Aplicação de Ferramentas de Data Science na Avaliação de Desempenho de Alunos do Ensino Médio por meio do ENEM

Emanuel M. Lima
Discente do curso de Engenharia Mecatrônica
Campus Alto Paraopeba (CAP)
Universidade Federal de São João del-Rei (UFSJ)
emanuel.lima@ufsj.edu.br

Michel C. R. Leles
Departamento de Tecnologia (DTECH)
Campus Alto Paraopeba (CAP)
Universidade Federal de São João del-Rei (UFSJ).
mleles@ufsj.edu.br

***Resumo*—Este trabalho apresenta uma aplicação de ferramentas de Ciência de Dados para uma análise dos microdados do ENEM - Exame Nacional do Ensino Médio - entre 2017 e 2019, extraindo informações acerca dos impactos de diversas características geográficas e socioeconômicas no desempenho e na probabilidade de desistência do candidato e fornecendo análises pormenorizadas através de técnicas de visualização de dados e *Big Data*. Apesar de uma significante ausência de dados sobre as escolas, foi possível explorar características importantes, dentre as quais as mais impactantes foram renda familiar, escolaridade dos pais e disponibilidade de acesso à *Internet*, ao passo que o sexo dos participantes apresentou pouca discrepância a princípio. Tais análises almejam fundamentar discussões e subsidiar o processo de tomada de decisão de gestores públicos quanto a políticas direcionadas à educação.**

***Index Terms*—Ciência de Dados, Aprendizado de Máquina, Avaliação Microdados do ENEM**

1. Introdução

A Ciência de Dados (*Data Science*) é uma nova área interdisciplinar que, devido ao crescente número de dados gerados no mundo, tem se tornado cada vez mais determinante na busca de soluções inovadoras para problemas que impactam a sociedade. O objetivo primordial é transformar dados brutos em conhecimento que possa servir de suporte para o processo de tomada de decisões. Diz-se interdisciplinar pois, para alcançar esse objetivo, são utilizadas técnicas de programação para conduzir uma aplicação de conceitos estatísticos a fim de se explorar os dados e obter um modelo matemático adequado ao problema em questão por meio de *Machine Learning* - Aprendizado de Máquina.

Essa combinação versátil cria inúmeras possibilidades de aplicações nas mais diversas áreas da sociedade e onde quer que seja conveniente tomar alguma decisão fundada em dados, até mesmo em um contexto socioeconômico. Nesse contexto, circundado de interpretações e discussões subjetivas inerentes à sua natureza, muitas vezes é um desafio obter uma amostra de dados significativa que represente a população de forma fiel. Entretanto, em determinados momentos como, por exemplo, na implementação de políticas públicas, cujo objetivo é atender às necessidades públicas, se torna algo necessário para monitorar e entender a situação daqueles aos quais se deseja atender de forma eficiente.

No presente projeto, esse arsenal de habilidades será utilizado para análise comparativa a partir da base de microdados recentes do Exame Nacional do Ensino Médio (ENEM).

Em 2015 e 2016, respectivamente, Silveira et al [1] e Kleinke [2], da sociedade Brasileira de Física, publicaram seus papéis contendo análises bem fundamentadas estatisticamente a respeito do impacto socioeconômico sobre o desempenho dos candidatos exclusivamente nas questões de Física do ENEM. Já em 2018, Santana [3] publicou sua dissertação de mestrado pela Universidade Federal de Pernambuco (UFPE) abrangendo a técnica de mineração de dados com o intuito de prever o desempenho dos participantes, fazendo um uso mais explícito de técnicas de Aprendizado de Máquina. Mais adiante, em 2019, Nascimento [4] realizou, em sua tese de Doutorado, uma análise dos microdados do ENEM com métricas voltadas para o campo da Educação, mostrando que o número de publicações no campo de Ensino que continham citações ao ENEM aumentou consideravelmente desde o início do século XX!.

A maioria dos trabalhos, no entanto, aborda apenas os impactos nas notas dos candidatos e, como consta no relatório do Inep sobre o ENEM 2019 [5], o custo para a realização do Exame foi de 105,52 reais por inscrito, já descontada a taxa de inscrição, implicando em uma perda de mais de 122 milhões de reais aos cofres públicos devido à abstenção dos candidatos. Isso demonstra que a abstenção ou desistência dos participantes se mostra, também, um fator que merece uma maior atenção de estudos e análises.

No presente trabalho, pretende-se fazer uma análise de desempenho médio tanto das notas do ENEM quanto do perfil dos desistentes em diferentes anos por meio de algumas características: i) raça declarada; ii) renda familiar; iii) nível de escolaridade dos familiares; iv) disponibilidade de *Internet* em casa; v) tempo de conclusão do Ensino Médio; vi) sexo ou gênero; vii) idade; e viii) macrorregião política de residência.

Entretanto, ao se aplicar Ciência de Dados nos microdados do ENEM, não se espera alcançar uma análise

exaustiva de sua complexidade tampouco negar a pluralidade e subjetividade de qualquer interpretação que possa surgir. Nesse contexto, o presente projeto é direcionado a uma aplicação que envolve um problema de natureza pública, de forma que a solução computacional desenvolvida possa eventualmente vir a auxiliar os gestores públicos em suas decisões.

O texto está organizado de acordo com: Seção 2 trata da fundamentação dos principais conceitos envolvidos; Seção 3 discute a metodologia proposta; Seção 4 mostra a análise dos resultados; enquanto a Seção 5 aborda as considerações finais e direções futuras de pesquisa.

2. Fundamentação Teórica

Nessa seção serão brevemente descritos os principais conceitos a serem utilizados ao longo do presente trabalho.

2.1. Exame Nacional do Ensino Médio

Desde 1998, o governo brasileiro realiza anualmente o Exame Nacional do Ensino Médio - ENEM - que é uma ferramenta de avaliação de desempenho da Educação e, hoje, funciona como o principal meio de ingresso ao Ensino Superior, substituindo grande parte dos tradicionais vestibulares, promovendo diversidade e mobilidade para estudantes e egressos do Brasil inteiro [1].

Ao longo do tempo, as provas do ENEM foram aprimoradas e recebendo mais atenção, assim como seu processo de inscrição que apresenta um questionário socioeconômico de preenchimento obrigatório cujo objetivo é conhecer alguns aspectos socioeconômicos daqueles candidatos a uma vaga no Ensino Superior. Dessa forma, o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), autarquia responsável pela organização do Exame, tornou possível a coleta de dados importantes para serem analisados, interpretados e, então, utilizados em prol da melhoria da Educação.

2.2. Ciência de Dados

A Ciência de Dados ou *Data Science* pode ser entendida como uma área interdisciplinar que busca, através da união entre Estatística e Computação, fazer uso de grandes quantidade de dados para levar soluções mais sólidas para problemas científicos, sociais ou econômicos de relevância para a indivíduos, empresas e sociedade [6]. Uma área em constante crescimento que tem sido adotada desde micro-empresas até gigantes multinacionais de tecnologia com o objetivo de realizar alguma tarefa de forma mais eficiente.

A partir de um problema a ser otimizado, utiliza-se da Estatística para se obter uma visualização dos seus dados para uma interpretação humana das características que esses dados apresentam (Análise dos Dados) para, então, utilizar-se de algoritmos computacionais de otimização e automação de conceitos estatísticos para classificar dados já existentes e/ou prever eventos futuros (*Machine Learning*).

2.3. Testes Estatísticos

Na Análise Exploratória dos Dados, a Estatística é utilizada para validar qualquer análise ou interpretação através de Testes de Hipóteses que descrevem a natureza das amostras testadas e revelam a significância de cada característica dessas amostras [7]. Para cada um dos diversos tipos de testes para diversos tipos de dados e finalidades, há exigências a serem cumpridas para que sejam realizados de maneira correta. Caso contrário, pode ser necessário considerar a realização de outro teste para validar a hipótese desejada ou até mesmo executar uma reamostragem.

Para as análises deste trabalho, que envolvem variáveis categóricas com dois ou mais grupos sendo comparados, como sexo ou grupos de renda familiar, o tipo de teste desejado é de comparação de médias ou variâncias e a decisão por qual teste a executar vai passar, principalmente, pelos resultados de testes de cada amostra. Tais testes de comparação podem revelar quais características são determinantes para cada análise, podendo servir de base para o algoritmo de Aprendizado de Máquina.

Portanto, na Ciência de Dados, uma análise pautada em Estatística tende a levar a interpretações mais sólidas e consistentes acerca do problema enfrentado, além de algoritmos mais eficazes e menos enviesados [7].

2.4. *Big Data*

Big Data é um termo designado para descrever tecnologias e métodos de se trabalhar a partir de grandes volumes de dados para fornecer informação e conhecimento de qualidade a quem deva tomar decisões no tempo certo [8]. Tem se tornado cada vez mais objeto de discussão e estudo à medida que mais e mais dados são gerados por segundo e que precisam ser tratados de forma a retornar valor para a sociedade.

Os microdados do ENEM contém as informações acerca dos inscritos, sendo constituído de mais de 100 colunas para cada um dos milhões de inscritos por ano. Entre 2017 e 2019, somam-se mais de 15 milhões de linhas obtidas de forma confiável por meio do Inep com as mais variadas características e tipos que necessitam ser tratados de forma computacionalmente eficiente a fim de agregar algum valor no processo de tomada de decisão. Segundo Taurion [9], essa situação pode atender aos fatores da fórmula que resume *Big Data*: Volume, Variedade, Velocidade, Veracidade e Valor.

