Relatório Final - Treinamento *Big Data Science*

Carlos Roberto de Souza Camilo

Análise da Base de Dados do ENEM 2021:
Predição da nota final do aluno a partir das informações preenchidas no questionário de inscrição

São José do Rio Preto - SP
Dez. 2022

SUMÁRIO

1. INTRODUÇÃO

O Exame Nacional do Ensino Médio (Enem) consiste num conjunto de provas para admissão ao Ensino Superior em instituições de ensino públicas e privadas. O Enem é organizado anualmente pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), vinculado ao Ministério da Educação (MEC). Em sua forma atual, o Enem consiste de 180 (cento e oitenta) questões de múltipla escolha e uma proposta de redação (INEP, 2022). As questões são agrupadas em quatro provas objetivas que abordam os Componentes Curriculares do Ensino Médio:

- Linguagens, Códigos e suas tecnologias (LC);
- Ciências Humanas e suas tecnologias (CH);
- Ciências da Natureza e suas tecnologias (CN);
- Matemática e suas tecnologias (MT).

Considerando fins de transparência da aplicação para com os candidatos inscritos e com a sociedade em geral, o INEP divulga anualmente a base de dados "Microdados do Enem", https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem. Estes dados trazem as informações preenchidas por cada candidato ou candidata durante o processo de inscrição, respeitando-se a Lei Geral de Proteção de Dados (LGPD). São disponibilizadas portanto informações sobre cada candidato inscrito, mas eliminados dados que poderiam possibilitar a eventual identificação indevida do mesmo.

Pelos dados apresentados para o Enem 2021, as provas foram aplicadas nos dias 21 (LC, CH e Redação) e 28 (CN e MT) de novembro de 2021, em formato impresso ou digital. O total de inscritos foi de 3.389.832 candidatos, com as provas sendo aplicadas em 1.712 municípios dos 26 Estados e do Distrito Federal. A nota final de cada candidato é calculada como a média da nota obtida nas quatro provas e redação, podendo assumir valores entre 0 e 1.000 pontos cada (a Tabela 1 apresenta a maior nota obtida para cada prova no Enem 2021).

Tabela 1 - Maiores notas obtidas no Enem 2021 (não necessariamente pelo mesmo candidato).

Prova	Maior nota no Enem 2021
CN	867,1
CH	846,9
LC	826,1
MT	953,1
Redação	1000,0
Nota final	862,7

Ao efetuar a inscrição para o Enem, o candidato ou candidata deve preencher uma série de informações pessoais, além de um *Questionário Socioeconômico* de 25 questões. Estes dados permitem uma análise ampla de quem prestou o exame, e, aliadas aos resultados obtidos (suas notas), possibilitam investigar se e quais condições (renda, idade, ter acesso à internet, etc.) podem ser preponderantes para um melhor resultado nas provas. Neste estudo, buscou-se construir um modelo através de algoritmos de aprendizado de máquinas (*machine learning*) que, a partir somente dos dados preenchidos pelo candidato, pudesse predizer a nota final que ele irá obter. A situação problema discutida neste trabalho portanto pode ser resumida em termos da seguinte questão: é possível estimar a nota que um candidato irá obter, antes que ele faça qualquer prova, somente a partir de suas informações pessoais, incluindo dados socioeconômicos?

Considerando as enormes desigualdades e barreiras sociais estabelecidas, a análise aqui desenvolvida pode fornecer resultados para a avaliação da educação no Brasil, justamente um dos objetivos propostos para o Enem (INEP, 2022). As seções a seguir descrevem a análise e tratamento dos dados do Enem 2021, a implementação do modelo de *machine learning* e os resultados obtidos.

2. METODOLOGIA

Algoritmos de *machine learning* são métodos numéricos para modelagem de problemas a partir de um tratamento estatístico de um conjunto de dados. Uma classe destes métodos são os algoritmos de predição, que uma vez modelados (treinados) sobre um grande conjunto de dados, são capazes de inferir a classificação ou o valor que um novo objeto/registro, desconhecido do modelo até então, irá assumir. Neste estudo, foram utilizados algoritmos de *machine learning* para analisar os dados do Enem 2021 e criar um modelo de regressão numérica que estime a nota final que um candidato irá obter a partir somente das informações pessoais que ele respondeu no processo de inscrição.

