Projeto Final Big Data Science

Análise dos dados coletados no ENEM 2021

Autor: Leonardo Shimizu Yojo

Sumário

Este estudo tem como objetivo avaliar os microdados sobre o ENEM 2021 divulgados pelo INEP e está divido da seguinte maneira: primeiro é apresentado uma introdução sobre o tema; em seguida os dados em si são explorados para melhor compreender o tema e tirar algumas conclusões; por fim duas propostas de emprego de algoritmos de aprendizado de máquina são propostas. Na primeira proposta, algoritmos de clusterização foram utilizados para avaliar se é possível distinguir diferentes grupos de participantes no ENEM 2021 e suas características. Na segunda proposta, algoritmos de classificação foram utilizados para se averiguar a possibilidade de se determinar se uma pessoa frequentou ou não escola pública no ensino médio, com a ideia de criar um sistema antifraude para cotas no ingresso ao ensino superior.

Introdução

O ENEM (Exame Nacional do Ensino Médio) é uma avaliação em nível nacional dos alunos concluintes do ensino médio no Brasil. Segundo o INEP (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira), órgão federal responsável pela prova, dentre os principais objetivos podemos citar: avaliação do aprendizado dos alunos; aperfeiçoamento do currículo escolar do ensino médio; utilização do exame como forma de ingresso no ensino superior e programas de financiamento estudantil; e desenvolvimento de estudos sobre a educação brasileira (1). Portanto essa ferramenta vem se tornando cada vez mais importante tanto para a avaliação individual dos concorrentes, como para avaliar a educação no país.

Desde sua primeira edição em 1998, o formato da prova se adequou ao longo do tempo. Atualmente o exame é composto por 180 questões objetivas de múltipla escolha divididas em 4 grandes áreas: 1) Ciências da Natureza e suas Tecnologias; 2) Ciências Humanas e suas Tecnologias; 3) Linguagens, Códigos e suas Tecnologias e; 4) Matemática e suas Tecnologias. Além disso também há uma prova de redação.

Em 2021, as provas foram aplicadas nos dias 21 e 28 de novembro. Além das provas objetivas e a redação, os alunos inscritos no exame responderam um questionário contendo 25 perguntas sobre seu nível socioeconômico, família, educação e trabalho.

Análise dos dados

O INEP disponibiliza os dados recolhidos por pesquisas, avaliações e exames realizados. O arquivo contendo todos os dados da edição do ano de 2021 do ENEM possui dados de todos os participantes anonimizados, para que não seja possível reconhecer o indivíduo a partir dessas informações. Portanto houve algumas modificações nos dados divulgados, em comparação com edições passadas, tais como: exclusão do campo escola do participante, exclusão do campo referente aos pedidos de atendimento especializado e substituição do campo idade pelo campo faixa etária. Os dados divulgados sobre o ENEM 2021 possibilita construir uma grande variedade de análises, algumas das quais foram abordadas neste estudo.

Uma observação relevante é sobre o método utilizado pelo INEP para calcular as notas dos participantes, a chamada Teoria de Resposta ao Item (TRI) (2). Nesse método, cada questão respondida é avaliada seguindo 3 parâmetros: parâmetro de discriminação, parâmetro de dificuldade e parâmetro de acerto casual. Portanto o TRI considera a particularidade de cada questão e as notas finais não dependem do total de acertos de cada participante, ou seja, duas pessoas que acertaram a mesma quantidade de questões podem ter notas finais distintas.

O arquivo contendo os dados do ENEM 2021 possui 76 campos com dados referentes ao participante (faixa etária, sexo, estado civil, etc.), dados referentes à escola que o participante frequentou, dados referentes ao local de realização da prova, dados das provas objetivas e redação, além do questionário socioeconômico. Nem todos as entradas possuem todos os dados dos 76 campos. O arquivo possui 3389832 entradas e a exploração inicial dos dados mostraram algumas curiosidades:

- 1299306 inscritos do sexo masculino;
- 2090526 inscritos do sexo feminino;
- 84582 inscritos tiveram redação anulada;
- 393 inscritos tiveram prova de Ciências da Natureza zerado;
- 4632 inscritos tiveram prova de Ciências Humanas zerado;
- 2248 inscritos tiveram prova de Linguagens e Códigos zerado;
- 453 inscritos tiveram prova de Matemática zerado;
- 192244 inscritos informaram vir de escola particular;
- 958611 inscritos informaram vir de escola pública.

A Figura 1 mostra a distribuição da quantidade de inscritos no exame nas 5 regiões do país. O maior número de participantes veio do Nordeste, seguido por Sudeste, Norte, Sul e Centro-Oeste. Esse dado é curioso pois não segue a proporção da população brasileira nas 5 regiões (a ordem da região mais populosa para a menos é Sudeste, Nordeste, Sul, Norte e Centro-Oeste (3)). O número de inscritos é menor em comparação com anos anteriores, possivelmente devido a problemas socioeconômicos, desinteresse dos jovens com o exame (4) e a situação de pandemia de COVID 19 (5).