3. Abordagem Proposta

Comumente, um problema de Ciência de Dados é abordado seguindo uma ordem de etapas, sendo:

1) Entendimento do problema;
2) Obtenção e tratamento dos dados;
3) Análise Exploratória dos Dados;
4) Aplicação de algoritmo de Aprendizado de Máquina; e
5) Comunicação dos resultados para tomada de decisão.

Para o tema apresentado, propôs-se uma abordagem similar. Haja vista que a ideia fora previamente discutida e esclarecida, portanto, tomar-se-á de pressuposto um entendimento prévio da natureza do que aqui se propõe a analisar.

Para um maior detalhamento dos processos realizados, algumas dessas etapas foram subdividas para que seja transmitidas as ideias com maior clareza.

3.1. Obtenção dos Dados

Os microdados relativos aos candidatos do ENEM são disponibilizados de forma anônima no portal do Inep àqueles que lhes forem de interesse. Os dados são dispostos em arquivos de texto no formato *csv* separados por ano de realização da prova acompanhados de arquivos de planilha (*.xlsx*) contendo seus respectivos dicionários de variáveis, com os quais é possível obter o tipo de dado esperado e a descrição de cada variável.

Neste trabalho, serão utilizados os microdados referentes aos anos de 2017, 2018 e 2019. Julga-se importante ressaltar que a maioria dos arquivos anuais dispostos no portal Inep anteriores a 2017 apresentaram problemas com *download*, alguns sequer completaram-no mesmo após diversas tentativas em dias também diversos. Tendo em vista que cada ano houve algo em torno de 5 ou 6 milhões de inscritos (linhas) por ano e descritos em mais de 120 características (colunas), decidiu-se por suficiente para uma boa modelagem os dados destes três anos.

Devido ao tamanho dos arquivos, notou-se que seria inviável manipulá-los da forma mais convencional com a biblioteca *pandas*, visto que haveriam operações muito ineficientes em um conjunto de dados tão grande. Foi necessário, então, o uso de ferramentas mais robustas e eficientes para arquivos maiores. Para isto, foram analisadas algumas possibilidades, como a biblioteca *pandas* com o recurso de dividir os dados em *chunks*, a biblioteca *dask* e *pyspark*.

Com ***pandas*** e o recurso de se dividir o arquivo em repartições (*chunks*), é possível se aproveitar da facilidade e amigabilidade de uso de uma das bibliotecas mais populares para manipulação de dados em Ciência de Dados. Entretanto, por se fazer uso de espaço da memória RAM para se armazenar o conjunto inteiro de dados, faz-se necessária a realização de um pré-processamento precoce dos dados, de forma a excluir linhas e colunas antes mesmo de saber seu conteúdo. Isso aumenta as chances de imprecisões na modelagem e uma possível perda de dados de forma precoce.

Já com relação à biblioteca ***dask***, a qual mantém uma familiaridade com a biblioteca anterior, é possível ler um conjunto de dados maior de forma que armazene em memória RAM somente uma parte do mesmo. Ao se utilizar operações vetoriais e adaptar ao uso "colunar"dos dados, é possível se obter uma maior eficiência na manipulação de maiores *datasets*.

Além disso, há o ***pyspark***. Trata-se de uma API do Apache Spark de armazenamento de dados de forma "colunar"em sistemas Hadoop e serviços de nuvem. É bastante indicado para trabalhos com *Big Data* por ser mais eficiente, já que os dados costumam ser agrupados em colunas e, portanto, não sendo necessário ler todo o conjunto de dados quando se deseja as informações de somente algumas colunas.

Visando trabalhar proporcionando escalabilidade ao projeto, transformou-se os dados, ainda antes de se começar a tratá-los, de *csv* para o formato *parquet*, o qual possibilita um armazenamento e manipulação eficientes para quaisquer bibliotecas mencionadas anteriormente. Tal procedimento foi realizado com cada um dos arquivos anuais obtidos diretamente do portal do Inep e concatenados posteriormente.

3.2. Tratamento dos Dados

Para que seja possível uma análise e a modelagem dos dados para o algoritmo de Aprendizado de Máquina, é necessária uma adequação dos dados obtidos do Inep. Com os dados já no formato *parquet* e lidos com *dask*, pode-se dar início ao tratamento de dados.

O primeiro passo foi analisar as colunas e reduzir o máximo possível, centralizando, por exemplo, a informação de nota, visto que é fornecida distribuída em diversas colunas representando as notas de cada prova e as competências da Redação. Assim, para se facilitar o uso de algum algoritmo preditivo ou até mesmo uma análise gráfica, as notas serão representadas pela média de suas componentes em uma única coluna.

Similarmente, são providos quatro indicadores de presença (um para cada prova) mais um de *status* da Redação. Entretanto, diferentemente das notas, determinou-se como candidatos desistentes ou ausentes aqueles que não obtiveram uma média válida em alguma das provas. Somente este novo indicador já é suficiente para funcionar como coluna-alvo numa classificação, por exemplo.

Os candidatos, no momento da inscrição, têm a possibilidade de solicitar algum recurso especializado ou específico. Como há um número muito diversificado de recursos e aqueles que os solicitam são um número minoritário, é razoável agrupá-los em indicadores gerais sem que seja dado um peso muito diferente a essas informações. As colunas por eles representadas podem, então, serem excluídas para o fim deste trabalho.

A última "pré-análise"realizada individualmente nos arquivos de cada ano deu-se pelo tratamento dos valores faltantes. Preferiu-se realizar este passo - importantíssimo - individualmente para uma concatenação mais limpa e melhor escalabilidade. Sendo assim, ao verificar a incidência de valores nulos, observou-se que, em todos os anos analisados, em torno de 70% dos dados referentes às escolas (como código, tipo de ensino, localidade, situação de funcionamento, etc.) estavam faltantes. Estas colunas foram devidamente excluídas por não fornecerem dados em proporções significativas em relação ao tamanho do conjunto dos dados.

Com as informações mais limpas, pode-se realizar sua preparação para o algoitmo de ML. Classificou-se, pois, as variáveis quanto a seus tipos entre quantitativas, qualitativas dicotômicas, qualitativas ordinais e qualitativas nominais. As quantitativas já se encontram, a esse ponto, prontas.

Quanto às dicotômicas, uma simples codificação, como a variável (*SEXO*) que categoriza o sexo do participante, que originalmente possuía os valores "M"e "F"passaram a ter os valores "0"e "1", mais facilmente entendidos por algoritmos de regressão, por exemplo. Todavia, ao se aprofundar nas variáveis categóricas, é entendido que se deve tomar muito cuidado e é necessária uma análise humana mais criteriosa.

Isto porque há as categóricas ordinais, as quais podem representar ordens e/ou discretizações de valores contínuos (por exemplo, categorias de renda e categorias de nível de escolaridade), e as categóricas nominais, que representam categorias independentes entre si (Unidades da Federação ou profissões, entre outros).

As categóricas ordinais foram ordenadas por sua natureza e codificadas em números inteiros (*Label Encoding*), sendo, por exemplo, "A- que representa quem recebe nenhuma renda familiar - , se torna zero, "B- quem recebe até 998 reais por mês - recebe o valor um e assim por diante. Já as categóricas nominais, optou-se por realizar o *Dummy Encoding* ou *One Hot Enconding*, obtendo, portanto, colunas representando cada uma das categorias com valores inteiros ("0"ou "1").

Desta forma, o *dataset* se torna utilizável como entrada para a maioria dos algoritmos de ML.

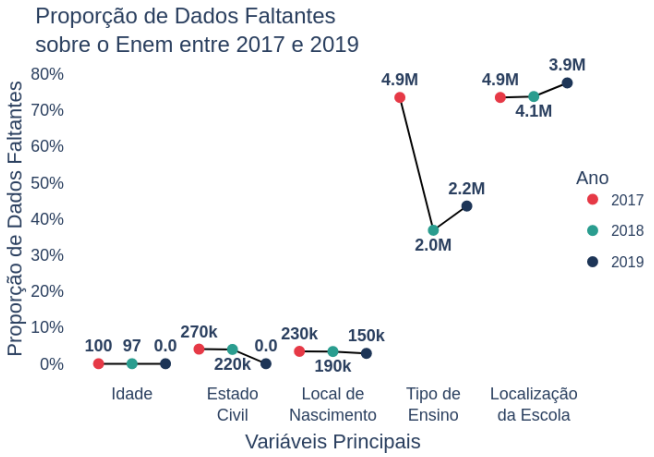


Figura 1: Proporção de dados faltantes em relação ao total de inscritos no ENEM entre 2017 e 2019 nas variáveis com maiores níveis de omissão. É perceptível a deficiência na coleta dos dados acerca das escolas, como na variável TP_ENSINO, que descreve se a educação é pública ou privada, bem como nas demais variáveis que possam especificar o município, UF, tipo de localização (urbana ou rural) e até mesmo a situação de funcionamento das escolas.

3.3. Análise Exploratória dos Dados

Nesta etapa do processo, foram realizadas análises de algumas variáveis-chave a fim de se mensurar o impacto geográfico, econômico e social no desempenho dos candidatos do ENEM. Para isso, utilizou-se as bibliotecas Plotly, Seaborn, Pandas e a API para Python do Apache Spark - PySpark - para se trabalhar com o volume total de dados de forma mais eficiente.

Entre as principais expectativas de análises estavam comparações de impacto nas médias das notas ou no índice de desistência dos candidatos com base no tipo de ensino (público, privado ou em casa). Entretanto, a Figura 1 mostra a disponibilidade dos dados apresentados pelo Inep nos grupos de variáveis mais afetadas por falta de dados e nota-se que, apesar de ter havido melhoras desde 2017 - quando se havia dados faltantes até a respeito da idade de alguns inscritos -, a taxa de dados que estão faltando com as informações sobre as escolas é ainda surpreendentemente alta.

