Var02_Mod_EAD_Pres.
- j. Variável(is) inserida(s) na etapa 10: Var10_Oportunidade_Iniciacao.

Adicionalmente, também é apresentada a Tabela 18, referente à classificação do referido modelo, em que houve acerto de 63,5% na Base de Treinamento e 65,7% na Base de Validação.

Considerando que o objetivo é identificar o perfil dos estudantes que obtiveram nota acima de 60 pontos, representados pelo P86%, é possível concluir que a acurácia de 63,5% é satisfatória para o modelo proposto.

A partir da Tabela 19, é possível construir a fórmula de probabilidade de determinado estudante obter nota acima de 60 pontos no ENADE do ano de 2017, sendo que:

$$\text{Probabilidade (Evento)} = \frac{1}{1 + e^{-(\alpha + \beta_1 x_1 + \beta_2 x_2 + \dots)}}$$

onde α é a constante do modelo, B é o valor correspondente a cada faixa/categoria da variável, e “x” o valor indicador: 1 se o cliente estiver classificado na referida categoria e 0, se não estiver.

4. Conclusão e considerações finais

Por meio do presente estudo, foi possível construir um modelo estatístico com técnica de regressão logística para identificação do perfil dos estudantes que obtiveram nota acima de 60 pontos no ENADE de 2017.

Considerando que esse perfil de estudante corresponde ao percentil P86% da distribuição das notas finais do referido exame, o acerto de 65,7% da Base de Validação pode ser considerado satisfatório para o modelo proposto.

Adicionalmente, o peso de cada variável no modelo, bem como as subdivisões ou categorias desenvolvidas para cada variável preditora, permite identificar a relevância dentre essas informações no contexto do perfil do estudante

Tabela 16. Bloco 0/Bloco Inicial (Tabela de classificação^{a,b}).

Observado			Previsto					
			Casos selecionados ^c			Casos não selecionados ^d		
			Target_60		% correta	Target_60		% correta
			0	1		0	1	
Etapa 0	Target_60	0	0	30000	,0	0	360304	,0
		1	0	30000	100,0	0	28902	100,0
% global			50,0			7,4		

^a A constante está incluída no modelo^b O valor de recorte é ,500^c Casos selecionados Publico_MDLG = 1^d Casos não selecionados Publico_MDLG ≠ 1**Tabela 17.** Extrato das etapas de inclusão de variáveis.

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. para exp(B)	
								Inferior	Superior
Etapa 1 ^a	Var01			2782,227	3	,000			
	Var01(1)	,335	,022	230,254	1	,000	1,398	1,339	1,460
	Var01(2)	,617	,027	511,662	1	,000	1,853	1,756	1,954
	Var01(3)	1118	,022	2633,511	1	,000	3,059	2,931	3,193
	Constante	-,501	,015	1089,053	1	,000	,606		
Etapa 2 ^b	Var01			2444,497	3	,000			
	Var01(1)	,286	,022	140,432	1	,000	1,304	1,148	1,363
	Var01(2)	,611	,028	488,130	1	,000	1,842	1,745	1,945
	Var01(3)	1047	,022	2237,756	1	,000	2,849	2,728	2,975
	Var08			1635,088	2	,000			
	Var08(1)	,468	,020	572,326	1	,000	1,592	1,537	1,660
	Var08(2)	,861	,022	1536,645	1	,000	2,366	2,266	2,470
	Constante	-,810	,018	2064,273	1	,000	,445		
:	:	:	:	:	:	:	:	:	:

^a Variável inserida na etapa 1: Var01_Cod_Categ_EIS^b Variável inserida na etapa 2: Var08_Salario_minimo

...

que obteve nota acima de 60 pontos.

O perfil encontrado para o estudante classificado como de melhor performance, por ordem de importância, foi em aluno de: IES Pública Federal, renda familiar superior a 6 salários mínimos, IES localizada no Sul ou Sudeste, idade do estudante até 24 anos, estudante mora sozinho, estuda em turno integral, professores com domínio do conteúdo, possui familiar com curso superior, modalidade de ensino presencial e teve oportunidades em projetos de iniciação científica.

Cabe destacar um aumento de variáveis capturadas pelo presente modelo em relação às obtidas no trabalho de Silva et al. (2017), com dados do ENADE 2014, bem como a reduzida

aderência entre elas.

Tais análises podem ensejar em ações governamentais específicas para o fomento de melhoria do estudante na sua formação e entrada no mercado de trabalho, refletindo naturalmente em uma maior quantidade de estudantes com nota elevada no exame.

É de se ressaltar que tais conclusões estão restritas ao público de estudantes que realizaram o exame ENADE do ano de 2017.

Como trabalho futuro, sugere-se a aplicação desse modelo nos exames de anos anteriores e no ENADE de 2018 - assim que disponibilizado, a fim de avaliar se o perfil dos estudantes com notas elevadas também pode ser explicado pelo modelo proposto.

