PREVENDO O DESEMPENHO NO ENADE: UMA APLICAÇÃO DE ALGORITMOS DE APRENDIZAGEM DE MÁQUINA

Eduardo Domingues Bordieri

RESUMO

O objetivo deste estudo foi prever o desempenho de novos alunos do curso de administração de empresa do Centro Universitário Cidade Verde (UNIFCV), com base no Questionário do Estudante e dados do Exame Nacional de Desempenho do Estudantes (ENADE) 2018. Para tanto, foi utilizado base de dados formada por informações socioeconômicas e acadêmicas submetida aos classificadores de Regressão Logística, Randon Forest, Arvores de decisão e Naive Bayes. Os resultados apontaram vantagens no uso da Regressão Logística nos dados analisados, apresentado uma acurácia de 63,04% e significância estatística. Além disso, diferentes características individuais apresentaram, segundo o algoritmo de Regressão Logística, probabilidades diferentes de o aluno obter desempenho acima da média no ENADE.

Palavras-chave: ENADE; Algoritmos de Classificação.

ABSTRACT

The objective of this study was to predict the performance of new students in the business administration course at Centro Universitario Cidade Verde (UNIFCV), based on the Student Questionnaire and data from the National Student Performance Exam (ENADE) 2018. used a database formed by socioeconomic and academic information submitted to Logistic Regression, Randon Forest, Decision Trees and Naive Bayes classifiers. The results showed advantages in the use of Logistic Regression in the analyzed data, presenting an accuracy of 63.04% and statistical significance. In addition, different individual characteristics presented, according to the Logistic Regression algorithm, different probabilities of the student achieving above average performance in ENADE.

Keywords: ENADE; Classification Algorithms.

INTRODUÇÃO

De acordo com a Nota Técnica nº 20/2019/CGCQES/DAES do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), o Exame Nacional de Desempenho dos Estudantes (ENADE) é um dos pilares da avaliação do Sistema Nacional de Avaliação da Educação Superior (SINAES), criado pela Lei nº 10.861, de 14 de abril de 2004.

Já o SINAES é composto pelos processos de Avaliação de Cursos de Graduação e de Avaliação Institucional que, junto com o ENADE, formam a base do processo avaliativo que permite conhecer em profundidade o modo de

funcionamento e a qualidade dos cursos e instituições de educação superior (IES) de todo o Brasil.

Dentro deste processo de avaliação do ensino superior no Brasil, o Conceito Preliminar de Curso (CPC) é um importante indicador de qualidade que as IES estão submetidas. A Nota Técnica nº 58/2020/CGCQES/DAES do INEP conceitua que o CPC é um indicador de qualidade que combina, em uma única medida, diferentes aspectos relativos aos cursos de graduação.

Neste sentido, o CPC é constituído por oito componentes, agrupados em quatro dimensões que se destinam a avaliar a qualidade dos cursos de graduação. Uma dessas dimensões engloba o desempenho dos Estudantes; mensurado a partir das notas dos estudantes concluintes no ENADE.

Cabe ressaltar que o desempenho dos estudantes, mensurado pelo ENADE, possui peso de 20% no CPC. Entretanto, é razoável supor que o peso do desempenho dos estudantes no CPC não se limita à nota obtida no ENADE, pois, alunos com bom desempenho, além de obterem notas melhores nas avaliações, são alunos engajados, que procuram se aprimorar e cobrar por melhores ambientes, estruturas e organização didático-pedagógica. Ademais, alunos com bom desempenho melhoram a imagem da IES na sociedade e, consequentemente, atraem novos alunos que buscam melhores desempenhos.

Portanto, dada a importância do desempenho dos estudantes para o êxito das IES, torna-se necessário que as IES não só acompanhem o desenvolvimento dos alunos ao longo da graduação, mas também devem considerar o maior número de informações sobre seus alunos, a fim de encontrar padrões existentes em seu corpo discente e propor ações específicas para seu desenvolvimento.

Desta forma, visando a melhoria contínua no desempenho dos alunos e, consequentemente, da IES, surge a necessidade de monitoramento e desenvolvimento dos alunos de forma mais eficiente e ampla possível. Dentro deste contexto, o presente trabalho busca prever o desempenho dos novos alunos do curso de Administração de Empresas do Centro Universitário Cidade Verde (UNIFCV) no ENADE com base em dados disponibilizados no Questionário dos Estudantes pelo INEP para o ano de 2018.

Por meio de técnicas e classificadores proporcionados pela aprendizagem de máquina, estimou-se o desempenho esperado do aluno com base em suas características socioeconômicas e, consequentemente, a previsão do desempenho para o aluno recém ingresso no curso.

