Previsão do Desempenho no Exame Nacional do Ensino Médio (ENEM) Baseado no Questionário Socioeconômico

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

No Brasil, o desempenho dos estudantes ao fim da escolaridade básica se mede pelo Exame Nacional do Ensino Médio (ENEM). Por ser de alcance nacional, é possível de observar candidatos de diversas regiões com diferentes perfis socioeconômicos. No cenário atual do país em que o nível socioeconômico afeta a qualidade da educação, o presente trabalho visa a avaliar quais características socioeconômicas têm mais impacto na nota do ENEM. Para isso, foram utilizadas as respostas dos questionários socioeconômicos do Enem, disponibilizadas pelo INEP, e foi treinado o algoritmo de árvore de decisão com base nesses dados. A avaliação do classificador foi realizada por sua acurácia, precisão, revocação e medida F1. Como resultado, obteve-se 0,74 de acurácia. Desse modo, o método mostra que há uma relação entre fatores socioeconômicos e o desempenho no Enem.

Introdução

Um dos objetivos da Agenda de 2030 da Organização das Nações Unidas (ONU) é de 'Assegurar a educação inclusiva e equitativa e de qualidade, e promover oportunidades de aprendizagem ao longo da vida para todos' (https://brasil.un.org/pt-br/sdgs/4). Com padrões estabelecidos e expectativas definidas, um dos indicadores utilizados para medir o progresso é obtido por uma prova nacional na maior parte dos países do mundo.

No Brasil, o indicador de como está o desempenho do estudante ao fim da escolaridade básica é o Exame Nacional do Ensino Médio (ENEM). Desde a sua criação em 1998 até hoje, todo ano há um número de participantes que engloba toda a dimensão nacional.

No entanto, essa alcance nacional fez tornar visível como há uma discrepância de desempenho entre regiões segundo o desenvolvimento socioeconômico. Estudos e pesquisas prévios já mostram que o nível de desempenho tende a se relacionar com a desigualdade social.

Desse modo, o presente trabalho propõe-se em identificar esta possível correlação do resultado da prova com as condições socioeconômicas e culturais do estudante. Para isso, foi utilizada a base de dados do INEP, definindo as respostas do questionário socioeconômico como variáveis preditoras. O trabalho propõe estimar o desempenho no ENEM de dado aluno e também identificar a utilidade de cada variável para o valor obtido.

Arcabouço Teórico

A aprendizagem supervisionada é uma área de aprendizado de máquina que aprende com rótulos, ou seja, os dados são oferecidos com respostas corretas. A classificação, que é um dos tipos do problema de aprendizado supervisionado de máquina, aprende a partir dos *features* (ou, características) e rótulos do conjunto de dados de treinamento e cria um modelo com base nele. E quando é fornecido um dado cuja classe é desconhecida, ele prevê o valor do rótulo. Ou seja, após reconhecer o padrão dos dados de cada rótulo, determina a classe a qual pertence a nova entrada.

Os principais algoritmos de classificação de aprendizado supervisionado são

- Naive Bayes: modelo baseado no teorema de Bayes.
- Regressão Logística: modelo baseado na relação linear entre variáveis independentes e dependentes.
- Árvore de decisão: modelo baseado nas regras.
- Support Vector Machine (SVM): modelo que busca uma margem entre cada classe.
- KNN: modelo baseado na distância de proximidade.
- Redes Neurais: modelo baseado em nós interconectados.
- Ensemble: um conjunto de algoritmos distintos ou iguais de aprendizado de máquina.

Dentre eles, a árvore de decisão é um dos algoritmos que pode ser aplicado de forma mais simples e flexível. A influência de pré-processamento como normalização e redução de dimensão é baixa. No entanto, para melhorar o desempenho, o modelo precisa ter regras complexas que podem levar ao overfitting e, consequentemente, não gerar uma boa generalização.

A árvore de decisão cria uma regra de classificação que toma um vetor de valores como entrada e no fim retorna um único valor de decisão (RUSSEL, NORVIG, 2009, p.

698). Para chegar nesse valor é realizado uma série de testes, cada nó na árvore contém um teste e cada ramificação corresponde a um possível valor do atributo, já cada nó folha da árvore equivale a um valor de decisão a ser retornado. A maneira mais simples é expressar essa regra em uma sequências de *if* (se) e *else* (senão).

Assim, a eficiência da classificação depende de quais critérios foram escolhidos a partir dos dados para compor a regra. Para maximizar a eficiência da classificação escolhese um critério que prioriza os atributos de maior utilidade, ou seja, obtém o maior sucesso em dividir os exemplos em classificações exatas.