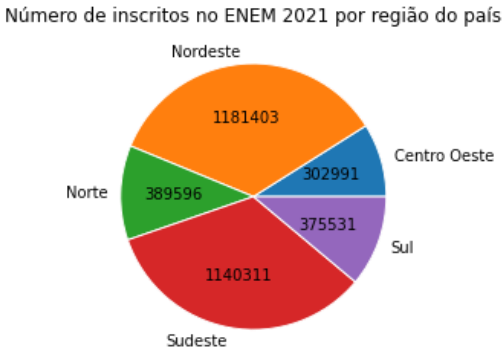

Figura 1 – Distribuição dos inscritos no ENEM 2021.

A seguir é possível observar a distribuição dos inscritos por faixa etária (Figura 2). A maioria é composta por jovens de 17 e 18 anos, como é de se esperar pois é a idade em que se finaliza o ensino médio, porém também há uma parcela considerável de participantes acima dos 25 anos de idade.

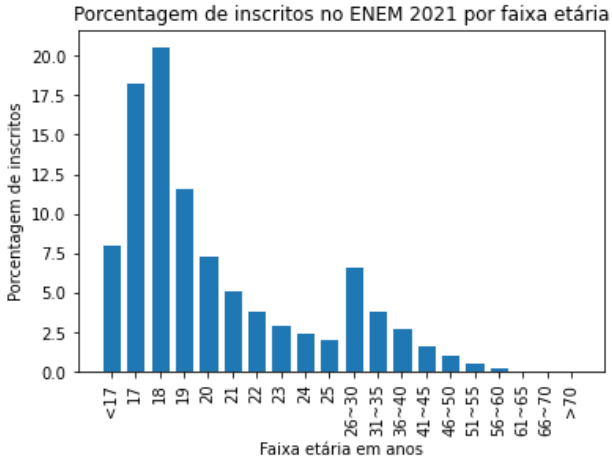

Figura 2 – Porcentagem de inscritos no ENEM 2021 por faixa etária.

Em relação à presença nas provas objetivas, é possível observar na Figura 3 que cerca de 30% dos inscritos não compareceram nas provas e uma pequena parcela foi eliminada de cada prova. As provas de Ciências da Natureza e Matemática foram realizadas no mesmo dia, assim como as provas de Ciências Humanas e Linguagens e Códigos.

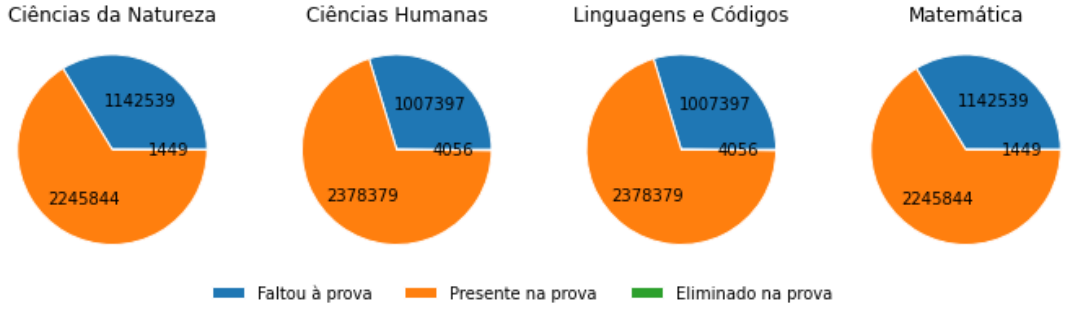

Figura 3 – Presença em cada prova objetiva.

Com relação à prova de redação, nem todas as entradas possuíam dado na categoria de situação da redação. Do total dos inscritos com esse campo preenchido, a grande maioria não teve problemas na redação.

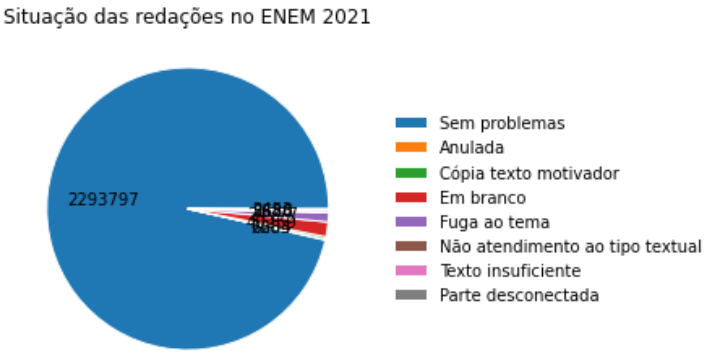

Figura 4 – Situação da prova de redação.

A Figura 5 mostra as médias das notas obtidas pelos candidatos divididas por região, nas 5 provas (4 provas objetivas e redação). Os gráficos possuem mesma escala no eixo vertical. É possível observar que a média é ligeiramente mais alta para os inscritos da região Sudeste,

enquanto a menor média em cada prova foi dos inscritos da região Norte. Esse resultado pode ser um indício da desigualdade na educação brasileira, já relatada em estudos passados (6).