O alto índice de omissão de dados descrito na Figura 1 eleva expressivamente a imprecisão de qualquer análise comparativa utilizando os dados sobre as escolas ou até mesmo o tipo de ensino dos candidatos. Contudo, com estas variáveis já devidamente eliminadas, ainda restam diversas análises que podem ser de interesse para um melhor entendimento acerca de outros eventuais impactos no desempenhos dos inscritos.

Foram obtidos, então, alguns gráficos para demonstrar os dados agrupados de forma mais visualmente amigável, como gráficos de barras e mapas coropléticos comparando as macrorregiões e Unidades Federativas do Brasil, assim como municípios de Minas Gerais e de outros Estados que possam ser de interesse em cada análise além de testes estatísticos para fundamentar cada uma.

Para tanto, realizou-se diferentes amostragens. Como a população é de 16 milhões de inscritos entre os anos de 2017 e 2019, considera-se uma população infinita. Para garantir a independência das amostras (uma exigência de muitos testes), ou seja, que não haja nenhuma chance de duas amostras serem de alguma forma relacionadas - uma mesma pessoa inscrita em dois anos distintos, por exemplo - reduziu-se o tamanho da população para apenas o ano de 2019, o mais recente com dados disponíveis. Isto não altera a natureza das amostras, visto que é idêntica em todos os anos observados.

A todas as variáveis foram realizadas amostragens estratificadas adequadas a cada variável com proporção α de 10% da população referente ao ano de 2019 para submissão de testes, somando algo em torno de 510 mil elementos ao todo, minimizando, assim, as chances de ocorrências de erros nos testes de hipótese. Dentre as métricas mais gerais estavam, por exemplo, a medida de assimetria (*skewness*) em relação a uma distribuição normal e três testes de normalidade (Shapiro-Wilk, D'Agostino e Anderson-Darling) para consenso acerca da normalidade ou não das amostras.

A variável TP_SEXO, por meio da qual os inscritos declararam seu gênero biológico em masculino ou feminino, podendo ser classificada, portanto, como uma variável categórica nominal dicotômica, teve como resultado a hipótese nula de normalidade das amostras rejeitada, significando que não se tem certeza sobre a curva de distribuição à qual as amostras pertencem. Portanto, foi utilizado o teste não-paramétrico Mann-Whitney para avaliar se a diferença entre as medianas dos grupos comparados é significativa estatisticamente, visto que o pré-requisito de normalidade

para um teste paramétrico como teste T não foi atendido, mesmo com eventual correção de Welch.

Para as demais variáveis categóricas politômicas, isto é, com três ou mais grupos a serem analisados na variável independente, como TP_COR_RACA e TP_ANO_CONCLUIU, através das quais o inscrito manifestou a cor à qual autodeclara pertencer e há quanto tempo concluiu o Ensino Médio, respectivamente, o resultado também foi ter as respectivas hipóteses nulas H_0 de normalidade rejeitada.

Entretanto, foram realizados testes mais específicos para análises com mais de dois grupos categóricos. Após os testes de normalidade, foram realizados testes de equivalência de variância, como o teste de Brown-Forsythe, os quais também tiveram essas hipóteses rejeitadas, significando então que as variâncias dos grupos comparados são distintas e, portanto, torna-se inválida a realização de um teste paramétrico como ANOVA. Desta forma, a alternativa restante para o objetivo à vista foi a utilização do teste não paramétrico de Kruskal-Wallis.

Para finalizar a etapa de Análise Exploratória e Visualização dos Dados, foram analisadas algumas variáveis presentes no formulário socioeconômico de preenchimento mandatório por parte dos inscritos, contendo informações acerca do nível de escolaridade do pai, da mãe ou responsável, renda mensal familiar e disponibilidade de internet em casa.

4. Resultados e Discussão

Para tratar dos dados de forma eficiente, faz-se necessário um melhor conhecimento acerca do que representam por meio de características como, por exemplo, de onde são, qual seu gênero, como se declaram em relação à sua cor e em quanto tempo concluíram o Ensino Médio, bem como características ao seu redor como o nível de escolaridade dos seus familiares, sua renda familiar e a disponibilidade de *Internet* em casa. O objetivo é visualizar esses dados e seus eventuais impactos na nota média dos inscritos e/ou na razão de desistência na prova.

4.1. Informações Gerais

É um importante passo obter primeiramente informações de cunho mais geral abrangendo todo o conjunto de dados para depois se realizar as amostragens necessárias para os testes estatísticos. Tendo isso em vista, a primeira análise foi saber de onde os inscritos realizaram a prova entre o ENEM 2017 e 2019 através da variável SG_UF_PROVA a qual contem a sigla da Unidade Federativa a partir da qual foi realizado o Exame, apresentado no mapa da Figura (ref).

A partir do mapa da Figura 2, é possível notar uma predominância já esperada do estado de São Paulo seguido de Minas Gerais, já que representam as duas maiores populações entre os estados brasileiros. É possível agrupar, então, as informações de acordo com as Macrorregiões do Brasil, isto é, Centro-Oeste, Norte, Nordeste, Sudeste e Sul, em ordem alfabética, e ter uma ideia da origem dos inscritos

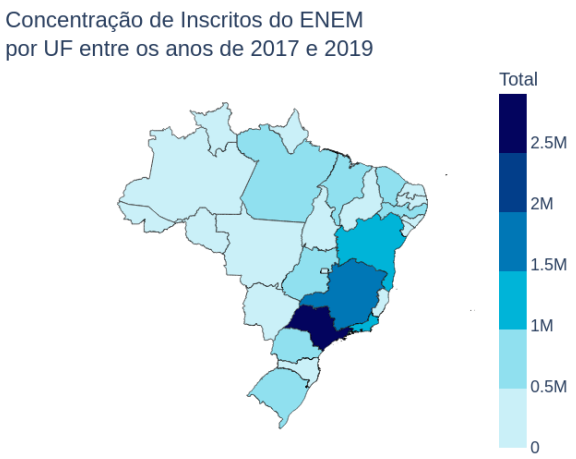


Figura 2: Inscritos no ENEM entre 2017 e 2019 por Estado do Brasil.

que vá além da proporção de povoamento de cada estado individualmente.

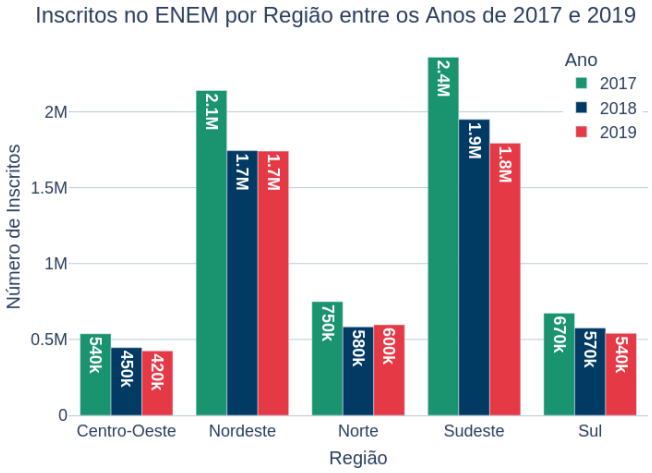


Figura 3: Inscritos no ENEM entre 2017 e 2019 por Macrorregião.

De acordo com o gráfico da Figura 3, é possível inferir que a proporção inscritos provenientes das regiões Sudeste e Nordeste entre os anos de 2017 e 2019 são bem próximas e predominantes no território nacional, principalmente no ano de 2019 no qual o número de inscritos reduziu em comparação a 2018 e 2017 e as duas regiões registraram um número por volta de 1,7 milhões de inscritos.

De acordo com o IBGE, na PNAD Contínua de 2019 consta que o número de mulheres é superior ao de homens no Brasil, sendo 51,8% e 48,2%, respectivamente. Esta diferença, porém, é ainda mais acentuada entre os inscritos do ENEM entre 2017 e 2019, visto na Tabela 1.

Tabela 1: Distribuição dos Inscritos no ENEM entre 2017 e 2019

Sexo	Quantidade	Porcentagem
Feminino	9.949.413	59,04%
Masculino	6.901.469	40,96%

Além disso, se separar os inscritos por blocos de idades, como na Figura 4, nota-se que o número de pessoas do sexo feminino é constante e consistentemente maior que as pessoas de sexo masculino.

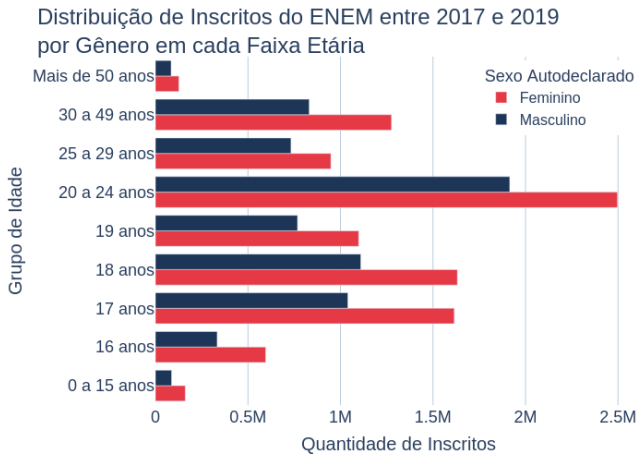


Figura 4: Distribuição dos Inscritos do ENEM entre 2017 e 2019 por Gênero e Faixa Etária.