A primeira etapa para o desenvolvimento do modelo é a aquisição dos dados, isto é, obter as informações sobre o problema a ser tratado, sendo que quanto maior for a quantidade de dados com que o modelo for treinado, melhor será sua capacidade em predizer resultados para novos registros (melhor a qualidade do modelo). Os dados utilizados neste estudo estão disponíveis gratuita e publicamente no portal Microdados do Enem. A etapa seguinte é a análise e tratamento dos dados, visando tratar registros com dados ausentes ou discrepantes (erros de formação, falhas ao salvar o dado, etc.); este processo é apresentado de forma mais detalhada no *notebook* em python '1.Analise_tratamento_testes_dos_dados.ipynb'. Conforme comentado neste, dados com quantidade significativa de valores ausentes não contribuem para análise, sendo necessário seu pré-processamento. Assim, por exemplo, embora a informação do tipo de escola (pública ou privada) em que o aluno estuda(ou) seja muito relevante para o tratamento do problema proposto neste trabalho, como 66% dos registros não contém tal informação, não foi possível utilizá-la; a mesma consideração vale para a informação da localização da escola (se em região urbana ou rural), ausente em 76% dos registros. Outros dados, como a nacionalidade, foram desconsiderados por não apresentarem variabilidade significativa para contribuírem com informações ao modelo (mais de 98% dos candidatos nasceram no Brasil); a mesma consideração vale para o tipo de ensino (regular ou EJA).

Após as análises e tratamento dos dados, foram identificados 8 (oito) campos com informação significativa (preenchimento e qualidade dos dados), sendo estes: faixa etária, sexo, estado civil, cor/raça, se já havia concluído ou não o Ens. Médio, se o candidato era treineiro, UF do local de prova (a maioria não preencheu UF do município onde mora), e

idioma escolhido para prova de língua estrangeira (inglês ou espanhol). Além destas, foram utilizadas as 25 questões do *Questionário Socioeconômico*, pois praticamente a totalidade de alunos as havia respondido, sem ocorrência de valores nulos (“não informado”, “não sabe”, etc.). Este conjunto de 33 características consistiu na base para o treinamento do modelo.

Outro procedimento importante antes da implementação dos algoritmos de *machine learning* é o tratamento de dados alfanuméricos. A resposta ao *Questionário Socioeconômico* consistiu de classes categorizadas em letras (A, B, C, etc.), sendo adequado sua conversão para valores numéricos (1, 2, 3, etc.), pois alguns algoritmos apresentam melhor desempenho ao trabalhar somente com dados numéricos. As informações detalhadas sobre os campos de dados e suas respostas categorizadas são fornecidas juntos com a base de dados original no arquivo ‘Dicionário_Microdados_Enem_2021.xlsx’, sendo utilizada a conversão {‘A’: 1, ‘B’: 2, ‘C’: 3, … }, conforme descrito no *notebook*.

Para compor a nota final de cada candidato, a variável a ser predita pelo modelo, utilizou-se a média das notas individuais obtidas nas quatro provas e na redação. Após testes, verificou-se a necessidade de excluir da análise os 34% de registros de candidatos que não compareceram em um ou nos dois dias de prova; as provas que estes não fizeram contavam como nota zero, afetando sua nota final (também zero, ou considerando apenas parte das provas, se o candidato compareceu em apenas um dia do exame). Os resultados para estes candidatos prejudicavam significativamente o modelo, não sendo de qualquer forma realista prever a nota de um aluno que não prestou todas as provas. Assim, o modelo foi aplicado aos 2.238.106 de candidatos que compareceram aos dois dias de prova.

Após a análise e tratamento dos dados, foi gerado uma versão resumida da base de dados, contendo apenas os registros e colunas a serem de fato utilizadas no modelo, ‘Dados_ENEM2021_filtrados.csv’, disponível para download neste link.

3. RESULTADOS

A base de dados tratada foi aplicada aos algoritmos de *machine learning Decision Tree* e *K-Nearest Neighbors*, com parâmetros definidos a partir de testes apresentados no *notebook* '1.Analise_tratamento_testes_dos_dados.ipynb'. Na implementação, 80% dos registros foram empregados para o treinamento do modelo (treino), sendo este então aplicado aos 20% de registros restantes (teste da predição). A nota final predita pelo modelo pode então ser comparada com a nota real obtida pelo candidato, podendo apresentar um erro para mais ou para menos, conforme exemplificado na Tabela 2. Para detalhes sobre a implementação dos modelos ver o *notebook* em python '2.Implementacao_algoritmos_ML.ipynb'.