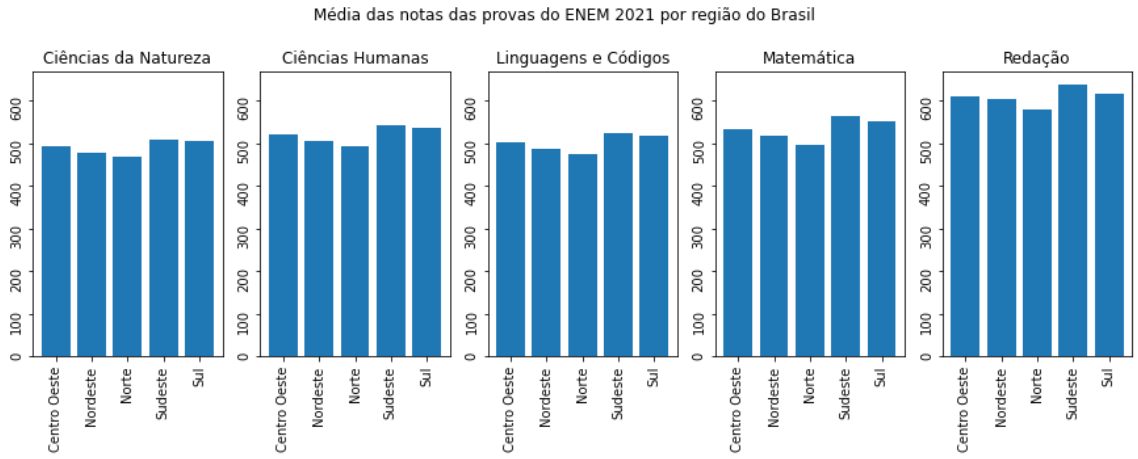

Figura 5 – Média das notas das provas por região.

A Figura 6 complementa a figura anterior ao mostrar a distribuição das notas nas provas objetivas para todo o Brasil. A maior média nas provas objetivas foi obtida para a prova de Matemática, enquanto a menor foi para a prova de Ciências da Natureza. Esse resultado pode ser um indício da deficiência do ensino nas matérias que envolvem este tema (química, física e biologia). Porém também é possível notar uma distribuição mais desigual para essas duas grandes áreas. A prova de redação possui maior desvio padrão.

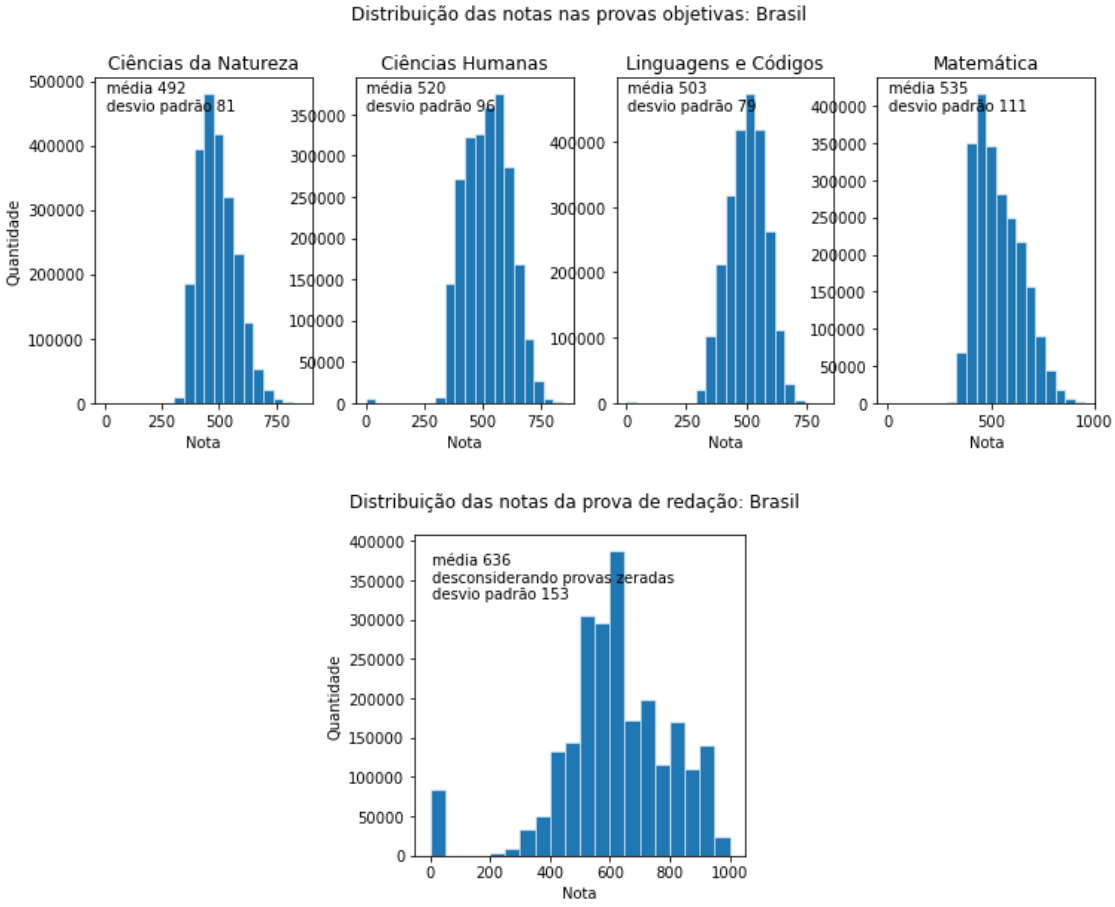

Figura 6 – Distribuição das notas nas provas objetivas e redação.