Como já esperado, o ENEM é realizado, em sua maioria, por jovens entre 17 e 19 anos haja vista que o objetivo do ENEM é ingressar no Ensino Superior, sendo de costume realizar o Exame logo após a conclusão do Ensino Médio. Entretanto, é possível notar, porém, a partir da Figura 4, que há uma presença significativa de inscritos entre 2017 e 2019 no grupo de idade entre 30 e 49 anos, além de que o número de inscritos com mais de 50 anos de idade é semelhantes àqueles até 15 anos, com detalhe curioso: no período observado, 98 pessoas do sexo masculino e 75 do sexo feminino com 80 anos ou mais realizaram o Exame, inclusive uma pessoa de 98 anos de idade.

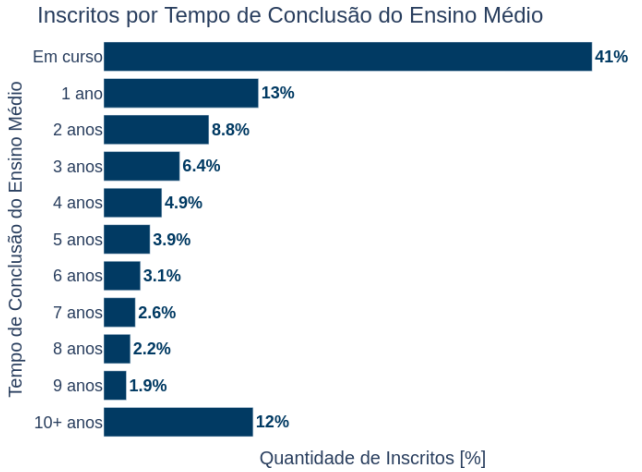


Figura 5: Distribuição dos Inscritos do ENEM 2019 de acordo com o Tempo de Conclusão do Ensino Médio.

Com a Figura 5, percebe-se que em torno de 41% dos inscritos no ENEM entre 2017 e 2019 não completaram o Ensino Médio, sendo composto de treineiros ou simplesmente por estarem finalizando o Terceiro Ano do Ensino Médio no momento da aplicação do Exame.

Além disso, a Figura 5 endossa aquilo mencionado anteriormente a respeito da presença significativa de inscritos entre 30 e 49 anos de idade ao revelar que há uma presença muito grande de inscritos que concluíram o Ensino Médio há dez anos ou mais. Trata-se, na realidade, de um grupo que representa, aproximadamente, 12,5% de todos os inscritos no ENEM entre 2017 e 2019, ou seja, uma proporção semelhante àqueles com 1 ano de conclusão do Ensino Médio. É uma proporção relevante e significa que muitas pessoas encontram no ENEM uma oportunidade de retornar aos estudos e, talvez, até uma oportunidade de recolocação no mercado através do Ensino Superior mesmo após tanto tempo.

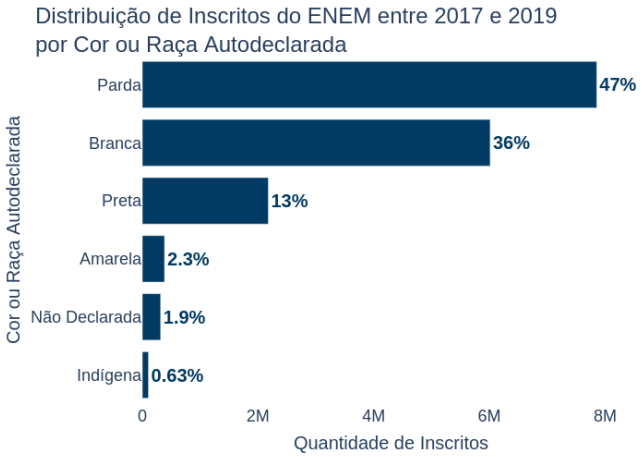


Figura 6: Distribuição da Cor ou Raça Autodeclarada dos Inscritos no ENEM entre 2017 e 2019.

Outro passo importante para conhecer mais a respeito dos inscritos é saber a cor ou raça à qual eles se identificam e autodeclaram pertencer. Como é possível observar na Figura 6, e como esperado, há uma predominância de inscritos que se autodeclaram da cor parda, tendo em vista todo o processo histórico de formação do País, apesar do Censo 2010 do IBGE apontar que há mais pessoas que se autodeclaram brancas que pardas no Brasil.

Outro reflexo desse processo, é ter, no período observado, apenas 0,63% dos inscritos que se identificam como indígena, o que, para um País que uma grande parte da população possui ascendência indígena, é um percentual muito pequeno. No entanto, considerando que nos Censos de 2000 e 2010 apenas 0,4% da população se autodeclarava indígena, quer dizer que os percentual de 0,63% dos inscritos é maior que o da população em geral em 2010.

Há outras características que foram armazenadas por meio de um formulário socioeconômico de preenchimento obrigatório a todos os inscritos, entre elas é a respeito do nível de escolaridade do pai ou homem responsável e da mãe ou mulher responsável.

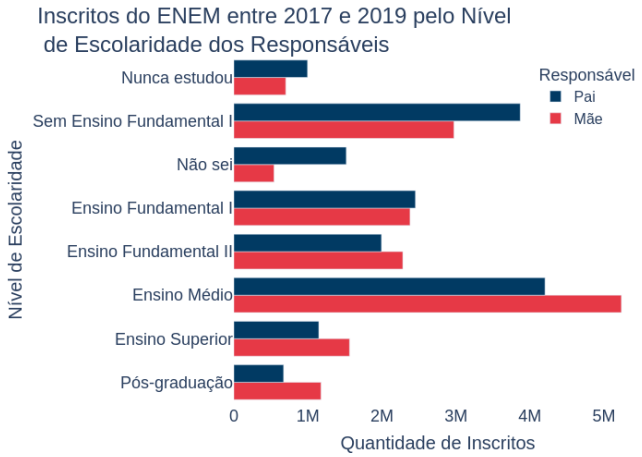


Figura 7: Distribuição dos Inscritos de Acordo com o Nível de Escolaridade do Pai, da Mãe ou Responsável.

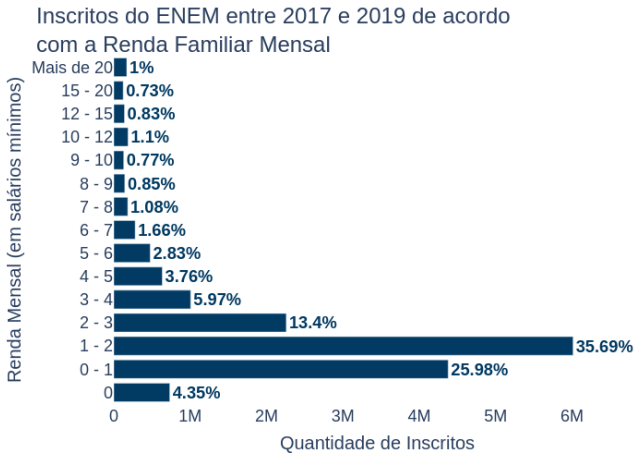


Figura 8: Distribuição dos Inscritos no ENEM entre 2017 e 2019 de acordo com a Renda Familiar Mensal em Salários Mínimos.

No gráfico da Figura 7, há diversos pontos a se considerar à primeira vista. O primeiro - e já esperado - é que há mais que o dobro de inscritos que não sabem sobre os pai que aqueles que não sabem sobre suas mães, por qualquer que seja a razão, mas evidencia um cenário de abandono muito maior por parte dos pais no Brasil.

Outro ponto a destacar é que a maior parte dos inscritos possuem pais que estudaram até o Ensino Médio imediatamente sucedidos por aqueles cujos pais não completaram o que hoje é o Ensino Fundamental I (até a 5ª série). Um retrato da cultura bem como das condições de gerações passadas com pouco acesso ao estudo, seja porque não quiseram ou porque não puderam, visto que muitos começavam a trabalhar desde muito pequenos e acabavam largando os estudos para ajudar em casa. Ao mesmo tempo, o fato de muitos inscritos apresentarem estes dados dá indícios de mudanças para as gerações em formação e vindouras. Um detalhe é que a quantidade de pais que não estudaram é muito semelhante aos que concluíram o Ensino Médio, enquanto o número de mães há uma diferença drástica entre esses grupos.

Ainda com relação ao gráfico da Figura 7, o último ponto a descrever é que, nos maiores níveis de escolaridade - Ensino Fundamental II (até 8ª série), Ensino Médio, Ensino Superior e Pós-Graduação -, há um número superior de mães dos inscritos. Enquanto isso, as categorias daqueles que nunca estudaram, que não completaram a 5ª série ou que não completaram a 8ª série, bem como aqueles dos quais não se tem informações a respeito, há uma predominância dos pais. Isso quer dizer que as mulheres têm um nível de escolaridade muito mais elevado que os homens e tende a continuar essa proporção já que a proporção de mulheres realizando o ENEM são maiores.

Já com relação à distribuição de renda dentre os inscritos, nota-se pelo gráfico da Figura 8, que a maior frequência, um pouco mais de um terço (36%), é de inscritos com renda familiar mensal entre 1 e 2 salários mínimos e que dois terços (66%) vivem com com até 2 salários mínimos. Um cenário muito complexo e que, talvez, carrega diversas dificuldades de acesso à educação a uma parte tão grande dos inscritos que podem refletir nas médias das notas e no índice de desistência geral.

Uma dessas dificuldades pode incluir o acesso à *Internet* em casa, visto que esse meio de acesso a informações, mesmo cada vez mais comum no cotidiano, ainda continua relativamente caro para aqueles que não possuem condições, principalmente quando se olha os custos periféricos associados à *Internet* como o custo de um dispositivo para acessá-la que, considerando que o salário mínimo em 2019 era de 998 reais, uma pessoa que recebesse um salário mínimo precisaria desembolsar algo entre um terço e metade (300 e 450 reais) para adquirir um aparelho celular básico à época.