Um dos critérios que podem ser utilizados é a entropia, que é a medida de incerteza em uma variável aleatória, uma redução na entropia corresponde a aquisição de informação (RUSSEL, NORVIG, 2009, p. 703), utilizando a entropia pode-se calcular o ganho de informação dos atributos, e assim, concluir qual possui maior utilidade.

A entropia pode ser calculada pela seguinte fórmula:

$$H(V) = -\sum_{i=0}^{k} Pr(v_k) \log_2 Pr(v_k)$$

E o ganho de informação pode ser calculado pela seguinte fórmula:

$$Gain(A) = B\left(\frac{p}{p+n}\right) - \sum_{k=1}^{d} \frac{p_k + n_k}{p+n} B\left(\frac{p_k}{p_k + n_k}\right)$$

Outro possível critério, criado por Corrado Gini em 1912, é o índice Gini, que pode ser usado para medir a heterogeneidade de um atributo, se igual a zero significa que todos os registros fazem parte da mesma classe (puro), caso se aproxime de 1 os registros são distribuídos uniformemente entre as classes (impuro), os atributos mais puros possuem a maior utilidade. O índice pode ser calculado por:

$$Gini = 1 - \sum_{i=1}^{n} (P_i)^2$$

Descrição e Modelagem do Problema

Problema

Segundo Caprara (2017), há os efeitos da classe social na conformação dos rendimentos acadêmicos. Assim, o problema a ser tratado neste artigo é buscar as características socioeconômicas que tenha mais relevância para nota do ENEM.

Dados

O conjunto de dados utilizado no presente trabalho são os microdados do ENEM de 2020 disponibilizado pelo INEP, disponível em: https://download.inep.gov.br/microdados/microdados_enem_2020.zip.

Trata-se de dados realísticos que contém informações de cada candidato, com respostas aos questionários que devem ser preenchidos no ato da inscrição. O estudo realizado faz uso das questões socioeconômicas disponíveis no conjunto de dados. Essas características, que incluem desde informações particulares do aluno como faixa etária e sexo até informações familiares como renda mensal e quantidade de eletrodomésticos na residência, são relacionadas e detalhadas a seguir.

- TP_FAIXA_ETARIA: Faixa etária
- TP_SEXO: Sexo
- TP_ESTADO_CIVIL: Estado Civil
- TP_COR_RACA: Cor/raça
- TP_NACIONALIDADE: Nacionalidade
- TP_ST_CONCLUSAO: Situação de conclusão do Ensino Médio
- TP_ANO_CONCLUIU: Ano de Conclusão do Ensino Médio
- TP_ESCOLA: Tipo de escola do Ensino Médio
- TP_ENSINO: Tipo de instituição que concluiu ou concluirá o Ensino Médio
- IN_TREINEIRO: Indica se o inscrito fez a prova com intuito de apenas treinar seus conhecimentos
- TP_DEPENDENCIA_ADM_ESC: Dependência administrativa (Escola)
- TP_LOCALIZACAO_ESC: Localização (Escola)
- TP_SIT_FUNC_ESC: Situação de funcionamento (Escola)
- Q001: Até que série seu pai, ou o homem responsável por você, estudou?
- Q002: Até que série sua mãe, ou a mulher responsável por você, estudou?
- Q003: A partir da apresentação de algumas ocupações divididas em grupos ordenados, indique o grupo que contempla a ocupação mais próxima da ocupação do seu pai ou do homem responsável por você. (Se ele não estiver trabalhando, escolha uma ocupação pensando no último trabalho dele).
- Q004: A partir da apresentação de algumas ocupações divididas em grupos ordenados, indique o grupo que contempla a ocupação mais próxima da ocupação da sua mãe ou da mulher responsável por você. (Se ela não estiver trabalhando, escolha uma ocupação pensando no último trabalho dela).
- Q005: Incluindo você, quantas pessoas moram atualmente em sua residência?
- Q006: Qual é a renda mensal de sua família? (Some a sua renda com a dos seus familiares.)
- Q007: Em sua residência trabalha empregado(a) doméstico(a)?
- Q008: Na sua residência tem banheiro?
- Q009: Na sua residência tem quartos para dormir?
- Q010: Na sua residência tem carro?
- Q011: Na sua residência tem motocicleta?
- Q012: Na sua residência tem geladeira?
- Q013: Na sua residência tem freezer (independente ou segunda porta da geladeira)?

