Experimentos preliminares para predição do conceito ENADE dos cursos de computação no Brasil

Preliminary experiments to predict the ENADE concept for computing courses in Brazil

Renan Guensuke Aoki Sakashita[1], Diego Bertolini[2], André Luis Schwerz[3]

RESUMO

O ENADE é uma avaliação de larga escala aplicada aos concluintes dos cursos de graduação do Brasil, cuja nota é usada como métrica de qualidade do ensino superior e conhecida como Conceito ENADE. Um alto índice proporciona maior visibilidade e investimentos às Instituições de Ensino Superior. Este trabalho tem como objetivo avaliar o desempenho de técnicas de Aprendizado de Máquina para predizer o Conceito ENADE dos cursos de computação no Brasil, usando microdados do exame.
PALAVRAS-CHAVE: predição; ENADE; Conceito ENADE; Aprendizado de Máquina.

ABSTRACT

ENADE is a large-scale assessment applied to graduates of undergraduate courses in Brazil, whose grade is used as a metric of quality in higher education and known as the ENADE Concept. A high index provides greater visibility and investments to Higher Education Institutions. This work aims to evaluate the performance of Machine Learning techniques to predict the ENADE Concept of computing courses in Brazil, using microdata from the exam.
KEYWORDS: prediction; ENADE; ENADE Concept; Machine Learning.

INTRODUÇÃO

O Exame Nacional de Desempenho dos Estudantes (ENADE) é uma prova realizada pelos concluintes de graduação no Brasil cujo objetivo é avaliar seus cursos numa escala de 1 a 5 conhecida como Conceito ENADE (CE) (BRASIL, 2023). Esta avaliação é realizada em um ciclo avaliativo de três anos no qual, em cada ciclo, uma das três grandes áreas de conhecimento é avaliada. A prova apresenta tanto questões específicas relacionadas à Diretriz Curricular Nacional do curso quanto questões de escopo geral. Junto a ela, também é solicitado aos estudantes, o preenchimento de um questionário socioeconômico.

Os dados coletados do exame, do questionário socioeconômico e do CE são disponibilizados de forma aberta em microdados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) (INEP, 2022a). Dessa forma é possível investigar o impacto das informações disponíveis dos cursos e seus estudantes sobre o desempenho alcançado no exame. As descobertas podem ser úteis na elaboração de políticas de educação superior que fomentem o investimento de formação de recursos humanos (MARINHO, 1998).

[1] Bolsista da UTFPR. Universidade Tecnológica Federal do Paraná, Campo Mourão, Paraná, Brasil. E-mail: renansakashita@alunos.utfpr.edu.br. ID Lattes: 8489483541695284.
[2] Docente no Curso de Ciência da Computação. Universidade Tecnológica Federal do Paraná, Campo Mourão, Paraná, Brasil. E-mail: diegobertolini@utfpr.edu.br. ID Lattes: 2264701855770284.
[3] Docente no Curso de Ciência da Computação. Universidade Tecnológica Federal do Paraná, Campo Mourão, Paraná, Brasil. E-mail: andreluis@utfpr.edu.br. ID Lattes: 4954414332524750.

No melhor do conhecimento, nenhuma contribuição anterior investiga a predição do CE com uma grande quantidade de características extraídas dos microdados e o desempenho do conjunto entre vários algoritmos de classificação.

O objetivo deste trabalho é propor uma abordagem baseada em Aprendizado de Máquina para avaliar o impacto das características coletadas no microdados do ENADE para a predição do Conceito ENADE. Para os experimentos, são utilizados os microdados das edições de 2011, 2014, 2017 e 2021 dos cursos na área de computação. Um extenso pré-processamento foi realizado para extração de características dos cursos e criação do conjunto de dados usado no experimento de predição.

MATERIAIS E MÉTODOS

Neste trabalho utilizou-se o método Knowledge Discovery in Databases (KDD) (FRAWLEY, 1992) para orientar o processo de extração de informações não-triviais, desconhecidas e potencialmente relevantes em banco de dados. O método inclui as etapas: pré-processamento, mineração de dados e avaliação.

PRÉ-PROCESSAMENTO

O pré-processamento inclui a seleção, limpeza, transformação e integração de dados. Como resultado, gerou-se um conjunto de dados com características dos cursos de graduação da área de computação no Brasil processadas a partir dos microdados do ENADE.