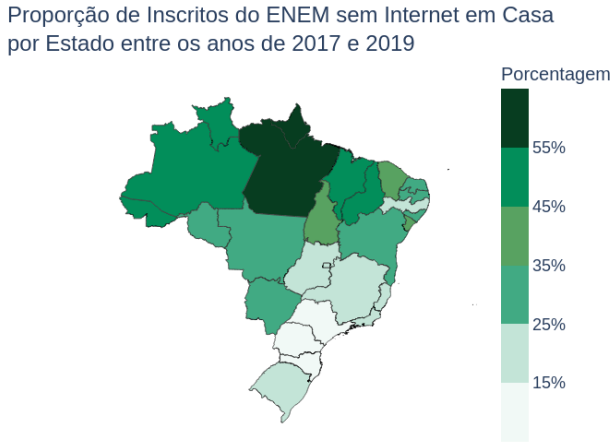


Figura 9: Distribuição dos Inscritos no ENEM entre 2017 e 2019 sem acesso a Internet em Casa.

Com base no mapa da Figura 9, é notória a discrepância de acesso à *Internet* entre as regiões e até chocante a proporção de pessoas sem acesso a um meio tão importante de busca por conhecimento e complemento do aprendizado da escola.

Com essas informações básicas acerca dos inscritos no ENEM entre 2017 e 2019, pode-se avaliar possíveis impactos geográficos ou socioeconômicos no desempenho dos mesmos no Exame, basicamente, de duas formas: diferença nas médias das notas e na proporção de desistência.

4.2. Médias das Notas

Para analisar a significância estatística na diferença das médias das notas em relação a cada uma das características apresentadas anteriormente, é necessário que seja cumprido o requisito de independência dos elementos da amostra e, como já se é sabido de antemão - e, também, pode ser inferido a partir do gráfico da Figura 5 -, os inscritos podem ser reincidentes de aplicações anteriores do Exame, o que faria descumprir tal requisito.

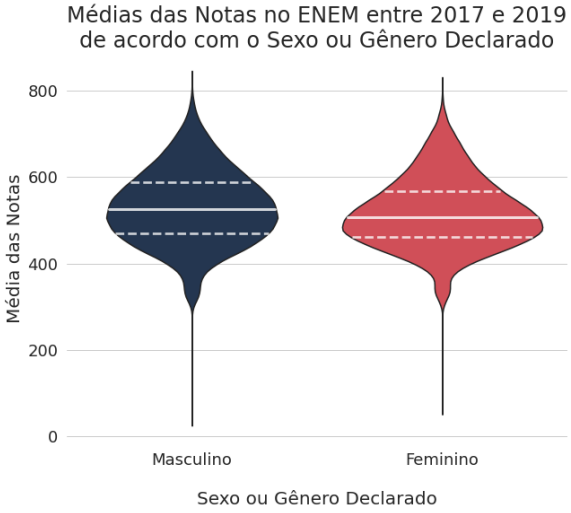


Figura 10: Média das Notas dos Inscritos do ENEM entre 2017 e 2019 de acordo com o Sexo ou Gênero Declarado. Apesar de parecer, visualmente, que as linhas dos quartis são semelhantes, o resultados do teste estatístico de Mann-Whitney sinalizou que as médias (ou medianas) são estatisticamente distintas.

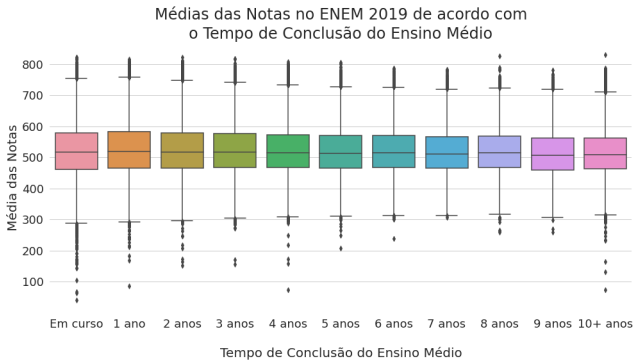
Tabela 2: Dados Estatísticos para TP_SEXO

Sexo	Média	Desvio Padrão	Assimetria
Masculino	530,2	86,0	0,1538
Feminino	517,1	81,39	0,3596

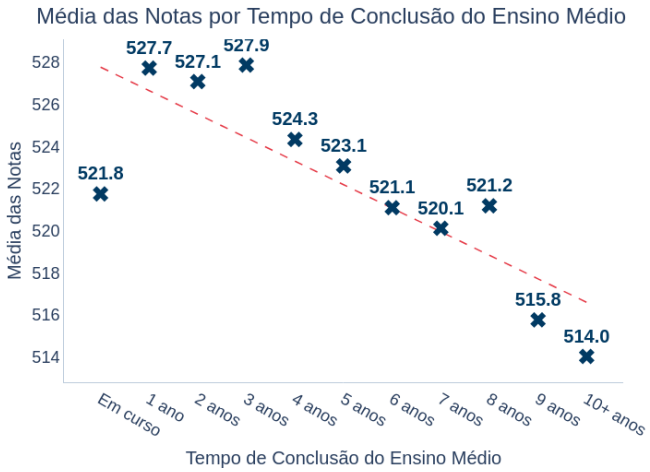
Portanto, a fim de garantir a independência dos elementos, aplicou-se os testes de hipótese apenas sobre os inscritos do ENEM 2019 sob uma amostragem estratificada com um tamanho de amostra significativa equivalente a 10% da população dos inscritos presentes, mantendo, assim, uma boa representação da população e redução de chances de erros.

Ao se ter a variável TP_SEXO sob análise, por exemplo, extraiu-se alguns dados e características estatísticas de uma amostra de tamanho 509.214 elementos garantidamente independentes. Dentre essas características estão as médias de notas, os desvios padrão em torno das médias e sua assimetria em relação à curva normal (*skewness*) de cada categoria pertencente à variável. Esta assimetria se trata de um valor que varia entre -1 e 1, sendo que quanto mais perto de zero mais simétrico ou semelhante a uma curva normal, enquanto que quanto mais próximo de +1 ou -1, mais assimétrica é a amostra para a direita ou para a esquerda, respectivamente, em relação a uma amostra proveniente de uma distribuição normal.

Vê-se, por meio da Figura 10 e da Tabela 2, que possuem diferenças tanto nas médias de notas dos dois grupos quanto em suas características, tendo o sexo masculino uma média ligeiramente maior, um maior desvio em torno da média e uma distribuição mais simétrica que o sexo feminino. Observa-se, também, que ambas amostras possuem uma assimetria positiva (para a direita) e que podem ser um sinal de que não são provenientes de distribuições normais.



(a) Distribuição das Notas: tempo decorrido desde a conclusão do Ensino Médio.



(b) Média das Notas: tempo decorrido desde a conclusão do Ensino Médio.

Figura 11: Análise das Notas dos Inscritos do ENEM 2019 de acordo com o Tempo de Conclusão do Ensino Médio. Apesar de aparentarem semelhantes, o Teste de Kruskal-Wallis apontou diferença significativa entre as medianas de, pelo menos, dois grupos categóricos.

Como se pode observar pela Figura 11a, novamente as amostras, após apresentaram distribuições não-normais

e variâncias não equivalentes, obtiveram como resultado medianas distintas por meio do Teste de Kruskal-Wallis, mesmo com as médias variando num intervalo pequeno, o resultado foi que ao menos duas categorias não apresentam medianas equivalentes, algo comum ao analisar diversos grupos como neste caso.

Já com relação às médias ao tempo de conclusão do Ensino Médio, as quais são representadas pela Figura 11b, vê-se que, apesar de se concentrarem as médias de todas as categorias em um intervalo muito curto em relação às outras variáveis analisadas, as médias das notas seguem uma regressão linear que descende à medida que o tempo de conclusão do Ensino Médio aumenta. Regressão essa, aliás, que fica mais clara e tende a diminuir o erro quanto mais elementos a amostra possuir.

Vale destacar que a média dos inscritos que não concluíram o Ensino Médio foge à curva para um valor menor, um comportamento recorrente ao longo de todo o conjunto de dados, talvez, por conter treineiros e levando em consideração, também, o fator nervosismo ou pressão muito comum a quem está terminando o Terceiro Ano do Ensino Médio e quer ingressar no Ensino Superior. Visto que muitos, ao não passarem no Exame assim que concluem o Ensino Médio, fazem cursos preparatórios focados exclusivamente para o ENEM, vê-se que os inscritos que concluíram o Ensino Médio entre 1 e 3 anos possuem as melhores notas. Por variarem em um intervalo relativamente curto, infere-se que, para um algoritmo de Aprendizado de Máquina, esta variável possivelmente terá um peso menor que outras variáveis.

Da mesma forma, obteve-se as notas médias, os desvios padrão e a assimetria da amostra estratificada em relação à variável TP_COR_RACA que descreve a cor ou raça à qual os inscritos autodeclaram pertencer.