Além desta introdução este trabalho está divido em quatro seções. Na primeira seção é descrito o referencial teórico que contempla os conceitos mais importantes para o desenvolvimento deste trabalho. Na segunda seção é descrita a metodologia utilizada, apresentando a base de dados e os algoritmos de aprendizado de máquina utilizados. Na terceira seção é demostrado os experimentos realizados e os resultados obtidos por meio destes. Por fim, são descritas as conclusões e considerações para trabalhos futuros.

REFERENCIAL TEÓRICO

Para responder às perguntas formuladas anteriormente, utilizaremos um referencial teórico que abrange dois grupos de conhecimento, sendo o primeiro, estudos realizados utilizando de aprendizado de máquina aplicados na área da educação, modelos de previsão de resultados educacionais e mineração de dados educacionais. Por outro lado, complementaremos esta análise com uma revisão bibliográfica formada de trabalhos sobre os sistemas de avalição educacional e desenvolvimento dos alunos, para buscarmos compreender sua aplicabilidade e funcionalidades, dentro do tema abordado pela pesquisa.

A diversos trabalhos na literatura que buscam realizar a previsão do desempenho acadêmico no âmbito das modalidades de ensino presencial e a distância. Manhães et al. (2011) utilizam-se de técnicas de mineração de dados educacionais na tentativa de prever a evasão de estudantes do ensino presencial da Escola Politécnica da Universidade do Rio de Janeiro, sendo testados mais de dez modelos diferentes, com o nível de acurácia média de 75% e 80%, sendo o classificador denominado Floresta Aleatória apresentando um dos melhores resultados.

Gottardo et al. (2012) demostraram a aplicação de dois classificadores com um nível de precisão acima de 75%, para os classificadores de Floresta Aleatória e Multilayer Perceptron, para prever o desempenho dos alunos da modalidade a distância, utilizandose de uma grande quantidade de atributos para representação dos estudantes.

Com um modelo de regressão linear, para prever o desempenho dos estudantes, Rodrigues et al. (2013) utilizam do ambiente virtual Moodle e do uso de series temporais, com diversas variáveis, onde inclui a quantidade de interações dos estudantes ao longo da semana, com vídeos, materiais de apoio e fóruns, disponibilizados no ambiente virtual em seu modelo.

Conforme demostra o estudo de Minaei-Bidgoli et al. (2003), onde foi realizado a aplicação do Combination of Multiple Classifiers para classificar os estudantes conforme a previsão de sua nota final, a técnica aplicada apresentou melhores resultados quando comparado com a aplicação de classificadores individualmente.

Detoni et al. (2015) utilizam de classificadores para predição de reprovação de acadêmicos para cursos de educação a distância a partir da contagem de interações no ambiente virtual Moodle. Segundo o referido estudo, o classificador de Redes Bayesianas apresenta melhor acurácia, conforme são disponibilizadas mais semanas o nível de precisão do classificador apresenta uma melhor eficiência na predição dos alunos reprovados.

Conforme Santos (2012), analisando o desempenho dos alunos do curso de Contabilidade no Exame Nacional de Cursos, para os anos de 2002 e 2003, e nos exames do ENADE em 2006 e 2009, há uma correlação significativa entre as variáveis estudadas (renda familiar, escolaridade dos pais, horas dedicadas ao

estudo e domínio do professor sobre o conteúdo) e as notas finais dos alunos, nos respectivos exames. Para Escobar, Dalfovo e Verdinelli (2010) as instituições de ensino superior que apresentam programas de (Mestrado/Doutorado) obtém um desempenho superior à média na prova do ENADE.

Cruz & Dutt Ross (2012) verificam a relação entre a quantidade percentual de disciplinas quantitativas ofertadas na grade curricular do curso de administração, com a performance dos alunos no exame do ENADE, assim obtendo resultados com correlação positiva e significativa entre as variáveis. Naercio Meneze-Filho (2012) mostra com base em modelos econométricos, que as variáveis com maior poder explicativo para o desempenho escolar são as características familiares e do aluno, educação da mãe, cor, atraso escolar e reprovação prévia, número de livros, presença de computador em casa e trabalho fora de casa.

METODOLOGIA

O presente estudo buscou a aplicação das técnicas e classificadores de aprendizagem de máquina para a criação de um modelo de previsão de notas para o curso de Administração do Centro Universitário Cidade Verde (UNIFCV), no Exame Nacional de Desempenho dos Estudantes (ENADE). Para tanto, esta seção esta subdivida em duas subseções, nas quais são abordados a construção e descrição da base de dados e a especificação dos classificadores utilizados no modelo de previsão.