SELEÇÃO DE DADOS

A etapa de seleção levou em consideração a reformulação dos microdados do ENADE, disponível a partir de 2022, devido à Lei Geral de Proteção de Dados (LGPD) (INEP, 2022a). As respostas das questões do questionário socioeconômico e outras informações foram aleatorizadas entre si e distribuídas por diferentes arquivos de forma que não é possível identificar uma série de informações de um aluno, mas sim informações gerais de um curso de graduação. Por exemplo, a partir dos microdados não pode-se afirmar que um determinado aluno de cor/raça branca tenha 19 anos, mas pode-se inferir que 60% dos alunos de um curso têm cor/raça branca e que a idade média é de 19 anos.

Originalmente, para cada edição do exame, o conjunto de dados é dividido em arquivos que contém o código do curso acompanhado por alguma informação específica como, por exemplo, as respostas do questionário socioeconômico ou a cor/raça declaradas pelos participantes. Por causa dessa fragmentação, foi preciso ordenar o conteúdo de cada arquivo pelo código do curso, para junção e, então, formação do conjunto de dados usado neste trabalho.

Os questionários socioeconômicos não estão padronizados ao longo dos anos. Os dados das edições de 2005 e 2008 apresentam inúmeras diferenças com as edições mais recentes e, por esse motivo, optou-se por não usá-los neste trabalho. Para as demais edições, selecionou-se um subgrupo de atributos comuns em todas as edições (2011, 2014, 2017 e 2021). No total obteve-se 36 atributos: 6 pertencentes aos cursos, 24 do questionário socioeconômico e 6 derivados.

O enquadramento é o processo pelo qual cada curso é vinculado à sua respectiva área de avaliação no ENADE (INEP, 2022b). Neste trabalho, utilizou-se as seguintes áreas de enquadramento relacionadas à computação: Análise e Desenvolvimento de Sistemas (ADS), Bacharelado em Ciência da Computação (BCC), Bacharelado em Engenharia da Computação (EC), Gestão da Tecnologia da Informação (GTI), Licenciatura em Ciência da Computação (LCC), Redes de Computadores (RC) e Bacharelado em Sistemas de Informação (SI).

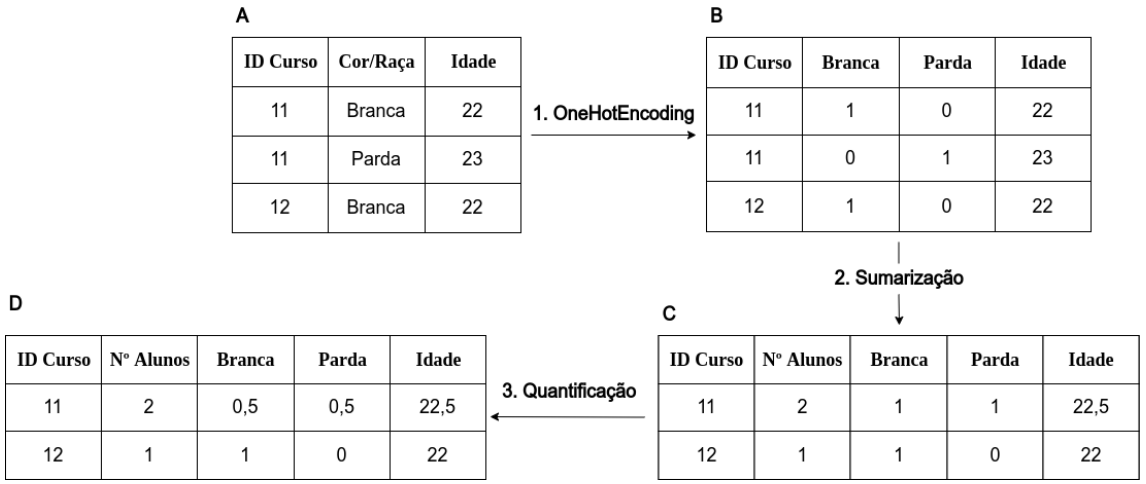
LIMPEZA DE DADOS

Os conjuntos de dados das edições foram analisados em relação à existência de valores não informados a fim de mensurar seu impacto na qualidade dos dados. No entanto, como citou-se anteriormente, as informações foram aleatorizadas entre si, implicando que a exclusão de uma instância por causa de um valor defeituoso levaria à provável eliminação de valores de diferentes alunos de um curso. Desta forma, valores nulos foram tratados como valores não informados.