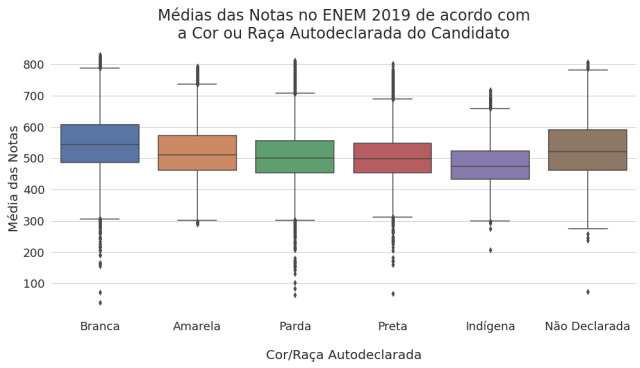


Figura 12: Distribuição das Notas dos Inscritos Presentes no ENEM 2019 de acordo com a Cor ou Raça Autodeclarada. Visualmente já se nota que há uma diferença entre as médias dos grupos e o Teste de Kruskal-Wallis confirmou a significância estatística da diferença entre, pelo menos, dois grupos analisados.

Após obter, como resultado dos testes, rejeições das hipóteses de distribuição normal e de equivalência de variâncias, foi possível observar numericamente na Tabela 3 e visualmente na Figura 12 que há, pelo menos, dois grupos

Tabela 3: Dados Estatísticos para TP_COR_RACA

Cor/Raça	Média	Desvio Padrão	Assimetria
Branca	548,2	85,61	0,1091
Amarela	521,0	84,02	0,4353
Parda	507,2	78,52	0,3256
Preta	502,6	73,77	0,2485
Indígena	479,4	72,92	0,2613
Não Declarada	528,7	91,59	0,1789

com medianas estatisticamente distintas através do Teste de Kruskal-Wallis.

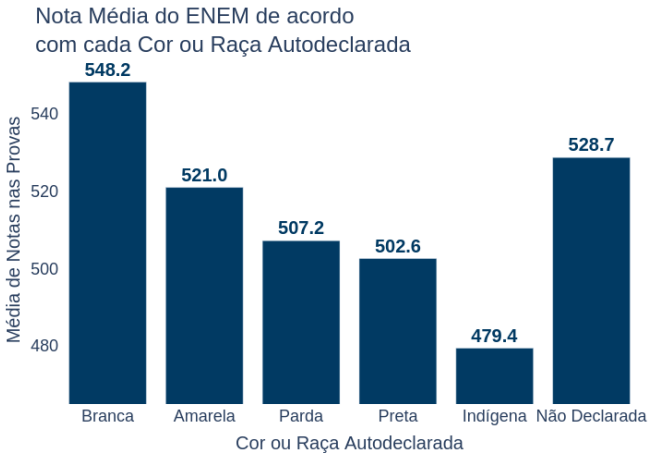


Figura 13: Média das Notas dos Inscritos do ENEM 2019 de acordo com a Cor ou Raça Autodeclarada.

A partir da Tabela 3, nota-se uma diferença mais brusca da média das notas entre a cor Branca e Amarela, as quais representam as duas maiores médias entre das cores ou raças autodeclaradas, se acentuando ainda mais quando comparadas com a categoria Indígena que possui a menor média de notas. Visualmente, percebe-se uma diferença entre as médias, com desvios padrão parecidos e todas assimétricas, mesmo aqueles de cor Branca que apresentaram a menor assimetria.

Entrando nos dados fornecidos através do formulário socioeconômico, pode-se medir o impacto das características dessa natureza sobre as médias das notas dos inscritos no Exame.

Como exemplo, a Figura 15 mostra a relação entre as médias das notas dos inscritos e o nível de escolaridade do pai e da mãe ou responsáveis. Ela mostra que o quanto a educação proveniente dos familiares pode influenciar de alguma forma no desempenho das próximas gerações. Além da evidente relação que demonstra que quanto maior o nível de escolaridade dos pais maior a média das notas, também revela que, se os pais tiverem feito um Ensino Superior ou Pós-graduação, tem um salto de distância mais determinante em comparação com aqueles cujos pais concluíram os estudos apenas no Ensino Médio.

Da mesma forma, o gráfico da Figura 17 revela uma forte tendência de aumentar a nota quanto maior for o

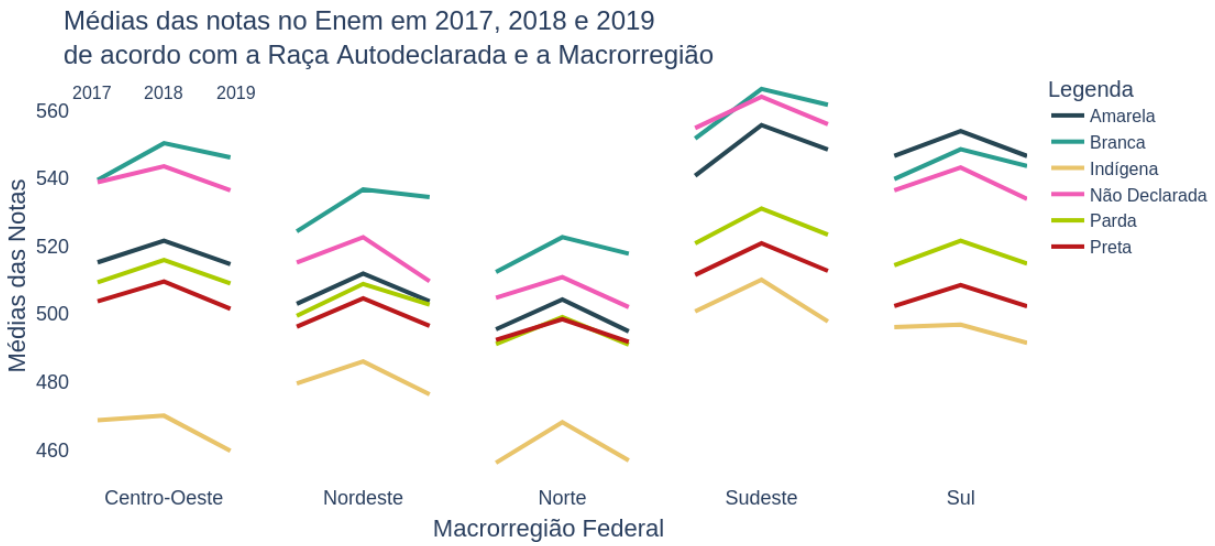


Figura 14: Média das Notas dos Inscritos do ENEM 2019 de acordo com a Cor ou Raça Autodeclarada e Macrorregião Federal.

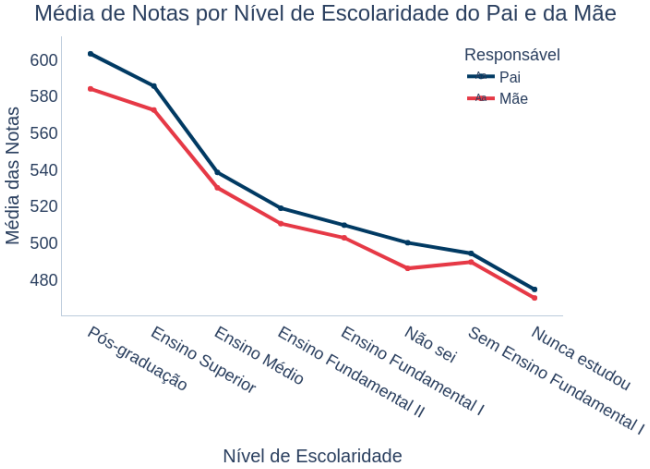


Figura 15: Média das Notas dos Inscritos do ENEM 2019 de acordo com o Nível de Escolaridade do Pai, da Mãe ou Responsável.

poder aquisitivo da família, com destaque para a região que corresponde àqueles com renda familiar mensal de até 1 salário mínimo até aqueles entre 6 e 7 salários mínimos, pois é a região onde se nota uma maior variação das médias em decorrência da renda familiar.

Percebe-se ainda, um salto significativo entre aqueles com renda de até 1 salário mínimo para aqueles que recebem até 2 salários mínimos, sendo a maior diferença observada no gráfico da Figura 17 ao passo que, a partir de uma renda familiar maior que 7 salários mínimos, a taxa passa a ter menor sensibilidade ao aumento da renda.

Toda essa disparidade e desigualdade social também pode ser refletida no gráfico da Figura 18, no qual é possível ver representada uma diferença significativa entre as médias das notas dos inscritos que dispões de acesso à *Internet* em casa e aqueles que não.

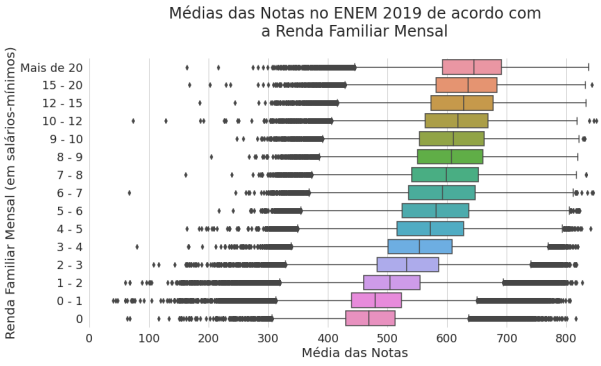


Figura 16: Distribuição das Média das Notas dos Inscritos do ENEM 2019 de acordo com a Renda Familiar Mensal.

4.3. Índice de Desistência

Outro ponto importante a se analisar sobre as características apresentadas é saber a razão de desistência dos inscritos, visto que, além do prejuízo ao próprio inscrito, aqueles que não comparecem ao Exame também têm suas provas impressas. Logo, é interessante ao Governo, tanto no âmbito do desenvolvimento da Educação quanto no setor da Economia, minimizar o índice de desistência.

No mapa da Figura 19 é possível observar os índices de desistência associados a cada Unidade da Federação. Nele, é possível notar, a princípio, um índice muito elevado de desistência nos estados da região Norte, mais acentuadamente no Amazonas, onde há uma média de 40% de desistência entre os inscritos.

