Previsão do Desempenho no Exame Nacional do Ensino Médio (ENEM) Baseado em Fatores Socioeconômicos

Matheus P.C. Santos,[1] **Ryan B. Ramos,**[2] **Sungwon Yoon,**[3] **Wendel F.Lana**[4]
Escola de Artes, Ciência e Humanidades - Universidade de São Paulo
matheuspecoraro@usp.br,[1] ryanramos@usp.br,[2] sungwon.yoon@usp.br,[3] wendel.lana@usp.br[4]

Resumo

Este artigo tem por objetivo apresentar as propostas iniciais da análise de diferença de desempenho no Exame Nacional de Ensino Médio existente em conta das questões socioeconômicas, utilizando um modelo de inteligência artificial.

Introdução

O Exame Nacional do Ensino Médio (Enem) é uma prova nacional que tem como objetivo avaliar o desempenho do estudante ao fim da escolaridade básica. Desde a sua criação em 1998 até hoje, todo ano tem número de participantes que engloba toda dimensão nacional.

Porém, essa alcance nacional fez tornar visível como há discrepância de desempenho entre regiões segundo o desenvolvimento socioeconômico. Vários indicadores e pesquisas mostram que o nível de desempenho tende a se relacionar à desigualdade social.

Assim, a proposta deste trabalho é mapear quais são os fatores mais determinantes no desempenho de candidatos do ENEM e detectar os alunos que possam apresentar rendimentos baixos na vida acadêmica e que, posteriormente, possam ser oferecidos assistências adequadas para que apresentem desempenhos melhores.

Metodologia

Dados

O conjunto de dados utilizado no presente trabalho é o conjunto de microdados do ENEM de 2020 disponibilizado pelo Inep em 2020, que pode ser obtido em https://download.inep.gov.br/microdados/microdados_enem_2020.zip.

Trata-se de dados realísticos que contêm informações de diversos participantes, como respostas às perguntas do questionário socioeconômico que devem ser autodeclarados na inscrição; dados da prova realizada como ano, nota das provas, gabarito e respostas da parte objetiva das provas; e informações pessoais como idade, faixa etária, etnia e situação escolar.

Os dados dos participantes estão anonimizados, sendo diferenciados por um número de inscrição que impossibilita a identificação de um candidato no exame ou em microdados de pesquisas diferentes.

Arcabouço PEAS

O agente (modelo projetado para a predição), que é único (não sofre influência de algum outro agente), pode ser descrito em termos dos seguintes itens compreendidos no arcabouço PEAS:

- **Ambiente:** conjunto de participantes do ENEM e seus fatores socioeconômicos autodescritos.

  Parcialmente observável: há outros fatores a serem levados em consideração que têm influência no desempenho do aluno no exame que não são descritos pelos dados coletados, tais como empenho, situações e aspirações pessoais.

  Estocástico: partindo apenas do conjunto de fatores possíveis de serem observados e descritos, alunos nas mesmas condições descritas pelas observações podem apresentar desempenhos diferentes. A depender do algoritmo escolhido na implementação do modelo, será necessário tomar uma decisão determinística ou não no momento da predição.

  Episódico: cada entrada terá uma predição específica que não será influenciada por predições passadas (a menos que o modelo seja retreinado após cada nova predição).

  Estático: cada observação se mantém a mesma após a coleta dos dados.

  Discreto: as variáveis de entrada são todas categóricas (variáveis potencialmente contínuas foram discretizadas na coleta de dados) e a saída também é discreta. Assim, temos que o conjunto de estados e ações possíveis também são discretos.

  Desconhecido: embora possa-se ter uma noção intuitiva *a priori* de como e quais fatores têm influência no desempenho do candidato, a função de transição exata não é conhecida. Em decorrência do treinamento do modelo, será possível encontra uma função que tenta aproximar a função real.

- **Sensor:** entrada do modelo com os fatores socioeconômicos do candidato.

- **Atuador:** saída do modelo com a predição da nota do candidato.
- **Medida de desempenho:** se dá pela acurária das predições do modelo proposto.

Solução segundo abordagem

Com a finalidade de detectar os estudantes que possam estar em risco educacional, será avaliado como as questões socioeconômicas impactam no desempenho do ENEM. Para isso, os dados que revelam os contextos sociais serão escolhidos como *features* para a pesquisa.

A correção da prova de ENEM é realizada por metodologia de Teoria de Respota ao Item (TRI) em que cada alternativa tem peso diferente, porém essas informações não são disponibilizadas no conjunto de dados, tampouco são conhecidos propriamente. Assim, esse fato pode dificultar a previsão das notas e comprometer o desempenho do modelo.

Desse modo, serão testados dois modelos: com saída final sendo a quantidade de respostas corretas e média de notas das quatro provas.

Possíveis algoritmos

O problema abordado neste trabalho pode ser resolvido pelo aprendizado de máquina supervisionado, uma vez que, no conjunto de treinamento, há valores de saída esperados — as notas ou quantidade de acertos. Como a tarefa de classificação não é binária, mas sim de classificação multiclasse, os algoritmos de repressão logísitca, de regressão linear e alguns algoritmos de redes neurais como *Support Vector Machine* (SVM) não foram considerados como possíveis soluções do problema, apesar da existência de bibliotecas que forneçam essa suporte para classificação multiclasse ou da possibilidade do problema ser decomposto em vários pequenos problemas binários.

Assim, possíveis algoritmos que podem ser implementados para solucionar o problema são *k-nearest neighbor*, árvore de decisão, *random forest*, *Naive Bayes* e redes neurais.

Resultados

O modelo desenvolvido para o presente trabalho será avaliado pela sua precisão e revocação e, ao todo, por seu F1 *Score* das saídas produzidas.

As notas do ENEM são valores contínuos que variam de 0 a 1000 e há no total 180 questões ao longo de todas as quatro provas, o que abrange uma quantidade muito grande de possibilidades de saídas. Assim, não será avaliado se o modelo realizada a previsão exata, mas se ela faz parte de uma faixa de notas dividida em 100 em 100 ou de acertos dividida em 10 em 10.

Referências

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. **Enem 2020**. Disponível em: ¡https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem¿. Acesso em: 26 abr. 2022

MINISTÉRIO DA EDUCAÇÃO. **ENEM - MEC**. Disponível em: ¡http://portal.mec.gov.br/enem-sp-2094708791¿. Acesso em: 03 mai. 2022

RUSSELL, S.; NORVIG, P. **Inteligência Artificial**. Tradução de Regina Célia Simille. 3º. ed. Rio de Janeiro: Elsevier, 2013