BASE DE DADOS

A partir dos dados socioeconômicos obtidas pelo Questionário dos Estudantes no ano de 2018 disponibilizados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), foram selecionadas 308 observações do curso de Administração de Empresas do Centro Universitário Cidade Verde (UNIFCV), sendo 164 do sexo feminino e 144 do sexo masculino.

Para tanto, conforme estudos apresentados na seção anterior, optou como variável dependente a nota do ENADE e 7 variáveis independentes sendo elas sexo, raça, escolaridade da mãe, renda familiar, tempo fora das instituições de ensino, forma de ingresso do curso e nota em português no exame do ENADE. A variável dependente nota do ENADE foi transformada em variável binária. Primeiramente foi calculada a média da nota no ENADE da base de dados em estudo. Em seguida, cada observação foi classificada se está acima da média e abaixo da média.

Com relação a variável escolaridade da mãe, com base no trabalho de Naercio Meneze-Filho (2012), identifica-se que a educação da mãe tem relação com o desempenho escolar do aluno. Desta forma a variável deve apresentar efeito positivo sobre o desempenho dos alunos no exame.

Para variável renda familiar, busca apresentar efeitos positivos sobre o desemprenho escolar, visto que quanto maior o nível de renda maior tende a ser a qualidade de vida, acesso à educação, saúde e alimentação.

Em relação a variável tempo fora das instituições de ensino, foi realizado uma proxy entre o tempo de formação do ensino médio e o tempo para conclusão da graduação em Administração de Empresas. Logo a relação entre os períodos nos informa a quantidade de tempo que o aluno ficou fora das instituições de ensino. Desta forma, quanto mais tempo fora, espera-se que maior as chances de um aluno obter nota abaixo na média no ENADE.

A variável nota em português busca representar a capacidade de interpretação, argumentação e escrita. Espera-se que quanto maior a nota em português maior será a probabilidade de o aluno ter nota acima da média do ENADE. Após a escolha das variáveis dos modelos, foi realizado o tratamento da base de dados bem como o escalonamento das variáveis numéricas. Para as variáveis, Sexo, Raça, Escolaridade da Mãe, Renda Familiar e Forma de ingresso no curso, optou-se em transformá-las de variáveis categóricas em fatores visando a melhor análise do modelo.

Após a escolha das variáveis e o tratamento dos dados, foi realizada a divisão da base de dados em duas partes, sendo uma base Treinamento que corresponde a 70% da base de dados total ou 216 observações e uma base como Teste que corresponde a 30% da total ou 92 observações, desta forma, será realizado o treinamento e o teste dos algoritmos de classificação.

CLASSIFICADORES

REGRESSÃO LOGÍSTICA

A regressão logística consiste em um modelo linear generalizado, cuja resultante é a estimativa da probabilidade de um evento ocorrer em função de um conjunto de variáveis independentes, que podem ser do caráter qualitativo/quantitativo. Redunda a uma classificação dicotômica de ocorrência ou não de um evento, onde os resultados de probabilidade ficam contidos no intervalo de 0 a 1, conforme (MONTGOMERY, D., PECK, E., VINING, G. 2012).

Desta forma o classificador de Regressão Logística considera a variável resposta (y) como dicotômica, ou seja, são lhe atribuídos dois valores: 1 para acometimento de interesse, denominado como sucesso, e 0 para acontecimento complementar, o fracasso. A probabilidade de sucesso é dada por πi, e a de fracasso por 1 - πi. Considerando-se uma série de variáveis aleatórias independentes x1, x2, x3, ..., xn, e um vetor β = β0, β1, β2, ..., βp, formado por parâmetros desconhecidos do modelo, a probabilidade de sucesso é dada por:

$$\pi = \frac{\exp(\beta + \beta x \quad + \beta x \quad +. . + \beta x \quad)}{1 + \exp(\beta + \beta x \quad + \beta x \quad +. . + \beta x \quad)}$$

A probabilidade de fracasso é dada por:

$$1 - \pi = \frac{1}{1 + \exp(\beta + \beta x \quad + \beta x \quad +. . + \beta x \quad)}$$

O logit para o modelo de Regressão Logística é dado por:

$$g(x) = ln\left[\frac{\pi}{1-\pi}\right] x\ \beta = \beta + \qquad \beta\ x$$

NAIVE BAYES

Naive Bayes é um método de classificação probabilística. Baseia-se no teorema de Thomas Bayes, que trata de problemas em que se deseja determinar a probabilidade de um evento B ocorrer na condição de que A já tenha ocorrido, conforme a definição: $P(A \mid B)\ P(B)$

$$P(B \mid A) = \frac{}{P(A)}$$

em que P(B | A) é a probabilidade condicional de ocorrer B dado que A ocorreu; P(B) e P(A) são as probabilidades de ocorrência de B e A, respectivamente; e P(A | B) é a probabilidade condicional de ocorrer A dado que B ocorreu (ZHANG, Z. 2016).