No gráfico da Figura 20, demonstra o nível de desistência agrupado por Macrorregião do Brasil e é possível inferir que a região Norte continua com o índice de desistência muito elevado, porém acompanhado mais de perto do Centro-Oeste com índices de 32,85% e 32,66%,

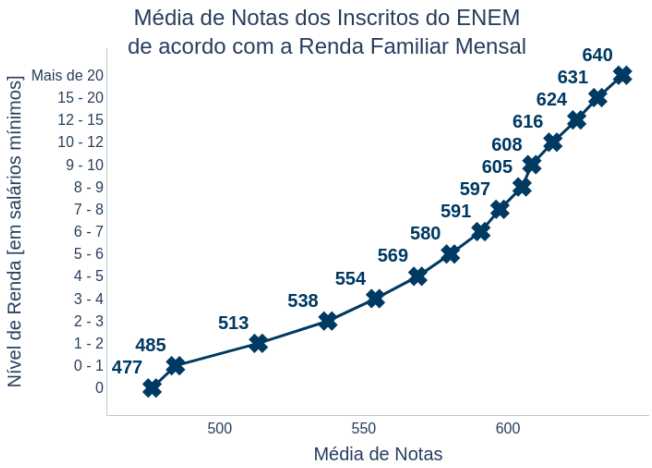


Figura 17: Média das Notas dos Inscritos do ENEM 2019 de acordo com a Renda Familiar Mensal.

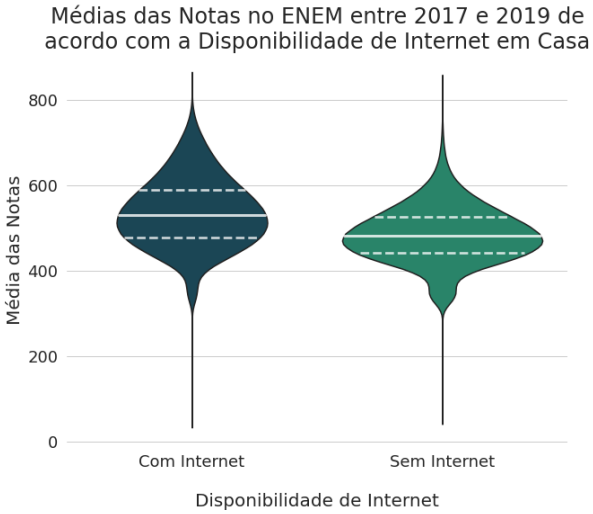


Figura 18: Média das Notas dos Inscritos do ENEM entre 2017 e 2019 de acordo com a Disponibilidade de Internet em Casa.

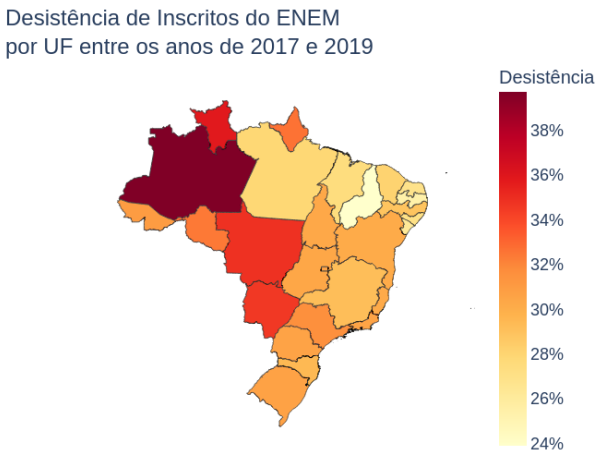


Figura 19: Índice de Desistência no ENEM entre 2017 e 2019 por Estado brasileiro.

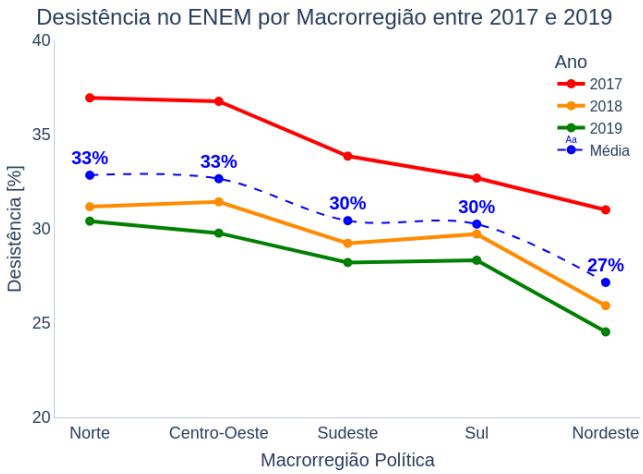


Figura 20: Índice de Desistência no ENEM entre 2017 e 2019 por Região.

respectivamente. Importante destacar, por outro lado, que o índice de desistência no Nordeste é consistentemente abaixo da média nacional com uma razão média de 27,16% de desistência em relação ao número de inscritos, do qual vale ressaltar o estado do Piauí com o menor índice nacional por volta de 24% de inscritos que não compareceram a algum dia do Exame.

Por outro lado, na Figura 21, é possível ver uma relação não-linear com uma forma semelhante a uma curva logarítmica entre o índice de desistência e o tempo que os inscritos declaram ter concluído o Ensino Médio.

Isso torna essa característica possivelmente importante de se levar em conta quando o objetivo for determinar a probabilidade de desistência do candidato, já que o índice de desistência aumenta à medida que o inscrito possui mais tempo que concluiu o Ensino Médio seguindo um padrão, tendendo a estabilizar em torno de 45% de desistência para aqueles que concluíram o Ensino Médio há mais que 10 anos.

Já com relação à cor ou raça autodeclarada dos indivíduos, como é possível observar (na Figura 22), há um índice de desistência mais acentuado entre os inscritos que se autodeclaram indígenas seguido daqueles que se autodeclaram da cor preta, com 32,50% e 30%, respectivamente. Apresenta-se com o menor índice de desistência aqueles que se autodeclaram brancos, com 25,1% de desistência.

O nível de escolaridade dos pais, no entanto, apresenta uma relação mais linear, já que a variável independente é categórica mas ordinal - é possível ordenar as categorias - e o índice de desistência aumenta à medida que o nível de escolaridade dos pais "abaixa".

De acordo com a Figura 23, nota-se, também, que o impacto do nível de escolaridade do pai ou homem responsável e da mãe ou mulher responsável no índice de desistência do inscrito são idênticos ao longo de toda a curva. A parte mais saliente dessa curva, ainda que muito mais tímida que

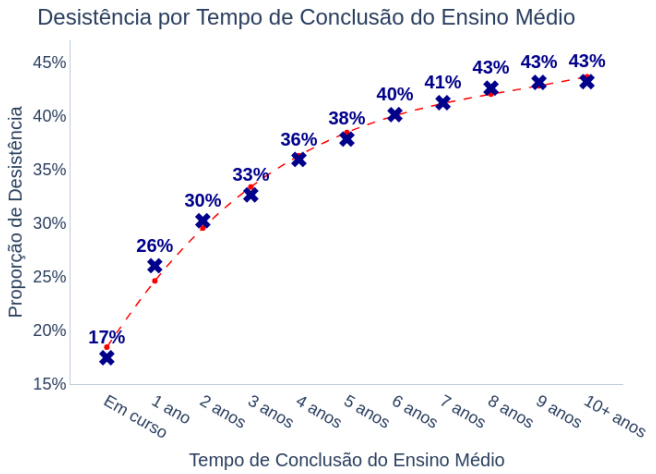


Figura 21: Índice de Desistência dos Inscritos do ENEM 2019 de acordo com o Tempo de Conclusão do Ensino Médio.

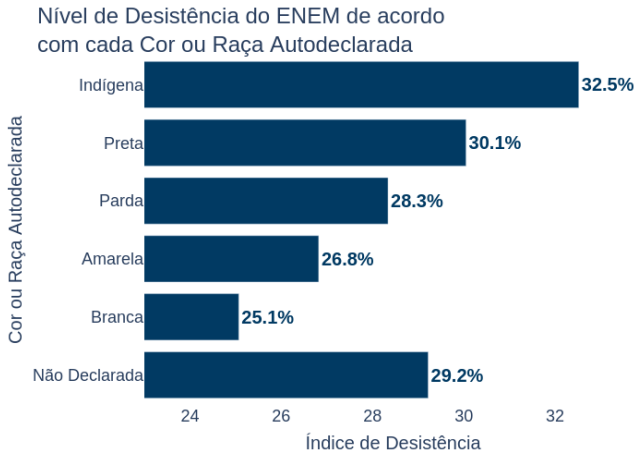


Figura 22: Desistência por Cor ou Raça dos Inscritos do ENEM 2019.

na Figura 15, é, novamente, da classe de Ensino Superior para a classe de Ensino Médio, ressaltando que o nível de escolaridade dos pais exercem um papel significante no desempenho dos inscritos no ENEM, principalmente se tiver concluído um Ensino Superior ou, ainda, uma Pós-graduação.

Outro fator presente nos dados do formulário socioeconômico é a distribuição da renda dos inscritos e, como mostrado no gráfico da Figura 24, é possível extrair, também, o índice de desistência dos candidatos pertencentes a cada categoria. Percebe-se, portanto, que, para aqueles com renda familiar mensal de até 2 salários mínimos - a maior parte dos candidatos -, o índice de desistência é superior aos 30% e, para aqueles cuja renda familiar mensal ultrapassa os 5 salários mínimos, a desistência permanece assintoticamente mais baixa que 20%, mesmo para a categoria de inscritos com uma renda maior que 20 salários mínimos que foge à curva.