- Q014: Na sua residência tem máquina de lavar roupa? (o tanquinho NÃO deve ser considerado)
- Q015: Na sua residência tem máquina de secar roupa (independente ou em conjunto com a máquina de lavar roupa)?
- Q016: Na sua residência tem forno micro-ondas?
- Q017: Na sua residência tem máquina de lavar louça?
- Q018: Na sua residência tem aspirador de pó?
- Q019: Na sua residência tem televisão em cores?
- Q020: Na sua residência tem aparelho de DVD?
- Q021: Na sua residência tem TV por assinatura?
- Q022: Na sua residência tem telefone celular?
- Q023: Na sua residência tem telefone fixo?
- Q024: Na sua residência tem computador?
- Q025: Na sua residência tem acesso à Internet?

Os dados dos participantes estão anonimizados, sendo diferenciados por um número de inscrição que impossibilita a identificação de um candidato em outro exame ou em microdados de pesquisas diferentes.

Pré-Processamento

Nesta fase, os dados obtidos pelo Inep foram tratados para utilizar no modelo. As variáveis como número de inscrição, ano, localização da realização da prova e outros foram retirados da análise e as variáveis categóricas representadas por caractéres foram substituídas por valores numéricos.

Muitas vezes, é necessária a normalização dos dados para garantir que todos os dados estejam no mesmo intervalo de 0 a 1. No entanto, como a árvore de decisão decide baseado nas informações, não houve a necessidade da normalização.

Além disso, as instâncias que apresentassem atributos nulos foram descartados, uma vez que o volume de dados era grande suficiente, totalizando em 520.728 entradas.

Modelagem

Inicialmente as classes propostas para a predição seriam divididas por faixa de acertos, dado que a nota se trata de uma variável contínua que não depende apenas da quantidade de acertos. No entanto, depois de alguns testes, chegou-se à conclusão de que seriam um excesso de classes, não apresentando uma diferença significativa de desempenho entre as duas classes consecutivas e apenas dificultando a predição. Portanto, foi dividida somente em três classes, numeradas de 1 a 3 — a primeira engloba a quantidade de acertos de 0 a 69, a segunda classe de 70 a 129 e, por último, a terceira classe de 130 a 180.

Os dados dos inscritos que estiveram ausentes em pelo menos uma das provas foram desconsiderados para a pesquisa. Assim, os dados consistem em 325.601 instâncias da classe 1, 187.764 instâncias da classe 2 e 7.363 instâncias da classe 3, estando desbalanceados. Por fim, foram divididos em dados de treino e de teste, com proporção de 7 para 3, utilizando o método *train_test_split* da biblioteca *sklearn* do Python.

Dentre as opções de algoritmos de classificação supervisionada, foi escolhida a árvore de decisão para o problema tratado nesta pesquisa.

Experimentos

Otimização de parâmetros

Para o classificador árvore de decisão, os melhores hiperparâmetros foram buscados por Grid Search, método *GridSearchCV* da *sklearn*. Como resultado obteve-se:

'criterion': 'entropy', 'max_depth': 6, 'min_samples_leaf': 2, 'min_samples_split': 7

Dessa forma, a árvore terá profundidade máxima de 6 nós com o minímo de amostras em um nó folha sendo 2 e o mínimo para dividir um nó folha definido como 7 amostras.

Métricas de Avaliação

A perfomance do modelo de classificação proposto foi avaliada utilizando acurácia, precisão, revocação e a medida F que é a média harmônica entre precisão e revocação.

A matriz de confusão resume o resultado da previsão para dado problema de classificação em termos de *true positive* (TP), *true negative* (TN), *false positive* (FP) e *false negative* (FN), que são os exemplos positivos corretamente previstos pelo classificador, os exemplos negativos corretamente classificados pelo modelo, as instâncias negativas incorretamente classificadas, e os exemplos positivos incorretamente classificados pelo modelo, respectivamente. A tabela 2 mostra as métricas usadas no artigo.

Métrica	Fórmula
Precisão (P)	$\frac{TP}{TP+FP}$
Revocação (R)	$\frac{TP}{TP+FN}$
Acurária	$\frac{TP+TN}{TP+TN+FN+FP}$
F-score	$2 \times \frac{R \times P}{R+P}$

Tabela 1: As métricas utilizadas na classificação.

Resultados

O modelo completo que foi treinado com todas as 38 características descritas na seção anterior resultou em 0,74 de acurácia. A tabela 2 a seguir apresenta os scores das características do modelo completo.

Todas as outras características que foram citadas na seção de dados, porém que não constam na tabela foram omitidas por apresentar score 0.

É perceptível que a renda mensal familiar, a dependência administrativa da escola (federal, estadual, municipal e privada) e ter ou não computador na residência foram as três características mais decisórias para a classificação. A questão de raça (etnia) e sexo (gênero), dois fatores que se desconsiderassem o contexto social não teriam uma relação com rendimento escolar, mas também foram considerados pelo modelo.