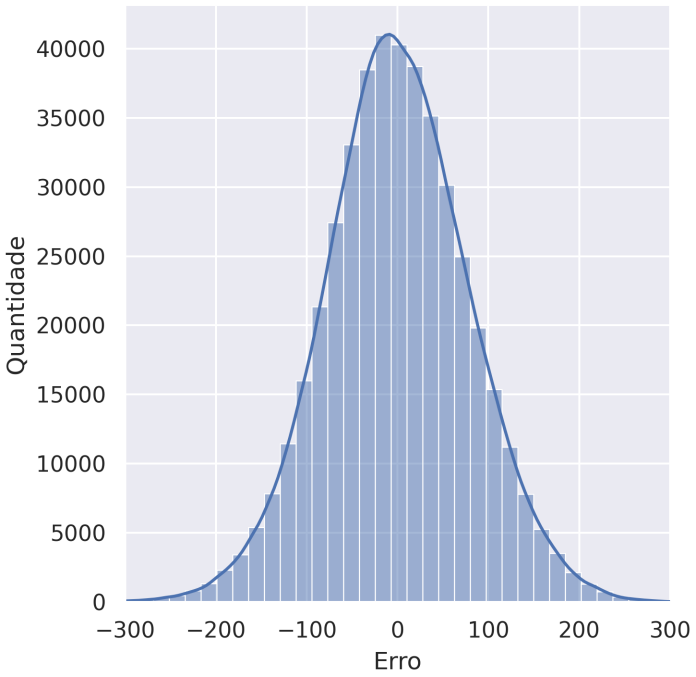
Tabela 2 - Comparação entre a nota final obtida pelo candidato e a nota predita pelo modelo de *machine learning* (Decision Tree) a partir das informações pessoais respondidas na inscrição.

	NOTA_FINAL	Nota predita	Erro
827649	616.06	640.405000	-24.345000
1851255	509.04	568.041818	-59.001818
840175	560.98	612.829000	-51.849000
1966909	484.66	480.198000	4.462000
2017334	483.42	585.967778	-102.547778
1182317	651.54	607.653750	43.886250
1254188	651.60	539.566316	112.033684
1152341	448.42	574.205000	-125.785000
1376302	439.14	507.121538	-67.981538
1209967	430.50	529.267273	-98.767273
1761672	603.66	609.841333	-6.181333
334011	742.70	613.816000	128.884000
1478970	507.58	527.131000	-19.551000
738989	645.26	632.192000	13.068000
1230675	510.04	526.268333	-16.228333

A partir de um histograma dos erros calculados, ver Figura 1, é possível avaliar que erros menores são os mais frequentes, pois a distribuição está centralizada no zero. Observa-se também que os erros para mais ou para menos são equivalentes, uma vez que a

distribuição é essencialmente simétrica (se ocorressem mais erros positivos do que negativos, ou vice-versa, seria um indicativo de viés no modelo, o que não ocorre).

Figura 1 - Histograma do erro verificado no resultados preditos utilizando *Decision Tree*.