Consequentemente, isso acarreta em um maior índice

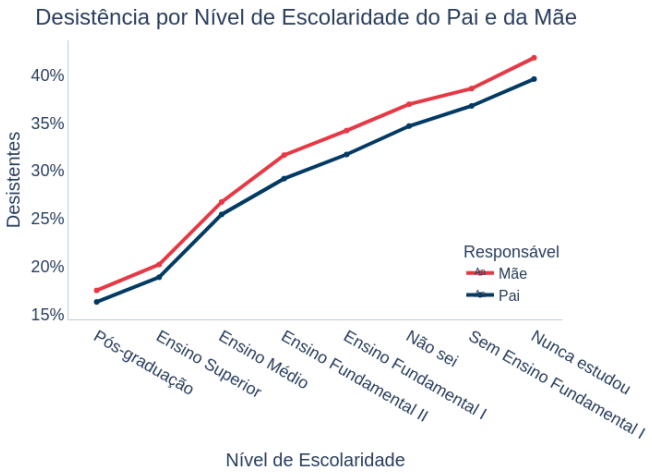


Figura 23: Índice de Desistência dos Inscritos do ENEM 2019 de acordo com o Nível de Escolaridade do Pai, da Mãe ou Responsável.

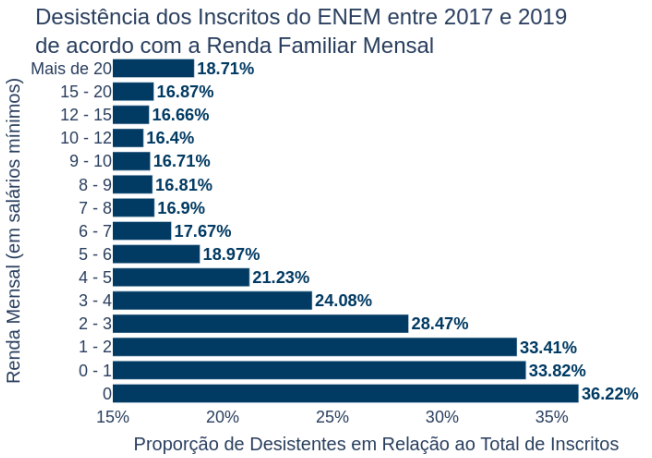


Figura 24: Índice de Desistência dos Inscritos do ENEM entre 2017 e 2019 de acordo com a Renda Familiar Mensal.

de desistência, também, para aqueles que declararam não ter *Internet* disponível em casa.

Tabela 4: Proporção de Desistência de acordo com a Disponibilidade de Internet em Casa

Disponibilidade	Presentes	Ausentes	Desistência
Com Internet	8.811.903	3.473.246	28.27%
Sem Internet	2.941.652	1.624.081	35.57%

A Tabela 4 demonstra a diferença entre os índices de desistência de quem possui e quem não possui acesso à *Internet*. Nela, vê-se o salto entre as categorias e é possível inferir o quão importante tem se tornado a conectividade no âmbito da Educação, oferecendo acesso a informações que podem ser determinantes para os candidatos terem um bom desempenho no ENEM.

5. Considerações Finais e Trabalhos Futuros

Com as discussões promovidas neste trabalho, fundamentadas nos gráficos a partir dos dados oficiais fornecidos pelo Inep, tem-se um panorama mais claro acerca do perfil dos candidatos e a média dos seus resultados no ENEM, assim como a probabilidade de se ausentar do Exame.

É perceptível que há uma certa relação entre alguns fatores geográficos e socioeconômicos dos participantes e os resultados obtidos, assim como com a probabilidade dos mesmos não comparecerem ao ENEM. Dentre tais fatores, é evidente a problemática referente ao nível acesso do participante a um sistema de educação decente e, também, ao ambiente domiciliar no qual se está inserido, observável através de características como a renda familiar, o nível de escolaridade dos pais ou responsáveis e a disponibilidade de *Internet* em casa, para citar alguns desses fatores.

De acordo com o analisado, aproximadamente dois terços dos participantes possuem renda familiar mensal de até dois salários-mínimos, faixa esta cujas médias das notas se encontram bem abaixo das demais e cujos níveis de desistência são destacadamente os maiores, sendo todos acima de 30%. Notou-se um degrau de desempenho íngreme já em relação àqueles com uma renda familiar entre dois e três salários-mínimos e evidenciou-se um vão se comparados com aqueles com rendas muito superiores.

Concomitantemente, o nível de escolaridade dos pais ou responsáveis pode estar correlacionado com esse problema. Observou-se uma relação muito clara deste fator com o desempenho dos candidatos, destacando um desempenho médio muito mais elevado caso os pais possuam, pelo menos, o Ensino Médio completo, e uma queda de rendimento daqueles cujos pais não estudaram. O nível de desistência se deu praticamente de forma linear chegando a mais de 40% para estes cujos pais não estudaram. Dentro os que marcaram a opção "Não sei", a suposta ausência da mãe ou mulher responsável causa mais impacto se comparado à ausência paterna, possivelmente por ser um cenário - infelizmente - mais comum.

Contudo, o fator mais impactante observado foi a proporção da indisponibilidade de *Internet* em casa, principalmente na região Norte e, mais especificamente, nos Estados do Pará e Amapá. A diferença de desempenho para aqueles com acesso à *Internet* é enorme, o que, claramente, resulta numa desvantagem competitiva e se enquadra em um cenário de acesso precário a uma educação decente e põe em xeque prováveis efeitos catastróficos após a pandemia de Covid-19 tendo em vista o modelo implementado no período e, portanto, é imprescindível um acompanhamento e apoio mais reforçados pós-pandemia.

O tempo de conclusão do Ensino Médio também foi uma característica observada e se relaciona de forma clara com o desempenho e desistência do candidato. À medida que o tempo de conclusão do Ensino Médio aumenta, o desempenho reduz, ao contrário do nível de desistência que aumenta até estabilizar em torno de 43%. Todavia, notou-se uma forte presença daqueles com mais de 10 anos que concluíram o Ensino Médio, indicando que o ENEM pode ser uma porta para um retorno aos estudos ou para novas oportunidades profissionais ou de ingresso ao Ensino Superior.

Além destes fatores, alguns outros como região geográfica, cor ou raça autodeclarada e gênero foram analisados. Com base em estatística, a região Norte apresenta o menor percentual de inscritos e, ainda assim, o maior índice de desistência, principalmente o Estado do Amazonas. Por conseguinte, e por razões históricas, nota-se uma menor participação dos povos indígenas, refletindo o acesso à educação que lhes é, em geral, insuficientemente concedido.

Felizmente, apesar de possuir uma diferença estatisticamente relevante, o desempenho e o índice de desistência entre homens e mulheres são idênticos a olho nu, o que demonstra uma menor distinção e maior igualdade de capacidades entre os sexos dos participantes, tornando a competição, a princípio, um pouco mais justa.

Importante ressaltar que correlação não implica em causalidade. No entanto, no contexto deste trabalho independente, as relações observadas podem auxiliar em um eventual mapeamento de consequências de um desnível que vai além da chamada meritocracia. Isto porque, apesar da significante ausência de dados a respeito das escolas, oferece discussões pormenorizadas para cada característica apresentada como um resultado da aplicação prática de ferramentas de Ciência de Dados e *Big Data*, áreas crescentes principalmente em Tecnologia, voltadas para um contexto de relevância social.

Por fim, como sugestão de trabalhos futuros, recomenda-se a análise, e comparação com os anos anteriores, dos microdados provenientes do ENEM 2020, o qual foi realizado em meio à pandemia e em condições conflituosas. É conveniente citar, também, uma possível implementação de ferramentas mais sofisticadas e, talvez, menos passíveis de pré-julgamento humano, como, por exemplo, ferramentas de *Deep Learning* e *Cloud Computing*, obtendo-se visões mais sistemáticas acerca do problema.

Agradecimentos

O presente trabalho foi realizado com apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico e da Universidade Federal de São João del-Rei, por meio do programa PIBIC/CNPq - Edital 003/2020/PROPE.

Referências

[1] F. L. D. Silveira, M. C. B. Barbosa, and R. D. Silva, "Exame nacional do ensino médio (enem): Uma análise crítica," *Revista Brasileira de Ensino de Física*, vol. 37, p. 1101, Mar 2015.

[2] M. U. Kleinke, "Influência do status socioeconômico no desempenho dos estudantes nos itens de física do enem 2012," *Revista Brasileira de Ensino de Física*, vol. 39, p. 1–19, Oct 2016.

[3] W. M. d. Santana, *Mineração em dados do ENEM para a predição do desempenho acadêmico no âmbito da rede federal de educação tecnológica*. PhD thesis, UFPE, 2018.

[4] M. M. Nascimento, *O acesso ao ensino superior público brasileiro: um estudo quantitativo a partir dos microdados do Exame Nacional do Ensino Médio*. PhD thesis, Universidade Federal do Rio Grande do Sul, 2019.

[5] "Maior índice de participação em 10 anos; custo com processo foi menor do que em 2018," Jan 2020.
[6] W. van der Aalst, *Data Science in Action*, pp. 3–23. Berlin, Heidelberg: Springer Berlin Heidelberg, 2016.
[7] S. Rahman, "Elucidation and dominance of hypothesis analogies in data science," *International Journal for Research in Applied Science and Engineering Technology*, vol. 8, p. 539–548, Sep 2020.
[8] M. Schermann, H. Hemsen, C. Buchmüller, T. Bitter, H. Krcmar, V. Markl, and T. Hoeren, "Big data," *Business & Information Systems Engineering*, vol. 6, p. 261–266, Sep 2014.
[9] C. Taurion, *Big Data*. BRASPORT, 1 ed., 2013.