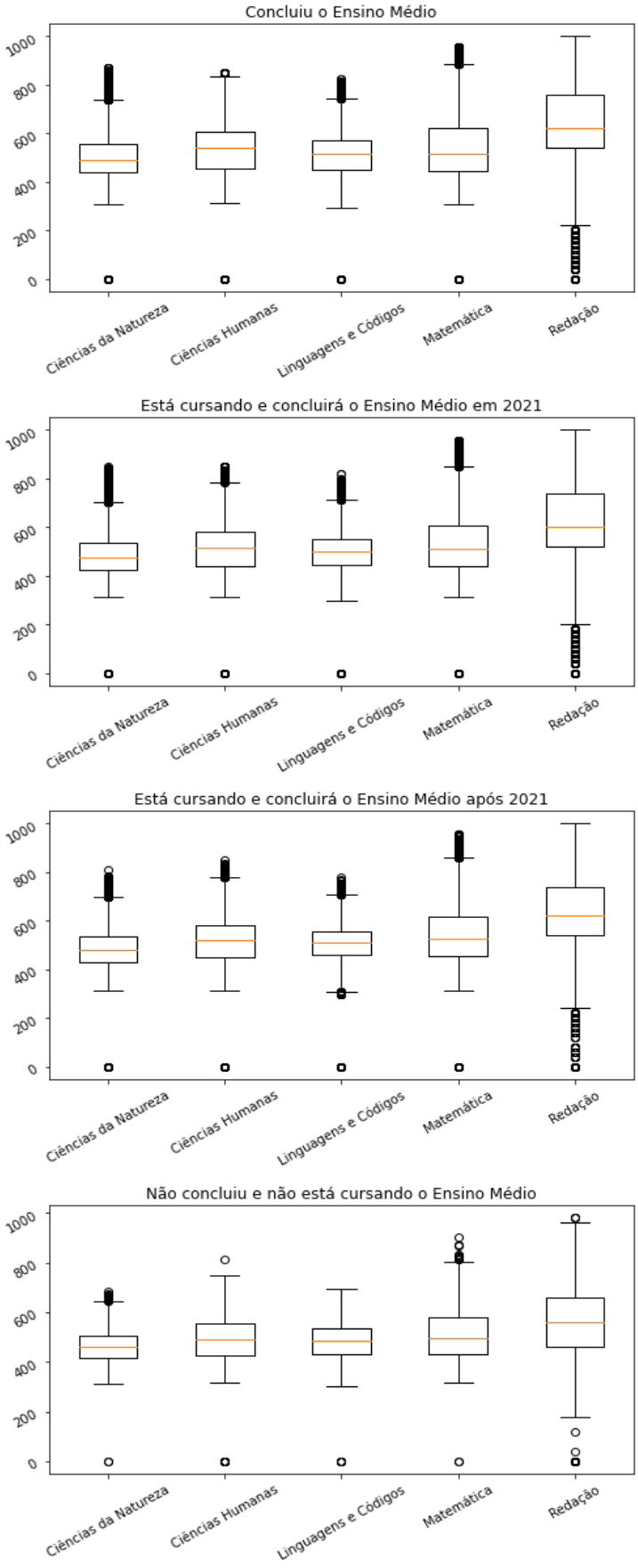

Figura 7 – Distribuição das notas nas provas objetivas e redação em relação a situação de conclusão do ensino médio.

A distribuição das notas nas provas objetivas e redação em relação à situação de conclusão do ensino médio pode ser observada na Figura 7. Tanto as médias quanto as notas máximas são ligeiramente maiores para os inscritos que concluíram o ensino médio, e as menores médias e notas máximas são dos inscritos que não concluíram e nem estão cursando o ensino médio. Entretanto não é possível notar grande diferença nas distribuições das notas entre os grupos de inscritos.