TRANSFORMAÇÃO DE DADOS

A granularidade original dos microdados do ENADE refere-se às informações dos estudantes. Neste estudo, para obtenção de instâncias referentes aos cursos, houve a combinação da técnica de *one-hot encoding* com procedimentos de sumarização e de quantificação ilustrados na Figura 1. Por exemplo, dado uma amostra A extraída dos microdados, obtém-se o conjunto B por meio do *one-hot encoding* de variáveis categóricas; em seguida, o conjunto C é obtido por meio da soma dos valores dos atributos nominais e média dos numéricos; por fim, obtém-se um conjunto D por meio da divisão das somas dos valores categóricos pelo número de alunos efetivos do curso. Após estas operações de transformação, fez-se também a normalização dos dados utilizando o método Z-Score (ROUT, 2023).

Figura 1. Processo ilustrativo de transformação de dados de alunos em cursos



Fonte: Elaborado pelos autores (2023).

Inicialmente, o conjunto possuía 205.837 instâncias considerando todos os estudantes participantes das edições de 2011, 2014, 2017 e 2021. Após o processo de transformação descrito acima, obteve-se 5.838 instâncias representando cada curso presente no conjunto.

INTEGRAÇÃO DE DADOS

Originalmente, os microdados não contêm o CE dos cursos. Por este motivo, após a etapa de transformação, tornou-se necessária a integração dos conjuntos de dados disponibilizados pelo ENADE que informam o CE (INEP, 2023). A proporção de cursos com CE 1, 2, 3, 4 e 5 pelas edições de 2011, 2014, 2017 e 2021 é apresentada na Tabela 1. Percebe-se um desbalanceamento entre as classes. Há, por exemplo, uma predominância do CE 3 representando por volta de 40% dos cursos nas quatro edições, enquanto os CE 1 e CE 5 possuem um pequeno número de instâncias.

Tabela 1. Distribuição do CE ao longo das edições

Edição	CE 1	CE 2	CE 3	CE 4	CE 5
2011	58 (4,7%)	317 (25,9%)	549 (44,9%)	201 (16,4%)	97 (7,9%)
2014	69 (4,6%)	469 (31,2%)	593 (39,4%)	300 (19,9%)	74 (4,9%)
2017	60 (3,7%)	465 (29%)	636 (39,6%)	349 (21,7%)	95 (5,9%)
2021	64 (4,2%)	502 (33,3%)	604 (40,1%)	281 (18,7%)	55 (3,7%)

Fonte: Elaborado pelos autores (2023).

MINERAÇÃO DE DADOS

Para avaliar a predição do CE, os seguintes classificadores foram utilizados: Árvore de Decisão (DT, do inglês Decision Tree), k-Vizinhos Mais Próximos (k-NN, do inglês k-Nearest Neighbors), Floresta Aleatória (RF, do inglês Random Forest), Máquina de Vetores de Suporte (SVM, do inglês Support Vector Machine) e Perceptron Multicamadas (MLP, do inglês Multilayer Perceptron). Estes classificadores são bem conhecidos na área de Aprendizado de Máquina e usados para uma grande variedade de problemas. Além disso, todos eles já foram usados em trabalhos relacionados com os microdados do ENADE de acordo com (BARBOSA, 2023).

AVALIAÇÃO

Durante o treinamento e avaliação dos modelos foram aplicadas as seguintes técnicas: HalvingGridSearchCV (SCIKIT-LEARN, 2023) para busca de hiperparametros; F1-Score para a avaliação dos modelos; e validação cruzada com 5 divisões. Nota-se que antes da avaliação do conjunto houve a normalização dos dados de forma separada para treino e teste.