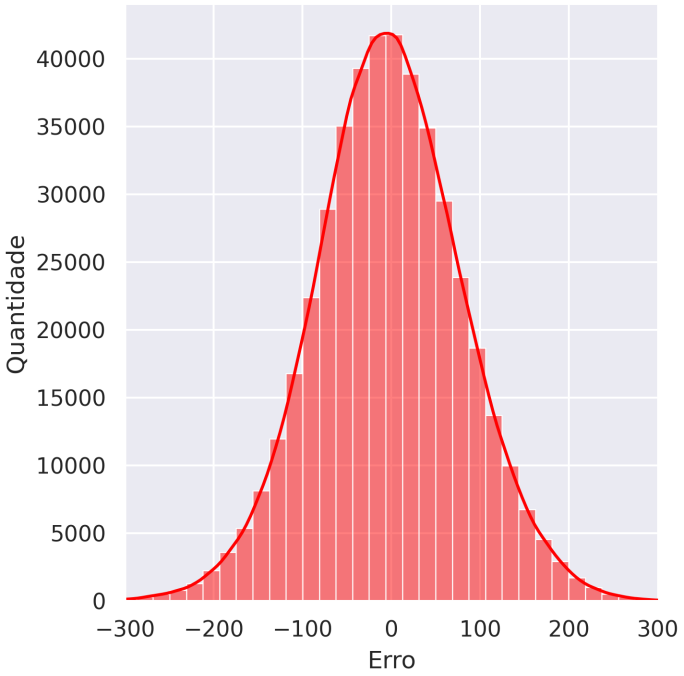


A principal métrica utilizada para avaliar modelos de *machine learning* de regressão é o erro absoluto médio, calculado como a média do módulo dos erros individuais. Para o algoritmo *Decision Tree* (Árvore de Decisão), o erro absoluto médio foi de 62,2 pontos, valor considerado satisfatório considerando que o espectro de notas possíveis varia entre 0 e 1.000 pontos, e a simplicidade do modelo aqui utilizado.

Uma característica importante a ser analisada em modelos de machine learning é se não está havendo overfitting, isto é, se o modelo não é especializado demais aos dados com que treinou, mas incapaz de lidar com novos dados. Uma forma de avaliar se está havendo overfitting é aplicando a validação cruzada, aplicando o modelo várias vezes ao conjunto de dados, mas trocando os dados de teste (20%) a cada nova rodada. O resultado obtido para uma validação cruzada com cinco réplicas no modelo deste estudo foi um *score* médio de -0,06, indicando que não haver overfitting significativo no modelo (quanto mais próximo de zero o *score*, melhor)

Para avaliação da qualidade do modelo, a base de dados foi aplicada ao algoritmo *K-Nearest Neighbors* (K-Vizinhos mais próximos), ou KNN, sendo o histograma dos erros calculados apresentado na Figura 2.

Figura 2 - Histograma do erro verificado no resultados preditos utilizando *KNN*.



O histograma indica que o modelo treinado com o algoritmo KNN também é capaz de predizer a nota final do aluno, apresentando tanto erros para mais (positivos) quanto para menos (negativos), quando feita a diferença entre o valor predito pelo modelo e a nota que o candidato efetivamente obteve (o erro é calculado somente sobre os registros separados para o conjunto de treino).

O valor obtido para o erro absoluto médio foi de 65,2 pontos, indicando que o KNN pode oferecer resultados equivalentes aos obtidos no primeiro modelo. Apesar dos dois modelos oferecerem resultados próximos, a implementação utilizando *Decision Tree* foi desenvolvida de forma significativamente mais rápida (melhor desempenho computacional), sendo portanto este algoritmo considerado o mais adequado para a solução do problema aqui proposto.

4. CONSIDERAÇÕES FINAIS

Neste trabalho, implementou-se modelos de *machine learning* para, a partir somente de respostas à questões pessoais realizadas no processo de inscrição, antes de fazer qualquer prova, estimar a nota final que um candidato viria a obter no exame. O erro médio dos resultados obtidos em dois algoritmos diferentes de *machine learning* foi de 62 e 65 pontos. Considerando que a nota final no exame é um valor entre 0 e 1.000, pode-se avaliar que os resultados obtidos são satisfatórios.

A partir de informações pessoais e de respostas às questões socioeconômicas é portanto possível estimar a nota final que um candidato irá obter na prova do Enem 2021. Este resultado comprova que informações como faixa etária, sexo, cor/raça, quando aliadas à questões socioeconômicas, podem sim definir os resultados que uma pessoa irá obter nas provas. Obviamente o modelo apresenta erros (há candidatos que conseguem superar as barreiras impostas por suas condições sociais), mas os resultados neste simples estudo evidenciam quão significativas questões socioeconômicas podem ser para pessoas buscando acesso à universidade a fim obter uma melhor formação.

Possíveis passos para o aprofundamento do estudo seriam analisar quais questões ou informações são mais significativas para o resultado (a nota final do candidato) – o algoritmo *Decision Tree* possivelmente permite este tipo de análise – e aplicar o modelo à base de dados de anos anteriores do Enem para avaliar se o modelo continua válido, se há flutuação ao longo dos anos (em 2020, o modelo seria sensível ao agravamento das desigualdades sociais devido às restrições impostas pela pandemia de COVID-19?)

REFERÊNCIAS

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. **Microdados do Enem 2021**. Brasília: Inep, 2022. Disponível em: < https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem >. Acesso em: 11 dez. 2022.