FLORESTA ALEATÓRIA (Randon Forest)

O algoritmo Floresta Aleatória (RF) é um exemplo de método que utiliza classificadores do tipo árvore de decisão (BREIMAN, L. 2001). O algoritmo RF adiciona aleatoriedade ao modelo quando da criação das árvores, na medida que busca as melhores características para fazer a partição dos nodos, a partir de subconjuntos aleatórios das variáveis. Este procedimento gera diversidade, o que normalmente leva à formação de melhores preditores ensemble (BREIMAN, L. 2001). Ao final, cada árvore classificadora é apontada como um componente preditor. Neste sentido, a RF constrói sua decisão por meio da contagem dos votos dos 21 componentes preditores em cada classe e, em seguida, seleciona a classe vencedora em termos de número de votos acumulados dentre todas as “árvores da floresta” (BREIMAN, L. 2001).

ARVORE DE DECISÃO

A árvore de decisão é uma técnica que utiliza a estratégia de analisar um problema complexo para subdividi-lo em problema menores (Neves e Souza, 2009).

RESULTADO

Para realização da avaliação do melhor classificador para o modelo, optou-se pelos testes dos seguintes classificadores: Regressão Logística, Randon Forest,

Arvores de decisão e Naive Bayes, com seus respectivos resultados apresentados na Tabela 1.

Tabela 1 – Métricas para avaliação dos classificadores

Classificadores	Acurácia	P-valor
REGRESSÃO LOGISTICA	0,6304	0,0231
RANDON FOREST	0,5870	0,3779
ARVORES DE DECISÃO	0,5652	0,1257
NAIVE BAYES	0,5978	0,2321

Fonte: Elaborado pelo autor.

De modo geral, conforme pode ser observado na Tabela 1, o classificador que apresentou o melhor resultado em termos de acurácia foi o algoritmo de Regressão Logística. O resultado obtido por este classificador apresentou acurácia acima da média, 0,6304, sendo ainda o único a apresentar significância estatística, p-valor 0,0231.

Encontrado o algoritmo de melhor resultado, aplicou-se em seguida o algoritmo de Regressão Logística a fim de prever desempenho de novos alunos no ENADE segundo características socioeconômicos hipotéticas.

Conforme a Tabela 2, o experimento 1 considera as seguintes características hipotéticas: aluno do sexo feminino, de cor amarela, tendo a mãe cursado o ensino superior, com uma renda familiar mensal de até 10 salários-mínimos nacionais, ingressado no curso sem ação afirmativa e racial, não ter ficado fora de instituição de ensino (entre conclusão do ensino médio e a realização do ENADE foram 4 anos) e nota na prova de português igual a 80. Conforme o algoritmo de Regressão Logística, para o perfil hipotéticos descrito acima, apresenta uma chance de 65% de o aluno apresentar desempenho acima da média no ENADE.

Tabela 2 – Características e resultados dos experimentos

Variável	Experimento 01	Experimento 02
SEXO	FEMININO	MASCULINO
RAÇA	AMARELA	PRETA
ESCOLARIDADE DA MAE	SUPERIOR	FUNDAMENTAL
RENDA FAMILIAR	10 SM	1 a 5 SM
INGRESSO NO CURSO	SEM AÇÃO	AÇÃO COTA RACIAL
TEMPO FORA IE	4 ANOS	6 ANOS
NOTA PORTUGUÊS	80 PONTOS	50 PONTOS
RESULTADO	65%	24%

Fonte: Elaborado pelo autor.

Para o experimento 02, considera-se as seguintes características hipotéticas: aluno do sexo masculino, de cor preta, tendo a mão cursado apenas o ensino fundamental, apresentando uma renda familiar mensal de 1 a 5 salários-mínimos nacionais, ingressado no curso por cota racial, ter ficado fora 2 anos da instituição de ensino (entre conclusão do ensino médio e a realização do ENADE foram 6 anos) e nota na prova de português igual a 50. Desta forma, considerando estas características socioeconômicas hipotéticas, o algoritmo de Regressão Logística prevê uma chance de 24% de o aluno apresentar desempenho acima da média no ENADE.