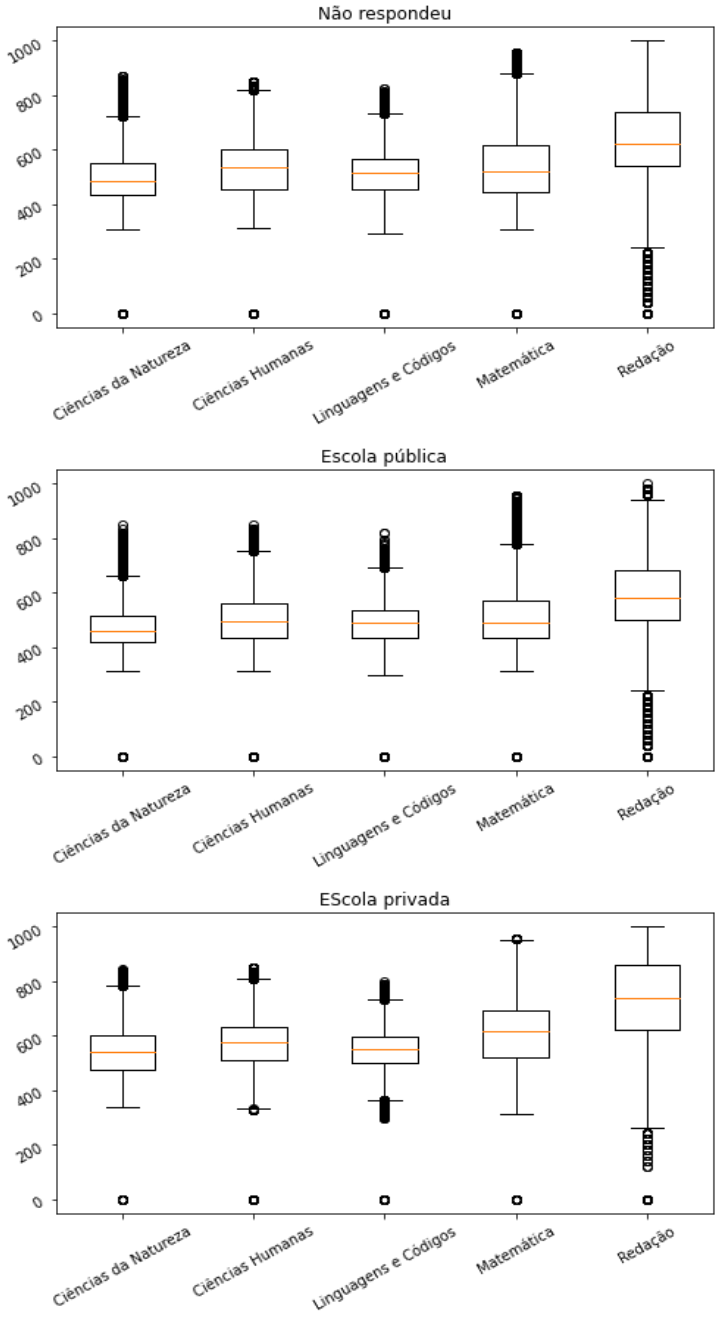

Figura 8 - Distribuição das notas nas provas objetivas e redação por tipo de escola no ensino médio.

Em relação ao tipo de escola frequentada pelo inscrito no ensino médio (escola pública, escola privada ou não informada), as distribuições das notas podem ser observadas na Figura 8. Neste caso é possível notar as maiores notas para os inscritos vindo de escolas privadas em comparação com os inscritos vindo de escolas públicas.

Dentre as questões socioeconômicas, também foram avaliadas as distribuições das notas em função de algumas das perguntas. Em destaque, o resultado para a pergunta Q006 é mostrada na Figura 9. A pergunta no questionário era a seguinte: “Qual é a renda mensal de sua família? (Some a sua renda com a dos seus familiares.)”. As respostas eram dadas por categorias, conforme mostrado na Tabela 1. É possível notar um aumento das notas e das médias em função da renda mensal da família para todas as provas. O grande número de *outliers* também mostra que há inscritos que fogem a essa regra. Senkevics discute amplamente essas diferenças entre alunos advindos de escolas privadas e escolas públicas no acesso ao ensino superior, assim como a relação entre a renda e as notas dos participantes no ENEM (7).