Característica	Importância
Q006	0,54049
TP_DEPENDENCIA_ADM_ESC	0,17779
Q024	0,13687
TP_ESCOLA	0,08530
TP_SEXO	0,02742
TP_FAIXA_ETARIA	0,02249
Q002	0,00420
TP_COR_RACA	0,00410
Q001	0,00114
Q025	0,00007
Q005	0,00006
TP_NACIONALIDADE	0,00006

Tabela 2: Scores das características do modelo completo.

Levando em consideração apenas as 25 questões socioeconômicas autodeclaradas, que não incluem dados escolares, obteve-se a acurácia de 0,71. Os detalhes do modelo são apresentados na tabela 3 a seguir.

Característica	Importância
Q006	0,74590
Q024	0,19589
Q002	0,02303
Q003	0,00854
Q005	0,00827
Q007	0,00542
Q001	0,00536
Q013	0,00525
Q025	0,00106
Q022	0,00056
Q011	0,00055
Q004	0,00017

Tabela 3: Scores das características do modelo socioeconômico.

As *features* que não estão listadas acima obtiveram score 0.

Similar ao modelo completo, a renda mensal e possuir ou não computador foram as duas características determinantes na classificação. A seguir, o grau da escolaridade da mãe apareceu como terceira característica mais importante, enquanto que o grau da escolaridade do pai não foi considerado. Ao contrário, em relação à ocupação, a do pai teve uma importância mais alta do que a da mãe.

Um modelo mais simples e leve pode ser obtido utilizando quantidades menores de características. Há algumas características que podem ser removidas sem prejudicar o resultado em relação ao do modelo completo. Para isso, selecionou-se as características com maiores importâncias nos experimentos anteriores. A tabela 4 a seguir apresenta os resultados.

Características Utilizadas	Acurácia	Precisão
TP_DEPENDENCIA_ADM_ESC, Q006	**0,72942**	**0,47**
TP_DEPENDENCIA_ADM_ESC	0,71584	0,46
TP_ESCOLA, TP_SIT_FUNC_ESC	0,71918	0,45
Q006, Q024	0,71294	0,46
Q024	0,69117	0,45
Q006	0,70449	0,45
TP_ESCOLA	0,70020	0,44
Q007, Q008, Q009, Q010, Q011, Q012, Q013, Q014, Q015, Q016, Q017, Q018, Q019, Q020, Q021, Q022, Q023	0,69262	0,44

Tabela 4: Acurácia e precisão dos modelos reduzidos. O modelo com melhor desempenho está destacado.

Conclusão

No artigo apresentado, foi proposta uma metodologia para prever o desempenho no Exame Nacional do Ensino Médio (ENEM) realizado no Brasil. O objetivo desta abordagem foi obter a melhor característica de previsão para que possa ser buscadas as formas de melhorar tais condições para os estudantes. Dessa forma, foi trabalhado com os dados da amostra no modelo de aprendizagem supervionada, que oferece visões sobre o problema.

Como pode-se ver nos resultados apresentados pelo algoritmo, a acurácia, ou seja, a probabilidade do valor obtido ser igual o valor esperado, é de aproximadamente 74%. Dessa forma, é possível estimar, com certa precisão, o desempenho de dado aluno com base em seus fatores socioeconômicos.

Portanto, pode-se afirmar que há uma relação nítida entre a renda mensal familiar e o desempenho obtido no ENEM. Outros fatores que podem ser citados estão correlacionados com a renda, como o acesso a computadores pessoais, qual a dependência administrativa da escola e qual o tipo de escola frequentada no ensino médio.

Ademais, para futuros trabalhos, o estudo pode ser aprimorado com o uso de métodos de seleção híbrida de características para que assim cada característica se torne mais otimizada e significante de acordo com a previsão de desempenho do estudante. O algoritmo de comitê de classificação

(*ensemble-based algoritm*), principalmente a técnica de *extreme gradient boosting*, pode ser utilizada nesse sentido.

Referências

CAPRARA, B. *Classes sociais e desempenho educacional no Brasil*. 2017. Tese (Doutorado) - Instituto de Filosofia e Ciências Humanas, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2017.

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

INSTITUTO DE PESQUISA ECONÔMICA APLICADA *Objetivos de Desenvolvimento Sustentável: 4. Educação de Qualidade*. Disponível em: https://www.ipea.gov.br/ods/ods4.html. Acesso em: 19 jul. 2022.

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

RUSSELL, Stuart Jonathan; NORVIG, Peter. *Artificial Intelligence: A Modern Approach*. 3. ed. [S. l.]: Prentice Hall, 2009.