Conforme os experimentos realizados, observa-se a que o primeiro experimento apresenta melhores chances de obter uma melhor nota no exame em relação ao segundo experimento. Assim pode-se verificar a importância das características individuais e socioeconômicas, como renda familiar, escolaridade da mãe e tempo fora das instituições de ensino, para o desempenho dos alunos na vida escolar e em exames nacionais como no caso do Exame Nacional de Desempenho do Estudantes (ENADE).

CONCLUSÃO

Os resultados obtidos por esta pesquisa apontam a importância de as instituições de ensino realizarem previsões relativas ao desempenho de estudantes em provas como a do Exame Nacional de Desempenho do Estudantes (ENADE).

Com base em inferências e previsões realizadas por modelos de aprendizagem de máquina, os professores e instituições podem acompanhar de maneira individual os estudantes e adotar medidas pedagógicas especificas que maximizem os resultados dos alunos em exames nacionais de avaliação.

Especificamente quanto ao algoritmo selecionado no presente estudo, o algoritmo de classificação Regressão logística apresentou resultados expressivos tanto em termos de acurácia quanto de significância estatística. Apresentando, portanto, possibilidade de sua utilização em um conjunto amplo de atributos e, consequentemente, a estimação do desempenho do corpo discente da instituição de ensino.

A partir destes resultados, futuros trabalhos podem utilizar bases de dados mais extensas e com maior quantidade de atributos. Podendo tanto serem específicos com base curricular de cada curso, quanto a possível comparação entre diferentes exames de avaliação.

REFERÊNCIAS

BREIMAN, L. “Random Forests”, Machine Learning v. 45, n. 1, pp. 5–32, Out. 2001.

CRUZ, Breno de Paula Andrade; SHARLAND, Elisa Maria Rodrigues, FREITAS JR. Antônio de Araújo de. Estrutura Curricular e Enade: há uma Correlação Positiva e

Significativa entre o Percentual de Disciplinas Quantitativas em um Curso de Administração e a nota do curso no Enade? In: Revista de Administração do Gestor, n. 2, v. 2, 2012, p. 61-84.

D. Detoni, R. M. Araujo, and C. Cechinel, "Predição de Reprovação de Alunos de Educação a Distância Utilizando Contagem de Interações," in Anais do Simpósio Brasileiro de Informática na Educação, 2014, pp. 896-905.

ESCOBAR, Maria Andrea Rocha ; DALFOVO, Michael Samir ; VERDINELLI, Miguel Angel . OS ÍNDICES IGC, ENADE E CAPES NOS CURSOS DE ADMINISTRAÇÃO. In: IX Coloquio Internacional sobre Gestão Universitária na America do Sul, 2010, Florianopolis.

E. Gottardo, C. Kaester, and R. V. Noronha, "Previsão de Desempenho de Estudantes em Cursos EAD Utilizando Mineração de Dados: uma Estratégia Baseada em Séries Temporais.," Anais do XXIII SBIE, 2012.

MANHÃES, L.M.B., Cruz, S.M.S., Costa, R.J.M, Zavaleta, J., Zimbrão, G. (2011) “Previsão de Estudantes com Risco de Evasão Utilizando Técnicas de Mineração de Dados”. Anais do XXII SBIE-XVII WIE, p. 150-159.

MENEZES FILHO, Naercio Aquino. Os determinantes do desempenho escolar do Brasil. In: O Brasil e a ciência econômica em debate[S.l: s.n.], v. 1. , 2012.

MONTGOMERY, D., PECK, E., VINING, G., Introduction to Linear Regression Analysis. 5 ed. Nova Iorque, John Wiley & Sons, 2012.

MINAEI-Bidgoli, B., Kashy, A.D., Kortemeyer, G., Punch, F.W. (2003) “Predicting Student Performance: An Application of Data Mining Methods with the Educational

Web-Based System LON-CAPA”. 33ª ASEE/IEEE Frontiers in Education Conference, p. 1-6.

NEVES, A. P.; Souza, D.; Desempenho dos Estudantes das Instituições Públicas e Privadas no ENADE: Um estudo no Estado de Roraima, 2009.

SANTOS, Nalbia de Araujo (2012), Determinantes do desempenho acadêmico dos alunos dos cursos de ciências contábeis, São Paulo, 2012.

ZHANG, Z. “Naive Bayes Classification in R”, Ann Transl Med v. 4, n. 12, pp. 241–245, Jun. 2016.