Tabela 18. Classificação do modelo^a.

Observado			Previsto					
			Casos selecionados ^c			Casos não selecionados ^d		
			Target_60		% correta	Target_60		% correta
			0	1		0	1	
Etapa 10	Target_60	0	19702	10298	65,7	237827	122477	66,0
		1	11611	18389	61,3	11175	17727	61,3
% global			63,5					

^a O valor de recorte é ,500^b Casos selecionados Publico_MDLG = 1^c Casos não selecionados Publico_MDLG ≠ 1

Tabela 19. Estatísticas das variáveis do modelo.

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I. para exp(B)	
								Inferior	Superior
Etapa 10	Var01			1317,192	3	,000			
	Var01(1)	,119	,024	25,235	1	,000	1,127	1,075	1,180
	Var01(2)	,496	,031	262,786	1	,000	1,127	1,075	1,180
	Var01(3)	,932	,028	1146,220	1	,000	2,540	2,507	2,681
	Var02(1)	-,053	,027	3,995	1	,046	,948	,900	,999
	Var03				2	,000			
	Var03(1)	,425	,042	100,329	1	,000	1,530	1,408	1,663
	Var03(2)	,778	,041	355,154	1	,000	2,177	2,008	2,361
	Var04			396,297	2	,000			
	Var04(1)	,080	,024	10,688	1	,001	1,083	1,032	1,136
	Var04(2)	,408	,023	307,076	1	,000	1,504	1,437	1,574
	Var05			198,477	2	,000			
	Var05(1)	,302	,023	168,250	1	,000	1,353	1,293	1,416
	Var05(2)	,241	,025	90,972	1	,000	1,272	1,211	1,337
	Var07			294,542	2	,000			
	Var07(1)	,316	,021	223,731	1	,000	1,372	1,316	1,430
	Var07(2)	,463	,031	219,843	1	,000	1,589	1,494	1,639
	Var08			1013,842	2	,000			
	Var08(1)	,377	,020	340,343	1	,000	1,458	1,401	1,518
	Var08(2)	,742	,024	977,483	1	,000	2,101	2,005	2,201
	Var09(1)	,161	,019	68,729	1	,048	1,174	1,131	1,220
	Var10(1)	-,084	,043	3,904	1	,000	0,919	,846	,999
	Var11(1)	,403	,032	158,614	1	,000	1,497	1,406	1,594
	Constante	-2,252	,066	1179,598	1	,000	0,105		

Referências

- Anbar Neto, T., Pereira, P. F., Nogueira, M. L., Gody, J. M. P. & Moscardini, A. C. (2018). Factors that affect the national student performance examination grades of Brazilian undergraduate medical programs, *GMS Journal for Medical Education*, 35(1).
- Bezerra, M. E. G. & Tassigny, M. M. (2018). The relationship between student funding policy and academic achievement of management students in ENADE, *Education Policy Analysis Archives*, 26.
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: step-bystep

- data mining guide', CRISP-DM Consortium.
<ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf>.
- Knop, M. & Collares, A. C. M. (2019). The influence of social origin on the probability of obtaining a college degree, *Sociedade e Estado*, 34(2), 351-380.
- Leão, H. A. T., Canedo, E. D., Ladeira, M. & Fagundes, F. (2018). Mining ENADE Data from the ULBRA Network Institution, *Advances in Intelligent Systems and Computing*, 738, 287-294.
- Rossetto, C. B. S., Gonçalves, F. O. (2015). Equity in higher education in brazil: A multinomial analysis of the access to entry public policies, *Dados*, 58(3), 791-824.
- Santos Neto, A. S., Dantas, M. J. P., Machado, R. L. (2017). Structural equation modeling applied to assess industrial engineering students' satisfaction according to ENADE 2011, *Producao*, 27.
- Silva, L. F., Rocha, M. E. P. S. da, & Fagundes, R. A. de A. (2017) ENADE: Math and Science Students Performance Analysis, *IEEE Latin America Transactions*, 15(9), 1742-1746.

Sobre os autores

Eduardo Barbosa Rigueira Cavalcanti

É graduado em Engenharia Civil pela Universidade Federal de Juiz de Fora (UFJF, 1995). É mestrando em Governança, Tecnologia e Inovação na UCB/DF, na linha de pesquisa em Governança. Possui 20 anos de experiência no mercado bancário e financeiro, em especial nas áreas de gestão de riscos, crédito e administração de ativos de terceiros.

Joel Mateus Lopes Ferreira

É graduado em Matemática pela Universidade de Campinas (Unicamp, 2005), possui MBA em Gestão Econômica e Financeira de Empresas pela Fundação Getúlio Vargas (FGV/SP, 2008) e especialização em Estatística Aplicada pela Universidade do Distrito Federal (UDF, 2013). É mestrando em Governança, Tecnologia e Inovação na Universidade Católica de Brasília. Possui 17 anos de experiência no mercado bancário, atuando com modelos de gestão de risco e de crédito.