RESULTADOS E DISCUSSÕES

Para análise dos microdados do ENADE, realizou-se um experimento utilizando validação cruzada com cinco partições para todos os classificadores destacados anteriormente. Para cada um deles, registrou-se a média e o desvio padrão do F1-Score das iterações em cada modelo de classificação. Utilizou-se o conjunto de dados com as edições de 2011, 2014, 2017 e 2021, sem quaisquer alterações, obtendo os resultados exibidos na Tabela 2. As melhores taxas foram atingidas pelos modelos SVM e MLP, com os valores de 46,84% e 46,55%, respectivamente.

Tabela 2. Comparativo das médias do F1-Score obtidas usando cinco classificadores

DT	k-NN	RF	SVM	MLP
38,17 ± 1,14	43,47 ± 2,06	43,11 ± 2,30	46,84 ± 2,84	46,55 ± 2,00

Fonte: Elaborado pelos autores (2023).

CONCLUSÕES

Neste trabalho apresentou-se um estudo abrangente com várias características extraídas dos microdados do ENADE com o objetivo de predizer o CE dos cursos. Os resultados apontam que o desempenho dos modelos varia de 38,17% a 46,84%. Além disso, para realização do experimento, foi produzido um conjunto de dados com características de cursos da área de computação, que pode ser usado pela comunidade para realizar outras investigações.

Como trabalhos futuros, espera-se melhorar o desempenho da predição com o enriquecimento dos dados com outros conjuntos de dados públicos, e com a combinação dos classificadores.

Uma versão estendida deste trabalho foi aceita para publicação no Simpósio Brasileiro de Informática na Educação (2023), que ocorrerá entre os dias 6 e 10 de novembro em Passo Fundo, RS.

Agradecimentos

Os autores agradecem a Universidade Tecnológica Federal do Paraná (Edital PROPPG nº 02/2022) pelo apoio financeiro.

Disponibilidade de código

Para o desenvolvimento dos modelos, foi utilizado a linguagem de programação Python e a biblioteca scikit-learn (PEDREGOSA, 2011). Para manter a reprodutibilidade da pesquisa, o código-fonte e os dados dos experimentos podem ser encontrados no repositório: https://github.com/RenanGAS/ML-ENADE.

Conflito de interesse

Não há conflito de interesse.

REFERÊNCIAS

BARBOSA, P. O sucesso não é apenas uma questão de sorte: um mapeamento sistemático sobre técnicas de análise do ENADE da área de computação. **Anais do III Simpósio Brasileiro de Educação em Computação**, pp. 59-68, 2023.

BRASIL. Exame Nacional de Desempenho dos Estudantes (Enade). **Instituto nacional de estudos e pesquisas educacionais anísio teixeira**, 2023. Disponível em: https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enade. Acesso em: 01 jul. 2023.

FRAWLEY, W. J. Knowledge discovery in databases: An overview. **AI Magazine**, v. 13, p. 57, 1992.

INEP. ENADE. **Instituto nacional de estudos e pesquisas educacionais anísio teixeira**, 21 jun. 2022a. Disponível em: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enade. Acesso em: 01 jul. 2023.

INEP. Verificação do enquadramento automático vai até 31/8. **Instituto nacional de estudos e pesquisas educacionais anísio teixeira**, 2022b. Disponível em: https://www.gov.br/inep/pt-br/assuntos/noticias/enade/verificacao-do-enquadramento-automatico-vai-ate-31-8. Acesso em: 08 jul. 2023.

INEP. Indicadores de qualidade da educação superior. **Instituto nacional de estudos e pesquisas educacionais anísio teixeira**, 2023. Disponível em: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/indicadores-educacionais/indicadores-de-qualidade-da-educacao-superior. Acesso em: 08 jul. 2023.

MARINHO, A. O aporte de recursos públicos para as instituições federais de ensino superior. **Revista de Administração Pública**, pp. 83-93, 1998.

PEDREGOSA, F. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v.12, pp. 2825–2830, 2011.

ROUT, A. R. Z-score in statistics. **GeeksforGeeks**, 2023. Disponível em: https://www.geeksforgeeks.org/z-score-in-statistics. Acesso em: 05 jul. 2023.

SCIKIT-LEARN. 2. tuning the hyper-parameters of an estimator. **Scikit-Learn**, 2023. Disponível em: https://scikit-learn.org/stable/modules/grid_search.html#searching-for-optimal-parameters-with-successive-halving. Acesso em: 09 jul. 2023.