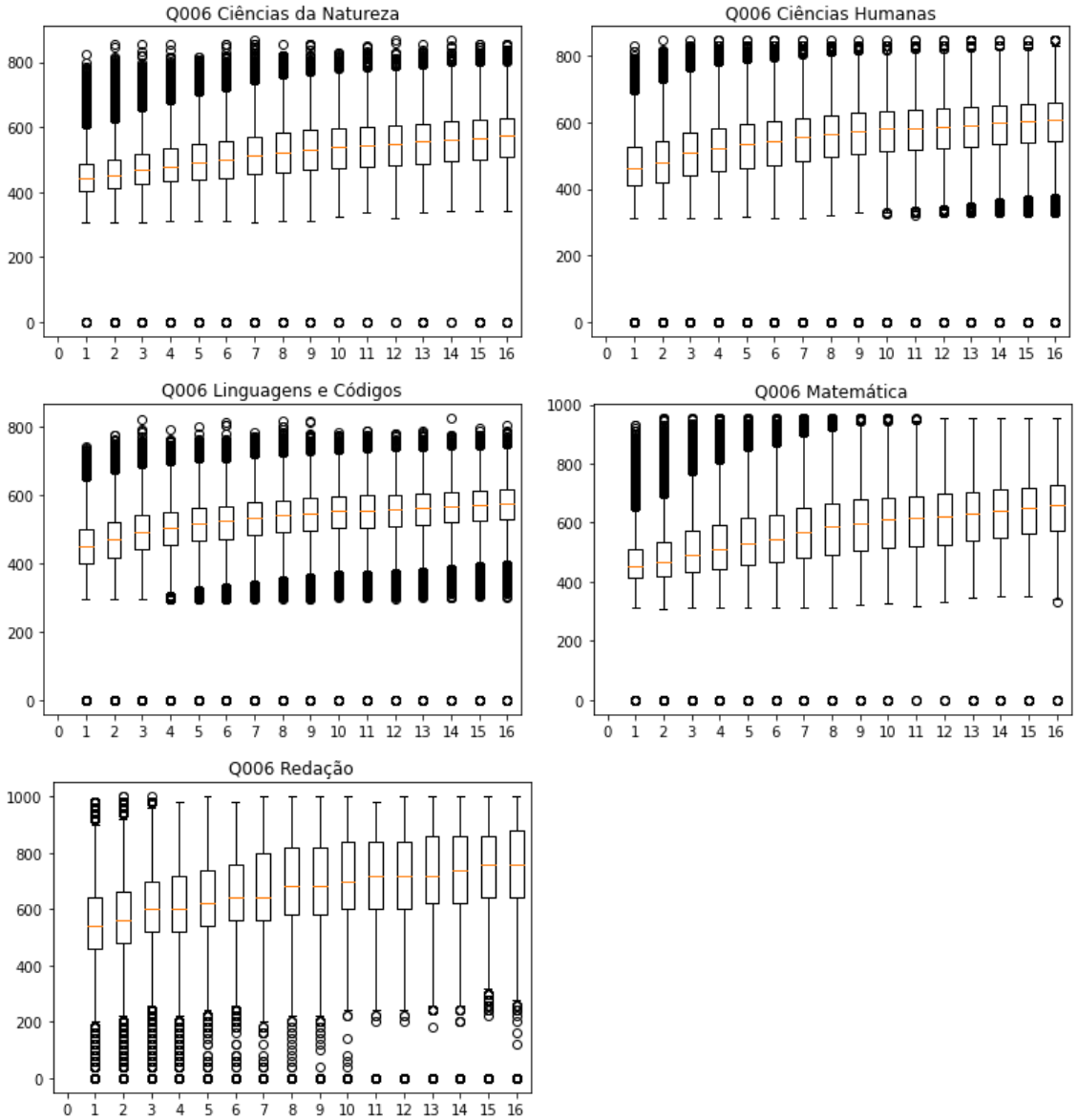

Figura 9 – Distribuição das notas nas provas objetivas e redação por renda mensal da família.

Tabela 1 – Categorias da pergunta Q006.

Categoria	Renda mensal da família
0	Nenhuma Renda
1	Até R$ 1.100,00
2	De R$ 1.100,01 até R$ 1.650,00.
3	De R$ 1.650,01 até R$ 2.200,00.
4	De R$ 2.200,01 até R$ 2.750,00.
5	De R$ 2.750,01 até R$ 3.300,00.
6	De R$ 3.300,01 até R$ 4.400,00.
7	De R$ 4.400,01 até R$ 5.500,00.
8	De R$ 5.500,01 até R$ 6.600,00.
9	De R$ 6.600,01 até R$ 7.700,00.
10	De R$ 7.700,01 até R$ 8.800,00.
11	De R$ 8.800,01 até R$ 9.900,00.
12	De R$ 9.900,01 até R$ 11.000,00.
13	De R$ 11.000,01 até R$ 13.200,00.
14	De R$ 13.200,01 até R$ 16.500,00.
15	De R$ 16.500,01 até R$ 22.000,00.
16	Acima de R$ 22.000,00.

Por fim, a Figura 10 mostra a relação entre a média das notas obtidas pelos participantes em cada prova em função da faixa de renda da família. Para todas as provas (objetivas e redação) há um crescimento da nota média para os inscritos com maior renda familiar, evidenciando a desigualdade do ensino no Brasil.

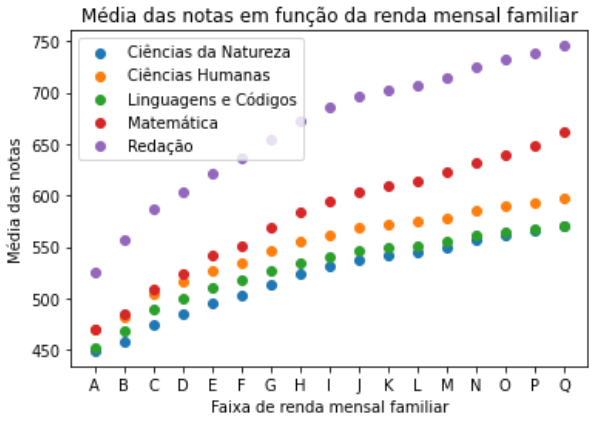

Figura 10 – Média das notas em função da renda mensal da família.

Emprego de algoritmos de aprendizado de máquina

Duas propostas foram avaliadas. Num primeiro cenário, algoritmos de clusterização foram utilizados para avaliar a possibilidade de se distinguir diferentes grupos de participantes no ENEM 2021 e suas características. Na segunda proposta, algoritmos de classificação foram utilizados para se averiguar a possibilidade de se determinar se uma pessoa frequentou ou não escola pública no ensino médio, com a ideia de criar um sistema antifraude para cotas no ingresso ao ensino superior.

Algoritmos de Clusterização:

A ideia por trás desse estudo é verificar se é possível separar o conjunto de dados em grupos com características semelhantes. Observando as principais características de cada grupo seria possível comparar as notas de cada grupo e assim determinar os possíveis aspectos que devem ser melhorados para que o grupo com as piores notas melhore sua performance. Porém, deve-se tomar cuidado com os resultados pois não é possível estabelecer uma dependência direta dos parâmetros analisados com o desempenho no ENEM sem um estudo mais aprofundado.

Os atributos escolhidos foram a faixa etária de cada inscrito, a situação de conclusão do ensino médio, o ano em que concluiu ou pretende concluir o ensino médio, o tipo de escola (privado ou pública) e as respostas das questões 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?”), 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.)”) e Q025 (“Na sua residência tem acesso à Internet?”) do questionário socioeconômico. As entradas que possuíam valores faltando foram descartadas, porém um estudo mais aprofundado do impacto do descarte desses dados seria necessário.

As questões escolhidas possuem respostas que são letras, cada uma representado uma categoria ou faixa de resposta. Para se trabalhar apenas com valores numéricos, foi adotado a simplificação de cada letra ser trocada pelo número correspondente em ordem alfabética. Além disso, foi realizado uma normalização dos dados. As categorias e faixas das respostas de todas as perguntas estão disponibilizadas em (8).

Dois algoritmos de clusterização do pacote scikit-learn foram utilizados, o *K-Means* e o *Agglomerative Clustering* para separar o conjunto de dados em dois grupos.

A Tabela 2 mostra o resultado obtido utilizando o algoritmo *K-Means*. É possível observar a separação dos inscritos entre dois grupos. Ao se excluir entradas com dados faltantes, apenas inscritos que estão cursando o ensino médio e que concluíram em 2021 foram considerados. A média e o desvio padrão foram adotados para se comparar os atributos dos dois grupos. É possível notar que o grupo 2 apresentou maior média de notas em todas as provas comparado ao grupo 1. A média de idade do grupo 2 é ligeiramente menor que a do grupo 1. Os pais dos inscritos do grupo 2 possuem maior grau de estudo em relação aos inscritos do grupo 1. A renda mensal familiar do grupo 2 é maior que a renda mensal do grupo 1. Além disso, a maioria dos inscritos do grupo 2 estudou em escola particular, enquanto a maioria do grupo 1 estudou em escola pública.

Tabela 2 – Resultado do algoritmo *K-Means*.

		Grupo				Grupo	
		1	2			1	2
Faixa etária	mean	2.87	2.41	**Nota da prova de redação**	mean	580.76	727.27
	std	1.27	0.61		std	184.88	158.01
Situação de conclusão do Ensino Médio	mean	2.00	2.00	**Q001**	mean	4.29	5.81
	std	0.00	0.00		std	1.81	1.21
Ano de conclusão do ensino médio	mean	0.00	0.00	**Q002**	mean	4.54	6.00
	std	0.00	0.00		std	1.52	1.01
Nota da prova de Ciências da Natureza	mean	466.56	538.60	**Q005**	mean	3.96	3.76
	std	67.13	83.36		std	1.31	1.05
Nota da prova de Ciências Humanas	mean	495.33	567.15	**Q006**	mean	3.68	9.88
	std	83.62	91.51		std	2.14	4.28
Nota da prova de Linguagens e Códigos	mean	481.92	546.59	**Q025**	mean	1.89	2.00
	std	71.21	70.52		std	0.31	0.04
Nota da prova de Matemática	mean	503.92	608.23	**Tipo de escola do Ensino Médio***	mean	-0.99	0.65
	std	92.60	114.44		std	0.17	0.76

*foi adotado os valores para escola pública de -1 e escola particular +1.

Tabela 3 – Resultado do algoritmo *Agglomerative Clustering*.

		Grupo				Grupo	
		1	2			1	2
Faixa etária	mean	2.85	2.39	**Nota da prova de redação**	mean	585.79	736.57
	std	1.25	0.59		std	185.27	154.44
Situação de conclusão do Ensino Médio	mean	2.00	2.00	**Q001**	mean	4.41	5.65
	std	0.00	0.00		std	1.84	1.28
Ano de conclusão do ensino médio	mean	0.00	0.00	**Q002**	mean	4.66	5.85
	std	0.00	0.00		std	1.55	1.07
Nota da prova de Ciências da Natureza	mean	469.64	540.93	**Q005**	mean	3.96	3.73
	std	69.18	83.59		std	1.31	1.02
Nota da prova de Ciências Humanas	mean	498.49	569.15	**Q006**	mean	4.12	9.42
	std	85.19	91.17		std	2.83	4.51
Nota da prova de Linguagens e Códigos	mean	484.86	548.03	**Q025**	mean	1.90	2.00
	std	72.37	70.11		std	0.30	0.00
Nota da prova de Matemática	mean	508.43	611.44	**Tipo de escola do Ensino Médio***	mean	-1.00	1.00
	std	95.74	114.39		std	0.08	0.00

*foi adotado os valores para escola pública de -1 e escola particular +1.

O resultado para o algoritmo de *Agglomerative Clustering* é apresentado na Tabela 3. Os números apresentados seguem as mesmas tendências obtidas para o algoritmo *K-Means*, ou seja, maiores notas médias em todas as provas para o grupo 2, maior escolaridade dos pais dos inscritos do grupo 2, maior renda mensal familiar para os inscritos do grupo 2. Ambos resultados estão de acordo com o estudo apresentado em (7), indicando que a performance no ENEM pode estar associada ao nível socioeconômico dos participantes. Porém vale ressaltar que são necessários estudos mais aprofundados para que seja possível obter conclusões mais exatas.

Algoritmos de Classificação:

A lei de cotas é uma ação afirmativa que garante parte das matrículas em instituições de ensino superior federais para alunos oriundos do ensino médio público (9). A ideia deste estudo é avaliar a possibilidade de construir um modelo para prever se um aluno é realmente oriundo de uma escola pública ou não, com base nos dados fornecidos na inscrição do ENEM. Assumindo que todas as respostas contidas nos microdados do ENEM 2021 são verídicas, foi utilizado o atributo tipo de escola (pública ou privada) para treinar algoritmos de classificação.

Assim como no caso anterior, foram realizados os mesmos procedimentos de pré-processamento dos dados. Foram escolhidos os seguintes atributos para treinamento dos algoritmos: faixa etária de cada inscrito, a situação de conclusão do ensino médio, o ano em que concluiu ou pretende concluir o ensino médio, o tipo de escola (privado ou pública), as respostas das questões Q001, Q002, Q005, Q006 e Q025 do questionário socioeconômico, e as notas nas provas objetivas e redação. Além disso, para o treinamento foram utilizadas quantidades iguais de entradas da categoria escola pública e da categoria escola privada.

Diferentes modelos clássicos foram avaliados, sendo eles: *Logistic Regression*, *K-Nearest Neighbors*, *Random Forest*; e *Support Vector Machine*. Foram utilizados os parâmetros padrões em cada modelo e como medida de comparação foi adotada a acurácia média de cada modelo, utilizando-se o algoritmo de validação cruzada. Foram obtidos os seguintes resultados:

- O modelo *Logistic Regression* resultou em um acurácia de 0.80
- O modelo *K-Nearest Neighbors* resultou em um acurácia de 0.72
- O modelo *Random Forest* resultou em um acurácia de 0.80
- O modelo *Support Vector Machine* resultou em um acurácia de 0.80

O modelo K-Nearest Neighbors apresentou a pior performance e os demais apresentaram resultado próximos entre si. Uma otimização ou troca dos modelos poderá ser necessária para que o resultado seja melhor. Entretanto é preciso ter cautela ao analisar os resultados obtidos já que um modelo mal treinado pode gerar previsões tendenciosas (10).

Conclusão

Neste relatório foram utilizados os microdados disponibilizados pelo INEP a respeito do ENEM do ano de 2021. Através da exploração inicial deste conjunto de dados foram observadas algumas informações tais como a quantidade de inscrições por regiões do país e as distribuições das notas, nas provas objetivas e na redação, em função de parâmetros socioeconômicos. Em seguida foram utilizados algoritmos de aprendizado de máquina para separar o conjunto de dados em grupos com características próprias. Observou-se uma relação entre fatores socioeconômicos e as notas nas provas. Também foi proposto a criação de um modelo para prever o tipo de escola frequentada pelo aluno, privada ou pública, como sistema antifraude para o ingresso no ensino superior pelo sistema de cotas. Esse conjunto de dados possibilita uma infinidade de análises, algumas da quais foram exploradas neste relatório, entretanto vale ressaltar que estudos mais profundos são necessários para que conclusões mais assertivas possam ser obtidas.

Referências

1. **Teixeira, Instituto Nacional de Estudos e Pesquisas Educacionais Anísio.** Microdados do Enem 2021. [Online] 2022. https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem.

2. —. *Entenda a sua nota no Enem: guia do participante.* Brasilia : s.n., 2021.

3. **Estatística, Instituto Brasileiro de Geografia e.** [Online] https://www.ibge.gov.br/estatisticas/sociais/populacao/9103-estimativas-de-populacao.

4. **News, BBC.** G1. [Online] 2021. https://g1.globo.com/educacao/noticia/2021/08/02/enem-o-que-explica-menor-numero-de-inscritos-na-prova-em-mais-de-uma-decada.ghtml.

5. **(Inep), Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira.** Divulgados dados sobre impacto da pandemia na educação. [Online] 2022. https://www.gov.br/inep/pt-br/assuntos/noticias/censo-escolar/divulgados-dados-sobre-impacto-da-pandemia-na-educacao.

6. **Castro, Jorge Abrahão de.** Evolução e desigualdade na educação brasileira. *Educação & Sociedade.* 2009.

7. **Senkevics, Adriano Souza.** *O acesso, ao inverso: desigualdades à sombra da expansão do ensino superior brasileiro, 1991-2020.* Universidade de São Paulo (USP). São Paulo : s.n., 2021. Tese de Doutorado.

8. **(inep), Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira.** [Online] https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem.

9. **Educação, Ministério da.** Lei de cotas para o ensino superior. [Online] http://portal.mec.gov.br/cotas/perguntas-frequentes.html.

10. **Lira, Ana Lídia.** *Discriminação em algoritmos de inteligência artificial: uma análise acerca da LGPD como instrumento normativo mitigador de vieses discriminatórios.* Universidade Federal do Ceará. 2021. TCC.